Statistical modelling in biology : with reference to the sheep and wool industries and medicine

Author:
Tallis, G. M. (George Michael)

Publication Date:
1976

DOI:
https://doi.org/10.26190/unsworks/6468

STATISTICAL MODELLING IN BIOLOGY

WITH REFERENCE TO

THE SHEEP AND WOOL INDUSTRIES

AND MEDICINE

I GENERAL PRELIMINARIES

Introduction

The work submitted here was completed over a period of fifteen years. Some of the research was done while I was employed by C.S.I.R.O., at first in the Division of Animal Genetics and subsequently in the Division of Mathematical Statistics.

Early in my career my background as a Geneticist and general Agriculturalist made me acutely aware of the need to apply sensible mathematics and statistics to a wide variety of biological problems. In those days these ideas were somewhat inovative, at least in Australia. However, with the recent escalation and popularisation of statistical education, more competent statisticians are turning to the biologist with a view to helping him make order from chaos.

Because of my various changes of employment, I have not devoted all my time to any one field. This has been advantageous as each new biological system has required its own special statistical approach. In turn, this has kept my interest alive.

Nevertheless, the work falls comfortably into two general categories:

(A) research associated with or arising from problems in the Sheep and Wool Industries;

(B) research on medical problems arising from consultations at the Cancer Institute, Melbourne.

Almost every result has been practically motivated since the pressure of working closely with biologists usually keeps the statistician on the rails. There is little time for models which cannot be readily harnessed for use, no matter how beautiful the mathematics. I have grown to realise over the years that applicability really is the essence of the contract.

Sometimes mathematical results had to be established to back up the modelling work. When this occurred, there was a satisfying interplay between motivated theoretical research and application.

General Format

The papers have been gathered under headings. Under these a brief description of the type of work reported in each paper is given. These comments are to assist in co-ordinating the material and the system used is self explanatory.

Fuller descriptions on the results can be obtained from the summaries which precede each paper.

II CLASSIFICATION OF THE WORK

(A) Research into Some Specific Problems of the Sheep and Wool Industries.

Research work completed under this heading falls into four main sub-headings:

(a) Research associated with the genetics of sheep breeding.

(i) Mass selection theory, papers [1] - [6].

(ii) Truncation procedures related to mass selection, papers [1] - [4].

(iii) The sampling errors of estimators of certain genetic parameters and predictors of genetic gain, papers [1] - [6].

(iv) Special models for discrete character selection; estimation under grouping and truncation, papers [1] - [5].

(b) The modelling of host-parasite cycles with an emphasis on nematode parasites in sheep, papers [1] - [7].

(c) Research into fertility and meat production in sheep, papers [1] - [4].

(d) Theoretical developments pertinent to A(b) and A(c), papers [1] - [10].

(B) Research into Some Specific Problems in Medicine with Special Reference to Breast Cancer.

Subsections are as follows:

(a) General comments.

(b) The initial problem and preliminary work, papers [1] - [6].

(c) The concept of continuous response, papers [7], [8].

(d) Actuarial work, papers [9], [10].

(e) Time to reporting breast cancer, papers [11] - [13].

(f) The concept of cure and screening trials, papers [9], [11], [12], [13].

(g) Summary and conclusions.

III COMMENTS ON PAPERS

(A) Research into Some Specific Problems of the Sheep and Wool Industries.

General Comment

A large proportion of the material in this section is concerned with matters of statistical genetics, with selection as the central theme. Later papers propose models for host parasite cycles in sheep with an emphasis on parasitological and immunological concepts. A lesser amount of work on ewe fertility and mutton production is also included.

Certain theoretical back-up results are needed to complement the above work. A great deal of this material has been utilized in the applications which motivated it.

(a) Research associated with the genetics of sheep breeding.

(i) Mass selection theory.

Paper (a)(i)[1] introduces the idea of selecting animals for life time performance by basing the selection index on early records of a number of characters. The methods used are a straightforward application of standard theory; it is the concept which is new.

[2] proposes a special form of selection index which will allow breeders to select for m characters in such a way that they make general genetic progress while moving the means of a subset of k characters to predetermined levels. This is an important exercise in some animal production situations.

[3] and [4] look at the effects of selection and migration on a set of k alleles. The treatment is different in that discrete problems are examined, whereas the other selection work involves continuous character theory.

[5] extends the current ideas of selection to the selection of growth curves. Stochastic processes with a continuous index set are introduced and analysed in a genetic context. Classical discrete procedures give way to continuous methods and the whole problem is discussed within a more realistic framework.

[6] examines in a general setting an extension of the problem of [2]. It is shown that it is immaterial whether one finds the optimal solution by maximising the correlation between the index and the weighted economic genotype or minimises the usual mean square error expression. This holds whether or not the index is subject to the usual form of linear constraint.

Note

The optimality of the selection index theory is generally accepted. Cochran (1950) (reference given in A(ii)(d) showed that this form of selection maximises expected genetic gains, and that is essentially that.

(ii) Truncation procedures related to mass selection.

Paper (a)(ii)[1] discusses practical methods of applying truncated selection in situations where it is uneconomical to permanently identify each member of the group. The actual point of truncation may be unknown until all individuals are measured. In these cases it may be important to have a procedure which allows selection to proceed without undue inefficiency. A non-parametric solution to this problem is proposed.

[2] develops general formulae for the moment generating function and the first and second order moments under truncated selection on a multi-normal distribution. The results are applied in a genetic context to show how this type of truncation affects certain genetic parameters.

[3] extends the work of [2] by considering the effects of index type selection on the moments of a multi-normal distribution. These results have immediate application to statistical genetics.

[4] investigates a different type of truncation to the usual rectangular system. It is found that by truncating on the "contours" of the multi-normal distribution simple formulae for the moments are obtained. This form of selection has useful application and it is used in combination with a radial form of selection to solve particular problems in the construction of experimental selection groups and control groups.

(iii) The sampling errors of estimators of certain genetic parameters and predictors of genetic gain.

Paper (a)(iii)[1] investigates optimum statistical designs for estimating intra-class correlation, ρ, from one-way analysis of variance tables. The asymptotic formula for the variance of $\hat{\rho}$ is manipulated to give suitable subclass numbers to minimise the variance.

[2] looks at the effect of certain controllable errors on biases in hesitability estimates and the efficiency with which genetic correlation coefficients are estimated. Some surprisingly large deleterious effects on genetic parameters estimation were observed due to delays in weighing newly born lambs.

[3], [4] and [5] develop large sample approximations to the sampling errors inherent in estimates of genetic correlation, selection index coefficients, genetic gain and family selection procedures. Standard δ methods were used to obtain the results, although the algebra became strenuous. Monte Carlo simulations have shown these approximations can be used with confidence.

[6] The results in [3] [4] and [5] relied on formulae for the second moments of estimates of covariance components based on balanced designs. In this paper general formulae for the unbalanced case are developed. The use of these are demonstrated on two types of selection indexes and the results correlated with those of [3].

Note

The above work seems to resolve the problem of establishing the order of errors associated with the estimation of functions of phenotypic and genetic parameters. At least the manipulative procedures are established so that other cases should follow easily.

(iv) Special models for discrete character selection; estimation under grouping and truncation.

Papers A(a)(iv)[1] and [2] The genetic analysis of discrete data runs into problems which can be partially overcome by the use of suitable discrete models. This obviates the necessity of applying continuous variate type analysis to such data. These two papers propose models specifically constructed to assist with the estimation of correlation structure in discrete data.

[3], [4] propose methods for estimating parameters of the log-normal and bivariate normal distributions under grouping. In [3] the methods are used to estimate correlation between lambing performance of ewes over two consecutive years.

[5] General methods are developed here to allow maximum likelihood estimates to be obtained under grouping. Provided the grouping intervals are small, these results lead to accurate estimates and a great saving of labour.

(b) The modelling of host-parasite cycles with an emphasis on nematode parasites in sheep.

Paper A(b)[1] develops the first tentative models to describe the distribution on pasture of the larvae of Nematode parasites of sheep. Basic biological postulates were set up and the models were tailored to these. Later the assumptions were tested explicitly in the field and as a result the models were modified in [5].

[2] extends some of the ideas of [1] to include a migration model for the parasites movements away from faecal deposits on the pasture. This model is of the diffusion type.

[3] sets up general models for studying host-parasite relationships. Immunological mechanisms are considered and the results specialised to sheep-worm relationships.

[4] flows from the results of [3] and develops special models for Helmenthic parasites. The work is oriented towards humans but would apply equally to sheep Helmenthiasis.

[5] revises and strengthens some of the results of [1] to conform with experience gained. More detailed modelling of larvae on pasture results in more information as to their habits once on the ground.

[6], [7] These two papers concern themselves with deriving full models for the whole life cycle of internal parasites of sheep. [6] concentrates on the deterministic theory while [7] suggests stochastic extensions.

Note

The prime aim of the projects reported above was to produce a workable model for the full life cycle of a nematode type parasite in sheep. Paper [6] contains the necessary pieces, and the steady state behaviour of the system was investigated closely here and in another paper A(d)[7]. It is unfortunate that, due to difficulties of obtaining enough data in the correct form and due to staff movements, the entire model has not yet been tuned into its practical setting.

(c) Research into fertility and meat production in sheep.

Papers A(c)[1], [2] and [3] report the results of some extensive studies involving the artificial insemination of sheep. Both problems of efficiency of the procedure and the influences of various factors on matters of fertility have been examined.

[4] In an effort to determine important characters to use in the selection of mutton production, a slaughter trial was performed on 150 merino sheep. The main result of the study was that, due to the high correlation of body weight with the weight of edible meat, other measurements were redundant.

(d) Theoretical developments pertinent to A(b) and A(c)

Paper A(d)[1] discusses general methods to assist with an interpretative analysis of some classes of contingency tables. These results were required urgently for the work in A(c)[1], [2]. There appear to be a number of situations where a full logit model analysis is not just unwarranted, but inappropriate.

[2], [3], [4] In the work of A(b) there were many "mixing" operations performed. Mixing invariably raises the problem of identifiability and this is discussed in [3], while [4] puts the problem in a general setting and the results of [3] appear as special cases. In [2] the estimation of a particular type of mixture is accomplished by the use of fractional moments in an effort to reduce the sampling errors associated with the use of integer moments.

[5] Some work on aspects of the chemistry of wool production raised the question of how to construct stochastic models for r-molecular reactions. This problem is dealt with in this paper where general deterministic solutions are developed and two stochastic approaches investigated.

[6] During the study of wool growth, the distribution of wax glands in the skin became an important characteristic. Theoretical consideration of the problems raised by investigators in this area lead to the old corpuscle problem and its extensions. These are discussed in this paper.

[7] The application of the models of A(b) requires accurate estimates of egg counts in faecal deposits. The whole estimation procedure is examined here and controlled sampling schemes suggested.

[8] Discussions of the stability of the models in A(b)[6] hinge on the stability or otherwise of a certain polynomial equation $f(x) = a$. This mathematical problem is discussed in detail in this paper.

[9] A problem raised in [2] was "under what conditions is a moment estimator equal to a maximum likelihood estimator". The question was answered in part in [2] and the complete solution is presented here. The simple result is that the two types of estimators are equal if and only if the parent density belongs to the exponential family and [9] is included here for completeness.

[10] That the exponential family is closely related to the existence of sufficient statistics is well known. This relationship is reexamined here, following [9] to give a simple result relatively free from restrictive assumptions.

(B) Research into Some Specific Problems in Medicine with Special Reference to Breast Cancer.

(a) General Comments

In this section the formal method of presentation and description will be dropped in favour of a chronological development of the total problem. This is feasible in this case since there has been logical progression through a number of phases and I was brought into the work from the outset. What started as a rather limited clinical trial, escalated to a project of substantial proportions.

The various research papers are mentioned in the report at the pertinent places. Little mention of applicability of the results outside the specific area of use is made. However, most of the ideas have relevance to other fields of Medicine and Biology. Almost all are pertinent to other forms of cancer disease.

(b) The initial problem and preliminary work.

In 1967 I was approached by Dr. G. Sarfaty to assist him in the interpretation of some English work on breast cancer. The main papers were those of Bulbrook and his co-authors (Ref. see B[1]), who were applying discriminant analysis techniques to help in the selection of women with advanced breast cancer for endocrine ablation. The predictor variables used in the discriminant were the levels of certain hormone related compounds excreted in the urine of the patients.

From an examination of this work two points emerged:

(1) the hormone levels were age dependent and this effect needed attention in any application of a discriminant function;

(2) there were errors of estimation and prediction associated with the use of discriminant procedures and these required investigation.

A clinical trial was proposed at that time which was along the lines of the Bulbrook experiments. This was to be run at the Peter McCallum clinic and the main aim was to tie in the relevant hormone picture with the response to ablation, in particular adrenalectomy. From a detailed examination of data generously made available by Bulbrook, it became clear that the trial stood the best chance of showing up useful predictor variables for remission if the chemical sampling and assay work were as tight as practicable.

Early efforts were made to eliminate errors inherent in urine sampling procedures. Diurnal as well as daily fluctuations in the levels of the excreted compounds made it desirable to set up a five day collection system, [1]. The aim at each stage was to reduce those parts of the variances of proposed predictor variables which were due to noise.

Part of the problem of tidying up laboratory procedures was the calibration of numerous pieces of equipment. A satisfactory philosophy for the purpose at hand was needed and is described in [2]. This approach to calibration has been used subsequently.

From the start there was dissatisfaction with the discriminant method since the allocation of candidates for ablation to either a higher than average or a lower than average remission group did not seem satisfactory. What was needed, it seemed, was a specific probability for remission for each patient.

In any case a pressing theoretical problem was to investigate (1) and (2) of the second paragraph. This was done in [3] where the delta method was used to estimate the order of the various errors. Some extensions to personal probabilities and a program implementing [3], called DISCRIM, was given in [4]. These methods were eventually applied to early results of the Peter McCallum trial in [5]. The trial itself is discussed in [6] .

The upshot of this work was that problems (1) and (2) had been dealth with and personal probabilities of remitting as a consequence of adrenalectomy could be predicted on the basis of certain measurements, $\underset{\sim}{x}$, made on patients before the operations. Let $P(\underset{\sim}{x})$ be this probability.

The effect of sampling errors on predictions was to lower $P(\underset{\sim}{x})$ towards the group average probability of remission, .32, uniformly in $\underset{\sim}{x}$. This is compatible with intuition since, if noise is introduced into a predictive system depending on $\underset{\sim}{x}$, say, the value of the prediction must diminish to the group average situation where $\underset{\sim}{x}$ is not known.

There were disadvantages with the system which has been developed:

(1) it was enormously complicated computationally and introduced a great strain on the data in terms of estimation;

(2) the theory was heavily dependent on the assumption that variables $\underset{\sim}{x}$ were normally distributed;

(3) the model was purely predictive and it was of no interpretative value at all.

In due course, the above approach was dropped entirely.

(c) <u>The concept of continuous response</u>.

From discussions and further deliberation it became clear that the idea of a dichotomy, as required for the discriminant model, was unsatisfactory. It seemed reasonable to postulate response to adrenalectomy as a continuous phenomenon, patients reacting to the operation across a spectrum. With this in mind, a latent response variable, X_0, was postulated and clinical remission associated with the event $X_0 \geqslant a$, non-remission with the event $X_0 < a$, where a was to be estimated from suitable data. This idea was developed in [7], where predictor variables $\underset{\sim}{X}$ and X_0 were assumed to have a multivariate normal distribution, and the effect of truncating X_0 at a on the expectation of $\underset{\sim}{X}$ was calculated.

The probability of remission given $\underset{\sim}{X} = \underset{\sim}{x}$, under this model, is simply $P_r\{X_0 \geqslant a | \underset{\sim}{X} = \underset{\sim}{x}\}$ and this was also investigated in [7]. To ensure that the multinormal distribution assumption was justified, a generalisation of a bivariate transformation proposed by Moran (Ref. see [7]) was used.

In order to implement the above model, it is necessary to obtain all the cross correlations between the individual components of $\underset{\sim}{X}$, and between these components and X_o. Thus the relationship between hormone and other physiological measurements are given in [7] together with their individual relationship with the response variable X_o. These are intrinsically interesting in their own right.

In order to test the predictive power of the model, it was fitted to the first 60 of the 130 women in the trial. Then using these estimates, the probability of remission for each of the next five patients was predicted and compared with the realised response. The estimates were then up-dated to be based on the first 65, and the process was repeated until predictions were available for the last 70 patients to enter the trial. This process simulates the real situation and is free from the criticism that back-predictions over data used to estimate the model produce "euphoric results".

The results were as follows:

	Model Pred. Prob.	Realised Prob.
$\leqslant$ med	.142	.176
> med	.444	.500

In order to construct the above table, trial patients were ordered according to their predicted probability of remission. Those below the median probability estimate were put into one group and those above the median into the second group. The average probability under the two models was calculated for each group and compared with the actual outcome.

A discussion of how to optimally apply treatments when several are available is given in [8]. If treatment i, T_i, has a probability $P_i(\underset{\sim}{x})$ of success depending on a set of observations $\underset{\sim}{x}$ made on patients prior to treatment then the procedure of choosing T_i if $P_i(\underset{\sim}{x}) \geqslant P_j(\underset{\sim}{x})$ for all j maximises the overall expected probability of response.

This fact has obvious applications when adrenalectomy is used in connection with other types of treatment. Assuming a remission rate α to radiotherapy, say, which essentially does not depend on $\underset{\sim}{x}$, the obvious, simple and mathematically correct thing to do is to use adrenalectomy when $P(\underset{\sim}{x}) \geqslant \alpha$ and radiotherapy otherwise. That is, "everthing else being equal" of course.

(d) Actuarial work

At the same time as the above described work was proceeding, a close study was made of records at The Victorian Cancer Registry concerning breast cancer. By 1970 there were about 9,000 women who had registered the disease and, from follow-up records, it is possible to determine the survival time of individual registrants. From these figures the survival pattern of breast cancer victims was to be studied as a function of the age of first reporting the disease.

What at first seemed a modest assignment turned into an enormous job. A main problem was the determination of a suitable family of models. Eventually a competing risk model was developed in which the force of mortality function consisted of additive components; one due to natural mortality, φ_0, and one due to the disease, φ_c. Certain parametric forms were proposed for φ_0 and φ_c and special fitting techniques devised, [9].

The parameters of φ_c, θ, were made to depend on age of reporting, a, $\theta(a)$, say. This enabled the construction of a complete set of life tables for women reporting the disease at age a = 35, (1), 85. Once $\theta(a)$ was estimated, the force of mortality was plotted as a function of age. The ensuing patterns may one day be helpful in understanding aspects of the biology of the disease.

Various other influences on mortality were assessed and we now have:

(a) a complete description of the effect of age of first reporting on mortality probabilities and life expectancy;

(b) a study of the effects of

(i) tumour histology

(ii) stage of the disease at reporting

(iii) combinations of (i) and (ii)

(iv) treatment

(v) combinations of (i) and (iv)

(vi) combinations of (ii) and (iv)

on survival probabilities and life expectations at age 60.

Estimates were supplied with standard errors to give some idea of the precision of the various estimation procedures.

Theoretical work arising from the estimation and testing problems in this work has led to quite general results which encompass many standard statistical procedures as special cases [10]. These findings are extensions of work in [9] and will be published elsewhere.

(e) Time to reporting breast cancer.

One of the consequences of the above research was that the effect of the stage of the disease at reporting on subsequent survival parameters was quantified. Moreover, stage was related to the delay time to reporting, and both are clearly associated with the degree of tumour growth.

There are four clinical stages recognised for this disease which are labelled S_1, S_2, S_3, and S_4.

A model was proposed in [11] which allowed the delay to reporting, R, to be random with a distribution depending on tumour volume, V. More explicitly, the conditional rate of reporting was assumed to be proportional to the rate of tumour growth. Under a standard assumption of exponential growth, this is equivalent to assuming that the rate of reporting is directly proportional to tumour volume.

Using the model which is implied by these assumptions and some results from [9], it is possible to estimate boundaries for V by stage. Thus, for S_1, S_2, S_3 and S_4 respectively, $V \leq 50$ ccs, $50 \leq V \leq 133$, $133 \leq V \leq 217$ and $217 \leq V$. The average volumes in each of the stages is given in [12] to be 25, 85, 170 and 325 respectively. In [13], an estimate of the average volume of tumour at death is quoted as about 1,000 ccs.

Also a consequence of the model in [11] it is possible to estimate the average tumour growth rate over all classes of tumours, and for particular tumour grades. The average doubling time from Registry data is three months, and this agrees with internationally quoted average figures. For anaplastic tumours the doubling time is 1.5 months and for less virulent grades, 5 months.

The idea of a minimal detectable volume for breast cancer is important, and clearly this volume is a function of the detection procedure. According to tumour grade then, the average delay to reporting given a minimal detectable volume of .5 ccs ranges from 1 to 3 years with a combined average of 2 years. If delay to reporting is measured from the time of onset of the disease, the average delays vary from 5 to 15 years with an overall average of 8 years.

(f) The concept of cure and screening trials.

In [12] and [13] the idea of an actuarial "cure" and its relation to tumour volume at reporting are introduced. The first part of [13] considers the problem of patients arriving at random into a trial and the effect of this on estimates of survival. It appears that we are the first to propose a specific model which allows a distribution for survival time to be fitted and tested. Presently, obvious conditional likelihood methods are used which do not allow direct tests of goodness of fit.

With the new model we are able to obtain a good estimate of the distribution of survival time for patients in the Peter McCallum Trial. In particular, the unbiased estimates of survival expectation for remitters and non-remitters are about 4 years and 1 year respectively.

Among the other results, however, the most important concern the proportion of actuarial "cures". A woman will be regarded as cured following treatment if her life expectancy is normal.

Now in [13] the life expectancies of women reporting the disease in S_1, S_2, S_3 and S_4 are estimated and these can be compared with the population breast cancer figures in [9]. Since no woman in the trial is cured and, essentially, in every case the disease results in death, the expectations in [13] are assumed to represent the order of life expectancy for those patients who are not cured, whether or not treatment is via adrenalectomy. The figures in [9] are uniformly larger than those in [13] and by solving a suitable equation estimates of cure, α_i, by stages are $\alpha_1 = .378$, $\alpha_2 = .268$, $\alpha_3 = .107$ and $\alpha_4 = .018$.

These values are used in [12] where a model relating tumour volume to cure rates is proposed. In this paper it is shown how the expected cure rate can be estimated for various screening designs for early detection.

In fact, if v_c is the minimal detectable volume of tumour and the screening interval is t_I, then the cure rate is written as a function of v_c and t_I. It seems that good results to screening can be anticipated even for relatively large v_c of 8 - 10 ccs provided t_I is made small, say .1 years. Of course, first class results should follow screenings where both v_c and t_I are small. A comparison of our predictions with results obtained empirically in a New York trial showed excellent agreement.

(g) Summary and conclusions

From work completed to date we have quantified a number of factors associated with breast cancer. For example:

(a) we have a complete actuarial description of the disease as a function of the age at reporting;

(b) the effects of stage at reporting, tumour type and treatment on survival probabilities and life expectancy are known;

(c) the approximate volumes of the tumours at the four stages and at death have been estimated;

(d) the growth rate of tumours in relation to tumour type is known;

(e) the delay patterns of women reporting the disease are available;

(f) we know how to predict remission for patients undergoing adrenalectomy in the advanced stages of the disease and how to optimise the use of various treatment procedures;

(g) we have examined the relationship between response to adrenalectomy and clinical remission;

(h) and finally, we can suggest how designs for early detection procedures can be evaluated.

The prime statistical aim has been to find out the maximum amount of information from the two data sources; the Peter McCallum Trial and The Victorian Cancer Registry. This was to be done by careful mathematical and statistical modelling.

This approach has its dangers and drawbacks. An inadequate model, or an adequate model fitted incorrectly to data can mislead. For these reasons great care in the selection of models has been taken and, where possible, an effort made to ensure that the models fit the data well. At every stage, results have been cross-checked with other known work and, in general, the agreement has been pleasing.

For some of the modelling of [11] [12] and [13] it was not possible to engage in rigorous fitting and testing calculations. More of an applied mathematics attitude was adopted to encourage the models to yield at least the order of magnitude of the various processes. Again, the only insurance against idiot results is careful intuition and cross-checking with other work where possible.

We now have a lot more information about the disease, although we are no nearer elucidating the inter-play of hormones. But, unless our modelling is totally misleading it appears that a satisfactory control of breast cancer may be achievable by:

(1) relatively frequent screenings using sophisticated technology, applied possibly to high risk groups;

or

(2) very frequent screenings or self-examination using insensitive techniques;

or

(3) a combination of both (1) and (2).

IV WORK SUBMITTED IN PREVIOUS DISSERTATIONS

Some of the papers cited above have been used as partial fulfillment of the requirements for M.Sc. and Ph.D. degrees. This work has been included here to provide essential coherence. The relevant papers are;

M.Sc. U.N.S.W. (1962)
a(i)[2]; a(iii)[1],[3],[4],[5].

Ph.D. U.N.S.W. (1964)
a(ii)[2],[3],[4]; a(iv)[1],[2],[3],[4],[5].

V STATEMENT OF PERSONAL CONTRIBUTION

It is not easy to partition contribution in some collaberative studies. However, I believe the following to be fair and accurate statements.

(1) All the comments made above indicate the nature of the original contributions.

(2) I was senior statistician on all research reported here. Therefore, it was my responsibility to give direction to the mathematical and statistical research. In collaborative work involving other statisticians, I endeavoured to more than pull my weight in developing the technical results.

(3) Where I was sole author, I did all the work apart from acknowledged assistance.

(4) In Section (A), except where explicitly mentioned below, the results are essentially all mine.

Paper	Contribution to Results.	Contribution to Research Direction
(a)(i)[1]	50%	50%
(a)(i)[6]	50%	100%
(a)(ii)[6]	50%	100%
(b)[6]	33% (3 authors)	100%
(c)[2],[3]	30-40%	20-30%
(d)[5]	85%	85%
(d)[7]	50%	100%
(d)[8]	50%	100%

(5) In Section (B) I am entirely responsible for all the modelling, theoretical work and for the general direction of the statistical research. The exception is B[9] where my contribution would be about 80%.

(6) Dr. A.A. Donald, who is a parasitologist, collaborated on some of the work in Section A. He provided the detailed biological information necessary for the construction of meaningful models. By discussion, he also indicated where modelling could be of practical use.

(7) Dr. G. Sarfaty played a similar role in the case of the work in (B). He provided essential medical information and stimulated a lot of the work through his enquiries.

(8) A great deal of lengthy numerical calculation has been organised and programmed by Mr. P. Leppard for parts of Section B. His comments have been helpful on some aspects of the modelling and he has checked a number of the results.

A

RESEARCH INTO SOME SPECIFIC PROBLEMS OF THE SHEEP AND WOOL INDUSTRIES

(a)

RESEARCH ASSOCIATED WITH THE GENETICS OF SHEEP BREEDING

(i)

MASS SELECTION THEORY

A (a) (i) [1]

Reprinted from JOURNAL OF ANIMAL SCIENCE, Vol. 20, No. 3, August, 1961

PERFORMANCE INDEX FOR LIFETIME PRODUCTION

S. S. Y. YOUNG AND G. M. TALLIS [1]

Division of Animal Genetics, C.S.I.R.O., McMaster Laboratory, Glebe, Sidney, N.S.W.

VARIOUS methods of selecting breeding animals for several characters have been investigated by a number of workers (Fairfield Smith 1936; Hazel and Lush, 1942; Hazel, 1943; Young and Weiler, 1960) and different techniques such as tandem selection, independent culling levels and index selection have been used. Relatively little attention, however, has been given to the theory of selection for lifetime production. Lush (1945) discussed the use of the estimate of repeatability in predicting the producing ability of cows, but no work has been reported on the theory of selection for lifetime performance on more than one trait.

Genetic gains through selection, measured per unit of time, are governed by the generation interval as well as by the selection differentials and levels of heritability of the characters under selection. For farm animals such as sheep and dairy cattle, where production can be measured on a number of occasions and where the generation intervals are long, it is possible to imagine a number of situations where lifetime performance is relatively more important than its additive genetic merit. For example, when economic weights of production characteristics are stable over short but unstable over long periods, or when the estimates of repeatability are much higher than those of heritability, it may be worthwhile to select for superiority in lifetime performance.

It is the aim of this paper to show that the theory of the selection index (Fairfield Smith, 1936; Hazel, 1943) can be used to develop an index for lifetime production, which has been called a "performance index". This index should be used in selection based on one record for each trait.

[1] Thanks are due to Professor P. R. McMahon of the School of Wool Technology, University of New South Wales who suggested to us the subject of this investigation, to Miss Helen Newton Turner of the Division of Animal Genetics for her comments on the manuscript and to Miss Turner and Mr. C. H. S. Dolling of the same Division for making available the data used in the numerical example.

Theory

Performance Index. The theory of the performance index can be adapted directly from that of the selection index. If several traits, $X_1, X_2, \ldots . X_n$, are under selection the observed value of X_i, x_i, can be written as:

$$x_i = g_i + f_i$$

where g_i is due to genetic and f_i to environmental factors. A selection index in the form of:

$$Z = \Sigma_i \, b_i \, x_i \qquad (1)$$

can then be constructed. The b_i values in the selection index are calculated such that Z can best discriminate between the total breeding values of animals $H = \Sigma_i \, a_i \, g_i$, the a_i being the predetermined economic weight of X_i. If x_i is rewritten as:

$$x_i = s_i + e_i$$

where s_i is due to factors causing permanent differences among animals and e_i to temporary environmental effects (s_i and e_i are assumed to be statistically independent) a performance index analogous to the selection index can be constructed. The performance index can be written as:

$$W = \Sigma_i \, k_i \, x_i \qquad (2)$$

where k_i are values calculated such that W can best discriminate among the total production values, $M = \Sigma_i \, a_i \, s_i$. Our problem is to find the appropriate values of k_i.

If W has variance σ^2_W, then the gain in lifetime production value per animal after selection is:

$$\Delta M = B_{M.W} \, \sigma_W \, I \qquad (3)$$

where $B_{M.W}$ is the regression coefficient of M on W, and I is the selection differential of W in standard deviation units.

To obtain the greatest gain in M, we require therefore to maximize ΔM for any given selection differential I, which is equivalent to maximizing $B_{M.W} \, \sigma_W$. By partial differentiation of the latter expression with re-

spect to the k_i and equating to zero, the optimum k values can be calculated as:

$$k=P^{-1} S a, \text{ or } k_i=\Sigma_{qr} P^{iq} S_{qr} a_r,$$
$$i=1, 2, \ldots n \qquad (4)$$

where: k=column vector of k_i
P^{-1}=inverse matrix of P with elements P^{ij}
P_{ij}=phenotypic covariance between x_i and x_j
S=matrix of S_{ij}
S_{ij}=covariance between s_i and s_j
a=column vector of a_i.

Matrices P and a are well known and S may be constructed by estimates of the between-animal components S_{ij} calculated in the usual manner (table 1).

An interesting special case of (4) is obtained when the traits under selection are independent. In this instance we have $S_{ij}=P^{ij}=0$ and $P^{ii}=\frac{1}{P_{ii}}$ so that:

$$k_i=a_i\frac{S_{ii}}{P_{ii}}$$

The quantity $\frac{S_{ii}}{P_{ii}}$ is generally known as the repeatability (t) of the trait X_i. When one record for each trait is available on the animals to be selected the index reduces to:

$$W=\sum_{i=1}^{n} a_i t_i x_i \qquad (5)$$

Equation (5) is also useful as an approximate index when the phenotypic correlations between traits are small.

As a decision will frequently have to be made between this performance index and the standard selection index, it is of interest to note that the correlation between the two may be written as:

$$r_{WZ}=\sum_{i=1}^{n}\sum_{j=1}^{n}\frac{b_i k_j P_{ij}}{\sigma_Z \sigma_W} \qquad (6)$$

where σ_Z is the standard deviation of the selection index.

TABLE 1. ANALYSIS OF VARIANCE AND COVARIANCE WHEN n ANIMALS ARE MEASURED FOR TWO TRAITS X_i AND X_j IN q YEARS

Source of variation	Degrees of freedom	Variance * and covariance components
Between years	q−1	$E_{ij}+nY_{ij}$
Between animals	n−1	$E_{ij}+qS_{ij}$
Years x animals	(q−1) (n−1)	E_{ij}

* For variance i=j.

Expected Phenotypic and Genetic Gains. It is always desirable to predict both the phenotypic and genetic changes in individual traits which are likely to follow the use of any index, and in the present case it is also useful to compare the total economic gains, so that a choice can be made between the selection and performance indices. The following formulas for these gains may be verified.

(i) Expected gains by using W:

(1) (a) Gain in M,
$$\Delta M=B_{M.W}\,\sigma_W\, I=\sigma_W I \qquad (7)$$

(b) Gain in s_i,
$$\Delta s_i=\Sigma_j\frac{k_j S_{ij}}{\sigma_W} I \qquad (8)$$

(2) (a) Gain in H,
$$\Delta G_W=\Sigma_i \Sigma_j\frac{a_i k_j G_{ij}}{\sigma_W} I \qquad (9)$$

where G_{ij} is the covariance between g_i and g_j

(b) Gain in g_i,
$$\Delta g_{iW}=\Sigma_j\frac{k_j G_{ij}}{\sigma_W} I \qquad (10)$$

(ii) Expected gains by using Z:

(1) (a) Gain in H,
$$\Delta G=\sigma_Z I \qquad (11)$$

(b) Gain in g_i,
$$\Delta g_i=\Sigma_j\frac{b_j G_{ij}}{\sigma_Z} I \qquad (12)$$

(2) (a) Gain in M,
$$\Delta M_Z=\Sigma_i \Sigma_j\frac{a_i b_j S_{ij}}{\sigma_Z} I \qquad (13)$$

(b) Gain in s_i,
$$\Delta s_{iZ}=\Sigma_j\frac{b_j S_{ij}}{\sigma_Z} I \qquad (14)$$

Numerical Illustration

Data from 62 unselected Merino ewes of a medium Peppin strain, run at the National Field Station "Gilruth Plains" Cunnamulla were used in this illustration. The ewes were born in 1948 in the "Control Group" described by Turner (1958), and records of clean wool weight (X_1) and crimps per inch (X_2) for each animal were first taken at 15–16 months of age on 10–11 months' wool growth. Observations of these characters were made annually for four consecutive years after the first sampling. From these data a performance index can be constructed for the selection of similar ewes at 15–16 months of age.

TABLE 2. ANALYSES OF VARIANCE AND COVARIANCE FOR RECORDS OF CLEAN WOOL WEIGHT (X_1) AND CRIMPS PER INCH (X_2) TAKEN FROM 62 EWES IN 5 YEARS

Source of variation	Degrees of freedom	Mean square (X_1)	Covariance (X_1X_2)	Mean square (X_2)	Components		
					X_1	X_1X_2	X_2
Between years	4	18.201	−12.274	10.957	.....	.....	.....
Between ewes	61	3.140	−3.513	16.304	0.5617	−.6845	2.8339
Years x ewes	244	0.332	−.090	2.134	(S_{11})	(S_{12})	(S_{22})

Records of clean wool weight were first corrected to a constant growth period of 365 days. From the records at first shearing values of $P_{11}=0.7777$, $P_{12}=-.7863$ and $P_{22}=3.7438$ were obtained. Estimates of S_{ij} were made from analyses of variance and covariance on the records for five years. The results are summarized in table 2.

Having obtained the estimates of P_{ij} and , the next step is to find suitable economic .eights for X_1 and X_2. Dunlop and Young (1960) estimated the economic weight for crimp to be approximately 25 units if the economic weight of clean wool is to be taken as 100 units. Using these estimates the k_i values can be calculated as:

$$\begin{bmatrix} k_1 \\ k_2 \end{bmatrix} = \begin{bmatrix} 0.7777 & -0.7863 \\ -0.7863 & 3.7438 \end{bmatrix}^{-1} \begin{bmatrix} 0.5617 & -0.6845 \\ -0.6845 & 2.8339 \end{bmatrix} \begin{bmatrix} 100 \\ 25 \end{bmatrix}$$

and we have $k_1=64.6$ and $k_2=14.2$. The performance index is thus:

$$W=64.6\,x_1+14.2\,x_2 \qquad (15)$$

In actual application (15) may be written as:

$$W=4.5\,x_1+x_2$$

since only the relative values of index scores for individual animals are of interest.

The expected phenotypic gains, for 1 standard deviation of selection differential, when (15) is used were calculated and are lown under "Case 1" in table 3.

In the calculations of the accompanying increase in genetic values, the heritability estimate for both clean wool weight and crimp was taken to be 0.4 (Tallis, 1959; Young *et al.*, 1960) and the genetic correlation between them was taken to be −.6 (Tallis, 1959). From these estimates together with the estimates of P_{ij} used previously, it is estimated that $G_{11}=0.3111$, $G_{12}=-.4095$ and $G_{22}=1.4975$. The resulting estimates of genetic gains are shown in table 3 (Case 1). It can be seen that, when the performance index (15) is used, the selected sheep may be expected to show a substantial phenotypic increase in economic value but a decrease in crimp number. The accompanying genetic changes are about half the value of the phenotypic changes.

Using the same estimates of a_i, P_{ij} and G_{ij} a selection index was calculated, the coefficients being $b_1=32.9$ and $b_2=6.0$. The expected genetic gains and their accompanying phenotypic gains are also shown in table 3 (Case 1). The gains from the two indices are similar, but this is not surprising in view of the fact that the correlation between them is 0.997 (Equation 6). This high correlation means that the majority of animals selected by one index will also be selected by the other.

TABLE 3. EXPECTED PHENOTYPIC AND GENETIC GAINS BY THE APPLICATION OF DIFFERENT INDICES IN TWO SITUATIONS [a]

Type of gain	Trait	Case 1		Case 2	
		With performance index	With selection index	With performance index	With selection index
Phenotypic	Total, (economic units)	50.6	50.3	50.6	42.1
	Wool weight (lb.)	0.525	0.557	0.525	0.234
	Crimps per in. (number)	−.079	−.214	−.079	0.747
Genetic	Total, (economic units)	25.7	25.8	35.7	42.8
	Wool weight (lb.)	0.282	0.301	0.220	0.209
	Crimps per in. (number)	−.103	−.174	0.548	0.878

[a] Figures in Cases 1 and 2 were computed with two different sets of genetic parameters.

Consider now an hypothetical case where P_{ij} and S_{ij} have the same values as in Case 1, but the heritability estimates for X_1 and X_2 are 0.2 and 0.4 respectively and the genetic correlation between them is 0.2. Then $G_{11}=0.15$, $G_{12}=0.10$ and $G_{22}=1.50$. The k_i values are the same as in Case 1, but now $b_1=44.8$ and $b_2=22.1$. The expected gains from both indices are shown in table 3 (Case 2).

The correlation between the two indices is 0.833 for Case 2 which is lower than for Case 1. The expected gains from the use of the two indices this time differ markedly.

Discussion

It is reasonable to assume that the permanent differences among animals are partly caused by the effects of early environment, which persist throughout the lives of animals. There is no justification, however, in assuming that the values of S_{ii} and S_{ij} will not vary with the age of the animals. The S_{ii} and S_{ij} estimated from the first two years' records, for example, may differ from the values estimated from records covering four years or more. Since the aim of using a performance index is to select animals for lifetime production, it is suggested that, for higher efficiency, S_{ij} should be estimated from records at as many ages as possible.

From the numerical examples summarized in table 3, it is clear that in some situations the use of either kind of index will lead to similar phenotypic and genetic gains in economic value. It is interesting to note that although gains in values are almost identical, changes in individual traits are not necessarily similar. For example, in Case 1, the use of a selection index will lead to a slightly greater gain in wool weight and a more appreciable decrease in crimp than when a performance index is used. In other situations, such as the hypothetical Case 2, the application of different kinds of indices will lead to quite different expected gains in values as well as in individual traits, and one index may be superior to the other. The suitability of different kinds of indices would, of course, depend on the relative importance of genetic and phenotypic gains to particular breeders of different livestock.

It is of interest to point out that the method of independent culling levels may also be used in the selection of animals for lifetime production. Formulae developed by Hazel and Lush (1942) and Young and Weiler (1960) can readily be adapted for this purpose by substituting the parameters S_{ii} and S_{ij} for G_{ii} and G_{ij}.

Summary

The theory of constructing a performance index which may be used in the selection of animals for phenotypic gains is presented. This index is compared with the genetic selection index and numerical examples are used to illustrate how the use of each index is likely to affect genetic and phenotypic gains.

Literature Cited

Dunlop, A. A. and S. S. Y. Young. 1960. Selection of Merino sheep: An analysis of the relative economic weights applicable to some wool traits. Empire J. Exp. Agr. 28:201.

Hazel, L. N. 1943. The genetic basis for constructing selection index. Genetics 28:476.

Hazel, L. N. and J. L. Lush. 1942. The efficiency of three methods of selection. J. Heredity 33:393.

Lush, J. L. 1945. Animal Breeding Plans. Collegiate Press, Ames, Iowa.

Smith, H. Fairfield. 1936. A discriminant function for plant selection. Ann. Eug. 7:240.

Tallis, G. M. 1959. Sampling errors of genetic correlation coefficients calculated from analyses of variance and covariance. Australian J. Stat. 1:35.

Turner, Helen Newton. 1958. Relationships among clean wool weight and its components. I. Changes in clean wool weight related to changes in the components. Australian J. Agr. Res. 9:521.

Young, S. S. Y., Helen Newton Turner and C. H. S Dolling. 1960. Comparison of estimates of repeatability and heritability for some production traits in Merino rams and ewes. II. Heritability. Australian J. Agr. Res. 11:604.

Young, S. S. Y. and H. Weiler. 1960. Selection for two correlated traits by independent culling levels. J. Genetics (in press).

A (a) (i) [2]

Reprinted from
BIOMETRICS
THE BIOMETRIC SOCIETY, Vol. 18, No. 1, March 1962

171 NOTE: A Selection Index For Optimum Genotype

G. M. TALLIS

Division of Animal Genetics, C.S.I.R.O., McMaster Laboratory Glebe, N.S.W., Australia.

The theory of the general, unrestricted selection index as applied to animal breeding is now well known. However, little attention has been given to the problems of conditional selection which may occur in practice. Recently Kempthorne and Nordskog [1959] presented an interesting method of maximising genetic progress under the restrictions that progress in certain linear genetic functions be zero. It is the purpose of the present note to extend these methods to the case of selection for an optimum genotype.

In the subsequent derivations the assumptions and notation, with

trivial changes, will be those of Kempthorne and Nordskog. For convenience, the relevant notation is summarised below:

p_i = phenotype of ith character,
g_i = additive genotype of ith character,
e_i = non-additive genotypic plus additive environmental contributions to phenotype,
a_i = economic weight of ith character,
$I = \sum_{i=1}^{m} b_i p_i = \mathrm{b'p}$ = selection index,
$H = \sum_{i=1}^{m} a_i g_i = \mathrm{a'g}$ = genetic value of an individual,
$\mathbf{P}$ = phenotypic variance-covariance matrix,
$\mathbf{G}$ = genetic variance-covariance matrix,
$\mathbf{G}_r$ = $r \times m$ matrix obtained from G by deleting the last $m-r$ rows,
Δ_I = selection differential of I,
σ_I^2 = variance of I.

The usual genetic model $p_i = \mu_i + g_i + e_i$, where $E(g) = E(e) = E(ge) = 0$, will be used in the following development.

Suppose now that m characters, p_i, $i = 1, \cdots, m$, are to be used in a selection index, $I = \mathrm{b'p}$, and that ultimately r of these are required to be altered by an amount k_j, $j = 1, \cdots, r$, to bring them to their optimum values. More explicitly, if μ_i is the population mean of p_i prior to selection, then an aim of the breeder may be to change μ_i to $\mu_i + k_i$, $j = 1, \cdots, r$, while allowing $\mu_{r+1}, \cdots, \mu_m$ to increase without limit. A mathematical solution to this problem is obtained by constructing a selection index to maximise gain in $H = \mathrm{a'g}$ subject to the r restrictions

$$\mathrm{Cov}\,(g_i, I)\,\Delta_I/\sigma_I^2 = \alpha k_i, \qquad j = 1, \cdots, r$$

where α is a constant of proportionality. The expression $\mathrm{Cov}\,(g_i, I)/\sigma_I^2$ is the regression coefficient of g_i on I, and hence the restrictions require that, for a given Δ_I, the expectation of g_i be changed by an amount proportional to k_i as a result of selection.

The solution to this problem is obtained as follows: Let $Q = E[(\mathrm{p'b} - \mathrm{g'a})'(\mathrm{p'b} - \mathrm{g'a})]$, then Q is to be minimised subject to the restrictions $\mathrm{Cov}\,(g_i, I) = k_i$, $j = 1, \cdots, r$. Thus, introducing the r Lagrange Multipliers λ_i, the required expression to minimise is

$$\begin{aligned} u &= Q + 2\sum_{i=1}^{r} \lambda_i \,\mathrm{Cov}\,(g_i, I), \\ &= Q + 2\sum_{j=1}^{r} \lambda_j \left(\sum_{i=1}^{m} G_{ji} b_i\right), \\ &= Q + 2\mathrm{b'G}_r'\lambda, \end{aligned}$$

with respect to the b_i . Vector differentiation gives

$$\frac{1}{2}\frac{\partial u}{\partial \mathbf{b}} = \mathbf{Pb} - \mathbf{Ga} + \mathbf{G}_r'\boldsymbol{\lambda} = 0,$$

and solving for b, using the equations of condition, $\mathbf{G}_r\mathbf{b} = \mathbf{k}$, we obtain

$$\mathbf{G}_r\mathbf{b} = \mathbf{k} = \mathbf{G}_r\mathbf{P}^{-1}[\mathbf{Ga} - \mathbf{G}_r'\boldsymbol{\lambda}].$$

If the above equation is solved for $\boldsymbol{\lambda}$ and the result substituted in the equation for b, the final matrix solution is

$$\mathbf{b} = [\mathbf{I} - \mathbf{P}^{-1}\mathbf{G}_r'(\mathbf{G}_r\mathbf{P}^{-1}\mathbf{G}_r')^{-1}\mathbf{G}_r]\mathbf{P}^{-1}\mathbf{Ga} + \mathbf{P}^{-1}\mathbf{G}_r'[\mathbf{G}_r\mathbf{P}^{-1}\mathbf{G}_r']^{-1}\mathbf{k}.$$

Since this index has been constructed so that Cov $(g_i , I) = k_i$, it follows that $\alpha = \Delta_I/\sigma_I^2$. Moreover, it is also clear than when $\mathbf{k} = 0$, the above expression reduces to the formula of Kempthorne and Nordskog and when $r = m$, $\mathbf{b} = \mathbf{G}^{-1}\mathbf{k}$. The degenerate case $r = 0$ gives the ordinary unrestricted selection index, $\mathbf{b} = \mathbf{P}^{-1}\mathbf{Ga}$.

After approximately $1/\alpha$ applications of this index the characters $p_1 , \cdots , p_r$ are expected to reach their optimum values while characters $p_{r+1} , \cdots , p_m$ will have made near maximum genetic advance under the specified restrictions. Because of the accumulated errors in the use of such indices (see Tallis [1960]), this ideal will probably not be achieved. Therefore, the index should be recalculated from time to time as actual progress is assessed. This will involve suitable adjustments to the k_i and/or the estimated G and P matrices as more information comes to hand.

There are several examples in animal industry where the index presented in this paper may be of use. For instance, in wool production, the main selection character is fleece weight, although other features of the fleece may also be important. Thus, a selection index may be required to increase the average fleece weight of the flock as much as possible while stabilizing staple length and crimps per inch at optimum values determined by the market. Another example is in meat production where maximum progress in amount and economy of production is required, accompanied by minor changes in certain carcass characteristics.

REFERENCES

Kempthorne, O., and Nordskog, A. W. [1959]. Restricted selection indices. *Biometrics 15*, 10–19.

Tallis, G. M. [1960]. The sampling errors of estimated genetic regression coefficients and the errors of predicted genetic gains. *Aust. J. Statist. 2*, 66–77.

A (a) (i)[3]

Reprinted from
BIOMETRICS
THE BIOMETRIC SOCIETY, VOL. 22, NO. 1, MARCH 1966

EQUILIBRIA UNDER SELECTION FOR k ALLELES

G. M. TALLIS
The Johns Hopkins University, Baltimore, Md., U. S. A.

INTRODUCTION

The theory of selection involving one locus and two alleles, A and a say, is well established and has been discussed by Li [1955]. The case when there are three alleles has received the attention of Owen [1954] and Li [1955], while the general multi-allelic situation has been investigated by Kimura [1956], Mandel [1959], and Kingman [1961].

It is the purpose of the present paper to propose an algorithm for finding all the equilibria of a k-allelic system and for testing for stationarity. By identifying the situation with the problem of the maximization of a quadratic form with a simplex as domain, elementary methods of quadratic programming can be employed to generate the required solutions.

Before embarking on the general case, it appears advisable to review briefly some of the basic results of the two allelic situation. We consider the situation illustrated in Table I. Notice that the

TABLE I

SELECTION IN A 2-ALLELIC SYSTEM

Genotype	Proportion f	Fitness W	fW
AA	p^2	1	p^2
Aa	$2pq$	$1 - s_1$	$2pq(1 - s_1)$
aa	q^2	$1 - s_2$	$q^2(1 - s_2)$
Total	1.00		$\bar{W}$

population is assumed to be panmictic and that the average fitness of all genotypes is $\bar{W} = p^2 + 2pq(1 - s_1) + q^2(1 - s_2) = 1 - 2s_1pq - s_2q^2$. The coefficients s_1 and s_2 can take any values less than or equal to 1, and therefore any selective situation can be obtained by choosing them appropriately. Now, Δq, the change in the gene frequency of a, can

be expressed in the form

$$\Delta q = \tfrac{1}{2}pq \frac{d \log \bar{W}}{dq}$$

which is due to Wright [1942]. By using the form $\bar{W} = p^2 + 2pq(1 - s_1) + q^2(1 - s_2)$ and taking the partial derivative with respect to q, we have the alternative expression

$$\Delta q = q\tfrac{1}{2}\left\{\frac{\partial \log \bar{W}}{\partial q} - 2\right\},$$

which finds a generalization when we consider k alleles.

In order to determine equilibria, we equate Δq to zero and solve for q. Thus, if we let $\Delta q = f(q)$, say, then $f(\hat{q}) = 0$ defines an equilibrium value of q. In order to classify $\hat{q}$ as to its stability, $df(\hat{q})/dq$ is calculated and

1) if $df(\hat{q})/dq < 0$, then $\hat{q}$ is a stable equilibrium,
2) if $df(\hat{q})/dq > 0$, then $\hat{q}$ is an unstable equilibrium,
3) if $\hat{q} = 0, 1$, then the equilibrium is trivially stable.

These ideas will be used in a more general form in the following development.

THE CASE OF k ALLELES

We now consider the case where we have k alleles, A_i, of frequency p_i $(i = 1, 2, \cdots k)$, and define $\gamma_{ij} = 1 - s_{ij}$ as the coefficient of fitness of genotype A_iA_j, where s_{ij} is the coefficient of selection and $i, j = 1, 2, \cdots, k$. Under panmixia and prior to selection, we have the genotypic array $\{\sum p_iA_i\}^2$, while after selection it is

$$\sum_{i,j} p_ip_j\gamma_{ij}A_iA_j \tag{1}$$

which can be written in matrix form as $\mathrm{p}'\gamma\mathrm{p} = \bar{W}$. The quantity $\bar{W}$ is, by definition, the average fitness of the population. The new frequency of A_i, $p_i^{(1)}$, after selection is given by

$$p_i^{(1)} = \sum_j p_ip_j\gamma_{ij}/\bar{W}$$

$$= \tfrac{1}{2}p_i\partial \log \bar{W}/\partial p_i ,$$

treating all the p_i as functionally independent for the purpose of differentiation. Hence we have the matrix result

$$\mathrm{p}^{(1)} = \tfrac{1}{2}D(\mathrm{p})\partial \log \mathrm{p}'\gamma\mathrm{p}/\partial\mathrm{p} \tag{2}$$

where $D(\mathrm{p}) = \mathrm{diag}\ (p_1, p_2, \cdots, p_k)$.

Now for an equilibrium we must have $\mathbf{p}^{(1)} - \mathbf{p} = \Delta\mathbf{p} = 0$ and therefore

$$\Delta\mathbf{p} = \tfrac{1}{2}D(\mathbf{p})\partial \log \mathbf{p}'\gamma\mathbf{p}/\partial\mathbf{p} - \mathbf{p} = 0 \tag{3}$$

is the required equation to be solved for $\mathbf{p}$. Thus

$$\partial \log \mathbf{p}'\gamma\mathbf{p}/\partial\mathbf{p} = 2$$

and, if we formally complete the differentiation we obtain

$$\gamma\mathbf{p}/\mathbf{p}'\gamma\mathbf{p} = 1 \tag{4}$$

whence

$$\hat{\mathbf{p}} = \gamma^{-1}1\lambda \tag{5}$$

where λ is chosen so that $1'\hat{\mathbf{p}} = 1 = (1'\gamma^{-1}1)\lambda$, $1' = (1, 1, \cdots 1)$. Therefore $\lambda = (1'\gamma^{-1}1)^{-1}$ and

$$\hat{\mathbf{p}} = \gamma^{-1}1/1'\gamma^{-1}1.$$

But, although $\mathbf{p}$ is normalized, there is in general no guarantee that all its components are greater than zero and the problem of finding the equilibrium points must therefore be examined more closely.

We note here also that the equilibrium fitness is

$$\hat{\mathbf{p}}'\gamma\hat{\mathbf{p}} = \lambda^2 1'\gamma^{-1}1 = \lambda.$$

We consider next the following stationary value problem:

Find all the stationary values of $Q(\mathbf{p}) = \mathbf{p}'\gamma\mathbf{p}$ subject to the constraint $\mathbf{p}'1 = 1$. Using the standard Lagrange procedure we find

$$\partial R(\mathbf{p})/\partial\mathbf{p} = 0 = \partial Q(\mathbf{p})/\partial\mathbf{p} - 2\lambda 1$$

where

$$R(\mathbf{p}) = Q(\mathbf{p}) - 2\lambda(\mathbf{p}'1 - 1),$$

and this equation leads to

$$\mathbf{p} = \gamma^{-1}1\lambda$$

which is clearly of the form (5) with $\lambda = (1'\gamma^{-1}1)^{-1}$. Thus, the problem of finding equilibrium points can be formally identified with the above stationary value problem.

However, in order to find permissible stationary values, that is values which lie in the $(k - 1)$ dimensional simplex with vertices $(1, 0, 0, \cdots 0)$, $(0, 1, 0, \cdots, 0)$, $(0, 0, 1, \cdots, 0)$, $\cdots$, $(0, 0, 0, \cdots 1)$, it is necessary to impose the further constraint $\mathbf{p} \geq 0$, i.e. $p_i \geq 0$ for

all i. We proceed by writing

$$\Delta \mathrm{p} = D(\mathrm{p})\{\gamma \mathrm{p} - 1\lambda\} = 0. \tag{6}$$

and we introduce the following notation. Let $\gamma(i)$ be the 'reduced' matrix obtained from γ by deleting the ith row and column, with a similar definition for $\gamma(i, j)$, $i \neq j$. Then it is clear from (6) that some of the components of p may be zero and, therefore, all possible solutions are found by first solving (6) for $\mathrm{p} \geq 0$, then deleting p_i to obtain the vector $\mathrm{p}(i)$ and using $\gamma(i)$ for all i to obtain k more solutions. The process is continued using $\gamma(i, j)$ for all $i \neq j$, and so on until all possible combinations have been investigated. In this fashion

$$\binom{k}{0} + \binom{k}{1} + \binom{k}{2} + \cdots + \binom{k}{k-1} = 2^k - 1$$

solutions are found, and we call the set of these solutions P. P may contain some vectors with negative components and these are eliminated to give the reduced set of solutions $\bar{P}$.

We have now found all the points of equilibrium, $\bar{P}$, and it is only necessary now to determine which ones are stable. We observe that, if $\hat{\mathrm{p}}$ is a vector corresponding to a stable equilibrium, then $\Delta p_i > 0$ when $p_i < \hat{p}_i$ and $\Delta p_i < 0$ when $p_i > \hat{p}_i$ for all $i = 1, 2, \cdots, k-1$. It is then clear that $\hat{\mathrm{p}}$ corresponds to a local maximum of $Q(\mathrm{p})$, and p_k is now written as $1 - \sum_{i=1}^{k-1} p_i$ by use of the constraint relation.

Alternatively, we can write $\gamma = \mathrm{U} - \mathrm{S}$, where $\mathrm{S} = (s_{ij})$ and $\mathrm{U} = 11'$ to obtain a minimization problem. Thus

$$\bar{W} = Q(\mathrm{p}) = \mathrm{p}'(\mathrm{U} - \mathrm{S})\mathrm{p} = 1 - \mathrm{p}'\mathrm{S}\mathrm{p}$$

and (6) becomes

$$D(\mathrm{p})\{\mathrm{S}\mathrm{p} - 1\lambda\} = 0. \tag{7}$$

Clearly any vector maximizing $Q(\mathrm{p})$ minimizes $\mathrm{p}'\mathrm{S}\mathrm{p}$ and the whole procedure of obtaining $\bar{P}$ can be applied using S instead of γ. This has a slight advantage when applying tests for the type of extrema.

In order to investigate the nature of the stationary values in $\bar{P}$ we calculate the $(k-1) \times (k-1)$ matrix $\theta = [\theta_{ij}]$, where

$$\theta_{ij} = \tfrac{1}{2}\, \partial^2 \mathrm{p}'\mathrm{S}\mathrm{p}/\partial p_i\, \partial p_j = s_{ij} - s_{ik} - s_{ik} + s_{kk}\,.$$

Thus θ can be obtained from S by subtracting the last column from the first $(k-1)$ columns and then subtracting the last row from the first $(k-1)$ rows. The kth row and column are then deleted from the resulting matrix to give θ.

To illustrate the method of testing for a stable equilibrium (a

minimum of p'Sp) suppose the vector $\hat{p} \in \bar{P}$ where $\hat{p}$ has all components greater than zero. Then to determine whether or not $\hat{p}$ is stable, we form all the $(k-1)$ determinants of the principle minors of 0. If these determinants are all positive, $\hat{p}$ is stable. This same procedure is followed for all the reduced vectors $p(i) \in \bar{P}$ by using the appropriate reduced form of 0 obtained from $S(i)$. In this fashion all elements of $\bar{P}$ can be tested and classified as to whether or not the equilibrium is a stable one.

If either γ or S is singular, then equations (6) or (7) can be solved for p using generalized inverses of γ and S. Thus, for instance, $\hat{p} = S^{c}1\lambda$ and $\lambda = (1'S^{c}1)^{-1}$ where S^{c} is a matrix such that $SS^{c}S = S$. As before, the test as to whether or not $\hat{p}$ is stable depends on 0. However, note that it is necessary that

$$\operatorname{rank}\begin{pmatrix}\gamma\\ 1'\end{pmatrix} = \operatorname{rank}\begin{pmatrix}S\\ 1'\end{pmatrix} = k$$

if a stable equilibrium is to exist with k positive gene frequencies.

We will now illustrate the above methods by establishing a result due to Wright (Li [1955], p. 260) on selection for heterozygotes. The assumption is that all heterozygotes (A_iA_i) have equal fitness, unity, and the homozygotes (A_iA_i) have fitness $(1 - s_i)$. The problem is to find the stable equilibrium (if one exists), $\hat{p}$ with all components of $\hat{p}$ positive.

The solution is obtained almost immediately. The matrix $S = \operatorname{diag}(s_1, s_2, \cdots s_k)$ and we have from (7)

$$\hat{p} = \lambda S^{-1}1, \qquad \lambda = (1'S^{-1}1)^{-1}.$$

It is now obvious that $\hat{p}_i = s_i^{-1}/\sum_{i=1}^{k} s_i^{-1}$ and the equilibrium is stable provided $s_i > 0$ for all i.

Suppose, now, $k = 3$. Then in the above case there are $2^3 - 1 = 7$ equilibrium values. These are

$$\hat{p}_1 = 1,$$

$$\hat{p}_2 = 1,$$

$$\hat{p}_3 = 1,$$

$$\hat{p}_1 = s_1^{-1}/(s_1^{-1} + s_2^{-1}) \quad \text{and} \quad \hat{p}_2 = s_2^{-1}/(s_1^{-1} + s_2^{-1}),$$

$$\hat{p}_1 = s_1^{-1}/(s_1^{-1} + s_3^{-1}) \quad \text{and} \quad \hat{p}_3 = s_3^{-1}/(s_1^{-1} + s_3^{-1}),$$

$$\hat{p}_2 = s_2^{-1}/(s_2^{-1} + s_3^{-1}) \quad \text{and} \quad \hat{p}_3 = s_3^{-1}/(s_2^{-1} + s_3^{-1})$$

and finally

$$\hat{p}_i = s_i^{-1} \Big/ \sum_{i=1}^{3} s_i^{-1} \qquad \text{for} \quad i = 1, 2, 3.$$

These equilibria are all stable under the obvious condition that the s_i involved are greater than zero.

We consider finally the situation where all homozygotes have unit fitness and all heterozygotes have fitness $\gamma = 1 - s$. Under these circumstances γ is of the form

$$\gamma = \begin{bmatrix} 1 & \gamma & \gamma & \cdots & \gamma \\ \gamma & 1 & \gamma & \cdots & \gamma \\ \vdots & \vdots & \vdots & & \vdots \\ \gamma & \gamma & \gamma & \cdots & 1 \end{bmatrix}_{k \times k}$$

and it is now necessary to calculate γ^{-1}. We assume that γ^{-1} is of the form

$$\gamma^{-1} = \begin{bmatrix} a & b & b & \cdots & b \\ b & a & b & \cdots & b \\ \vdots & \vdots & \vdots & & \vdots \\ b & b & b & \cdots & a \end{bmatrix}_{k \times k}$$

and solve the two equations

$$a + (k - 1)\gamma b = 1$$
$$b + [(k - 2)b + a]\gamma = 0$$

for a and b. These equations have a unique solution and, since we require $\gamma^{-1}1 = 1(a + (k - 1)b)$, we obtain $\hat{p} = 1(1 + (k - 1)\gamma)^{-1}\lambda$ after some algebra. However, $\lambda = (1'\gamma^{-1}1)^{-1} = [k(1 + (k - 1)\gamma)^{-1}]^{-1}$ and hence $\hat{p}_i = 1/k$ for all i. (This neat method of inverting γ was suggested to me by Dr. Charles Rohde.)

It turns out that in this case 0 has the form

$$0 = \begin{bmatrix} -2s & -s & -s & \cdots & -s \\ -s & -2s & -s & \cdots & -s \\ \vdots & \vdots & \vdots & & \vdots \\ -s & -s & -s & \cdots & -2s \end{bmatrix}_{(k-1)\times(k-1)}$$

and it can be shown that the determinant of 0, $|0|$, is $k(-s)^{k-1}$. This

shows clearly that the determinants of the principal minors of S alternate in sign if $s > 0$, are all zero if $s = 0$ and are all positive if $s < 0$. Thus, for $s < 0$ and $k = 3$, say, the non-trivial stable equilibria are $(\frac{1}{3}, \frac{1}{3}, \frac{1}{3})$, $(\frac{1}{2}, \frac{1}{2}, 0)$, $(\frac{1}{2}, 0, \frac{1}{2})$ and $(0, \frac{1}{2}, \frac{1}{2})$.

The author is grateful for a number of references which were brought to his attention by a referee and Dr. Mandel.

REFERENCES

Kimura, M. [1956]. Rules for testing stability of a selective polymorphism. *Proc. U. S. Nat. Acad. Sci 42*, 336–40.

Kingman, J. F. C. [1961]. A mathematical problem in population genetics. *Proc. Camb. Phil. Soc. 57*, 574–82.

Li, C. C. [1955]. *Population genetics*. The University of Chicago Press.

Mandel, S. P. H. [1959]. The stability of a multiple allelic system. *Heredity 13*, 289–302.

Owen, A. R. G. [1954]. Balanced polymorphism of a multiple allelic series. *Caryologia, Supp. 6*, 1240–41.

Wright, S. [1942]. Statistical genetics and evolution. *Bull. Am. Math. Soc. 48*, 223–46.

A(a)(i)[4]

Reprinted from
BIOMETRICS
THE BIOMETRIC SOCIETY, Vol. 22, No. 2, June 1966

220 NOTE:

A Migration Model

G. M. TALLIS
Johns Hopkins University, Baltimore, Md., U.S.A.

1 INTRODUCTION

A model for migration, first introduced by Wright, is discussed by Li [1955], Chapter 21. Basically, the assumption is that the total population is subdivided into k isolates $\pi^{(i)}$, $i = 1, 2, \cdots k$, each

undergoing random mating. The size of the groups is considered to be sufficiently large and of equal size so that the genotypic array for the ith group for a single locus and two alleles is

$$\pi^{(i)} = \{p_i^2 AA + 2p_iq_iAa + q_i^2aa\},$$

where p_i is the gene frequency of A in $\pi^{(i)}$. Under these assumptions the mean and variance of group gene frequencies are

$$\bar{q} = \sum q_i/k \quad \text{and} \quad \sigma_q^2 = \sum (q_i - \bar{q})^2/k.$$

It follows immediately that the genotypic array in the whole population is

$$\pi^{(\cdot)} = \{(\bar{p}^2 + \sigma_q^2)AA + (2\bar{p}\bar{q} - 2\sigma_q^2)Aa + (\bar{q}^2 + \sigma_q^2)aa\}.$$

If the groups are of different sizes, the above results are suitably modified by introducing appropriate weights. If w_i is the correct weight for group i, $\sum_1^k w_i = 1$, then define

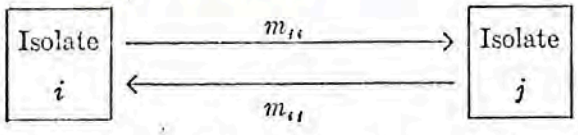
$$\bar{q} = \sum_1^k w_iq_i \quad \text{and} \quad \sigma_q^2 = \sum_1^k w_iq_i^2 - \bar{q}^2$$

and the above representation of $\pi^{(\cdot)}$ holds.

To introduce the notion of migration, suppose now that each group exchanges a proportion m of its members with a random sample of the total population every generation. If q_i is the gene frequency of the particular sub-population under consideration then $q_i' = (1-m)\ q_i + m\bar{q}$. Moreover, $q_i - \bar{q}$ is the deviation of gene frequency prior to migration which after migration becomes $q_i' - \bar{q} = (1 - m)(q_i - \bar{q})$. This shows that the variance σ_q^2 is altered to $(1 - m)^2\sigma_q^2$ as a result of migration, emphasizing the fact that, under the assumed model, migration tends to make all the q_i equal.

II THE NEW MODEL

As above we consider k isolates with gene frequencies $q_i^{(n)}$, $i = 1, 2, \cdots, k$ at the beginning of generation n. Let the quantities m_{ij} and m_{ji} be the proportions of migrants from isolate j to isolate i and from isolate i to isolate j respectively:



We will also assume initially that the matrix $\mathbf{M} = [m_{ij}]$ is doubly stochastic. Then, since $q_i^{(n+1)} = \sum_{j=1}^k m_{ij}q_j^{(n)}$, we have the equation

$$\mathbf{q}^{(n+1)} = \mathrm{M}\mathbf{q}^{(n)} \tag{1}$$

whence

$$\mathbf{q}^{(n)} = \mathrm{M}^n\mathbf{q}^{(0)} \tag{2}$$

where $\mathbf{q}^{(0)}$ is the vector of gene frequencies for the k isolates at generation zero.

The mean gene frequency at generation n is $k^{-1}1'\mathbf{q}^{(n)} = \bar{q}^{(n)} = \bar{q}^{(0)}$, where 1 is a vector with k unit components and the variance $V(q^{(n)})$ can be put in the form

$$V(q^{(n)}) = k^{-1}\mathbf{q}^{(n)\prime}[\mathrm{I} - k^{-1}11']\mathbf{q}^{(n)},$$

where I is the $(k \times k)$ identity matrix. If $\mathbf{q}^{(n)}$ tends to a vector of the form $1q_\infty$, with q_∞ some constant, then $V(q^{(n)})$ will tend to zero. On the other hand, if $q^{(n)}$ tends to some vector $\mathbf{q}^{(\infty)}$ which is not of the form $1q_\infty$, then $V(q^{(n)})$ will not tend to zero. Finally, it may happen that $\mathbf{q}^{(n)}$ does not approach a limit as n gets large and in this case no equilibrium is reached.

III EQUILIBRIA

In order to make some progress with the problem of examining equilibrium values, we will assume that M, in addition to being doubly stochastic, is non-singular and similar to a diagonal matrix. Under these conditions, we have the following result.

Theorem

(a) If M has complex latent roots of modulus unity, $\mathbf{q}^{(n)}$ does not tend to a limit.

(b) When M has no complex roots of modulus one, and the multiplicity r of the largest latent root, $(\lambda_1 = 1)$ is one, $\mathbf{q}^{(n)}$ tends to the limit $1\bar{q}^{(0)}$ and $V(q^{(n)})$ tends to zero.

(c) Under the conditions of (b) and with $r > 1$, $\mathbf{q}^{(n)}$ tends to a limit $\mathbf{q}^{(\infty)}$ which is not necessarily of the form $1\bar{q}^{(0)}$ and $V(q^{(n)})$ may not tend to zero.

The proofs of these statements will not be given in detail since they follow directly from the spectral theory of stochastic matrices. In fact, (a) is obvious and (c) can be verified by examples. In the case of (b), M^n tends to the matrix $k^{-1}11'$ and $\mathbf{q}^{(\infty)} = k^{-1}11'\mathbf{q}^{(0)} = 1\bar{q}^{(0)}$.

The above theorem emphasizes the dependence of the equilibrium gene frequencies on the matrix M. This is in contrast to the more elementary model discussed earlier where, always, $\mathbf{q}^{(n)} \xrightarrow[n]{} 1\bar{q}^{(0)}$.

IV UNEQUAL POPULATION GROWTH

Suppose that the k isolates are not all of the same size at generation zero, and let these sizes be given by the vector $\mathbf{w}^{(0)}$. Moreover, if $\mathbf{M}$ is only singly stochastic with $\sum_{i=1}^{k} m_{ij} = 1$ and $\sum_{j=1}^{k} m_{ij} = \theta_i$, say, then we have the situation of differential growth among the isolates.

To discuss this case briefly, we assume for simplicity that $\mathbf{M}$ satisfies the conditions of (b) of the theorem. The population sizes at generation n are given by $\mathbf{w}^{(n)} = \mathbf{M}^n\mathbf{w}^{(0)}$ and $\mathbf{w}^{(\infty)} = k^{-1}\mathbf{x}_1\mathbf{1}'\mathbf{w}^{(0)}$, where $\mathbf{x}_1$ is the right eigenvector corresponding to $\lambda_1 = 1$ and $\mathbf{x}_1'\mathbf{1} = k$.

Similarly, if we let $\gamma_j^{(0)} = w_j^{(0)}q_j^{(0)}$, $j = 1, 2, \cdots, k$, the distribution of 'gene mass' for allele a at generations zero and n are given by $\boldsymbol{\gamma}^{(0)}$ and $\boldsymbol{\gamma}^{(n)} = \mathbf{M}^n\boldsymbol{\gamma}^{(0)}$ respectively. Clearly, $q_i^{(n)} = \gamma_i^{(n)}/w_i^{(n)}$ and $q_i^{(\infty)} = \gamma_i^{(\infty)}/w_i^{(\infty)} = k^{-1}x_{1i}k\bar{\gamma}^{(0)}/k^{-1}x_{1i}k\bar{w}^{(0)} = \bar{q}^{(0)}$. Notice also that $\bar{q}^{(n)} = \sum_{i=1}^{k} w_i^{(n)}q_i^{(n)}/\sum_{i=1}^{k} w_i^{(n)} = \mathbf{1}'\mathbf{M}^n\boldsymbol{\gamma}^{(0)}/\mathbf{1}'\mathbf{M}^n\mathbf{w}^{(0)} = \bar{q}^{(0)}$, as it should.

The author is indebted to the referee for valuable comments on the original draft of this note.

REFERENCE

Li, C. C. [1955]. *Population Genetics.* The University of Chicago Press.

A(a)(i)[5]

Commonwealth of Australia
COMMONWEALTH SCIENTIFIC & INDUSTRIAL RESEARCH ORGANIZATION
"BIOMETRICS" Vol. 24, No. 1 Pages 169–177 March

SELECTION FOR AN OPTIMUM GROWTH CURVE

G. M. TALLIS[1]

The Johns Hopkins University, Baltimore, Maryland, U.S.A.

SUMMARY

Growth and development can be regarded as a stochastic process in continuous time. Moreover, in some situations of primary production, certain growth patterns may be more economical, or otherwise more desirable, than others. In this paper an attempt is made to develop techniques which could be used to exert some selection pressure for optimal growth curves.

A discrete solution to this problem is suggested which should be relatively easy to apply in practice. However, the complete discussion of the situation requires the introduction of certain integral equations to replace the usual matrix equations of the classical theory. The conventional phenotypic and additive genetic covariance matrices give way to continuous kernels and, as expected, it is found that the continuous and the discrete theories are similar. The usual case of unrestricted selection is also developed for the continuous time model and the selection of several characters towards respective optimal curves simultaneously is also treated.

Two different numerical procedures for solving the integral equations are proposed.

INTRODUCTION

In this paper we consider the growth curves of animals. These may be regarded as the realization of a stochastic process $P(t)$ with expectation function $\mu(t)$, $t \geq 0$. Suppose now there is an optimum curve $\alpha(t)$, then the problem is to construct a selection index, I, which will gradually move $\mu(t)$ to $\alpha(t)$ at some or all points $t \in [0, a]$.

DISCRETE SOLUTION

We first consider the problem of changing $\mu(t)$ to $\alpha(t)$ at well-defined points t_i, $i = 1, 2, \cdots, n$. These points may be considered as defining critical stages in the growth cycle of the animals. In order to make progress, we assume the usual additive model $P(t) = \mu(t) + g(t) + e(t)$, where $g(t)$ is the additive genetic and $e(t)$ the environmental contribution to phenotype at time t. The components $g(t)$ and $e(t)$ are considered to satisfy the relation $E\{g(s)e(t)\} = 0$ for all $s, t \in [0, a]$.

---

[1]Now at Division of Mathematical Statistics, C.S.I.R.O. Alpha House, 60 King Street, Newtown, N.S.W., Australia

For convenience we write $p_i \equiv p(t_i) \equiv P(t_i) - \mu(t_i)$, $\mu(t_i) \equiv \mu_i$, $g_i \equiv g(t_i)$, and $e_i \equiv e(t_i)$. In this notation then the problem is to construct a selection index $I = \sum_{i=1}^{n} \beta_i p_i = \beta' \mathrm{p}$ such that cov $(g_i, I) = k_i$, $i = 1, 2, \cdots, n$, where $k_i = (\alpha_i - \mu_i)$. Since under standard assumptions Δ_i, the expected change in μ_i after selection on I is Δ_I cov $(g_i, I)/\sigma_I^2$, where Δ_I is the selection differential in I, suitably defined, we would have $\Delta_i = \Delta_I k_i/\sigma_I^2$. Hence, after approximately σ_I^2/Δ_I selections, the total change in μ_i should be about k_i and the optimum values will have been reached.

In order to find β, notice that the above conditions can be written in matrix form as $E\{\mathrm{gp}'\beta\} = \mathrm{k}$ and, since $E\{\mathrm{gp}'\} = \mathrm{G}$, the genetic covariance matrix $\beta = \mathrm{G}^{-1}\mathrm{k}$. This is a special case of a restricted selection index reported by the author [1962].

For a rather wide class of procedures for estimating G, $\hat{\beta} = \hat{\mathrm{G}}^{-1}\mathrm{k}$ is a consistent estimator of β. Moreover, the analysis can be put in the multivariate analysis of covariance form

source	D.F.	estimate	expectation
between families	f_b	$\hat{\mathrm{B}}$	$\mathrm{B} = \mathrm{W} + (r/m)\mathrm{G}$
error	f_w	$\hat{\mathrm{W}}$	W

where r is the number per family group, and $m = 2$ or $m = 4$ according to the type of family relationship. Thus, $\hat{\mathrm{G}} = (m/r)[\hat{\mathrm{B}} - \hat{\mathrm{W}}]$.

The covariance matrix for $\hat{\beta}$, $\mathrm{V}(\hat{\beta})$, is obtained by noticing that $\mathrm{G}(d\beta) + (d\mathrm{G})\beta = 0$, whence $\mathrm{G}(d\beta) = -(d\mathrm{G})\beta$. Thus, proceeding in the standard fashion

$$\mathrm{GV}(\hat{\beta})\mathrm{G} = E\{d\mathrm{G}\beta\beta' \, d\mathrm{G}\}, \tag{1}$$

and letting $\beta\beta' = \gamma$, then the (q, r) element of $\mathrm{GV}(\hat{\beta})\mathrm{G}$ is $\sum_s \sum_t \gamma_{st}$ cov $(\hat{G}_{tr}, \hat{G}_{qs})$. But it can be shown that

$$\text{cov } (\hat{G}_{tr}, \hat{G}_{qs}) = \frac{m^2}{r^2}\{(B_{tq}B_{rs} + B_{ts}B_{rq})f_b^{-1} + (W_{tq}W_{rs} + W_{ts}W_{rq})f_w^{-1}\},$$

and substituting this into (1) we have after simplification

$$\mathrm{V}(\hat{\beta}) \simeq \frac{m^2}{r^2}\mathrm{G}^{-1}\{(\mathrm{B}\gamma\mathrm{B} + \mathrm{B}\,\mathrm{Trace}\,\mathrm{B}\gamma)f_b^{-1} + (\mathrm{W}\gamma\mathrm{W} + \mathrm{W}\,\mathrm{Trace}\,\mathrm{W}\gamma)f_w^{-1}\}\mathrm{G}^{-1}. \tag{2}$$

It is now possible to calculate the approximate variance of the estimate $\hat{I} = \hat{\beta}'\mathrm{p}$. We write $\hat{\beta} = \beta + \Delta\beta$, then since $E\{\Delta\beta\mathrm{p}'\} = 0$,

$E(\hat{I}) = 0$ and

$$V(\hat{I}) = E(\beta'\mathrm{p} + \Delta\beta'\mathrm{p})^2 = \beta'\mathrm{P}\beta + \mathrm{Trace}\,(\mathrm{V}(\hat{\beta})\mathrm{P}), \tag{3}$$

where $\mathrm{P} = E(\mathrm{pp}')$, the phenotypic covariance matrix.

THE CONTINUOUS SOLUTION

Although the index constructed above, $\beta = \mathrm{G}^{-1}\mathrm{k}$, may be of some practical importance, it is only of minor theoretical interest. What is really required is some index which will apply selection pressure to all points of the curve simultaneously. In this section we propose to discuss such an index.

We introduce now the covariance function $E\{p(s)p(t)\} = \Gamma_p(s, t) = \Gamma_g(s, t) + \Gamma_e(s, t)$, and in the following argument $\Gamma_p(s, t)$ will replace the matrix G of the discrete solution. For this treatment we let $k(t) = \alpha(t) - \mu(t)$, a continuous function of time. Our index, instead of being of the form $\beta'\mathrm{p}$, is $I = \int_0^a \beta(t)p(t)\,dt$, where the integral is to be interpreted as a stochastic integral. It is well known that a sufficient condition for the latter to exist is that $\int_0^a \int_0^a \Gamma_p(s, t)\beta(s)\beta(t)\,ds\,dt = \sigma_I^2$ exists.

The condition that cov $(g_i, I) = k_i$ is now replaced by $E\{g(t)I\} = k(t)$. But, $E\{g(t)\int_0^a p(s)\beta(s)\,ds\} = \int_0^a \beta(s)\Gamma_g(s, t)\,ds = k(t)$ (see appendix), and in order to find $\beta(s)$ the integral equation

$$\int_0^a \beta(s)\Gamma_g(s, t)\,ds = k(t) \tag{4}$$

must be solved. Equation (4) is a Fredholm integral equation of the first kind, and exact inversion, in general, is not an easy task.

If $p(t)$ is a normal process, then $\int_0^a p(t)\beta(t)\,dt = I$ will also be normal (Loève [1963] page 485). Thus the expected change at the point $t \in [0, a]$ after selection will be $\Delta(t) = \Delta_I k(t)/\sigma_I^2$. Hence, as for the discrete case, after σ_I^2/Δ_I selections, $\mu(t)$ should be near $\alpha(t)$ at all points $t \in [0, a]$.

Some points pertaining to the solution of the selection integral equation will now be discussed. First, in order to make some progress, some parametric form can be assigned to the covariance kernel $\Gamma_g(s, t)$ and, for purposes of illustration, we will let $\Gamma_g(s, t) = \sum_{i=1}^n \omega_i e^{\eta_i|s-t|}$, ω_i and η_i real, $\omega_i > 0$. Although the form of Γ_g has been chosen for mathematical convenience, this model should be satisfactory whenever the kernel can be assumed strictly positive.

In order to find a solution to (4) one can attempt to write $\beta(s) = \sum_{i=0}^{\infty} \theta_i\phi_i(s)$, where the θ_i are real constants and the $\phi_i(s)$ are the eigenfunctions of the kernel $\Gamma_g(s, t)$. However, with the particular form

assumed above we can calculate an approximation to $\beta(s)$ in another way.

We have $\int_0^a (\sum_{i=1}^n \omega_i e^{\eta_i t})\beta(s)\,ds = k(t)$ and, by differentiating both sides with respect to t and setting $t = 0$, we obtain $\mu_1 (\sum_{i=1}^n \omega_i \eta_i) = k'(0)$, where $\mu_1 = \int_0^a s\beta(s)\,ds$. Similarly,

$$\mu_i\left(\sum_{i=1}^n \omega_i \eta_i^j\right) = k^{(i)}(0), \qquad \mu_i = \int_0^a s^i \beta(s)\,ds,$$

and provided the $k^{(i)}(0)$ exist, these equations can be solved for μ_i if $\sum_{i=1}^n \omega_i \eta_i^i \neq 0$.

If, now, $\beta(s)$ can be expressed approximately as a polynomial of degree $m-1$, $\beta(s) = \sum_{i=0}^{m-1} b_i s^i$, then $\mu_i = \sum_{i=0}^{m-1} b_i a^{i+j+1}/(j+i+1)$ and the vector of coefficients, b, can be found from the equation

$$\mathbf{b} = \mathbf{A}^{-1}\mathfrak{y},$$

where $\mathrm{A} = (a^{i+j+1}/(i+j+1))$ and $\mathfrak{y}' = (\mu_1, \mu_2, \cdots, \mu_n)$. A higher degree polynomial $\beta(s)$ can be constructed by taking more moments and if, for some j, $\sum \omega_i \eta_i^j = 0$, then the jth moment is omitted and the process is completed using an additional moment of higher order than μ_j.

From the Weierstrass approximation theorem it follows that if a continuous solution to (4) exists, then the above procedure leads to a uniformly close approximation to $\beta(s)$. That is, given an ϵ there exists an $m = M$ such that for all $m > M$, $|\sum_{i=0}^{m-1} b_i s^i - \beta(s)| < \epsilon$ for all $s \in [0, a]$.

In order to obtain estimates of the parameters ω_i and η_i, $\hat{\omega}_i$ and $\hat{\eta}_i$, $i = 1, 2, \cdots, n$, we may obtain estimates of genetic covariance, cov $(g(t_s), g(t_r)) = G_{sr}$, as described earlier, and use these in a standard non-linear least squares analysis to calculate $\hat{\omega}_i$ and $\hat{\eta}_i$. The details of such an analysis will not be given here since they are well established.

In the above case it is not difficult to obtain an expression for $V(\hat{I})$ which is entirely analogous to (3). It will be possible, in general, to obtain approximate sampling variances and covariances for the $\hat{\omega}_i$ and $\hat{\eta}_i$; hence we can find $\mathrm{V}(\hat{\mathfrak{u}})$. Now, $\hat{\mathbf{b}} = \mathrm{A}^{-1}\hat{\mathfrak{y}}$, $\hat{\beta}(s) = \hat{\mathbf{b}}'\mathbf{s}$ when $\mathbf{s}' = [1, s, s^2, \cdots, s^{m-1}]$ and $V(\hat{\mathbf{b}}) = \mathrm{A}^{-1}\mathrm{V}(\hat{\mathfrak{y}})\mathrm{A}^{-1}$.

If we let $\hat{\beta}(s) = \beta(s) + \Delta\beta(s)$, then

$$V(\hat{I}) = V\left(\int_0^a p(t)\hat{\beta}(t)\,dt\right)$$

$$= E\left\{\int_0^a p(t)(\beta(t) + \Delta\beta(t))\,dt \cdot \int_0^a p(s)(\beta(s) + \Delta\beta(s))\,ds\right\},$$

and using the assumption that $E\{p(t)\Delta\beta(s)\} = 0$ we find that

$$V(\hat{I}) = \int_0^a \int_0^a \beta(s)\beta(t)\Gamma_p(s, t)\, ds\, dt + \int_0^a \int_0^a B(s, t)\Gamma_p(t, s)\, ds\, dt, \qquad (5)$$

where $B(s, t) = E\{\Delta\beta(s)\ \Delta\beta(t)\}$.

It is clear that in the example $B(s, t) = \mathrm{s'A^{-1}V(\hat{u})A^{-1}t}$ and hence we have the required extension of (3) to the continuous case. Another method of solving (4) is given in the Discussion.

EXTENSIONS

We consider next the continuous analogue of the general genetic selection index. In the discrete case the problem is to find a vector β such that $\beta'\mathrm{p}$ is in some sense the best predictor of $\mathrm{a'g}$, when a is a vector of economic weights. By using a least squares argument, we minimize

$$\begin{aligned} E\{(\mathrm{g'a} - \mathrm{p'}\beta)'(\mathrm{g'a} - \mathrm{p'}\beta)\} \\ &= \mathrm{a'Ga} - 2\beta'\mathrm{Ga} + \beta'\mathrm{P}\beta \\ &= \mathrm{a'Ga} - \mathrm{a'GP^{-1}Ga} + (\beta - \mathrm{P^{-1}Ga})'\mathrm{P}(\beta - \mathrm{P^{-1}Ga}) \qquad (6) \end{aligned}$$

with respect to β. It is clear that equation (6) is at a minimum when $\beta = \mathrm{P^{-1}Ga}$, which is the required solution.

Introducing the economic weight function $a(t)$, in the continuous case we must minimize

$$\begin{aligned} I(\beta) &= E\left\{\int_0^a g(t)a(t)\, dt - \int_0^a p(t)\beta(t)\, dt\right\}^2 \\ &= \int_0^a \int_0^a a(s)a(t)\Gamma_g(s, t)\, ds\, dt - 2\int_0^a \int_0^a a(t)\beta(s)\Gamma_g(s, t)\, ds\, dt \\ &\qquad + \int_0^a \int_0^a \beta(s)\beta(t)\Gamma_p(s, t)\, ds\, dt \end{aligned}$$

with respect to $\beta(t)$. This is a variational problem and we write $I(\beta + \epsilon\xi) = F(\epsilon)$, where $\xi(t)$ is an arbitrary continuous function vanishing at 0 and a. Then

$$\delta I = 0 = \epsilon \frac{dF(0)}{d\epsilon} = 2\epsilon \int_0^a \xi(t)\left[\int_0^a \Gamma_g(s, t)a(s)\, ds - \int_0^a \Gamma_p(s, t)\beta(s)\, ds\right] dt$$

implies, since $\xi(t)$ is arbitrary, that

$$\int_0^a \Gamma_p(s, t)\beta(s)\, ds = \int_0^a \Gamma_g(s, t)a(s)\, ds. \qquad (7)$$

Thus, (7) is another Fredholm integral equation of the first kind which must be solved for $\beta(t)$.

Finally, suppose that, instead of just a single growth curve, we are interested in different curves which describe the overall phenotypic growth of the animal. Thus, we have a vector valued random process $\mathrm{P}(t)' = [P_1(t), P_2(t), \cdots, P_q(t)]$ and we let

$$E\{\mathrm{P}(t)\} = \mathfrak{y}(t), \qquad E\{[\mathrm{P}(t) - \mathfrak{y}(t)][\mathrm{P}(s) - \mathfrak{y}(s)]'\} = \Gamma_p(s, t) = [\Gamma_p^{ij}(s, t)],$$

where

$$\Gamma_p^{ij}(s, t) = E\{p_i(t)p_j(s)\} = \Gamma_g^{ij}(s, t) + \Gamma_e^{ij}(s, t).$$

The problem now is to move $\mathfrak{y}(t)$ to $\mathfrak{y}(t) + \mathrm{k}(t)$.

Consider $I = \sum_{i=1}^{q} \int_0^a p_i(s)\beta_i(s)\, ds$ as the prospective index. Then we must have

$$E\{g_i(t)I\} = k_i(t) = \sum_{j=1}^{q} \int_0^a \Gamma_g^{ij}(s, t)\beta_j(s)\, ds, \qquad i = 1, 2, \cdots q,$$

or in obvious matrix form

$$\int_0^a \Gamma_g(s, t)\beta(s)\, ds = \mathrm{k}(t). \tag{8}$$

In the finite case these results specialize as follows. Let $I = \sum_{i=1}^{q} \mathrm{p}_i'\beta_i$, where $\mathrm{p}_i' = [p_i(t_1), p_i(t_2), \cdots, p_i(t_n)]$, then the condition is that

$$E\{\mathrm{g}_i I\} = \mathrm{k}_i, \qquad \mathrm{k}_i' = [k_i(t_1), k_i(t_2), \cdots, k_i(t_n)].$$

If $\mathrm{G}(ij) = E\{\mathrm{g}_i\mathrm{g}_j'\}$, then we obtain the equation

$$\sum_{j=1}^{q} \mathrm{G}(ij)\beta_j = \mathrm{k}_i; \qquad i = 1, 2, \cdots, q$$

or, in full matrix notation

$$\mathrm{G}\beta = \mathrm{k}, \tag{9}$$

where $\beta' = [\beta_1', \beta_2', \cdots, \beta_q']$, $\mathrm{k}' = [\mathrm{k}_1', \mathrm{k}_2', \cdots, \mathrm{k}_q']$ and

$$\mathrm{G} = \begin{bmatrix} \mathrm{G}(11) & \mathrm{G}(12) & \cdots & \mathrm{G}(1q) \\ \mathrm{G}(21) & \mathrm{G}(22) & \cdots & \mathrm{G}(2q) \\ \vdots & \vdots & & \vdots \\ \mathrm{G}(q1) & \mathrm{G}(q2) & \cdots & \mathrm{G}(qq) \end{bmatrix}$$

From (9) it is clear that a unique solution to the more general problem is guaranteed.

DISCUSSION

There are a number of practical situations to which the techniques of the previous sections could be applied. For instance, in fat lamb production one desirable characteristic is rapid and early increase in body weight. On the other hand, if these weight increments are too great then some production and marketing problems may arise. Hence, some optimal growth curve can perhaps be specified towards which the average flock performance is to be pushed.

Different breeds of livestock are characterized not only by the quality of the associated primary product, but also by how rapidly, in what quantity, and how efficiently it is produced. Thus, early maturity in lambs may be a desirable feature under some systems of management and marketing, whereas under entirely different conditions late maturing sheep may be optimal. Under any specific set of circumstances it makes sense to use the particular breed which has, among other things, the correct growth pattern. The concept of an optimal growth curve, therefore, appears to have genuine and important practical implications.

In most cases one feels that the discrete solution suggested above would provide an adequate selection tool. Practically, one would probably be satisfied to have the population growth curve approach the optimal curve at a finite number of points since intuition suggests that intermediate points will also be brought near optimality automatically. Provided n is a reasonable number, the vector β can be estimated with little trouble.

However, it is not just of academic interest to investigate possible means of applying selection pressure throughout a continuum. It is worthwhile to see where the theory extends and to observe the similarity between the general and the specific results. If very many measurements were recorded on each individual it is possible that the discrete methods would become unwieldy, and the continuous solution may be the most appropriate approximation to use. Of course, with enough points numerical quadrature methods could be used to estimate accurately $\hat{I} = \int_0^a \hat{\beta}(t)p(t)\,dt$.

Nevertheless, the theoretical problems associated with the solution of (4) have not been adequately emphasized. For any particular kernel Γ_σ and function $k(t)$ there may in fact be no exact solution. To see this for a very simple case, in the parametric representation of Γ_σ let $n = 1$, $\omega_1 = 1$, $\eta_1 = -1$ and $\beta(s) \equiv 0$ for $s > a$. Then (4) takes the form

$$\int_0^\infty e^{-ts}\beta(s)\,ds = k(t)$$

and it is clear that unless $k(t)$ is the Laplace Transform of some function for $0 \leq t \leq a$, the equation has no solution. Hence, for the moment method of solving (4) it is implicitly assumed that $k(t)$ is such that a solution does exist. Moreover, this solution is presumed continuous.

From the practical viewpoint, these matters are probably of little consequence. The suggested procedure should produce satisfactory results in most cases.

In the final analysis, however, fully numerical procedures for finding $\beta(s)$ may give the best results. To illustrate one such method suppose, in the notation of the discrete solution, we let $t_1 = \Delta/2$, $t_2 = 3\Delta/2$, $\cdots$, $t_n = (2n - 1)\Delta/2$, where $\Delta = a/n$. Then, provided n is sufficiently large, (4) can be approximated by

$$\sum_{j=1}^{n} G_{ij}\beta_j = k_i/\Delta, \qquad i = 1, 2, \cdots, n. \tag{10}$$

Thus the numerical calculation of the continuous solution reduces to a special case of the discrete solution, although the philosophy is quite different. In (10) we are trying to approximate an integral whereas in the solution $\beta = G^{-1}k$ this is not so.

Suppose now that the β_i has been calculated according to (10), $\beta = G^{-1}k/\Delta$, then if in the future all animals are measured at the same time points t_i, then I can be approximated by

$$I = \sum_{i=1}^{n} \beta_i p_i \Delta.$$

In spite of all the theoretical problems involved in solving (4), therefore, in any particular case approximations are obtainable from suitable data. Moreover, the procedure outlined above may be more satisfactory than the one discussed earlier in the paper, since no parameterization of the kernel Γ_s is required.

ACKNOWLEDGEMENT

The author is indebted to Dr. Earle Klosterman of the Ohio Agricultural Experiment Station, Wooster, for a discussion leading to the formulation of the problems discussed above.

This work was supported in part by Research Grant GM-13225 from the National Institutes of Health while the author was a Visiting Associate Professor in the Biometrics Unit, Plant Breeding Department, Cornell University, in July–August, 1966.

SELECTION POUR UNE COURBE DE CROISSANCE OPTIMALE

RESUME

La croissance et le développement peuvent être considérés comme un processus stochastique à temps continu. De plus, dans quelques situations de production primaire, certains schémas de croissance peuvent être plus économiques, ou d'une autre façon plus désirables que d'autres. Dans ce papier on essaie de développer des techniques qui pourraient être utilisées pour exercer quelque pression de sélection en vue d'obtenir des courbes de croissances optimales.

Une solution discrète de ce problème est suggérée, qu'on pourrait mettre en pratique de façon relativement aisée. Cependant, la discussion complète de la situation implique l'introduction de certaines équations intégrales pour remplacer les habituelles équations matricielles de la théorie classique. Les matrices de covariance conventionnelles conduisent à des noyaux continus et, comme prévu, on trouve que les théories continue et discrète sont semblables. Le cas habituel de sélection sans contrainte est également développé pour le modèle à temps continu et la sélection de plusieurs caractères en vue d'une optimisation simultanée de leurs courbes respectives est également traitée.

Deux procédures numériques différentes pour résoudre les équations intégrales sont proposées.

REFERENCES

Loève, M. [1963]. *Probability Theory.* 3rd Edn. D. Van Nostrand Company, Inc., Princeton, New Jersey.

Tallis, G. M. [1962]. A selection index for optimum genotype. *Biometrics 18*, 120-2.

APPENDIX

Prior to equation (3) it is asserted that $E\{g(t)I\} = \int_0^a \beta(s)\Gamma_g(s, t)\,ds$ and the purpose here is to establish this result. Let $t_0, t_1, t_2, \cdots, t_n$ be a partition of the interval $[0, a]$ and define

$$I_n = \sum_{i=1}^{n} \beta(t_i)p(t_i)(t_i - t_{i-1})$$

then

$$|E\{g(t)(I - I_n)\}| \leq E\{|g(t)(I - I_n)|\} \leq [E\{g^2(t)\} \cdot E\{(I - I_n)^2\}]^{1/2}$$

by the Schwarz inequality. Taking limits on both sides of the above inequality we find, since $\lim_{n\to\infty} E\{(I - I_n)^2\} = 0$, that

$$E\{g(t)I\} = \lim_{n\to\infty} \sum_{i=1}^{n} \beta(s_i)\Gamma_g(s_i, t)(s_i - s_{i-1})$$

$$= \int_0^a \beta(s)\Gamma_g(s, t)\,ds.$$

The last integral exists by assumption.

A(a)(i)[6]

TECHNICAL PAPER NO. 5
APRIL 1975

TWO OPTIMISATION PROBLEMS OF CONSTRAINED INDEX SELECTION THEORY

BY

G.M. TALLIS AND P. CHESSON

UNIVERSITY OF ADELAIDE
DEPARTMENT OF STATISTICS

TWO OPTIMISATION PROBLEMS OF CONSTRAINED INDEX SELECTION THEORY

Tallis, G.M., Chesson, P.

I INTRODUCTION

We consider random variables

$$Y_0, Y_1, \ldots, Y_q, X_1, \ldots, X_p, \qquad q \leq p$$

where $E[Y_i] = E[X_j] = 0$, $V[Y_i] = 1$, $V[\underset{\sim}{X}] = V$, is of full rank, $C[\underset{\sim}{Y}\underset{\sim}{X}'] = [\underset{\sim}{a}_0, \underset{\sim}{a}_1, \ldots \underset{\sim}{a}_q]$. Put $A = [\underset{\sim}{a}_1, \ldots, \underset{\sim}{a}_q]$ and assume rank $[\underset{\sim}{a}_0, A] = q+1$.

Problem 1

Minimise $E[Y_0 - \underset{\sim}{\beta}'\underset{\sim}{X}]^2$ with respect to $\underset{\sim}{\beta}$ subject to $A\underset{\sim}{\beta} = \alpha\underset{\sim}{w}$, α arbitrary.

Problem 2

Maximise $Cor[Y_0, \underset{\sim}{\beta}'\underset{\sim}{X}]$ with respect to $\underset{\sim}{\beta}$ subject to $A\underset{\sim}{\beta} = \alpha\underset{\sim}{w}$, α arbitrary.

II RESULTS

Problems 1 and 2 have common solutions. This can be seen from the following lemma.

Lemma

Let B be a closed set in R^p such that, for all $\underset{\sim}{x}$ in B, $\alpha\,\underset{\sim}{x}$ is in B. If $\underset{\sim}{\beta}_0$ is a point in B at which $E[Y - \underset{\sim}{X}'\underset{\sim}{\beta}]^2$ in a minimum then $\underset{\sim}{\beta}_0$ maximises $Cor(Y_0, \underset{\sim}{\beta}'\underset{\sim}{X})$.

Proof

Suppose $\underset{\sim}{\beta}_0$ is a point in B which minimises $E[Y-\underset{\sim}{X}'\underset{\sim}{\beta}]^2$. Let $\underset{\sim}{\beta}_1$ be a point in B and define $\underset{\sim}{\beta}_2$ as

$$\left(\frac{\underset{\sim}{\beta}_0' V \underset{\sim}{\beta}_0}{\underset{\sim}{\beta}_1' V \underset{\sim}{\beta}_1}\right)^{\frac{1}{2}} \underset{\sim}{\beta}_1$$

With this definition $\beta_0' V \beta_0 = \beta_2' V \beta_2$ and $\mathrm{Cor}\,(Y_0, \beta_2' \underset{\sim}{X}) = \mathrm{Cor}\,(Y_0, \beta_1' \underset{\sim}{X})$. However

$$E[Y_0 - \beta_0' \underset{\sim}{X}]^2 \leqslant E[Y_0 - \beta_2' \underset{\sim}{X}]^2;$$

and noting that

$$E[Y - \underset{\sim}{X}' \beta]^2 = 1 + \beta' V \beta - 2\,\mathrm{Cor}(Y_0, \beta' X)(\beta' V \beta)^{\frac{1}{2}}$$

it is seen that

$$\begin{aligned} \mathrm{Cor}\,(Y_0, \beta_0' \underset{\sim}{X}) &\geqslant \mathrm{Cor}\,(Y_0, \beta_2' \underset{\sim}{X}) \\ &= \mathrm{Cor}\,(Y_0, \beta_1' \underset{\sim}{X}). \end{aligned}$$

Hence β_0 maximises the correlation.

General Solution

Note that any β satisfying $A\beta = \alpha \underset{\sim}{w}$ can be written

$$\beta = \alpha A^g \underset{\sim}{w} + N \underset{\sim}{v}\ ,\ v \text{ arbitrary},$$

where A^g is a generalised inverse of A (hence $\alpha A^g \underset{\sim}{w}$ is a particular solution of $A\underset{\sim}{\beta} = \alpha \underset{\sim}{w}$) and the columns of N are a basis for the null space of A. Now put $\underset{\sim}{\lambda}' = (\alpha, \underset{\sim}{v}')$ and $M = [A^g \underset{\sim}{w} \vdots N]$, M having full rank, then

(1) $$\beta = M \underset{\sim}{\lambda},\quad \lambda \text{ arbitrary.}$$

With this observation

$$\begin{aligned} \min_{A\underset{\sim}{\beta} = \alpha \underset{\sim}{w}} E[Y_0 - \beta' \underset{\sim}{X}]^2 &= \min_{\underset{\sim}{\lambda}} E[Y_0 - \underset{\sim}{\lambda}' M' \underset{\sim}{X}]^2 \\ &= \min_{\underset{\sim}{\lambda}} [1 - 2\underset{\sim}{\lambda}' M \underset{\sim}{a}_0 + \underset{\sim}{\lambda}' M' V M \underset{\sim}{\lambda}]. \end{aligned}$$

Differentiating with respect to $\underset{\sim}{\lambda}$ and equating to $\underset{\sim}{0}$ gives

(2) $$M' V M \underset{\sim}{\lambda} = M' \underset{\sim}{a}_0$$

$$\underset{\sim}{\lambda} = (M' V M)^{-1} M' \underset{\sim}{a}_0$$

and $$\beta_1 = M (M' V M)^{-1} M' \underset{\sim}{a}_0\ .$$

Note that if any of the restrictions, $q \leqslant p$, rank $[\underset{\sim}{a}_0 \vdots A] = q + 1$ and $\underset{\sim}{w} \neq \underset{\sim}{0}$ do not hold an optimal β is still given by equations (1) and (2) but $M' V M$ may not be invertible.

An alternative method of solution, which leads to an explicit representation for the minimising vector, is to perform the sequential minimisation

$$\min_{\alpha} \min_{A\underset{\sim}{\beta}=\alpha\underset{\sim}{w}} E[Y_0-\underset{\sim}{\beta}'\underset{\sim}{X}]^2.$$

Put

$$\phi(\underset{\sim}{\beta}) = 1-2\underset{\sim}{\beta}'\underset{\sim}{a}_0+\underset{\sim}{\beta}'V\underset{\sim}{\beta}+\underset{\sim}{\beta}'A'\underset{\sim}{\lambda}$$

then

$$\tfrac{1}{2}\frac{\partial\phi}{\partial\underset{\sim}{\beta}} = -\underset{\sim}{a}_0+V\underset{\sim}{\beta}+A'\underset{\sim}{\lambda} = 0$$

$\underset{\sim}{\beta} = V^{-1}[\underset{\sim}{a}_0-A'\underset{\sim}{\lambda}]$, and $\underset{\sim}{\lambda} = (AV^{-1}A')^{-1}[AV^{-1}\underset{\sim}{a}_0-\alpha\underset{\sim}{w}]$

using $A\underset{\sim}{\beta} = \alpha\underset{\sim}{w}$, and finally

$$\underset{\sim}{\beta} = [I-V^{-1}A'(AV^{-1}A')^{-1}A]\,V^{-1}\underset{\sim}{a}_0+V^{-1}A'(AV^{-1}A')^{-1}\alpha\underset{\sim}{w}$$

$\underset{\sim}{\beta}(\alpha)$, say. Now $1-2\underset{\sim}{\beta}'(\alpha)\underset{\sim}{a}_0+\underset{\sim}{\beta}'(\alpha)V\underset{\sim}{\beta}(\alpha)$ is a quadratic in α and has a minimum at $\alpha^X = \underset{\sim}{a}^{0\prime}V^{-1}A'(AV^{-1}A')^{-1}\underset{\sim}{w}/\underset{\sim}{w}'(AV^{-1}A')^{-1}\underset{\sim}{w}$ and hence the vector minimising $E[Y_0-\underset{\sim}{\beta}'\underset{\sim}{X}]^2$ subject to $A\underset{\sim}{\beta}=\alpha\underset{\sim}{w}$, α arbitrary, is $\underset{\sim}{\beta}(\alpha^X)$.

III DISCUSSION

The application of this result to genetic index selection is immediate. The variable Y_0 plays the role of the weighted economic genotype while $Y_1,\ldots Y_q$, represent additive genotypes for q characters. The X_i, then, are the full set of p phenotypes and the selection problem is to design an index I, using the X_i, which maximises the expected gain in Y_0. This maximisation is to be subject to restricted genetic progress in characters 1 to q i.e. Problem 2.

It is well known that, in the unconstrained situation, the two formulations of the problem given in II lead to the same index, $\beta = V^{-1}\underset{\sim}{a}_0$. However, when there are constraints involving Cor $[I,Y_i]$, it is not a priori clear that the two formulations still lead to a common result.

We have shown here that they do.

(ii)

TRUNCATION PROCEDURES RELATED TO MASS SELECTION

A (a)(ii)[1]

Commonwealth of Australia
COMMONWEALTH SCIENTIFIC AND INDUSTRIAL
RESEARCH ORGANIZATION

Reprinted from "Applied Statistics"
Vol. 10 No. 2 Page Nos. 77-82 June 1961

An Application of Non-Parametric Statistics to Truncated Selection

G. M. TALLIS

AN APPLICATION OF NON-PARAMETRIC STATISTICS TO TRUNCATED SELECTION

G. M. TALLIS
C.S.I.R.O., Division of Animal Genetics, McMaster Laboratory, Glebe, New South Wales

In this article Dr Tallis describes a procedure which saves time and labour when a proportion of a large population has to be selected on the basis of some measurement. Details of the calculations necessary are indicated for the general case, and for the particular cases likely to be of practical interest the results of these calculations are presented in a table to facilitate application.

Introduction

In animal and plant breeding, selection for a single character is frequently effected by truncation. More specifically, if x is a particular metrical character of interest, then the usual procedure is to discard all $x<b$, where b is a constant chosen in such a manner that the proportion of individuals retained is p, say.

Unfortunately, the point of truncation, b, is not usually known until all measurements are made, and hence all individuals must be tagged and identified with their particular scores. If the size of the group is large the latter operation may be expensive and inconvenient. Therefore in some cases it may be desirable to estimate b from a random sample taken from the group prior to measurement, so that individuals can be selected, at the time of measuring, on the basis of the estimate. It is the purpose of this paper to develop a method to accomplish this and to indicate the expected efficiency associated with the technique.

A good example of a situation where some prior estimate of b is desirable occurs in large flocks of sheep at shearing time. Recently, many commercial wool growers in Australia have been selecting their replacements on fleece weights and usually selection is completed at the conclusion of shearing. If the grower requires to save the best sheep available for replacements and he needs a proportion p of the total replacements flock, then obviously he is faced with the problem of truncated selection discussed above. In these circumstances the usual procedure is to tag each sheep and identify it with its fleece weight so that the best $100p\%$ animals may be determined from the records and sorted out of the main flock after shearing. If the flock is large, the amount of work involved in this operation is considerable and the methods of this paper have been devised with the aim of reducing the amount of labour of such selection programmes, while keeping in mind the limited facilities for performing computations in the field.

Methods

Let a sample of size n be drawn at random from a group which is to undergo truncated selection for the character x, and let the proportion to be retained be p. If the sample is ranked in descending order of x, then the n observations may be indicated by $x_1, x_2, \ldots x_n$, where subscripts refer to particular rankings. For instance, x_s is the measurement corresponding to the sth member of the ranked sample. Now, if $f(x)$, a continuous density function, is used to approximate the distribution of x in the group from which selection is to be made, then it is possible to estimate the percentage points of $f(x)$ from the ordered sample, $x_1, x_2, \ldots x_n$. In fact it is easily verified that the area of $f(x)$ lying to the left of x_s, w say, is distributed as a beta density with parameters s and $N-s$, where $N=n+1$. This result is independent of the type of density function, $f(x)$.

The beta density function may now be used to solve the following problems:

(*a*) What value of s, $s=l$ say, satisfies the equation $\Pr(w>p)=1-\alpha$, given a fixed n. This amounts to solving

$$\frac{1}{B(s, N-s)}\int_p^1 w^{s-1}(1-w)^{N-s-1}dw = 1-\alpha \qquad \ldots . (1)$$

for s.

(*b*) Once $s=l$ is determined, what value of p, p_1 say, satisfies the equation $\Pr(w<p_1)=1-\beta$ for fixed n and l. The required equation is

$$\frac{1}{B(l, N-l)}\int_0^{p_1} w^{l-1}(1-w)^{N-l-1}dw = 1-\beta \qquad \ldots . (2)$$

which must be solved for p_1.

The details of the solution of these equations are rather uninteresting. Briefly, (1) is transformed to Fisher's z distribution by the change of variable

$$z = \tfrac{1}{2}\ln\left\{\frac{(1-q)w}{q(1-w)}\right\} \qquad q = \frac{s}{N}$$

It is found that under this transformation z is distributed with parameters $n_1=2s$ and $n_2=2(N-s)$. From the asymptotic expansion of the z distribution presented by Fisher and Cornish (1960) it is possible to find an $s=l$ such that

$$\Pr\left[2z>\ln\left\{\frac{(1-q_1)p}{q_1(1-p)}\right\}\right] \simeq 1-\alpha \qquad q_1 = l/N$$

The latter equation must be solved by iterative techniques. Once l is found, it is a simple matter to calculate a z_1, such that $\Pr(z<z_1)=1-\beta$. The required quantity p_1 satisfying (2) is then obtained immediately from the inverse transformation

$$p_1 = \frac{q_1 e^{2z_1}}{1 + q_1(e^{2z_1} - 1)}$$

Equations (1) and (2) have been solved for various values of n and p for the case when $\alpha = \beta = 0{\cdot}05$ and the results of these calculations are recorded in Table I (p. 82). Next to the entries for p_1 there is a column for the 'efficiency', E_1, the derivation of which will now be given.

From equations (1) and (2), it is clear that Table I has been constructed so that, given the desired proportion of individuals to be retained, p, and the sample size, n, then if x_l is used as the estimate of the true truncation point b, then $\Pr(p < w < p_1) \simeq 0{\cdot}90$, where w is the actual proportion retained. Now, since $\mathrm{Av}(w) = l/N = q_1$, it is clear that instead of the desired proportion p being saved, on the average q_1 will be retained. Moreover, $q_1 > p$ and hence in order to finally obtain p selected individuals, on the average $(q_1 - p)$ will have to be discarded at random from the selected group. This procedure must obviously result in a certain loss of efficiency. In order to obtain some idea of the magnitude of this loss, it is convenient to assume $\mathrm{f}(x)$ approximately normal and to compare the selection differentials associated with p and q_1. (The selection differential is defined as the difference between the means of the selected and unselected groups.) If Δ and Δ_1 represent the selection differentials associated with p and q_1 respectively, then for the present purpose efficiency is defined as

$$E = \frac{\Delta_1}{\Delta} = \frac{\mathrm{Z}(q_1)p}{\mathrm{Z}(p)q_1}$$

where $\mathrm{Z}(p)$ and $\mathrm{Z}(q_1)$ are the ordinates of the normal curve associated with proportions p and q_1. Of course, if $\mathrm{f}(x)$ deviates from normality the values of E_1 in Table I may not necessarily apply. However, these figures are probably sufficient to obtain a good indication of the way the efficiency at the average truncation percentage changes with varying p and n. In order to avoid undue loss of efficiency it seems desirable to tag all sheep in the sample so that these sheep may also be selected on the basis of the estimate of b.

For example, suppose 60% of a group of individuals is to be retained, and an efficiency of 80% is considered tolerable. Then $p = 0{\cdot}6$, $E_1 = 80$, and from the Table, $n = 75$, $l = 52$. Thus 75 individuals would need to be measured and ranked, the truncation point being the measurement of the individual of 52nd rank. The 90% confidence interval for w is 0·6 to 0·768, that is, truncation at the 52nd rank would result in retaining from 60% to 77% of the individuals. The selected group may be adjusted later to the required size by discarding individuals at random. When the odd one in twenty chance occurs and $w < 0{\cdot}6$, then the additional numbers required are obtained at random from the unselected individuals; here again some loss of efficiency may occur.

Alternative Procedure

In order to complete the required truncation by the above method, it is only necessary to divide the population into two groups. The top group is retained and is adjusted to the required p value by discarding individuals at random. However, it can be seen that, when n is small, this procedure may result in a considerable loss of efficiency.

In order to overcome this, a slightly more complicated method of selection can be adopted. If we solve the equation,

$$\frac{1}{\mathrm{B}(s, \mathcal{N}-s)} \int_0^p w^{s-1}(1-w)^{\mathcal{N}-s-1}\mathrm{d}w = 1-\gamma \qquad \ldots . (3)$$

for s, we can find an $s=u$, say, such that if truncation is effected at x_u then $\Pr(w<p)=1-\gamma$. If the results of (1) and (3) are combined it can be seen that if the true percentage point of $\mathrm{f}(x)$ corresponding to p is x_p, then $\Pr(x_l<x_p<x_u)=1-\alpha-\gamma$. Equation (3) has been solved for various combinations of n and p with $\gamma=0{\cdot}05$ and the values of u are also given in Table I.

Note: In order to solve (3) for s with $\gamma=0{\cdot}05$, no calculations are necessary. By the change of variable $y=1-w$ in (3) we find that

$$\frac{1}{\mathrm{B}(u, \mathcal{N}-u)} \int_0^p w^{u-1}(1-w)^{\mathcal{N}-u-1}\mathrm{d}w$$

$$= \frac{1}{\mathrm{B}(\mathcal{N}-u, u)} \int_{1-p}^1 y^{\mathcal{N}-u-1}(1-y)^{u-1}\mathrm{d}y \simeq 1-\gamma$$

The second integral is in the form of (1) with $\mathcal{N}-u=l$ and $1-p$ replacing p. Therefore, for a particular n and p, u is found by $u=\mathcal{N}-l$, where l is the value which has already been calculated for n, $1-p$.

The efficiency, E_2, is calculated on the same assumptions associated with E_1. If all individuals with x values greater than x_u are retained, then $\mathrm{Av}(w)=q_2=u/\mathcal{N}$. Now $p>q_2$, and hence, on the average, it will be necessary to obtain the required proportion p by adding $(p-q_2)$ individuals from the group where $x_l<x<x_u$. On the assumption of normality, the expected selection differential of the group selected in this manner is

$$\Delta_2 = \frac{1}{p}\left\{\mathrm{Z}(q_2) + \frac{(p-q_2)[\mathrm{Z}(q_1)-\mathrm{Z}(q_2)]}{q_1-q_2}\right\}$$

and

$$E_2 = \Delta_2/\Delta = \frac{(p-q_2)\mathrm{Z}(q_1)-(p-q_1)\mathrm{Z}(q_2)}{q_1-q_2}$$

As previously, the practical application of this procedure requires that p and E_2 be set in advance so that the appropriate n can be found from the table. A random sample is then drawn from which x_l and x_u are obtained. The remainder of the population may now be measured and put into three groups:

(A) individuals with measurements less than x_l (discards),
(B) individuals with measurements less than x_u and greater than x_l (reserves),
(C) individuals with measurements greater than x_u (retained).

If this is done, then l and u have been determined in such a way that nine times out of ten some individuals from group B will have to be taken at random and added to group C to bring the numbers of selected individuals to requirement. In the odd case when this is not so, adjustment must be made by either discarding animals at random from C when C is too large, or by adding animals at random from group A when the combined numbers of groups B and C are too small.

As an illustration, take the sample of 75 individuals previously considered. By retaining all above the individual 38th in rank, and putting those between the 38th and 52nd rank temporarily in reserve for later adjustment, the efficiency of the procedure is increased from 80% to 97%.

The choice between the two methods of selection presented here must be made on practical grounds. If it is inconvenient to break the population into three groups, then the first method must be used with its associated lower efficiency. However, whenever three groups can be handled satisfactorily, the high efficiency of the alternative procedure makes it the obvious choice.

In the particular example of selection for fleece weight mentioned in the introduction, a great reduction in clerical and physical labour can be expected by the use of either methods. The tedious procedure of writing 'roll call' lists and the even more unpleasant task of roll calling large flocks are avoided and this may represent a saving of a considerable number of man-days. The choice of which estimation procedure to use may be determined by the type of facilities and labour available.

As a final point, it is interesting to consider the relation of n to the efficiencies E_1 and E_2. These efficiency ratios simply reflect changes in the average selection differential as a result of taking larger samples; in the particular case of sheep selection, by the application of rather elementary genetical arguments, it is possible to approximate the annual loss of income due to the use of any selection method for any n from the efficiencies listed in Table I. Since it is suggested that individuals in the sample be tagged and identified with their scores, it would be possible to draw up cost curves by plotting economic gains against n. It would only be necessary to cost the labour of tagging and sorting 25, 50, 75, . . . sheep, and to take into account annual loss of income due to the statistical inefficiency of the method in order to draw up such a cost chart. From the graph it would then be possible to determine the optimum selection procedure to adopt.

Acknowledgements

I am indebted to Miss Elaine Smith for her assistance with the pre-

paration of Table I, and should like to acknowledge the helpful advice of Mr H. Weiler of the Division of Mathematical Statistics.

REFERENCE

FISHER, R. A. and CORNISH, E. A. (1960). 'The percentile points of distributions having known cumulants', *Technometrics*, **2**, 209.

TABLE I

Values of l, u, p_1, E_1, and E_2 for various combinations of p and n

p	0·1					0·2					0·3				
n	l	u	p_1	E_1	E_2	l	u	p_1	E_1	E_2	l	u	p_1	E_1	E_2
25	6	1	0·375	75·1	87·5	9	2	0·504	76·2	88·2	12	4	0·621	74·2	90·1
50	9	2	0·270	83·6	92·3	15	6	0·403	83·7	95·1	21	10	0·526	81·5	95·2
75	12	4	0·231	87·0	95·7	21	10	0·364	86·5	96·7	30	17	0·488	84·1	97
100	16	6	0·221	87·0	95·9	27	14	0·342	88·0	97·4	38	23	0·456	87·1	97
150	22	10	0·196	89·4	97·5	39	23	0·318	89·5	98·1	55	36	0·430	89·0	98·3
200	28	14	0·181	90·8	98·1	50	31	0·300	91·0	98·6	71	50	0·409	90·8	98·9
500	62	40	0·149	94·1	99·4	115	86	0·261	94·3	99·6	167	134	0·368	94·2	99·6
1000	116	85	0·133	95·9	99·7	221	180	0·243	96·0	99·8	324	277	0·348	95·7	99·8
p	0·4					0·5					0·6				
n	l	u	p_1	E_1	E_2	l	u	p_1	E_1	E_2	l	u	p_1	E_1	E_2
25	15	7	0·730	70·3	92·2	17	9	0·798	70·7	92·5	19	11	0·860	70·1	92·2
50	26	15	0·624	81·0	96·1	31	20	0·717	79·2	96·3	36	25	0·805	75·8	96·1
75	33	24	0·594	82·6	97·2	45	31	0·683	82·2	97·3	52	38	0·768	80·8	97·2
100	49	33	0·567	85·1	97·9	59	42	0·664	83·7	97·8	68	52	0·748	83·3	97·9
150	70	51	0·530	88·7	98·7	86	65	0·635	86·4	98·5	100	81	0·724	85·8	98·7
200	92	69	0·516	89·7	98·9	112	89	0·614	88·9	99·0	132	109	0·711	86·9	98·9
500	219	182	0·474	93·3	99·5	269	232	0·573	92·7	99·6	318	282	0·670	91·9	99·6
1000	426	375	0·451	95·3	99·8	526	475	0·551	94·9	99·8	626	575	0·650	94·2	99·8
p	0·7					0·8					0·9				
n	l	u	p_1	E_1	E_2	l	u	p_1	E_1	E_2	l	u	p_1	E_1	E_2
25	22	14	0·943	56·5	90·1	24	17	0·985	44·7	88·2	25	20	0·998	44·1	87·5
50	41	30	0·887	69·3	95·2	45	36	0·946	64·1	95·1	49	42	0·993	45·1	92·3
75	59	46	0·850	77·6	97·5	66	55	0·926	70·4	96·7	72	64	0·982	58·5	95·7
100	78	63	0·837	78·8	97·7	87	74	0·913	73·5	97·4	95	85	0·974	64·1	95·9
150	115	96	0·816	81·8	98·3	128	112	0·893	79·3	98·1	141	129	0·963	70·5	97·
200	151	130	0·800	85·0	98·9	170	151	0·886	80·1	98·6	187	173	0·957	74·0	98·1
500	367	334	0·761	90·6	99·6	415	386	0·855	88·0	99·6	461	439	0·939	82·9	99·4
1000	724	677	0·746	93·3	99·8	821	780	0·840	91·4	99·8	916	885	0·929	87·2	99·7

p = proportion of individuals to be retained.
n = number of individuals which are measured, their measurements then being ranked in order of merit as $x_1, x_2, \ldots x_u, \ldots x_l, \ldots x_n$.
l = rank whose measurement (x_l) is used to split the population into two groups.
u = rank whose measurement (x_u) is used to split the population into three groups. In this case u and l define the limits of the reserve group.
p_1 = upper 95% limit of proportion actually retained using the two group method.
E_1 = efficiency (%) of two group method at average truncation point.
E_2 = efficiency (%) of three group method at average truncation point.

Printed in Great Britain by
ROBERT CUNNINGHAM AND SONS LTD.
Alva

A (a) (ii) [2]

The Moment Generating Function of the Truncated Multi-normal Distribution

BY

G. M. TALLIS

Reprinted from
THE JOURNAL OF THE ROYAL STATISTICAL SOCIETY,
SERIES B (METHODOLOGICAL)
Volume 23, No. 1, 1961
(pp. 223 - 229)

PRINTED FOR PRIVATE CIRCULATION
1961

The Moment Generating Function of the Truncated Multi-normal Distribution

By G. M. Tallis

Division of Animal Genetics, C.S.I.R.O., Glebe, N.S.W.

[Received December 1960]

Summary

In this paper the moment generating function (m.g.f.) of the truncated n-dimensional normal distribution is obtained. From the m.g.f., formulae for $E(X_i)$ and $E(X_i X_j)$ are derived, and are used to investigate certain special cases. Some applications of these results to statistical genetics are also discussed.

1. Introduction

The problem of finding the means, variances and covariances of a standardized n-dimensional normal distribution (here abbreviated to standard n-normal) truncated in $p \leqslant n$ coordinates was solved by Birnbaum and Meyer (1953). The solutions were obtained by direct integration and the general results left in a somewhat difficult form for explicit evaluation.

It is the purpose of this paper to present a different method of solving the same problem. Since the moment generating function (m.g.f.) approach is used, the required moments are obtained by differentiation rather than integration. General formulae for computing $E(X_i)$ and $E(X_i X_j)$ $(i, j = 1, 2, \ldots, n)$ are given as well as explicit formulae for the same moments for the special case $n = 3$. Two examples are used to illustrate the methods of evaluating the general formulae.

2. Notation

It is convenient in the following development to let ϕ represent the frequency function of an arbitrary number of standardized normal variates. Thus, if X_s $(s = 1, 2, \ldots, n)$ are n such variates with correlation matrix R (assumed positive definite), we have

$$\phi_n(x_1, x_2, \ldots, x_n; \mathrm{R}) = \phi_n(x_s; \mathrm{R}) = (2\pi)^{-n/2} |\mathrm{R}|^{-\frac{1}{2}} \exp\{-\tfrac{1}{2}\mathbf{x}' \mathrm{R}^{-1} \mathbf{x}\}, \qquad (1)$$

where x is the column vector of the X_s. This distribution for $X_q = b_q$ and $X_q = b_q$, $X_r = b_r$ may be written

$$\phi_n(x_s, x_q = b_q; \mathrm{R}) = \phi(b_q)\, \phi_{n-1}(y_s; \mathrm{R}_q) \quad (s \neq q),$$

$$\phi_n(x_s, x_q = b_q, x_r = b_r; \mathrm{R}) = \phi(b_q, b_r; \rho_{qr})\, \phi_{n-2}(z_s; \mathrm{R}_{qr}) \quad (s \neq q \neq r),$$

where R_q and R_{qr} are the matrices of first- and second-order partial correlation coefficients of X_s for $s \neq q$, and for $s \neq q$, $s \neq r$ respectively, and

$$Y_s = (X_s - \rho_{qs} b_q)/\sqrt{(1-\rho_{qs}^2)},$$

$$Z_s = (X_s - \beta_{sq.r} b_q - \beta_{sr.q} b_r)/\sqrt{\{(1-\rho_{sq}^2)(1-\rho_{sr.q}^2)\}}.$$

In the above formulae $\beta_{sq.r}$ and $\beta_{sr.q}$ are the partial regression coefficients of X_s on X_q and X_r respectively and $\rho_{sr.q}$ is the partial correlation coefficient between X_s and X_r for fixed X_q.

Now, if the operator

$$\int_{b_1}^{\infty} \cdots \int_{b_n}^{\infty} (\quad)\, dx_1 \ldots dx_n$$

is abbreviated to

$$^{(n)}\int_{b_s}^{\infty} (\quad)\, dx_s,$$

and if we let

$$\Phi_n(b_s; \mathrm{R}) = {}^{(n)}\int_{b_s}^{\infty} \phi_n(x_s; \mathrm{R})\, dx_s,$$

then it follows from the above formulae that

$$^{(n-1)}\int_{b_s}^{\infty} \phi_n(x_s, x_q = b_q; \mathrm{R})\, dx_s = \phi(b_q)\, \Phi_{n-1}(B_{qs}; \mathrm{R}_q) \quad (s \neq q)$$

and

$$^{(n-2)}\int_{b_s}^{\infty} \phi_n(x_s, x_q = b_q, x_r = b_r; \mathrm{R})\, dx_s = \phi(b_q, b_r;\, \rho_{qr})\, \Phi_{n-2}(B_{rs}^q; \mathrm{R}_{qr}) \quad (s \neq q \neq r),$$

where

$$B_{qs} = (b_s - \rho_{qs} b_q)/\sqrt{(1-\rho_{qs}^2)},$$

$$B_{rs}^q = (b_s - \beta_{sq.r} b_q - \beta_{sr.q} b_r)/\sqrt{\{(1-\rho_{sq}^2)(1-\rho_{sr.q}^2)\}}.$$

3. General Results

Let W_s $(s = 1, 2, \ldots, n)$ have the standard n-normal distribution with correlation matrix R and let W_s be truncated at a_s so that

$$\alpha = \text{prob}\,(W_1 > a_1, W_2 > a_2, \ldots, W_n > a_n) = \Phi_n(a_s; \mathrm{R}).$$

The joint m.g.f. of the truncated population $W_1 > a_1, W_2 > a_2, \ldots, W_n > a_n$ is

$$m(t_s) = m = \alpha^{-1}\, {}^{(n)}\int_{a_s}^{\infty} e^{t'w}\, \phi_n(w_s; \mathrm{R})\, dw_s$$

$$= \alpha^{-1}(2\pi)^{-n/2} |\mathrm{R}|^{-\frac{1}{2}}\, {}^{(n)}\int_{a_s}^{\infty} \exp\left[-\tfrac{1}{2}\{w'\,\mathrm{R}^{-1}\,w - 2t'\,w\}\right] dw_s,$$

where t is the column vector of the t_s $(s = 1, 2, \ldots, n)$. Now the identity

$$-\tfrac{1}{2}\{w'\,\mathrm{R}^{-1}\,w - 2t'\,w\} \equiv \tfrac{1}{2}t'\,\mathrm{R}t - \tfrac{1}{2}(w-\zeta)'\,\mathrm{R}^{-1}(w-\zeta)$$

is easily verified by noticing that $\zeta = \mathrm{Rt}$ and then expanding the right-hand side. It can then be shown that the above integral for m may be written

$$m = \alpha^{-1}(2\pi)^{-n/2}|\mathrm{R}|^{-\frac{1}{2}} e^{T} \int_{a_s}^{\infty} \exp\{-\tfrac{1}{2}(\mathrm{w}-\zeta)'\,\mathrm{R}^{-1}(\mathrm{w}-\zeta)\}\, dw_s,$$

where $T = \frac{1}{2}\mathrm{t}'\,\mathrm{Rt}$, t and $\mathrm{w}-\zeta$ are column vectors of t_s and $W_s-\zeta_s$, and $\zeta_s = \Sigma\rho_{sv}t_v$. By the change of variables $X_s = W_s-\zeta_s$, we obtain immediately

$$\alpha m = e^{T}\Phi_n(b_s;\ \mathrm{R}) \quad (b_s = a_s-\zeta_s). \tag{2}$$

With the results of section 2, equation (2) may be readily differentiated, first with respect to t_i and then with respect to t_j. It can be verified that, when the derivatives are calculated with all $t_s = 0$,

$$\alpha\frac{\partial m}{\partial t_i} = \alpha E(X_i) = \sum_{q=1}^{n} \rho_{iq}\phi(a_q)\Phi_{n-1}(A_{qs};\ \mathrm{R}_q), \tag{3}$$

$$\alpha\frac{\partial^2 m}{\partial t_j\,\partial t_i} = \alpha E(X_i X_j) = \rho_{ij}\alpha + \sum_{q=1}^{n} \rho_{qi}\rho_{qj}a_q\phi(a_q)\Phi_{n-1}(A_{qs};\ \mathrm{R}_q)$$

$$+ \sum_{q=1}^{n}\left\{\rho_{qi}\sum_{r\neq q}\phi(a_q, a_r;\ \rho_{qr})\Phi_{n-2}(A^q_{rs};\ \mathrm{R}_{qr})(\rho_{rj}-\rho_{qr}\rho_{qj})\right\}, \tag{4}$$

where

$$A_{qs} = (a_s-\rho_{sq}a_q)/\sqrt{\{1-\rho^2_{sq}\}},$$

$$A^q_{rs} = (a_s-\beta_{sq.r}a_q-\beta_{sr.q}a_r)/\sqrt{\{(1-\rho^2_{sq})(1-\rho^2_{sr.q})\}}$$

and $s\neq q$ in Φ_{n-1} and $s\neq q\neq r$ in Φ_{n-2}.

The expressions (3) and (4) necessitate the evaluation of such integrals as $\Phi_n(a_s;\ \mathrm{R})$. These integrals have been tabulated for $n = 1$ and $n = 2$ (Pearson, 1931), $n = 2$ (Owen, 1956). For $n \geqslant 3$, see Plackett (1954) and Steck (1958).

Some special cases for $E(X_i)$ and $E(X_i X_j)$ arise when certain $a_s = -\infty$. In these instances, the appropriate modifications to (3) and (4) may be obtained by noticing that:

(a) all terms involving ϕ, where ϕ is a function of any $a_s = -\infty$, are zero since $\phi = 0$;

(b) by definition, if $a_s = -\infty$, then $A_{qs} = A^q_{rs} = -\infty$. Hence all integrals involving A_{qs} or A^q_{rs} have their dimension reduced by one for each negatively infinite parameter;

(c) obviously if $a_s = -\infty$, $\Phi(A_{qs}) = \Phi(A^q_{rs}) = 1$.

4. Special Cases in Two and Three Dimensions

In order to illustrate the use of expressions (3) and (4), $E(X_1)$, $E(X_1^2)$ and $E(X_1 X_2)$ will be evaluated for the special case $n = 3$. The expression for $E(X_1)$ is obtained by setting $i = 1$ in (3), and $E(X_1^2)$ and $E(X_1 X_2)$ are obtained by setting $i = j = 1$ and $i = 1, j = 2$ in (4) respectively.

The results are

$$\alpha E(X_1) = \phi(a_1)\,\Phi(A_{12}, A_{13};\ \rho_{23.1}) + \rho_{12}\,\phi(a_2)\,\Phi(A_{21}, A_{23};\ \rho_{13.2}) + \rho_{13}\,\phi(a_3)\,\Phi(A_{31}, A_{32};\ \rho_{12.3}),$$

$$\begin{aligned}\alpha E(X_1^2) = \alpha &+ a_1\,\phi(a_1)\,\Phi(A_{12}, A_{13};\ \rho_{23.1}) + \rho_{12}^2\,a_2\,\phi(a_2)\,\Phi(A_{21}, A_{23};\ \rho_{13.2}) \\ &+ \rho_{13}^2\,a_3\,\phi(a_3)\,\Phi(A_{31}, A_{32};\ \rho_{12.3}) + \rho_{12}(1-\rho_{12}^2)\,\phi(a_1, a_2;\ \rho_{12})\,\Phi(A_{13}^2) \\ &+ \rho_{13}(1-\rho_{13}^2)\,\phi(a_1, a_3;\ \rho_{13})\,\Phi(A_{12}^3) \\ &+ \phi(a_2, a_3;\ \rho_{23})\{\Phi(A_{31}^2)\,\rho_{12}(\rho_{13}-\rho_{12}\rho_{23}) \\ &+ \Phi(A_{21}^3)\,\rho_{13}(\rho_{12}-\rho_{23}\rho_{13})\},\end{aligned}$$

$$\begin{aligned}\alpha E(X_1 X_2) = \alpha\rho_{12} &+ \rho_{12}\,a_1\,\phi(a_1)\,\Phi(A_{12}, A_{13};\ \rho_{23.1}) \\ &+ \rho_{12}\,a_2\,\phi(a_2)\,\Phi(A_{21}, A_{23};\ \rho_{13.2}) \\ &+ \rho_{13}\rho_{23}\,a_3\,\phi(a_3)\,\Phi(A_{31}, A_{32};\ \rho_{12.3}) + (1-\rho_{12}^2)\,\phi(a_1, a_2;\ \rho_{12})\,\Phi(A_{23}^1) \\ &+ \rho_{13}(1-\rho_{23}^2)\,\phi(a_2, a_3;\ \rho_{23})\,\Phi(A_{21}^3) \\ &+ \phi(a_1, a_3;\ \rho_{13})\{(\rho_{23}-\rho_{13}\rho_{12})\,\Phi(A_{32}^1) + \rho_{13}(\rho_{12}-\rho_{13}\rho_{23})\,\Phi(A_{12}^3)\},\end{aligned}$$

where A_{qs} and A_{rs}^q are as defined in the previous section. If now

$$a_3 = A_{13}^2 = A_{23}^1 = A_{13} = A_{23} = -\infty,$$

the first and second moments of the truncated standard bi-normal distribution are obtained. These formulae agree with those presented by Weiler (1959).

5. Applications

Example 1. Young and Weiler (1961) have considered the case of the selection of animals (or plants) by the method of independent culling levels, using two bi-normally distributed characters W_1 and W_2, with frequency function

$$N(\mu_1, \mu_2, P_{11}, P_{22}, \rho_p).$$

This technique involves the simultaneous truncation of W_1 and W_2 at p_1 and p_2 in such a manner that $\text{prob}(W_1 > p_1, W_2 > p_2) = \alpha$. From their formulae for the first moments of the truncated bivariate distribution, it is possible to compute the phenotypic advance due to selection. However, it is also of interest to calculate the total genetic gains.

In order to make further progress with the latter problem, we assume the usual genetic models $W_1 = G_1 + E_1$ and $W_2 = G_2 + E_2$, where the G_i and E_i are the additive genetic and environmental contributions to phenotype respectively. The components G_i and E_i $(i = 1, 2)$ of the models are assumed to be independently and normally distributed. In this notation, the genetic value of an animal, relative to the population, may be defined as

$$G = \gamma_1\{G_1 - E(G_1)\} + \gamma_2\{G_2 - E(G_2)\},$$

where γ_1 and γ_2 are the economic weights for W_1 and W_2. Now let the variance of G be σ_G^2, let $X_1 = (W_1-\mu_1)/\sqrt{P_{11}}$, $X_2 = (W_2-\mu_2)/\sqrt{P_{22}}$ and $X_3 = G/\sigma_G$. Then X_i $(i = 1, 2, 3)$ are assumed to have a standard tri-normal distribution with correlation coefficients,

$$\rho_{12} = \rho_p, \rho_{13} = \frac{\gamma_1 V(G_1)+\gamma_2 C(G_1, G_2)}{\sigma_G\sqrt{P_{11}}}, \quad \rho_{23} = \frac{\gamma_1 C(G_1, G_2)+\gamma_2 V(G_2)}{\sigma_G\sqrt{P_{22}}}.$$

Here ρ_p is the phenotypic correlation between W_1 and W_2 and V and C denote variance and covariance. Since the truncation points of X_1, X_2 and X_3 are given by

$$a_1 = (p_1-\mu_1)/\sqrt{P_{11}}, \quad a_2 = (p_2-\mu_2)/\sqrt{P_{22}} \quad \text{and} \quad a_3 = -\infty \text{ respectively,}$$

it is possible to deduce from the formula for $E(X_1)$ in section 4 (by symmetry) that

$$\alpha E(X_3) = \rho_{13}\phi(a_1)\Phi(A_{12})+\rho_{23}\phi(a_2)\Phi(A_{21}).$$

From the definition of X_3 it is clear that $E(G) = \sigma_G E(X_3)$. It can be shown that this result is algebraically equivalent to the results obtained by Young and Weiler (1961) by a method analogous to linear interpolation. With the aid of the formula for $E(X_1^2)$, it can also be verified that

$$\alpha E(X_3^2) = 1+\rho_{23}^2 a_2\phi(a_2)\Phi(A_{21})+\rho_{13}^2 a_1\phi(a_1)\Phi(A_{12}) + \{2\rho_{23}\rho_{13}-\rho_{12}(\rho_{23}^2+\rho_{13}^2)\}\phi(a_1, a_2; \rho_{12}).$$

Therefore, the new variance of G, $\sigma_{G'}^2$, is

$$\sigma_{G'}^2 = \sigma_G^2[E(X_3^2)-\{E(X_3)\}^2].$$

Now, if a sample of N animals is taken from the truncated population $W_1 > b_1$, $W_2 > b_2$, then $\bar{G} = \Sigma G/N$. Although G cannot be measured directly, by virtue of the Central Limit Theorem we have for N sufficiently large

$$\text{prob}\{E(G)-t_\beta\sigma_{G'}/\sqrt{N} < \bar{G} < E(G)+t_\beta\sigma_{G'}/\sqrt{N}\} \simeq 1-\beta,$$

where t_β is the standard normal deviate corresponding to the 100β per cent., two-tailed probability level. Thus, although in practice the required parameters for calculating $E(G)$ and $\sigma_{G'}^2$ have to be estimated, it is possible to obtain some idea of the interval in which $\bar{G}$ is expected to lie with given probability.

Example 2. As a final illustration of these methods, consider the n variables $W_s = Y_s+Z_s$ $(s = 1, 2, \ldots, n)$, where the Y_s and Z_s are normally and independently distributed with zero expectations. Now, if all $W_s < a_s\{V(W_s)\}^{\frac{1}{2}}$ are discarded, it may be of interest to investigate the changes in the means, variances and covariances of the $2n$ variables W_s and Y_s.

In order to proceed with the problem, it is convenient to let

$$W_1/[V(W_1)]^{\frac{1}{2}} = X_1, \ldots, W_n/[V(W_n)]^{\frac{1}{2}} = X_n, Y_1/[V(Y_1)]^{\frac{1}{2}} = X_{n+1}, \ldots, Y_n/[V(Y_n)]^{\frac{1}{2}} = X_{2n}$$

and let R be the $2n \times 2n$ correlation matrix of X_s $(s = 1, 2, \ldots, 2n)$. Then it is possible to write R as the partitioned matrix

$$\mathbf{R} = \begin{bmatrix} \mathbf{K} & \mathbf{L} \\ \mathbf{M} & \mathbf{N} \end{bmatrix}.$$

where K, L, M and N are $n \times n$. We thus have for $s, t = 1, 2, \ldots, n$

$$\mathbf{K} = [k_{st}] = \left[\frac{C(W_s W_t)}{\{V(W_s)\, V(W_t)\}^{\frac{1}{2}}}\right], \qquad \mathbf{L} = [l_{st}] = \left[\frac{C(Y_s Y_t)}{\{V(W_s)\, V(Y_t)\}^{\frac{1}{2}}}\right],$$

$$\mathbf{M} = [m_{st}] = \left[\frac{C(Y_s Y_t)}{\{V(Y_s)\, V(W_t)\}^{\frac{1}{2}}}\right], \qquad \mathbf{N} = [n_{st}] = \left[\frac{C(Y_s Y_t)}{\{V(Y_s)\, V(Y_t)\}^{\frac{1}{2}}}\right].$$

With the above notation and the rules for special cases, it is now possible to write down formulae for $E(X_i)$ and $E(X_i X_j)$, remembering $a_s = -\infty$ $(n < s \leqslant 2n)$. For $i, j = 1, 2, \ldots, 2n$, we have that

$$\alpha E(X_i) = \sum_{q=1}^{n} \rho_{iq}\, \phi(a_q)\, \Phi_{n-1}(A_{qs};\, \mathrm{K}_q),$$

$$\alpha E(X_i X_j) = \rho_{ij}\, \alpha + \sum_{q=1}^{n} \rho_{qi}\, \rho_{qj}\, a_q\, \phi(a_q)\, \Phi_{n-1}(A_{qs};\, \mathrm{K}_q) + \sum_{q=1}^{n} \left(\rho_{qi} \sum_{\substack{r=q \\ \leqslant n}} \phi(a_q, a_r;\, \rho_{qr})\, \Phi_{n-2}(A^{q}_{rs};\, \mathrm{K}_{qr})\, (\rho_{rj} - \rho_{qr}\, \rho_{qj}) \right),$$

where $s \neq q$ in Φ_{n-1}, $s \neq q \neq r$ in Φ_{n-2} and $s \leqslant n$ in all cases.

As a particular illustration let W_1 and W_2 be two phenotypic characters (as in Example 1), then Y_1 and Y_2 represent the additive genetic contributions and Z_1 and Z_2 the environmental contributions to phenotype respectively. In this instance

$$\mathbf{R} = \left[\begin{array}{cc:cc} 1 & \rho_p & h_1 & h_1 \rho_g \\ \rho_p & 1 & h_2 \rho_g & h_2 \\ \hdashline h_1 & h_2 \rho_g & 1 & \rho_g \\ h_1 \rho_g & h_2 & \rho_g & 1 \end{array}\right],$$

where ρ_p and ρ_g are the phenotypic and genetic correlations between the two characters and $h_1 = \{V(Y_1)/V(W_1)\}^{\frac{1}{2}}$ and $h_2 = \{V(Y_2)/V(W_2)\}^{\frac{1}{2}}$. In this case we have, for $i, j = 1, 2, 3, 4$, that

$$\alpha E(X_i) = \rho_{i1}\, \phi(a_1)\, \Phi(A_{12}) + \rho_{i2}\, \phi(a_2)\, \Phi(A_{21}),$$

$$\alpha E(X_i X_j) = \rho_{ij}\, \alpha + \rho_{1i}\, \rho_{1j}\, a_1\, \phi(a_1)\, \Phi(A_{12}) + \rho_{2i}\, \rho_{2j}\, a_2\, \phi(a_2)\, \Phi(A_{21}) + \phi(a_1, a_2;\, \rho_{12}) \{\rho_{1i}(\rho_{2j} - \rho_{12}\, \rho_{1j}) + \rho_{2i}(\rho_{1j} - \rho_{12}\, \rho_{2j})\}.$$

Therefore, it is clear from the last results that, by evaluating $E(X_i)$ and $E(X_i X_j)$ for appropriate i, j, it is possible to study the effects of phenotypic truncation on heritability, h_i^2, and genetic correlation, ρ_g. Moreover, the work required to accomplish this for two characters is relatively small and can be completed with the aid of existing tables for the bivariate normal distribution.

6. Extensions

The methods of section 3 may be used to investigate certain additional problems related to the truncation of multi-normally distributed variates. For instance, the evaluation when all $t_s = 0$ of $\partial^3 m / \partial t_i^3$ and $\partial^4 m / \partial t_i^4$ would provide the third and fourth

moments of the marginal distributions of the X_i. Moreover, a further generalization is achieved by considering the X_s $(s = 1, 2, \ldots, n)$ as doubly truncated so that $\text{prob}(a_1 \leqslant X_1 \leqslant c_1, \ldots, a_n \leqslant X_n \leqslant c_n) = \alpha$. In this case the required m.g.f. is

$$\alpha m = e^{T^{(n)}} \int_{b_s}^{d_s} \phi(x_s; \mathrm{R})\, dx_s,$$

where $$b_s = a_s - \sum_{r=1}^{n} \rho_{sr} t_r \quad \text{and} \quad d_s = c_s - \sum_{r=1}^{n} \rho_{sr} t_r.$$

For example, the m.g.f. for the bi-normal distribution under double truncation is

$$\alpha m = e^T \int_{b_1}^{d_1} \int_{b_2}^{d_2} \phi(x_1, x_2; \rho_{12})\, dx_1\, dx_2$$

which may be written with advantage

$$\alpha m = e^T \{\Phi(b_1, b_2; \rho_{12}) + \Phi(d_1, d_2; \rho_{12}) - \Phi(d_1, b_2; \rho_{12}) - \Phi(b_1, d_2; \rho_{12})\}$$

and hence it is clear that

$$\alpha m = \alpha_1 m_1 + \alpha_2 m_2 - \alpha_3 m_3 - \alpha_4 m_4,$$

where the subscripts 1, 2, 3 and 4 refer to the bi-normal distribution truncated at $(a_1, a_2), (c_1, c_2), (c_1, a_2)$ and (a_1, c_2) respectively. The first and second moments may now be obtained in an obvious way from the formulae of Weiler (1959).

ACKNOWLEDGEMENT

The author is extremely grateful to Dr D. J. Finney and Dr R. N. Curnow for valuable suggestions in connection with this work.

REFERENCES

BIRNBAUM, Z. W. and MEYER, P. L. (1953), "On the effect of truncation in some or all co-ordinates of a multi-normal population", *J. Indian Soc. agric. Statist.*, 5, 17–28.

OWEN, D. B. (1956), "Tables for computing bivariate normal probabilities", *Ann. math. Statist.*, **27**, 1075–1090.

PEARSON, K. (1931), *Tables for Statisticians and Biometricians*. Cambridge University Press.

PLACKETT, R. L. (1954), "A reduction formula for normal multivariate integrals", *Biometrika*, **41**, 351–360.

STECK, G. P. (1958), "A table for computing trivariate normal probabilities", *Ann. math. Statist.*, **29**, 780–800.

WEILER, H. (1959), "Means and standard deviations of a truncated normal bivariate distribution", *Aust. J. Statist.*, **1**, 73–81.

YOUNG, S. S. Y. and WEILER, H. (1961), "Selection for two correlated traits by independent culling levels", *J. Genet.*, **58** (to be published).

A (a) (ii) [3]

Plane Truncation in Normal Populations

BY

G. M. TALLIS

Reprinted from
THE JOURNAL OF THE ROYAL STATISTICAL SOCIETY
SERIES B (METHODOLOGICAL)
Volume 27, No. 2, 1965
(pp. 301–307)

PRINTED FOR PRIVATE CIRCULATION
1965

Plane Truncation in Normal Populations

By G. M. TALLIS

The Johns Hopkins University

[Received December 1964. Revised March 1965]

SUMMARY

This paper considers the truncation of normal distributions by means of planes. The moment-generating function for the truncated distribution is obtained, and it is shown that, by suitable transformations, the problem reduces to the case of rectangular truncation.

1. INTRODUCTION

THE problem of subjecting n-dimensional normal populations to arbitrary rectangular truncation has been considered by Birnbaum and Meyer (1953), Tallis (1961) and Finney (1962). However, rectangular truncation is not the only type of truncation procedure which is of practical interest. For instance, when animals are chosen for breeding purposes, selection is often made by means of a linear compound of variables, or index. Such a selection procedure has been shown to be optimal from the point of view of maximizing genetic gains, suitably defined. Other situations where selection is based on some linear combination of random variables can also be envisaged, and the purpose of this paper is to investigate the effects of such truncation on the moments of the original joint distribution of the variables, given that this distribution is multivariate normal.

The system of notation adopted in Tallis (1961) will be closely followed here. Briefly, $\phi_n(\mathrm{x}; \mathrm{R})$ will be used to specify the frequency function of n standardized normal variates with correlation matrix R, and $\Phi_n(\mathrm{b}; \mathrm{R})$ is defined by

$$\Phi_n(\mathrm{b}; \mathrm{R}) = \int_{b_1}^{\infty}\int_{b_2}^{\infty}\cdots\int_{b_n}^{\infty}\phi_n(\mathrm{x}; \mathrm{R})\,dx_1\,dx_2\ldots dx_n = \int_{\mathrm{b}}^{\infty}\phi_n(\mathrm{x}; \mathrm{R})\,d\mathrm{x}.$$

It was shown in the earlier paper that if the random variables $X_1, X_2, \ldots, X_n$ are subjected to the rectangular truncation $X_1 \geqslant a_1, X_2 \geqslant a_2, \ldots, X_n \geqslant a_n$, then the moment-generating function (m.g.f.) of the truncated distribution is given by

$$\alpha m(\mathrm{t}) = e^{T}\Phi_n(\mathrm{b}; \mathrm{R}) \tag{1}$$

where

$$\alpha = \int_{\mathrm{a}}^{\infty}\phi_n(\mathrm{x}; \mathrm{R})\,d\mathrm{x},$$

$T = \frac{1}{2}\mathrm{t}'\mathrm{R}\mathrm{t}$ and $\mathrm{b} = \mathrm{a} - \mathrm{R}\mathrm{t}$. Equation (1) was then used to obtain the first- and second-order moments for the truncated distribution.

In the present paper, the special case of truncation by a single plane will be considered first and it will be shown that this situation can be simply reduced to rectangular truncation. An example which utilizes these results will then be discussed. Subsequently, the more general question of truncation by means of $q \leqslant n$ planes is considered and these results applied to two special cases.

2. METHODS

In the case of rectangular truncation, the appropriate truncation set is specified by

$$A_n(\mathbf{x}) = \{\mathbf{x};\ x_1 \geqslant a_1, x_2 \geqslant a_2, \ldots, x_n \geqslant a_n\}, \tag{2}$$

or

$$A_n(\mathbf{x}) = \{\mathbf{x};\ a_1 \leqslant x_1 \leqslant b_1, a_2 \leqslant x_2 \leqslant b_2, \ldots, a_n \leqslant x_n \leqslant b_n\}. \tag{3}$$

Obviously, (2) is the special case of (3) with all $b_i = \infty$. Under plane truncation, we consider first the set

$$A_n(\mathbf{x}) = \{\mathbf{x};\ \mathbf{c}'\mathbf{x} \geqslant p\} \tag{4}$$

where $\mathbf{c}$ is a normalized vector of coefficients which may be regarded as the direction cosines of the normal to the plane $\mathbf{c}'\mathbf{x} = p$. The more general case is

$$A_n(\mathbf{x}) = \{\mathbf{x};\ \mathbf{c}_1'\mathbf{x} \geqslant p_1, \mathbf{c}_2'\mathbf{x} \geqslant p_2, \ldots, \mathbf{c}_q'\mathbf{x} \geqslant p_q\} \tag{5}$$

or

$$A_n(\mathbf{x}) = \{\mathbf{x};\ p_{11} \leqslant \mathbf{c}_1'\mathbf{x} \leqslant p_{12}, p_{21} \leqslant \mathbf{c}_2'\mathbf{x} \leqslant p_{22}, \ldots, p_{q1} \leqslant \mathbf{c}_q'\mathbf{x} \leqslant p_{q2}\} \tag{6}$$

with similar definitions for the $\mathbf{c}_j$ $(j = 1, 2, \ldots, q)$. Notice that inequalities of the form $\mathbf{c}'\mathbf{x} \leqslant p$ can be converted to the type of inequality above by multiplying through by -1. The problem will be to find suitable linear transformations which will change (4) to (2) and (5) to (2) respectively. It is found that the same transformation which changes (5) to the form (2) also changes (6) to (3), so that no generality is lost by working with the set (5). This leads to some simplification in notation.

3. SINGLE PLANE TRUNCATION

Consider now the multinormal frequency function $\phi_n(\mathbf{x}; \mathrm{R})$ introduced above. We calculate first, with $A_n(\mathbf{x})$ specified by (4),

$$\alpha m(\mathbf{t}) = \int_{A_n(\mathbf{x})} e^{\mathbf{t}'\mathbf{x}} \phi_n(\mathbf{x}; \mathrm{R})\, d\mathbf{x} \tag{7}$$

$$= (2\pi)^{-\frac{1}{2}n} \int_{A_n(\mathbf{x})} e^{\mathbf{t}'\mathbf{x}} |\mathrm{R}|^{-\frac{1}{2}} \exp\{-\tfrac{1}{2}\mathbf{x}'\mathrm{R}^{-1}\mathbf{x}\}\, d\mathbf{x}.$$

Make the orthogonal transformation $\mathbf{x} = \mathrm{H}\mathbf{y}$ where H has $\mathbf{c}$ as its first column, the remaining columns being orthogonal to $\mathbf{c}$ and orthonormal amongst themselves. Then

$$\alpha m(\mathbf{t}) = K \int_{A_n(\mathbf{y})} \exp\{\mathbf{t}'\mathrm{H}\mathbf{y} - \tfrac{1}{2}\mathbf{y}'\mathrm{V}^{-1}\mathbf{y}\}\, d\mathbf{y},$$

where $\mathrm{V}^{-1} = \mathrm{H}'\mathrm{R}^{-1}\mathrm{H}$, $K = (2\pi)^{-\frac{1}{2}n}|\mathrm{R}|^{-\frac{1}{2}}$ and $A_n(\mathbf{y})$ is given by (2) with

$$a_2 = a_3, = \ldots = a_n = -\infty$$

and $a_1 = p$. Now let $\theta = \mathrm{H}'\mathbf{t}$ and use the identity

$$-\tfrac{1}{2}(\mathbf{y}'\mathrm{V}^{-1}\mathbf{y} - 2\theta'\mathbf{y}) \equiv \tfrac{1}{2}\theta'\mathrm{V}\theta - \tfrac{1}{2}(\mathbf{y} - \beta)'\mathrm{V}^{-1}(\mathbf{y} - \beta), \tag{8}$$

writing $\beta = \mathrm{V}\theta$, to obtain

$$\alpha m(\mathbf{t}) = K e^{T} \int_{A_n(\mathbf{z})} \exp\{-\tfrac{1}{2}\mathbf{z}'\mathrm{V}^{-1}\mathbf{z}\}\, d\mathbf{z} = e^{T}\, \Phi\left(\frac{p - \beta_1}{\gamma}\right), \tag{9}$$

where $\beta_1 = \mathbf{c}'\mathbf{R}\mathbf{t}, \gamma = (\mathbf{c}'\mathbf{R}\mathbf{c})^{\frac{1}{2}}$ and $A_u(\mathbf{z}) = \{\mathbf{z}; z_1 + \beta_1 \geqslant p\}$. Upon setting $\mathbf{t} = 0$, it is found that $\alpha = \Phi(p/\gamma)$.

The cumulant generating function is

$$k(\mathbf{t}) = T + \ln \Phi\left(\frac{p-\beta_1}{\gamma}\right) - \ln \alpha, \tag{10}$$

whence the mean vector, $\boldsymbol{\mu}$, is

$$\boldsymbol{\mu} = (\alpha\gamma)^{-1}\phi(p/\gamma)\mathbf{R}\mathbf{c} \tag{11}$$

and the dispersion matrix, $\mathbf{M}$, is found to be

$$\mathbf{M} = \mathbf{R} + \mathbf{R}\mathbf{c}\mathbf{c}'\mathbf{R}(\alpha\gamma^2)^{-1}\phi(p/\gamma)\{p/\gamma - \phi(p/\gamma)/\alpha\}. \tag{12}$$

It is important to notice that, by a suitable normalization of the coefficients, the above results apply to all cases of index selection. For suppose all individuals satisfying $\mathbf{a}'\mathbf{x} \geqslant b$ are to be retained, then by defining $c_i = a_i/\sqrt{(\mathbf{a}'\mathbf{a})}$, and $p = b/\sqrt{(\mathbf{a}'\mathbf{a})}$, this expression assumes the form $\mathbf{c}'\mathbf{x} \geqslant p$, where $\mathbf{c}'\mathbf{c} = 1$.

Example. We now apply the above methods to a problem in two dimensions. Let X_1 and X_2 have a bivariate normal distribution with means μ_1, μ_2, variances σ_1^2, σ_2^2 and correlation coefficient ρ, and form the ratio $Y = X_1/X_2$. Selection is to be practised for Y in such a way that a proportion α is retained and it is required to know the effect of this selection on marginal moments. It is also important to determine the correct value y_α of Y to use as a point of truncation so that the desired proportion of individuals is saved.

The first step is to write Y in terms of the standardized variables x_1 and x_2 in the equation $X_1 = X_2 y_\alpha$. Thus

$$x_1\sigma_1 + \mu_1 = y_\alpha(x_2\sigma_2 + \mu_2),$$

and

$$\sigma_1 x_1 - y_\alpha\sigma_2 x_2 = y_\alpha\mu_2 - \mu_1,$$

which can be put in the form

$$c_1 x_1 + c_2 x_2 = p,$$

with

$$c_1 = \sigma_1/(\sigma_1^2 + \sigma_2^2 y_\alpha^2)^{\frac{1}{2}}, \quad c_2 = -y_\alpha\sigma_2/(\sigma_1^2 + \sigma_2^2 y_\alpha^2)^{\frac{1}{2}},$$

and

$$p = (y_\alpha\mu_2 - \mu_1)/(\sigma_1^2 + \sigma_2^2 y_\alpha^2)^{\frac{1}{2}}.$$

Now $\Phi(p/\gamma) = \alpha$, and hence we set $p/\gamma = t_\alpha$, say, where $t_\alpha = \Phi^{-1}(\alpha)$. For this problem

$$\gamma^2 = \mathbf{c}'\mathbf{R}\mathbf{c} = (\sigma_1^2 - 2\rho y_\alpha\sigma_1\sigma_2 + y_\alpha^2\sigma_2^2)/(\sigma_1^2 + \sigma_2^2 y_\alpha^2),$$

and y_α can be found as a root of the equation $y_\alpha^2 T_{22} - 2y_\alpha T_{12} + T_{11} = 0$, where $T_{ii} = \sigma_i^2 - \mu_i^2/t_\alpha^2$ $(i = 1, 2)$, and $T_{12} = \rho\sigma_1\sigma_2 - \mu_1\mu_2/t_\alpha^2$.

The correct root may be determined by substitution in the equation $p = \gamma t_\alpha$. Suppose the correct value of γ is γ_α, then we have

$$E(X_1) = \frac{\sigma_1\phi(p/\gamma_\alpha)(\sigma_1 - y_\alpha\sigma_2\rho)}{\alpha\gamma_\alpha(\sigma_1^2 + \sigma_2^2 y_\alpha^2)^{\frac{1}{2}}} + \mu_1$$

$$E(X_2) = \frac{\sigma_2\phi(p/\gamma_\alpha)(\rho\sigma_1 - y_\alpha\sigma_2)}{\alpha\gamma_\alpha(\sigma_1^2 + \sigma_2^2 y_\alpha^2)^{\frac{1}{2}}} + \mu_2,$$

and

$$\mathbf{M} = \begin{bmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{bmatrix} + \alpha^{-1} \begin{bmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{bmatrix} \begin{bmatrix} \gamma_\alpha^2 \sigma_2^2 & \gamma_\alpha \sigma_1 \sigma_2 \\ \gamma_\alpha \sigma_1 \sigma_2 & \sigma_1^2 \end{bmatrix}^{-1} \begin{bmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{bmatrix}$$
$$\times \phi(p/\gamma_\alpha)\{p/\gamma_\alpha - \phi(p/\gamma_\alpha)/\alpha\}.$$

4. TRUNCATION WITH $q \leqslant n$ PLANES

The more general case of plane truncation is specified by (5), i.e.

$$A_n(\mathbf{x}) = \{\mathbf{x};\ \mathbf{C}'\mathbf{x} \geqslant \mathbf{p}\},$$

where $\mathbf{C}$ is the $n \times q$ matrix with columns $\mathbf{c}_j$ $(j = 1, 2, \ldots, q)$ and $\mathbf{p}' = (p_1, p_2, \ldots, p_q)$. We now make the transformation $\mathbf{y} = \mathbf{B}\mathbf{x}$, where

$$\mathbf{B} = \begin{bmatrix} \mathbf{C}' \\ \mathbf{H}'\mathbf{R}^{-1} \end{bmatrix}.$$

The columns of the matrix $\mathbf{H}$ are chosen to be orthogonal to those of $\mathbf{C}$ and orthonormal amongst themselves. Since there are $2^{-1}(n-q)(1+q+n)$ constraint relations and $n(n-q)$ independent elements, such a matrix can always be constructed for $q \leqslant n-1$. When $q = n$, $\mathbf{B} = \mathbf{C}'$ since $\mathbf{C}'$ is then $n \times n$.

The m.g.f. for the truncated distribution is now given by

$$\alpha m(\mathbf{t}) = K \int_{A_n(\mathbf{y})} \exp\{\boldsymbol{\theta}'\mathbf{y} - \tfrac{1}{2}\mathbf{y}'\mathbf{V}^{-1}\mathbf{y}\}\, d\mathbf{y},$$

where

$$K = (2\pi)^{-\frac{1}{2}n}|\mathbf{V}|^{-\frac{1}{2}}, A_n(\mathbf{y}) = \{\mathbf{y};\ y_1 \geqslant p_1, \ldots, y_q \geqslant p_q\}, \quad \boldsymbol{\theta} = (\mathbf{B}^{-1})'\mathbf{t}$$

and

$$\mathbf{V}^{-1} = (\mathbf{B}^{-1})'\mathbf{R}^{-1}\mathbf{B}^{-1}.$$

We again use identity (8) and notice that $\boldsymbol{\theta}'\mathbf{V}\boldsymbol{\theta} = \mathbf{t}'\mathbf{R}\mathbf{t} = 2T$ to establish

$$\alpha m(\mathbf{t}) = K e^T \int_{A_n(\mathbf{y})} \exp\{-\tfrac{1}{2}(\mathbf{y}-\boldsymbol{\beta})'\mathbf{V}^{-1}(\mathbf{y}-\boldsymbol{\beta})\}\, d\mathbf{y},$$

with $\boldsymbol{\beta} = \mathbf{V}\boldsymbol{\theta} = \mathbf{B}\mathbf{R}\mathbf{t}$.

Suitable matrix multiplication shows that

$$\boldsymbol{\beta} = \begin{bmatrix} \mathbf{C}'\mathbf{R}\mathbf{t} \\ \mathbf{H}'\mathbf{t} \end{bmatrix} \quad \text{and} \quad \mathbf{V} = \begin{bmatrix} \mathbf{C}'\mathbf{R}\mathbf{C} & \mathbf{O}' \\ \mathbf{O} & \mathbf{H}'\mathbf{R}^{-1}\mathbf{H} \end{bmatrix},$$

and therefore

$$\mathbf{V}^{-1} = \begin{bmatrix} (\mathbf{C}'\mathbf{R}\mathbf{C})^{-1} & \mathbf{O}' \\ \mathbf{O} & (\mathbf{H}'\mathbf{R}^{-1}\mathbf{H})^{-1} \end{bmatrix}.$$

It follows, therefore, that the expression for $m(\mathbf{t})$ can be put in the form

$$\alpha m(\mathbf{t}) = (2\pi)^{-\frac{1}{2}q}|\mathbf{C}'\mathbf{R}\mathbf{C}|^{-\frac{1}{2}} F \times (2\pi)^{-\frac{1}{2}(n-q)}|\mathbf{H}'\mathbf{R}^{-1}\mathbf{H}|^{-\frac{1}{2}} G$$

where

$$F = \int_{y_1=p_1}^{\infty} \cdots \int_{y_q=p_q}^{\infty} \exp\{-\tfrac{1}{2}(\mathbf{y}_q - \mathbf{C}'\mathbf{R}\mathbf{t})'(\mathbf{C}'\mathbf{R}\mathbf{C})^{-1}(\mathbf{y}_q - \mathbf{C}'\mathbf{R}\mathbf{t})\}\, d\mathbf{y}_q,$$

and

$$G = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \exp\{-\tfrac{1}{2}(\mathbf{y}_{n-q} - \mathbf{H}'\mathbf{t})'(\mathbf{H}'\mathbf{R}^{-1}\mathbf{H})^{-1}(\mathbf{y}_{n-q} - \mathbf{H}'\mathbf{t})\}\, d\mathbf{y}_{n-q},$$

$\mathbf{y}_q$ representing the first q components of $\mathbf{y}$ and $\mathbf{y}_{n-q}$ the last $(n-q)$ components. The last factor on the right of the first equation is unity and a further transformation $z_i = (y_i - \mathbf{c}_i'\mathbf{R}\mathbf{t})/(\mathbf{c}_i'\mathbf{R}\mathbf{c}_i)^{\frac{1}{2}}$ $(i = 1, 2, \ldots, q)$ reduces the m.g.f. to

$$\alpha m(\mathbf{t}) = e^{T}\, \Phi_q(\mathbf{d};\, \mathbf{R}_c), \tag{13}$$

where $\mathbf{R}_c = \mathbf{D}\mathbf{C}'\mathbf{R}\mathbf{C}\mathbf{D}$, $\mathbf{d} = \mathbf{D}(\mathbf{p} - \mathbf{C}'\mathbf{R}\mathbf{t})$ and $\mathbf{D}^{-2} = \mathrm{diag}(\mathbf{c}_1'\mathbf{R}\mathbf{c}_1, \ldots, \mathbf{c}_q'\mathbf{R}\mathbf{c}_q)$. The moments are now obtainable from the general formulae developed in Tallis (1961).

5. Some Special Cases

It is of interest to examine two special cases of plane truncation which are of practical importance in selection theory. In the first instance, we consider the vector $\mathbf{x}$ partitioned as

$$\mathbf{x} = \begin{bmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \\ \vdots \\ \mathbf{x}_q \end{bmatrix}$$

for $q \leqslant n$. The components of $\mathbf{x}_i$ and the truncation set are specified by

$$\mathbf{x}_i' = (x_{1i}, x_{2i}, \ldots, x_{n_i i})$$

and

$$A_n(\mathbf{x}) = \{\mathbf{x};\ \mathbf{c}_1'\mathbf{x}_1 \geqslant p_1, \mathbf{c}_2'\mathbf{x}_2 \geqslant p_2, \ldots, \mathbf{c}_q'\mathbf{x}_q \geqslant p_q\}.$$

This type of truncation occurs when the population is subjected to q different selections, each one involving a different set of characters, $\mathbf{x}_i$. The above system will be called group truncation.

The required m.g.f. can be inferred directly from (13). However, it is instructive to derive it from first principles. Make the single orthonormal transformation $\mathbf{x} = \mathbf{H}\mathbf{y}$ where $\mathbf{H}$ is the direct sum

$$\mathbf{H} = \mathrm{diag}(\mathbf{H}_1, \mathbf{H}_2, \ldots, \mathbf{H}_q),$$

and each $\mathbf{H}_i$ is $(n_i \times n_i)$, orthonormal and with first column $\mathbf{c}_i$. We obtain for the m.g.f.

$$\alpha m(\mathbf{t}) = K \int_{A_n(\mathbf{y})} \exp\{(\mathbf{H}'\mathbf{t})'\mathbf{y} - \tfrac{1}{2}\mathbf{y}'\mathbf{H}'\mathbf{R}^{-1}\mathbf{H}\mathbf{y}\}\, d\mathbf{y},$$

where

$$K = (2\pi)^{-\frac{1}{2}n}\,|\mathbf{R}|^{-\frac{1}{2}}$$

and

$$A_n(\mathbf{y}) = \{\mathbf{y};\ y_{11} \geqslant p_1, y_{12} \geqslant p_2, \ldots, y_{1q} \geqslant p_q\}.$$

This can be put in the form

$$\alpha m(\mathbf{t}) = K e^{T} \int_{A_n(\mathbf{w})} \exp\{-\tfrac{1}{2}\mathbf{w}'\mathbf{H}'\mathbf{R}^{-1}\mathbf{H}\mathbf{w}\}\, d\mathbf{w},$$

where

$$A_n(\mathbf{w}) = \{\mathbf{w};\ w_{11} + \mathbf{c}_1'\mathbf{R}_{1.}\mathbf{t} \geqslant p_1, \ldots, w_{1q} + \mathbf{c}_q'\mathbf{R}_{q.}\mathbf{t} \geqslant p_q\}$$

and R is partitioned according to x

$$\mathbf{R} = \begin{bmatrix} \mathbf{R}_{1.} \\ \mathbf{R}_{2.} \\ \vdots \\ \mathbf{R}_{q.} \end{bmatrix}.$$

Upon integrating out the $(n-q)$ unconditional variables, we are left with a matrix

$$\mathbf{V}_c = \begin{bmatrix} \mathbf{c}_1'\mathbf{R}_{11}\mathbf{c}_1 & \mathbf{c}_1'\mathbf{R}_{12}\mathbf{c}_2 & \ldots & \mathbf{c}_1'\mathbf{R}_{1q}\mathbf{c}_q \\ \mathbf{c}_2'\mathbf{R}_{21}\mathbf{c}_1 & \mathbf{c}_2'\mathbf{R}_{22}\mathbf{c}_2 & \ldots & \mathbf{c}_2'\mathbf{R}_{2q}\mathbf{c}_q \\ \mathbf{c}_q'\mathbf{R}_{q1}\mathbf{c}_1 & \mathbf{c}_q'\mathbf{R}_{q2}\mathbf{c}_2 & \ldots & \mathbf{c}_q'\mathbf{R}_{qq}\mathbf{c}_q \end{bmatrix}$$

where R has been partitioned as

$$\mathbf{R} = \begin{bmatrix} \mathbf{R}_{11} & \mathbf{R}_{12} & \ldots & \mathbf{R}_{1q} \\ \mathbf{R}_{21} & \mathbf{R}_{22} & \ldots & \mathbf{R}_{2q} \\ \mathbf{R}_{q1} & \mathbf{R}_{q2} & \ldots & \mathbf{R}_{qq} \end{bmatrix}$$

in conformity with $E(\mathbf{x}\mathbf{x}')$. Now let $\mathbf{R}_c = \mathbf{D}\mathbf{V}_c\mathbf{D}$ with

$$\mathbf{D}^{-2} = \operatorname{diag}\{(\mathbf{c}_1'\mathbf{R}_{11}\mathbf{c}_1), \ldots, (\mathbf{c}_q'\mathbf{R}_{qq}\mathbf{c}_q)\},$$

then we make the final change of variable $\mathbf{w} = \mathbf{D}\mathbf{z}$ and

$$\alpha m(\mathbf{t}) = e^{T}(2\pi)^{-\frac{1}{2}q}|\mathbf{R}_c|^{-\frac{1}{2}} \int_{A_q(\mathbf{z})} \exp\{-\tfrac{1}{2}\mathbf{z}'\mathbf{R}_c^{-1}\mathbf{z}\}\, d\mathbf{z},$$

with

$$A_q(\mathbf{z}) = \{\mathbf{z};\ z_1 \geqslant (p_1 - \mathbf{c}_1'\mathbf{R}_{1.}\mathbf{t})/(\mathbf{c}_1'\mathbf{R}_{11}\mathbf{c}_1)^{\frac{1}{2}}, \ldots, z_q \geqslant (p_q - \mathbf{c}_q'\mathbf{R}_{q.}\mathbf{t})/(\mathbf{c}_q'\mathbf{R}_{qq}\mathbf{c}_q)^{\frac{1}{2}}\}.$$

Thus, $\alpha m(\mathbf{t})$ can be put in the form (13); however, in this case d_s takes the form

$$d_s = (p_s - \mathbf{c}_s'\mathbf{R}_{s.}\mathbf{t})/(\mathbf{c}_s'\mathbf{R}_{ss}\mathbf{c}_s)^{\frac{1}{2}} \quad (s = 1, 2, \ldots, q). \tag{14}$$

As a final example we consider a sequential type of truncation where x is partitioned into two subvectors,

$$\mathbf{x} = \begin{bmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \end{bmatrix}$$

and

$$A_n(\mathbf{x}) = \{\mathbf{x};\ \mathbf{c}_1'\mathbf{x}_1 \geqslant p_1, \mathbf{c}'\mathbf{x} \geqslant p\}.$$

We now use (13) to write the appropriate m.g.f. as

$$\alpha m(\mathrm{t}) = e^{T} \Phi_2\{(p_1 - \mathrm{c}_1'\mathrm{Rt})/(\mathrm{c}_1'\mathrm{Rc}_1)^{\frac{1}{2}}, (p - \mathrm{c}'\mathrm{Rt})/(\mathrm{c}'\mathrm{Rc})^{\frac{1}{2}}; \mathrm{R}_c\} \tag{15}$$

where

$$\mathrm{R}_c = \begin{bmatrix} 1 & \rho_c \\ \rho_c & 1 \end{bmatrix}$$

and $\rho_c = \mathrm{c}_1\mathrm{Rc}/(\mathrm{c}_1'\mathrm{Rc}_1 . \mathrm{c}'\mathrm{Rc})^{\frac{1}{2}}$. It is easily seen that

$$\alpha = \Phi_2\{p_1/(\mathrm{c}_1'\mathrm{Rc}_1)^{\frac{1}{2}}, p/(\mathrm{c}'\mathrm{Rc})^{\frac{1}{2}}; \mathrm{R}_c\}$$

and the correct formulae for the first and second moments after truncation are obtained from the formulae for rectangular truncation. Thus for instance,

$$\alpha E(X_i) = \phi(p_1')\Phi(P_{12})\Sigma\rho_{iv}c_{1v}/(\mathrm{c}_1'\mathrm{Rc}_1)^{\frac{1}{2}} + \phi(p')\Phi(P_{21})\Sigma\rho_{iv}c_v/(\mathrm{c}'\mathrm{Rc})^{\frac{1}{2}},$$

where

$$P_{12} = (p' - \rho_c p_1')/(1-\rho_c^2)^{\frac{1}{2}},$$

$$P_{21} = (p_1' - \rho_c p')/(1-\rho_c^2)^{\frac{1}{2}},$$

$$p_1' = p_1/(\mathrm{c}_1'\mathrm{Rc}_1)^{\frac{1}{2}} \quad \text{and} \quad p' = p/(\mathrm{c}'\mathrm{Rc})^{\frac{1}{2}}.$$

From a practical viewpoint, group and sequential truncation systems are important in the theory of selection. For instance, when animals are selected for breeding, a first selection may be made on the subvector x_1 of x, the vector of all economically important characters. In fact, there are q stages to the selection programme and each set of measurements is used in turn. The types of selection differ in that, although both consider an initial selection in x_1, in the case of sequential truncation the final selection is for the total vector x. . These procedures lead to numerous questions of efficiency. Thus, in many situations in genetic selection, it is desirable to maximize some linear compound $\Sigma\beta_i E(X_i)$ subject to the constraint that $\Pi\alpha_i = \alpha$. This is to be accomplished by suitable choices of the coefficient vectors c_i. However, these problems will not be investigated further here and reference is made to Cochran (1950) for a fuller discussion of these topics.

ACKNOWLEDGEMENT

The author is grateful to the referees who made a number of comments to improve the original form of this paper.

REFERENCES

BIRNBAUM, Z. W. and MEYER, P. L. (1953), "On the effect of truncation in some or all co-ordinates of a multi-normal population", *J. Ind. Soc. agric. Statist.*, 5, 17–28.

COCHRAN, W. G. (1950), "Improvement by means of selection", *Proc. Second Berkeley Symp. Math. Statist. and Prob.*, 449–470. University of California Press.

FINNEY, D. J. (1962), "Cumulants of truncated multinormal distributions", *J. R. statist. Soc.* B, **24**, 535–536.

TALLIS, G. M. (1961), "The moment generating function of the truncated multi-normal distribution", *J. R. statist. Soc.* B, **23**, 223–229.

A (a) (ii) [4]

Reprinted from THE ANNALS OF MATHEMATICAL STATISTICS
Vol. 34, No. 3, September, 1963
Printed in U.S.A.

ELLIPTICAL AND RADIAL TRUNCATION IN NORMAL POPULATIONS

BY G. M. TALLIS

McMaster Laboratory, C.S.I.R.O., Glebe, N.S.W.

1. Introduction. Over the past few years, considerable attention has been devoted to problems of truncation in normal (and other) parent populations, [see Birnbaum and Meyer (1953), Weiler (1959) and Tallis (1961)]. This work has been useful in the general theory of selection and has provided the basis for a number of selection techniques. It is the purpose of this note to introduce the concept of elliptical truncation in normal populations and to derive the moment generating function, m.g.f., for the resulting distribution. Some applications of the results to selection are given in the last section, where also, combined elliptical and radial truncation is discussed by means of problems in two dimensions.

2. The multinormal distribution under elliptical truncation. Consider the standardised, n-dimensional multinormal distribution

$$\phi(\mathbf{x}) = (2\pi)^{-\frac{1}{2}n}|\mathbf{R}|^{-\frac{1}{2}} \exp(-\tfrac{1}{2}\mathbf{x}'\mathbf{R}^{-1}\mathbf{x}), \tag{1}$$

where R is positive definite, and define a set E in n-space by

$$E = \{\mathbf{x} \mid a \leqq \mathbf{x}'\mathbf{R}^{-1}\mathbf{x} \leqq b\}, 0 \leqq a < b.$$

That is, E is the set of points which lie inside or on the boundary of the ellipsoid $\mathbf{x}'\mathbf{R}^{-1}\mathbf{x} = b$ and outside or on the boundary of the ellipsoid $\mathbf{x}'\mathbf{R}^{-1}\mathbf{x} = a$.

The problem now is to find the m.g.f. for the n variables in the subspace E.

By definition

$$\alpha m(\mathbf{t}) = (2\pi)^{-\frac{1}{2}n} |\mathbf{R}|^{-\frac{1}{2}} \int_E \exp(-\tfrac{1}{2}\mathbf{x}'\mathbf{R}^{-1}\mathbf{x} + \mathbf{t}'\mathbf{x})\, d\mathbf{x}, \tag{2}$$

which can be reduced by the non-singular transformation $\mathbf{y} = \mathbf{P}^{-1}\mathbf{x}$ $(\mathbf{PP}' = \mathbf{R})$ to

$$\begin{aligned} \alpha m(\mathbf{t}) &= (2\pi)^{-\frac{1}{2}n} \int_F \exp(-\tfrac{1}{2}\mathbf{y}'\mathbf{y} + (\mathbf{P}'\mathbf{t})'\mathbf{y})\, d\mathbf{y} \\ &= (2\pi)^{-\frac{1}{2}n} e^T \int_F \exp[-\tfrac{1}{2}(\mathbf{y} - \mathbf{P}'\mathbf{t})'(\mathbf{y} - \mathbf{P}'\mathbf{t})]\, d\mathbf{y}, \end{aligned} \tag{3}$$

where $T = \frac{1}{2}\mathbf{t}'\mathbf{R}\mathbf{t}$ and $F = \{\mathbf{y} \mid a \leqq \mathbf{y}'\mathbf{y} \leqq b\}$. From (3) it is clear that the variable $\mathbf{Y}'\mathbf{Y} = W$, say, has a non-central chi-square distribution with parameters n and T. Hence, if $F_{n+2i}(\cdot)$ represents the chi-square distribution function with parameter $n + 2i$,

$$e^{-T}\alpha m(\mathbf{t}) = \sum_{i=0}^{\infty} [F_{n+2i}(b) - F_{n+2i}(a)]T^i/i! \tag{4}$$

since the distribution function of W, $H(w)$, is $H(w) = \sum_{i=0}^{\infty} F_{n+2i}(w)T^i/i!$

Received April 23, 1962; revised April 15, 1963.

It is now obvious that $\alpha = F_n(b) - F_n(a)$ and that the mean vector $\mathbf{u}$ and moment matrix M are given by $\mathbf{u} = 0$ and $\mathrm{M} = \alpha^{-1}[F_{n+2}(b) - F_{n+2}(a)]\mathrm{R}$. In fact, all odd order moments vanish and even moments of order $2k$ are obtained from those of the multinormal distribution by multiplication by $\alpha^{-1}[F_{n+2k}(b) - F_{n+2k}(a)]$.

3. Applications and extensions.

(a) *General applications in n-dimensions.* A direct application of elliptical truncation to selection would require, for instance, that all individuals in the population satisfying $0 \leqq \mathrm{x}'\mathrm{R}^{-1}\mathrm{x} \leqq a$ be retained and the rest discarded. This procedure would ensure that a proportion α be retained without altering the means of the n variates concerned. Such a situation may be desirable, for instance, if a breeding population applying zero selection pressure to all characters is required as a control group. In practice then, the population from which the selections are made is treated as multinormally distributed with correlation matrix $\bar{\mathrm{R}}$. If the population is accessible and finite, $\bar{\mathrm{R}}$ can be calculated; otherwise $\bar{\mathrm{R}}$ may be an estimate of the true matrix R which is usually unknown. Now, all individuals with measurement vectors satisfying $\mathrm{x}'\bar{\mathrm{R}}\mathrm{x} > a$ are discarded and, in the remaining group, the desired condition $E(X_i) = 0$ for all i will be approximately satisfied.

Not only can selection be performed without altering the means of the n variates, but the following argument shows that a selected group can be formed such that the covariance matrix also remains unchanged. If selection is carried out in such a way that individuals with measurement vectors, x, satisfying $a \leqq \mathrm{x}'\mathrm{R}^{-1}\mathrm{x} \leqq b$ are retained, then it follows from (4) that if M is to equal R,

$$F_{n+2}(b) - F_n(b) = F_{n+2}(a) - F_n(a) \tag{5}$$

is a necessary and sufficient condition. Let $G_n(x) = F_{n+2}(x) - F_n(x)$, then since $G_n(0) = 0$ and $G_n(x)$ decreases monotonically and continuously to a minimum at $x = n$ and thereafter increases monotonically to $G_n(\infty) = 0$, it follows that for every $a \varepsilon [0, n]$ there exists a $b \varepsilon [n, \infty]$ such that $G_n(a) = G_n(b)$. Moreover, b is a strictly monotone decreasing and continuous function of a and, as a moves continuously from 0 to n, b moves continuously from ∞ to n. Thus, $F_n(b) - F_n(a)$ takes all values of α from 1 to 0. We have shown, therefore, that selection can in fact be carried out in such a way as to have the first and second moments of the selected group the same as the parent population.

Values of a and b are given for $\alpha = .1(.1).9$ and $n = 2$ in Table 1. The figures in the table were found as the non-trivial, simultaneous solution to the equations $ye^{-y} - xe^{-x} = 0$ and $e^{-x} - e^{-y} = \alpha$, where $x = a/2$ and $y = b/2$.

(b) *Extensions for n = 2.* The following two problems in two dimensions are considerably more interesting than the general applications given in (a).

PROBLEM 1. Let S be the sub-space of the plane defined by $x_1^2 + x_2^2 - 2\rho x_1 x_2 \geqq (1 - \rho^2)a$, where X_1 and X_2 have joint frequency function

$$\phi(x_1, x_2; \rho) = (2\pi)^{-1}(1 - \rho^2)^{-\frac{1}{2}} \times \exp\{-[2(1 - \rho^2)]^{-1}(x_1^2 + x_2^2 - 2\rho x_1 x_2)\}. \tag{6}$$

TABLE 1

Values of a and b for n = 2 and various α

α	a	b	α	a	b
0.1	1.740	2.285	0.6	0.684	4.411
0.2	1.500	2.601	0.7	0.506	5.144
0.3	1.277	2.956	0.8	0.335	6.161
0.4	1.068	3.361	0.9	0.171	8.632
0.5	0.871	3.836			

Consider the sub-space S' of S enclosed in the sector $\theta = \theta_1$, $\theta = \theta_2$, $\theta_2 > \theta_1$, and let $\Pr\{(X_1, X_2) \varepsilon S'\} = \alpha$. Then it is required to determine the values of θ_1, θ_2 and a, $(\bar{\theta}_1, \bar{\theta}_2, \bar{a})$ which maximise $[\beta_1 E(X_1) + \beta_2 E(X_2)]$, where β_1 and β_2 are arbitrary real numbers.

Such a maximisation is desirable, for instance, when animals are selected for breeding. In this case, the β's are the appropriate regression functions of the X's on the particular genotype considered and the problem posed above is analogous to the one discussed by Young and Weiler (1960). These authors investigated the problem of the maximisation of $[\beta_1 E(X_1) + \beta_2 E(X_2)]$ under rectangular truncation in X_1 and X_2 and published several charts for this purpose. From the point of view of maximisation, the system of combined radial and elliptical truncation is much more easily handled, since the maxima can be obtained directly from a single table such as Table 2. With rectangular truncation, maximisation in general can only be achieved iteratively with the aid of a complicated six-dimensional chart.

In order to find $(\bar{\theta}_1, \bar{\theta}_2, \bar{a})$, make the transformation $\mathbf{x} = \mathbf{P}\mathbf{y}$ where

$$\mathbf{P} = \begin{bmatrix} 2^{-\frac{1}{2}}(1-\rho)^{\frac{1}{2}} & 2^{-\frac{1}{2}}(1+\rho)^{\frac{1}{2}} \\ -2^{-\frac{1}{2}}(1-\rho)^{\frac{1}{2}} & 2^{-\frac{1}{2}}(1+\rho)^{\frac{1}{2}} \end{bmatrix}.$$

Now $\beta_1 E(X_1) + \beta_2 E(X_2) = \gamma_1 E(Y_1) + \gamma_2 E(Y_2)$, $\gamma_1 = (\beta_1 - \beta_2)2^{-\frac{1}{2}}(1-\rho)^{\frac{1}{2}}$, $\gamma_2 = (\beta_1 + \beta_2)2^{-\frac{1}{2}}(1+\rho)^{\frac{1}{2}}$, and the new angle θ_i', $(i = 1, 2)$, are given by the formula

$$\tan\theta_i' = [(1-\rho)/(1+\rho)]^{\frac{1}{2}}[(1+\tan\theta_i)/(1-\tan\theta_i)].$$

Another transformation, this time orthogonal, subsequently simplifies the problem. Let $\mathbf{z} = \mathbf{H}\mathbf{y}$, where

$$\mathbf{H} = \begin{bmatrix} \gamma_1(\gamma_1^2+\gamma_2^2)^{-\frac{1}{2}} & \gamma_2(\gamma_1^2+\gamma_2^2)^{-\frac{1}{2}} \\ -\gamma_2(\gamma_1^2+\gamma_2^2)^{-\frac{1}{2}} & \gamma_1(\gamma_1^2+\gamma_2^2)^{-\frac{1}{2}} \end{bmatrix}.$$

Upon making the above two transformations in (6) and letting $z_1 = r\cos\theta$, $z_2 = r\sin\theta$ we obtain finally

$$\alpha E[\beta_1 X_1 + \beta_2 X_2] = (2\pi)^{-1}(\gamma_1^2 + \gamma_2^2)^{\frac{1}{2}} \int_{\theta_1''}^{\theta_2''} \int_a^\infty \cos\theta r^2 e^{-r^2/2}\, dr\, d\theta$$

$$= (\gamma_1^2 + \gamma_2^2)^{\frac{1}{2}} E(z_1) = F(\theta_1'', \theta_2'', a), \text{ say,} \tag{7}$$

and by assumption

$$(2\pi)^{-1} \int_\theta^{\theta_2''} \int_a^\infty r e^{-r^2/2}\, dr\, d\theta = (2\pi)^{-1}(\theta_2'' - \theta_1'')e^{-a^2/2} = \alpha.$$

Let

$$H(\theta_1'', \theta_2'', a) = (2\pi)^{-1}(\theta_2'' - \theta_1'')e^{-a^2/2} - \alpha = 0 \tag{8}$$

and $G = F + \lambda H$, then

$$\text{(a)}\quad \partial G/\partial\theta_1'' = (2\pi)^{-1}(\gamma_1^2 + \gamma_2^2)^{\frac{1}{2}} \left(\int_a^\infty r^2 e^{-r^2/2}\, dr \right) \cos\theta_1''$$

$$+ (2\pi)^{-1} \lambda\, e^{-a^2/2} = 0$$

$$\text{(b)}\quad \partial G/\partial\theta_2'' = (2\pi)^{-1}(\gamma_1^2 + \gamma_2^2)^{\frac{1}{2}} \left(\int_a^\infty r^2 e^{-r^2/2}\, dr \right) \cos\theta_2'' \tag{9}$$

$$+ (2\pi)^{-1}\lambda e^{-a^2/2} = 0$$

$$\text{(c)}\quad \partial G/\partial a = (2\pi)^{-1}(\gamma_1^2 + \gamma_2^2)^{\frac{1}{2}} a^2 e^{-a^2/2}(\sin\theta_2'' - \sin\theta_1'')$$

$$+ (2\pi)^{-1}\lambda(\theta_2'' - \theta_1'')ae^{-a^2/2} = 0$$

Substract 9(b) from 9(a) to obtain $\cos\theta_1'' = \cos\theta_2''$ or $\theta_2'' = -\theta_1''$, and for $a > 0$, divide 9(c) through by $a(\theta_2'' - \theta_1'')$ and subtract it from 9(b) to give

$$\{1 + [1 - \Phi(\bar{a})]/\bar{a}\phi(\bar{a})\} \cos\bar{\theta}_2'' - \sin\bar{\theta}_2''/\bar{\theta}_2'' = 0. \tag{10}$$

By using the relation $\phi(a) = \alpha/\theta_2'' \cdot (\pi/2)^{\frac{1}{2}}$, (10) can be solved iteratively for $\bar{\theta}_2''$. The quantities $\bar{\theta}_1''$ and $\bar{a}$ are obtained immediately and back substitution gives $\bar{\theta}_1$ and $\bar{\theta}_2$. If $a = 0$, it is found that $\theta_2'' = -\theta_1''$, as previously, and $\theta_2'' = \alpha\pi$.

If the constraints (8) and $\theta_2'' = -\theta_1''$ are introduced into (7), F becomes a function of a only and

$$F(a) = K \sin(\pi\alpha e^{a^2/2}) \int_a^\infty r^2 e^{-r^2/2}\, dr \tag{11}$$

TABLE 2

Values of θ_2'', a and $E(z_1)$ for various α

α	θ_2''	a	$E(z_1)$	α	θ_2''	a	$E(z_1)$
.1	0.877	1.433	1.722	.6	1.885	0	0.632
.2	1.044	1.008	1.375	.7	2.199	0	0.461
.3	1.196	0.691	1.144	.8	2.513	0	0.293
.4	1.357	0.393	0.960	.9	2.827	0	0.137
.5	1.571	0.000	0.798				

where $K = \pi^{-1}(\gamma_1^2 + \gamma_2^2)^{\frac{1}{2}}$. For $\alpha < \frac{1}{2}$, (10) has a unique root, $0 < \bar{a} < (-2 \ln 2\alpha)^{\frac{1}{2}}$, and from an inspection of (11), $F(\bar{a})$ is clearly a maximum. When $\alpha \geqq \frac{1}{2}$ (10) has no solution, but $F(a)$ is monotone decreasing in a and hence attains its maximum when $a = 0$. Points of maximisation and values of $E(z_1)$ are given for various values of α in Table 2.

A referee has pointed out that (7) can be maximised readily without using the Lagrange procedure. First make the transformation $\epsilon = 2^{-1}(\theta_1'' + \theta_2'')$ and $\Delta = 2^{-1}(\theta_1'' - \theta_2'')$ to show that, for all Δ and a, (7) is maximum when $\epsilon = 0$. By introducing the constraint $\Delta = \pi\alpha e^{a^2/2}$, (7) can be written in the form (11) and the extreme points investigated in the usual manner. Both methods lead to the same result.

Problem 2. It was shown above that, from an original population, a control population can be constructed so that no changes in means or second order moments occur, provided the radii a and b are suitably chosen. However, the problem of simultaneously establishing a control group of proportion α and a selection group of proportion $\delta < 1 - \alpha$ from a single base population often arises. In this case it may be desirable to leave the first and second moments in the control group the same as the base population and, at the same time, maximise $E[\beta_1 X_1 + \beta_2 X_2]$ in the selection group to obtain the greatest possible selection differential using a single sector.

The control group is established by means of the elliptical truncation $a \leqq \mathbf{x}'\mathbf{R}^{-1}\mathbf{x} \leqq b$, where a and b are determined from Table 1. In order to find the region from which the selection group is formed, notice that

$$\delta E[\beta_1 X_1 + \beta_2 X_2] = (2\pi)^{-1}(\gamma_1^2 + \gamma_2^2)^{\frac{1}{2}} \int_{\theta_1^*}^{\theta_2^*} \cos\theta \left(\int_0^a + \int_b^\infty \right) r^2 e^{-r^2/2}\, dr\, d\theta$$

and $H(\theta_1'', \theta_2'') = (2\pi)^{-1}(\theta_2'' - \theta_1'')(1 - \exp(-a^2/2) + \exp(-b^2/2)) - \delta = 0$. It is found immediately that $\bar{\theta}_2'' = -\bar{\theta}_1''$, as previously, and

$$\bar{\theta}_2'' = \delta\pi/(1 - e^{-a^2/2} + e^{-b^2/2})$$

$$E(z_1) = (2/\pi)^{\frac{1}{2}} \sin(\bar{\theta}_2'')[b\phi(b) - a\phi(a) + 1 + \Phi(a) - \Phi(b)].$$

Thus, all those individuals lying in the sector $(\bar{\theta}_1'', \bar{\theta}_2'')$ and outside the control group form the selection group.

4. Acknowledgement. The author gratefully acknowledges the valuable suggestions of Mr. George Brown and the referees.

REFERENCES

Birnbaum, Z. W. and Meyer, P. L. (1953). On the effect of truncation in some or all coordinates of a multi-normal population. *J. Indian Soc. Agric. Statist.* **5** 17–28.

Tallis, G. M. (1961). The moment generating function of the truncated multi-normal distribution. *J. Roy. Statist. Soc. Ser B* **23** 223–229.

Weiler, H. (1959). Means and standard deviations of a truncated normal bivariate distribution. *Austral. J. Statist.* **1** 73–81.

Young, S. S. Y. and Weiler, H. (1960). Selection for two correlated traits by independent culling levels. *J. Genetics* **57** 329–338.

NOTE

The methods of A(a)(ii)[4] are useful for evaluating certain integrals related to the uniform distribution. Some of these results are developed here as illustration.

As earlier, put $E_n = \{\underset{\sim}{x};\ \underset{\sim}{x}'V^{-1}\underset{\sim}{x} \leqslant a\}$ and define

$$M_a(\underset{\sim}{t}) = \int\cdots\int_{E_n} e^{\underset{\sim}{t}'\underset{\sim}{x}} d\underset{\sim}{x}.$$

Clearly $M_a(\underset{\sim}{t})$ is proportional to the m.g.f. of a uniform density defined over the ellipsoid $\underset{\sim}{x}'V^{-1}\underset{\sim}{x} = a$.

Now

$$M_a(\underset{\sim}{t}) = \lim_{\sigma\to\infty} \int\cdots\int_{E_n} e^{\underset{\sim}{t}'\underset{\sim}{x} - \frac{1}{2\sigma^2}\ \underset{\sim}{x}'V^{-1}\underset{\sim}{x}}$$

and putting $V^{-1} = P'P$, $\underset{\sim}{y} = P\underset{\sim}{x}$, $F_n = \{y;\ y'y \leqslant a\}$

$$M_a(\underset{\sim}{t}) = \lim_{\sigma\to\infty} \int\cdots\int_{F_n} e^{(P^{-1}{}'\underset{\sim}{t})'\underset{\sim}{y} - \frac{1}{2\sigma^2}\ \underset{\sim}{y}'\underset{\sim}{y}} |P|^{-1} d\underset{\sim}{y}.$$

Note that $|P| = |V^{-1}|^{\frac{1}{2}} = |V|^{-\frac{1}{2}}$ and let $\underset{\sim}{u} = \frac{1}{\sigma}\underset{\sim}{y}$, $P^{-1}{}' = R$ then

$$M_a(\underset{\sim}{t}) = (2\pi)^{\frac{n}{2}} |V|^{\frac{1}{2}} \lim_{\sigma\to\infty} \sigma^n \int\cdots\int_{G_n} e^{\sigma(Rt)'\underset{\sim}{u} - \frac{1}{2}\underset{\sim}{u}\underset{\sim}{u}} \frac{d\underset{\sim}{u}}{(2\pi)^{n/2}}$$

$$= (2\pi)^{\frac{n}{2}} |V|^{\frac{1}{2}} \lim_{\sigma\to\infty} \sigma^n e^{\sigma^2 T} \int\cdots\int_{G_n} e^{-\frac{1}{2}Q(\underset{\sim}{u})} \frac{d\underset{\sim}{u}}{(2\pi)^{n/2}}$$

where $G_n = \{\underset{\sim}{u};\underset{\sim}{u}'\underset{\sim}{u} \leqslant a/\sigma^2\}$, $T = \frac{1}{2}\underset{\sim}{t}'V\underset{\sim}{t}$ and

$$Q(\underset{\sim}{u}) = (\underset{\sim}{u}-\sigma R\underset{\sim}{t})'(\underset{\sim}{u}-\sigma R\underset{\sim}{t}).$$

Since $W_n = \underset{\sim}{U}'\underset{\sim}{U}$ has a $\chi^2(n,\lambda)$ distribution with $2\lambda = \sigma^2 t'R'Rt = \sigma^2 t'Vt = 2\sigma^2 T$, the required integral is

$$\Pr\{W_n \leqslant a/\sigma^2\} = \int_0^{a/\sigma^2} e^{-\lambda} \sum_{i=0}^{\infty} g_{n+2i}(x)\ \lambda^i/i!\ dx$$

and

$$M_a(\underset{\sim}{t}) = (2\pi)^{\frac{n}{2}} |V|^{\frac{1}{2}} \lim_{\sigma\to\infty} \sum_{i=0}^{\infty} G_{n+2i}(a/\sigma^2)\ \sigma^{n+2i}\ T^i/i!$$

where $G_k(x) = \int_0^x g_k(y)\,dy$.

It is easily verified that

$$\lim_{\sigma\to\infty} \sigma^k \, G_k(a/\sigma^2) = \frac{a^{k/2}}{\Gamma\left(\frac{k+2}{2}\right)2^{k/2}}$$

and hence

$$M_a(\underset{\sim}{t}) = (2\Pi)^{\frac{n}{2}} \, |V|^{\frac{1}{2}} \sum_{i=0}^{\infty} \frac{a^{(n+2i)/2} \, T^i}{2^{(n+2i)/2} \, \Gamma\left(\frac{n+2i+2}{2}\right) i!}$$

With $i = 0$ we get the volume of the n-dimensional ellipsoid as $(\Pi)^{\frac{n}{2}} \, |V|^{\frac{1}{2}} \, a^{n/2}/\Gamma\left(\frac{n+2}{2}\right)$ which specialises to $(\Pi)^{n/2} \, a^{n/2}/\Gamma\left(\frac{n+2}{2}\right)$ for the n-dimensional sphere.

In order to get the m.g.f. for the uniform density on $\underset{\sim}{x}'V^{-1}\underset{\sim}{x} = a$, $M(\underset{\sim}{t})$ is normalised and

$$M_a(\underset{\sim}{t}) = \sum_{i=0}^{\infty} \frac{a^i T^i \, \Gamma\left(\frac{n+2}{2}\right)}{i! \; 2^i \Gamma\left(\frac{n+2i+2}{2}\right)}$$

If we wish to choose a such that the covariance matrix of this uniform density is the same as V, put $i = 1$ and set

$$\frac{a \, \underset{\sim}{t}'V\underset{\sim}{t} \, \Gamma\left(\frac{n+2}{2}\right)}{2\times 2 \left(\frac{n+2}{2}\right) \Gamma\left(\frac{n+2}{2}\right)} = \tfrac{1}{2}\underset{\sim}{t}'V\underset{\sim}{t}$$

and $a = n+2$, see Cramér (1946), Mathematical Methods of Statistics, page 120.

(iii)

THE SAMPLING ERRORS OF ESTIMATORS OF CERTAIN GENETIC PARAMETERS AND PREDICTORS OF GENETIC GAIN

A (a) (iii) [1]

EFFICIENT ESTIMATES OF HERITABILITY FROM PATERNAL HALF-SIB CORRELATIONS

G. M. TALLIS AND EARLE W. KLOSTERMAN
Ohio Agricultural Experiment Station

Reprinted from JOURNAL OF ANIMAL SCIENCE, Vol. 18, No. 2, May, 1959

Reprinted from JOURNAL OF ANIMAL SCIENCE, Vol. 18, No. 2, May, 1959

EFFICIENT ESTIMATES OF HERITABILITY FROM PATERNAL HALF-SIB CORRELATIONS [1, 2]

G. M. TALLIS AND EARLE W. KLOSTERMAN [3]
Ohio Agricultural Experiment Station

OVER the past years, many estimates of heritability have been obtained from paternal half-sib correlations. Such estimates are subject to large sampling errors and extensive data must be used if accurate heritabilities are to be calculated.

One formula expressing the standard error of an estimate of an intra-class correlation coefficient is discussed by Fisher (1952) viz:

$$S_r = \frac{(1-r)\left[1+r\,(k-1)\right]}{\sqrt{\frac{1}{2}(k-1)\,bk}} \qquad (1)$$

where S_r = the standard error
r = the estimated intra-class correlation coefficient
b = the number of classes
k = the number of individuals within classes (same in all the b classes, $k>1$)

Clearly, the total number of individuals in the sample, n, is given by b x k. It is thus possible to deduce from formula 1

$$n = \frac{2(1-r)^2\left[1+r\,(k-1)\right]^2}{(k-1)\,S_r^2} \qquad (2)$$

Formula 1 has been used to estimate the standard error of heritability estimates, Sh, obtained from half-sib relationships (Hazel and Terrill, 1945). In this case

$$h^2 = \frac{4r}{1+F'} \quad \text{and} \quad S_h = \frac{4S_r}{1+F'}$$

[1] Published with the approval of the Associate Director as Journal Article No. 62-58.
[2] This manuscript was developed by G. M. Tallis and was taken in part from a dissertation presented by him to the Graduate School, The Ohio State University, in partial fulfillment of the Ph.D. degree in 1957. Present address is C.S.I.R.O., McMaster Animal Health Laboratory, Glebe, Sydney, Australia.
[3] Acknowledgment: The authors wish to acknowledge the helpful criticisms of Dr. J. A. Morris and Dr. F. E. Binet, of the Division of Animal Health and Production, C.S.I.R.O., Poultry Research Center, Werribee, Victoria.

where F' is the average coefficient of inbreeding of the sires, and h^2 is the estimated heritability.

With the aid of formulae 1 and 2, it is possible to investigate the following questions: (1) What is the minimum number of animals which must be studied in order to estimate heritability to a given accuracy? (2) When the number, n, of animals studied must be restricted to N, what combination of b and k gives the most efficient estimate of heritability?

Problem 1. An inspection of formula 2 reveals that, when Sr is held constant, n is a function of k and r. If environmental correlation is disregarded, r could have a theoretical range of

$$0 \leq r \leq .5$$

in half-sib data when the inbreeding of the sires is considered. However, the range which is of prime interest is

$$0 \leq r \leq .25.$$

$$\text{Now } n\Big|_{r=.25} = \frac{9(k+3)^2}{128(k-1)S_r^2} > n\Big|_{r=0} = \frac{2}{(k-1)S_r^2} \quad [k > 2]$$

Moreover, the value of r which makes n a maximum, r', is given by the equation

$$r' = \frac{k-2}{2(k-1)} \quad [k > 1]$$

and consequently $r' \geq .25$, $k > 2$..

Therefore, over the specified range of r, n is a monotone increasing function of r for all values of k greater than 2. This means that the number of animals required to estimate heritability to a given degree of accuracy depends on the heritability level, and this number increases as heritability goes from 0 to 1.

This result is more plausible when consideration is given to the ratio $\frac{r}{Sr}$. Clearly, if Sr is fixed and r is allowed to vary from 0 to .25, the value of the ratio increases. Thus, under the null hypothesis that r is zero, the larger values of r are statistically more significant (in terms of standard deviations) than smaller values. It is therefore not surprising that greater numbers of animals are required to obtain the higher significance levels of large r values.

A similar analysis may be made with respect to the variable k. When $\frac{\partial n}{\partial k}$ is calculated and equated to 0, it is found that n is minimized when $k = \frac{1+r}{r} = k'$. This means that each level of heritability has its own

optimum value of k,k'. If other values of k are used, n must inevitably be larger if heritability is to be estimated with the same accuracy. Values of k' are given in table 1.

TABLE 1. VALUES OF n', k' AND b' FOR DIFFERENT LEVELS OF HERITABILITY (F'=0)

		Sh=·05		Sh=·10		Sh=·15	
h^2	k'	b'	n'	b'	n'	b'	n'
·1	41	30	1230	7	287	3	123
·2	21	110	2310	28	588	12	252
·3	14	235	3290	60	840	27	378
·4	11	377	4147	94	1034	42	462
·5	9	544	4896	136	1224	60	540
·6	8	694	5552	165	1320	73	584
·7	7	871	6097	207	1449	92	644
·8	6	1092	6552	273	1638	121	726
·9	5	1388	6940	384	1920	171	855

The optimum number of offspring per sire decreases with increasing heritability. However, it is now possible to calculate the minimum number, n', of animals necessary to estimate heritability with a certain standard error, Sh. We summarize now the relationships used above:

$$k' = \frac{1+r}{r}, \quad r = \frac{(1+F')h^2}{4}, \quad S_r = \frac{(1+F')S_h}{4}$$

$$\text{and } n' = \frac{2(1-r)^2\left[1+r(k'-1)\right]^2}{(k'-1)S_r^2}$$

For example, let $F' = 0$

$h^2 = .4$

$S_h = .05$

then $r = .1$, $k' = 11$, $S_r = .0125$ and

$$n' = \frac{2(1-.1)^2\left[1+.1(11-1)\right]^2}{(11-1)(.0125)^2}$$

$$= 4,147$$

As $n = kb$, the optimum number of sires is $b' = \frac{n'}{k'}$. In this manner table 1 may be constructed. Attention is drawn to the fact that the values appearing in the table are only close approximations to the true values because k′ has been calculated to the nearest integer.

The results of this section agree well with the findings of Koch (1957), who examined the efficiency of different values of k and b for estimating heritability to a given degree of accuracy. Such a study is made possible by setting Sr and r, allowing k to vary and computing values of n from formula (2). As $b = \frac{n}{k}$, different combinations of b and k may be calculated and, together with n, compared with n′, b′ and k′ for the particular level of heritability under investigation.

Problem 2. In usual practice, the highest possible value of n is determined by economic factors. Thus, Sr becomes the independent variable (formula 1) and just like n, Sr increases as h^2 goes from 0 to 1 and is minimized for a given heritability when $k = \frac{1+r}{r} = k'$.

Hence, when n is set at a specific value, N, the number of offspring per sire to study for a given degree of heritability, h^2_i, is k'_i, and the optimum number of sires to use is $\frac{N}{k'_i} = b'_i$

The discussion thus far has neglected the fact that, usually, the heritabilities of several characteristics are to be estimated individually from data obtained on one group of animals. As the heritability of these characteristics is likely to vary widely, and because k′ changes with the level of heritability, the question arises as to which value of k, k'_o, to use under these circumstances. A solution to this problem may be obtained as follows, if it is assumed that

$$\sum_{i=1}^{a} \sum_{j=1}^{a} \mathrm{Cov}(r_i r_j),\ i \neq j,$$

is insignificantly small. Let h^2_i ($i = 1, 2, \cdots a$) be the heritability of the i^{th} characteristic, then $r_i = \frac{(1+F')}{4} h^2_i$. One criterion for selecting an efficient k value would be such that $\sum_{i=1}^{a} S^2_{ri}$ is a minimum. For this to obtain, $\frac{a \sum_{i=1}^{a} S^2_{ri}}{\partial k}$ must be equated to zero and solved for k.

We have $$\sum_{i=1}^{a} S_{ri}^2 = \sum_{i=1}^{a} \frac{2(1-r_i)^2\left[1+r_i(k-1)\right]^2}{(k-1)n}$$

and $$\frac{\partial \sum_{i=1}^{a} S_{ri}^2}{\partial k} = \sum_{i=1}^{a} \frac{2(1-r_i)^2}{n}\left[\frac{2r_i\left[1+r_i(k-1)\right]}{(k-1)} - \frac{\left[1+r_i(k-1)\right]^2}{(k-1)^2}\right] = 0$$

$$= \sum_{i=1}^{a} (1-r_i)^2\left[r_i^2(k-1)^2 - 1\right] = 0$$

whence $$k'_o = \sqrt{\frac{\sum_{i=1}^{a}(1-r_i)^2}{\sum_{i=1}^{a}(1-r_i)^2 r_i^2}} + 1 \qquad (3a)$$

If it were found desirable to weight individual characteristics (according to economic value and/or according to the reciprocal of the variances of the individual r estimates used in computing k'_o), some weighting factor, v, could be calculated for each characteristic. The formula for k' would then be,

$$k'_o = \sqrt{\frac{\sum_{i=1}^{a}(1-r_i)^2 v_i}{\sum_{i=1}^{a}(1-r_i)^2 r_i^2 v_i}} + 1 \qquad (3b)$$

A certain amount of circumlocution is unavoidable if use is to be made of formulae 3a and 3b. For instance, an investigator who can study 1,000 animals intends to estimate the heritability of seven characteristics. Moreover, he wishes to know the number of offspring per sire which will maximize the efficiency of his experiment. His first task is to obtain a rough estimate of the heritability of each trait from previous work. Once this information is available, he may proceed as shown below.

Characteristic	Estimated h^2 (Literature)	r, (F'=0)
A	.2	.05
B	.2	.05
C	.4	.10
D	.4	.10
E	.8	.20
F	.8	.20
G	.8	.20

From these figures k'_o may be calculated.

$$k'_o = \sqrt{\frac{2(.95)^2 + 2(.9)^2 + 3(.8)^2}{2(.95)^2(.05)^2 + 2(.9)^2(.1)^2 + 3(.8)^2(.2)^2}} + 1 \leq 9$$

This example assumes a constant economic value for all characteristics and a constant standard error of r.

Apparently, 9 offspring per sire are desirable and he will need approximately 111 sires $\left(\frac{1.000}{9}\right)$. Furthermore, he can expect to estimate the heritability of the i^{th} characteristic with a standard error of

$$Sh_i = \frac{4 S_{ri}}{1+F'} = \frac{4(1-r_i)[1+r_i(k'_o-1)]}{(1+F')\sqrt{\frac{1}{2}n(k'_o-1)}}$$

Thus, Sh for A and B is likely to be close to

$$\frac{4(1-.05)[1+.05(9-1)]}{\sqrt{\frac{1}{2} \times 1{,}000(9-1)}} \leq .08$$

Similarly, expected Sh's for the second group (C, D) and for the third group (E, F, G) are 0.10 and 0.14, respectively. Analogous but somewhat simpler reasoning is required to utilize table 1.

As a final illustration of these methods, suppose we wish to know what minimum value of n, n'_o, satisfies the conditions

$$\sum_{i=1}^{a} S_{ri}^2 = ac^2$$

where $i = 1, 2, \cdots a$ and c^2 is an arbitrary mean variance chosen by the investigator.

Now $$\sum_{i=1}^{a} S_{ri}^2 = \sum_{i=1}^{a} \frac{2(1-r_i)^2[1+r_i(k-1)]^2}{(k-1)n} = ac^2.$$

Hence

$$n = \frac{2}{ac^2(k-1)} \sum_{i=1}^{a} (1-r_i)^2 [1+r_i(k-1)]^2$$

and

$$n'_o = \frac{2}{ac^2(k'_o-1)} \sum_{i=1}^{a} (1-r_i)^2 [1+r_i(k'_o-1)]^2$$

Summary

Factors influencing the efficiency and accuracy of heritability estimates based on paternal half-sib correlations have been examined. It is concluded that, in any one experiment, the number of offspring per sire, k, plays an important role in determining the sizes of the errors of estimate. Optimum values of k, k′, for different heritability levels are presented together with a method for calculating the best k value, k'_o, to use when heritability is estimated for several characteristics. Minimum numbers of offspring necessary for estimating heritability to a given accuracy are also discussed.

Literature Cited

Fisher, R. A. 1952. Statistical Methods for Research Workers. Oliver and Boyd, London.

Hazel, L. N. and C. E. Terrill. 1945. Heritability of weaning weight and staple length in range rambouillet lambs. J. Animal Sci. 5:55.

Koch, R. 1957. Personal Communication.

A (a) (iii) [2]

Reprinted from JOURNAL OF ANIMAL SCIENCE, Vol. 19, No. 4, November, 1960

EFFECT OF SOME CONTROLLABLE ERRORS ON ESTIMATES OF GENETIC PARAMETERS, WITH SPECIAL REFERENCE TO EARLY POST-NATAL GROWTH IN MERINO SHEEP

G. M. TALLIS [1, 2]

Glebe, N.S.W., Australia

IN animal-breeding work, it is not always possible either to make observations on animals at a standard time or age, or even to adjust the recorded measurements to a standard basis before estimating genetic parameters. As an example, consider the estimation of heritability of birth weight of lambs, and its genetic correlation with other characters. Facilities and labour may not be available for collecting and weighing lambs immediately after birth, and errors result, depending on the lapse of time between birth and weighing. Labour may be saved by inspecting the lambing flock at less frequent intervals, but in the process the expected error term is increased. It is the purpose of this paper to investigate the effect of such "controllable" errors on estimates of heritability and genetic correlation.

Methods

It will be assumed, firstly, that the genetic parameters are to be estimated from half-sib data. Later, the case of estimation from parent-offspring regression will also be considered.

The genetic model for character x of the jth offspring of the ith sire group (x_{ij}) may be written:

$$x_{ij}=\mu+\tfrac{1}{2}g_i+c_{ij}+f_{ij} \qquad (1)$$

where μ is the population mean, $\frac{1}{2}g_i$ is the gene contribution from the ith sire, c_{ij} is some controllable error term and f_{ij} is the random error. It is assumed that the terms of equation (1) are independent. The relevant analysis of variance and covariance model for the qth and rth traits is given in table 1. It is appropriate to point out here that situations arise in

TABLE 1. ANALYSIS OF VARIANCE OR COVARIANCE

Source	d.f.	M.S. or Cov [a]	E(M.S.) or E(Cov)
Between sires	d_s	V_{qr}	$C_{qr}+F_{qr}+\frac{k}{4}G_{qr}$
Within sires	d_i	v_{qr}	$C_{qr}+F_{qr}$

[a] For mean square q=r.

[1] Division of Animal Genetics, C.S.I.R.O., McMaster Laboratory.

[2] Some data used in the example were obtained from records of experimental sheep maintained at the National Field Station, "Gilruth Plains", Cunnamulla, N.S.W.

practice where the components of (1) cannot be considered as statistically independent and, in these cases, the following analyses are inadequate.

It can be shown that $F_{qr}=\frac{1}{4}G_{qr}+E_{qr}$, where E_{qr} is an environmental variance or covariance, and hence G_{qr} can be estimated without bias in the usual way by

$$\hat{G}_{qr}=(4/k)\,[V_{qr}-v_{qr}] \tag{2}$$

As, by definition, the genetic correlation between x_q and x_r is estimated by

$$\hat{r}_g=\frac{\hat{G}_{qr}}{(\hat{G}_{qq}\cdot\hat{G}_{rr})^{1/2}} \tag{3}$$

it may be seen that the controllable error terms in no way bias $\hat{r}_g$.

On the other hand, the heritability of x_q is defined to be

$$h^2_q=\frac{G_{qq}}{G_{qq}+E_{qq}} \tag{4}$$

and, because v_{qq} is usually used as an estimate of $F_{qq}=\frac{1}{4}G_{qq}+E_{qq}$, h^2_q is estimated by

$$\hat{h}^2_q=\frac{4(V_{qq}-v_{qq})}{V_{qq}+(k-1)v_{qq}} \tag{5}$$

Hence, in this case, if $C_{qq}>0$, h^2_q is biased downwards and the average amount of bias is

$$\frac{G_{qq}+E_{qq}}{G_{qq}+E_{qq}+C_{qq}}.$$

In the case of the parent-offspring method of estimating genetic parameters, similar results may be derived. For any character, the following models may be written

$$\begin{aligned} w_i&=\mu_w+g_i+c_i+e_i \\ x_{ij}&=\mu_x+\tfrac{1}{2}g_i+d_{ij}+f_{ij} \end{aligned} \tag{6}$$

where w and x represent parent and offspring phenotypes respectively, μ_w and μ_x are parent and offspring population means, g_i is the genetic deviation of ith parent from the mean parental genotype, c_i and d_{ij} are controllable error terms and e_i and f_{ij} are random errors. It is assumed that all components of the model are independent and that g_i, e_i and f_{ij} have zero expectations.

From (6) it may be shown that

$$E[w_{qi}-E(\bar{w}_{q.})]\;[x_{rij}-E(\bar{x}_{ri.})]=\tfrac{1}{2}G_{qr}$$

where w_q and x_r represent parent phenotype for the qth character and offspring phenotype for the rth character respectively. Again, the G_{qr} can be estimated unbiasedly. The conventional formulae for $\hat{r}_g$ in this case are

$$\hat{r}_g=\frac{\hat{G}_{qr}+\hat{G}_{rq}}{2(\hat{G}_{qq}\cdot\hat{G}_{rr})^{1/2}} \quad\text{or}\quad \hat{r}_g=\left[\frac{\hat{G}_{qr}\cdot\hat{G}_{rq}}{\hat{G}_{qq}\cdot\hat{G}_{rr}}\right]^{1/2} \tag{7}$$

where $\hat{G}_{qr}$ and $\hat{G}_{rq}$ stand for 2 Cov $(\hat{w}_q x_r)$ and 2 Cov $(\hat{w}_r x_q)$ respectively. Because $\hat{r}_g$ is a function of the $\hat{G}_{qr}$, it is in no way biased by the controllable error terms.

Heritability is usually estimated from the formula

$$\hat{h}^2_q = \frac{2 \text{ Cov } (\hat{w}_q x_q)}{\text{Var } (\hat{w}_q)} \qquad (8)$$

since Var $(\hat{w}_q)$ is taken as an estimate of $G_{qq}+E_{qq}$. This, of course, is only true when $C_{qq}=0$, and for $C_{qq}>0$ a biased estimate of h^2_q is obtained. The amount of bias is again

$$\frac{G_{qq}+E_{qq}}{G_{qq}+E_{qq}+C_{qq}}$$

However, it is worth stressing that C_{qq} is the controllable error variance component in the parent's data.

Formulae for computing the sampling variance of $\hat{r}_g$, Var $(\hat{r}_g)$, estimated from half-sib data have been presented (Tallis, 1959), and similar formulae for parent-offspring data are also available (Reeve, 1955). From a close investigation of these formulae it is clear that, in both types of data, controllable error terms tend to inflate Var $(\hat{r}_g)$. This is due, primarily, to the negative bias incurred by heritability estimates as a result of large C values. Thus, if two systems of data collection are to be compared and the second system (2) involves greater controllable error terms than the first (1), then, provided sufficient data are available Var $(\hat{r}_g)_1$ and Var $(\hat{r}_g)_2$ can be computed and compared. Moreover, if we let

$$\frac{\text{Var } (\hat{r}_g)_2}{\text{Var } (\hat{r}_g)_1} = R > 1$$

and s is the number of parent groups (sire groups or parent-offspring pairs), then if (2) is to provide as much information as (1), $s_2 = Rs_1$ parent groups will be necessary. The loss of efficiency as a result of using (2) instead of (1) is therefore $\frac{R-1}{R}$. This loss of efficiency, together with other factors, could help decide which system of data collection to adopt.

It is concluded from the above analyses that controllable error terms generally:

(i) tend to bias heritability estimates downwards.
(ii) do not bias estimates of genetic correlation.
(iii) increase the errors of estimate of genetic correlations.

Example. In order to illustrate the above results, the two characters birth weight (x_1) and weaning weight (x_2) of Merino sheep will be examined. However, before progress can be made, good estimates of the

genetic variances of x_1 (G_{11}) and x_2 (G_{22}) and the genetic covariance between them (G_{12}) are necessary. These estimates have been obtained from data on the experimental flock of medium Peppin Merino sheep maintained at the C.S.I.R.O. National Field Station, "Gilruth Plains", Cunnamulla. This flock has been described by Turner (1958). The data used were drawn from observations made on lambs and weaners in three mating groups and born during the three years 1954 to 1956. Genetic variance and covariance components were calculated in the conventional way from analyses of variance and covariance tables which had been computed on a within year, mating group and sex basis. These estimates, obtained with 82 degrees of freedom between-sires, and 1282 within-sires, were:

$$\hat{G}_{11}=0.1963 \quad \hat{G}_{22}=3.4526 \quad \hat{G}_{12}=0.2919$$

TABLE 2. VARIANCES OF ESTIMATES OF BIRTH WEIGHT * OBTAINED BY FOUR DIFFERENT METHODS

Group	Degrees of freedom	Variance between lambs
Lambs weighed immediately after birth	34	0.5820
Lambs collected once every 24 hours	39	0.6657
Lambs collected once every 48 hours	76	1.1934
Lambs collected once every 72 hours	108	1.3513

* Pounds.

So that heritability estimates could be calculated based on different procedures of data collection, phenotypic variances for birth weight were estimated during the 1958 lambing at the C.S.I.R.O. field station at Armidale, N.S.W. Four systems of data collection were considered:

a. The weighing of lambs at birth.
b. The weighing of all lambs born during a 24 hour period. Weighings were made daily at 8 a.m.
c. The weighing of all lambs born during a 48 hour period. Weighings were made every alternate day at 8 a.m.
d. The weighing of all lambs born during a 72 hour period. Weighings were made every three days at 8 a.m.

Clearly, the longer the interval between each set of weighings, the larger is the expected phenotypic variance because of the effect of the regression of weight on age. This fact is illustrated by the figures in table 2. From $\hat{G}_{11}$ and the phenotypic variances of table 2, it is possible to obtain a rough idea of the effect on heritability if the estimates of birth weights of lambs are not obtained immediately after birth. For instance, the estimate of heritability under system b of data collection is $\frac{0.1963}{0.6657}=0.295$ and the "bias factor" is calculated as $\frac{0.295}{0.337}=0.875$. In this manner, table 3 has been constructed.

The approximations in table 3 give some idea of the actual magnitude of the bias that can obviously be expected under the four systems of data collection. From these results it seems that undesirably large biases may be incurred if lambs are weighed at intervals longer than 24 hours.

Since system (b) of obtaining birth weights is used at "Gilruth Plains", it was possible to obtain data on weaning weights for a similar analysis. It was found that, if weaning weights are not corrected for age differences at weaning (system f), the estimate of the phenotypic variance is 44.2218. However, this estimate is reduced to 39.1376 after correction for age

TABLE 3. ESTIMATES OF HERITABILITY OF x_1, x_2 and x_2-x_1 UNDER DIFFERENT SYSTEMS OF DATA COLLECTION AND CORRECTION

Character	Type of collection or correction	Heritability	Bias factor
Birth weight (x_1)	a (at birth)	0.337	1.000[a]
	b (every 24 hours)	0.295	0.875
	c (every 48 hours)	0.164	0.487
	d (every 72 hours)	0.145	0.430
Weaning weight (x_2)	corrected for age (e)	0.038	1.000
	uncorrected for age (f)	0.078	0.886
x_2-x_1	a, e	0.091	1.000
	d, f	0.078	0.857

[a] 1.000=no bias.

(system e) is made. Estimates of the heritability of weaning weight under these two systems of data correction are also given in table 3. Thus, although the non-correction of the data for age tends to bias heritability of weaning weight downwards, the bias is not large.

Consider now the heritability (h^2_D) of gains from birth to weaning (x_2-x_1). Estimates of this parameter may be obtained from the expression

$$\hat{h}^2_D=\frac{\hat{G}_{11}+\hat{G}_{22}-2\hat{G}_{12}}{\hat{P}_{11}+\hat{P}_{22}-2\hat{P}_{12}}$$

Since controllable error terms are not expected to be correlated, P_{12} may be estimated and used here for all systems of data collection and correction. An estimate of P_{12} from the Gilruth Plains data is 3.0626, and, with this result, the estimates of h^2_D under systems a and e, and d and f have been calculated and recorded in the last two lines of table 3. Again, a relatively small bias is evident.

Now the genetic correlation between x_1 and x_2 is estimated from equation (3) to be 0.355. Under the two conditions of data collection described for h^2_D, the estimate of r_g is not affected by Var $(\hat{r}_g)_1$ and Var $(\hat{r}_g)_2$ were computed to be

$$\text{Var}\,(\hat{r}_g)_1 = 0.05136$$

$$R = \frac{0.09655}{0.05136} = 1.880$$

$$\text{Var}\,(\hat{r}_g)_2 = 0.09655$$

(The variances of $\hat{r}_g$ have been computed from the formulae developed by Tallis (1959)).

From these figures we can conclude that weighing lambs at three day intervals and not correcting weaning weights for age would result in a loss of efficiency (as defined in Section II) of $\frac{R-1}{R} = 0.47$. Therefore, if the second system of data collection and correction is to yield as much information as the first system, approximately $Rs_1 = 1.880 \times 82 = 154$ sires must be used. This result is rather surprising and re-emphasizes the importance of examining experimental procedures.

Discussion

The above example emphasizes that the effect of controllable errors on the estimation of genetic parameters can be large or small and, therefore, their importance depends on what the investigator's interests are. If, for instance, he wishes to select for birth weight, either to make genetic gains in birth weight or weaning weight, it is clear from the equations

$$\Delta g_1 = i\frac{G_{11}}{(P_{11})^{\frac{1}{2}}}, \qquad \Delta g_2 = i\frac{G_{12}}{(P_{11})^{\frac{1}{2}}}$$

that by inflating P_{11} by means of controllable error terms, drastically reduced expected genetic gains are obtained. Here Δg_1 and Δg_2 are the genetic gains in birth weight and weaning weight respectively and i is the selection differential in standard units.

Moreover, genetic correlations of birth weight with some other characters are estimated very inefficiently when errors involved in the measurement of birth weight are large. On the other hand, if the investigator is only interested in selecting for weight gain from birth to weaning, the type of data collection and correction have little influence on heritability and, hence, on genetic gains.

Another point which is re-emphasized is the necessity, when making predictions of genetic progress, for using estimates of heritability made in the same context. In the present example, the higher estimates of heritability obtained when weighing at birth would not be applicable if lambs were weighed only every three days.

Therefore, it is clear that the design of any equipment, which aims at selection of animals or the estimation of genetic parameters, depends entirely on the characters of interest. Obviously, forethought with the aid of some preliminary investigations may increase the efficiency of the experiment considerably.

Summary

In this paper the effect of some controllable errors on the estimation of heritability and genetic correlation is investigated. It is concluded that, although these controllable errors generally do not bias estimates of genetic correlation, heritability estimates may sustain a severe negative bias. Controllable errors also tend to inflate the sampling variances of estimates of genetic correlation. These findings are illustrated numerically for the two characters birth weight and weaning weight in Merino sheep.

Literature Cited

Reeve, E. C. R. 1955. The variance of the genetic correlation coefficient. Biometrics 11:357.

Tallis, G. M. 1959. Sampling errors of genetic correlation coefficients calculated from analyses of variance and covariance. Australian J. Stat. 1:35.

Turner, Helen Newton. 1958. Relationships among clean wool weight and its components. Australian J. Agr. Res. 9:521.

A (a) (iii)[3]

Commonwealth of Australia.
COMMONWEALTH SCIENTIFIC AND INDUSTRIAL RESEARCH ORGANIZATION.

Reprinted from *The Australian Journal of Statistics*. Vol. 1, No. 2, pages 35-43, August, 1959.

SAMPLING ERRORS OF GENETIC CORRELATION COEFFICIENTS CALCULATED FROM ANALYSES OF VARIANCE AND COVARIANCE

G. M. Tallis
Division of Animal Health and Production, C.S.I.R.O., McMaster Laboratory, Sydney, N.S.W.

SAMPLING ERRORS OF GENETIC CORRELATION COEFFICIENTS CALCULATED FROM ANALYSES OF VARIANCE AND COVARIANCE

G. M. TALLIS*

Summary. In this paper a formula is developed for estimating the sampling variance of a genetic correlation estimated from analyses of variance and covariance. The formula holds provided the heritability estimate of neither character is zero. However, the development assumes a constant number of offspring per sire, k, and the effect of varying values of k is discussed briefly. The efficiency of experiments from which genetic parameters are to be estimated has also been investigated and optimum values of k are given for various combinations of phenotypic and genetic parameters.

I. Introduction. Over the past fifteen years, the pioneer paper of Hazel, Baker and Reinmiller (1943) has been used by workers wishing to obtain estimates of genetic correlation from full-sib and half-sib data. Analyses as outlined in the paper have been applied to livestock, and some estimates based on large numbers of sire groups have appeared recently (Koch and Clark, 1955). To date there has been no general method available for testing the reliability of genetic correlation coefficients estimated in this manner. It is the purpose of this paper to outline a method for calculating sampling variances of such correlation coefficients from the experimental data.

II. Procedure. In the following development it is assumed that the genetic correlation, r_g, has been estimated for two characters, x_1 and x_2, from observations on half-sibs. The additional assumption that x_1 and x_2 are distributed in the normal, bivariate manner is also made. Throughout the discussion, the genetic model used is :

$$x=\mu+g+e, \quad E(g)=E(e)=E(ge)=0$$

where x is phenotype, g and e are respectively genetic and environmental contributions to phenotype, and μ is the population mean. The standard analysis of variance and covariance model for half-sibs is given in Table 1. The model assumes that the number of offspring per sire, k, is constant for all sire groups.

TABLE 1

Analysis of Variance and Covariance

Source	*df*	MS or Cov	*E* (MS) or *E* (Cov)
Between sires ..	d_s	V_{qr}*	$\sigma_{qri}+k\sigma_{qrs}$
Within sires	d_i	v_{qr}	σ_{qri}

* q and r designate two characters, and for mean squares $q=r$.

Received for publication March 20, 1959.

* Division of Animal Health and Production, C.S.I.R.O., McMaster Laboratory, Parramatta Road, Glebe, N.S.W., Australia.

The genetic interpretations of the variance and covariance components are (Hazel and Terrill, 1945):

$$\sigma_{qri} = \tfrac{3}{4}G_{qr} + E_{qr}$$
$$\sigma_{qrs} = \tfrac{1}{4}G_{qr}$$

where G_{qr} and E_{qr} represent genetic and environmental variances and covariances respectively.

Now, by definition, the genetic correlation, r_g, between x_1 and x_2, is estimated by

$$\hat{r}_g = \frac{\hat{G}_{12}}{\sqrt{\hat{G}_{11} \cdot \hat{G}_{22}}} \tag{1}$$

Therefore, the problem is to find an expression for calculating the sampling variance of $\hat{r}_g$, Var $(\hat{r}_g)$.

Taking logarithms of (1), we obtain

$$\log \hat{r}_g = \log \hat{G}_{12} - \tfrac{1}{2}\log \hat{G}_{11} - \tfrac{1}{2}\log \hat{G}_{22}. \tag{2}$$

When (2) is expressed in the form of differentials, squared and expected values taken, the result is

$$\frac{\text{Var}\,(\hat{r}_g)}{r_g^2} = \frac{\text{Var}\,(\hat{G}_{12})}{G_{12}^2} + \frac{\text{Var}\,(\hat{G}_{11})}{4G_{11}^2} + \frac{\text{Var}\,(\hat{G}_{22})}{4G_{22}^2} - \frac{\text{Cov}\,(\hat{G}_{12},\hat{G}_{11})}{G_{12}G_{11}} - \frac{\text{Cov}\,(\hat{G}_{12},\hat{G}_{22})}{G_{12}G_{22}} + \frac{\text{Cov}\,(\hat{G}_{11},\hat{G}_{22})}{2G_{11}G_{22}}. \tag{3}$$

The G_{qr} are estimated from Table 1 in the usual way by means of the following expression:

$$\hat{G}_{qr} = \frac{1}{k}[V_{qr} - v_{qr}]. \tag{4}$$

Hence, the estimates are obtained from linear combinations of mean squares and covariances which are assumed to be independent of each other, i.e. V_{qr} is independent of v_{qr}. When the appropriate expressions for the $\hat{G}_{qr}$ are substituted in (3), it is clear that, in order to solve the problem, it is necessary to know how to calculate Var (V_{qr}) and Cov (V_{qr}, V_{st}); $q, r, s, t = 1, 2$ and a similar set of moments for v_{qr}. From formulae developed by Fisher (1928) it is possible to deduce

$$\text{Var}\,(m_{qr}) = \frac{(\sigma_{qq}\sigma_{rr} + \sigma_{qr}^2)}{(n-1)} \tag{5}$$

$$\text{Cov}\,(m_{qr}, m_{st}) = \frac{(\sigma_{qs}\sigma_{rt} + \sigma_{qt}\sigma_{rs})}{(n-1)}$$

where the m_{qr} are unbiased estimates of the population moments μ_{qr}. Notation used in (5) is not the usual moment notation, but is consistent with Table 1. These results indicate that

$$\text{Var}\,(V_{qr}) \sim \frac{(V_{qq}V_{rr} + V_{qr}^2)}{d_s} \tag{6}$$

$$\text{Cov}\,(V_{qr}, V_{st}) \sim \frac{(V_{qs}V_{rt} + V_{qt}V_{rs})}{d_s}$$

and that similar expressions hold for v_{qr}, in which case, of course, division is by d_i instead of d_s.

From (4) and (6) it is now possible to write terms estimating variances and covariances of genetic components:

$$(7)\qquad \mathrm{Var}\,(\hat{G}_{qr}) \sim \frac{16}{k^2}\left[\frac{V_{qq}V_{rr}+V_{qr}^2}{d_s}+\frac{v_{qq}v_{rr}+v_{qr}^2}{d_i}\right]$$

$$\mathrm{Cov}\,(\hat{G}_{qr},\hat{G}_{st}) \sim \frac{16}{k^2}\left[\frac{V_{qs}V_{rt}+V_{qt}V_{rs}}{d_s}+\frac{v_{qs}v_{rt}+v_{qt}v_{rs}}{d_i}\right]$$

When the expressions (7) are evaluated and substituted in (3), a formula for estimating Var $(\hat{r}_g)$ can be written.

$$(8)\quad \text{Est. Var}\,(\hat{r}_g)=\frac{32\hat{r}_g^2}{k^2}\left[\frac{V_{11}^2/d_s+v_{11}^2/d_i}{4\hat{G}_{11}^2}+\frac{V_{22}^2/d_s+v_{11}^2/d_i}{4\hat{G}_{22}^2}\right.$$
$$+\frac{(V_{11}V_{22}+V_{12}^2)/d_s+(v_{11}v_{22}+v_{12}^2)/d_i}{2\hat{G}_{12}^2}-\frac{V_{11}V_{12}/d_s+v_{11}v_{12}/d_i}{\hat{G}_{11}\hat{G}_{12}}$$
$$\left.-\frac{V_{22}V_{12}/d_s+v_{22}v_{12}/d_i}{\hat{G}_{22}\hat{G}_{12}}+\frac{V_{12}^2/d_s+v_{12}^2/d_i}{2\hat{G}_{11}\hat{G}_{22}}\right]$$

Equation (8) may be used to estimate Var $(\hat{r}_g)$ provided the components of variance and covariance are bounded away from zero. However, the amount of computation is considerable and (8) may be used to derive a more convenient equation for Var $(\hat{r}_g)$. It may be shown that

$$(9)\quad E(V_{11})=(1+(k-1)t_1)P_{11} \qquad E(v_{11})=(1-t_1)P_{11}$$
$$E(V_{22})=(1+(k-1)t_2)P_{22} \qquad E(v_{22})=(1-t_2)P_{22}$$
$$E(V_{12})=r_p+(k-1)r_g t_1^{\frac{1}{2}}t_2^{\frac{1}{2}})P_{11}^{\frac{1}{2}}P_{22}^{\frac{1}{2}} \qquad E(v_{12})=(r_p-r_g t_1^{\frac{1}{2}}t_2^{\frac{1}{2}})P_{11}^{\frac{1}{2}}P_{22}^{\frac{1}{2}}$$

where $t_q=\frac{1}{4}h_q^2$ represents the correlation between half-sibs for x_q, r_p is the phenotypic correlation between x_1 and x_2, and P_{qq} is the phenotypic variance of x_q. By substituting (9) in (8) it will be found that

$$\mathrm{Var}\,(\hat{r}_g)=\frac{1}{k^2d_st_1t_2}[A\{(1+(k-1)t_1)(1+(k-1)t_2)+(r_p+(k-1)B)^2\}$$
$$-2B(r_p+(k-1)B(C+2(k-1))+D]$$
$$(10)\qquad +\frac{1}{k^2d_it_1t_2}[A\{(1-t_1)(1-t_2)+(r_p-B)^2\}$$
$$-2B(r_p-B)(C-2)+D]$$

where $A=1+r_g^2$ $\qquad C=t_1^{-1}+t_2^{-1}$

$B=t_1^{\frac{1}{2}}t_2^{\frac{1}{2}}r_g$ $\qquad D=r_g^2(t_1-t_2)^2/(2t_1t_2)$

Equation (10) can be expressed in a slightly more elegant form by letting

$$r_b=\frac{r_p+(k-1)r_g t_1^{\frac{1}{2}}t_2^{\frac{1}{2}}}{(1+(k-1)t_1)^{\frac{1}{2}}(1+(k-1)t_2)^{\frac{1}{2}}} \qquad r_w=\frac{r_p-r_g t_1^{\frac{1}{2}}t_2^{\frac{1}{2}}}{(1-t_1)^{\frac{1}{2}}(1-t_2)^{\frac{1}{2}}}$$

TABLE 2

Optimum Sizes of Progeny Groups for the Estimation of Genetic Correlation
($n > 1,000$)

t_2	t_1 \ r_p / r_g	0·05					0·10					0·15					0·20					0·30					0·40				
		+0·6	+0·3	0	−0·3	−0·6	+0·6	+0·3	0	−0·3	−0·6	+0·6	+0·3	0	−0·3	−0·6	+0·6	+0·3	0	−0·3	−0·6	+0·6	+0·3	0	−0·3	−0·6	+0·6	+0·3	0	−0·3	−0·6
0·05	+0·6	19	26	36	46	56	15	20	27	34	41	14	18	24	29	35	14	18	22	27	32	14	17	21	25		14	17	20		
	+0·3	18	19	22	26	32	13	13	16	19	23	10	11	13	16	19	9	9	11	14	17	7	8	10	12			7	9	11	
	0	22	20	19	20	22	16	14	13	14	16	12	11	10	11	12	11	9	9	9	11	8	7	7	7	8		6	5	6	
	−0·3	32	26	22	19	18	23	19	16	13	13	19	16	13	11	10	17	14	11	9	9		12	10	8	7		11	9	7	
	−0·6	56	46	36	26	19	41	34	27	20	15	35	29	24	18	14	32	27	22	18	14		25	21	17	14			20	17	14
0·10	+0·6						9	13	17	22	28	7	11	14	18	23	7	10	13	16		6	9	11	14		6	8	11		
	+0·3						9	9	11	13	16	7	7	8	10	13	6	6	7	9	11	4	5	6	7			4	5		
	0						11	9	9	9	11	9	8	7	8	9	7	6	6	6	7		5	5	5			4	4		
	−0·3						16	13	11	9	9	13	10	8	7	7	11	9	7	6	6		7	6	5	4			5	4	
	−0·6						28	22	17	13	9	23	18	14	11	7		16	13	10	7		14	11	9	6			11	8	6
0·15	+0·6											6	8	11	15	*	5	7	10	13		4	6	8			4	6	8		
	+0·3											5	6	7	8	10	5	5	6	7		4	4	5	6			3	4		
	0											7	6	6	6	7	6	5	5	5	6		4	4	4			3	3		
	−0·3											10	8	7	6	5		7	6	5	5		6	5	4	4			4	4	
	−0·6												15	11	8	6		13	10	7	5			8	6	4			8	6	4
0·20	+0·6																4	6	8	11		3	5	7			3	4	6		
	+0·3																4	4	5	6		3	3	4	5			3	3		
	0																5	4	4	4	5		3	3	3				2	2	
	−0·3																	6	5	4	4			4	3	3				3	3
	−0·6																	11	8	6	4			7	5	3				6	4
0·30	+0·6																					2	4	5			2	3			
	+0·3																						2	3	3			2	2		
	0																							2	2	2			2	2	
	−0·3																								2	3				3	3
	−0·6																								3	4				4	4
0·40	+0·6																										2	2			
	+0·3																											2	2		
	0																												2	2	
	−0·3																													2	2
	−0·6																													3	3

NOTE.—t_q = intraclass correlation of character x_q. $\left(\text{In the case of full-sibs, } t_q = \frac{h_q^2}{2} \text{ and for half-sibs, } t_q = \frac{h_q^2}{4}.\right)$

r_g = genetic correlation between characters x_1 and x_2.

r_p = phenotypic correlation between characters x_1 and x_2.

n = total number of offspring.

* Blank entries in the squares on or above the diagonal correspond to combinations of r_p, r_g, t_1, t_2, which cannot satisfy the identity, $r_p = r_g h_1 h_2 + r_e \sqrt{\left(1 - h_1^2\right)\left(1 - h_2^2\right)}$.

The variance of $\hat{r}_g$ may now be written as follows

$$\text{(11)}\quad \text{Var}\,(\hat{r}_g)=\frac{1}{d_s k^2 t_1 t_2}\Big[(1+r_y^2)(1+r_b^2)(1+(k-1)t_1)(1+(k-1)t_2)$$
$$-2r_g r_b[t_1t_2(1+(k-1)t_1)(1+(k-1)t_2)]^{\frac{1}{2}}\left(\frac{1+(k-1)t_1}{t_1}+\frac{1+(k-1)t_2}{t_2}\right)$$
$$+\frac{r_y^2(t_1-t_2)^2}{2t_1t_2}\Big]+\frac{1}{d_i k^2 t_1 t_2}\Big[(1+r_y^2)(1+r_w^2)(1-t_1)(1-t_2)$$
$$-2r_g r_w[t_1t_2(1-t_1)(1-t_2)]^{\frac{1}{2}}\left(\frac{1-t_1}{t_1}+\frac{1-t_2}{t_2}\right)+\frac{r_y^2(t_1-t_2)^2}{2t_1t_2}\Big].$$

Estimates of r_b and r_w are given by

$$\hat{r}_b=\frac{V_{12}}{\sqrt{V_{11}V_{22}}}\qquad \hat{r}_w=\frac{v_{12}}{\sqrt{v_{11}v_{22}}}$$

The author is indebted to Dr. Robertson for suggesting this type of simplification. In fact, when $t_1=t_2$, the formula reduces to the one which he has developed for Var $(\hat{r}_g)$ (Robertson, 1958).

Provided neither t_1 nor t_2 is zero, Var $(\hat{r}_g)$ may be estimated from (11) by replacing the given parameters in the equation with sample estimates.

In the case of $r_g=0$, Var $(\hat{r}_g)$ reduces to

$$\text{Var}\,(\hat{r}_g)=\frac{1}{k^2 d_s t_1 t_2}[(1+(k-1)t_1)(1+(k-1)t_2)+r_p^2]$$
$$\text{(12)}\qquad +\frac{1}{k^2 d_i t_1 t_2}[(1-t_1)(1-t_2)+r_p^2].$$

Thus, if $\hat{r}_g$ is normally distributed with a mean of zero, (12) provides a quick check of the null hypothesis $r_g=0$.

For a given set of parameters, it is possible to calculate k', the value of k which minimizes Var $(\hat{r}_g)$, by letting $d_s=s-1$ and $d_i=s(k-1)$ where s is the number of sires. When $n=sk>1{,}000$, k' may be estimated satisfactorily from the formula

$$\text{(13)}\qquad k'=\sqrt{1+E},$$

where

$$E=\frac{L}{M}=\frac{A[(1-t_1-t_2+r_p(r_p-2B)]+2B[r_p(2-C)+BC]+D}{A(t_1t_2+B^2)-4B^2}>3$$

When $E<3$, no meaningful value of k' can be found. Values of k' for equation (12) may be computed from

$$\text{(14)}\qquad k'=\frac{(1+t_1t_2+r_p^2-t_1-t_2)^{\frac{1}{2}}}{t_1^{\frac{1}{2}}t_2^{\frac{1}{2}}}.$$

Table 2 gives k' for different values of r_g, r_p, t_1 and t_2. This table has been extended to cover the case of full-sibs as well.

It is of interest to note that k' computed from (14) for specific t_1 and t_2 agrees well with the value of k, which minimizes the sum of the sampling variances of t_1 and t_2 (Tallis, 1957). This is convenient because, should an experiment be designed to estimate the heritabilities of several characters as well as the genetic correlations between them then, usually, previous estimates of heritability will be the sole estimates available for computing k'. However, the statement above indicates that the k' value calculated from heritabilities will tend to maximize the probability of finding significant $\hat{r}_g$ values.

If it is desirable to know what minimum sample size, n' is necessary to estimate r_g to a given accuracy, i.e. for fixed Var $(\hat{r}_g)$, n' may be approximated by

$$n' = \frac{[(k'-1)F_1' + F_2']}{\text{Var}\,(\hat{r}_g)k'(k'-1)t_1t_2} \tag{15}$$

where F_1' and F_2' represent the expressions inside the two large brackets of (11), taken in order. The value of k to use in F_1' and F_2' is k'. Thus, n' may be computed for fixed Var $(\hat{r}_g)$ and any given values of t_1, t_2, r_g and r_p.

When c different r_g parameters are to be estimated from the one group of animals, a value for k' which minimizes $\sum_{i=1}^{c} \text{Var}\,(\hat{r}_g)i$ can be computed from

$$k' = \sum_{i=1}^{c} L_i t_{1i}^{-1} t_{2i}^{-1} \Big/ \sum_{i=1}^{c} M_i t_{1i}^{-1} t_{2i}^{-1} \tag{16}$$

where L and M are as defined in (13).

Equations (8) to (16) are readily applicable to full-sib analyses. Under these circumstances V_{qr} and v_{qr} represent dam and within dam mean squares and covariances, d_s and d_i are between and within dam degrees of freedom and $t_q = \frac{h_q^2}{2}$ is the within-dam full-sib correlation of x_q. With these modifications, and by changing the constant in front of the large bracket in (8) to $\frac{8\hat{r}_g^2}{k^2}$, the last seven equations can be used without further alteration.

Finally, it must be stressed that the distribution of $\hat{r}_g$ is unknown and, therefore, normal tests of significance are not necessarily appropriate. Moreover, throughout this development, k has been assumed constant. The work of Hammersley (1949) indicates that varying values of k should result in a larger estimate of Var $(\hat{r}_g)$. Hence, equations (8) and (11) give the limiting case and the assumption of a fixed k results in a minimum estimate.

Example. In order to illustrate the use of some of the above formulae, the genetic correlation between clean wool weight and number of crimps per inch of staple has been computed for data from a flock of medium-wool Peppin Merino sheep.

The flock is maintained at the C.S.I.R.O. National Field Station, "Gilruth Plains", Cunnamulla, and has been described by Turner

TABLE 3

Analyses of Variance and Covariance of Clean Wool Weight (x_1) and Number of Crimps per Inch (x_2)

Source of Variation	d.f.	Sums of Squares and Products			Variances and Covariances			E (MS) or E (COV)
		x_1^2	x_1x_2	x_2^2	x_1	x_1x_2	x_2	
Between rams (within years and mating groups) ..	51	74·1535	—111·3525	592·9779	1·4540	—2·1834	11·6270	$\sigma_{qri}+16\cdot75\sigma_{qrs}$
Within rams	802	407·2043	—432·8590	2998·9501	0·5077	—0·5397	3·7393	σ_{qri}

q, $r=1$, 2.

(1958). The data used were drawn from observations made on ewes 16 months old in three experimental mating groups in the four years 1954 to 1957. Sums of squares and cross products, which were computed on a within year and within mating group basis, are given in Table 3. Estimates of h_1^2, h_2^2, t_1, t_2, r_g and r_p were obtained from the phenotypic and genetic variances and covariances recorded in Table 4.

TABLE 4

Phenotypic and Genetic Variances and Covariances

Character	Phenotypic		Genetic	
	x_1	x_2	x_1	x_2
x_1	0·5642	−0·6379	0·2260	−0·3926
x_2		4·2103		1·8838

$\hat{h}_1^2 = 0{\cdot}45$ $\quad \hat{t}_1 = 0{\cdot}11$ $\quad d_s = 51$

$\hat{h}_2^2 = 0{\cdot}40$ $\quad \hat{t}_2 = 0{\cdot}10$ $\quad d_i = 802$

$\hat{r}_p = -0{\cdot}52$ $\quad \hat{r}_g = -0{\cdot}61$ $\quad k = 16{\cdot}75$

$\hat{r}_b = -0{\cdot}53$ $\quad \hat{r}_w = -0{\cdot}39$

For practical purposes $\hat{t}_1 = \hat{t}_2 = 0{\cdot}10$, and with this small alteration it will be found that Est. Var $(\hat{r}_g)$ is 0·0198. Hence the standard error of $\hat{r}_g$ is of the order of 0·14.

In order to test the hypothesis $r_g = 0$, Est. Var $(\hat{r}_g)$ can be computed from (12) to be approximately 0·052. Hence, this hypothesis could be rejected with considerable confidence.

If it were desired to design an experiment so that $\hat{r}_g$ would have a standard error of approximately 0·05, the optimum value of k, k', to use (from Table 2) is about 10. The minimum number of animals, n', necessary to obtain an estimate with this size error may be computed from (15). In this case

$$\hat{F}_1' = 1{\cdot}565, \quad \hat{F}_2' = 0{\cdot}393, \quad \text{Var}\,(\hat{r}_g) = 0{\cdot}0025 \quad \text{and} \quad n' = 6{,}435.$$

It is, therefore, concluded that approximately 640 rams and 6,400 ewes would be required for such an experiment.

Acknowledgments. The author wishes to thank Miss Helen Newton Turner of the McMaster Laboratory, Division of Animal Health and Production, C.S.I.R.O.; Dr. H. S. Konijn, Senior Lecturer in Statistics in the Faculty of Economics at the University of Sydney; and Dr. Alan Robertson, Institute of Genetics, Edinburgh, for their most valuable assistance in the preparation of this paper. Special acknowledgment is also made of the contributions of Mrs. Nancy Carter and Mrs. Fay Guinane of the McMaster Laboratory. Mrs. Carter checked the development of formula (13), while Mrs. Guinane undertook the laborious task of computing Tables 2 and 3.

Data used in the numerical example were obtained with the collaboration of Mr. C. H. S. Dolling and the staff of the National Field Station, " Gilruth Plains ", Cunnamulla, Queensland, where the sheep are maintained. Fleece measurements were made by the staff of the Fleece Analysis Section of the Sheep Biology Laboratory, Prospect, under the direction of the Officer-in-Charge, Mr. R. E. Chapman.

Dr. B. D. H. Latter, of the Division of Plant Industry, C.S.I.R.O., developed the same formulae independently just after this paper had been prepared. The author is grateful for his helpful comments on the manuscript.

References

Fisher, R. A. (1928). " Moments and Product Moments of Sampling Distributions." *Proc. Lond. math. Soc.*, 30, 199–238.

Hammersley, J. M. (1949). " The Unbiased Estimate and Standard Error of the Interclass Variance." *Metron*, 15, 189–205.

Hazel, L. N., Baker, M. L., and Reimmiller, C. F. (1943). " Genetic and Environmental Correlations between the Growth Rates of Pigs at Different Ages." *J. Anim. Sci.*, 2, 118–128.

Hazel, L. N., and Terrill, C. E. (1945). " Heritability of Weaning Weight and Stable Length in Range Rambouillets Lambs." *J. Anim. Sc.*, 4, 347–358.

Kendall, M. G. (1943). *The Advanced Theory of Statistics*, Vol. I. Griffin & Co., London.

Koch, R. M., and Clark, R. T. (1955). " Genetic and Environmental Relationships among Economic Characters in Beef Cattle. I. Correlation among Paternal and Maternal Half-Sibs." *J. Anim. Sci.*, 14, 775–785.

Robertson, A. (1958). " The Sampling Variance of the Genetic Correlation Coefficient." *Biometrics*. (In press.)

Tallis, G. M. (1957). Ph.D. Dissertation. The Ohio State University.

Turner, Helen Newton (1958). " Relationships among Clean Wool Weight and its Components." *Aust. J. Agric. Res.*, 9, 521–552.

AUSTRALASIAN MEDICAL PUBLISHING CO. LTD.
SEAMER AND ARUNDEL STS., GLEBE, SYDNEY

A (a) (iii)[4]

Commonwealth of Australia.
COMMONWEALTH SCIENTIFIC AND INDUSTRIAL RESEARCH ORGANIZATION.

Reprinted from *The Australian Journal of Statistics*, Vol. 2, No. 2, pages 66-77, August, 1960.

THE SAMPLING ERRORS OF ESTIMATED GENETIC REGRESSION COEFFICIENTS AND THE ERRORS OF PREDICTED GENETIC GAINS

G. M. Tallis
Division of Animal Genetics, C.S.I.R.O., McMaster Laboratory, Glebe, N.S.W.

THE SAMPLING ERRORS OF ESTIMATED GENETIC REGRESSION COEFFICIENTS AND THE ERRORS OF PREDICTED GENETIC GAINS[1]

G. M. TALLIS

* Division of Animal Genetics, C.S.I.R.O., McMaster Laboratory, Glebe, N.S.W.

Introduction. The theory of selection indices has been presented by Fairfield Smith (1936) and Hazel (1943) and later discussed by Lush (1945), Lerner (1950), Morley (1950) and many others. Moreover, Hazel and Lush (1942) and Young (1959) have shown that this method of selection, when applied to several characters, is never less efficient, in terms of genetic gains, than any other known selection technique. It is mainly for this reason that many selection indices have been calculated for livestock over the past 15 years in order to assist in establishing genetic progress.

To date relatively little attention has been given to the sampling variances of estimated genetic regression coefficients and estimates of genetic gain. Some aspects of the problem have been discussed by Bartlett (1939) and Nanda (1949), but their results are of little use to the animal breeder who requires a general theory which is relatively easy to put into practice. It is the purpose of this paper to develop such a theory and to apply it to certain special cases.

The genetic value of an animal, relative to the population as a whole, for n economically important characters may be written as

$$H=\sum_{i=1}^{n} a_i g_i, \tag{1}$$

where a_i and g_i are the relative economic weight and the genetic value of the i^{th} character, measured from the mean genotype of the population. In the subsequent development it will be assumed that it is possible to express H as a multiple regression model of the form

$$H=\sum_{i=1}^{n} \beta_i x_i+\varepsilon \tag{2}$$

where x_i is the phenotype of the i^{th} character (measured from the population mean phenotype) and ε is a normal random error component. A selection index, I, will now be defined as the best linear predictor of H, i.e.

$$I=\sum_{i=1}^{n} \beta_i x_i \tag{3}$$

The β_i are in effect genetic partial regression coefficients but, for simplicity, in this paper the β_i are referred to as genetic regression coefficients.

The function I can be found by determining the β_i which satisfy

$$\partial E(H-\sum_{i=1}^{n} \beta_i x_i)^2/\partial \beta_i=0, \quad i=1, 2, \ldots, n$$

By assuming the usual additive genetic model

$$x_i=g_i+e_i, \; E(g_i)=E(e_i)=E(g_i e_i)=0$$

[1] Received for publication February 11, 1960; revised June 13, 1960.

in which e_i is the normally distributed environmental component of the phenotype of the i^{th} character, it is readily verified that

$$\mathbf{P}\boldsymbol{\beta}=\mathbf{G}\mathbf{a}$$

(4) or

$$\boldsymbol{\beta}=\mathbf{P}^{-1}\mathbf{G}\mathbf{a}$$

where $\boldsymbol{\beta}$ and $\mathbf{a}$ are the column vectors of the β_i and a_i and P and G are the variance-covariance matrices for the x_i and g_i respectively. As a consequence of (4), it follows that

$$(4a) \qquad \beta_i=\sum_{qr} p^{iq}g_{qr}a_r$$

where p^{iq} and g_{qr} are the $i.q^{th}$ and $q.r^{th}$ elements of $\mathbf{P}^{-1}$ and G respectively. These results are in agreement with those of Fairfield Smith (1936).

There are two main practical methods of estimating the β_i from sample data :

Estimation Method (*a*).

The β_i may be estimated from relationships between parents and offspring. The form of analysis used in this method is equivalent to an ordinary multiple regression analysis.

Estimation Method (*b*).

In this method the matrices P and G are estimated from relationships between full-sibs or half-sibs. The $\hat{\beta}_i$ are then calculated as

$$(5) \qquad \hat{\boldsymbol{\beta}}=\hat{\mathbf{P}}^{-1}\hat{\mathbf{G}}\mathbf{a}.$$

The above two methods of constructing an estimate of I will now be considered in more detail.

1. Sampling Errors of Estimated Genetic Regression Coefficients.

(i) I estimated by method (*a*).

In this instance, the complete analysis is relatively simple. In order to show this we consider the specific case where data have been collected on dam-offspring pairs. In this instance the usual statistical models are

$$(6) \qquad \left.\begin{aligned} x_i&=g_i+e_i \\ y_i&=\tfrac{1}{2}g_i+f_i \end{aligned}\right\} E(g)=E(e)=E(f)=E(ge)=E(gf)=E(ef)=0$$

where x_i is the phenotypic measurement of the i^{th} character in the parent, y_i is the phenotypic measurement of the same character in the offspring, and e and f are independent, normal errors. In equations (6), all measurements are taken from the respective population means.

Now consider the variable $z=\sum_{i=1}^{n} y_i a_i$. If we try to estimate z from the x_i and write the appropriate multiple regression equation as

$$(7) \qquad z=\sum_{i=1}^{n}\gamma_i x_i+\delta,\ E(\delta)=0$$

where δ is a normal error component, it is found by computing

$$\partial E(z-\Sigma\gamma_i x_i)^2/\partial\gamma_i=0, \quad i=1, 2, \ldots, n$$

and using (6), that

$$\text{(8)} \qquad \mathbf{P}\gamma = \tfrac{1}{2}\mathbf{G}a \quad \text{or} \quad 2\gamma = \mathbf{P}^{-1}\mathbf{G}a.$$

For this reason the least squares estimates of the γ_i in fact estimate $\frac{1}{2}\beta_i$. As exact tests of significance are available for estimates of multiple regression coefficients, confidence limits for the β_i can be constructed simply.

More explicitly, if a sample of N parent-offspring pairs are available to estimate I, the procedure is to solve the n equations

$$\text{(9)} \quad \gamma_1\Sigma(X_i-\bar{X}_i)(X_1-\bar{X}_1)+\ldots+\gamma_n\Sigma(X_i-\bar{X}_i)(X_n-\bar{X}_n) = \Sigma(X_i-\bar{X}_i)(Z-\bar{Z}), \quad i=1, 2, \ldots, n$$

for the γ_i, where $X_i=\mu_i+x_i$, $Z=\mu_z+z$ and the summation extends over all N pairs. By noticing that

$$\Sigma(X_i-\bar{X}_i)(Z-\bar{Z}) = \Sigma(X_i-\bar{X}_i)(Y_1-\bar{Y}_1)a_1+\ldots+\Sigma(X_i-\bar{X}_i)(Y_n-\bar{Y}_n)a_n$$

and letting X and Y be the $n\times n$ matrices of the terms

$$\Sigma(X_i-\bar{X}_i)(X_j-\bar{X}_j) \text{ and } \Sigma(X_i-\bar{X}_i)(Y_j-\bar{Y}_j) \text{ respectively,}$$

the n equations for the γ_i can be written in matrix notation as $\mathbf{X}\gamma=\mathbf{Y}a$, and hence $\gamma=\mathbf{X}^{-1}\mathbf{Y}a$. If we denote the elements of $\mathbf{X}^{-1}$ as x^{ij}, then it is well known that the statistic

$$t=(\hat{\gamma}_i-\gamma_i)/s(x^{ii})^{\frac{1}{2}}, s^2=\Sigma[(Z-\bar{Z})-\sum_{i=1}^{n}\hat{\gamma}_i(X_i-\bar{X}_i)]^2/N-n-1$$

has the t distribution with $(N-n-1)$ degrees of freedom. Since $\hat{\beta}_i=2\hat{\gamma}_i$, the confidence intervals for the β_i are twice those of the γ_i.

(ii) I estimated by method (*b*).

The more complicated case of estimating I from full-sib or half-sib data will now be considered. The general analysis of variance and covariance model is given in Table 1, where the number of offspring within sub-groups, k, is assumed constant for all sub-groups. Genetic interpretations of the expected mean squares are:

$$\text{(10)} \qquad w_{qr}=(m-1)g_{qr}/m+e_{qr}$$
$$\sigma_{qrs}=g_{qr}/m$$

TABLE 1

Analysis of Variance and Covariance

Source	*df*	MS or Cov	E(MS) or E(Cov)
Between sires ..	df_b	$\hat{b}_{qr}$*	$b_{qr}=w_{qr}+k\sigma_{qrs}$
Within sires ..	df_w	$\hat{w}_{qr}$	w_{qr}

* q and r designate two characters, and for mean squares $q=r$.

where g_{qr} and e_{qr} represent genetic and environmental variances and covariances and m takes the values 2 and 4 for full-sibs and half-sibs respectively. From these relationships it is clear that the genetic parameter g_{qr} and the phenotypic parameter p_{qr} are estimated by

(11) $$\hat{g}_{qr}=[\hat{b}_{qr}-\hat{w}_{qr}]m/k$$
$$\hat{p}_{qr}=[\hat{b}_{qr}+(k-1)\hat{w}_{qr}]/k.$$

One easy method of obtaining the large sample variance-covariance matrix of the $\hat{\beta}_i$ is to differentiate the equation $\mathrm{P}\beta=\mathrm{Ga}$ directly, remembering that the a_i are constants. We have

$$d\mathrm{P}\beta+\mathrm{P}d\beta=d\mathrm{Ga}$$

and
$$d\beta=\mathrm{P}^{-1}[d\mathrm{Ga}-d\mathrm{P}\beta].$$

Some rearrangement of the last expression using equations (11) and replacing $\hat{b}_{qr}$ and $\hat{w}_{qr}$ by b_{qr} and w_{qr} gives

$$kd\beta=\mathrm{P}^{-1}[d\mathrm{Bu}-d\mathrm{Wv}]$$

where u and v are column vectors of the quantities $(a_jm-\beta_i)$ and $(a_im+(k-1)\beta_i)$ respectively, k is a scalar and B and W are $n\times n$ matrices with elements b_{qr} and w_{qr}. If $kd\beta$ is post-multiplied by its transpose, we obtain

$$k^2d\beta d\beta^{\mathrm{T}}=\mathrm{P}^{-1}[d\mathrm{Bu}-d\mathrm{Wv}][\mathrm{u}^{\mathrm{T}}d\mathrm{B}-\mathrm{v}^{\mathrm{T}}d\mathrm{W}]\mathrm{P}^{-1}$$

since P^{-1}, $d\mathrm{B}$ and $d\mathrm{W}$ are symmetrical. When expected values of both sides of the last expression are taken and $E(d\beta_i d\beta_j)$, $E(db_{qr}db_{st})$, $E(dw_{qr}dw_{st})$ and $E(db_{qr}dw_{st})$ are associated with Cov $(\hat{\beta}_i,\hat{\beta}_j)$, Cov $(\hat{b}_{qr},\hat{b}_{st})$, Cov $(\hat{w}_{qr},\hat{w}_{st})$ and Cov $(\hat{b}_{qr},\hat{w}_{st})=0$ respectively for all possible values of the subscripts, the result is

(12) $$k^2[\mathrm{Cov}\ (\hat{\beta}_i,\hat{\beta}_j)]=\mathrm{P}^{-1}\{E[d\mathrm{Buu}^{\mathrm{T}}d\mathrm{B}]+E[d\mathrm{Wvv}^{\mathrm{T}}d\mathrm{W}]\}\mathrm{P}^{-1}$$
$$=\mathrm{P}^{-1}\{E[d\mathrm{BC}d\mathrm{B}]+E[d\mathrm{WD}d\mathrm{W}]\}\mathrm{P}^{-1}$$

writing $\mathrm{C}=\mathrm{uu}^{\mathrm{T}}$ and $\mathrm{D}=\mathrm{vv}^{\mathrm{T}}$.

In order to evaluate (12) explicitly, consider firstly $E[d\mathrm{BC}d\mathrm{B}]$. Let

$$\mathrm{C}d\mathrm{B}=\mathrm{L}=[l_{sr}]=[\sum_{t=1}^{n}c_{st}db_{tr}]$$

and
$$d\mathrm{BL}=\mathrm{M}=[m_{qr}]=[\sum_{s}db_{qs}l_{sr}]$$

It is now clear that

$$m_{qr}=\sum_{s}db_{qs}\sum_{t}c_{st}db_{tr}$$
$$=\sum_{s}\sum_{t}c_{st}db_{tr}db_{qs}.$$

Hence, $E(m_{qr})=\sum_{s}\sum_{t}c_{st}$ Cov $(\hat{b}_{tr},\hat{b}_{qs})$ and the required expression must finally be

(13) $$k^2[\mathrm{Cov}\ (\hat{\beta}_i,\hat{\beta}_j)]=\mathrm{P}^{-1}[\mathrm{M}+\mathrm{N}]\mathrm{P}^{-1}$$

where $\mathrm{M}=[m_{qr}]=[\sum_{s}\sum_{t}c_{st}$ Cov $(\hat{b}_{tr},\hat{b}_{qs})]$

$\mathrm{N}=[n_{qr}]=[\sum_{s}\sum_{t}d_{st}$ Cov $(\hat{w}_{tr},\hat{w}_{qs})]$

Formulae for estimating such terms as Cov $(\hat{b}_{tr},\hat{b}_{qs})$ and Cov $(\hat{w}_{tr},\hat{w}_{qs})$ have been presented by Tallis (1959) and are given below.

(14) $$\mathrm{Cov}\ (\hat{b}_{tr},\hat{b}_{qs})\sim(\hat{b}_{tq}\hat{b}_{rs}+\hat{b}_{ts}\hat{b}_{rq})/df_b$$
$$\mathrm{Cov}\ (\hat{w}_{tr},\hat{w}_{qs})\sim(\hat{w}_{tq}\hat{w}_{rs}+\hat{w}_{ts}\hat{w}_{rq})/df_w$$

By identifying appropriate elements of $[\widehat{\mathrm{Cov}}(\hat{\beta}_i,\hat{\beta}_j)]$ with $\mathbf{P}^{-1}[\mathbf{M}+\widehat{\mathbf{N}}]\mathbf{P}^{-1}/k^2$, estimates of the sampling variances of individual $\hat{\beta}_i$ may be obtained.

2. Sampling Errors of Predicted Genetic Gains. A genetic gain, Δ_g, may be defined as the average genetic superiority of a selected sub-group of animals over the average genotype of the particular group from which selection is made. To be more specific, let N_p be the number in the original group before selection and suppose a sample of N_s animals is selected from this group. We have from (1) and (2) of the previous section

$$\bar{H}_p=\sum_{i=1}^{n} a_i\bar{g}_{pi}=\sum_{i=1}^{n}\beta_i\bar{x}_{pi}+\bar{\varepsilon}_p \tag{15}$$

$$\bar{H}_s=\sum_{i=1}^{n} a_i\bar{g}_{si}=\sum_{i=1}^{n}\beta_i\bar{x}_{si}+\bar{\varepsilon}_s$$

The subscripts p and s refer to the original group and the selected sub-group respectively. Now, by definition

$$\Delta_g=\bar{H}_s-\bar{H}_p=\sum_{i=1}^{n}\beta_i(\bar{x}_{si}-\bar{x}_{pi})+\bar{\varepsilon}_s-\bar{\varepsilon}_p \tag{16}$$

$$=\sum_{i=1}^{n}\beta_i(\bar{X}_{si}-\bar{X}_{pi})+\bar{\varepsilon}_s-\bar{\varepsilon}_p$$

where $\bar{X}_{pi}$ and $\bar{X}_{si}$ are the observed means of the i^{th} character in the original group and the selected group respectively. Upon substituting, $d_i=\bar{X}_{si}-\bar{X}_{pi}$ (not to be confused with the d_{ij} of the previous section) and $\bar{\varepsilon}=\bar{\varepsilon}_s-\bar{\varepsilon}_p$, (2) becomes

$$\Delta_g=\Sigma\beta_i d_i+\bar{\varepsilon} \tag{17}$$

and in this form Δ_g is analogous to the original model for H. The best linear estimate of Δ_g is

$$\Delta_g'=\sum_{i=1}^{n}\beta_i d_i \tag{18}$$

and an estimate of Δ_g is given by

$$\hat{\Delta}_g=\sum_{i=1}^{n}\hat{\beta}_i d \tag{19}$$

where the $\hat{\beta}_i$ are calculated by either of the two methods discussed in Section 1.

Estimation Method (*a*).

As above, this method will be discussed in relation to the analysis of dam-offspring pairs. It was found earlier that $\hat{\beta}_i=2\hat{\gamma}_i$, so that

$$E(\hat{\Delta}_g-\Delta_g)^2=\sigma_\tau^2=\mathrm{Var}\left(\sum_{i=1}^{n}\hat{\beta}_i d_i\right)+\mathrm{Var}\,(\bar{\varepsilon})$$

$$=4\mathrm{Var}\left(\sum_{i=1}^{n}\hat{\gamma}_i d_i\right)+[1/N_s+1/N_p]\sigma^2$$

$$=[4\mathbf{d}^{\mathrm{T}}\mathbf{X}^{-1}\mathbf{d}+1/N_s+1/N_p]\sigma^2$$

where $\mathbf{d}$ is the column vector of the d_i. Moreover, if s^2 is the unbiased estimate of σ^2, then the statistic

$$(20) \quad t=\frac{(\hat{\Delta}_g-\Delta_g)/\sigma_{\hat{\tau}}}{s/\sigma}=(\hat{\Delta}_g-\Delta_g)/s(1+\mathbf{d}^{\mathrm{T}}\mathbf{X}^{-1}\mathbf{d}+1/N_s+1/N_p)^{\frac{1}{2}}$$

is t distributed with $(N-n-1)$ degrees of freedom. Thus, confidence limits may be set in the usual way.

Estimation Method (b).

Unfortunately, it is not possible to obtain exact results in this case. We have from the previous section,

$$\hat{\boldsymbol{\beta}}=\dot{\mathbf{P}}^{-1}\dot{\mathbf{G}}\mathbf{a}$$

where the elements of $\dot{\mathbf{P}}$ and $\dot{\mathbf{G}}$ are obtained from analyses of variance and covariance tables. By proceeding in a similar way to Estimation Method (a), we obtain

$$\begin{aligned} E(\hat{\Delta}_g-\Delta_g)^2=\sigma_{\hat{\tau}}^2&=\mathrm{Var}\,(\Sigma_i\hat{\beta}_i d_i)+(1/N_s+1/N_p)\sigma^2 \\ &=\mathbf{d}^{\mathrm{T}}[\mathrm{Cov}\,(\hat{\beta}_i,\hat{\beta}_j)]\mathbf{d}+(1/N_s+1/N_p)\sigma^2, \end{aligned} \quad (21)$$

where $\sigma^2=\sigma_H^2(1-R_{IH}^2)$.

In the above equation R_{IH} is the correlation coefficient between I and H and σ_H^2 is the variance of H. In matrix notation,

$$\sigma_H^2=\mathbf{a}^{\mathrm{T}}\mathbf{G}\mathbf{a},\quad \sigma_I^2=\boldsymbol{\beta}^{\mathrm{T}}\mathbf{P}\boldsymbol{\beta},$$

$$\sigma_{IH}=\boldsymbol{\beta}^{\mathrm{T}}\mathbf{G}\mathbf{a}=\boldsymbol{\beta}^{\mathrm{T}}\mathbf{P}\mathbf{P}^{-1}\mathbf{G}\mathbf{a}=\boldsymbol{\beta}^{\mathrm{T}}\mathbf{P}\boldsymbol{\beta}=\sigma_I^2$$

and hence $\sigma^2=\mathbf{a}^{\mathrm{T}}\mathbf{G}\mathbf{a}(1-\boldsymbol{\beta}^{\mathrm{T}}\mathbf{P}\boldsymbol{\beta}/\mathbf{a}^{\mathrm{T}}\mathbf{G}\mathbf{a})$. An estimate of $\sigma_{\hat{\tau}}^2$, $\hat{\sigma}_{\hat{\tau}}^2$, is obtained by replacing $[\mathrm{Cov}\,(\hat{\beta}_i,\hat{\beta}_j)]$, $\mathbf{G}$ and $\mathbf{P}$ by their estimates computed from the sample data. Approximate confidence intervals for Δ_g can now be set by assuming

$$\hat{t}=(\hat{\Delta}_g-\Delta_g)/\hat{\sigma}_{\hat{\tau}}$$

is normally distributed with zero mean and unit variance.

3. Special Cases. The methods so far developed are entirely general. However, because of the wide interest and application of certain special cases, it seems desirable at this stage to consider two of these in some detail. The notation of previous sections is used here without further explanation.

Case 1—*Estimation of* g_i *from* x_i. The appropriate models for this case are obtained by setting $a_s=\beta_s=0$, $s\neq i$, and $a_i=1$ in (1). From (4 a) it is clear that $\beta_i=G_{ii}/P_{ii}$ and this ratio is known as the heritability of the i^{th} character and is written h_i^2. We have, therefore,

$$H=g_i=h_i^2x_i+\varepsilon.$$

Estimation Method (a).

Since, in this instance, y_i and x_i are assumed to be binormally distributed, with zero means, variances equal to P_{ii} and with a correlation coefficient of $h_i^2/2$, the problem reduces to one of ordinary simple linear regression. We have

$$\hat{h}_i^2=2\Sigma(Y_i-\bar{Y}_i)(X_i-\bar{X}_i)/\Sigma(X_i-\bar{X}_i)^2$$

$$\hat{\Delta}_g=\hat{h}_i^2(\bar{X}_{si}-\bar{X}_{pi})=\hat{h}_i^2d_i$$

$$t=(\hat{h}_i^2-h_i^2)[\Sigma(X_i-\overline{X}_i)^2]^{\frac{1}{2}}/2s,s^2=\frac{4\Sigma(Y_i-\overline{Y}_i)^2-(\hat{h}_i^2)^2\Sigma(X_i-\overline{X}_i)^2}{4(N-2)}$$

$$t=(\hat{\Delta}_g-\Delta_g)/s[4d_i^2/\Sigma(X_i-\overline{X}_i)^2+1/N_s+1/N_p]^{\frac{1}{2}}$$

Both t variates are distributed with $N-2$ degrees of freedom.

Estimation Method (*b*).

In order to consider special cases, it is convenient to rewrite (13) as

$$k^2\mathrm{P}[\mathrm{Cov}\ (\hat{\beta}_i,\hat{\beta}_j)]\mathrm{P}=\mathrm{M}+\mathrm{N}.$$

Now, if only certain β_q are to be considered, cross out all rows and columns of P, [Cov $(\hat{\beta}_i,\hat{\beta}_j)$], M and N which do not have q as a subscript and substitute $\beta_i=0$, $i\neq q$, in C and D. This procedure will now be demonstrated for Case 1.

Using the model discussed above, we have

$$c_{ii}=(m-h_i^2)^2, d_{ii}=[m+(k-1)h_i^2]^2,$$

and all other elements of the C and D matrices are zero because $a_s=\beta_s=0$, $s\neq i$. By use of the formulae for m_{qr} and n_{qr} it is found that

$$k^2P_{ii}\ \mathrm{Var}\ (\hat{h}_i^2)P_{ii}=\mathrm{Var}\ (\hat{b}_{ii})(m-h_i^2)^2+\mathrm{Var}\ (\hat{w}_{ii})[m+(k-1)h_i^2]^2$$

It is informative to simplify the above formula by means of the relations

$$\mathrm{Var}\ (\hat{b}_{ii})=2b_{ii}^2/df_b,\ \mathrm{Var}\ (\hat{w}_{ii})=2w_{ii}^2/df_w$$

$$E(\hat{b}_{ii})=b_{ii}=[1+(k-1)t_i]P_{ii},\ E(\hat{w}_{ii})=w_{ii}=(1-t_i)P_{ii}$$

where $t_i=h_i^2/m$ is the intraclass correlation among full-sibs ($m=2$) or half-sibs ($m=4$). After some simplification, the formula becomes

$$k^2\ \mathrm{Var}\ (\hat{h}_i^2)=2m^2[1+(k-1)t_i]^2(1-t_i)^2\{1/df_b+1/df_w\}$$

Setting $df_b=s-1$ (s=number of classes) and $df_w=s(k-1)$, the formula for Var $(\hat{h}_i^2)$ assumes the familiar form

$$\mathrm{Var}\ (\hat{h}_i^2)\simeq 2m^2(1-t_i)^2[1+(k-1)t_i]^2/(s-1)k(k-1).$$

The sampling variance for a predicted genetic gain is

$$E(\hat{\Delta}_g-\Delta_g)^2=\sigma_\tau^2=d_i^2\ \mathrm{Var}\ (\hat{h}_i^2)+[1/N_s+1/N_p](1-h_i^2)G_{ii}$$

An estimate of σ_τ^2, $\hat{\sigma}_\tau^2$, is obtained by substituting the relevant parameter estimates into the above equation. Approximate confidence limits may be set from

$$\hat{t}=(\hat{\Delta}_g-\Delta_g)/\hat{\sigma}_\tau$$

which is treated as a standard normal variate.

Case 2—Estimation of g_i from x_j. The correct model is obtained by setting $a_s=0$, $s\neq i$, $a_i=1$, $\beta_s=0$, $s\neq j$. This gives

$$H=g_i=\beta_j x_j+\varepsilon$$

where $\beta_j=G_{ij}/P_{jj}$ (4 *a*).

Estimation Method (*a*).

Case 2 is entirely analogous to Case 1 as x_j and y_i are assumed to be binormally distributed with zero means, variances P_{jj} and P_{ii}

respectively and correlation coefficient $\rho_j = G_{ij}/2(P_{ii}P_{jj})^{\frac{1}{2}}$. Hence the following results

$$\hat{\beta}_j = 2\Sigma(Y_i - \bar{Y}_i)(X_j - \bar{X}_j)/\Sigma(X_j - \bar{X}_j)^2$$
$$\hat{\Delta}_g = \hat{\beta}_j(\bar{X}_{sj} - \bar{X}_{pj}) = \hat{\beta}_j d_j$$
$$t = (\hat{\beta}_j - \beta_j)[\Sigma(X_j - \bar{X}_j)^2]^{\frac{1}{2}}/2s, s^2 = [\Sigma(Y_i - \bar{Y}_i)^2 - \hat{\beta}_j^2\Sigma(X_j - \bar{X}_j)^2]/N - 2$$
$$t = (\hat{\Delta}_g - \Delta_g)/s[\tfrac{1}{4}d_j^2\Sigma(X_j - \bar{X}_j)^2 + 1/N_s + 1/N_p].$$

Both t variables are distributed with $N-2$ degrees of freedom. It should be stressed that in the above formulae Δ_g refers to a genetic gain in the i^{th} character since $a_s = 0$, $s \neq i$.

Estimation Method (*b*)

In this instance it is readily verified that

$$\mathbf{C} = \begin{bmatrix} m^2 & -m\beta_j \\ -m\beta_j & \beta_j^2 \end{bmatrix} \qquad \mathbf{D} = \begin{bmatrix} m^2 & -m(k-1)\beta_j \\ -m(k-1)\beta_j & (k-1)^2\beta_j^2 \end{bmatrix}$$

and that

$$k^2 P_{jj}^2 \operatorname{Var}(\hat{\beta}_j) = [m^2(\operatorname{Var}(\hat{b}_{ij}) + \operatorname{Var}(\hat{w}_{ij}))$$
$$+ \beta_j^2(\operatorname{Var}(\hat{b}_{jj}) + (k-1)^2 \operatorname{Var}(\hat{w}_{jj}))$$
$$- 2m\beta_j(\operatorname{Cov}(\hat{b}_{ij}, \hat{b}_{jj}) + (k-1)\operatorname{Cov}(\hat{w}_{ij}, \hat{w}_{jj}))].$$

(The author has been unable to obtain an interesting expression for the above formula in terms of genetic parameters.)

The sampling variance of a predicted genetic gain in g_i is

$$E(\hat{\Delta}_g - \Delta_g)^2 = \sigma_{\bar{z}}^2 = d_j^2 \operatorname{Var}(\hat{\beta}_j) + (1 - \rho_j^2)G_{ii}[1/N_s + 1/N_p)],$$

where $\rho_j = G_{ij}/(G_{ii}P_{jj})^{\frac{1}{2}}$.

Approximate confidence intervals may again be obtained from the formula

$$\hat{t} = (\hat{\Delta}_g - \Delta_g)/\hat{\sigma}_{\bar{z}}$$

where $\hat{\sigma}_{\bar{z}}$ is an estimate of $\sigma_{\bar{z}}$.

4. Extensions. In practice, it is often desirable to estimate the mean of the progeny from selected parents. The above formulae are easily modified to take care of this situation. These modifications will only be indicated in the case of formula (20) as the results are readily applied to special cases.

Let the genetic deviation of a selected group of males from the unselected group be

$$\Delta_{g1} = \sum_{i=1}^{n} \beta_i(\bar{X}_{si}^1 - \bar{X}_{pi}^1) + \bar{\varepsilon}_{s1} - \bar{\varepsilon}_{p1}$$

where $\bar{X}_{si}^1$ and $\bar{X}_{pi}^1$ are the means of the selected and unselected groups of males for the i^{th} character respectively. If a similar expression is written for a selected group of females by replacing 1 and 2, and if the selected males are mated randomly with the selected females, then the expected genetic gain of the offspring is

$$\bar{\Delta}_g = (\Delta_{g1} + \Delta_{g2})/2 = \tfrac{1}{2}\Big\{\sum_{i=1}^{n} \beta_i[(\bar{X}_{si}^1 + \bar{X}_{si}^2) - (\bar{X}_{pi}^1 + \bar{X}_{pi}^2)] +$$
$$+ \bar{\varepsilon}_{s1} + \bar{\varepsilon}_{s2} - \bar{\varepsilon}_{p1} - \bar{\varepsilon}_{p2}\Big\}.$$

TABLE 2

Analyses of Variance and Covariance of Clean Wool Wt.(X_1) and Number of Crimps per Inch (X_2)

Source of Variation	d.f.	Sums of Squares and Products			Variances and Covariances			E(MS) or E(Cov)
		X_1^2	X_1X_2	X_2^2	X_1^2	X_1X_2	X_2^2	
Between rams (within years and mating groups)	51	74·1535	−111·3525	592·9779	1·4540	−2·1834	11·6270	$w_{qr}+16\cdot75\sigma_{qrs}$
Within rams	802	407·2043	−432·8590	2998·9501	0·5077	−0·5397	3·7393	w_{qr}

$q, r=1, 2$

Writing $\bar{d}_i=[(\bar{X}^1_{si}+\bar{X}^2_{si})-(\bar{X}^1_{pi}+\bar{X}^2_{pi})]$, it follows from previous results that if the β_i are estimated by method (a)

$$E(\hat{\bar{\Delta}}_g-\bar{\Delta}_g)^2=\sigma^2_{\bar{\tau}}=(4\bar{\mathbf{d}}^{\mathrm{T}}\mathbf{X}^{-1}\bar{\mathbf{d}}+1/N_{s1}+1/N_{s2}+1/N_{p1}+1/N_{p2})\sigma^2/4$$

If Estimation Method (b) has been used,

$$E(\hat{\bar{\Delta}}_g-\bar{\Delta}_g)^2=\sigma^2_{\bar{\tau}}=\{\bar{\mathbf{d}}^{\mathrm{T}}[\mathrm{Cov}(\hat{\beta}_i,\hat{\beta}_j)]\bar{\mathbf{d}}+\sigma^2(1/N_{s1}+1/N_{s2}+1/N_{p1}+1/N_{p2})\}/4$$

where $\sigma^2=(1-R^2_{IH})\sigma^2_H$. In the above formulae $\bar{\mathbf{d}}$ is the column vector of the $\bar{d}_i$. Confidence intervals may now be set by the techniques of the previous sections.

5. Example. In order to illustrate some of the above results, a selection index will be calculated for clean wool weight (X_1) and the number of crimps per inch of the wool staple (X_2). The analyses of variance and covariance of X_1 and X_2 appear in Table 2 and the genotypic and phenotypic variances and covariances are recorded in Table 3. The data from which these estimates were computed come from the flock described by Turner, Dolling and Sheaffe (1959). Values of a_1 and a_2 will be taken as 4 and 1 respectively (Dunlop and Young (1960)).

TABLE 3

Phenotypic and Genetic Variances and Covariances

Character	Phenotypic		Genetic	
	X_1	X_2	X_1	X_2
X_1	0·564229	−0·637853	0·225972	−0·392516
X_2		4·210245		1·883624

As the determinant of $\hat{\mathbf{P}}$, $|\hat{\mathbf{P}}|$, is 1·968,686, the matrix $\hat{\mathbf{P}}^{-1}$ is

$$\hat{\mathbf{P}}^{-1}=\begin{bmatrix} 2\cdot138607 & 0\cdot323999 \\ 0\cdot323999 & 0\cdot286602 \end{bmatrix}$$

and we have

$$\begin{bmatrix}\hat{\beta}_1\\ \hat{\beta}_2\end{bmatrix}=\begin{bmatrix}2\cdot138607 & 0\cdot323999\\ 0\cdot323999 & 0\cdot286602\end{bmatrix}\begin{bmatrix}0\cdot225972 & -0\cdot392516\\ -0\cdot392516 & 1\cdot883624\end{bmatrix}\begin{bmatrix}4\\ 1\end{bmatrix}$$

Multiplication of these matrices gives $\hat{\beta}_1=1\cdot1952$ and $\hat{\beta}_2=0\cdot2556$. Hence,

$$I=1\cdot20X_1+0\cdot26X_2.$$

In order to evaluate $[\mathrm{Cov}(\hat{\beta}_i,\hat{\beta}_j)]$ it is first necessary to evaluate the M and N matrices for the particular case $n=2$. If the expression for the individual elements of M is expanded, we obtain

$$\begin{aligned}
m_{11}&=\mathrm{Var}\,(\hat{b}_{11})c_{11}+2\,\mathrm{Cov}\,(\hat{b}_{12},\hat{b}_{11})c_{12}+\mathrm{Var}\,(\hat{b}_{12})c_{22}\\
m_{12}=m_{21}&=\mathrm{Cov}\,(\hat{b}_{11},\hat{b}_{12})c_{11}+\mathrm{Cov}\,(\hat{b}_{11},\hat{b}_{22})c_{12}\\
&\quad+\mathrm{Var}\,(\hat{b}_{12})c_{12}+\mathrm{Cov}\,(\hat{b}_{12},\hat{b}_{22})c_{22}\\
m_{22}&=\mathrm{Var}\,(\hat{b}_{12})c_{11}+2\,\mathrm{Cov}\,(\hat{b}_{12},\hat{b}_{22})c_{12}+\mathrm{Var}\,(\hat{b}_{22})c_{22}
\end{aligned}$$

The N matrix is obtained by replacing $\hat{b}_{qr}$ and c_{qr} by $\hat{w}_{qr}$ and d_{qr} in the M matrix above. We may now proceed:

$$\hat{\mathbf{C}}=\begin{bmatrix}16-\hat{\beta}_1\\ 4-\hat{\beta}_2\end{bmatrix}[16-\hat{\beta}_1,\ 4-\hat{\beta}_2]=\begin{bmatrix}219\cdot181600 & 55\cdot435755\\ 55\cdot435755 & 14\cdot020898\end{bmatrix}$$

$$\hat{\mathbf{D}}=\begin{bmatrix}16+(k-1)\hat{\beta}_1\\ 4+(k-1)\hat{\beta}_2\end{bmatrix}[16+(k-1)\hat{\beta}_1,\ 4+(k-1)\hat{\beta}_2]=$$

$$=\begin{bmatrix}1212\cdot757484 & 279\cdot465460\\ 279\cdot465460 & 64\cdot399473\end{bmatrix}$$

since $k-1=15\cdot75$, $\hat{\beta}_1=1\cdot195$ and $\hat{\beta}_2=0\cdot256$.

From formulae (13) it is possible to calculate

Var $(\hat{b}_{11})$	$=0\cdot082905$	Var $(\hat{w}_{11})$	$=0\cdot000643$
Var $(\hat{b}_{12})$	$=0\cdot424955$	Var $(\hat{w}_{12})$	$=0\cdot002731$
Var $(\hat{b}_{22})$	$=5\cdot301472$	Var $(\hat{w}_{22})$	$=0\cdot034869$
Cov $(\hat{b}_{11},\hat{b}_{22})$	$=-0\cdot124495$	Cov $(\hat{w}_{11},\hat{w}_{12})$	$=-0\cdot000683$
Cov $(\hat{b}_{11},\hat{b}_{22})$	$=0\cdot186947$	Cov $(\hat{w}_{11},\hat{w}_{22})$	$=0\cdot000726$
Cov $(\hat{b}_{12},\hat{b}_{22})$	$=-0\cdot995538$	Cov $(\hat{w}_{12},\hat{w}_{22})$	$=-0\cdot005033$

and matrices $\hat{\mathbf{M}}$ and $\hat{\mathbf{N}}$ become

$$\hat{\mathbf{M}}=\begin{bmatrix}10\cdot326669 & -7\cdot324003\\ -7\cdot324003 & 57\cdot096995\end{bmatrix}\qquad \hat{\mathbf{N}}=\begin{bmatrix}0\cdot573541 & -0\cdot186791\\ -0\cdot186791 & 2\cdot743998\end{bmatrix}$$

Finally, adding $\hat{\mathbf{M}}$ and $\hat{\mathbf{N}}$ we have

$$(16\cdot75)^2[\mathrm{Cov}\,\widehat{(\hat{\beta}_i,\hat{\beta}_j)}]=$$

$$\begin{bmatrix}2\cdot138607 & 0\cdot323999\\ 0\cdot323999 & 0\cdot286602\end{bmatrix}\begin{bmatrix}10\cdot900211 & -7\cdot510794\\ -7\cdot510294 & 59\cdot840993\end{bmatrix}\times\begin{bmatrix}2\cdot138607 & 0\cdot323999\\ 0\cdot323999 & 0\cdot286602\end{bmatrix}$$

and $[\mathrm{Cov}\,\widehat{(\hat{\beta}_i,\hat{\beta}_j)}]=\begin{bmatrix}0\cdot162983 & 0\cdot027507\\ 0\cdot027507 & 0\cdot016626\end{bmatrix}$

From these results the estimated 95% confidence intervals for β_1 and β_2 are

$$\beta_1=1\cdot20\pm0\cdot78\qquad \beta_2=0\cdot26\pm0\cdot25$$

In order to complete this example, let 10 rams be selected from a flock of 100. Suppose that for the two groups

$$\bar{X}_{s1}=9\qquad \bar{X}_{p1}=8\qquad d_1=1$$
$$\bar{X}_{s2}=9\qquad \bar{X}_{p2}=10\qquad d_2=-1$$

where the subscripts 1 and 2 again refer to wool weight and number of crimps per inch respectively. An estimate of Δ_g is

$$\hat{\Delta}_g=1\cdot20-0\cdot26=0\cdot94$$

and

$$\hat{\sigma}_{\bar{\tau}}^2=[1,\ -1]\begin{bmatrix}0\cdot1629S3 & 0\cdot027507\\ 0\cdot027507 & 0\cdot016626\end{bmatrix}\begin{bmatrix}1\\ -1\end{bmatrix}+\left(\frac{1}{10}+\frac{1}{100}\right)\hat{\sigma}_H^2(1-\hat{R}_{IH}^2)$$

The estimates $\hat{\sigma}_H^2$ and $\hat{R}_{IH}^2$ are most easily computed from the formulae

$$\hat{\sigma}_H^2=\mathbf{a}^{\mathrm{T}}\hat{\mathbf{G}}\mathbf{a} \text{ and } \hat{R}_{IH}^2=\hat{\boldsymbol{\beta}}^{\mathrm{T}}\hat{\mathbf{P}}\hat{\boldsymbol{\beta}}/\hat{\sigma}_H^2.$$

Appropriate calculations give

$$\hat{\sigma}_{\bar{\tau}}^2=0\cdot1246+0\cdot11\times2\cdot3590(1-0\cdot2931)=0\cdot3080.$$

Hence, the estimated 95% confidence interval for Δ_g is $-0\cdot15$ to $2\cdot03$.

Acknowledgements. The author is indebted to Dr. S. S. Y. Young of the Division of Animal Genetics for his assistance with notation and checking of formulae. Special acknowledgement is also made of the work of Mrs. Fay Guinane, who performed most of the computations in the numerical example.

Data used in the numerical example were obtained with the collaboration of Mr. C. H. S. Dolling and the staff of the National Field Station, "Gilruth Plains", Cunnamulla, Queensland, where the sheep are maintained. Fleece measurements were made by the staff of the Fleece Analysis Section of the Ian Clunies Ross Animal Research Laboratory, Prospect, under the direction of the Officer-in-Charge, Mr. R. E. Chapman.

References

Bartlet, M. S. (1939). The standard errors of discriminant function coefficients. *J. roy. statist. Soc.*, Suppl. 6, 169–173.

Dunlop, A. A., and Young, S. S. Y. (1960). Selection of Merino sheep : An analysis of the relative economic weights applicable to some wool traits. *Emp. J. Exp. Agric.*, 28 (in press).

Hazel, L. N. (1943). The genetic basis for constructing selection indexes. *Genetics*, 28, 476–490.

Hazel, L. N., and Lush, J. L. (1942). The efficiency of three methods of selection. *J. Hered.*, 33, 393–399.

Lush, J. L. (1945). "Animal Breeding Plans." (Collegiate Press : Ames, Iowa.)

Mood, A. M. (1950). "Introduction to the Theory of Statistics." (McGraw-Hill : New York.)

Morley, F. H. W. (1950). Ph.D. Thesis, Iowa State College.

Nanda, D. N. (1949). The standard errors of discriminant function coefficients in plant breeding experiments. *J. roy. statist. Soc.*, B, 11, 283–290.

Smith, Fairfield H. (1936). A discriminant function for plant selection. *Ann. Eugen. Lond.*, 7, 240–250.

Tallis, G. M. (1959). Sampling errors of genetic correlation coefficients calculated from analyses of variance and covariance. *Aust. J. Statist.*, 1, 35–43.

Turner, Helen Newton, Dolling, C. H. S., and Sheaffe, P. H. G. (1959). Vital statistics for an experimental flock of Merino sheep. *Aust. J. Agric. Res.* 10, 581–590.

Young, S. S. Y. (1959). Ph.D. Thesis, University of Sydney.

AUSTRALASIAN MEDICAL PUBLISHING CO. LTD.
SEAMER AND ARUNDEL STS., GLEBE, SYDNEY

A (a) (iii) [5]

SAMPLING ERRORS ASSOCIATED WITH FAMILY SELECTION

G. M. Tallis

Division of Animal Genetics, McMaster Laboratory, C. S. I. R. O., Glebe, New South Wales, Australia.

BIOMETRICS (1964) 20, 118-121

The theory of family selection has been investigated from various points of view by many workers over the past few years. This work was recently reviewed by Young [1961] who further developed some aspects of this form of selection. However, the coefficients of the family selection index have to be estimated and, to date, the errors of estimation have not been examined. It is the aim here to apply the methods of a previous paper, Tallis [1960], to family selection.

We consider the variates $\bar{I}$, $\bar{D}$, $\bar{F}$, $\bar{S}$ and $\bar{O}$, where

$\bar{I}$ = mean of m records of the individual
$\bar{D}$ = mean of k records of the individual's dam
$\bar{F}$ = mean of j records of the individual's sire
$\bar{S}$ = mean of n half-sibs each with m records
$\bar{O}$ = mean of q offspring

and construct an index, using the above five variates, to best estimate g, the additive genotype of an individual for a particular character. To avoid unnecessary constants, all measurements are assumed to be made from respective population means. We then let

$$g = b_1\bar{I} + b_2\bar{D} + b_3\bar{F} + b_4\bar{S} + b_5\bar{O} + \epsilon \qquad (1)$$

and proceed to calculate the b_i by least squares.

We now need the usual models for parent and offspring

$$x = g + e$$

$$y = \tfrac{1}{2}g + f$$

where x and y are the phenotypic measurements of parent and offspring respectively, e and f are random normal error components, g is the additive genetic contribution to phenotype, and

$$E(g) = E(e) = E(f) = E(ge) = E(ef) = 0. \qquad (2)$$

Both x and y are measured from their respective population means. With the aid of (2) it is possible to verify that the covariance matrix for the six variates, g, $\bar{I}$, $\bar{D}$, $\bar{F}$, $\bar{S}$ and $\bar{O}$ is

$$\begin{array}{c} \\ \bar{I} \\ \bar{D} \\ \bar{F} \\ \bar{S} \\ \bar{O} \\ g \end{array} \begin{array}{c} \begin{array}{cccccc} \bar{I} & \bar{D} & \bar{F} & \bar{S} & \bar{O} & G \end{array} \\ \begin{bmatrix} PM & \frac{1}{2}G & \frac{1}{2}G & \frac{1}{4}G & \frac{1}{2}G & G \\ \frac{1}{2}G & PK & 0 & 0 & \frac{1}{4}G & \frac{1}{2}G \\ \frac{1}{2}G & 0 & PJ & \frac{1}{2}G & \frac{1}{4}G & \frac{1}{2}G \\ \frac{1}{4}G & 0 & \frac{1}{2}G & \frac{P}{n}\{M + (n-1)t\} & \frac{1}{8}G & \frac{1}{4}G \\ \frac{1}{2}G & \frac{1}{4}G & \frac{1}{4}G & \frac{1}{8}G & PQ & \frac{1}{2}G \\ G & \frac{1}{2}G & \frac{1}{2}G & \frac{1}{4}G & \frac{1}{2}G & G \end{bmatrix} \end{array}$$

where

$$M = \{1 + (m-1)\rho\}/m \qquad J = \{1 + (j-1)\rho\}/j$$
$$K = \{1 + (k-1)\rho\}/k \qquad Q = \{1 + (q-1)t\}/q,$$

G and P represent the additive genetic and phenotypic variances for the character considered, ρ is the correlation between repeated records (repeatability) and $t = h^2/4$ is the correlation between half-sibs.

From the above dispersion matrix we define another matrix

$$\mathbf{A} = \mathbf{D} + h^2\mathbf{B}$$

with

$$\mathbf{D} = \text{diag}\,(M, K, J, M/n, 1/q)$$

and

$$\mathbf{B} = \begin{bmatrix} 0 & \frac{1}{2} & \frac{1}{2} & \frac{1}{4} & \frac{1}{2} \\ \frac{1}{2} & 0 & 0 & 0 & \frac{1}{4} \\ \frac{1}{2} & 0 & 0 & \frac{1}{2} & \frac{1}{4} \\ \frac{1}{4} & 0 & \frac{1}{2} & \frac{n-1}{4n} & \frac{1}{8} \\ \frac{1}{2} & \frac{1}{4} & \frac{1}{4} & \frac{1}{8} & \frac{q-1}{4q} \end{bmatrix}.$$

Now let $\mathbf{x} = \text{col}(1, \frac{1}{2}, \frac{1}{2}, \frac{1}{4}, \frac{1}{2})$ and $\mathbf{b}$ be the vector of the b_i, $i = 1, 2, \cdots 5$, then the least squares solution for the b_i is given by

$$\mathbf{b} = h^2\mathbf{A}^{-1}\mathbf{x}. \qquad (3)$$

In order to obtain $[C(\hat{b}_i, \hat{b}_i)]$ we write

$$\hat{\mathrm{A}}\hat{\mathrm{b}} = \hat{h}^2\mathrm{x} = (\mathrm{A} + \Delta\mathrm{A})(\mathrm{b} + \Delta\hat{\mathrm{b}}) = (h^2 + \Delta h^2)\mathrm{x},$$

whence

$$\Delta\mathrm{b} = \mathrm{A}^{-1}\{\Delta h^2\mathrm{x} - \Delta\mathrm{A}\mathrm{b} - \Delta\mathrm{A}\Delta\mathrm{b}\}.$$

Now

$$\Delta\mathrm{A} = \Delta\mathrm{D} + \Delta h^2\mathrm{B} \quad \text{and} \quad \Delta\mathrm{D} = \Delta\rho\mathrm{R}$$

where

$$\mathrm{R} = \mathrm{diag}\left(\frac{m-1}{m}, \frac{k-1}{k}, \frac{j-1}{j}, \frac{m-1}{mn}, 0\right)$$

and hence, neglecting the term $\Delta\mathrm{A}\Delta\mathrm{b}$,

$$\Delta\mathrm{b} = \mathrm{A}^{-1}\{\Delta h^2[\mathrm{x} - \mathrm{B}\mathrm{b}] - \Delta\rho\mathrm{R}\mathrm{b}\}$$

and

$$[C(\hat{b}_i, \hat{b}_i)] = E(\Delta\mathrm{b}\Delta\mathrm{b}')$$
$$= \mathrm{A}^{-1}[V(\hat{h}^2)(\mathrm{x} - \mathrm{B}\mathrm{b})(\mathrm{x} - \mathrm{B}\mathrm{b})' + V(\hat{\rho})\mathrm{R}\mathrm{b}\mathrm{b}'\mathrm{R}]\mathrm{A}^{-1}, \qquad (4)$$

since it is assumed that h^2 and ρ are independently estimated. Formulae for $V(\hat{h}^2)$ and $V(\hat{\rho})$ are well known or else readily accessible, no matter how h^2 and ρ are estimated. Thus, if $\hat{g} = \hat{\mathrm{b}}'\mathrm{y}$, $\mathrm{y} = \mathrm{col}(\bar{I}, \bar{D}, \bar{F}, \bar{S}, \bar{O})$, then

$$E(\hat{g} - g)^2 = \mathrm{y}'[C(\hat{b}_i, \hat{b}_i)]\mathrm{y} + \sigma^2,$$

where $\sigma^2 = G(1 - R^2)$ and R is the correlation between the index $\mathrm{b}'\mathrm{y}$ and g.

Since the amount of information on family performance may vary with the individual, separate vectors b may have to be computed for each animal. This could represent a great deal of computation and may, in fact, be uneconomical unless good computing facilities are readily accessible. In the next paragraph it will be assumed that equal information is available for each individual considered. Finally, this restriction will be lifted in order to obtain a more general formula.

One of the main uses of $[C(\hat{b}_i, \hat{b}_i)]$ is in the investigation of errors of predicted genetic gains. Following the notation of Tallis [1960], a genetic gain, Δg, is defined as the average genetic superiority (or inferiority if Δg is negative) of a selected sub-group of animals over the average genotype of the particular group from which selection is made. To be more specific, let N_p be the number in the original group before selection, and suppose a sample of N_s animals is selected from this group.

Then

$$\Delta g = \bar{g}_s - \bar{g}_p$$

where

$$\bar{g}_s = \sum_{q=1}^{N_s} g_q/N_s \quad \text{and} \quad \bar{g}_p = \sum_{q=1}^{N_p} g_q/N_p .$$

Now

$$\hat{\Delta g} = \hat{\bar{g}}_s - \hat{\bar{g}}_p = \hat{\mathrm{b}}'(\bar{\mathrm{y}}_s - \bar{\mathrm{y}}_p) = \hat{\mathrm{b}}'(\bar{\mathrm{Y}}_s - \bar{\mathrm{Y}}_p) = \hat{\mathrm{b}}'\bar{\mathrm{y}}$$

where $\bar{\mathrm{Y}}_s$ and $\bar{\mathrm{Y}}_p$ are the mean vectors for the two groups uncorrected for the population mean and $\bar{\mathrm{y}} = \bar{\mathrm{Y}}_s - \bar{\mathrm{Y}}_p$. Now,

$$E(\hat{\Delta g} - \Delta g)^2 = \mathrm{y}'[C(\hat{b}_i , \hat{b}_j)]\mathrm{y} + \sigma^2\left(\frac{1}{N_s} - \frac{1}{N_p}\right). \tag{5}$$

If the amount of family information varies with the individual the results are less pleasant. Let $\hat{g}_q = \hat{\mathrm{b}}_q'\mathrm{y}_q$ be the estimate of the additive genotype of the qth animal, then

$$V(\hat{g}_q) = \mathrm{y}_q'[C(b_i , b_j)]\mathrm{y}_q + G(1 - R_q^2) = \sigma_q^2$$

where R_q is the correlation of g with $\mathrm{b}_q'\mathrm{y}$. Moreover, since it is readily verified that $C(\hat{g}_q , \hat{g}_r) = 0$, $q \neq r$, we have

$$V(\hat{\Delta g}) = V(\hat{\bar{g}}_s - \hat{\bar{g}}_p) = \sum_{q=1}^{N_p} \sigma_q^2/N_p^2 + \sum_{r=1}^{N_s} \sigma_r^2/N_s^2 - 2\sum_{r=1}^{N_s} \sigma_r^2/N_sN_p . \tag{6}$$

If N_p is large, then $V(\hat{\Delta g}) \simeq \sum_{r=1}^{N_s} \sigma_r^2/N_s^2$. By assuming that $\hat{\Delta g}$ is approximately normally distributed, confidence intervals for Δg may be calculated from (5) and (6) by standard methods.

In the development of (6) it was assumed that all measurements were made from respective population means. However, in practice, these would have to be estimated in order to calculate the vectors y_q. It is readily verified that this procedure introduces an error of order N^{-1}, where N is the minimum population size from which the means are estimated.

Obviously special cases using two or more dependent variables may be considered by deleting appropriate elements from b, D, B, x and R. However, there seems to be little point in carrying out the operations algebraically since the matrix formulae are probably the most satisfactory for numerical computation.

REFERENCES

Tallis, G. M. [1960]. The sampling errors of estimated genetic regression coefficients and the errors of predicted genetic gains. *Aust. J. Statist. 3*, 66–77.

Young, S. S. Y. [1961]. The use of sire's and dam's records in animal selection. *Heredity 16*, 91–102.

A (a) (iii)[6]

Biometrika (1969), 56, 3, *p.* 517
Printed in Great Britain

Exact first- and second-order moments of estimates of components of covariance

BY C. A. ROHDE
Johns Hopkins University

AND G. M. TALLIS
C.S.I.R.O., Newtown, New South Wales

SUMMARY

This paper develops formulae for the first two moments of estimates of covariance for the general multivariate 'one-way' and 'two-way' models. The results are used to obtain the large sample dispersion matrix for estimated coefficients of two types of genetic selection indexes. These dispersion matrices provide the necessary extension of known results in balanced models to the unbalanced case.

1. INTRODUCTION

The problems of estimation associated with variance and covariance component analysis with unbalanced data have been of major concern to the statistical geneticist. This is primarily because most of the selection procedures applied to livestock require a knowledge of genetic variances and covariances which usually have to be estimated from the analyses of hierarchical models. Invariably, there is a marked lack of balance in the data thus rendering standard formulae for the variances of the estimates inapplicable.

Serious consideration to these problems has been given by Henderson (1953), Searle (1956) and Hartley & Rao (1967). Searle gave particular attention to the one-way analysis of variance and covariance and used matrix methods to calculate the moments of the various estimators. Other work in this area concerns analysis of variance models of varying complexity; see Searle (1958, 1961), Mahamunulu (1963) and Blischke (1966).

It is the purpose of the present paper to extend and complement existing results. With the advent of high speed computers, matrix operations can be handled with great speed and hence formulae for expectations and covariances of sums of squares and products can be left in a general computable form. Thus, explicit algebraic evaluation of each case is, in most cases, not only time-consuming but unnecessary.

We consider here the general one-way and two-way analysis of covariance model with fixed and random effects. The number of variables included in the analysis is assumed to be arbitrary and this seems to lead to somewhat involved notation and algebra. However, general results are required in order to solve a number of practical problems. We give two examples from statistical genetics.

In the theory of animal breeding interest centres around certain phenotypic and genetic parameters. Suppose that k characters of a particular breed of animal are relevant from the point of view of a selection programme. Then we let P and G be the phenotypic and additive genetic covariance matrices for the k characters and we consider two types of selection index which are based on these matrices.

We first consider the case where the breeder wishes to move the means of the k characters in his group of animals to certain predetermined optimum values. It can be shown that an index of the form $I = \beta'X$, $\beta = G^{-1}\alpha$, where X is the vector of the values of the k characters and α is a vector of k known constants which are the distances of the current means of the characters from their optimal values, has the required properties (Tallis, 1968). Unfortunately, G is unknown and usually it is estimated from an analysis of data on family groups. In such cases it is often the 'one-way' model which is appropriate and it turns out in fact that $\hat{G} = m\hat{C}$, where $\hat{C}$ is given by (4) of § 3 and m is a constant depending on the type of family group studied.

The estimate of β is given by $\hat{\beta} = \hat{G}^{-1}\alpha$ and the large sample covariance matrix for $\hat{\beta}$ in the balanced case is given by Tallis (1968). However, in most cases the data are unbalanced and we give the general formula for var $(\hat{\beta})$ in § 3·1.

A second type of index has been developed to assist the breeder to make the maximum economic advance by selection. Let a be the vector of economic weights pertaining to the k characters, i.e. the weights that specify the relative importance of each character to the breeder. Then, it is well known that an index $\beta'X$, $\beta = P^{-1}Ga$, leads to an optimal selection procedure. This result was first proved by Fairfield Smith (1936) but a simpler derivation is given by Tallis (1968). Both P and G can be estimated from a one-way analysis of covariance and the balanced case has been treated by Tallis (1960). Again, we give the general expression for var $(\hat{\beta})$ in § 3·1.

In order to cope with cases where the design matrix associated with a particular set of data is singular, the methods appropriate to the solution of least squares equations subject to constraints have been used in § 3·1. However, in § 3·2 standard results employing generalized inverses are involved. Both these techniques are discussed briefly in § 2·1.

There are a number of reasons for writing the results in terms of the two techniques. Some users may be more at home with, for instance, the method of linear constraints and such readers will have little difficulty in writing all the results in those terms. Moreover, this technique may be more manageable computationally and, in fact, may be preferable for other reasons as well.

On the other hand, analysis of variance theory is most conveniently discussed in terms of generalized inverses. These concepts have been developed by Rao (1965) and his notation is used subsequently. However, as pointed out above, the most general framework is not always useful for particular applications.

2. METHODS

2·1. *Notation*

It is well known that the least squares estimate of β for the linear model $\mathbf{Y} = \mathbf{X}\beta + \epsilon$, where X is an $n \times p$ matrix of known coefficients, $E(\epsilon) = 0$ and $E(\epsilon\epsilon') = \sigma^2\mathbf{I}$, is given by

$$\mathbf{b} = \mathbf{S}^{-1}\mathbf{X}'\mathbf{y},$$

when $\mathbf{S} = \mathbf{X}'\mathbf{X}$ is non-singular. However, when the rank of X, $\rho(\mathbf{X})$, is such that $\rho(\mathbf{X}) = t < p$, S does not have an inverse in the usual sense and other methods must be resorted to.

Consider the $(p-t) \times p$ matrix H, $\rho(\mathbf{H}) = p - t$. When X is a design matrix, the least squares equations are usually solved employing a set of constraints $\mathbf{Hb} = \mathbf{c}$, say, where H is as above and $\rho(\mathbf{X}'\mathbf{H}') = p$. In fact it can be shown that, under the above conditions, the system of equations

$$\mathbf{X}'\mathbf{X}\mathbf{b} = \mathbf{X}'\mathbf{y}, \quad \mathbf{H}\mathbf{b} = \mathbf{c}$$

is consistent and has the unique solution

$$\mathbf{b} = \mathbf{S}^{-1}\mathbf{X}'\mathbf{y} + \mathbf{S}^{-1}\boldsymbol{\gamma},$$

where $\mathbf{S} = \mathbf{X}'\mathbf{X} + \mathbf{H}'\mathbf{H}$ and $\boldsymbol{\gamma} = \mathbf{H}'\mathbf{c}$. It follows easily that, if $\mathbf{H}\boldsymbol{\beta} = \mathbf{c}$, regarding $\mathbf{b}$ as a random vector $E(\mathbf{b}) = \boldsymbol{\beta}$, $\text{var}(\mathbf{b}) = \sigma^2\mathbf{S}^{-1}\mathbf{X}'\mathbf{X}\mathbf{S}^{-1}$ and $\rho\{\text{var}(\mathbf{b})\} = t$. Usually, $\mathbf{c} = 0$; this will be assumed subsequently.

Under normality assumptions it can also be shown that, since $\mathbf{X}\mathbf{S}^{-1}\mathbf{X}'$ is idempotent, the sum of squares $\mathbf{Y}'\mathbf{Y}$ can be decomposed into two independently distributed quadratic forms; $\mathbf{Y}'\mathbf{X}\mathbf{S}^{-1}\mathbf{X}'\mathbf{Y}$, the sum of squares due to regression and $\mathbf{Y}'(\mathbf{I} - \mathbf{X}\mathbf{S}^{-1}\mathbf{X}')\mathbf{Y}$, the error sum of squares. The degrees of freedom associated with the first sum of squares is t and, with the second, $(n-t)$.

Suppose now that X, H and $\boldsymbol{\beta}$ are partitioned as $\mathbf{X} = [\mathbf{X}_1, \mathbf{X}_2]$, $\mathbf{H} = [\mathbf{H}_1, \mathbf{H}_2]$ and $\boldsymbol{\beta}' = [\boldsymbol{\beta}_1', \boldsymbol{\beta}_2']$, where $[\mathbf{X}_i', \mathbf{H}_i']$ is of full rank for $i = 1, 2$. Again, this would be the case in most design situations and we would have the obvious additional property that $\mathbf{H}_1'\mathbf{H}_2 = 0$.

For example, in a randomized block experiment, $\mathbf{X}_1$ and $\mathbf{X}_2$ could be the incidence matrices for blocks and treatments respectively (Graybill, 1961, p. 225). In this case, if there are b blocks and t treatments,

$$\mathbf{H}_1 = \begin{bmatrix} 0 & 1_b' \\ 0 & 0_b' \end{bmatrix}, \quad \mathbf{H}_2 = \begin{bmatrix} 0_t' \\ 1_t' \end{bmatrix},$$

where, for instance, Y_b' is a row vector consisting of b 1's and $0_b'$ is a row vector of b 0's. These $\mathbf{H}_i$ matrices impose the usual constraints,

$$\sum_{i=1}^{b} \beta_i = \sum_{i=1}^{t} \tau_i = 0$$

and clearly $\mathbf{H}_1'\mathbf{H}_2 = 0$.

With the above notation it is found that, in partitioned form,

$$\begin{bmatrix} \mathbf{X}_1'\mathbf{X}_1 + \mathbf{H}_1'\mathbf{H}_1 & \mathbf{X}_1'\mathbf{X}_2 \\ \mathbf{X}_2'\mathbf{X}_1 & \mathbf{X}_2'\mathbf{X}_2 + \mathbf{H}_2'\mathbf{H}_2 \end{bmatrix} \begin{bmatrix} \mathbf{b}_1 \\ \mathbf{b}_2 \end{bmatrix} = \begin{bmatrix} \mathbf{X}_1'\mathbf{y} \\ \mathbf{X}_2'\mathbf{y} \end{bmatrix}.$$

On putting

$$\mathbf{X}_1'\mathbf{X}_1 + \mathbf{H}_1'\mathbf{H}_1 = \mathbf{S}_{11}, \quad \mathbf{X}_1'\mathbf{X}_2 = \mathbf{S}_{12}, \quad \mathbf{X}_2'\mathbf{X}_1 = \mathbf{S}_{21}, \quad \mathbf{X}_2'\mathbf{X}_2 + \mathbf{H}_2'\mathbf{H}_2 = \mathbf{S}_{22},$$

$$\mathbf{V}' = \mathbf{X}_2' - \mathbf{S}_{21}\mathbf{S}_{11}^{-1}\mathbf{X}_1', \quad \mathbf{U} = \mathbf{S}_{22} - \mathbf{S}_{21}\mathbf{S}_{11}^{-1}\mathbf{S}_{12},$$

the analysis of variance of Table 1 is easily inferred from the full rank case. In the table, $\boldsymbol{\beta}_2|\boldsymbol{\beta}_1$ represents the effect of $\boldsymbol{\beta}_2$ adjusted for $\boldsymbol{\beta}_1$. It can be verified that $\mathbf{X}_1'\mathbf{X}_1\mathbf{S}_{11}^{-1}\mathbf{X}_1' = \mathbf{X}_1'$ and $\mathbf{X}_2\mathbf{V}\mathbf{U}^{-1}\mathbf{V}' = \mathbf{V}'$ and these are the only results required to show that the matrices of the three quadratic forms are idempotent and mutually orthogonal.

Table 1. *Analysis of variance*

(Single partitioning of X)

Source	D.F.	Sum of squares
$\boldsymbol{\beta}_1$	$\rho(\mathbf{X}_1'\mathbf{X}_1)$	$\mathbf{y}'\mathbf{X}_1\mathbf{S}_{11}^{-1}\mathbf{X}_1'\mathbf{y}$
$\boldsymbol{\beta}_2|\boldsymbol{\beta}_1$	$t - \rho(\mathbf{X}_1'\mathbf{X}_1)$	$\mathbf{y}'\mathbf{V}\mathbf{U}^{-1}\mathbf{V}'\mathbf{y}$
Error	$n-t$	$\mathbf{y}'(\mathbf{I} - \mathbf{X}_1\mathbf{S}_{11}^{-1}\mathbf{X}_1' - \mathbf{V}\mathbf{U}^{-1}\mathbf{V}')\mathbf{y}$

Most of the ideas discussed above are given, for instance, by Scheffé (1959, Chapter 1). These results have been kept separate from the general treatment since the whole analysis

can be carried out within the framework of standard matrix algebra. All inverses employed are computable by the usual procedures and, where machine programmes are employed, this may be advantageous. Of course, $S^{-1} = (X'X + H'H)^{-1}$ is a generalized inverse of $X'X$ and it is a particularly manageable one computationally.

The general theory for solving (1) when $\rho(X) < p$ is given by Rao (1965, p. 26 and Chapter 4). In fact, if b is any solution,

$$b = (X'X)^{-}X'y + \{I - (X'X)^{-}(X'X)\}z,$$

where z is arbitrary and $(X'X)^{-}$ is a generalized inverse of $X'X$. It turns out that $X(X'X)^{-}X'X = X$, $X(X'X)^{-}X'$ is unique and $\rho\{X(X'X)^{-}X'\} = \rho(X'X)$ and hence $y'X(X'X)^{-}X'y$ and $y'\{I - X(X'X)^{-}X'\}y$ are uniquely determined and are independent of which generalized inverse $(X'X)^{-}$ is used.

If X is partitioned as $X = [X_1, X_2, X_3]$ and $\beta' = [\beta_1', \beta_2', \beta_3']$, then the analysis of variance takes the form given in Table 2 where

$$D_1 = I - X_1(X_1'X_1)^{-}X_1', \quad D_{12} = D_1 - D_1X_2(X_2'D_1X_2)^{-}X_2'D_1$$

and $y'X(X'X)^{-}X'y$ is the sum of the first three quadratic forms.

Table 2. *Analysis of variance*

(Double partitioning of X)

Source	D.F.	Sum of squares
β_1	$\rho(X_1'X_1)$	$y'X_1(X_1'X_1)^{-}X_1'y$
$\beta_2 \mid \beta_1$	$\rho(X_2'D_1X_2)$	$y'D_1X_2(X_2'D_1X_2)^{-}X_2'D_1y$
$\beta_3 \mid \beta_1, \beta_2$	$\rho(X_3'D_{12}X_3)$	$y'D_{12}X_3(X_3'D_{12}X_3)^{-}X_3'D_{12}y$
Error	$n - \rho(X'X)$	$y'\{I - X(X'X)^{-}X'\}y$

Methods of computing generalized inverses are discussed in Chapter 4 of Rao's book, while the necessary formula for the generalized inverse of a partitioned matrix which is used to construct Table 2 is given by Rohde (1965).

2·2. *First and second moments of bilinear forms of normal variates*

Since expectations and covariances of bilinear forms involving normal variables are required in the next sections, we derive below the required general expressions. In fact, two procedures are indicated for obtaining the covariance formula.

Let Y_i, Y_j, Y_k and Y_l be jointly normally distributed with means and covariances

$$\mu = \begin{bmatrix} \mu_i \\ \mu_j \\ \mu_k \\ \mu_l \end{bmatrix}, \quad V = \begin{bmatrix} V_{ii} & V_{ij} & V_{ik} & V_{il} \\ V_{ji} & V_{jj} & V_{jk} & V_{jl} \\ V_{ki} & V_{kj} & V_{kk} & V_{kl} \\ V_{li} & V_{lj} & V_{lk} & V_{ll} \end{bmatrix},$$

then we have the following results:

$$E(Y_i'FY_j) = \mu_i'F\mu_j + \mathrm{tr}\,(FV_{ji}), \tag{2·1}$$

$$\begin{aligned}\mathrm{cov}\,(Y_i'FY_j, Y_k'GY_l) = {} & \mu_i'FV_{jl}G'\mu_k + \mu_i'FV_{jk}G\mu_l + \mu_j'F'V_{il}G'\mu_k \\ & + \mu_j'F'V_{ik}G\mu_l + \mathrm{tr}\,(FV_{jl}G'V_{ki}) + \mathrm{tr}\,(FV_{jk}GV_{li}).\end{aligned} \tag{2·2}$$

A suitable expression for $\text{var}(Y_i'FY_j)$ is obtained from (2·2) by setting $F = G$, $i = k$ and $j = l$.

To establish (2·1) no assumption of normality is required. Thus

$$E(Y_i'FY_j) = E\{\text{tr}(FY_jY_i')\} = \text{tr}\{F(V_{ji}+\mu_j\mu_i')\},$$

which gives (2·1). In order to obtain (2·2) we can make use of the fact that, under normal theory,

$$\text{cov}(Y'B_1Y, Y'B_2Y) = 4\mu'B_1VB_2\mu + 2\,\text{tr}(B_1VB_2V).$$

Now, with the same partitioning as V let

$$B_1 = \frac{1}{2}\begin{bmatrix} 0 & F & 0 & 0 \\ F' & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}, \quad B_2 = \frac{1}{2}\begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & G \\ 0 & 0 & G' & 0 \end{bmatrix};$$

then $Y'B_1Y = Y_i'FY_j$ and $Y'B_2Y = Y_k'GY_l$. Matrix multiplication now gives

$$B_1VB_2 = \frac{1}{4}\begin{bmatrix} 0 & 0 & FV_{jl}G' & FV_{jk}G \\ 0 & 0 & F'V_{il}G' & F'V_{ik}G \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix},$$

$$4\mu'B_1VB_2\mu = \mu_i'FV_{jl}G'\mu_k + \mu_j'F'V_{il}G'\mu_l + \mu_i'FV_{jk}G'\mu_l + \mu_j'F'V_{ik}G\mu_l,$$

$$\begin{aligned} 2\,\text{tr}(B_1VB_2V) &= \tfrac{1}{2}\{\text{tr}(FV_{jl}G'V_{ki}) + \text{tr}(FV_{jk}GV_{li}) + \text{tr}(F'V_{il}G'V_{kj}) + \text{tr}(F'V_{ik}GV_{lj})\} \\ &= \text{tr}(FV_{jl}G'V_{ki}) + \text{tr}(FV_{jk}GV_{li}). \end{aligned}$$

An alternative way of establishing (2·2) is to start from first principles. Define $M = t_1B_1 + t_2B_2$, then the joint moment generating function of $Y_i'FY_j$ and $Y_k'GY_l$ is given by

$$m(t_1, t_2) = E\{\exp(\tfrac{1}{2}Y'MY)\}$$

and by standard techniques we find the cumulant generating function can be written as

$$\log m(t_1, t_2) = \phi(t_1, t_2) = \tfrac{1}{2}\mu' \sum_{n=1}^{\infty} (MV)^n V^{-1}\mu + \frac{1}{2}\sum_{n=1}^{\infty} \text{tr}(MV)^n/n,$$

from which the coefficient of t_1t_2 can be found with relative ease.

3. Results

3·1. *The 'one-way' model*

We consider initially the simplest situation of the one-way multivariate analysis of variance. This term is used to describe the situation where the outcome of a particular experiment is vector valued and the experimental structure is such that analysis is conducted on a between group and within group basis. Classical examples of this situation exist in statistical genetics where, for instance, from progeny studies several characters of a particular breed of animal are analysed jointly for between parent and within parent effects; see § 1.

The appropriate mixed model for this case is

$$Y_i = X_1\beta_i + X_2\gamma_i + \epsilon_i \quad (i = 1, \ldots, k), \tag{3·1}$$

where $X = [X_1, X_2]$ is $(n \times p)$, X_r is $(n \times p_r)$ $(r = 1, 2)$, $p_1 + p_2 = p$, Y_i and ϵ_i are $(n \times 1)$, β_i is $(p_1 \times 1)$ and γ_i is $(p_2 \times 1)$. The coefficients β_i are fixed effects and γ_i is a random vector with zero expectation and covariance matrix $c_{ii} I_{p_2}$. The parameter c_{ii} is a component of variance in the standard terminology of the analysis of variance. As usual,

$$E(\epsilon_i) = 0, \quad \text{var}(\epsilon_i) = e_{ii} I_n.$$

Analogously, we assume that $\text{cov}(\gamma_i, \gamma_j) = c_{ij} I_{p_2}$ and $\text{cov}(\epsilon_i, \epsilon_j) = e_{ij} I_n$ and define $Y' = [Y_1', Y_2', \ldots, Y_k']$ with similar definitions for β', γ' and ϵ'. Then (3·1) can be written conveniently as

$$Y = [I_k \otimes X_1 | I_k \otimes X_2 | I_k \otimes I_n] \begin{bmatrix} \beta \\ \gamma \\ \epsilon \end{bmatrix}. \tag{3·2}$$

From (3·2) $\quad E(Y) = [I_k \otimes X_1]\beta, \quad \text{var}(Y) = [C \otimes X_2 X_2'] + [E \otimes I_n],$

where $C = \{c_{ij}\}$ and $E = \{c_{ij}\}$ are $(k \times k)$ and $E(\epsilon, \gamma') = 0$ by assumption.

Use will now be made of the results and notation of § 2·1. Specifically, S_{11}, V and U are defined as in that section and we stress that, therefore, all the required matrix inverses exist in the usual sense. Thus, setting

$$F_1 = X_1 S_{11}^{-1} X_1', \quad F_2 = VU^{-1}V', \quad F_3 = I - F_1 - F_2$$

and introducing the notation

$$Q_{ij}(s) = Y_i' F_s Y_j \quad \text{and} \quad Z' = [Y_1, Y_2, \ldots, Y_k],$$

we obtain the analysis of covariance of Table 3.

Table 3. *Analysis of covariance*

(One-way classification)

Source	Matrix	Expectation
β	$ZF_1Z' = Q(1)$	—
$\gamma \mid \beta$	$ZF_2Z' = Q(2)$	$\rho(V)E + \text{tr}(X_2'VU^{-1}V'X_2)\,C$
Error	$ZF_3Z' = Q(3)$	$[n - \rho(X'X)]E$

It is found that under standard assumptions of normality the general expressions for the covariances are

$$\begin{aligned} \text{cov}\{Q_{ij}(2), Q_{kl}(2)\} &= \text{tr}(X_2'VU^{-1}V'X_2)^2 (c_{jl}c_{ki} + c_{jk}c_{il}) \\ &\quad + \text{tr}(X_2'VU^{-1}V'X_2)(c_{jl}e_{ki} + c_{ki}e_{jl} + c_{il}e_{jk} + c_{jk}e_{il}) \\ &\quad + \rho(V)(e_{jl}e_{ki} + e_{jk}e_{il}), \end{aligned} \tag{3·3}$$

$$\text{cov}(Q_{ij}(2), Q_{kl}(3)) = 0,$$

$$\text{cov}(Q_{ij}(3), Q_{kl}(3)) = \{n - \rho(X'X)\}(e_{jl}e_{ki} + e_{jk}e_{il}).$$

The standard estimators are given by

$$\begin{aligned} \hat{E} &= \{n - \rho(X'X)\}^{-1} Q(3), \\ \hat{C} &= \{\text{tr}(X_2'VU^{-1}V'X_2)\}^{-1}[Q(2) - \rho(V)\{n - \rho(X'X)\}^{-1}Q(3)]. \end{aligned} \tag{3·4}$$

Thus, from (3·4) we deduce that

$$\begin{aligned}
\operatorname{cov}(\hat{e}_{ij}, \hat{e}_{kl}) &= \{n-\rho(\mathrm{X}'\mathrm{X})\}^{-1}(e_{jl}e_{ki}+e_{jk}e_{il}),\\
\operatorname{cov}(\hat{c}_{ij}, \hat{e}_{kl}) &= -\{\operatorname{tr}(\mathrm{X}_2'\mathrm{V}\mathrm{U}^{-1}\mathrm{V}'\mathrm{X}_2)\}^{-1}\{n-\rho(\mathrm{X}'\mathrm{X})\}^{-1}\rho(\mathrm{V})(e_{jl}e_{ki}+e_{jk}e_{il}),\\
\operatorname{cov}(\hat{c}_{ij}, \hat{c}_{kl}) &= \{\operatorname{tr}(\mathrm{X}_2'\mathrm{V}\mathrm{U}^{-1}\mathrm{V}'\mathrm{X}_2)\}^{-2}(\operatorname{tr}(\mathrm{X}_2'\mathrm{V}\mathrm{U}^{-1}\mathrm{V}'\mathrm{X}_2)^2(c_{jl}c_{ki}+c_{jk}c_{il})\\
&\quad+\operatorname{tr}(\mathrm{X}_2'\mathrm{V}\mathrm{U}^{-1}\mathrm{V}'\mathrm{X}_2)(c_{jl}e_{ki}+c_{ki}e_{jl}+c_{il}e_{jk}+c_{jk}e_{il})\\
&\quad+\rho(\mathrm{V})[1+\rho(\mathrm{V})\{n-\rho(\mathrm{X}'\mathrm{X})\}^{-1}](e_{jl}e_{ki}+e_{jk}e_{il})).
\end{aligned} \tag{3·5}$$

In order to obtain expressions in terms of generalized inverses, defining D_1 and D_{12} as in § 2, let $\mathrm{F}_1 = \mathrm{X}_1(\mathrm{X}_1'\mathrm{X}_1)^-\mathrm{X}_1'$, $\mathrm{F}_2 = \mathrm{D}_1\mathrm{X}_2(\mathrm{X}_2'\mathrm{D}_1\mathrm{X}_2)^-\mathrm{X}_2'\mathrm{D}_1$ and $\mathrm{F}_3 = \mathrm{D}_{12}$. With this notation change $\rho(\mathrm{V})$ to $\rho(\mathrm{X}_2'\mathrm{D}_1\mathrm{X}_2)$, $\operatorname{tr}(\mathrm{X}_2'\mathrm{V}\mathrm{U}^{-1}\mathrm{V}'\mathrm{X}_2)$ to $\operatorname{tr}(\mathrm{X}_2'\mathrm{D}_1\mathrm{X}_2)$ and $\operatorname{tr}(\mathrm{X}_2'\mathrm{V}\mathrm{U}^{-1}\mathrm{V}'\mathrm{X}_2)^2$ to $\operatorname{tr}(\mathrm{X}_2'\mathrm{D}_1\mathrm{X}_2)^2$, and the above formulae apply.

We refer back to the two examples discussed in the introduction for application of the above formulae. First, we consider the estimated index $\hat{\beta} = \hat{\mathrm{G}}^{-1}\alpha$ and proceed in the usual way to find the large sample covariance matrix for $\hat{\beta}$. Thus, taking matrix differentials of both sides of the equation $\mathrm{G}\beta = \alpha$ we have that

$$\beta + \mathrm{G}(d\beta) = 0$$

and approximately, $\mathrm{G}\operatorname{var}$ $E\{(d\mathrm{G})\beta\beta'(d\mathrm{G})\}$.

Let $\beta\beta' = \gamma$. Then the (j, k)th elen $\operatorname{var}(\hat{\beta})\,\mathrm{G}$ is

$\hat{\mathrm{G}}_{ij}, \hat{\mathrm{G}}_{kl})$.

We notice that since $\hat{\mathrm{G}} = m\hat{\mathrm{C}}$,

$\operatorname{cov}($ $\operatorname{cov}(\hat{c}_{ij}, \hat{c}_{kl})$

and we use (3·5) to obtain

$\operatorname{var}\{\hat{\beta}\} = \{\operatorname{tr}(\mathrm{X}_2'\mathrm{V}\mathrm{U}^-$ $\mathrm{X}_2'\mathrm{V}\mathrm{U}^{-1}\mathrm{V}'\mathrm{X}_2)^2\{\mathrm{G}\gamma\mathrm{G}+\mathrm{G}\operatorname{tr}(\mathrm{G}\gamma)\}$

$+m\operatorname{tr}(\mathrm{X}$ $+\mathrm{E}\gamma\mathrm{G}+\mathrm{E}\operatorname{tr}(\mathrm{G}\gamma)+\mathrm{G}\operatorname{tr}(\mathrm{E}\gamma)\}$

$+m^2\rho(\mathrm{V})$ $\}^{-1}]\{\mathrm{E}\gamma\mathrm{E}+\mathrm{E}\operatorname{tr}(\mathrm{E}\gamma)\}]\,\mathrm{G}^{-1}$.

Before deriving general results fo lection index we write the results for the balanced case. Tallis (1960) mewhat awkward form and, in the notation of that paper, equatio ues

$\operatorname{var}(\hat{\beta}) = k^{-2}\mathrm{P}^{-1}[\{\mathrm{BCB}+$ $\mathrm{W}\operatorname{tr}(\mathrm{DW})\}/f_w]\,\mathrm{P}^{-1}$.

Notice that W in the above formula paper while C is not related to our C.

In the general case we let $k = \operatorname{tr}($ (3·5) to find that

$\operatorname{var}\{\hat{\beta}\} = \{\operatorname{tr}(\mathrm{X}_2'\mathrm{V}\mathrm{U}^{-1}\mathrm{V}'\mathrm{X}_2)\}^{-2}\mathrm{P}^{-1}[n$ $\operatorname{tr}(\mathrm{CG})\}$

$+m^{-1}\operatorname{tr}(\mathrm{X}_2'\mathrm{V}\mathrm{U}^{-1}\mathrm{V}'\mathrm{X}_2)($ $\operatorname{tr}(\mathrm{CW})\}$

$+\rho(\mathrm{V})\{\mathrm{WCW}+\mathrm{W}\operatorname{tr}(\mathrm{C}$ $\mathrm{V}+\mathrm{W}\operatorname{tr}(\mathrm{DW})\}]\,\mathrm{P}^{-1}$.

3·2. *The 'two-way' model*

As a direct generalization of (1) w

$$\mathbf{Y}_i = \mathrm{X}_1\beta_i + \quad (3·6)$$

where $$E(\mathrm{Y}_i) = \mathrm{X}_1\beta_i, \quad \mathrm{cov}\,(\mathrm{Y}_i, \mathrm{Y}_j) = \mathrm{X}_2\mathrm{X}_2' c_{ij} + \mathrm{X}_3\mathrm{X}_3' a_{ij} + \mathrm{I}_n e_{ij},$$

since by assumption

$$E(\mathrm{X}_2\gamma_i\gamma_j'\mathrm{X}_2') = \mathrm{X}_2\mathrm{X}_2' c_{ij}, \quad E(\mathrm{X}_3\alpha_i\alpha_j'\mathrm{X}_3') = \mathrm{X}_3\mathrm{X}_3' a_{ij}, \quad E\{\epsilon_i\epsilon_j'\} = \mathrm{I}_n e_{ij}.$$

Again, (3·6) can be more conveniently written as

$$\mathrm{Y} = [\mathrm{I}_k \otimes \mathrm{X}_1 | \mathrm{I}_k \otimes \mathrm{X}_2 | \mathrm{I}_k \otimes \mathrm{X}_3 | \mathrm{I}_k \otimes \mathrm{I}_n] \begin{bmatrix} \beta \\ \gamma \\ \alpha \\ \epsilon \end{bmatrix} \tag{3·7}$$

as in (3·2). Clearly,

$$E(\mathrm{Y}) = (\mathrm{I}_k \otimes \mathrm{X}_1)\beta \quad \text{and} \quad \mathrm{var}\,(\mathrm{Y}) = (\mathrm{C} \otimes \mathrm{X}_2\mathrm{X}_2') + (\mathrm{A} \otimes \mathrm{X}_3\mathrm{X}_3') + (\mathrm{E} \otimes \mathrm{I}_n).$$

If we now let

$$\mathrm{F}_1 = \mathrm{X}_1(\mathrm{X}_1'\mathrm{X}_1)^-\mathrm{X}_1', \qquad \mathrm{F}_2 = \mathrm{D}_1\mathrm{X}_2(\mathrm{X}_2'\mathrm{D}_1\mathrm{X}_2)^-\mathrm{X}_2'\mathrm{D}_1,$$

$$\mathrm{F}_3 = \mathrm{D}_{12}\mathrm{X}_3(\mathrm{X}_3'\mathrm{D}_{12}\mathrm{X}_3)^-\mathrm{X}_3'\mathrm{D}_{12}, \qquad \mathrm{F}_4 = \mathrm{I} - \mathrm{X}(\mathrm{X}'\mathrm{X})^-\mathrm{X}',$$

we get the analysis of covariance of Table 4.

Table 4. *Analysis of covariance*

(Two-way classification)

Source	Matrix	Expectation
β	$\mathrm{ZF}_1\mathrm{Z}' = \mathrm{Q}(1)$	—
$\gamma\|\beta$	$\mathrm{ZF}_2\mathrm{Z}' = \mathrm{Q}(2)$	$\rho(\mathrm{X}_2'\mathrm{D}_1\mathrm{X}_2)\mathrm{E} + \mathrm{tr}\{\mathrm{X}_3'\mathrm{D}_1\mathrm{X}_2(\mathrm{X}_2'\mathrm{D}_1\mathrm{X}_2)^-\mathrm{X}_2'\mathrm{D}_1\mathrm{X}_3\}\mathrm{A}$ $+ \mathrm{tr}\,(\mathrm{X}_2'\mathrm{D}_1\mathrm{X}_2)\mathrm{C}$
$\alpha\|\beta\gamma$	$\mathrm{ZF}_3\mathrm{Z}' = \mathrm{Q}(3)$	$\rho(\mathrm{X}_3'\mathrm{D}_{12}\mathrm{X}_3)\mathrm{E} + \mathrm{tr}\,(\mathrm{X}_3'\mathrm{D}_{12}\mathrm{X}_3)\mathrm{A}$
Error	$\mathrm{ZF}_4'\mathrm{Z}' = \mathrm{Q}(4)$	$\{n - \rho(\mathrm{X}'\mathrm{X})\}\mathrm{E}$

The expectations in the above table are calculated elementwise by use of (2·1) of § 2·2. The noncentrality terms are zero because of the relationship $\mathrm{D}_1\mathrm{X}_1 = 0$.

For convenience we let

$$\rho(\mathrm{X}_2'\mathrm{D}_1\mathrm{X}_2) = k_1, \quad \mathrm{tr}\{\mathrm{X}_3'\mathrm{D}_1\mathrm{X}_2(\mathrm{X}_2'\mathrm{D}_1\mathrm{X}_2)^-\mathrm{X}_2'\mathrm{D}_1\mathrm{X}_3\} = k_2, \quad \mathrm{tr}\,(\mathrm{X}_2'\mathrm{D}_1\mathrm{X}_2) = k_3,$$

$$\rho(\mathrm{X}_3'\mathrm{D}_{12}\mathrm{X}_3) = k_4, \quad \mathrm{tr}\,(\mathrm{X}_3'\mathrm{D}_{12}\mathrm{X}_3) = k_5, \quad n - \rho(\mathrm{X}'\mathrm{X}) = k_6.$$

Then we can calculate the standard unbiased estimators as follows:

$$\left.\begin{aligned} \hat{\mathrm{E}} &= k_6^{-1}\mathrm{Q}(4), \quad \hat{\mathrm{A}} = k_5^{-1}\{\mathrm{Q}(3) - k_4 k_6^{-1}\mathrm{Q}(4)\}, \\ \mathrm{C} &= k_3^{-1}\{\mathrm{Q}(2) - k_2 k_5^{-1}\mathrm{Q}(3) + (k_5 k_6)^{-1}(k_2 k_4 - k_5 k_1)\mathrm{Q}(4)\}. \end{aligned}\right\} \tag{3·8}$$

The method of estimation outlined above is essentially equivalent to Henderson's Method III (Henderson, 1953).

We seek general expressions for $\mathrm{cov}\,\{Q_{ij}(s), Q_{kl}(t)\}$. Again, since $\mathrm{D}_1\mathrm{X}_1 = 0$, from § 2 we have

$$\mathrm{cov}\,\{Q_{ij}(s), Q_{kl}(t)\} = \mathrm{tr}\,(\mathrm{F}_s\mathrm{V}_{jl}\mathrm{F}_t\mathrm{V}_{ik}) + \mathrm{tr}\,(\mathrm{F}_s\mathrm{V}_{jk}\mathrm{F}_t\mathrm{V}_{il}). \tag{3·9}$$

Using (3·9), and performing tedious algebra, it can be shown that

$$\begin{aligned}
\operatorname{cov}\{Q_{ij}(2), Q_{kl}(3)\} &= \operatorname{tr}\{(X_3' F_2 X_3)(X_3' D_{12} X_3)\}(a_{jl}a_{ik}+a_{il}a_{jk}),\\
\operatorname{cov}\{Q_{ij}(2), Q_{kl}(4)\} &= \operatorname{cov}(Q_{ij}(3), Q_{kl}(4)\} = 0,\\
\operatorname{cov}\{Q_{ij}(2), Q_{kl}(2)\} &= \operatorname{tr}(X_2' D_1 X_2)^2 (c_{jl}c_{ik}+c_{jk}c_{il})\\
&\quad + \operatorname{tr}\{(X_2' D_1 X_3)(X_3' D_1 X_2)\}(a_{ik}c_{jl}+a_{jl}c_{ik}+a_{il}c_{jk}+a_{jk}c_{il})\\
&\quad + k_3(c_{jl}e_{ik}+c_{ik}e_{jl}+c_{jk}e_{il}+c_{il}e_{jk})\\
&\quad + \operatorname{tr}(X_3' F_2 X_3)^2 (a_{jl}a_{ik}+a_{jk}a_{il})\\
&\quad + k_2(a_{jl}e_{ik}+a_{ik}e_{jl}+a_{jk}e_{il}+a_{il}e_{jk}) + k_1(e_{jl}e_{ik}+e_{jk}e_{il}), \qquad (3\cdot10)\\
\operatorname{cov}\{Q_{ij}(3), Q_{kl}(3)\} &= \operatorname{tr}(X_3' D_{12} X_3)^2 (a_{ik}a_{jl}+a_{il}a_{jk})\\
&\quad + k_5(a_{il}e_{jk}+a_{jk}e_{il}+a_{jl}e_{ik}+a_{ik}e_{jl}) + k_4(e_{ik}e_{jl}+e_{il}e_{jk}),\\
\operatorname{cov}\{Q_{ij}(4), Q_{kl}(4)\} &= k_6(e_{ik}e_{jl}+e_{il}e_{jk}).
\end{aligned}$$

Suitable terms for variances may be obtained from (3·10) by setting $i = k$ and $j = l$. Moreover, since estimators (3·8) are all linear combinations of the $Q(t)$, all required variances and covariances between the elements of $\hat{E}$, $\hat{A}$ and $\hat{C}$ can be calculated.

It is almost obvious that the above formulae quickly generalize although the complexity of the algebra is increased. Thus one can easily develop formulae for estimation of components in the model $\quad Y_i = X_1\beta_i + X_2\gamma_i + X_3\alpha_i + X_4(\gamma\alpha)_i + \epsilon_i,$

where $(\gamma\alpha)_i$ represents an interaction effect. The only change will be the addition of a new matrix,

$$D_{123} = D_{12} - D_{12}X_3(X_3' D_{12} X_3)^- X_3' D_{12},$$

to the computation of the sum of squares for $(\gamma\alpha)|\beta, \gamma, \alpha$ and the resulting modifications in the remaining formula. No new results are needed for computation of expectations, variances or covariances.

References

Blischke, W. R. (1966). Variances of estimates of variance components in a three-way classification. *Biometrics* **22**, 553–65.

Graybill, F. A. (1961). *An Introduction to Linear Statistical Models*, vol. 1. New York: McGraw-Hill.

Hartley, H. O. & Rao, J. N. K. (1967). Maximum-likelihood estimation for the mixed analysis of variance model. *Biometrika* **54**, 93–108.

Henderson, C. R. (1953). Estimation of variance and covariance components. *Biometrics* **9**, 226–52.

Mahamunulu, D. M. (1963). Sampling variances of the estimates of variance components in the unbalanced 3-way nested classification. *Ann. Math. Statist.* **34**, 521–7.

Rao, C. R. (1965). *Linear Statistical Inference and its Applications*. New York: Wiley.

Rohde, C. A. (1965). Generalized inverses of partitioned matrices. *SIAM Journal* **13**, 1033–5.

Scheffé, H. (1959). *The Analysis of Variance*. New York: Wiley.

Searle, S. R. (1956). Matrix methods in components of variance and covariance analysis. *Ann. Math. Statist.* **27**, 737–48.

Searle, S. R. (1958). Sampling variances of estimates of components of variance. *Ann. Math. Statist.* **29**, 167–78.

Searle, S. R. (1961). Variance components in the unbalanced 2-way nested classification. *Ann. Math. Statist.* **32**, 1161–6.

Smith, H. Fairfield (1936). A discriminant function for plant selection. *Ann. Eugen.* **7**, 240–50.

Tallis, G. M. (1960). The sampling errors of estimated genetic regression coefficients and the errors of predicted genetic gains. *Aust. J. Statist.* **2**, 66–77.

Tallis, G. M. (1968). Selection for an optimum growth curve. *Biometrics* **24**, 169–77.

[*Received November* 1967. *Revised June* 1969]

(iv)

SPECIAL MODELS FOR DISCRETE CHARACTER SELECTION; ESTIMATION UNDER GROUPING AND TRUNCATION

A (a) (iv) [1]

The Use of a Generalized Multinomial Distribution in the Estimation of Correlation in Discrete Data

BY

G. M. TALLIS

Reprinted from
THE JOURNAL OF THE ROYAL STATISTICAL SOCIETY,
SERIES B (METHODOLOGICAL)
Volume 24, No. 2, 1962
(pp. 530 – 534)

PRINTED FOR PRIVATE CIRCULATION
1962

The Use of a Generalized Multinomial Distribution in the Estimation of Correlation in Discrete Data

By G. M. Tallis

Division of Animal Genetics, C.S.I.R.O., Glebe, N.S.W.

[Received August 1961. Revised February 1962]

Summary

This paper presents a joint distribution for n identically distributed multinomial variates, X_q. The distribution is constructed so that $C(X_q, X_r) = \rho V(X)$ for all q and r, $q \neq r$, and applications of this model to the estimation of correlation in discrete data are discussed.

1. Introduction

In several branches of Biology, the analysis of discrete type data presents unpleasant statistical problems. This is particularly the case when it is desired to estimate intra-class correlation for discontinuous variates and to devise satisfactory tests of hypotheses. Some approximate methods of attacking this problem have been suggested; see, for example, Robertson and Lerner (1949). However, their efficiency remains questionable.

It is the purpose of this paper to introduce a generalized multinomial distribution incorporating an additional parameter, ρ. It will be shown that ρ is the correlation coefficient for two variables which are marginally distributed as multinomial variates with common means. Methods of estimating the relevant parameters are presented and some applications of the model to the estimation of intra-class correlation are given in section 4.

2. The General Distribution

Let the random variable X take the values $0, 1, 2, \ldots, k$, with probabilities $\Pr(X = 0) = p_0$, $\Pr(X = 1) = p_1, \ldots, \Pr(X = k) = p_k$ subject to the restriction $\Sigma p_i = 1$ and consider the n identically distributed variates $X_j (j = 1, 2, \ldots, n)$. A probability generating function, p.g.f., for the joint probabilities

$$\Pr\{X_1 = a, X_2 = b, X_3 = c, \ldots\} = \alpha_{abc\ldots} \quad (a, b, c = 0, 1, 2, \ldots, k),$$

may now be written, for $0 \leqslant \rho \leqslant 1$, as

$$G_n(\mathbf{s}) = \rho\left\{\sum_{i=0}^{k} p_i\left(\prod_{j=1}^{n} s_j\right)^i\right\} + (1-\rho)\prod_{j=1}^{n} P(s_j), \tag{1}$$

where $P(s_j) = \Sigma p_i s_j^i$ and $\mathbf{s} = \mathrm{col}\{s_1, s_2, \ldots, s_n\}$. The parameter ρ appearing in (1) is the correlation coefficient between X_q and X_r $(q \neq r; q, r, = 1, 2, \ldots, n)$.

The marginal distribution of X_q is obtained immediately from (1) by setting $s_r = 1$ $(r \neq q)$ and we have $G(s_q) = P(s_q)$, which is the generating function of a multinomial variate with parameters p_i.

In order to obtain the joint moment generating function, m.g.f., $M_n(\mathbf{t})$, $\mathbf{t} = \text{col}\{t_1, t_2, \ldots, t_n\}$, let $s_j = e^{t_j}$ in (1), then

$$M_n(\mathbf{t}) = \rho\left(\sum_{i=0}^{k} p_i e^{i\Sigma t_j}\right) + (1-\rho)\prod_{j=1}^{n} P(e^{t_j}). \tag{2}$$

Suitable differentiation of (2) gives

$$C(X_r, X_q) = E(X_r, X_q) - E^2(X) = \rho V(X),$$

where $E(X) = \Sigma i p_i$, $V(X) = \Sigma i^2 p_i - E^2(X)$, and this verifies the statement that the correlation coefficient for X_q and X_r is ρ. The joint probability $\alpha_{abc\ldots}$ may be obtained from (1), although, with a little experience, the required probabilities can be written down by inspection.

Case 1. Let $a = b = c = \ldots = i$, say, then $\alpha_{iii\ldots} = p_i^n(1-\rho) + \rho p_i$.

Case 2. Let at least two members of the array $a, b, c, \ldots$ take different values. Then $\alpha_{abc\ldots} = (1-\rho)p_a p_b p_c \ldots$.

Consider now the new variate $R = X_1 + \ldots + X_n$. The p.g.f. for R is given by

$$G_n(s) = \sum_{i=0}^{nk} \alpha_i s^i = \rho\left(\sum_{i=0}^{k} p_i s^{ni}\right) + (1-\rho)\{P(s)\}^n \tag{3}$$

and is obtained from (1) by setting $s_r = s$ $(r = 1, 2, \ldots, n)$. The m.g.f. for R is

$$M_R(t) = G_n(e^t) = \rho\left(\sum_{i=0}^{k} p_i e^{nit}\right) + (1-\rho)\{P(e^t)\}^n \tag{4}$$

and straightforward calculations show

$$E(R) = nE(X), \quad V(R) = nV(X)\{1 + (n-1)\rho\},$$

as may well have been anticipated.

Because of the wide interest in the binomial distribution, the above results will be specialized for $k = 1$. From (1) we obtain

$$G_n(\mathbf{s}) = \rho\left(p_0 + p_1 \prod_{j=1}^{n} s_j\right) + (1-\rho)\prod_{j=1}^{n} (p_0 + p_1 s_j).$$

Thus the two special cases give the following results.

Case 1. Here

$$\alpha_{000\ldots} = p_0^n(1-\rho) + \rho p_0, \quad \alpha_{111\ldots} = p_1^n(1-\rho) + \rho p_1.$$

Case 2. Suppose in the array $a, b, c, \ldots$, zero appears n_0 times and 1 appears n_1 times $(n_0, n_1 \neq n)$, then $\alpha_{abc\ldots} = p_0^{n_0} p_1^{n_1}(1-\rho)$.

Formula (3) reduces to

$$G_n(s) = \rho(p_0 + p_1 s^n) + (1-\rho)(p_0 + p_1 s)^n,$$

which gives probabilities

$$\alpha_0 = p_0^n(1-\rho) + \rho p_0, \quad \alpha_n = p_1^n(1-\rho) + \rho p_1,$$

$$\alpha_j = \frac{(1-\rho)n!}{j!(n-j)!} p_1^j p_0^{n-j} \quad (j \neq 0, n).$$

Finally, the appropriate formulae for the mean and variance are

$$E(R) = np_1, \quad V(R) = np_1 p_0\{1 + (n-1)\rho\}.$$

Now that the basic distribution has been presented, it is interesting to notice certain features. It is clear that (1) specifies the joint distribution of n variates, each margin being multinomial with parameters p_i $(i = 0, 1, ..., k)$. When $\rho = 0$, the distribution becomes the product of n identical multinomial distributions, and this shows that the variates are then independently distributed. On the other hand, when $\rho = 1$, the distribution degenerates in such a way that $\alpha_{abc...} = 0$ unless $a = b = c = ... = i$, say, when $\alpha_{iii...} = p_i$. Thus it seems that (1) is the weighted mean of a distribution with perfect correlation and one with complete independence, the weights being ρ and $1-\rho$ respectively.

The variate R is the sum of n, generally non-independent, variates X_j $(j = 1, 2, ..., n)$. In some work, R may be the only information available and its distribution is important, at least for estimation purposes.

3. Estimation

3.1. *General*

A central problem of practical interest is the estimation of the parameters p_i $(i = 0, 1, ..., k)$ and ρ. Suppose the n variates X_j are jointly observed on N independent occasions, then maximum likelihood estimation may be carried out in several ways.

3.2. *Estimation from the Joint Distribution*

Full maximum likelihood estimates of the unknown parameters may be obtained from the joint probabilities for the variates X_j. It was shown in the previous section how the probability of any outcome of the form $X_1 = a$, $X_2 = b$, $X_3 = c$, ..., can be derived and, therefore, it is possible to write the probability of the N outcomes as

$$L \propto \prod \alpha^r_{abc...}, \tag{5}$$

where $\alpha^r_{abc...}$ is the probability of the rth outcome with array $a^{(r)}, b^{(r)}, c^{(r)}, ...$ Now, if in the latter array 0 appears $n_0^{(r)}$ times, 1 appears $n_1^{(r)}$ and, in general, i appears $n_i^{(r)}$ times, then

$$\alpha^{(r)}_{abc...} = \prod_{i=0}^{k} p_i^{n_i^{(r)}} (1-\rho)$$

unless $a^{(r)} = b^{(r)} = c^{(r)} = ... = i$, say, when

$$\alpha^{(r)}_{iii...} = p_i^n(1-\rho) + \rho p_i.$$

With the aid of the above expressions (5) can readily be simplified and maximized with respect to the $k+2$ unknown parameters by standard techniques; see Aitchison and Silvey (1960).

3.3 *Estimation from the Distribution of R*

Since the expressions for $\alpha_j = \Pr(R = j)$ $(j = 0, 1, ..., nk)$ may be obtained from (3), it is necessary to count the number of times $R = j$ in the N observations, N_j, and write

$$L \propto \prod_{j=0}^{nk} \alpha_j^{N_j}. \tag{6}$$

However, it is readily verified that, in general, estimates calculated from (6) are not fully efficient. Only for the special case of $k = 1$ do methods (3.2) and (3.3) give identical estimates.

3.4. *Data with Varying n*

It may frequently occur that n does not remain constant from trial to trial. Fortunately, this does not cause a great deal of trouble from the point of view of estimation since the probability of any outcome can be written down for any n in terms of the parameters p_i $(i = 0, 1, ..., k)$ and ρ. Thus let P_q be the probability associated with the qth trial, then

$$L \propto \prod_{q=1}^{N} P_q. \tag{7}$$

The likelihood function (7) can be simplified and maximized in an analogous way to (5).

4. DISCUSSION

The valid application of the generalized multinomial distribution presented above is restricted to those cases where $E(X)$ can be assumed to remain constant throughout the experiment. Unfortunately, numerous situations arise where this assumption may not be made and in these instances the analysis breaks down. It is now proposed to give some examples where the model can be applied with reasonable confidence.

In statistical genetics it is often of interest to estimate the heritability, h^2, of birth records of livestock. The parameter h^2 may be defined as the ratio of the additive genetic variance to the total phenotypic variance of a particular character (Lush, 1945). In sheep, for instance, birth records generally take the values 0, 1 and 2, the occurrence of triplets, quadruplets, etc. being exceedingly rare in most flocks. For convenience, the ability to produce lambs will be referred to as "fertility" in the subsequent discussion.

An estimate of the heritability of fertility may be obtained in the following way. If we take the first lambing records of n daughters of each of N sires (i.e. ewes within sire groups are half-sibs), then we may regard fertility at first record as a multinomial variate, X. If ρ is the correlation between records within sire groups then, applying the usual genetical argument, $h^2 = 4\rho$. If in this case we are prepared to assume that $E(X)$ is the same for all ewes, we may apply the methods of the previous sections to estimate ρ.

Another example of a similar nature is the estimation of repeatability of fertility records from year to year. In this example, N ewes may be observed for a period of n years and, here, R corresponds to the number of lambs born to each ewe during the period. If it can be assumed that lambing performance is relatively unaffected by the age of the ewe and year to year environment, then the multinomial model suggested above may be used to estimate repeatability, ρ. However, in contrast to the previous example, close attention must be given to the assumption of constant $E(X)$ because of the year-to-year environmental effect on fertility.

Note that, where n is constant from trial to trial,

$$E\left\{\sum_{ij}\frac{(X_{ij}-\bar{X}_{i.})^2}{N(n-1)}\right\} = V(X)(1-\rho),$$

$$E\left\{\sum_{i}\frac{(\bar{X}_{i.}-\bar{X}_{..})^2}{N-1}\right\} = V(X)\{1+(n-1)\rho\},$$

where X_{ij} represents the jth observation of the ith trial,

$$n\bar{X}_{i.} = \sum_{j} X_{ij} \quad \text{and} \quad nN\bar{X}_{..} = \sum_{i,j} X_{ij}.$$

These results will be recognized as analogous to the expectations of within and between mean squares of the one-way analysis of variance, with ρ representing the intraclass correlation coefficient. Thus ρ can be estimated, $\tilde{\rho}$ say, by the usual procedure for a continuous variate satisfying the analysis of variance conditions. Obviously $\tilde{\rho}$ is relatively easily obtained and can serve as a first guess of $\hat{\rho}$, but it suffers from the disadvantages that it is an inefficient statistic and that appropriate tests of significance are lacking. Thus, by use of the distribution for the X_{ij}, these disadvantages are overcome and it was precisely these considerations which stimulated the work of this paper.

Unfortunately, so far no worth-while generalizations have been achieved by considering the joint distribution of n multinomial variates, each distributed with a different set of parameters, p_i.

ACKNOWLEDGEMENTS

The author is grateful for the valuable comments of Dr F. E. Binet during the early stages of this work and also to the referee, who suggested many improvements in the presentation of this paper.

REFERENCES

AITCHISON, J. and SILVEY, S. D. (1960), "Maximum-likelihood estimation procedures and associated tests of significance", *J. R. statist. Soc.* B, 22, 154–171.

LUSH, J. L. (1945), *Animal Breeding Plans*. Ames, Iowa: Iowa State College Press.

ROBERTSON, A. and LERNER, I. M. (1949), "The heritability of all-or-none traits: viability in poultry", *Genetics*, 34, 395–411.

A (a) (iv) [2]

Further Models for Estimating Correlation in Discrete Data

BY

G. M. TALLIS

Reprinted from

THE JOURNAL OF THE ROYAL STATISTICAL SOCIETY

SERIES B (METHODOLOGICAL)

Volume 26, No. 1, 1964

(pp. 82 – 85)

PRINTED FOR PRIVATE CIRCULATION

1964

Further Models for Estimating Correlation in Discrete Data

By G. M. TALLIS

Division of Animal Genetics, C.S.I.R.O., Glebe, N.S.W.

[Received May 1963. Revised August 1963]

SUMMARY

This note considers the distribution of the sum of n identically distributed multinomial variates X_i, the correlation coefficient for X_i and X_j being ρ for all $i, j, i \neq j$. The initial model, with n and $\mathrm{p}' = (p_1, p_2, \ldots, p_k)$ fixed, is generalized by allowing first n, then p and finally both n and p to be random variables.

1. INTRODUCTION

IN a previous paper, Tallis (1962), the author considered the random variable $R_n = X_1 + \ldots + X_n$, where each X_i was a multinomial variate with probability generating function, p.g.f.,

$$p(s) = \sum_{j=0}^{k} p_j s^j,$$

and R_n had p.g.f.

$$f_n(s) = \rho p(s^n) + (1-\rho)\{p(s)\}^n. \tag{1}$$

It was shown that the parameter ρ is the correlation coefficient for X_i and X_j $(i \neq j)$, and, by suitable differentiation of (1), it is readily verified that

$$E(R_n) = nE(X), \quad E(X) = \sum_{j=0}^{k} jp_j,$$

$$V(R_n) = nV(X)\{1+(n-1)\rho\}, \quad V(X) = \sum_{j=0}^{k} j^2 p_j - \{E(X)\}^2.$$

In this note, (1) is to be generalized by first of all allowing n to be a random variable and then, keeping n fixed, allowing $\mathrm{p}' = (p_1, p_2, \ldots, p_k)$ to be a random vector. Finally, both n and p are allowed to be random. Little attention is devoted to the general results for arbitrary k, but specific formulae are presented for the most important case, $k = 1$.

2. THE DISTRIBUTIONS

2.1. *n a Random Variable*

Suppose that n is itself a random variable N, with probabilities $g_n = P\{N = n\}$; it is then required to find the distribution of R_N. Now

$$f_n(s) = \sum_{i=0}^{\infty} f_{ni} s^i$$

is the conditional p.g.f. of R_N for fixed $N = n$, and the unconditional probability h_j that $R_N = j$ is given by

$$h_j = \sum_{n=0}^{\infty} g_n f_{nj}.$$

Since the p.g.f. of R_N is

$$h(s) = \sum_{j=0}^{\infty} h_j s^j,$$

we have, substituting for h_j,

$$h(s) = \sum_{j=0}^{\infty} \sum_{n=0}^{\infty} g_n f_{nj} s^j = \sum_{n=0}^{\infty} g_n f_n(s). \tag{2}$$

In order to obtain some explicit results, N is given a Poisson distribution with parameter λ and (2) becomes

$$\begin{aligned} h(s) &= e^{-\lambda} \sum_{n=0}^{\infty} \{\lambda^n f_n(s)/n!\} \\ &= e^{-\lambda} \left\{ \rho \sum_{i=0}^{k} p_i e^{\lambda s^i} + (1-\rho) e^{\lambda p(s)} \right\}. \end{aligned} \tag{3}$$

The task of identifying the coefficients of s^n for $k \geqslant 2$ is extremely unpleasant for large n. However, for $k = 1$,

$$h(s) = e^{-\lambda}\{\rho(p_0 e^{\lambda} + p_1 e^{\lambda s}) + (1-\rho) e^{\lambda(p_0 + p_1 s)}\}, \tag{4}$$

whence, by inspection,

$$\begin{aligned} h_0 &= \rho p_0 + \rho p_1 e^{-\lambda} + (1-\rho) e^{-\lambda p_1}, \\ h_j &= \rho p_1 \lambda^j e^{-\lambda}/j! + (1-\rho)(p_1 \lambda)^j e^{-\lambda p_1}/j! \quad (j > 0). \end{aligned} \tag{5}$$

From (3) the mean and variance for R_N are found to be

$$\begin{aligned} E(R_N) &= \lambda E(X), \quad \lambda = E(N), \\ V(R_N) &= \lambda^2 \rho V(X) + \lambda E(X^2), \end{aligned}$$

which, for $k = 1$, specialize to

$$\begin{aligned} E(R_N) &= \lambda p_1, \\ V(R_N) &= \lambda p_1 (1 + \rho \lambda p_0). \end{aligned}$$

2.2. p *a Random Vector*

In this case N is held fixed and p is assigned the frequency function $\phi(\mathrm{p})$. By a similar argument to that used in Section 2.1,

$$h_j = \int_0^1 f_{nj}(\mathrm{p})\, \phi(\mathrm{p})\, d\mathrm{p},$$

where

$$\int_0^1 (\cdot)\, d\mathrm{p} \quad \text{stands for} \quad \int_0^1 \cdots \int_0^{1-\sum_{i=1}^{k-1} p_i} \int_0^{1-\sum_{i=1}^{k-1} p_i} (\cdot)\, dp_1 \ldots dp_k.$$

Hence

$$h(s) = \int_0^1 f_n(\mathrm{p}, s)\, \phi(\mathrm{p})\, d\mathrm{p}. \tag{6}$$

Again, in order to obtain some useful results we let $k = 1$ and

$$\phi(p_1) = \{B(q, r)\}^{-1} p_1^{q-1} p_0^{r-1},$$

where

$$B(q,r) = \int_0^1 x^{q-1}(1-x)^{r-1}\,dx.$$

It is now found that

$$h(s) = \rho(\bar{p}_0 + \bar{p}_1 s^n) + (1-\rho)\{B(q,r)\}^{-1} \sum_{j=0}^{n} \binom{n}{j} B(q+j, r+n-j)\, s^j \tag{7}$$

and

$$h_0 = \rho\bar{p}_0 + (1-\rho)\{B(q,r)\}^{-1} B(q, r+n), \tag{8a}$$

$$h_j = (1-\rho)\binom{n}{j}\{B(q,r)\}^{-1} B(q+j, r+n-j) \quad (0<j<n), \tag{8b}$$

$$h_n = \rho\bar{p}_1 + (1-\rho)\{B(q,r)\}^{-1} B(q+n, r), \tag{8c}$$

where $E(p_1) = \bar{p}_1 = q/(q+r)$ and $\bar{p}_0 = 1-\bar{p}_1$. These results reduce to those of Skellam (1948) when $\rho = 0$. Incidentally, (7) establishes the relationship

$$\sum_{j=0}^{n} \binom{n}{j} B(q+j, r+n-j) = B(q,r)$$

and this can be used to show that

$$E(R_n) = n\bar{p}_1,$$

$$V(R_n) = n(n-1)(1-\rho)\,V(p_1) + n\bar{p}_1\,\bar{p}_0\{1+(n-1)\rho\},$$

where $V(p_1) = \bar{p}_1\,\bar{p}_0/(q+r+1)$.

Explicit results for arbitrary k may be obtained, for instance, by letting $\phi(\mathrm{p})$ take the form of a multivariate β-distribution (see Mosimann, 1962). However, the resulting expressions are more cumbersome and they will not be developed here.

2.3. *n and* p *Random Variables*

In this last case, both n and p are allowed to be random variables and we have

$$h(s) = \int_0^1 \sum_{n=0}^{\infty} g_n f_n(\mathrm{p}, s)\,\phi(\mathrm{p})\,d\mathrm{p}. \tag{9}$$

Results are obtained for $k = 1$ by defining g_n as in Section 2.1 and $\phi(p_1)$ as in Section 2.2, and

$$h(s) = e^{-\lambda}\left[\rho(\bar{p}_0\,e^{\lambda} + \bar{p}_1\,e^{\lambda s}) + (1-\rho)\{B(q,r)\}^{-1}\int_0^1 e^{\lambda(p_0+p_1 s)}\,p_1^{q-1}\,p_0^{r-1}\,dp_1\right]. \tag{10}$$

The problem of obtaining expressions for the h_j is now considerably simplified by introducing the factorial moment generating function $b(s) = h(s+1)$.

We have

$$b(s) = \rho(\bar{p}_0 + \bar{p}_1\,e^{\lambda s}) + (1-\rho)\{B(q,r)\}^{-1}\int_0^1 e^{\lambda p_1 s}\,p_1^{q-1}\,p_0^{r-1}\,dp_1 \tag{11}$$

and, identifying the coefficients of s^j,

$$b_0 = 1,$$

$$b_j = [\rho\bar{p}_1\,\lambda^j + (1-\rho)\,\lambda^j\{B(q,r)\}^{-1} B(q+j, r)]/j! \quad (j>0), \tag{12}$$

and the respective h_j may be obtained from the inversion formula

$$h_j = \sum_{i=j}^{\infty} (-1)^{i-j} \binom{i}{j} b_i. \tag{13}$$

Finally, it is found that

$$E(R_n) = \lambda \bar{p}_1 = b_1,$$
$$V(R_n) = \lambda \bar{p}_1 (1 + \rho \lambda \bar{p}_0) + (1 - \rho) \lambda^2 V(p_1) = 2b_2 + b_1 - b_1^2. \tag{14}$$

3. DISCUSSION

The results obtained in Section 2.1 should be useful for the analysis of data showing considerable variation in n. Although a solution of the likelihood equations can be obtained without assigning a distribution to n, the standard properties of maximum-likelihood estimation may no longer apply (Kendall and Stuart, 1961, p. 60). In a recent paper (Brown *et al.*, 1963) some estimates of ρ were calculated from a mixture of distributions, but the process was extremely tedious and long. However, for (5), and for the other distributions reported here, $\Sigma h_j = 1$ and Fisher's scoring methods (Rao, 1952, p. 165) can be used to obtain an iterative solution of the likelihood equations and to estimate the covariance matrix of the estimators.

The model of Section 2.2 attempts to counteract a deficiency of the original model (1) by allowing p_1 to vary from trial to trial. In the context of the discussion of Tallis (1962), for instance, biologists may be unwilling to assume that each ewe of a flock has an equal probability of giving birth to a single lamb in any year. If this is so, then (7) may provide a more satisfactory description of the data for fixed n. In this case the maximization of the likelihood equations can be organized according to the suggestions of Skellam (1948).

The last model allows both n and p_1 to be random variables. Nevertheless, the increased generality is obtained at a price since an application of even the scoring method obviously involves tedious algebra and computation. Although, in this case, h_j is expressed as an infinite sum of terms involving b_i, in practice it appears that the series converges fairly rapidly. However, these problems of estimation and evaluation will not be considered further here.

REFERENCES

BROWN, G., TALLIS, G. M. and YOUNG, S. S. Y. (1963), "Selection for fertility in Australian Merino sheep—Appendix: Maximum-likelihood estimation of heritability for all-or-none traits in half-sib data", *Aust. J. agric. Res.*, 14, 479–482.

KENDALL, M. G. and STUART, A. (1961), *The Advanced Theory of Statistics*, 2. London: Griffin.

MOSIMANN, J. E. (1962), "On the compound multinomial distribution, the multivariate β-distribution, and correlations among proportions", *Biometrika*, 49, 65–82.

RAO, C. R. (1952), *Advanced Statistical Methods in Biometric Research*. New York: Wiley.

SKELLAM, J. G. (1948), "A probability distribution derived from the binomial distribution by regarding the probability of success as variable between the sets of trials", *J. R. statist. Soc.* B, **10**, 257–261.

TALLIS, G. M. (1962), "The use of a generalized multinomial distribution in the estimation of correlation in discrete data", *J. R. statist. Soc.* B, 24, 530–534.

A(a)(iv)[3]

Commonwealth of Australia

COMMONWEALTH SCIENTIFIC AND INDUSTRIAL RESEARCH ORGANIZATION

BIOMETRICS (1962), 18, 342-353

THE MAXIMUM LIKELIHOOD ESTIMATION OF CORRELATION FROM CONTINGENCY TABLES

G. M. TALLIS

C. S. I. R. O. Division of Animal Genetics, McMaster Laboratory, Glebe, Sydney, N.S.W., Australia

I. INTRODUCTION

It is sometimes desirable in practice to estimate the degree of correlation existing in 2×2 or 3×3 contingency tables. In some instances an underlying bivariate normal distribution can be assumed and the problem reduces to estimating ρ from the observed frequencies in the table. In the case of the 2×2 tables, this may be accomplished with the aid of tetrachoric functions, although this technique does not seem to have been extended to $p \times q$ tables in general. It is the purpose of the present paper to show how Maximum Likelihood (M.L.) estimates of ρ, $\hat{\rho}$, may be obtained from such tables when, in fact, the parent distribution is bivariate normal.

A somewhat similar problem has been considered by Mosteller [1946] who investigated the efficiency of estimating ρ from punch card data. Mosteller considered the case when the cards are sorted with respect to the two co-ordinates, x and y, in a particular manner, and derives the M.L. estimate of ρ from an order statistics argument. However, his results are not generally applicable to the problem considered here because of the special sorting model which he employs.

This topic is part of a wider attempt to develop more satisfactory methods of analysing discrete and continuous data arising in some fields of quantitative genetics. For instance, in the study of heritability and repeatability of birth records of domestic animals, the required correlations are usually calculated by the product-moment or intra-class correlation techniques and hence satisfactory tests of hypotheses are lacking. In the particular case of fertility, a reasonable assumption may be that the potential to produce offspring is normally distributed, but, that, necessarily, phenotypic expression is only possible at distinct threshold values of this potential. Thus, for all potentials below a certain threshold no offspring result; for potentials above this critical value but below a second threshold level one offspring is produced

and for potentials greater than the second critical level two offspring are born, and so on.

Under this hypothesis we consider that fertility for two given years is binormally distributed with distinct threshold levels existing on both co-ordinates. In these circumstances it is possible by means of the techniques of this paper to obtain M.L. estimates of the correlation between fertility records in two different years and, in some cases, to test the hypothesis of an underlying binormal distribution.

By a slight change in argument, heritability may be estimated from the records of dam-daughter pairs. In this instance records of dams' and daughters' fertility are assumed to be binormally distributed with threshold levels and, applying the usual genetical argument, the correlation between these records estimates one half heritability. Undoubtedly, further applications can be found but these two examples are mentioned since they, in part, stimulated this investigation.

II. CASE 1. 2×2 TABLES

Let U and V be two standardised variates with a joint bivariate normal density function $\phi(u, v; \rho) = \phi(u, v)$, say. Now define two new variates X and Y in such a way that

$$\Pr(X = x_0) = \Pr(U < a) = \int_{-\infty}^{a} \frac{e^{-u^2/2}}{\sqrt{2\pi}}\, du = \int_{-\infty}^{a} \phi(u)\, du = \Phi(a) = P_0.$$

$$\Pr(X = x_1) = \Pr(U \geq a) = 1 - \Phi(a) = P_1.$$

$$\Pr(Y = y_0) = \Pr(V < b) = \Phi(b) = P_{.0}$$

$$\Pr(Y = y_1) = \Pr(V \geq b) = 1 - \Phi(b) = P_{.1},$$

and the joint distribution of X and Y is specified by

$$\Pr(X = x_0, Y = y_0) = \int_{-\infty}^{a}\int_{-\infty}^{b} \phi(u, v)\, dv\, du = \Phi(a, b) = P_{00}$$

$$\Pr(X = x_0, Y = y_1) = \Phi(a) - \Phi(a, b) = P_{01}$$

$$\Pr(X = x_1, Y = y_0) = \Phi(b) - \Phi(a, b) = P_{10}$$

$$\Pr(X = x_1, Y = y_1) = 1 - \Phi(a) - \Phi(b) + \Phi(a, b) = P_{11}.$$

The quantities x_0, x_1, y_0 and y_1 need not be numerical, but can refer to any type of discrete classification where the events x_0, x_1 and y_0, y_1 are, pairwise, mutually exclusive.

The problem now is to estimate ρ from a random sample of size n

classified with respect to X and Y. Such data may be conveniently summarised in a 2×2 table as shown below.

	x_0	x_1
y_0	n_{00}	n_{10}
y_1	n_{01}	n_{11}

In the table n_{ij} is the observed number with attributes x_i and y_j, and $\sum_{i,j=0}^{1} n_{ij} = n$.

The appropriate likelihood function is given by

$$L = CP_{00}^{n_{00}}P_{10}^{n_{10}}P_{01}^{n_{01}}P_{11}^{n_{11}}, \tag{1}$$

where C is a constant which does not depend on the parameters to be estimated. It is now possible to obtain the M.L. estimates of a, b and ρ in the usual way.

However, in order to complete the necessary differentiation it is first of all convenient to evaluate $\partial\Phi(a, b)/\partial\rho$. Putting $R = \sqrt{1 - \rho^2}$ we have

$$\begin{aligned}\Phi(a, b) &= \int_{-\infty}^{a}\int_{-\infty}^{b} \phi(u, v)\, dv\, du \\ &= \int_{-\infty}^{a} \phi(u)\Phi\left(\frac{b - \rho u}{R}\right) du,\end{aligned}$$

and differentiating with respect to ρ we obtain

$$\begin{aligned}\frac{\partial\Phi(a, b)}{\partial\rho} &= R^{-2}\int_{-\infty}^{a} \phi(u)\phi\left(\frac{b - \rho u}{R}\right)\left(\frac{\rho b - u}{R}\right) du \\ &= -R^{-1}\phi(b)\int_{-\infty}^{(a-\rho b)/R} l\phi(l)\, dl \\ &= \phi(a, b).\end{aligned}$$

Differentiation under the integral sign is obviously justified for $|\rho| < 1$ and for $\rho = \pm 1$, the problem is degenerate. Putting l equal to $\ln L$, the following equations may be now verified,

$$\begin{aligned}\frac{\partial l}{\partial\rho} &= \phi(a, b)\left\{\frac{n_{00}}{P_{00}} - \frac{n_{10}}{P_{10}} - \frac{n_{01}}{P_{01}} + \frac{n_{11}}{P_{11}}\right\} \\ \frac{\partial l}{\partial a} &= \phi(a)\left[\Phi(B)\left\{\frac{n_{00}}{P_{00}} - \frac{n_{10}}{P_{10}}\right\} + [1 - \Phi(B)]\left\{\frac{n_{01}}{P_{01}} - \frac{n_{11}}{P_{11}}\right\}\right] \qquad (2) \\ \frac{\partial l}{\partial b} &= \phi(b)\left[\Phi(A)\left\{\frac{n_{00}}{P_{00}} - \frac{n_{01}}{P_{01}}\right\} + [1 - \Phi(A)]\left\{\frac{n_{10}}{P_{10}} - \frac{n_{11}}{P_{11}}\right\}\right]\end{aligned}$$

where $A = (a - \rho b)/R$ and $B = (b - \rho a)/R$. Requisite elements of the information matrix, I, are given by

$$
\begin{aligned}
-E\left(\frac{\partial^2 l}{\partial \rho^2}\right) &= I_{\rho\rho} = n[\phi(a, b)]^2\{P_{00}^{-1} + P_{10}^{-1} + P_{01}^{-1} + P_{11}^{-1}\} \\
-E\left(\frac{\partial^2 l}{\partial a^2}\right) &= I_{aa} = n[\phi(a)]^2\{[\Phi(B)]^2[P_{00}^{-1} + P_{10}^{-1}] \\
&\quad + [1 - \Phi(B)]^2[P_{01}^{-1} + P_{11}^{-1}]\} \\
-E\left(\frac{\partial^2 l}{\partial b^2}\right) &= I_{bb} = n[\phi(b)]^2\{[\Phi(A)]^2[P_{00}^{-1} + P_{01}^{-1}] \\
&\quad + [1 - \Phi(A)]^2[P_{10}^{-1} + P_{11}^{-1}]\} \\
-E\left(\frac{\partial^2 l}{\partial a\, \partial \rho}\right) &= I_{a\rho} = n\phi(a)\phi(a, b)\{\Phi(B)[P_{00}^{-1} + P_{10}^{-1}] \\
&\quad - [1 - \Phi(B)][P_{01}^{-1} + P_{11}^{-1}]\} \\
-E\left(\frac{\partial^2 l}{\partial b\, \partial \rho}\right) &= I_{b\rho} = n\phi(b)\phi(a, b)\{\Phi(A)[P_{00}^{-1} + P_{01}^{-1}] \\
&\quad - [1 - \Phi(A)][P_{10}^{-1} + P_{11}^{-1}]\} \\
-E\left(\frac{\partial^2 l}{\partial a\, \partial b}\right) &= I_{ab} = n\phi(a)\phi(b)\{\Phi(A)\Phi(B)P_{00}^{-1} \\
&\quad + [1 - \Phi(A)][1 - \Phi(B)]P_{11}^{-1} \\
&\quad - \Phi(B)[1 - \Phi(A)]P_{10}^{-1} - \Phi(A)[1 - \Phi(B)]P_{01}^{-1}\}.
\end{aligned}
\tag{3}
$$

Since I is symmetric, it is completely determined by the above six expressions.

III CASE 2. 3×3 TABLES

The extension of the above techniques to 3×3 contingency tables is immediate. In this instance X and Y take on the additional values x_2 and y_2 and we have

$$
\begin{aligned}
\Pr(X = x_0) &= \Phi(a_1) = P_{0.}, \\
\Pr(X = x_1) &= \Phi(a_2) - \Phi(a_1) = P_{1.}, \\
\Pr(X = x_2) &= 1 - \Phi(a_2) = P_{2.}, \\
\Pr(Y = y_0) &= \Phi(b_1) = P_{.0}, \\
\Pr(Y = y_1) &= \Phi(b_2) - \Phi(b_1) = P_{.1}, \\
\Pr(Y = y_2) &= 1 - \Phi(b_2) = P_{.2},
\end{aligned}
$$

TABLE 1

NECESSARY FORMULAE FOR ESTIMATING ρ FROM 3 × 3 CONTINGENCY TABLES

Observed No., n_{ij}	Expected Proportion, P_{ij}	$\partial P_{ij}/\partial\rho$
n_{00}	$P_{00} = \Phi(a_1, b_1)$	$\phi(a_1, b_1)$
n_{01}	$P_{01} = \Phi(a_1, b_2) - \Phi(a_1, b_1)$	$\phi(a_1, b_2) - \phi(a_1, b_1)$
n_{10}	$P_{10} = \Phi(a_2, b_1) - \Phi(a_1, b_1)$	$\phi(a_2, b_1) - \phi(a_1, b_1)$
n_{11}	$P_{11} = \Phi(a_1, b_1) + \Phi(a_2, b_2) - \Phi(a_1, b_2) - \Phi(a_2, b_1)$	$\phi(a_1, b_1) + \phi(a_2, b_2) - \phi(a_1, b_2) - \phi(a_2, b_1)$
n_{02}	$P_{02} = \Phi(a_1) - \Phi(a_1, b_2)$	$-\phi(a_1, b_2)$
n_{20}	$P_{20} = \Phi(b_1) - \Phi(a_2, b_1)$	$-\phi(a_2, b_1)$
n_{12}	$P_{12} = \Phi(a_2) - \Phi(a_1) - \Phi(a_2, b_2) + \Phi(a_1, b_2)$	$\phi(a_1, b_2) - \phi(a_2, b_2)$
n_{21}	$P_{21} = \Phi(b_2) - \Phi(b_1) - \Phi(a_2, b_2) + \Phi(a_2, b_1)$	$\phi(a_2, b_1) - \phi(a_2, b_2)$
n_{22}	$P_{22} = 1 - \Phi(a_2) - \Phi(b_2) + \Phi(a_2, b_2)$	$\phi(a_2, b_2)$

TABLE 1—(*Continued*)

Observed No., n_{ij}	$\partial P_{ij}/\partial a_1$	$\partial P_{ij}/\partial b_1$	$\partial P_{ij}/\partial a_2$	$\partial P_{ij}/\partial b_2$
n_{00}	$\phi(a_1)\,\Phi(B_{11})$	$\phi(b_1)\,\Phi(A_{11})$	0	0
n_{01}	$\phi(a_1)[\Phi(B_{21}) - \Phi(B_{11})]$	$-\phi(b_1)\,\Phi(A_{11})$	0	$\phi(b_2)\,\Phi(A_{12})$
n_{10}	$-\phi(a_1)\,\Phi(B_{11})$	$\phi(b_1)[\Phi(A_{21}) - \Phi(A_{11})]$	$\phi(a_2)\,\Phi(B_{12})$	0
n_{11}	$\phi(a_1)[\Phi(B_{11}) - \Phi(B_{21})]$	$\phi(b_1)[\Phi(A_{11}) - \Phi(A_{21})]$	$\phi(a_2)[\Phi(B_{22}) - \Phi(B_{12})]$	$\phi(b_2)[\Phi(A_{22}) - \Phi(A_{12})]$
n_{02}	$\phi(a_1)[1 - \Phi(B_{21})]$	0	0	$-\phi(b_2)\,\Phi(A_{12})$
n_{20}	0	$\phi(b_1)[1 - \Phi(A_{21})]$	$-\phi(a_2)\,\Phi(B_{12})$	0
n_{12}	$-\phi(a_1)[1 - \Phi(B_{21})]$	0	$\phi(a_2)[1 - \Phi(B_{22})]$	$\phi(b_2)[\Phi(A_{12}) - \Phi(A_{22})]$
n_{21}	0	$-\phi(b_1)[1 - \Phi(A_{21})]$	$\phi(a_2)[\Phi(B_{12}) - \Phi(B_{22})]$	$\phi(b_2)[1 - \Phi(A_{22})]$
n_{22}	0	0	$-\phi(a_2)[1 - \Phi(B_{22})]$	$-\phi(b_2)[1 - \Phi(A_{22})]$

where $A_{ij} = (a_i - \rho b_j)/\sqrt{(1-\rho^2)}$, $B_{ij} = (b_i - \rho a_j)/\sqrt{(1-\rho^2)}$.

and the joint distribution of X and Y is specified in an obvious way by the P_{ij} in Table 1.

Since there are two additional parameters to be estimated (a_2 and b_2), the work of solving the likelihood equations must be put into a computationally more manageable form. The required formulae are set out in Table 1.

It now becomes convenient for subsequent notation to let $\rho = \theta_1$, $a_1 = \theta_2$, $b_1 = \theta_3$, $a_2 = \theta_4$ and $b_2 = \theta_5$. With these changes we obtain $\partial l/\partial\theta_s = \sum_{ij} n_{ij} P_{ij}^{-1} (\partial P_{ij}/\partial\theta_s)$ and $\mathrm{I} = [I_{st}]$, where

$$I_{st} = N \sum_{ij} P_{ij}^{-1} \left(\frac{\partial P_{ij}}{\partial \theta_t}\right)\left(\frac{\partial P_{ij}}{\partial \theta_s}\right).$$

Hence the elements $\partial l/\partial\theta_s$ and I_{st} can be calculated from Table 1 by multiplying appropriate elements and summing over i and j.

IV. SOLUTION OF THE LIKELIHOOD EQUATIONS

In order to present the general method of solving the likelihood equations for both Cases 1 and 2, we let $\rho = \theta_1$, $a = \theta_2$ and $b = \theta_3$ in Case 1 in conformity with the notational changes introduced above. If $\theta_s^{(1)}$ is a first guess of $\hat{\theta}_s$, the M.L. estimate of θ_s ($s = 1, 2, 3$ for Case 1 and $s = 1, 2, 3, 4, 5$ for Case 2), then numerical substitution of these values in $\partial l/\partial\theta_s$ will give some quantity which we symbolise by $\delta_s^{(1)}$. Denoting the column vector of the $\delta_s^{(1)}$ by $\boldsymbol{\delta}^{(1)}$, then a better estimate of $\hat{\boldsymbol{\theta}}$, $\boldsymbol{\theta}^{(2)}$, may be obtained from the equation

$$\boldsymbol{\theta}^{(2)} = \boldsymbol{\theta}^{(1)} + \mathrm{V}_{(1)}\, \boldsymbol{\delta}^{(1)}, \tag{4}$$

where $\hat{\boldsymbol{\theta}}$, $\boldsymbol{\theta}^{(2)}$ and $\boldsymbol{\theta}^{(1)}$ are column vectors of the $\hat{\theta}_s$, $\theta_s^{(2)}$ and $\theta_s^{(1)}$ respectively and $\mathrm{V}_{(1)} = \mathrm{I}_{(1)}^{-1}$ is the inverse of I with $\boldsymbol{\theta}^{(1)}$ substituted for $\boldsymbol{\theta}$. This process may be repeated until $\boldsymbol{\delta}^{(k+1)}$ is sufficiently small. Successive approximations to $\hat{\boldsymbol{\theta}}$ are $\boldsymbol{\theta}^{(1)}$, $\boldsymbol{\theta}^{(2)}$, $\cdots$ which may be obtained from the relation

$$\boldsymbol{\theta}^{(k+1)} = \boldsymbol{\theta}^{(k)} + \mathrm{V}_{(k)}\, \boldsymbol{\delta}^{(k)}. \tag{5}$$

First approximations to the $\hat{\theta}_s$, $s \neq 1$, are easily obtained. For example, in Case 1, $\hat{a}$ and $\hat{b}$ can be estimated from the marginal frequencies of the 2×2 table giving $a^{(1)} = \Phi^{-1}(\bar{P}_{0.})$ and $b^{(1)} = \Phi^{-1}(\bar{P}_{.0})$. In these expressions $\bar{P}_{0.} = (n_{00} + n_{01})/n$, $\bar{P}_{.0} = (n_{00} + n_{10})/n$ and Φ^{-1} is the inverse function of Φ. These methods are easily extended to 3×3 tables.

A satisfactory preliminary estimate of $\hat{\rho}$ is not so readily arrived at. Luckily, however, in numerical work it will frequently be found that the terms of the information matrix I_{1s}, ($s = 2, 3$ for Case 1 and

$s = 2, 3, 4, 5$ for Case 2) are relatively small, and usually good approximations may be achieved by setting them equal to zero. If this is done, then any guess of $\hat{\theta}_1$, $\theta_1^{(1)}$, may be taken and a better estimate obtained from the formula

$$\theta_1^{(2)} = \theta_1^{(1)} + \delta_1^{(1)}/I_{11} .$$

A third approximation, θ_1' , can now be calculated as

$$\theta_1' = \theta_1^{(2)} + \delta_1'/I_{11}' ,$$

where $\delta_1' = \partial l/\partial\theta_1$ and $I_{11}' = -E(\partial^2 l/\partial\theta_1^2)$ evaluated at $\theta_1 = \theta_1^{(2)}$ and $\theta_s = \theta_s^{(1)}$, $s \neq 1$. The value of $\rho' = \theta_1'$ will usually be close to $\hat{\rho}$ and $V(\rho') = (I_{11}')^{-1}$ will differ only slightly from $V(\hat{\rho})$. Using θ_1' and $\theta_s^{(1)}$, $s \neq 1$, as trial values, a full matrix iteration may be calculated to obtain a closer estimate of $\hat{\theta}$. For most purposes, one matrix adjustment should be sufficient.

It will be noticed that for the solution of the likelihood equations, it is necessary to compute some bivariate-normal volumes. This work is greatly facilitated by the tables presented by Owen [1957] from which the required volumes can be calculated with relatively little effort.

V. EXAMPLE

In order to illustrate the above methods the correlation between first and second lambing records of a flock of 227 Merino ewes will be estimated. These lambings were recorded in 1952 and 1953 and the flock has been described by Turner et al. [1958].

The distribution of lambings for both years is given in Table 2. From the marginal totals we obtain $a_1^{(1)} = -0.2397$, $a_2^{(1)} = 1.5779$, $b_1^{(1)} = -0.0276$, $b_2^{(1)} = 1.1369$. As a first approximation to $\hat{\rho}$, 0.15 was taken as the value for $\rho_1^{(1)}$, and these five estimates were then used

TABLE 2

	1952			
1953	No Lambs	1 lamb	2 lambs	Total
No lambs	58	52	1	111
1 lamb	26	58	3	87
2 lambs	8	12	9	29
Total	92	122	13	227

TABLE 3

n_{ij}	P_{ij}	$\partial P_{ij}/\partial\rho$
$n_{00} = 58$	$P_{00} = 0.2214$	0.1561
$n_{01} = 26$	$P_{01} = 0.1440$	−0.0790
$n_{10} = 52$	$P_{10} = 0.2464$	−0.1117
$n_{11} = 58$	$P_{11} = 0.2145$	0.0649
$n_{02} = 8$	$P_{02} = 0.0399$	−0.0774
$n_{20} = 1$	$P_{20} = 0.0212$	−0.0447
$n_{12} = 12$	$P_{12} = 0.0765$	0.0468
$n_{21} = 3$	$P_{21} = 0.0247$	0.0141
$n_{22} = 9$	$P_{22} = 0.0114$	0.0306
227	1.0000	

to calculate P_{ij} and $\partial P_{ij}/\partial\rho$ from the formulae of Table 1; these values are shown in Table 3, and are used to calculate

$$\sum n_{ij}P_{ij}^{-1}\frac{\partial P_{ij}}{\partial\rho} = 36.268108, \qquad 227\sum P_{ij}^{-1}\left(\frac{\partial P_{ij}}{\partial\rho}\right)^2 = 133.318689$$

$$\rho^{(2)} = 0.15 + \frac{36.268108}{133.318689} = 0.4220.$$

A third approximation to $\hat{\rho}$, ρ', using this technique and $\rho^{(2)}$, $a_1^{(1)}$, $a_2^{(1)}$, $b_1^{(1)}$ and $b_2^{(1)}$ gave $\rho' = 0.4212$. This represents a negligible change from $\rho^{(2)}$.

One matrix iteration was then calculated using the first estimates

TABLE 4

NUMERICAL EVALUATION OF TABLE 1 USING $a_1^{(1)}$, $a_2^{(1)}$, $b_1^{(1)}$, $b_2^{(1)}$ and ρ'

Observed No. n_{ij}	Proportion P_{ij}	$\partial P_{ij}/\partial\rho$	$\partial P_{ij}/\partial a_1$	$\partial P_{ij}/\partial b_1$	$\partial P_{ij}/\partial a_2$	$\partial P_{ij}/\partial b_2$
58	0.2654	0.1700	0.2063	0.1598	0	0
26	0.1198	−0.1028	0.1479	−0.1598	0	0.0448
52	0.2138	−0.1322	−0.2063	0.2231	0.0256	0
58	0.2374	0.1091	−0.1479	−0.2231	0.0517	0.1407
8	0.0201	−0.0672	0.0334	0	0	−0.0448
1	0.0098	−0.0378	0	0.0159	−0.0256	0
12	0.0862	0.0231	−0.0334	0	0.0346	−0.1407
3	0.0260	−0.0063	0	−0.0159	−0.0517	0.0236
9	0.0215	0.0141	0	0	−0.0346	−0.0236

of the a's and b's and ρ'. These calculations are set out in Table 4, from which it may be verified that

$$\delta_1 = \sum n_{ij} \frac{\partial P_{ij}}{\partial \rho} = -0.311962, \qquad \delta_2 = \sum n_{ij} \frac{\partial P_{ij}}{\partial a_1} = -0.483038$$

$$\delta_3 = \sum n_{ij} \frac{\partial P_{ij}}{\partial b_1} = -0.215312, \qquad \delta_4 = \sum n_{ij} \frac{\partial P_{ij}}{\partial a_2} = 0.999523$$

$$\delta_5 = \sum n_{ij} \frac{\partial P_{ij}}{\partial b_2} = -0.476079,$$

and that the information matrix, I, is

$$\mathrm{I} = \begin{bmatrix} 181.061783 & -12.664684 & -13.273162 & 13.530943 & 19.105645 \\ -12.664684 & 159.489818 & -33.902372 & -16.386339 & -11.866292 \\ -13.273162 & -33.902372 & 178.730512 & -7.439942 & -46.856292 \\ 13.530943 & -16.386339 & -7.439942 & 60.652756 & -8.110261 \\ 19.105645 & -11.866292 & -46.856292 & -8.110261 & 108.274246 \end{bmatrix}$$

By calculating $\mathrm{I}^{-1}\delta = \mathrm{V}\delta$, it is found that the appropriate corrections for the five estimates ρ', $a_1^{(1)}$, $b_1^{(1)}$, $a_2^{(1)}$, $b_2^{(1)}$ are -0.0028, -0.0024, -0.0023, 0.0156, -0.0040 respectively. The correction to ρ' does not affect the second decimal place and hence we have

$$\hat{\rho} = 0.42 \pm 0.076$$

since

$$\mathrm{V} = \begin{bmatrix} 0.005761 & 0.000279 & 0.000159 & -0.001326 & -0.001016 \\ 0.000279 & 0.007064 & 0.001909 & 0.002311 & 0.001724 \\ 0.000159 & 0.001909 & 0.006909 & 0.001769 & 0.003303 \\ -0.001326 & 0.002311 & 0.001769 & 0.017972 & 0.002599 \\ -0.001016 & 0.001724 & 0.003303 & 0.002599 & 0.011228 \end{bmatrix}$$

The first trial value of $\hat{\rho}$ was clearly unsatisfactory. However, the convergence of the short method appears to be so rapid that the accuracy of the first guess may be relatively unimportant. This leads to the suggestion that $\rho^{(1)}$ be set equal to zero since this leads to a great reduction in the computations necessary to obtain $\rho^{(2)}$ by the short method. One additional iteration by the latter method before the final matrix iteration should then be sufficient.

VI. DISCUSSION

It frequently happens that the contingency table from which ρ is to be estimated is incomplete with respect to some row or column. This type of situation is easily handled by pooling the relevant P_{ij} and proceeding in exactly the same manner. For instance, in Case 2, if no details are available for the Y classification of $X = x_0$, we obtain $P_{0.} = P_{00} + P_{01} + P_{02} = \Phi(a_1)$ and $n_{0.} = n_{00} + n_{01} + n_{02}$. Hence, in order to obtain the appropriate table, the three lines corresponding to P_{00}, P_{01} and P_{02} in Table 1 are deleted and are replaced by a single line with entries $n_{0.}$, $P_{0.} = \Phi(a_1)$, 0, $\phi(a_1)$, 0, 0, 0. The M.L. estimate of ρ may now be calculated as usual.

Sometimes it may be desirable to estimate ρ from several independent sets of records. If it can be assumed that each set provides estimates of the same parameters, the analysis may be completed with no additional trouble. Let the number of different data sets be m, and let L_q be the likelihood of the qth set, then the likelihood function for the m sets is

$$L \propto \prod_{q=1}^{m} L_q . \tag{6}$$

It is clear from the above expression that numbers in various classes, i.e. (x_0, y_0), (x_1, y_0), $\cdots$ etc., are simply pooled over the m sets of data and the estimation of ρ then proceeds as usual.

Unfortunately, it frequently happens that the m sets of data cannot be assumed to be samples from the same population. In this case each set of data may be analysed separately to obtain m estimates of ρ, $\hat{\rho}_i$, as well as estimates of sampling variances Var $(\hat{\rho}_i)$, where $i = 1, 2, \cdots m$. Once these estimates are available, it is possible to test the homogeneity of the correlation coefficients in the m populations by means of the formula

$$\chi^2_{m-1} \simeq \sum_{i=1}^{m} w_i(\hat{\rho}_i - \bar{\rho})^2 \tag{7}$$

where $w_i = 1/\text{Var}\ (\hat{\rho}_i)$ and $\bar{\rho} = \sum w_i\hat{\rho}_i / \sum w_i$. If the χ^2 value computed from (7) is not significantly large, then $\bar{\rho}$ may be used as the pooled estimate of ρ from the m sets of data with estimated variance

$$\text{Var}\ (\bar{\rho}) = 1/\sum w_i .$$

It is interesting to notice that, for a $p \times q$ table, whenever $p \times q - 1$ is greater than the number of parameters to be estimated, it is possible to test the assumption of an underlying binormal distribution. If the

M.L. estimate of P_{ij} is written as $\hat{P}_{ij}$, then this may be done by computing the quantity

$$Q^2 = \sum_{ij} \frac{(n_{ij} - n\hat{P}_{ij})^2}{n\hat{P}_{ij}},$$

which is distributed asymptotically as χ^2 with $(p \times q - 1 - t)$ degrees of freedom, where t is the number of parameters to be estimated. Should Q^2 be significantly large, then the above procedure for estimating ρ may be unsatisfactory.

The quantity Q^2 was not calculated for the numerical example in the previous section. It was felt that the small numbers in some of the classes would not allow a satisfactory χ^2 test to be performed and would not warrant the additional labour of computing the $\hat{P}_{ij}$.

Finally, it is clear that the methods which have been presented in this paper are easily extended to two dimensional contingency tables of any size. Moreover, since volumes of the trivariate normal distribution have been tabulated by Steck [1958], there is in principle no difficulty in extending the results to three dimensional tables. However, these modifications will not be considered here.

ACKNOWLEDGEMENTS

The author wishes to thank Mrs. J. Williams and Mrs. M. Tonkin for valuable assistance with the numerical example.

Thanks are also extended to Miss H. N. Turner for making available the data used in the numerical example.

REFERENCES

Mosteller, F. [1946]. On some useful "inefficient" statistics. *Ann. Math. Statist. 17*, 377–408.

Owen, D. B. [1956]. Tables for computing bivariate normal probabilities. *Ann. Math. Statist. 27*, 1075–90.

Pearson, K. [1931]. *Tables for Statisticians and Biometricians*, Cambridge University Press.

Steck, G. P. [1958]. A table for computing trivariate normal probabilities. *Ann. Math. Statist. 29*, 780–800.

Turner, Helen Newton, Hayman, R. H., and Prunster, R. W. [1958]. Repeatability of twin births. *Proc. Aust. Soc. Anim. Prod. 2*, 106–7.

A (a)(iv)[4]

Commonwealth of Australia.
COMMONWEALTH SCIENTIFIC AND INDUSTRIAL RESEARCH ORGANIZATION.

Reprinted from *The Australian Journal of Statistics*, Vol. 4, No. 2, pages 49-54, August, 1962.

MAXIMUM LIKELIHOOD ESTIMATION OF PARAMETERS OF THE NORMAL, LOG-NORMAL, TRUNCATED NORMAL AND BIVARIATE NORMAL DISTRIBUTIONS FROM GROUPED DATA

G. M. Tallis and S. S. Y. Young

Division of Animal Genetics, C.S.I.R.O., McMaster Laboratory Glebe, N.S.W.

MAXIMUM LIKELIHOOD ESTIMATION OF PARAMETERS OF THE NORMAL, LOG-NORMAL, TRUNCATED NORMAL AND BIVARIATE NORMAL DISTRIBUTIONS FROM GROUPED DATA[1]

G. M. Tallis and S. S. Y. Young
Division of Animal Genetics, C.S.I.R.O., McMaster Laboratory Glebe, N.S.W.

1. Introduction

The estimation of population moments from grouped data has received considerable attention in the past. It has long been established that for samples classified into numerous equally spaced groups, the class centre may be used to calculate the various moments, and that the bias introduced by this procedure can be corrected by the use of Sheppard's corrections.

In an interesting paper Lindley (1950) investigated the effect of grouping on the Maximum Likelihood (ML) estimation of parameters. In particular, he showed the equivalence of the method of moments and the method of ML in the estimation of the variance of a normal distribution, from data grouped into numerous equally spaced classes. It is clear, however, that when classes are few and of unequal width Sheppard's corrections as well as Lindley's results are inapplicable.

The ML procedure has been used by many workers (e.g. Gupta (1952), Cohen (1957, 1959)) in the estimation of the mean and variance of the normal distribution from samples falling into two or three classes. Samples of these types are referred to as singly or doubly censored samples. The case of many censor classes was considered by Gjeddebaek (1949), who later considered the loss of information due to grouping (Gjeddebaek (1956)).

This paper discusses the ML estimation of parameters for (*a*) the log-normal distribution, (*b*) the truncated normal distribution, and (*c*) the bivariate normal distribution from random samples whose members fall into an arbitrary number of censor classes. A short discussion on asymptotic tests of hypotheses is also presented.

[1] Manuscript received November 27, 1961; revised March 8, 1962.

2. Methods

In the subsequent development it is first of all convenient to consider the normal distribution in order to establish notation as well as some basic formulae. It will be noticed that the second equation of (5) differs from the results of Gjeddeback since differentiation has been with respect to σ^2 instead of σ.

Suppose a random sample of size N is drawn from a normal parent population and the numbers of observations falling into various measurement classes are recorded. The normal population with parameters μ and σ^2 is divided into $(k+1)$ censor classes at the points a_i, $i=1, 2, \ldots k$; $a_{i+1}>a_i$. Let us define

$$(1) \qquad \Phi(b_i)=(2\pi)^{-\frac{1}{2}}\int_{-\infty}^{b_i}\exp\{-\tfrac{1}{2}t^2\}dt=\int_{-\infty}^{b_i}\varphi(t)dt,$$

where $b_i=(a_i-\mu)/\sigma$, then the probability that any single observation will fall into the i^{th} class is given by

$$(2) \qquad p_i=\Phi(b_{i+1})-\Phi(b_i).$$

If the number of observations falling into the i^{th} class is denoted by n_i, where $\sum_{i=0}^{k} n_i=N$, and if, moreover, there is no further information with regard to individual measurements of members of the sample, then the likelihood function, L, is given by

$$(3) \qquad L=C\prod_{i=0}^{k} p_i^{n_i} \text{ and}$$

$$(4) \qquad \ln L\equiv l=K+\Sigma n_i \ln p_i.$$

It may now be verified that

$$(5) \qquad \begin{aligned}\frac{\partial l}{\partial\mu}&=\sigma^{-1}\sum_{i=0}^{k} n_i\{\varphi(b_i)-\varphi(b_{i+1})\}/p_i\\ \frac{\partial l}{\partial\sigma^2}&=(2\sigma^2)^{-1}\sum_{i=0}^{k} n_i\{b_i\varphi(b_i)-b_{i+1}\varphi(b_{i+1})\}/p_i.\end{aligned}$$

These two equations follow from the relations

$$(6) \qquad \begin{aligned}\frac{\partial p_i}{\partial\mu}&=\sigma^{-1}\{\varphi(b_i)-\varphi(b_{i+1})\}\\ \frac{\partial p_i}{\partial\sigma^2}&=(2\sigma^2)^{-1}\{b_i\varphi(b_i)-b_{i+1}\varphi(b_{i+1})\}.\end{aligned}$$

We may obtain $\hat{\mu}$ and $\hat{\sigma}^2$, estimates of μ and σ^2, by equating the right of (5) to zero and solving iteratively. For this purpose the usual scoring method is particularly convenient since the best estimate of the variance-covariance matrix for $\hat{\mu}$ and $\hat{\sigma}^2$ is obtained as part of the computational routine (see Aitchison and Silvey (1960)).

This iterative technique necessitates first estimates of μ and σ^2, $\bar{\mu}$ and $\bar{\sigma}^2$, which may be obtained from the following formulae

$$(7) \qquad \bar{\mu}=\frac{1}{N}\sum_{i=0}^{k} n_i d_i;\quad \bar{\sigma}^2=\frac{1}{N}\sum_{i=0}^{k} d_i^2 n_i-\bar{\mu}^2,$$

where $d_0=a_0$, $d_i=\frac{1}{2}(a_{i+1}+a_i)$, $(i=1, \ldots k-1)$, $d_k=a_k$.

(*a*) *Log-Normal Population*

When a censored sample is taken from a log-normally distributed parent population, ML methods can also be used in the estimation of μ and σ^2. By making the transformation $Y=ln\,X$, where X is the variate under consideration, the problem may be solved in two different ways.

(i) *Direct Method.* By direct integration we have

$$\mu_x=\exp\{\tfrac{1}{2}(\sigma_y^2+2\mu_y)\};\quad \mu'_{2(x)}=\exp\{2(\sigma_y^2+\mu_y)\} \tag{8}$$

and solving these equations we have the relations

$$\mu_y=2ln\,\mu_x-\tfrac{1}{2}ln\,\mu'_{2(x)};\quad \sigma_y^2=ln\,\mu'_{2(x)}-2ln\,\mu_x. \tag{9}$$

Let the points of censoring of X be c_i, $(i=1, 2, \ldots k)$, then we can proceed to estimate μ_x and σ_x^2 by ML methods as before. Now

$$p_i=\Phi(d_{i+1})-\Phi(d_i), \tag{10}$$

where $d_i=(ln\,c_i-\mu_y)/\sigma_y$. Hence we may proceed to differentiate p_i with respect to μ_x and σ_x^2 using (9) to obtain $\partial p_i/\partial\mu_x$ and $\partial p_i/\partial\sigma_x^2$. Using these formulae it is then possible to estimate μ_x and σ_x^2 by iteration as before. However, this procedure is laborious.

(ii) *Indirect Method.* From (8) and (9) it is seen that each pair of μ_y and σ_y^2 uniquely determines another pair of parameters μ_x and σ_x^2. In order to make use of this relationship we require the following well-known lemma.

Lemma: *Let* E *and* F *be two point sets and let* f *be a one to one mapping from* E *onto* F. *If* h *is a real valued and bounded point function defined for all* $x\varepsilon$E *and if* h *attains a unique maximum at* $x_0\varepsilon$E, *then* $g=h(f^{-1})$ *has a unique maximum at* $y_0=f(x_0)\varepsilon$F.

If we identify the sample spaces of $\hat{\mu}_y$ and $\hat{\sigma}_y^2$ and $\hat{\mu}_x$ and $\hat{\sigma}_x^2$ with the sets of E and F respectively, then since the condition of a one to one mapping relating elements in the two spaces is satisfied, we have, from the above lemma,

$$\hat{\mu}_x=\exp\{\tfrac{1}{2}(\hat{\sigma}_y^2+2\hat{\mu}_y)\};\quad \hat{\sigma}_x^2=\exp\{2(\hat{\mu}_y+\tfrac{1}{2}\hat{\sigma}_y^2)\}[\exp\{\hat{\sigma}_y^2\}-1] \tag{11}$$

where $\hat{\mu}_y$ and $\hat{\sigma}_y^2$ are the ML estimates of μ_y and σ_y^2 calculated for the normal distribution, using $a_i=ln\,c_i$ as censor points. The large sample variances of $\hat{\mu}_x$ and $\hat{\sigma}_x^2$ are given by

$$\begin{aligned} V(\hat{\mu}_x)&=\mu_x^2\{\tfrac{1}{4}V(\hat{\sigma}_y^2)+V(\hat{\mu}_y)+C(\hat{\sigma}_y^2,\hat{\mu}_y)\}.\ \text{and}\\ V(\hat{\sigma}_x^2)&=4\sigma_x^2V(\hat{\mu}_y)+4\sigma_x^2[\sigma_x^2+\exp\{\mu_y+\sigma_y^2\}]C(\hat{\mu}_y,\hat{\sigma}_y^2)\\ &\quad+[\sigma_x^2+\exp\{\mu_y+\sigma_y^2\}]^2V(\hat{\sigma}_y^2). \end{aligned} \tag{12}$$

(*b*) *Truncated Normal Distribution*

We now consider the case where we have a censored sample of size N drawn from a normal distribution truncated at a_0 and the portion lying to the left of a_0 is α. The estimation of μ and σ^2 for the

full distribution now proceeds in a straightforward fashion. Using the notation of the previous sections we have

$$q_i=p_i/\alpha, \text{ where } \alpha=1-\Phi(b_0). \text{ Hence}$$

$$l=ln\, L=K+\Sigma n_i\, ln\, q_i=K+\Sigma n_i\, ln\, p_i-N\, ln\, \alpha.$$

The latter relation gives us immediately

$$\frac{\partial l}{\partial \mu}=\sigma^{-1}\left\{\sum_{i=0}^{k} n_i\{\varphi(b_i)-\varphi(b_{i+1})\}/p_i-N\varphi(b_0)/\alpha\right\},$$

(13)

$$\frac{\partial l}{\partial \sigma^2}=(2\sigma^2)^{-1}\left\{\sum_{i=0}^{k} n_i\{b_i\varphi(b_i)-b_{i+1}\varphi(b_{i+1})\}/p_i-Nb_0\varphi(b_0)/\alpha\right\}.$$

In order that these equations may be solved iteratively, it is useful to have the quantities $\partial q_i/\partial\mu$ and $\partial q_i/\partial\sigma^2$ given below

$$\frac{\partial q_i}{\partial \mu}=\alpha^{-1}\left\{\frac{\partial p_i}{\partial \mu}-\frac{p_i}{\alpha}\frac{\partial \alpha}{\partial \mu}\right\}=\alpha^{-1}\left\{\frac{\partial p_i}{\partial \mu}+q_i\frac{\varphi(b_0)}{\sigma}\right\}$$

(14)

$$\frac{\partial q_i}{\partial \sigma^2}=\alpha^{-1}\left\{\frac{\partial p_i}{\partial \sigma^2}+q_i\frac{b_0\varphi(b_0)}{2\sigma^2}\right\},$$

where $\frac{\partial p_i}{\partial \mu}$ and $\frac{\partial p_i}{\partial \sigma^2}$ are as given previously.

(c) *Bivariate Normal Distribution*

As an extension of the above methods, suppose a censored random sample is drawn from a bivariate normal population with parameters μ_x, μ_y, σ_x^2, σ_y^2, and ρ. Let the censor points on the x axis be a_i, $i=1, 2, \ldots k$, and on the y axis c_j, $j=1, 2, \ldots l$. Then, the probability that an observation fall in the sub-space $a_i<X<a_{i+1}$, $c_j<Y<c_{j+1}$ is given by

$$(15)\quad P_r\{a_i<X<a_{i+1},\, c_j<Y<c_{j+1}\}=\int_{b_i}^{b_{i+1}}\int_{f_j}^{f_{j+1}}\varphi(u,v)dvdu=P_{ij},$$

where $\varphi(u,v)$ is the standardized bivariate normal density function, $b_i=(a_i-\mu_x)/\sigma_x$ and $f_j=(c_j-\mu_y)/\sigma_y$. Briefly, then, the appropriate likelihood function is

$$(16)\qquad L=C\prod_{i,j}P_{ij}^{n_{ij}}$$

and the ML estimates for the required parameters can be obtained by the scoring method discussed earlier. It is now only necessary to calculate $\partial P_{ij}/\partial\mu_x$, and so on. This may best be achieved by writing

$$(17)\quad P_{ij}=\Phi(b_i,f_j)+\Phi(b_{i+1},f_{j+1})-\Phi(b_i,f_{j+1})-\Phi(b_{i+1},f_j)$$

where, for instance,

$$\Phi(b_i,f_j)=\int_{-\infty}^{b_i}\int_{-\infty}^{f_j}\varphi(u,v)dvdu.$$

Now, it may be verified that

$$
\begin{aligned}
&\frac{\partial \Phi}{\partial \mu_x} = -\sigma_x^{-1}\varphi(b_i)\Phi(F_{ji}), \quad \frac{\partial \Phi}{\partial \mu_y} = -\sigma_y^{-1}\varphi(f_j)\Phi(B_{ij}) \\
&\frac{\partial \Phi}{\partial \sigma_x^2} = -\frac{b_i}{2\sigma_x^2}\varphi(b_i)\Phi(F_{ji}), \quad \frac{\partial \Phi}{\partial \sigma_y^2} = -\frac{f_j}{2\sigma_y^2}\varphi(f_j)\Phi(B_{ij})
\end{aligned} \tag{18}
$$

and $\partial\Phi/\partial\rho = \varphi(b_i, f_j)$, where $F_{ji} = \dfrac{f_j - \rho b_i}{(1-\rho^2)^{\frac{1}{2}}}$, $B_{ij} = \dfrac{b_i - \rho f_j}{(1-\rho^2)^{\frac{1}{2}}}$.

From these results it is clear that

$$
\begin{aligned}
\frac{\partial P_{ij}}{\partial \mu_x} &= -\sigma_x^{-1}\{\varphi(b_i)\Phi(F_{ji}) + \varphi(b_{i+1})\Phi(F_{j+1,i+1}) \\
&\quad -\varphi(b_i)\Phi(F_{j+1,i}) - \varphi(b_{i+1})\Phi(F_{j,i+1})\} \\
\frac{\partial P_{ij}}{\partial \mu_y} &= -\sigma_y^{-1}\{\varphi(f_j)\Phi(B_{ij}) + \varphi(f_{j+1})\Phi(B_{i+1,j+1}) \\
&\quad -\varphi(f_j)\Phi(B_{i+1,j}) - \varphi(f_{j+1})\Phi(B_{i,j+1})\} \\
\frac{\partial P_{ij}}{\partial \sigma_x^2} &= -(2\sigma_x^2)^{-1}\{b_i\varphi(b_i)\Phi(F_{ji}) + b_{i+1}\varphi(b_{i+1})\Phi(F_{j+1,i+1}) \\
&\quad -b_i\varphi(b_i)\Phi(F_{j+1,i}) - b_{i+1}\varphi(b_{i+1})\Phi(F_{j,i+1})\} \\
\frac{\partial P_{ij}}{\partial \sigma_y^2} &= -(2\sigma_y^2)^{-1}\{f_j\varphi(f_j)\Phi(B_{ij}) + f_{j+1}\varphi(f_{j+1})\Phi(B_{i+1,j+1}) \\
&\quad -f_{j+1}\varphi(f_{j+1})\Phi(B_{i,j+1}) - f_j\varphi(f_j)\Phi(B_{i+1,j})\} \\
\frac{\partial P_{ij}}{\partial \rho} &= \varphi(b_i f_j) + \varphi(b_{i+1}, f_{j+1}) - \varphi(b_i, f_{j+1}) - \varphi(b_{i+1}, f_j).
\end{aligned} \tag{19}
$$

From the above expressions the ML estimates of the five parameters may be obtained iteratively.

(*d*) *Asymptotic Tests*

It is interesting to notice that the assumption that one of the three distributions discussed above describes the data satisfactorily can be tested provided $k>3$ in the univariate cases and $kl>6$ in the bivariate case. A suitable test can be carried out with the aid of the following formula

$$
Q^2 = \sum_{i=0}^{k} \frac{(n_i - N\hat{\omega}_i)^2}{N\hat{\omega}_i} \tag{20}
$$

where $\hat{\omega}_i$ is the estimated probability associated with the i^{th} class. For example, in the case of the normal distribution,

$$
\hat{\omega}_i = p_i = \int_{a_i}^{a_{i+1}} \varphi\left(\frac{x-\hat{\alpha}}{\hat{\sigma}}\right)\frac{dx}{\hat{\sigma}}. \tag{21}
$$

The statistic Q^2 is asymptotically distributed as χ^2 with $(k-3)$ or $(kl-6)$ degrees of freedom, and significantly large values of Q^2 indicate discrepancies in the hypothesis of the underlying distribution.

Another problem which may arise is the testing of the homogeneity of several independent estimates of μ, σ^2 or ρ. If we let θ stand for either of these parameters and if there are J estimates of θ, then

$$(22) \qquad W^2 = \sum_{j=1}^{J} w_j(\hat{\theta}_j - \bar{\theta})^2, \; w_j = 1/\mathrm{Var}\,(\hat{\theta}_j), \; \bar{\theta} = \Sigma w_j \hat{\theta}_j / \Sigma w_j$$

is approximately χ^2 distributed with $(J-1)$ degrees of freedom. Significant values of W^2 indicate heterogeneity.

3. Acknowledgement

Some of the problems investigated in this work were brought to our attention by Dr. Yvonne Cossart in connection with her population studies of immunity to influenza viruses.

References

Aitchison, J., and Silvey, S. D. (1960). "Maximum-likelihood procedures and associated tests of significance." *J. Roy. Statist. Soc.*, B 22, 154–171.

Cohen, A. C. (1957). "On the solution of estimating equations for truncated and censored samples from normal populations." *Biometrika*, 44, 225–236.

Cohen, A. C. (1959). "Simplified estimators for the normal distribution when samples are singly censored or truncated." *Technometrics*, 1, 217–237.

Gjeddebaek, N. F. (1949). "Contribution to the study of grouped observations. Application of the method of Maximum Likelihood in case of normally distributed observations." *Skand. Aktuarietidskr.*, 32, 135–159.

Gjeddebaek, N. F. (1956). "Contribution to the study of grouped observations. II. Loss of information caused by grouping of normally distributed observations." *Skand. Aktuarietidskr.*, 39, 154–159.

Gupta, A. K. (1952). "Estimation of the mean and standard deviation of a norma population from a censored sample." *Biometrika*, 39, 260–273.

Lindley, D. V. (1950). "Grouping corrections and maximum likelihood equations." *Proc. Camb. Philos. Soc.*, 46, 106–110.

AUSTRALASIAN MEDICAL PUBLISHING CO. LTD.
SEAMER AND ARUNDEL STS., GLEBE, SYDNEY

A (a) (iv) [5]

TECHNOMETRICS VOL. 9, NO. 4 NOVEMBER 1967

Approximate Maximum Likelihood Estimates From Grouped Data

G. M. TALLIS*

C. S. I. R. O., Newtown, Australia

In this paper methods are developed for obtaining approximate maximum likelihood, m.l., estimates of parameters of distribution functions, d.f., under conditions where the data are grouped. Both the cases of a multivariate d.f. under equal grouping on all co-ordinates and a one-dimensional d.f. under unequal grouping are considered. These are extensions to Lindley's results of 1949.

The procedure requires that for any particular d.f. and grouping set-up, a correction factor, usually depending on the grouping width, be calculated. This factor when added to a specific initial estimate should provide reasonably close approximations to the m.l. estimates under grouping. The sampling variances of these approximations are also obtained and they are found to depend, in part, on the squares of the interval widths.

1. INTRODUCTION

This paper considers the problem of obtaining approximate maximum likelihood, m.l., estimates of parameters from grouped data, where the grouping intervals are set in advance. The interval widths are assumed to be under the control of the experimenter so that they may be kept small enough to enable the methods developed below to be applied.

This situation often arises in the determination of the potency of some chemical or biological substance. If mass titration techniques are used, it will be possible to state that an individual's titre lies between two well-defined boundaries, i.e. the individual belongs to a certain titre class. From a random sample whose members have been put into these titre classes it is often required to estimate the parameters of some assumed background distribution.

Another more specific example, which will be considered later in greater detail, concerns the failure of certain objects which are subject to the exponential failure law. The precise time of failure of each object may be unknown and the only available information may be that failure occurred between two well-defined time points, two inspection times. If data are collected on the failure of a number of such objects, then these can be classified according to the inspection period during which failure occurred. It may then be required to estimate the failure rate from the numbers in the various failure groups.

Received August 1965; revised January 1967.

* This work was supported in part by Research Grant GM-13225 from the National Institutes of Health while the author was a Visiting Associate Professor in the Biometrics Unit, Plant Breeding Department, Cornell University, in July-August 1966 while on leave from the Johns Hopkins University.

The purpose of this paper is to extend the results of Lindley (1), who established procedures for obtaining approximate m.l. estimates for parameters of a one-dimensional distribution function, d.f., under equal interval grouping. The generalizations include the cases of k-dimensional d.f.s. under equal grouping and one-dimensional d.f.s. under unequal grouping.

The organization of the material is as follows. In the next section the essential notation and results are presented with the aid of a numerical illustration. Further algebraic examples include the univariate and bivariate normal distributions. The appendix contains an outline of some derivations of the various formulas.

2. Results and Examples

Consider the frequency function $f(x, \theta)$ depending on a single parameter θ. Let the real line, R, be partitioned into intervals of equal width h and centres $x_{(i)}$, i.e., $U_i[x_{(i)} - h/2, x_{(i)} + h/2] = R$. A random sample of size N is now drawn from a population with frequency function f and the numbers falling in each interval counted. Let N_i be the number of observations falling in $[x_{(i)} - h/2, x_{(i)} + h/2]$, then it is required to obtain the m.l. estimate of θ from the grouped sample.

If we let

$$p_i(\theta) = \int_{x_{(i)}-h/2}^{x_{(i)}+h/2} f(x, \theta)\, dx,$$

then an application of standard techniques would lead us to maximize

$$L(\theta) = C \prod_i [p_i(\theta)]^{N_i} \tag{2.1}$$

with respect to θ. However, the work required to accomplish this is often considerable and, if many estimates are required using the same intervals, it is reasonable to search for approximate methods requiring less effort.

One such approximation is as follows. Let θ_0 be the m.l. estimate of θ calculated on the basis of no grouping using the class centres, $x_{(i)}$, as the observed values of the variable. Thus, $N_1 x_{(1)}$'s, $N_2 x_{(2)}$'s, etc. constitute the sample from which θ_0 is computed. Then under certain conditions the approximate m.l. estimate under grouping is given by

$$\hat{\theta} = \theta_0 + \delta, \qquad \delta = V(\theta_0) e(\theta_0) \tag{2.2}$$

where

$$[V(\theta_0)]^{-1} = -E\{\partial^2 \ln f(x, \theta_0)/\partial\theta^2\}$$

and

$$e(\theta_0) = E\{\partial[h^2 f''(x, \theta_0)/24 f(x, \theta_0)]/\partial\theta\}, \qquad f'' = \partial^2 f/\partial x^2.$$

This is a slight, but convenient, modification of the result of Lindley (1).

The variance of $\hat{\theta}$ can be calculated from the formula

$$NV(\hat{\theta}) = [I - h^2/24 E\{\partial^2 A/\partial\theta^2 + A\partial^2 \ln f/\partial\theta^2 - f^{-1} D_x^2[(\partial^2 \ln f/\partial\theta^2) f]\}]^{-1}.$$

In the above formula $I = I(\theta) = -E\{\partial^2 \ln f(x, \theta)/\partial\theta^2\}$, $A = f''/f$ and $D_x = d/dx$

and it is clear that both $\hat{\theta}$ and its variance may depend on h^2, the square of the interval width.

We now illustrate these results by means of the exponential failure example discussed in the Introduction. Suppose the objects which are failing have life times distributed as $f(x, \theta) = \theta \exp\{-\theta x\}$ and we are required to estimate θ from the grouped data.

In the present case, $L(\theta) \propto \theta^N \exp\{-N\theta\bar{x}\}$ and the m.l. estimate (ungrouped) of θ is $\bar{x}^{-1}$. Hence, θ_0 is simply $[\sum_i N_i x_{(i)}/N]^{-1}$.

Further calculations show that $V(\theta) = \theta^2$ and $f''(x, \theta) = \theta^3 \exp\{-\theta x\}$ and therefore

$$c = \partial[E\{f''/f\}h^2/24]/\partial\theta = \theta h^2/12$$

and

$$\hat{\theta} = (\sum_i N_i x_{(i)}/N)^{-1}[1 + (\sum_i N_i x_{(i)}/N)^{-2}h^2/12].$$

We now require the variance of $\hat{\theta}$, $V(\hat{\theta})$. The quantity $I(\theta)$ is easily shown to be equal to θ^{-2} and $\partial^2(f''/f)/\partial\theta^2 = 2$, $(f''/f)\partial^2 \ln f/\partial\theta^2 = -1$. Moreover, $E\{f^{-1} D_x^2[\partial^2 \ln f/\partial\theta^2 f]\} = D_x[\partial^2 \ln f/\partial\theta^2 f]_0^\infty = e^{-\theta x}|_0^\infty = -1$ and finally therefore

$$V(\hat{\theta}) = \theta^2/N[1 - (\theta h)^2/12].$$

As a numerical example a random sample of size 200 was drawn from an exponential population with $\theta = 1$. This sample was put into classes of width 1.0 and class centres $x_{(i)} = .5 + (i - 1)$ for $i = 1, 2, \cdots, 6$. The N_i, in their respective order, were as follows: 126, 42, 22, 6, 3, 1. For these data $\theta_0 = .905$ and

$$\hat{\theta} = .905 + (.905)^3/12 = .967$$

while

$$V(\hat{\theta}) = \{200[1 - 1/12]\}^{-1} = .00545.$$

The above figure for $V(\hat{\theta})$ emphasizes that little information has been lost as a result of grouping since the minimum variance bound is .005.

In the Appendix the above techniques are extended to multivariate distributions which are functions of several parameters, and the case of unequal class widths is also considered. We summarize here the multivariate extensions and refer to the appendix for the formulas of unequal grouping, specifically formulas 4.8 and 4.9.

Let $f(\mathbf{x}, \boldsymbol{\theta})$ be a frequency function for the k-dimensional random vector $\mathbf{X}$ depending on the s-dimensional parameter vector $\boldsymbol{\theta}$. If each axis is divided into equal intervals, h_i being the width of the intervals on the ith axis, the sample space can be partitioned into k-dimensional boxes of identical shape and volume. If $\boldsymbol{\theta}_0$ is the m.l. estimate of $\boldsymbol{\theta}$ using the class centres as the observed values of $\mathbf{X}$, then the approximate m.l. estimates of $\boldsymbol{\theta}$ under grouping is given by

$$\hat{\boldsymbol{\theta}} = \boldsymbol{\theta}_0 + \mathbf{V}(\boldsymbol{\theta}_0)\mathbf{e}(\boldsymbol{\theta}_0) \tag{2.3}$$

where $\mathbf{V}^{-1}(\boldsymbol{\theta}_0) = \mathbf{I}(\boldsymbol{\theta}_0) = [-E\{\partial^2 \ln f(\mathbf{x}, \boldsymbol{\theta}_0)/\partial\theta_i\partial\theta_j\}]$ and

$$\mathbf{e}(\boldsymbol{\theta}_0)' = [e_1(\boldsymbol{\theta}_0), \cdots, e_s(\boldsymbol{\theta}_0)], \quad e_i(\boldsymbol{\theta}_0) = E\left\{\partial\left[\sum_{j=1}^{k} h_j^2 f_{ij}(\mathbf{x}, \boldsymbol{\theta}_0)/24 f(\mathbf{x}, \boldsymbol{\theta}_0)\right] \Big/ \partial\theta_i\right\},$$

$f_{jj} = \partial^2 f/\partial x^{(j)2}$, where $x^{(j)}$ is the variable of the jth co-ordinate. This is completely analogous to the simpler case discussed above.

The approximate covariance matrix for $\hat{\theta}$, $V(\hat{\theta})$, can be obtained from the formula $NV(\hat{\theta}) = [I(h)]^{-1}$, where

$$I_{qr}(h) = I_{qr} - \sum_{i=1}^{k} (h_i^2/24)\ E\{\partial^2 A_i/\partial\theta_r\,\partial\theta_q + A_i\,\partial^2 \ln f/\partial\theta_r\,\partial\theta_q - f^{-1} D_i^2[\partial^2 \ln f/\partial\theta_r\,\partial\theta_q f]\} \tag{2.4}$$

$$I_{qr} = -E\{\partial^2 \ln f/\partial\theta_r\,\partial\theta_q\}, \qquad A_i = f_{ii}/f \quad \text{and} \quad D_i = \partial/\partial x^{(i)}.$$

In order to illustrate the use of (2.3) and (2.4) two algebraic examples are worked in the next section.

3. Further Algebraic Examples

(a) *The univariate normal distribution*

The above results are easily applied to the normal frequency function

$$f(x, \theta) = (2\pi\sigma^2)^{-\frac{1}{2}} \exp\{-(x-\mu)^2/2\sigma^2\}.$$

In this case

$$f''/f = \sigma^{-2}\{(x-\mu)^2/\sigma^2 - 1\}$$

and

$$E\{\partial(f''/f)/\partial\mu\} = 0, \qquad E\{\partial(f''/f)/\partial\sigma^2\} = -\sigma^{-4}$$

and from (2.3)

$$\begin{bmatrix}\delta_1\\ \delta_2\end{bmatrix} = \begin{bmatrix}\sigma^2 & 0\\ 0 & 2\sigma^4\end{bmatrix}\begin{bmatrix}0\\ -(h\sigma^{-2})^2/24\end{bmatrix} = \begin{bmatrix}0\\ -h^2/12\end{bmatrix} \tag{3.1}$$

as expected. Therefore, the approximate m.l. estimates of μ and σ^2 are $\bar{x}$ and $s^2 - h^2/12$, where these are calculated from the class centres. It is easy to show using (2.4) that the variance of these estimators is $\sigma^2\{N(1 - h^2/12\sigma^2)\}^{-1}$ and $2\sigma^4\{N(1 - h^2/6\sigma^2)\}^{-1}$ respectively.

(b) *The bivariate normal distribution*

As a second illustration of these techniques we use the bivariate normal distribution

$$f(x_1, x_2, 0) = K \exp\{-Q/2\}$$

where $K = (2\pi)^{-1}(\sigma_1^2\sigma_2^2 - \sigma_{12}^2)^{-\frac{1}{2}}$ and Q is the quadratic form

$$Q = C^{-1}\{\sigma_2^2(x_1-\mu_1)^2 + \sigma_1^2(x_2-\mu_2)^2 - 2\sigma_{12}(x_1-\mu_1)(x_2-\mu_2)\}$$

writing $C = \sigma_1^2\sigma_2^2 - \sigma_{12}^2$. The parameter σ_{12} is the covariance between X_1 and X_2.

Some calculations show that

$$f_{11}/f = \sigma_2^4 C^{-2}\{x_1 - \mu_1 - \sigma_{12}\sigma_2^{-2}(x_2-\mu_2)\}^2 - \sigma_2^2 C^{-1}$$

$$f_{22}/f = \sigma_1^4 C^{-2}\{x_2 - \mu_2 - \sigma_{12}\sigma_1^{-2}(x_1-\mu_1)\}^2 - \sigma_1^2 C^{-1}$$

and it is readily verified that

$$E\{\partial[f_{11}h_1^2/24f + f_{22}h_2^2/24f]/\partial\mu_i\} = 0, \qquad i = 1, 2$$

and hence $e_1 = e_2 = 0$. Now

$$E\{\partial(f_{11}/f)/\partial\sigma_1^2\} = -\sigma_2^4C^{-2}, \qquad E\{\partial(f_{22}/f)/\partial\sigma_1^2\} = -\sigma_{12}^2C^{-2}$$

and we have $e_3 = -(24C^2)^{-1}(\sigma_2^4h_1^2 + \sigma_{12}^2h_2^2)$, $e_4 = -(24C^2)^{-1}(\sigma_1^4h_2^2 + \sigma_{12}^2h_1^2)$. Similarly, it is found that $e_5 = \sigma_{12}(12C^2)^{-1}(\sigma_2^2h_1^2 + \sigma_1^2h_2^2)$.

In this case it is well known that the appropriate dispersion matrix, V, is given by

$$\mathrm{V} = \begin{bmatrix} \sigma_1^2 & \sigma_{12} & 0 & 0 & 0 \\ \sigma_{12} & \sigma_2^2 & 0 & 0 & 0 \\ 0 & 0 & 2\sigma_1^4 & 2\sigma_{12}^2 & 2\sigma_1^2\sigma_{12} \\ 0 & 0 & 2\sigma_{12}^2 & 2\sigma_2^4 & 2\sigma_2^2\sigma_{12} \\ 0 & 0 & 2\sigma_1^2\sigma_{12} & 2\sigma_2^2\sigma_{12} & \sigma_1^2\sigma_2^2 + \sigma_{12}^2 \end{bmatrix}$$

and appropriate multiplication and reduction gives

$$\delta_1 = \delta_2 = 0, \qquad \delta_3 = -h_1^2/12, \qquad \delta_4 = -h_2^2/12, \qquad \delta_5 = 0. \tag{3.2}$$

These corrections agree with the bivariate Sheppard corrections given by Wold (3).

The means are estimated independently of the variances, and the appropriate elements of the information matrix for the means are

$$N^{-1}I_{11}(h) = \sigma_2^2C^{-1}[1 - (12C)^{-1}(h_1^2\sigma_2^2 + h_2^2\sigma_1^2\rho^2)]$$

$$N^{-1}I_{12}(h) = -\sigma_{12}C^{-1}[1 - (12C)^{-1}(h_1^2\sigma_2^2 + h_2^2\sigma_1^2)]$$

$$N^{-1}I_{22}(h) = \sigma_1^2C^{-1}[1 - (12C)^{-1}(h_1^2\sigma_2^2\rho^2 + h_2^2\sigma_1^2)]$$

The information matrix for the estimates of σ_i^2 and σ_{12} from grouped data is not easily obtainable from (2.4). Fortunately, however, the work may be done in a more direct manner. We use the well known formula

$$C(\bar{m}_{qr}, \bar{m}_{st}) = N^{-1}(\bar{\mu}_{q+s,r+t} - \bar{\mu}_{qr}\bar{\mu}_{st}),$$

where the bar indicates the distribution under grouping and substitute for $\bar{\mu}_{q+s,r+t}$ and $\bar{\mu}_{qr}$ in terms of $\mu_{q+s,r+t}$, μ_{qr}, h_1^2 and h_2^2. The required formulas are, neglecting fourth order terms in h,

$$\bar{\mu}_{20} = \mu_{20} + h_1^2/12, \qquad \bar{\mu}_{11} = \mu_{11}, \qquad \bar{\mu}_{31} = \mu_{31} + \mu_{11}h_1^2/4$$

$$\bar{\mu}_{22} = \mu_{22} + \mu_{20}h_2^2/12 + \mu_{02}h_1^2/12, \qquad \bar{\mu}_{40} = \mu_{40} + \mu_{20}h_1^2/2$$

the other quantities being obtainable by suitably permuting the subscripts. After some reductions we find the required dispersion matrix to be

$$\bar{\mathrm{M}} = \mathrm{M} + \mathrm{K} \tag{3.3}$$

where

$$\mathrm{K} = (12N)^{-1}\begin{bmatrix} 4\mu_{20}h_1^2 & 0 & 2\mu_{11}h_1^2 \\ 0 & 4\mu_{02}h_2^2 & 2\mu_{11}h_2^2 \\ 2\mu_{11}h_1^2 & 2\mu_{11}h_2^2 & \mu_{20}h_2^2 + \mu_{02}h_1^2 \end{bmatrix}$$

and M is the usual dispersion matrix for sample variances and covariances from a normal distribution (see the formula for V).

Because of the bias of the estimators $\bar{m}_{20} = \bar{s}_1^2$ and $\bar{m}_{02} = \bar{s}_2^2$, the usual estimate of the population correlation coefficient, ρ, from grouped data is

$$\bar{r} = \bar{m}_{11}(s_1 s_2)^{-1} \tag{3.4}$$

where $s_1^2 = \bar{s}_1^2 - h_1^2/12$ and $s_2^2 = \bar{s}_2^2 - h_2^2/12$. The sampling variance of $\bar{r}$ can be obtained from the general formula for the variance of a sample product-moment correlation coefficient by substituting for the grouped moments $\bar{\mu}_{sr}$ as above. The result is, again neglecting fourth order terms,

$$V(\bar{r}) = N^{-1}(1 - \rho^2)^2 + (12N)^{-1}(h_1^2/\sigma_1^2 + h_2^2/\sigma_2^2)(1 - \rho^4). \tag{3.5}$$

The above work can be extended to the multivariante normal distribution. The necessary algebra is long and tedious and it transpires that no corrections for the means and covariances are required and the corrected estimates of variances are $s_i^2 = \bar{s}_i^2 - h_i^2/12$, $i = 1, 2, \cdots, k$. The covariance matrix of the estimated means is $N^{-1}\{\Sigma + 12^{-1}\mathrm{H}^2\}$ where $\mathrm{H}^2 = \mathrm{diag}\,(h_1^2, h_2^2, \cdots, h_k^2)$, $\Sigma = \mathrm{Var}\,(\mathrm{X})$, while the variances and covariances of the estimated second central moments are obtainable by an extension of the argument for $k = 2$.

Appendix

(*a*) *Equal interval grouping*

Consider the subspace of Euclidean k-space, E_k, defined by

$$B_k = [a_1, b_1] \times [a_2, b_2] \times \cdots \times [a_k, b_k].$$

The end points of the intervals, a_i and b_i, need not necessarily be finite. A partition of B_k is set up as follows. If a_i and b_i are both finite, then divide $[a_i, b_i]$ into n_i subintervals of length $h_i = (b_i - a_i)/n_i$, the first subinterval being centred at $a_i + h_i/2$ and the last at $b_i - h_i/2$. Now label the class centres $a + h_i/2, a + 3h_i/2, \cdots, b - h_i/2$ by $x_1^{(i)}, x_2^{(i)}, \cdots, x_{n_i}^{(i)}$ and let $A_i = \{x_j^{(i)}\}$. If, say, $b_i = \infty$, then $x_1^{(i)} = a_i + h_i/2$, where h_i is arbitrary, and A_i is now an infinite set. Similar, obvious modifications are made if $a_i = -\infty$ and b_i is finite or $a_i = -\infty$ and $b_i = \infty$.

It will be assumed subsequently that there are k variates X_i, $i = 1, 2, \cdots, k$, with distribution and frequency functions $F(\mathrm{x}, \theta)$ and $f(\mathrm{x}, \theta)$ respectively, where θ is a column vector of s parameters. It is further assumed that $A = A_1 \times A_2 \times \cdots \times A_k$, the cartesian product of the sets A_i, is given and that a partition has been set up according to the above procedure. Then, the probability that a random vector X fall in a particular k-dimensional box is given by

$$p(\mathrm{x}, \theta) = \int_{x^{(1)}-h_1/2}^{x^{(1)}+h_1/2} \cdots \int_{x^{(k)}-h_k/2}^{x^{(k)}+h_k/2} f(\mathrm{y}, \theta)\, d\mathrm{y} \tag{4.1}$$

In (4.1) the subscipt j on $x_j^{(i)}$ has been dropped, but this should lead to no confusion if it is understood that $\mathrm{x} \in A$. It can be shown that (4.1) can be written as

$$p(\mathrm{x}, \theta) = \prod_{i=1}^{k} h_i \left[f + \sum_{i=1}^{k} f_{ii} h_i^2/24 + O(h^3) \right]_{y=x}. \tag{4.2}$$

Now let θ_0 be the m.l. estimate of θ for the function $f(\mathrm{x}, \theta)$ calculated from a sample of size N. Under grouping, θ_0 is calculated using class centres, since the actual values are unknown, and it is required to obtain an adjustment, δ, for grouping. In order to accomplish this we notice that, under suitable regularity conditions,

$$\partial \ln p(\mathrm{x}, \theta)/\partial\theta_s = \partial \ln f/\partial\theta_s + \partial\left[\sum_{i=1}^{k} f_{ii} h_i^2/24f \right]/\partial\theta_s + O(h^3). \tag{4.3}$$

Now applying the Newton–Raphson method of finding roots to equations, see e.g. Ralston (2), and summing over all sample values

$$\begin{aligned}
&\delta_1 \sum \partial^2 \ln f/\partial\theta_1^2 \quad + \cdots + \delta_s \sum \partial^2 \ln f/\partial\theta_s\, \partial\theta_1 = -\sum \partial\left[\sum_{i=1}^{k} f_{ii} h_i^2/24f \right]/\partial\theta_1 \\
&\qquad\vdots \qquad\qquad\qquad\qquad \vdots \qquad\qquad\qquad\qquad \vdots \\
&\delta_1 \sum \partial^2 \ln f/\partial\theta_1\, \partial\theta_s + \cdots + \delta_s \sum \partial^2 \ln f/\partial\theta_s^2 \quad = -\sum \partial\left[\sum_{i=1}^{k} f_{ii} h_i^2/24f \right]/\partial\theta_s
\end{aligned} \tag{4.4}$$

In (4.4) each function is evaluated at $\theta = \theta_0$ and all terms of order higher than h^2 have been dropped. In application, (4.4) may be somewhat tedious, even in simple cases and considerable improvement is achieved by replacing the various terms by their expectation. This can be done without altering the order of the neglected terms and we obtain the average bias from

$$\delta = \mathrm{V}(\theta_0)\mathrm{e}(\theta_0), \tag{4.5}$$

where $\mathrm{V}(\theta_0)$ is the inverse of the information matrix I for a sample of size one evaluated at $\theta = \theta_0$ and $\mathrm{e}(\theta_0)$ is an $(s \times 1)$ column vector with elements

$$e_i(\theta_0) = E\left\{ \partial\left[\sum_{i=1}^{k} f_{ii}(\mathrm{x}, \theta_0) h_i^2/24f(\mathrm{x}, \theta_0) \right]/\partial\theta_i \right\}.$$

The formula for the information matrix associated with these estimators is given by (2.4).

(*b*) *Unequal grouping*

Some progress can also be made even if the class widths vary. For definiteness we consider the case of one variable defined on the interval $[a, b]$ and grouped in such a way that $x_1 = a + h_1/2$, $x_2 = a + h_1 + h_2/2$, $\cdots$, $x_w = b - h_w/2$, where h_i is not necessarily equal to h_j, $i \neq j$.

The likelihood function for a sample of N can be written as

$$L(\theta) = C \prod_{i=1}^{v} [p_i(\theta)]^{N_i}, \qquad \sum N_i = N,$$

where C is a constant which does not depend on θ and

$$p_i(\theta) = \int_{x_i - h_i/2}^{x_i + h_i/2} f(x, \theta)\, dx.$$

By similar methods to those used above the following equations for the δ_i may be constructed;

$$\begin{aligned} \sum N_i\, \partial^2 \ln f(x_i, \theta_0)\, \delta_1/\partial\theta_1^2 \quad &+ \cdots + \sum N_i\, \partial^2 \ln f(x_i, \theta_0)\, \delta_s/\partial\theta_s\, \partial\theta_1 \\ &\vdots \\ &= -\sum N_i\, \partial[f''(x_i, \theta_0)h_i^2/24f(x_i, \theta_0)]/\partial\theta_1 \\ \sum N_i\, \partial^2 \ln f(x_i, \theta_0)\, \delta_1/\partial\theta_s\, \partial\theta_1 &+ \cdots + \sum N_i\, \partial^2 \ln f(x_i, \theta_0)\, \delta_s/\partial\theta_s^2 \\ &= -\sum N_i\, \partial[f''(x_i, \theta_0)h_i^2/24f(x_i, \theta_0)]/\partial\theta_s\,. \end{aligned} \tag{4.7}$$

After dividing through by N, replacing N_i/N by $p_i \approx h_i f(x_i, \theta_0)$, equations (4.7) may be solved approximately in matrix form as follows;

$$\boldsymbol{\delta} = \mathrm{V}(\theta_0)\, \mathrm{d}(\theta_0) \tag{4.8}$$

where the transpose of $\mathrm{d}(\theta_0)$, $\mathrm{d}'(\theta_0)$, is defined by

$$\mathrm{d}' = [d_1, d_2, \cdots, d_v], \quad d_s = (24)^{-1} \sum_{i=1}^{v} f(x_i, \theta_0)\, \partial[f''(x_i, \theta_0)/f(x_i, \theta_0)]h_i^3/\partial\theta_s\,.$$

From (4.8) the appropriate $\boldsymbol{\delta}$ may be calculated for the particular grouping scheme provided, of course, none of the h_i are too large. The advantage of (4.8) is that if a whole series of samples are drawn from the same population with the grouping scheme fixed, $\boldsymbol{\delta}$ may be obtained once and for all.

In the above case the information matrix $\mathrm{I}(h)$ turns out to be

$$\mathrm{I}(h) = \mathrm{I} - \mathrm{L} \tag{4.9}$$

where I is the ungrouped information matrix and $\mathrm{L} = [l_{sr}]$ with

$$\begin{aligned} l_{sr} = \sum_{i=1}^{v} \{ & f(x_i, \theta)\, \partial^2[f''(x_i, \theta)/f(x_i, \theta)]/\partial\theta_r\, \partial\theta_s \\ & + f''(x_i, \theta)\, \partial^2 \ln f(x_i, \theta)/\partial\theta_r\, \partial\theta_s - D_x^2[f(x, \theta)\, \partial^2 \ln f(x, \theta)/\partial\theta_r\, \partial\theta_s]_{x=x_i}\}h_i^3/24. \end{aligned}$$

Acknowledgement

The author wishes to thank Professor J. B. Douglas of the University of New South Wales for his helpful comments in connection with this work. He is also grateful to the referees for their detailed comments and for suggestions leading to an improvement in the presentation of the paper.

References

1. Lindley, D. V., 1949. Grouping corrections and maximum likelihood equations. *Proc. Camb. Philos. Soc, 46*, 106–10.
2. Ralston, A., 1965. *A First Course in Numerical Analysis*. McGraw-Hill, New York.
3. Wold, H., 1934. Sheppard's correction formulae in several variables. *Skand. Aktuarietidskr, 17*, 248–55.

(b)

THE MODELLING OF HOST-PARASITE CYCLES WITH AN EMPHASIS ON NEMATODE PARASITES IN SHEEP

A (b)[1]

Reprinted from the
AUSTRALIAN JOURNAL OF BIOLOGICAL SCIENCES
VOLUME 17, NUMBER 2, PAGES 504-13, MAY 1964

MODELS FOR THE DISTRIBUTION ON PASTURE OF INFECTIVE LARVAE OF THE GASTROINTESTINAL NEMATODE PARASITES OF SHEEP

By G. M. Tallis and A. D. Donald

Reprinted for the
Commonwealth Scientific and Industrial Research Organization
Australia

MODELS FOR THE DISTRIBUTION ON PASTURE OF INFECTIVE LARVAE OF THE GASTROINTESTINAL NEMATODE PARASITES OF SHEEP

By G. M. Tallis* and A. D. Donald†

[*Manuscript received September* 13, 1963]

Summary

Two models are proposed for the distribution on pasture of infective larvae of the gastrointestinal nematode parasites of sheep. These models were developed to include as many as possible of the known biological components. Procedures for estimating the parameters of the models are outlined and advantages of these models over earlier attempts to describe the distribution of infective larvae on pasture are briefly discussed.

I. Introduction

In studies on the population dynamics of gastrointestinal nematode parasitism in sheep it is important to obtain some measure of the rate at which the infective larvae of these parasites are ingested by the grazing animal. It is likely that the rate of larval intake by the host depends largely on the grazing behaviour of the sheep, and the distribution and abundance of infective larvae on the pasture. In this paper, specific models are developed for the distribution of infective larvae on pasture incorporating as many of the known biological components as possible.

A brief description of the biological processes requiring mathematical treatment follows:

(1) Eggs which are laid by the female parasites in the alimentary tract of the host are passed out in the faeces onto the pasture. Under favourable environmental conditions, the eggs undergo several stages of development, culminating in the appearance of third-stage larvae which are infective for the host. The infective third-stage larvae migrate only a small distance away from the faecal deposit to adjacent herbage (Dinaburg 1944; Furman 1944), where they may survive for a limited period. The host becomes infected either by penetration of the infective larvae through its skin, or, for the great majority of these parasite species, by ingestion of the infective larvae with the herbage.

(2) The rate of development to the third larval stage and the rate of mortality, both during development and in the third larval stage, depend on microclimatic factors, principally temperature and humidity.

(3) Observations by Crofton (1954) on fields being grazed by sheep have shown that the distribution of faecal deposits is not random, i.e. is not described by a Poisson law. He has also shown that sheep, while grazing,

* Division of Mathematical Statistics, CSIRO, McMaster Laboratory, Glebe, N.S.W.

† Division of Animal Health, CSIRO, McMaster Laboratory, Glebe, N.S.W.

move in a more or less well-integrated group, which at any point in time occupies an area rarely less than one-sixth and never more than one-third of the total pasture area. Assuming that the periods of grazing and defaecation are broadly coincident, it is likely that the distribution of faeces deposited per unit area of the total pasture area within a unit of time will be "overdispersed" statistically, i.e. the variance of the distribution will be appreciably greater than the mean (Bliss and Fisher 1953).

(4) It has been shown by Hunter and Quenouille (1952) that replicate faecal worm egg counts (eggs per gram of faeces) from the same sheep followed a Poisson series, but that the distribution of egg counts between sheep fitted reasonably well to the negative binomial distribution with $k \simeq 0{\cdot}7$. It is likely, therefore, that the distribution of total egg numbers in faecal masses deposited by a flock of sheep in a unit of time will also be overdispersed.

In the following development these biological aspects are important, since we require mathematical models to describe the distribution of third-stage larvae on the pasture. Thus the distribution of faecal deposits on the pasture, the distribution of egg output of the flock at a given point in time, and the rate of mortality of the larval stages of the parasite must be considered during construction of the models. These points will be emphasized in the next section.

II. The Distribution

We consider the situation where a fixed number, S, of sheep are introduced onto a pasture of total area A at time $t = 0$ and are removed at $t = t_1$. The probability generating function (p.g.f.) for the distribution of faecal deposits for the ith sheep is assumed to be of the form

$$\{p/(1-qs)\}^{a_i t}, \tag{1}$$

where $q = 1-p$ and $i = 1, 2, \ldots, S$. The expression (1) specifies a negative binomial distribution with parameters p and $a_i t$. Now, if the effects of sheep on the total distribution of faecal deposits are stochastically additive and independent, then the p.g.f. for the S sheep may be written as

$$\{p/(1-qs)\}^{at}, \tag{2}$$

with

$$a = \sum_{i=1}^{S} a_i.$$

Now let the time segment $[0, t]$ be partitioned into n intervals of equal length tn^{-1} and label the points of subdivision $t_0, t_1, t_2, \ldots, t_n$. Then, because of the infinite divisibility property of (2), the p.g.f. of total faecal deposits for any of the subintervals is given by

$$\{p/(1-qs)\}^{at/n}. \tag{3}$$

Furthermore, consider the p.g.f. of the number, N, of eggs dropped onto the pasture during the ith time interval $[t_{i-1}, t_i]$. We denote the average number of eggs per faecal deposit at some time $\bar{t}_i \epsilon [t_{i-1}, t_i]$ by $\lambda(\bar{t}_i)$ and assume that the distribution of the numbers of such eggs follows a Poisson law with parameter $\lambda(\bar{t}_i)$, which is a continuous function of time. (Initially, we assume, contrary to (4) of the Introduction, that $\lambda(\bar{t}_i)$ is the same for all sheep. Subsequently, this restriction will be removed.) Moreover, if the probability that a given egg is in the infective larval stage at time t after being dropped is denoted by the continuous function $f(t)$, then it is easily shown that at time t, the distribution of larvae developing from eggs deposited during the ith subinterval is again Poisson with parameter

$$\mu(t, \bar{t}_i) = \lambda(\bar{t}_i)\, f(t-\bar{t}_i).$$

Thus the p.g.f. for the number of larvae surviving at time t from the ith interval is given approximately by

$$\left\{p/[1-q.\exp\{-\mu(t, \bar{t}_i)(1-s)\}]\right\}^{at/n}. \tag{4}$$

Therefore, for the p.g.f. of the total number of larvae on the pasture at time t we have also approximately,

$$g_n(t, s) = \prod_{i=1}^{n} \left\{p/[1-q.\exp\{-\mu(t, \bar{t}_i)(1-s)\}]\right\}^{at/n} \tag{5}$$

The larger n becomes, the smaller is the interval width tn^{-1} and the closer does $g_n(t, s)$ approach to the conceptual p.g.f. with continuous time increments.

We therefore define the limiting p.g.f. as follows:

$$\begin{aligned} g(t, s) &= \lim_{n\to\infty} g_n(t, s) \\ &= \lim_{n\to\infty} \exp\left\{atn^{-1} \sum_{i=1}^{n} \ln\left[p/(1-q.\exp\{-\mu(t, \bar{t}_i)(1-s)\})\right]\right\} \\ &= \exp\left\{at\ln p - a\int_0^t \ln\left[1-q.\exp\{-\mu(t, x)(1-s)\}\right]dx\right\} \end{aligned} \tag{6a}$$

by the continuity of e^y and the definition of the Riemann integral.

In the above derivation it was implicitly assumed that $t \leqslant t_1$. However, if $t > t_1$ we consider the interval $[0, t_1]$ and apply entirely analogous reasoning to that used above. Thus, for $t > t_1$

$$g(t_1, s) = \exp\left\{at_1\ln p - a\int_0^{t_1} \ln\left[1-q.\exp\{-\mu(t, x)(1-s)\}\right]dx\right\}. \tag{6b}$$

Since $\lambda(t)$ and $f(t)$ are continuous functions of t, the integral $\int_0^t [\lambda(x)f(t-x)]^a \mathrm{d}x$, $a \geqslant 0$, exists, and if the expressions (6a) and (6b) are suitably differentiated with respect to s, it is found that

$$\begin{aligned} E_S(N) &= aqp^{-1}\int_0^t \lambda(x)\, f(t-x)\, \mathrm{d}x \qquad t \leqslant t_1 \\ &= aqp^{-1}\int_0^{t_1} \lambda(x)\, f(t-x)\, \mathrm{d}x \qquad t > t_1 \\ V_S(N) &= aqp^{-2}\int_0^t [\lambda(x)\, f(t-x)]^2 \mathrm{d}x + E_S(N) \qquad t \leqslant t_1 \\ &= aqp^{-2}\int_0^{t_1} [\lambda(x)\, f(t-x)]^2 \mathrm{d}x + E_S(N) \qquad t > t_1 \end{aligned} \tag{7}$$

where the subscript S is used to emphasize that these values refer to the application of S sheep to the area.

It is important to notice that (6a) and (6b) are not only sheep additive, but space additive as well. Although these formulae refer to a particular area of size A, they may be considered as the convolution of the effects of S sheep on L areas each of size A/L. In this case, each subplot has a negative binomial distribution with parameters p and at/L. These features emphasize the flexibility of the model.

It is clear from (7) that if the sheep are left on pasture indefinitely, then the following theoretical equilibrium is reached at $t = \infty$,

$$e(N) = aqp^{-1} \lim_{t\to\infty} \int_0^t \lambda(x)\, f(t-x)\mathrm{d}x.$$

In order to obtain bounds for $e(N)$, notice that

$$aqp^{-1}\inf_x[\lambda(x)]\int_0^\infty f(x)\mathrm{d}x \leqslant e(N) \leqslant aqp^{-1}\sup_x[\lambda(x)]\int_0^\infty f(x)\mathrm{d}x,$$

and therefore

$$e(N) = \lambda(\tilde{x})\int_0^\infty f(x)\mathrm{d}x\ ,\ \tilde{x}\epsilon[0,\infty].$$

However, when the sheep are removed at $t = t_1$, we obtain

$$\begin{aligned} e(N) &= aqp^{-1} \lim_{t\to\infty} \int_0^{t_1} \lambda(x)\, f(t-x)\mathrm{d}x \\ &= aqp^{-1}\int_0^{t_1} \lambda(x)[\lim_{t\to\infty} f(t-x)]\mathrm{d}x \\ &= 0, \end{aligned}$$

as required, since $f(\infty) = 0$.

If, in the above model, sheep additivity does not seem to be a valid assumption, then the model can be applied to varying flock sizes and different a values estimated for each flock size. This, in fact, would allow the null hypothesis of sheep additivity to be investigated, since under this hypothesis the a's should be proportional to the flock sizes.

However, as mentioned in the Introduction, it may not be safe to assume a constant $\lambda(t)$ for each sheep as was implicitly done in the above derivations. If this is so, and sheep additivity can be assumed, then each sheep must be given its own $g(t, s)$ and the parameters a_j and $\lambda_j(t)$, $j = 1, 2, \ldots, S$, must be separately estimated and the average number of larvae on the pasture would then be given by

$$E_S(N) = qp^{-1} \sum_{j=1}^{S} a_j \int_0^t \lambda_j(x) f(t-x)\mathrm{d}x. \tag{8}$$

Some investigation would be necessary in order to show whether or not expressions (7) are satisfactory approximations.

If the investigator is seriously concerned about both the assumptions of sheep additivity and constant $\lambda(t)$, then the whole model may have to be changed. One way of doing this is to assume that the numbers of larvae at time t which develop from eggs deposited on the pasture during the ith time interval has p.g.f.

$$\{u/1-v[1-f(t-\bar{t}_i)(1-s)]\}^{k(\bar{t}_i)}, \tag{9}$$

where $k(t)$ is an arbitrary function of time and $f(t-x)$ is as defined earlier. The p.g.f. (9) is the result of compounding a negative binomial distribution [parameters u and $k(\bar{t}_i)$] with a binomial distribution [parameter $f(t-\bar{t}_i)$]. Analogous reasoning to that used earlier shows that in this case, assuming the same distribution of faecal deposits,

$$g(t, s) = \exp\left\{at \ln p - a\int_0^t \ln\left[1-q\{u/(1-v[1-f(t-x)(1-s)])\}^{k(x)}\right]\mathrm{d}x\right\}, \tag{10}$$

which is an extremely complicated distribution. However, the mean turns out to be

$$E_S(N) = atqp^{-1}\, v\, u^{-1} \int_0^t k(x) f(t-x)\mathrm{d}x, \tag{11}$$

which is of the same form as (7). The increased generality is achieved by the introduction of the additional parameter u. Of course, the same type of modification for $t>t_i$ applies to (10) as for (6a).

It is interesting to notice that (11) can be written down directly from other considerations. If $f(s)$, $g(s)$, and $k(s)$ are p.g.f.'s of random variables X, Y, Z, then the mean of the compound variate specified by $f(g(k(s)))$ is simply $E(X)\cdot E(Y)\cdot E(Z)$. Thus for any time subinterval i, the mean number of third-stage larvae is given by

$$atn^{-1}qp^{-1}vu^{-1}k(\bar{t}_i)f(t-\bar{t}_i),$$

and summing this expression over all intervals, the contribution from intervals being stochastically independent, and letting $n\to\infty$, gives the result (11). Uniqueness is guaranteed by the Continuity Theorem (Feller 1960, p. 262).

We now turn to problems of estimation and consider model (6a) in detail. Suitable procedures for the other models can be worked out in a similar way.

III. Estimation

The quantity which is of major practical importance is the concentration of larvae per unit of area. In order to estimate this in an efficient manner from a given area A and a given number of sheep S, we subdivide A into L subunits of equal size a so that $La = A$. Moreover, the time interval $[0, t]$ is also subdivided into T equal intervals of length t/T. The p.g.f. for faecal deposits corresponding to any subinterval and for areas of size a, is given by

$$\{p/(1-qs)\}^{at/TL}.$$

Now S sheep are introduced at $t = 0$ and the distribution of faecal deposits in the L subareas recorded for each time interval. If r_{ij} is the number observed in the jth plot for time interval i, then, if we let $at = \beta$, $k' = \beta/LT$, and $m' = k'\gamma$, where $\gamma = (1-p)/p$, moment estimates of k' and m' are given by

$$\hat{m}_i' = \bar{r}_{i.},$$

$$\hat{k}_i' = \bar{r}_{i.}^2(S_{i.}^2 - \bar{r}_{i.})^{-1},$$

where

$$\bar{r}_{i.} = \sum_{j=1}^{L} r_{ij}/L,$$

and

$$S_i^2 = \sum_{j=1}^{L} (r_{ij} - \bar{r}_{i.})^2/(L-1).$$

From Anscombe (1950) the variances of $\hat{m}_i'$ and $\hat{k}_i'$ are given by

$$\begin{aligned} V(\hat{m}_i') &= m'(1+\gamma)/L \\ &= \beta\gamma(1+\gamma)/L^2T, \\ V(\hat{k}_i') &= 2k'(k'+1)(1+\gamma)^2/\gamma^2L \\ &= 2\beta(\beta+LT)(1+\gamma)^2/\gamma^2L^3T^2, \end{aligned}$$

and it can be shown that $C(\hat{m}_i', \hat{k}_i') = 0$.

Since we obtain T estimates of m' and k', one set for each subinterval, and since $V(\hat{\theta}_i) = V(\hat{\theta}_j)$, all i, j, and $\theta = m', k'$, it follows that the best linear estimates for the combined data are given by

$$\bar{m}' = \sum_{i=1}^{T} \hat{m}_i'/T, \quad \bar{k}' = \sum_{i=1}^{T} \hat{k}_i'/T,$$

with variances

$$\begin{aligned} V(\bar{m}') &= \beta\gamma(1+\gamma)/L^2T^2, \\ V(\bar{k}') &= 2\beta(\beta+LT)(1+\gamma)^2/\gamma^2L^3T^3. \end{aligned}$$

However, since we are interested in the distribution of faecal deposits at time t and not for a subinterval, the parameters of interest are $k = Tk' = \beta/L$, and $m = Tm' = \beta\gamma/L$. Estimates of these are given by $\bar{m} = T\bar{m}'$ and $\bar{k} = T\bar{k}'$ and the variances are

$$V(\bar{m}) = \beta\gamma(1+\gamma)/L^2,$$

$$V(\bar{k}) = 2\beta(\beta+LT)(1+\gamma)^2/\gamma^2L^3T.$$

As stated earlier, it is the concentration of larvae which is of interest and we consider the quantity $C = E(N)L/A$ (remembering that $E(N)$ is now referred to areas of size $a = A/L$)which is estimated by

$$\hat{C}_S = L\widehat{E_S(N)}/A = L\bar{m}A^{-1}\int_0^t \widehat{\mu(t,x)}\mathrm{d}x, \tag{12}$$

where

$$\int_0^t \widehat{\mu(t,x)}\mathrm{d}x = \widehat{\int_0^t}\cdot$$

is an estimate of the required convolution integral. We now find the variance of $\hat{C}_S$ as a function of L and T and determine for which values of these two parameters it is minimized. Straight forward calculations show that

$$V(\hat{C}_S) = (L^2/A^2)\left\{V(\bar{m})\left(\int_0^t \cdot\right)^2 + m^2V\left(\widehat{\int_0^t}\cdot\right)\right\} \tag{13}$$

$$= (\beta\gamma/A^2)\left\{(1+\gamma)\left(\int_0^t \cdot\right)^2 + \beta\gamma V\left(\widehat{\int_0^t}\cdot\right)\right\}.$$

It is shown below that for one method of estimating $\int_0^t \mu(t,x)\mathrm{d}x$, $V\left(\widehat{\int_0^t}\cdot\right) = O(T^{-1})$ and therefore the conclusion is that, in this case, $V(\hat{C}_S)$ is independent of L and decreases with increasing A and T. However, the estimate of k increases in precision with an increase in both L and T, while the variance of $\bar{m}$ only depends on L.

It is, of course, possible to estimate the parameters k and m more efficiently by maximum likelihood methods (see Anscombe, loc. cit.). In this case an expansion for $V(\hat{C}_S)$ is easily obtained in terms of estimated variances of $V(\hat{k})$ and $V(\hat{m})$ of the maximum likelihood estimates. However, the additional rather heavy computational work necessitated by the maximum likelihood procedure does not really seem warranted.

There remains the question of the estimation of $\int_0^t \mu(t,x)\mathrm{d}x$. There are numerous ways in which successive values of $\mu(t, x)$ can be estimated to provide ordinates for numerical integration. For ease of illustration, we consider just one direct

approach and we concern ourselves with the case $t \leqslant t_1$. The analysis for other situations would be analogous.

Assume now that $[0, t]$ has been subdivided, as described above, into T subintervals and let t_i be the upper boundary for the ith interval. Then during each interval i $(i = 1, 2, \ldots, T)$ F fresh faecal deposits are marked, and at time t the average number of larvae emanating from these deposits, in each of the T groups, is determined. If n_{ij} represents the number of larvae in the jth faecal deposit from the ith subinterval which survive to time t, then $\mu(t, t_i)$ is estimated by

$$\sum_{j=1}^{F} n_{ij}/F = \hat{\mu}(t, t_i).$$

Suppose now, that in order to estimate $\int_0^t \mu(t, x)\, dx$, we use the trapezoidal rule for numerical integration, then

$$\int_0^t \widehat{\mu(t,x)}dx = \frac{t}{2T} \sum_{i=1}^{T} \left\{\hat{\mu}(t, t_i) + \hat{\mu}(t, t_{i-1})\right\}$$

and, neglecting errors of integration,

$$V\left(\widehat{\int_0^t \cdot}\right) = \frac{t^2}{4T^2F} \sum_{i=1}^{T} \left\{\mu(t, t_i) + \mu(t, t_{i-1})\right\} \tag{14a}$$

$$\simeq \frac{t}{2TF} \int_0^t \mu(t, x)dx.$$

For the case of $t > t_1$, it is the interval $[0, t_1]$ which is subdivided and measurements of larvae numbers are made at time t. Thus, for this case

$$V\left(\widehat{\int_0^{t_1} \cdot}\right) \simeq \frac{t_1}{2TF} \int_0^{t_1} \mu(t, x)dx. \tag{14b}$$

If, finally, it is desired to bring the discussion down to a sheep per unit area basis, then since $a = S\bar{a}$, for an average sheep the expected concentration is $C = C_S/S$. Obviously, $\hat{C} = \hat{C}_S/S$ and $V(\hat{C}) = V(\hat{C}_S)/S^2$.

No detailed discussion of estimation procedures for model (10) will be presented here. Obviously, the faecal component can be estimated as for (6a) and the remaining expression,

$$vu^{-1} \int_0^t k(x)\, f(t-x)dx, \tag{15}$$

approximated in various ways. For instance, $f(t)$ can be obtained by a separate investigation, while $vu^{-1}\, k(t)$ can be calculated by establishing the egg output of

individual sheep in the flock at different points on the time scale. If T different time intervals are used, then the $(T+1)$ parameters u and $k(t_i)$, $i = 1, 2, \ldots, T$, can be estimated by maximum likelihood and the final expression obtained by numerical integration. Alternatively, the number of third-stage larvae, at time t, associated with faecal deposits dropped during previous time intervals can be obtained and the whole expression (15) approximated by numerical integration as for (6a). However, attention would have to be given to the variances of these estimates since they would not be of the same form as (14a) and (14b).

IV. Discussion

The distribution on pasture of the infective larvae of the gastrointestinal nematode parasites of sheep has been considered by Crofton (1952, 1954). He sampled the most evenly grazed portion of three pastures and showed that the observed frequencies of infective larval numbers agreed fairly closely with theoretical frequencies calculated according to Neyman's Contagious Distribution Type A (Neyman 1939). It is intrinsic in this distribution that the clumps of organisms are Poisson-distributed. Since Crofton (1954) has shown that this is unlikely to be true for the distribution of faecal deposits over a field being grazed by sheep, he has pointed out that this limits the usefulness of the Neyman model to small areas of pasture only.

Donald (unpublished data) has fitted the negative binomial to the distribution of infective larval numbers recovered from 50 4-in. quadrat samples of pasture collected from a ¼-acre field being grazed by five sheep, and has found $k \simeq 0{\cdot}2$. While this is consistent with a contagious distribution, several quite different hypothetical situations will give rise to a negative binomial distribution (Anscombe, loc. cit.) Thus, obvious difficulties of interpretation arise when attempts are made to compare the distribution of infective larvae of different species and to follow movements of the distributions with time.

The main purpose of this paper is to show how to construct models describing the distribution on pasture of the infective stages of parasites of grazing animals. Of the two models developed here, (10) is slightly more general since it incorporates component (4) of the Introduction. However, this increased generality introduces an extra parameter u, and the problems of estimation are increased. The simple properties of the Poisson distribution are lost and the rather natural interpretation of $E_S(N)$ is somewhat destroyed.

However, for most purposes (7) should provide a sufficiently accurate description of the distribution of the larvae on pasture. Once the faecal component has been estimated, a and p, different theoretical curves for $\lambda(t)$ and $f(t)$ can be used in (7) in order to investigate the effects such changes would have on infective larval populations on pasture. This would provide information, say, on the comparative behaviour of two different species of parasite or on the behaviour of a single species under different environmental conditions. Furthermore, the effect of each component of the model (faecal distribution, egg numbers per faecal deposit, and the mortality rate of the free-living larval stages) can be isolated and its ultimate influence on infective larval populations determined.

The introduction of a time element into the models seems advantageous. The influence of time on the total distribution of infective larvae is now clearly specified, and this enables theoretical questions, such as equilibrium values, to be settled. This was not possible in earlier studies when less specific models were fitted to estimates of infective larval populations on pasture.

V. References

Anscombe, F. J. (1950).—Sampling theory of the negative binomial and logarithmic series distributions. *Biometrika* **37**: 358–82.

Bliss, C. I., and Fisher, R. A. (1953).—Fitting the negative binomial distribution to biological data. *Biometrics* **9**: 176–200.

Crofton, H. D. (1952).—The ecology of immature phases of trichostrongyle nematodes. IV. Larval populations on lowland pastures. *Parasitology* **42**: 77–84.

Crofton, H. D. (1954).—The ecology of the immature phases of trichostrongyle parasites. V. The estimation of pasture infestation. *Parasitology* **44**: 313–24.

Dinaburg, A. G. (1944).—The survival of the infective larvae of the common ruminant stomach worm, *Haemonchus contortus*, on outdoor grass plots. *Amer. J. Vet. Res.* **5**: 32–7.

Feller, W. (1960).—"An Introduction to Probability Theory and its Applications." 2nd Ed. Vol. 1. (John Wiley and Sons: New York.)

Furman, D. P. (1944).—Effects of environment upon the free-living stages of *Ostertagia circumcincta* (Stadelmann). Trichostrongylidae. I. Laboratory experiments. *Amer. J. Vet. Res.* **5**: 79–86.

Hunter, G. C., and Quenouille, M. H. (1952).—A statistical examination of the worm egg count sampling technique for sheep. *J. Helminth.* **26**: 157–70.

Neyman, J. (1939).—On a new class of "contagious" distributions, applicable in entomology and bacteriology. *Ann. Math. Statist.* **10**: 35–57.

A (b) [2]

Short Communication reprinted from the Australian Journal of Biological Sciences, Volume 17, Number 4, pp. 1016-19, November 1964

A NOTE ON THE ESTIMATION OF LARVAL CONCENTRATIONS ON PASTURE

By G. M. Tallis

Reprinted for the
Commonwealth Scientific and Industrial Research Organization
Australia

A NOTE ON THE ESTIMATION OF LARVAL CONCENTRATIONS ON PASTURE*

By G. M. TALLIS†

Introduction

In a recent paper by Tallis and Donald (1964)‡ [which will be referred to subsequently as (T.D.)] models were developed to describe the distribution on pasture of the infective larvae of gastrointestinal nematode parasites of sheep. It was pointed out that the distribution of faecal deposits, the number of eggs per deposit, and the developmental rate to the infective larval stage were three important components determining the total larval distribution. Moreover, formulae for the expected concentration of larvae on the pasture lead to straightforward methods of estimation which are free from large biases.

To date, other pasture-sampling methods have been employed. In particular, small areas of pasture are often determined in some random fashion, clipped, and the numbers of larvae in each clipping estimated separately or from a combined sample. However, we are not concerned with the post-clipping procedure here. Instead, the intention is to investigate the validity of the method of collecting small representative samples of the pasture on which larval counts are to be made.

The notation will conform with that of (T.D.). Moreover, some of the derivations and assumptions will follow closely those leading to equations (6a) and (7) in the above paper and in all instances t will be less than t_1, where t is the time that the flock has been on pasture and t_1 is the time of their removal. Finally, since it will be quite sufficient to consider the simplest model (6a) in order to demonstrate the main points of this note, the reader is referred to Section I and the first few paragraphs of Section II of (T.D.) for a fuller discussion of the methods used below.

Methods

To be specific, we consider a rectangular field, R_1, of area $A = a \times b$ and concern ourselves with a small sampling plot, R_2, within R_1. It turns out subsequently that it is convenient to have the plot circular with radius r, although the shape is not important until we look for specific results. Further, as in (T.D.), we consider a fixed time interval $[0, t]$ which is subdivided into n intervals of equal length t/n, the ith subinterval being $[t_{i-1}, t_i]$. In addition we also consider a small rectangular subsection of the field of area $\Delta u \Delta v$.

As pointed out in (T.D.), the negative binomial distribution used in connection with this problem can be assumed to be time and space additive. Therefore, the

* Manuscript received June 22, 1964.

† Division of Mathematical Statistics, CSIRO, McMaster Laboratory, Glebe, N.S.W.

‡ TALLIS, G. M., and DONALD, A. D. (1964).—Models for the distribution on pasture of infective larvae of the gastrointestinal nematode parasites of sheep. *Aust. J. Biol. Sci.* **17**: 504–13.

distribution of faecal deposits in the small rectangle during the ith time interval is specified by the probability generating function, p.g.f.,

$$\{p/(1-qs)\}^{\alpha\Delta t\Delta v\Delta u/A}, \tag{1}$$

where α and p are parameters related to the size of the flock on R_1 and $\Delta t = t/n$. Moreover, we denote the average number of eggs per faecal deposit at some time $\bar{t}_i \epsilon [t_{i-1}, t_i]$ by $\lambda(\bar{t}_i)$ and assume that the distribution of the numbers of such eggs follows a Poisson law with parameter $\lambda(\bar{t}_i)$, which is a continuous function of time. If the probability that a given egg is in the infective larval stage at time t after being dropped is denoted by the continuous function $f(t)$, then the probability that a given egg is in the larval stage and is in R_2 is $f(t-\bar{t}_i)p(\bar{u}, \bar{v}, \bar{t}_i, t)$, where the point $(\bar{u}, \bar{v})$ is in the rectangle of area $\Delta u \Delta v$. The quantity $p(u, v, \bar{t}_i, t)$ is the probability that a larva which developed from an egg in the ith time interval will wander away from the point (u, v) into R_2. It is now easily shown that the distribution of larvae developing from eggs deposited during the ith time interval and which migrate to R_2 is again Poisson with parameter

$$\mu(\bar{u}, \bar{v}, \bar{t}_i, t) = \lambda(\bar{t}_i) f(t-\bar{t}_i) p(\bar{u}, \bar{v}, \bar{t}_i, t).$$

By an argument entirely analogous to that used in (T.D.) to derive (6a), it is found that the p.g.f. for larvae in R_2 at time t is given by

$$g(t, s) = \exp\left\{\alpha t \ln p - \frac{\alpha}{A} \int_0^t \int_0^a \int_0^b \ln[1-q \exp\{-\mu(u, v, z, t)(1-s)\}] \, dv \, du \, dz\right\}, \tag{2}$$

and the mean of this distribution is

$$E(N) = \alpha q (pA)^{-1} \int_0^t \int_0^a \int_0^b \lambda(z) f(t-z) p(u, v, z, t) \, dv \, du \, dz. \tag{3}$$

For $p(u, v, z, t) \equiv 1$, this expression reduces to (7) of (T.D.) and this corresponds to the case where $R_1 = R_2$.

If this method of sampling the field is to produce results which really estimate the concentration of larvae on the pasture, we must have

$$E(N) = BA^{-1} E_s(N),$$

where B is the area of R_2 and

$$E_s(N) = \alpha q p^{-1} \int_0^t \lambda(z) f(t-z) \, dz.$$

Thus

$$\int_0^t \int_0^a \int_0^b \lambda(z) f(t-z) p(u, v, z, t) \, dv \, du \, dz = B \int_0^t \lambda(z) f(t-z) \, dz \tag{4}$$

is the required condition.

In order to obtain some specific results we assume that R_2 is circular with radius r and centre $(0, 0)$ and that $p(u\ v, z, t)$ is specified by

$$p(u, v, z, t) = [2\pi\sigma^2(z, t)]^{-1} \iint_{x^2+y^2 \leq r^2} \exp\{-[2\sigma^2(z, t)]^{-1}[(x-u)^2+(y-v)^2]\} \, dx \, dy \tag{5}$$

where $\sigma^2(z,t)$ is a scaling parameter which can be expressed as a continuous function of z and t. Clearly the smaller σ^2 is, the smaller is the expected migration distance of larvae from a faecal deposit at (u,v). We now evaluate the required integrals of (4).

From (5) it can be seen that $w = (x^2+y^2)/\sigma^2$ may be regarded as the sum of two variables squared, which are each normally distributed with the means u/σ and v/σ respectively and both with unit variance. Therefore, w has a non-central χ^2 distribution with parameters 2 and $\gamma = (2\sigma^2)^{-1}(u^2+v^2)$. Hence,

$$p(u,v,z,t) = \sum_{i=0}^{\infty} F_{2(i+1)}(r^2/\sigma^2)\,\mathrm{e}^{-\gamma}\gamma^i/i!, \tag{6}$$

where $F_{2(i+1)}(.)$ is the distribution function of a χ^2 variable with $2(i+1)$ degrees of freedom.

It is now convenient to introduce another circular area $R_3 \subset R_1$ of radius θ, also centred at $(0,0)$. The radius θ is considerably greater than r and for the present it is taken to be arbitrary. We now wish to integrate $p(u,v,z,t)$ over R_3 and hence we have

$$\iint_{u^2+v^2\leq\theta^2} p(u,v,z,t)\,\mathrm{d}u\,\mathrm{d}v = q(t,z),$$

and a typical term of this integration (after a valid interchange of $\int \Sigma$ to $\Sigma \int$) is

$$(2^i i!\,\sigma^{2i})^{-1}F_{2(i+1)}(r^2/\sigma^2) \iint_{u^2+v^2\leq\theta^2} \mathrm{e}^{-(1/2\sigma^2)(u^2+v^2)}(u^2+v^2)^i\,\mathrm{d}u\,\mathrm{d}v.$$

Let $u = \rho\cos\phi$ and $v = \rho\sin\phi$, then we obtain

$$(2^i i!\,\sigma^{2i})^{-1}F_{2(i+1)}(r^2/\sigma^2) \iint_{\substack{0\leq\rho\leq\theta \\ 0\leq\phi\leq 2\pi}} \mathrm{e}^{-\rho^2/2\sigma^2}\rho^{2i+1}\,\mathrm{d}\rho\,\mathrm{d}\phi,$$

and another change of variable ($\eta = \rho^2/\sigma^2$) reduces this integral to

$$2\pi\sigma^2[2^{i+1}\Gamma(i+1)]^{-1}F_{2(i+1)}(r^2/\sigma^2)\int_0^{\theta^2/\sigma^2} \mathrm{e}^{-\frac{1}{2}\eta}\eta^i\,\mathrm{d}\eta = 2\pi\sigma^2 F_{2(i+1)}(r^2/\sigma^2)F_{2(i+1)}(\theta^2/\sigma^2),$$

and we obtain finally

$$q(t,z) = 2\pi\sigma^2 \sum_{i=0}^{\infty} F_{2(i+1)}(r^2/\sigma^2)F_{2(i+1)}(\theta^2/\sigma^2). \tag{7}$$

The radius θ may be made large as compared with σ which means that any faecal deposit lying outside R_3 has negligible influence on the number of larvae found in R_2. Under these conditions

$$q(t,z) \to 2\pi\sigma^2 \sum_{i=0}^{\infty} F_{2(i+1)}(r^2/\sigma^2) \tag{8}$$

since $F_{2(i+1)}(\theta^2/\sigma^2)$ tends to unity for all values of $F_{2(i+1)}(r^2/\sigma^2)$ which can contribute

significantly to the sum. By interchanging the order of integration and summation in (8) we find

$$q(t,z) \simeq 2\pi\sigma^2 \int_0^{r^2/\sigma^2} \frac{1}{2}\left\{\sum_{i=0}^{\infty} e^{-\frac{1}{2}t}(\tfrac{1}{2}t)^i/i!\right\} dt$$

$$= 2\pi\sigma^2 \times r^2/2\sigma^2 = \pi r^2, \tag{9}$$

which is independent of t and z.

We have therefore shown that, with the particular models chosen,

$$\int_0^t \int_0^a \int_0^b \lambda(z)\, f(t-z) p(u,v,z,t)\, du\, dv\, dz \simeq \pi r^2 \int_0^t \lambda(z)\, f(t-z)\, dz$$

which is the required form of (4) with $B = \pi r^2$. Hence, provided the models used here conform reasonably well with the situation in practice, the conclusion is that the clipping of numerous small areas of pasture should allow valid estimates of the concentration of larvae on the pasture to be made. This is certainly a surprise to the author who felt that this method of estimation of larval concentration would produce biased results, the bias being in some way related to σ and r.

In conclusion, therefore, it seems to be appropriate to select as many sample plots as possible, k say, and count the number of larvae on each plot after the sheep have been on the pasture for a time period of length t, say. Let n_i be the count of the ith plot, then a suitable estimator for the concentration of the larvae on pasture is $\hat{C} = \bar{n}/r^2\pi$, if the plots are circular with radius r and where $\bar{n} = \sum_{i=1}^{k} n_i/k$. Certainly

$$E(\hat{C}) = C = E(N)/\pi r^2$$

and

$$V(\hat{C}) \simeq \sum_{i=1}^{k} (n_i - \bar{n})^2/r^4\pi^2(k-1),$$

which can be made satisfactorily small by either increasing r or k, or both. By an appeal to the Central Limit Theorem, appropriate large-sample confidence limit procedures can be applied.

The above analysis holds provided the sample plots are sufficiently far apart to eliminate correlation between the n_i. Thus the shortest distance between two plots should be at least $4 \times$ [maximum value of $\sigma(z,t)$].

Note
Part of the argument above is greatly simplified by the use of finite sampling theory. But the work shows how a diffusion-migration process can be brought into the P.G.F. structure to complete the model. More general densities for r can be used e.g. $Ke^{-r^d/2\sigma^2}$ $d>0$, and quantities as the average area contaminated calculated.

A (b)[3]

J. Theoret. Biol. (1966) 13, 251–260

A Stochastic Approach to the Study of Parasite Populations

G. M. Tallis and Morley Leyton

Johns Hopkins University,
Baltimore, Maryland, U.S.A.

(*Received* 10 *June* 1966)

The aim of this paper is to develop a general framework for building stochastic models to describe some features of parasite populations. Starting from the few basic biological assumptions outlined below, it is shown that input mechanisms, whereby the parasite gains entrance to the host, can be defined in probabilistic terms. Once in the host, the female parasite is allowed to "mature" and produce offspring according to given probability laws. Moreover, the host is supposed to react to control the rate of maturation and/or the rate of production of offspring. The overall accumulation of parasites, male and female, in the host is also considered.

In the sequel, a host is defined to be any organism which is subjected to a burden or infection of "lesser" organisms which we will call parasites. The biological assumptions concerning the relationships between host and parasite are outlined below.

(i) The parasite gains entrance to the host, either orally, intradermally or otherwise. Entry may be as a continuous stream or in the form of administered doses.

(ii) Once in the host the female enters a period of maturation at the completion of which she is capable of producing offspring.

(iii) Each parasite in the host has an "antigenic information trajectory". This term is used to describe the phenomenon that at any fixed time, the parasite is releasing information to the host to the effect that he, the parasite, is there. It is further assumed that antigenic information is additive, in the sense that the information emitted by a number of parasites is the sum of the individual antigenic informations. The offspring are assumed to produce no relevant antigenic information.

(iv) Each host responds in his own way to the build up of antigenic information. He responds by, in some way, controlling or otherwise affecting the rate of maturation of the parasites and/or the rate of reproduction of the females.

A deliberate attempt has been made to keep the treatment as general as possible. Unfortunately, this may lead to some obscurities and, in order to demonstrate the ideas, examples will be discussed in conjunction with the general development.

In one case we consider sheep as the host, worms as the parasite and eggs of the female worm as the offspring. The sheep are given a massive

dose of larvae (immature worms) at zero time and the task is to describe the total egg output of the worm population in the sheep as a function of time.

As further examples we examine some highly simplified continuous models. The purpose of these exercises is purely illustrative and the components of the models have been chosen for mathematical convenience. Nevertheless, it is hoped that the resulting models are not too unrealistic.

Details leading to the biological assumptions used in this paper may be found in Dineen (1963*a*, *b*), Donald, Dineen, Turner & Wagland (1964) and Dineen, Donald, Wagland & Offner (1965), Dineen, Donald, Wagland & Turner (1965).

1. Introduction

The benefits accruing to the parasitologist from the type of modeling suggested here may be somewhat intrinsic. It is clear that before detailed statistical work can commence, the biological hypotheses must be crystalized and this in itself is of some merit.

In addition, by the very nature of the models, parts of the general structure can be examined independently of the rest and the adequacy of the assumptions assessed and modified. Thus, the overall model can be altered as experience accumulates until confidence in the whole mathematical formulation is established.

Once the model is accepted as being "reasonable", it may be possible to infer many interesting biological results algebraically. For instance, in some cases the influence of time on the process can be examined as well as the effects of changing some of the meaningful parameters related to reproduction and survival. In fact, in some instances, from measurements made external to the host, inferences with regard to parasite numbers inside the host car be made.

Obviously, the full potential of model building in parasitology has not yet been realized. However, it is clear that a sensible mixing of mathematical and biological concepts may lead to sensible and even useful results.

2. The Conditional Models

(A) CONTINUOUS INPUT MODELS

We consider the process of parasites gaining entrance to a host and we will concentrate, for the present, exclusively on the female parasites. Further, we observe the host at time t and we are interested in the number of progeny being produced by the parasites in the host at this time.

Once having gained entrance to the host, the parasites are considered to undergo a maturation process, whereby the female becomes capable of

reproduction. To describe this process probabilistically we let $\alpha(y,x)$ be the conditional frequency function of maturation time given that entrance occurred at time x, then the conditional probability that a female mature in the interval $[y, y+\Delta y]$ is approximately $\alpha(y, x)\Delta y$. We further assume that a parasite which matures at time y, produces offspring according to the probability generating function, p.g.f., $h(s, t, y)$ at time $t \geq y$,

Suppose now that the probability that an immature female parasite gains an entrance to the host during the time interval $[x, x+\Delta x]$ is approximately $\lambda(x)\Delta x$, independently of the number of parasites already present in the host, then we are led to the following postulate.

Postulate 1

The probability that a female parasite enters a particular host during the time interval $[x+\Delta x]$ and matures in the interval $[y, y+\Delta y]$, $x \leq y$, is $\lambda(x)\alpha(y,x)\Delta x\Delta y + o(\Delta x\Delta y)$, where $\lim_{\Delta x, \Delta y \to 0} o(\Delta x\Delta y)/\Delta x\Delta y = 0$. This probability is independent of the number of parasites already in the host.

From the above assumptions we can now write down the p.g.f.'s for the two variables $F(t)$ and $P(t)$, the number of mature females in the host at time t and the number of progeny being produced by these parasites. Thus the required p.g.f.'s are

$$f(s,t) = \exp\left\{\int_0^t\int_0^y \lambda(x)\alpha(y,x)\,dx\,dy\,(s-1)\right\} \tag{1}$$

and

$$p(s,t) = \exp\left\{\int_0^t\int_0^y \lambda(x)\alpha(y,x)[h(s,t,y)-1]\,dx\,dy\right\}, \tag{2}$$

respectively. The p.g.f. (1) specifies a Poisson distribution with parameter $\int_0^t\int_0^y \lambda(x)\alpha(y,x)\,dx\,dy$. It is also clear that the distribution of the total number of female parasites, mature and immature, is also Poisson with parameter $\int_0^t \lambda(x)\,dx$.

From (2) we find immediately that

$$E\{P(t)\} = \int_0^t\int_0^y \lambda(x)\alpha(y,x)h'(1,t,y)\,dx\,dy, \quad h'(1,t,y) = \frac{\partial}{\partial s}h(s,t,y)\Big|_{s=1}.$$

At this stage an actual example may be informative. Supposing we have a constant Poisson input $\lambda \equiv \lambda(x)$ and that the function $\alpha(y, x)$ takes the

form $\alpha e^{-\alpha(y-x)}$, $x \leq y$, where α is some positive constant. Then (1) becomes

$$\begin{aligned} (s,t) &= \exp\left\{\int_0^t\int_0^y \lambda\alpha e^{-\alpha(y-x)}\,dx\,dy\,(s-1)\right\} \\ &= \exp\{\Lambda(t)(s-1)\}, \end{aligned} \tag{3}$$

where $\Lambda(t) = \lambda t + (\lambda/\alpha)(e^{-\alpha t}-1)$. Further, if the generating function $h(s, t, y)$ is Poisson with parameter $\gamma > 0$, i.e. is independent of t and y then (2) assumes the form

$$\begin{aligned} p(s,t) &= \exp\left\{\int_0^t\int_0^y \lambda\alpha e^{-\alpha(y-x)}[e^{\gamma(s-1)}-1]\,dx\,dy\right\} \\ &= \exp\{\Lambda(t)[e^{\gamma(s-1)}-1]\}, \end{aligned} \tag{4}$$

which is a Neyman Type A distribution with parameters $\Lambda(t)$ and γ. Notice that, since $e^x > 1+x$ for $x \neq 0$ and $\Lambda(0) = 0$,

$$\Lambda(t) > \lambda t + \frac{\lambda}{\alpha}(1-\alpha t-1) = 0, t > 0,$$

and we have a genuine distribution for $t \geq 0$.

(B) DISCONTINUOUS INPUT MODELS

We now define the continuity and jump sets of $[0, t]$. If for $x \in [0, t]$, the process satisfies the conditions of postulate I, then x will be said to be a point of continuity of the process. The set of all continuity points, $\mathscr{C}$, will be called the continuity set.

On the other hand all points of $[0, t]$ at which another p.g.f. is defined are referred to as jump points, the set of all such points being designated by $\mathscr{J}$. The set $\mathscr{J}$ will always contain a finite number of points and $\mathscr{C} \cup \mathscr{J} = [0, t]$. It makes no difference to the final results if all (or some) of the points of $\mathscr{J}$ also belong to $\mathscr{C}$.

It is clear then that the contribution from $\mathscr{C}$ to the p.g.f.'s of $F(t)$ and $P(t)$ are results (1) and (2), which will now be written as $f_C(s, t)$ and $p_C(s, t)$. Now suppose that for $t_j \in \mathscr{J}$ the input p.g.f. is $g(s, t_j)$ then, since the probability that a female parasite is mature at time t given that she gained entrance at t_j, is

$$\int_{t_j}^{t} \alpha(y, t_j)\,dy = A(t, t_j),$$

the p.g.f. for the number of mature female parasites at t is

$$g[1+A(t, t_j)(s-1), t_j] = G(s, t_j).$$

Thus the contribution from $\mathscr{I}$ to the overall p.g.f. for $F(t)$ is

$$f_{\mathscr{I}}(s,t) = \prod_{t_j \in \mathscr{I}} G(s,t_j).$$

Similarly, the contribution of $\mathscr{I}$ to the p.g.f. of $P(t)$ is

$$p_{\mathscr{I}}(s,t) = \prod_{t_j \in \mathscr{I}} H(s,t_j)$$

where

$$H(s,t_j) = g[1-A(t,t_j)+\int_{t_j}^{t}\alpha(y,t_j)h(s,t,y)\,dy, t_j].$$

Thus finally the total p.g.f.'s for $F(t)$ and $P(t)$ are $f_{\mathscr{C}}(s,t)f_{\mathscr{I}}(s,t) = f(s,t)$ and $p_{\mathscr{C}}(s,t)p_{\mathscr{I}}(s,t) = p(s,t)$, respectively.

There is a considerable simplification if the input is Poisson, for then if $g(s,t_j) = \exp\{\lambda(t_j)(s-1)\}$

$$f(s,t) = \exp\left\{\left[\sum_{t_j\in\mathscr{I}}\lambda(t_j)A(t,t_j)+\int_0^t\int_0^y\lambda(x)\alpha(y,x)\,dx\,dy\right](s-1)\right\} \tag{5}$$

and

$$p(s,t) = \exp\left\{\sum_{t_j\in\mathscr{I}}\lambda(t_j)\left[-A(t,t_j)+\int_{t_j}^{t}\alpha(y,t_j)h(s,t,y)\,dy\right]+ \right.$$
$$\left.+\int_0^t\int_0^y\lambda(x)\alpha(y,x)[h(s,t,y)-1]\,dx\,dy\right\}. \tag{6}$$

As an example, consider the situation discussed in the Summary where the host (a sheep) is given a massive dose of parasites (worms) at time $t = 0$. In this case $\mathscr{C}$ is null and $\mathscr{I} = \{0\}$ and if the numbers of female larvae gaining entrance follow a Poisson distribution with parameter λ_0,

$$f(s,t) = \exp\left\{\lambda_0\int_0^t\alpha(y,0)\,dy\,(s-1)\right\}$$

$$p(s,t) = \exp\left\{\lambda_0\left[-\int_0^t\alpha(y,0)\,dy+\int_0^t\alpha(y,0)h(s,t,y)\,dy\right]\right\}.$$

In order to obtain some explicit results we can assume, as before, that $\alpha(y,0) = \alpha e^{-\alpha y}$ and $h(s,t,y) = e^{\gamma(s-1)}$. It is now easily verified that under these conditions

$$f(s,t) = \exp\{\lambda_0\Lambda(t)(s-1)\}$$

and

$$p(s,t) = \exp\{\lambda_0\Lambda(t)(e^{\gamma(s-1)}-1)\},$$

where

$$\Lambda(t) = 1-e^{-\alpha t}.$$

The primary objective of the above analysis has been to obtain general p.g.f.'s for $P(t)$. For this reason, the emphasis has been on the number of females gaining entrance to the host, the probability of maturation and the offspring production of mature females. If a prime interest was in the total number of parasites in the host at t, irrespective of sex, $N(t)$, we can let $\delta(x)$ be the rate at which parasites, male or female, gain entrance to the host at time x. Moreover $\beta(t, x)$ can be defined as the conditional probability that a parasite entering at time x is alive at time t. With these new definitions of the functions we can assume that the probability that a parasite enters during the period $[x, x+\Delta x]$ and is alive at time t is approximately $\delta(x)\beta(t, x)\Delta x$ and in this way find the p.g.f. for the total number of live parasites at t to be

$$n_C(s, t) = \exp\left\{\int_0^t \delta(x)\beta(t, x)\,dx\,(s-1)\right\}. \tag{7}$$

Equation (7) is the contribution of $\mathscr{C}$ to the p.g.f. of $N(t)$ and, by similar reasoning to that used to derive $f_J(s, t)$, $n_J(s, t)$ can be defined and we find that $n(s, t) = n_C(s, t)n_J(s, t)$. Moreover, (7) suggests the possibility that with a continuous input system, an equilibrium population would be maintained in the host as defined by

$$n(s, \infty) = \exp\left\{\lim_{t\to\infty}\int_0^t \delta(x)\beta(t, x)\,dx\,(s-1)\right\}.$$

If, for example, $\delta(x) \equiv \delta$ and $\beta(t, x) = e^{-\beta(t-x)}$, then

$$n(s, t) = \exp\left\{\frac{\delta}{\beta}(1-e^{-\beta t})(s-1)\right\}.$$

In this case it is clear that $n(s, \infty)$ is Poisson with parameter δ/β, the ratio of the input rate to the death rate.

3. Antigenic Information

Each parasite has an antigenic information trajectory (as described in the Summary) and we will assume here, for simplicity, that this trajectory is specified by the model

$$u(t) = m(t)+\varepsilon(t), t \geq 0,\ \mathrm{E}\{\varepsilon(t)\} = 0. \tag{8}$$

We assume that $u(t)$ has a frequency function $k(u, t)$ and, in order to calculate the properties of the total amount of antigenic information present in the host, we need

Postulate 2

The probability that a parasite enters the host during the time interval $[x, x+\Delta x]$ and that its information at time t lies in the interval $[u, u+\Delta u]$ is

$$\delta(x)\beta(t,x)k(u,t-x)\Delta x\Delta u+o(\Delta x\Delta u).$$

From these assumptions it follows that the characteristic function, c.f., for the total amount of antigenic information in the host at time t is

$$\chi_C(\theta,t)=\exp\left\{\int_0^t\delta(x)\beta(t,x)[\phi(\theta,t-x)-1]\,dx\right\} \tag{9}$$

where

$$\phi(\theta,t-x)=\int_{-\infty}^{\infty}e^{i\theta u}k(u,t-x)\,du.$$

Explicit inversion of equation (9) would be a difficult task in most cases, but the cumulants are relatively easily obtained. In fact, if $\kappa_j(t)$ are the cumulants specified by χ_C and $\mu_j(t)$ are the moments specified by ϕ, then

$$\kappa_j(t)=\int_0^t\delta(x)\beta(t,x)\mu_j(t-x)\,dx, j\geq 1. \tag{10}$$

Again we define a continuity set $\mathscr{C}$ and a jump set $\mathscr{J}$ and, clearly, the above results apply to $\mathscr{C}$. If now $g(s, t_j)$ is the input p.g.f. for female and male parasites, then the contribution of $\mathscr{J}$ to the overall c.f. is

$$\chi_J(\theta,t)=\prod_{t_j\in\mathscr{J}}\chi_j(\theta,t),$$

where

$$\chi_j(\theta,t)=g([1+\beta(t,t_j)(\phi(\theta,t-t_j)-1)],t_j).$$

The c.f. then takes the form $\chi(\theta,t)=\chi_J(\theta,t)\chi_C(\theta,t)$.

We turn again to specific examples. Suppose, for instance, $\delta(x)\equiv\delta$ and $\beta(t,x)=e^{-\beta(t-x)}$, $\beta>0$, and $k(u,t-x)=\eta\,e^{-\eta u}$ then we find that

$$\begin{aligned}\chi_C(\theta,t)&=\exp\left\{\int_0^t\delta\,e^{-\beta(t-x)}[(1-i\theta/\eta)^{-1}-1]dx\right\}\\&=\exp\{B(t)[(1-i\theta/\eta)^{-1}-1]\}.\end{aligned}$$

where

$$B(t)=(\delta/\beta)[1-e^{-\beta t}].$$

Since $\log\chi_C(\theta,t)=B(t)[(1-i\theta/\eta)^{-1}-1]$, it is clear that $\kappa_j=j!\eta^{-j}B(t)$.

On the other hand, suppose the input is not continuous but instead is as described for the sheep earlier. In this case $g(s,0)=\exp\{\lambda_0(s-1)\}$

$$\chi_J(\theta,t)=\exp\{\lambda_0\beta(t,0)[\phi(\theta,t)-1]\}$$

and substituting the parametric forms of $\beta(t, 0)$ and $\phi(0, t)$ used above

$$\chi_J(0, t) = \exp\{\lambda_0 e^{-\beta t}[(1 - i\theta/\eta)^{-1} - 1]\}.$$

If $Z(t)$ is the total amount of antigenic information being produced by parasites in the host at time t, then the jth cumulant of $Z(t)$ is $\lambda_0 e^{-\beta t} j! \eta^{-j}$.

4. The Unconditional Models

So far we have been considering conditional (or personal) p.g.f.'s since they are specific for one particular host. In order to take care of the fourth biological assumption of the Summary it is now necessary to introduce a further generality.

Once the arbitrary functions $\alpha(y, x)$, $h(s, t, y)$ and $\beta(t, x)$ of equations (1), (2) and (7) are parameterized for any specific problem, then the p.g.f.'s f, p and n will be functions of some parameter sets θ_f, θ_p and θ_n, say. The p.g.f.'s are individual in the sense that we will postulate that each host has his own particular values of the parameters concerned. To illustrate the ideas, we will take $f(s, t)$ and write $f(s, t|\theta_f)$ to emphasize the above point.

In order to obtain unconditional p.g.f.'s for an arbitrary host drawn at random from the population of all hosts, we assign a distribution function $\Phi_f(\theta)$ to the vector θ_f. The unconditional p.g.f. $f(s, t)$ is then given by

$$f(s, t) = \int_{\Omega_f} f(s, t|\theta_f)\, d\Phi_f(\theta),$$

where Ω_f is the parameter space for θ_f.

We now give specific examples to fix ideas. Equation (3), then, is written as

$$f(s, t|\alpha) = \exp\{[\lambda t + \lambda/\alpha(e^{-\alpha t} - 1)](s - 1)\}$$

and suppose that α has a distribution function

$$\Phi_f(\alpha) = \int_0^{\alpha} \rho_1 e^{-\rho_1 x}\, dx.$$

Then $\Omega_f = [0, \infty]$ and

$$f(s, t) = \int_0^{\infty} \exp\{[\lambda t + \lambda/\alpha(e^{-\alpha t} - 1)](s - 1)\}\rho_1 e^{-\rho_1 \alpha} d\alpha.$$

The above integral does not appear to allow explicit evaluation, although $E\{F(t)\}$ and $V\{F(t)\}$ can be found with little trouble. For example,

$$I(t) = E\{F(t)\} = \int_0^{\infty} [\lambda t + \lambda/\alpha(e^{-\alpha t} - 1)]\rho_1 e^{-\rho_1 \alpha} d\alpha$$
$$= \lambda t - \lambda \rho_1 \log(1 + t/\rho_1).$$

The above integral $I(t)$ is negotiated by first differentiating the integrand with respect to t and then integrating with respect to α. This procedure is

justified since the integral defining $I'(t)$ is uniformly convergent for all t. The required expression is then obtainable immediately from $I'(t)$ by a further integration.

Similarly, in the case where an initial dose of female larvae are given to sheep, as discussed earlier, we have

$$(s,t)=\int_0^\infty \exp\{\lambda_0[1-e^{-\alpha t}](s-1)\}\rho_1 e^{-\rho_1\alpha}\,d\alpha$$

and

$$E\{F(t)\}=\lambda_0\int_0^\infty (1-e^{-\alpha t})\rho_1 e^{-\rho_1\alpha}\,d\alpha$$

$$=\lambda_0[1-\rho_1/(t+\rho_1)]=\frac{\lambda_0 t}{t+\rho_1}.$$

Suppose, now we consider the distribution of $P(t)$ as specified by (4). Then if we let γ be distributed exponentially with parameter ρ_2, we obtain for $E\{P(t)\}$

$$E\{P(t)\}=\int_0^\infty\int_0^\infty [\lambda t+\lambda/\alpha(e^{-\alpha t}-1)]\gamma\rho_1 e^{-\rho_1\alpha}\rho_2 e^{-\rho_2\gamma}\,d\alpha\,d\gamma$$

$$=[\lambda t-\lambda\rho_1\log(1+t/\rho_1)]/\rho_2$$

and for the discrete input case

$$E\{P(t)\}=\lambda_0 t/(t+\rho_1)\rho_2.$$

5. Discussion

The primary objective of this paper is to indicate how general models describing the accumulation and reproduction of parasites in a host can be formulated. Moreover, by considering discrete and continuous input systems a variety of situations can be dealt with simultaneously. Apart from the algebraic examples, the formulae are in a general form and it is found that, for instance, by suitably specifying the functions $\alpha(y, x)$, $h(s, t, y)$ and $\beta(t, x)$, the concept that the control of parasite burdens is mediated around threshold levels of responsiveness as proposed by Dineen (*loc. cit.*) can be readily dealt with.

The next step will be to adopt the discrete input models to various studies involving the parasites of sheep. Careful choice of the parametric forms of the functions α, h and β will have to be made and the applicability of the models assessed from collected data. Once experience is gained with the simple discrete input models, a continuous input system for the grazing sheep can be examined.

Clearly, a great deal of careful work lies ahead, both in the collection of suitable data and in the construction of satisfactory models. There will be considerable trouble finding useable estimators of the unknown parameters of the models. Nevertheless, it appears likely that initial stochastic developments may follow the lines presented in this paper.

The authors wish to thank Dr G. S. Watson for his comments on an earlier draft of this paper. They are also grateful to Dr John Dineen for discussions leading to the formulation of the models.

REFERENCES

BARTLETT, M. S. (1960). "Stochastic Population Models in Ecology & Epidemiology." Methuen's Statistical Monographs. New York: John Wiley & Sons.

DINEEN, J. K. (1963*a*). *Nature, Lond.* **197**, 268.

DINEEN, J. K. (1963*b*). *Nature, Lond.* **197**, 471.

DINEEN, J. K., DONALD, A. D., WAGLAND, B. M. & OFFNER, J. (1965). *Parasitology*, **55**, 515.

DINEEN, J. K., DONALD, A. D., WAGLAND, B. M. & TURNER, J. H. (1965). *Parasitology*, **55**, 163.

DONALD, A. D., DINEEN, J. K., TURNER, J. H. & WAGLAND, B. M. (1964). *Parasitology* **54**, 527.

A (b) [4]

Stochastic Models of Populations of Helminthic Parasites in the Definitive Host. I*

G. M. TALLIS** AND M. K. LEYTON***
Department of Biostatistics, Johns Hopkins University, Baltimore, Maryland

Communicated by K. E. F. Watt

ABSTRACT

Since helminthic parasites must enter the human host from an external environment and cannot multiply within the definitive host, the parasite population at a given moment in the host may be characterized by an immigration-death process. The arrival (input) of worms may be of either a random or a contagious nature. A general model is developed and specific relevant discrete distributions are considered. The equilibrium distributions are regarded as the appropriate form to characterize worm populations in hosts inhabiting endemic regions. An application of the model to an epidemiological problem is discussed.

INTRODUCTION

By definition, a parasite passes its life cycle within one or more hosts. That organism in which the parasite matures to the adult form is called the definitive host, other hosts being intermediate. This article is primarily concerned with helminthic parasites, that is, parasitic worms (e.g., schistosomes, hookworms, tape worms) that enter the definitive host from the external environment and produce offspring in the form of eggs or larvae. These ultimately move on to complete the remainder of the cycle elsewhere. Since, subsequently, attention is restricted to that part of the

* Paper Number 429, Johns Hopkins University, Department of Biostatistics.

** Current address: Division of Mathematical Statistics, C.S.I.R.O., Alpha House, 60 King street, Newtown, N.S.W., Australia.

*** Current address: Division of Biostatistics, Department of Preventive Medicine, School of Medicine, University of Washington, Seattle, Washington.

Mathematical Biosciences 4 (1969), 39–48

life cycle involving the definitive host, the process to be described is a population of worms that is subject to immigration and mortality pressures.

It is obviously difficult to obtain detailed information on parasites affecting human populations. The main source of reliable data concerning the breeding and survival potential of the worms within human subjects is from autopsy studies. The limitations of these restricted studies require no emphasis.

There is a pressing need for reliable mathematical models to assist the parasitologist with his work. So far, few such models have been developed. Hairston [2] proposed a deterministic model of the complete life cycle of the digenetic trematode *Schistosoma japonicum*. In 1963, Tallis and Donald [7] developed stochastic models to describe the distribution on pasture of the larval forms of intestinal nematodes of sheep. Subsequently, Tallis and Leyton [8] reported general models describing host–parasite relationships and certain pertinent mathematical techniques were presented.

In this paper, specific stochastic models for the size of a helminthic parasite population in the definitive host are developed. These models consider several types of input and survival mechanisms. In some cases, the relevant discrete distributions are too complicated to allow individual probability terms to be obtained explicitly. However, from the derived probability generating function (pgf), the mean and variance of each process are calculated as well as the probability that there is no infection. The equilibrium distributions are often more tractable, and these are regarded as the appropriate form to characterize worm populations in hosts living in an endemic region. In the last section, an application of the models is discussed.

It is stressed that the models derived in the following apply to any homeotherm and, in particular, to man. The simplest meaningful set of biological postulates is used as a basis for individual models; and such factors as the death of the host, development of resistance by the host, competition effects on the parasite, and seasonal variations are disregarded. Undoubtedly, further complicating assumptions may be required as experience with the various formulas accumulates under the specific conditions of application.

RESULTS

In order to isolate and emphasize the various models discussed in this section, it is partitioned into short subsections. The techniques that are

used in the derivations of the equations are well known and most of them are presented in Feller [3].

Size of the Helminthic Parasite Population in the Definitive Host

Because of the form of the life cycle spent outside the definitive host by various forms of parasites, the rencontre of host and infective forms of the parasite is truly a random event. Let λ be the "exposure rate" of the host, that is, the average number of infective contacts per unit time; then we assume that the probability of an exposure occurring in time Δt is $\lambda\,\Delta t + o(\Delta t)$. This assumption implies that the number of infective contacts of the host by time t is Poisson with parameter λt.

Suppose now that the number of parasites that gain entrance at each infective contact is a random variable N with pgf $h(s)$; then the number of worms that have entered the host by time t has pgf

$$L(s, t) = \exp\{\lambda t[h(s) - 1]\}. \tag{1}$$

At this stage, it is appropriate to introduce the "death function" $f(t)$, which is the frequency function for the survival time of the parasite once it is inside the definitive host. Thus, $M(t) = 1 - F(t)$, $F(t) = \int_0^t f(t)\,dt$, is the probability that a parasite entering the host at zero time survives to time t. Consider the n-partition of the interval $[0, t]$, $t_0 = 0, t_1, \ldots, t_n = t$, where $t_i - t_{i-1} = \Delta t = t/n$. Concentrating on the ith subinterval, the pgf for the number of parasites entering the host during this interval and surviving to time t is

$$[1 + \Delta t\lambda\{h[1 + M(t - \bar{t}_i)(s - 1)] - 1\} + o(\Delta t)].$$

From the foregoing pgf, and using the independence of the exposures, we can show that the pgf for the total number of live worms in the host at time t is

$$\Pi(s, t) = \exp\left(\lambda\int_0^t \{h[1 + M(t - x)(s - 1)] - 1\}\,dx\right). \tag{2}$$

It is now easily verified that the first two factorial moments of the process are

$$\mu(t) = \lambda\bar{N}\int_0^t M(w)\,dw,$$

$$\mu_{[2]}(t) = \lambda h''(1)\int_0^t M^2(w)\,dw + \lambda^2\bar{N}^2\left(\int_0^t M(w)\,dw\right)^2 \tag{3}$$

where $\bar{N} = h'(1) = E(N)$.

The equilibrium situation is obtained by letting $t \to \infty$ and, in particular,

$$\mu(\infty) = \lambda \bar{N} \int_0^\infty M(w)\, dw$$

$$= \lambda \bar{N} \int_0^\infty [1 - F(w)]\, dw = \lambda \bar{N} \bar{M} \tag{4}$$

where $\bar{M}$ is the average life-span of the worms in the host. Thus, $\mu(\infty)$ is the average input times the average length of life of the worms. Certain special cases will now be considered.

(1) *Random input and age-independent death rate.* This is the most elementary model; it assumes that parasites enter independently with intensity λ and that the life-span is exponentially distributed with parameter μ. This situation corresponds to a time-homogeneous linear immigration-death process (Cox and Miller [2], p.168). Thus, (2) specializes by setting $h(s) = s$ and $M(t - x) = \exp[-\mu(t - x)]$ and it is found that

$$\Pi(s, t) = \exp\left\{\frac{\lambda}{\mu}[1 - \exp(-\mu t)](s - 1)\right\}, \tag{5}$$

which specifies a Poisson distribution with parameter $(\lambda/\mu)[1 - \exp(-\mu t)]$. Clearly, $\Pi(s, \infty)$ is Poisson with parameter λ/μ, which agrees with (4).

(2) *Random input with age-dependent death rate.* Here it is assumed that the life-span of the parasites has distribution function

$$F(t) = \int_0^t \frac{\mu^\gamma}{\Gamma(\gamma)} x^{\gamma-1} \exp(-\mu x)\, dx,$$

and, again, $h(s) = s$. Clearly, the process is still Poisson with parameter

$$\lambda \int_0^t \left[1 - \int_0^w \frac{\mu^\gamma}{\Gamma(\gamma)} y^{\gamma-1} \exp(-\mu y)\, dy\right] dw$$

and

$$\Pi(s, \infty) = \exp\left[\frac{\lambda\gamma}{\mu}(s - 1)\right].$$

(3) *Contagious input and age-independent death rates.* Equation (2) will now be examined by setting $M(w) = \exp(-\mu w)$ and using various forms of $h(s)$. Where necessary, the distributions have been truncated to remove the zero class. First, let $h(s) = ps/(1 - qs)$, $p + q = 1$, the truncated geometric distribution. Now

$$\Pi(s, t) = \exp\left(\lambda \int_0^t \left\{ \frac{p[1 + (s - 1)\exp(-\mu w)]}{1 - q[1 + (s - 1)\exp(-\mu w)]} - 1 \right\} dw\right)$$

$$= \left[\frac{p - q(s - 1)\exp(-\mu t)}{1 - qs}\right]^{\lambda/\mu q} \tag{6}$$

after some algebra. From (6) the mean and variance are found to be

$$\mu(t) = \frac{\lambda}{\mu p}[1 - \exp(-\mu t)],$$

$$\sigma^2(t) = \frac{\lambda}{\mu p^2}[1 - \exp(-\mu t)][1 + q\exp(-\mu t)].$$

The foregoing formulas are most easily obtained by substituting $s = e^\theta$ in $\ln \Pi(s, t)$ to obtain the cumulant generating function, and the first two derivatives evaluated at $\theta = 0$ give the required results. The equilibrium distribution is

$$\Pi(s, \infty) = \left(\frac{p}{1 - qs}\right)^{\lambda/\mu q},$$

which is negative binomial with mean $\lambda/\mu p$ and variance $\lambda/\mu p^2$. The probability that there are no parasites in the host at time t is

$$\Pi(0, t) = [p + q\exp(-\mu t)]^{\lambda/\mu q}.$$

In the case of the log series distribution, $h(s) = -\alpha \ln(1 - \beta s)$, $\alpha > 0$ and $\beta = 1 - \exp(-1/\alpha)$ but Π takes the somewhat unmanageable form

$$\Pi(s, t) = \exp\left[\lambda \int_0^t (-\alpha \ln\{1 - \beta[1 + \exp(-\mu w)(s - 1)]\} - 1)\, dw\right]. \tag{7}$$

However,

$$\mu(t) = \alpha\lambda\left[\exp\left(\frac{1}{\alpha}\right) - 1\right]\frac{1 - \exp(-\mu t)}{\mu}$$

and

$$\sigma^2(t) = \mu(t) + \alpha\lambda\left[\exp\left(\frac{1}{\alpha}\right) - 1\right]^2 \frac{1 - \exp(-2\mu t)}{2\mu},$$

as may be verified by converting Π to cumulant generating function form as above. The zero class has probability

$$\Pi(0, t) = \exp\left(\alpha\lambda t - \alpha\lambda \int_0^t \ln\{1 + \beta[1 + \exp(-\mu w)]\}\, dw\right).$$

It can be shown after some algebra that

$$\Pi(s, \infty) = \exp\left[-\frac{\alpha\lambda}{\mu} \int_0^{\lambda\{[\alpha^{-1}+1](s-\eta)\}} \ln(1 - y)\frac{dy}{y}\right],$$

from which the expression for $\Pi(0, \infty)$ is readily written down. It can also be verified that as α tends to infinity, (7) approaches (5).

Suppose that $h(s)$ is a truncated Poisson distribution

$$h(s) = \frac{\exp[\eta(s - 1)] - \exp(-\eta)}{1 - \exp(-\eta)};$$

then proceeding as earlier it is found that $\Pi(s, t)$ assumes an unmanageable form. However, it turns out that

$$\mu(t) = \frac{\lambda\eta}{\mu[1 - \exp(-\eta)]}[1 - \exp(-\mu t)],$$

$$\sigma^2(t) = \mu(t) + \frac{\lambda\eta^2}{2\mu[1 - \exp(-\eta)]}[1 - \exp(-2\mu t)],$$

$$\Pi(0, t) = \exp\left[\lambda \int_0^t \frac{\{\exp[-\eta \exp(-\mu w)] - 1\}\, dw}{1 - \exp(-\eta)}\right],$$

and as η tends to zero $\Pi(s, t)$ tends to (5). Instead of presenting an explicit formula for $\Pi(s, \infty)$, we will develop a general result.

Consider the expression

$$\lim_{t\to\infty} \exp\left[\lambda \int_0^t (h\{1 + \exp[-\mu(t - x)](s - 1)\} - 1)\, dx\right]$$

$$= \exp\left(\lambda \int_0^\infty \{h[1 + \exp(-\mu w)(s - 1)] - 1\}\, dw\right);$$

then, concentrating on the term in the exponent, make the transformation

$1 + \exp(-\mu w)(s - 1) = v$. The term becomes

$$\frac{\lambda}{\mu}\int_s^1 \frac{h(v) - 1}{1 - v}\, dv$$

and the integral exists if

$$E[N] = h'(1) < \infty$$

in which case

$$\Pi(s, \infty) = \exp\left[\frac{\lambda}{\mu}\int_s^1 \frac{h(v) - 1}{1 - v}\, dv\right]. \tag{8}$$

As an illustration, we return to the case where $h(s)$ is a truncated Poisson. It is found then that

$$\Pi(s, \infty) = \exp\left\{\frac{\lambda}{\mu}\int_s^1 \frac{\exp[\eta(v - 1)] - 1}{(1 - \exp(-\eta))(1 - v)}\, dv\right\}.$$

Moreover, from (8),

$$\mu(\infty) = \frac{\lambda}{\mu}\bar{N},$$

as may be verified by differentiating $\Pi(s, \infty)$ with respect to s and letting $s \to 1$. Thus, for the truncated Poisson, $\mu(\infty) = \lambda\eta/\mu[1 - \exp(-\eta)]$.

A general result can be established for (2) in a similar way. Suppose $\bar{N}$ and $\bar{M}$ are both finite, as assumed in the derivation of most of the foregoing formulas; then, working with exponent of (2),

$$\lambda\int_0^t \{h[1 + M(w)(s-1)] - 1\}\,dw$$

$$= \lambda\int_0^t \left\{\sum_{n=0}^{\infty}[1 + M(w)(s-1)]^n p_n - 1\right\} dw$$

$$\leqslant \lambda(1-s)\int_0^t \sum_{n=0}^{\infty}(1 - F(w))n p_n\,dw \leqslant \lambda(1-s)\bar{M}\bar{N}$$

since $(1 - |x|)^n \geqslant 1 + n|x|$. The inequality above holds for all t and, hence, the limiting pgf $\Pi(s, \infty)$ exists. By an obvious application of the continuity theorem ([3], p. 262), the limiting distribution also exists.

AN APPLICATION

The relationship of prevalence to disease severity

The parasitologist is careful to distinguish between infection, the presence of a parasitic organism in the host, and disease, which is recognized by certain symptoms in the host. A characteristic of most helminthic parasites is that the severity of the disease is proportional to the number of parasites in the host. Few hookworms need to be present for eggs to be found in a person's feces, the common diagnostic test, whereas for the subject to show hookworm symptoms, it appears that the infection must consist of at least a hundred worms.

The question arises: Under what circumstances can a population have almost 100% infection and exhibit no hookworm disease? This point was raised by Dr. E. Schiller after a health survey in the Bandipur Union, West Bengal, India. Schiller and Chowdhurry [5] found a high percentage of infected people but a low incidence of the disease.

The converse question is also of interest: Under what conditions can there be a low percentage of infection but severe disease among those infected? Parasitologists offer several biological interpretations of these phenomena. Here, however, we show that these situations can be explained by purely statistical arguments with the aid of models derived earlier.

As an example, if the Bandipur district under consideration can be considered an endemic region, then we can apply the negative binomial distribution, which is the limiting distribution of (6). It is subsequently convenient to find the average number of worms per infected individual and the truncated distribution specified by the pgf. For this purpose,

$$\frac{[p/(1 - qs)]^{\lambda/\mu q} - p^{\lambda/\mu q}}{1 - p^{\lambda/\mu q}}$$

will be used. The required mean and variance are

$$\mu(\infty) = \frac{\lambda/\mu}{p(1 - p^{\lambda/\mu q})}$$

and

$$\sigma^2(\infty) = \mu(\infty)\left[\frac{q}{p}\left(\frac{\lambda}{\mu q} + 1\right) + 1 - \mu(\infty)\right].$$

For a high proportion of infection but low severity, the probability of no infection $p^{\lambda/\mu q}$ should be small. Remembering that $1/p$ is the average number of worms entering per exposure, we would expect this average to be small, say 2. That is, $p = q = 1/2$ and $p^{\lambda/\mu q} = (1/2)^{2\lambda/\mu}$. If $\lambda/\mu > 5$, then the probability of no infection is $1/1024$ and more than 99% of the population would be infected. The average number of worms per infected individual would be $10[1 - (1/2)^{10}]^{-1} \simeq 10$. The variance is 20 and standard deviation about 4.5. Thus, there would be little disease present.

In the case of a low proportion of infection with high severity, the average number of worms per exposure could be high, say 256, and $\lambda/\mu < 1/8$. Now $p^{\lambda/\mu q} \simeq 1/2$, the average number of worms per infected individual would be about 64 and the standard deviation 120. Thus, infected people would often have the disease severely.

These examples are enough to emphasize the possible use of models to explain some of the stochastic behavior of parasitic disease. Obviously, the formulas are an adjunct to, not a substitute for, sound biological thinking.

REFERENCES

1 A. C. Chandler and C. P. Read, *Introduction to parasitology with special reference to the parasites of man* (10th ed.). Wiley, New York, 1961.
2 D. R. Cox and H. D. Miller, *The theory of stochastic processes*. Wiley, New York, 1965.
3 W. Feller, *An introduction to probability theory and its applications*, Vol. I (2nd ed.). Wiley, New York, 1957.
4 N. G. Hairston, Population ecology and epidemiological problems, *Bilharziasis* (CIBA Found. Symp.). Little, Brown, Boston, 1962.
5 E. L. Schiller and A. B. Chowdhurry, *Parasitological survey of the Bandipur health union*. Johns Hopkins Center Rept.: 39–47 (1964).
6 G. M. Tallis, A note on the estimation of larval concentrations on pasture, *Australian J. Biol. Sci.* 17(1964), 1016–1019.
7 G. M. Tallis and A. D. Donald, Models for the distribution on pasture of infective larvae of the gastrointestinal nematode parasites of sheep, *Australian J. Biol. Sci.* **17**(1964), 504–513.
8 G. M. Tallis and M. K. Leyton, A stochastic approach to the study of the immunological control of parasite populations, *J. Theoret. Biol.* **13**(1966), 251–260.

A (b)[5]

Further Models for the Distribution on Pasture of Infective Larvae of the Strongyloid Nematode Parasites of Sheep

G. M. TALLIS
Division of Mathematical Statistics,
C.S.I.R.O. Newtown, N.S.W., Australia

A. D. DONALD
Division of Animal Health,
C.S.I.R.O. McMaster Laboratory
Glebe, N.S.W., Australia

Communicated by K. E. F. Watt

ABSTRACT

This paper discusses a general approach to the problem of constructing useful models for the distribution on pasture of infective larvae of sheep nematode parasites. The work is related to earlier results obtained by the authors, and modifications are introduced as a result of practical experience with the original models. An explicit form for the function $f(t)$, the probability that an egg reaches the infective larval stage in time t, is derived. The new version of $f(t)$ has components with direct biological interpretation.

INTRODUCTION

In earlier papers [5, 6], stochastic models were developed to describe the distribution on pasture of the infective larvae of some nematode parasites of sheep. These models were based on current knowledge of the distribution of fecal deposits in a paddock and on the assumption that the distribution of eggs between fecal deposits is Poisson. Subsequent experience [2, 4] has led to modifications in the original postulates and it seems appropriate at this stage to redevelop some of the previous results in the light of the present ideas.

Studies of fecal distributions on paddocks of different sizes carrying varying numbers of sheep have emphasised that deposits are definitely

Mathematical Biosciences 7 (1970), 179–190

44

nonuniformly distributed over the total area. However, it seems possible to divide the paddock into subareas within which the distribution is, to a good approximation, uniform, although there may be considerable between area differences in concentration due to the grazing and resting habits of the flock.

With this information in mind, it is the primary purpose of this paper to develop completely general distributions for the numbers of larvae in the various subareas of a paddock carrying S sheep. From these expressions certain reasonable simplifications will be introduced to make the formulae useful in practical investigations. Most of the argument is carried through using the general forms for the means and variances of the various processes.

In the sequel we pay particular attention to the function $f(t)$, which is the probability that an egg dropped onto the pasture at time zero is in the infective larval stage on the herbage at time t. Explicit expressions for $f(t)$ are derived from underlying biological hypotheses to give the formulae further interpretational value.

RESULTS

Distribution of the Total Number of Infective Larvae on Pasture. Let the paddock under consideration be of area A and suppose S sheep are introduced at zero time, $T = 0$. In accordance with the Introduction, the paddock is divided into k plots of area A_j, $\sum_{j=1}^{k} A_j = A$. Within each plot the defecation pattern of the flock is such that fecal deposits are approximately uniformly distributed over the area. Initially we focus attention on plot j and sheep i.

Let the number of eggs per fecal deposit for the ith sheep at $T = x$ be a random variable with probability generating function, p.g.f., $g_i(s, x)$ with mean

$$\lambda_i(x) = g_i'(1, x)$$

and variance

$$\sigma_i^2(x) = g_i''(1, x) + \lambda_i(x) - \lambda_i^2(x).$$

Now define the step function $N_{ij}(t)$ as the total number of fecal deposits associated with the ith sheep on plot j by $T = t$, and let $f(t)$ be as defined in the Introduction. Then the p.g.f. for the number of live larvae at $T = t$ emanating from a deposit dropped at $T = x$ is

$$h_i(s, x, t) = g_i\{[1 + f(t - x)(s - 1)], x\},$$

and the p.g.f. for the total number of live larvae on plot j due to the

ith sheep at $T = t$ is

$$\phi_i(s_j, t) = \exp\left\{\int_0^t \log h_i(s_j, x, t)\, dN_{ij}(x)\right\}. \tag{1}$$

Formula (1) assumes stochastic independence between the numbers of eggs in the deposits. Although this assumption can be made plausible by some mathematical argument, the details will not be presented since they are somewhat involved. The restriction of independence is, of course, easily removed at the cost of increasing the complexity of the expressions and the estimation procedures.

Similarly, the p.g.f. for the number of larvae on plot j due to the S sheep is

$$\phi(s_j, t) = \exp\left\{\sum_{i=1}^{S}\int_0^t \log[h_i(s_j, x, t)]\, dN_{ij}(x)\right\}, \tag{2}$$

and the joint p.g.f. for all k plots is

$$\phi(s, t) = \prod_{j=1}^{k} \phi(s_j, t). \tag{3}$$

If the prime interest is the total number of larvae on pasture, then the appropriate p.g.f., $\phi(s, t)$ is obtained from (3) by setting $s_j = s$ for all j.

Equations (1), (2), and (3) give the required expressions in the most general form, but they are of limited practical value since, for instance, the step functions $N_{ij}(t)$ must be known. Below, some reasonable simplifications are introduced to illustrate the use of the various p.g.f.'s. However, before proceeding, general expressions for the means and variances of the various processes are obtained.

Let $L_j(t)$ be the number of live larvae on plot j at $T = t$, then

$$\begin{aligned} E\{L_j(t)\} &= \sum_{i=1}^{S}\int_0^t \lambda_i(x) f(t-x)\, dN_{ij}(x), \\ V\{L_j(t)\} &= \sum_{i=1}^{S}\int_0^t V_i(x, t)\, dN_{ij}(x), \end{aligned} \tag{4}$$

where $V_i(x, t) = \sigma_i^2(x) f^2(t-x) + \lambda_i(x) f(t-x)[1 - f(t-x)]$.

If $L(t)$ is the total number of larvae on the whole paddock, then

$$\begin{aligned} L(t) &= \sum_{j=1}^{k} L_j(t), \\ E\{L(t)\} &= \sum_{j=1}^{k} E\{L_j(t)\}, \end{aligned}$$

and

$$V\{L(t)\} = \sum_{j=1}^{k} V\{L_j(t)\}.$$

In order to simplify (4) to a manageable approximation set

$$S\bar{\lambda}(x) = \sum_{i=1}^{S} \lambda_i(x),$$

$$S\bar{\sigma}^2(x) = \sum_{i=1}^{S} \sigma_i^2(x),$$

and

$$N_{ij}(x) = n_j x,$$

where n_j is the daily fecal output per sheep on plot j averaged over all sheep; then

$$E\{L_j(t)\} \simeq Sn_j \int_0^t \bar{\lambda}(x) f(t-x)\, dx,$$

$$V\{L_j(t)\} \simeq Sn_j \int_0^t \bar{V}(x, t)\, dx, \tag{5}$$

where

$$\bar{V}(x, t) = \bar{\sigma}^2(x) f^2(t-x) + \bar{\lambda}(x) f(t-x)[1 - f(t-x)].$$

By suitable sampling, estimates of the functions n_j, $\bar{\lambda}(x)$, and $\bar{\sigma}^2(x)$ must be obtained for values of x in the range of interest. The integrals can then be approximated by numerical quadrature provided $f(t)$ is known.

The amount of work required to carry out the preceding estimation is considerable and, at best, tedious. Some further assumptions will now be made which should facilitate the application of (5).

Let $Z_i(x)$ be the random variable associated with $g_i(s, x)$, and suppose that the weight of each fecal deposit, W, is a random variable. Then for fixed $W = w$ let $g_i(s, x)$ be Poisson with parameter

$$\lambda_i(x) \mid w = w\gamma_i(x).$$

Then

$$E\{Z_i(x) \mid w\} = w\gamma_i(x)$$

and hence

$$\lambda_i(x) = E\{Z_i(x)\} = E(W)\gamma_i(x)$$

and, similarly,

$$\sigma_i^2(x) = V\{Z_i(x)\} = \gamma_i(x)E(W) + \gamma_i^2(x)V(W).$$

In order to use (5) under the assumptions above it is necessary to estimate $E(W)$ and $V(W)$, $\hat{E}(W)$, and $\hat{V}(W)$ say, for the flock of sheep. Then, since $\gamma_i(x)$ can be estimated for each sheep by standard methods, average values of $\lambda(x)$ and $\sigma^2(x)$ are easily obtained at different time points. The integrals in (5) can then be approximated by numerical quadrature.

A Model for $f(t)$. The function $f(t)$ is of some biological interest in its own right. Consider the elementary flow diagram in Fig. 1.

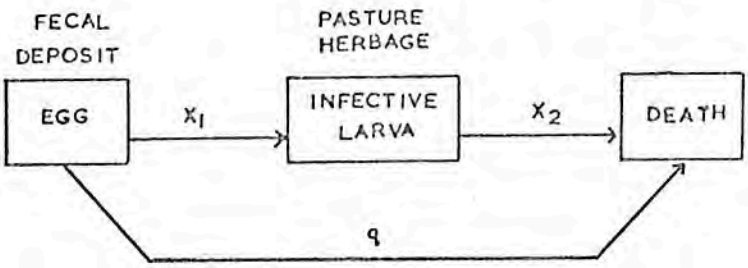


FIG. 1. Elementary flow diagram.

(i) q is the probability that an egg does not get to the herbage as an infective larva and $p = 1 - q$ is the probability that it will.

(ii) Given that an egg completes development to the infective stage and the resulting larva migrates to the pasture, X_1 is the time taken for development and migration from fecal deposit to herbage.

(iii) X_2 is the length of life as an infective larva on the herbage.

Suppose that X_1 and X_2 have distribution functions Φ_1 and Φ_2 and are independently distributed, then

$$f(t) = p\int_0^t [1 - \Phi(t - x)]\, d\Phi_1(x). \tag{6}$$

The assumption of independence is not necessary, but a more general approach leads to complexities of estimation which are not pursued here.

In order to establish (6) let A be the event that an egg never develops into a larva that completes the migration process to the grass. Thus $\Pr\{A\} = q$ and $\Pr\{\bar{A}\} = p$, where $\bar{A}$ is the complementary event to A. Define the set

$$S = \{x_1, x_2;\ 0 \leqslant x_1 \leqslant t,\ t - x_1 \leqslant x_2\},$$

then

$$f(t) = \Pr\{\text{egg develops into an infective larva alive at time } t \mid A\}q + \Pr\{\text{egg develops into an infective larva alive at time } t \mid \bar{A}\}p$$

$$= p\iint_S d\Phi_1(x)\, d\Phi_2(x)$$

$$= p\int_0^t [1 - \Phi_2(t-x)]\, d\Phi_1(x).$$

If $f^*(s)$ is the Laplace transform, L.T., of $f(t)$, then

$$f^*(s) = ps^{-1}\phi_1^*(s)[1 - \phi_2^*(s)], \tag{7}$$

where $\phi_i^*(s)$ is the L.T. of $\phi_i(x)$, the derivative of $\Phi_i(x)$ which is assumed to exist.

From (7) it is easily verified that

$$\mu_r = \int_0^\infty t^r f(t)\, dt$$

$$= pr! \sum_{n=0}^{r} \mu_n^{(1)}\mu_{r+1-n}^{(2)}/n!\,(r+1-n)! \tag{8}$$

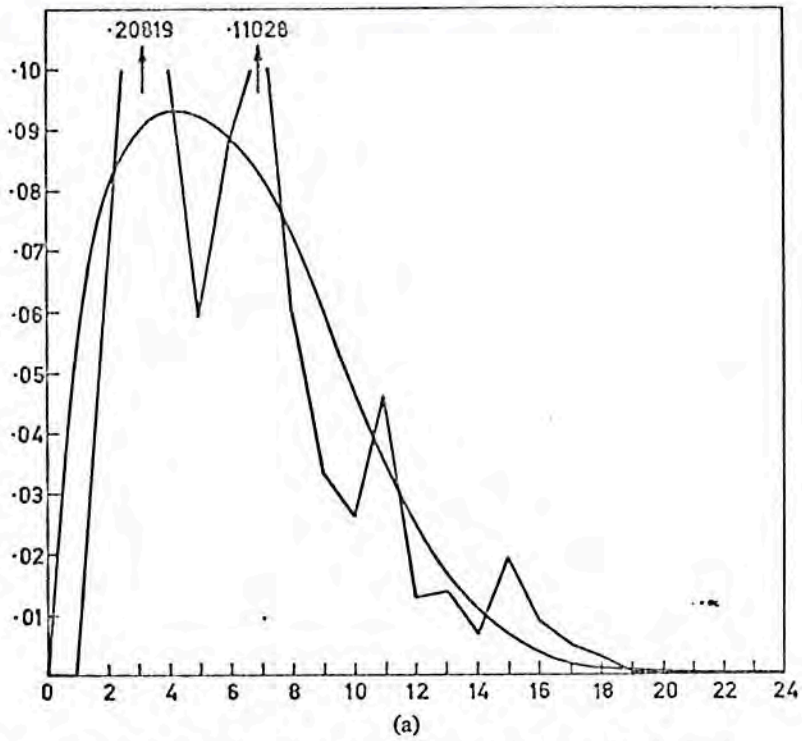
and also $\lim_{s\to 0} sf^*(s) = 0$, implying $\lim_{t\to\infty} f(t) = 0$. In (8),

$$\mu_n^{(i)} = \int_0^\infty x^n\, d\Phi_i(x),$$

and it is interesting to notice that the area under $f(t)$ is $p\mu_1^{(2)}$. The moment equations specified by (8) are useful for fitting $f(t)$ to data, as illustrated below.

Fitting the $f(t)$ Model to Empirical Data. Empirical estimates of the function $f(t)$ have been obtained in the course of field ecological studies, which will be reported in detail elsewhere, on the free-living stages of *Trichostrongylus colubriformis* and *Haemonchus contortus*, two important nematode parasites of sheep. Donald [4] has given a preliminary account of some of this work, including the methods used. Briefly, a known amount of sheep faeces containing parasite eggs is scattered on a small plot of pasture. Estimates are made initially of the total number of eggs of each species placed on the plot, and at weekly intervals the numbers of infective larvae present on the herbage are estimated from samples. Point estimates of the function $f(t)$ at weekly intervals are obtained by dividing the

numbers of infective larvae recovered at time t by the number of eggs exposed on the plot at $t = 0$. Beginning in June 1967, a fresh plot has been set up every four weeks and herbage sampling of each plot has continued until at least three consecutive zero recoveries have been obtained For the present purpose of illustrating the fitting of the model, data for *T. colubriformis* from each of four plots have been selected as representative.



(a)

FIG. 2. Plots of the observed and fitted values of $f(t)$ for the four groups.

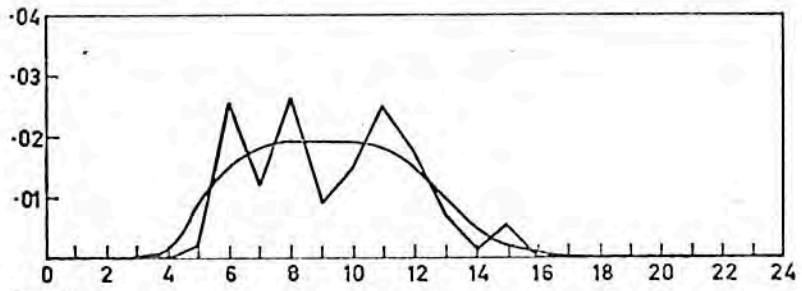


FIG. 2b

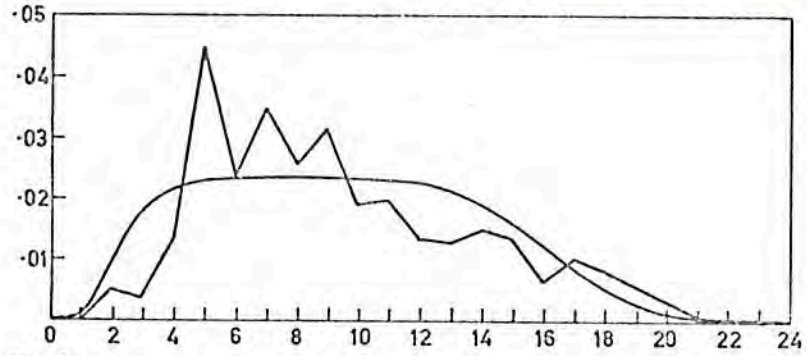


FIG. 2c

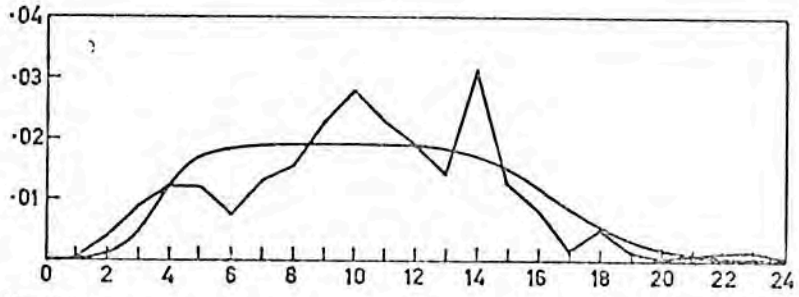


FIG. 2d

In order to fit (6) to these data assume that

$$d\Phi_i(x) = \frac{a^{\gamma_i}}{\Gamma(\gamma_i)} x^{\gamma_i - 1} e^{-ax}\, dx;$$

then

$$f(t) = p[I(at, \gamma_1) - I(at, \gamma_1 + \gamma_2)],$$

where $I(x, \gamma)$ is the incomplete gamma function. The first three moments are

$$\mu_0 = \frac{p\gamma_2}{a},$$

$$\mu_1 = \frac{p\gamma_2[\gamma_1 + (\gamma_2 + 1)/2]}{a^2},$$

$$\mu_2 = \frac{p\gamma_2[(\gamma_1 + 1)\gamma_1 + \gamma_1(\gamma_2 + 1) + (\gamma_2 + 1)(\gamma_2 + 2)/3]}{a^3}.$$

Trapezoidal approximation to μ_j can be calculated using the formula

$$\hat{\mu}_j = \tfrac{1}{2} \sum_i [\hat{f}(t_{i+1}) t_{i+1}^j + \hat{f}(t_i) t_i^j](t_{i+1} - t_i)$$

for $j = 0, 1, 2$. The values of $f(t)$ must be estimated at equally spaced time points as described above.

Set $\hat{a}\hat{\mu}_0 = 1$ and define $i(j) = \hat{\mu}_j/\hat{\mu}_0^{j+1}$, then

$$\hat{\gamma}_2 = [1 + 12[i(2) - i(1)^2 + i(1)]]^{1/2},$$

$$\hat{\gamma}_1 = i(1) - \frac{(\hat{\gamma}_2 + 1)}{2},$$

and

$$\hat{p} = \frac{\hat{\mu}_0 l}{\hat{\gamma}_2}.$$

Other methods of selecting $\hat{a}$ can be used, but the one chosen above is very convenient and seems to produce satisfactory results in practice. There does not appear to be a simple way of using another equation,

TABLE I

REQUIRED STEPS TO FIT $f(t)$ BY THE METHOD OF MOMENTS

	1	2	3	4
$\hat{\mu}_0$	0.866920	0.148050	0.321000	0.242340
$\hat{\mu}_1$	5.311610	1.380660	3.066130	2.523180
$\hat{\mu}_2$	43.240090	13.818520	35.586030	30.365180
$i(1)$	7.067538	62.989755	29.756408	42.963339
$i(2)$	66.366593	4258.300410	1075.881509	2133.538826
$\hat{\gamma}_2$	10.638966	52.270613	43.922382	54.200773
$\hat{\gamma}_1$	1.248055	36.354449	7.295217	15.362953
$\hat{p}$	0.093994	0.019131	0.022767	0.018500
$E\{\hat{X}_1\}$	1.081964	5.382276	2.341765	3.723058
$E\{\hat{X}_2\}$	9.223132	7.738664	14.099085	13.135015

μ_3, to obtain a better fit. Notice that $E\{X_i\} = \gamma_i/a$ is an important parameter in determining the relative influences of each stage of the pasture cycle on $f(t)$. The observed and fitted $f(t)$ function for the four plots are shown in Fig. 2 and the appropriate steps in the calculations of $\hat{p}$, $\hat{\gamma}_1$, $\hat{\gamma}_2$ and $E\{\hat{X}_i\}$ are given in Table I.

DISCUSSION

In their original models, Tallis and Donald [6] proposed a negative binomial model for the distribution of fecal deposits on whole paddocks grazed by sheep. From investigations on the distribution of fecal deposits,

Donald and Leslie [4] concluded that, although the negative binomial provided a reasonable empirical description of the distributions, there was some evidence of departure from the model at a low rate of stocking. More importantly, however, the hypothesis that the distribution of feces is additive and independent with respect to time was found to be unacceptable. This finding arose from the tendency shown by flocks of sheep to deposit heavy concentrations of feces in the same circumscribed area of a paddock during consecutive resting periods. Donald and Leslie [4] concluded that, in the presence of heterogeneities of pasture and topography, the known tendency for subflock formation in large flocks and the different behavior patterns of different age classes of sheep, it seemed doubtful whether any simple two-parameter probability distribution could adequately describe the distribution of fecal deposits in all situations.

The present model overcomes these difficulties by using the general forms for the means and variances of the different processes and is therefore much more flexible. It has the added advantage that particular subareas of paddocks are easily considered separately. For example, there is some evidence that sheep do not graze on resting areas (or "camps") while these areas carry heavy concentrations of freshly deposited feces but may do so later when such areas have ceased to be used for resting [1, 4]. Thus, when potential rates of infective larval intake by grazing sheep are being considered, it may be necessary to derive separate estimates of infective larval abundance for grazing and resting areas, respectively.

Turning to the model for $f(t)$, this is a considerable advance over the original models of Tallis and Donald [6] in which this component was left to be estimated empirically. The parameters p, $E(X_1)$, and $E(X_2)$ have simple biological meaning, and it would not be difficult to design experiments to estimate them independently. The fit of the model to the four sets of data, shown in Fig. 2, appears reasonable, particularly if the error variance of the estimates, which is unavoidably rather large, is taken into account. Each empirical point estimate of $f(t)$ is the mean of four samples and an average standard error of .01 for each sample mean has been calculated.

During the course of the field studies from which the data used to illustrate the fitting of the $f(t)$ model were derived, some independent evidence was also obtained relating to $E(X_1)$, namely, the average time taken for development to the infective stage and migration to the herbage. When herbage samples were collected each week, the associated fecal material was also collected, and estimates were made of the numbers of surviving eggs and preinfective larvae which had not yet completed development to the infective stage and of surviving infective larvae which had not yet migrated from the feces. These data are presented in Table II in the

TABLE II

ESTIMATES OF THE AVERAGE TIME TAKEN FOR DEVELOPMENT TO THE INFECTIVE STAGE AND MIGRATION TO THE HERBAGE DERIVED FROM FITTING THE $f(t)$ MODEL COMPARED WITH ESTIMATES OF MAXIMUM DEVELOPMENT AND MIGRATION TIMES OBTAINED FROM EXAMINATION OF FECES

Plot	$E\{\hat{X}_1\}$ (weeks)	"Development completed,"[a] by week	"Migration completed,"[a] by week
1	1.08	2	3
3	2.34	2	6
4	3.72	2	13
2	5.38	6	10

[a] For explanation, see text.

form of times recorded for the completion of development and of migration for the majority of eggs originally present in the sample. The former estimate represents the first weekly sampling at which no viable pre-infective stages were recovered, and the latter is the first sampling at which no infective larvae were found in the feces. Because the estimate of $E(X_1)$ derived from the fitting of $f(t)$ is an average value whereas the estimates of development and migration times obtained from fecal samples represent maximum values, they are not directly comparable. However, inspection of Table II reveals at least that there are no gross anomalies.

The $f(t)$ model may be particularly useful in two situations. First, estimates of the parameters derived from fitting the model might constitute suitable dependent variables for the application of such techniques as multiple regression analysis against components of environment. Second, simulation studies with the whole model will be greatly facilitated if various forms of the function $f(t)$ can be generated by choosing parameter values appropriate to particular sets of climatic and other environmental conditions.

REFERENCES

1 H. D. Crofton, Nematode parasite populations in sheep on lowland farms. VI. Sheep behaviour and nematode infections. *Parasitology* 48(1958), 251–260.

2 A. D. Donald, Population studies on the infective stage of some nematode parasites of sheep. III. The distribution of strongyloid egg output in flocks of sheep, *Parasitology* 58(1968), 951–960.

3 A. D. Donald, Ecology of the free living stages of nematode parasites of sheep, *Australian Vet. J.* 44(1968), 139–144.

4 A. D. Donald and R. T. Leslie, Population studies on the infective stage of some nematode parasites of sheep. II. The distribution of faecal deposits on fields grazed by sheep, *Parasitology* 59(1969), 141–157.

5 G. M. Tallis, A note on the estimation of larval concentrations on pasture, *Australian J. biol. Sci.* 17(1964), 1016–1019.

6 G. M. Tallis and A. D. Donald, Models for the distribution on pasture of infective larvae of the gastrointestinal nematode parasites of sheep. *Australian J. biol. Sci.* 17(1964), 504–513.

A(b)[6]

A Deterministic Model for the Life Cycle of a Class of Internal Parasites of Sheep

G. GORDON, M. O'CALLAGHAN, AND G. M. TALLIS
Division of Mathematical Statistics
C.S.I.R.O.
Newtown, N.S.W., Australia

Communicated by K. E. F. Watt

ABSTRACT

The life cycle of the sheep parasite *Haemonchus contortus* has two phases: the development to maturity of ingested larvae in the sheep, and the development on the pasture of eggs excreted in feces. The development in time of a parasite population of several larval stages is discussed for each of these phases. A system of linear differential equations with constant coefficients links the two phases in a complete life cycle. The egg-laying behavior of the parasites and the immunity reaction of the sheep to the presence of parasites is also discussed, the treatment being deterministic throughout.

1. INTRODUCTION

The development in sheep of the parasite *Haemonchus contortus* [1–4] takes place in several stages. The parasite is ingested from the pasture in the third larval stage L3. It develops to the early and late fourth larval stages L4, and finally to the adult stage. Sexual differentiation occurs at the late L4 stage, while the female adult stage can be further divided into nonegg-laying, or immature, and egg-laying, or mature, stages. These observable stages form the basis of a deterministic compartmental model that has been used in a study of the parasite.

This model is given here and is extended to describe the egg-laying behavior and the retardation of the development of the parasite population due to immunity acquired by the host. On the pasture the parasite develops from the egg to the L3 stage. This development is given mathematical treatment and a model is derived for the complete life cycle.

Mathematical Biosciences 8 (1970), 209–226

2. PARASITE IN THE SHEEP

Suppose that the parasite population in a sheep at time t is described by the numbers $x_i(t)$, $i = 1, \ldots, 5$, of parasites in the five stages L3, early L4, late L4, immature adult, and mature adult, respectively. For simplicity we ignore for the moment the splitting of the population into sexed stages. We suppose that the instantaneous rate at which parasites leave a stage to enter the next is proportional to the number in the stage, and that the rate at which parasites in a given stage die is also proportional to the number in the stage; i.e., the number of parasites "skipping" from the ith stage to the $(i+1)$th stage in a small time interval δt is $\lambda_i x_i(t)\,\delta t$ and the number of parasites dying is $\mu_i x_i(t)\,\delta t$. The behavior of the system is then described by the differential equations

$$\begin{aligned}
\dot{x}_1 &= -\kappa_1 x_1, \\
\dot{x}_2 &= \lambda_1 x_1 - \kappa_2 x_2, \\
\dot{x}_3 &= \lambda_2 x_2 - \kappa_3 x_3, \\
\dot{x}_4 &= \lambda_3 x_3 - \kappa_4 x_4, \\
\dot{x}_5 &= \lambda_4 x_4 - \kappa_5 x_5.
\end{aligned} \tag{1}$$

Here $\kappa_i = \lambda_i + \mu_i$ and $\kappa_5 = \mu_5$.

These equations may be written in matrix form

$$\dot{x} = \mathbf{A}\mathbf{x} \tag{2}$$

where A is the matrix of coefficients in Eqs. (1). We may give the parameters another interpretation. Thus, from (2)

$$\mathbf{x} = \mathbf{A}^{-1}\dot{\mathbf{x}}.$$

So

$$\begin{aligned}
\int_0^\infty \mathbf{x}(t)\,dt &= \mathbf{A}^{-1}[\mathbf{x}(\infty) - \mathbf{x}(0)]. \\
&= -\mathbf{A}^{-1}\mathbf{x}(0)
\end{aligned}$$

since the κ_i are positive. $\mathbf{A}^{-1}$ may be evaluated (see Section 7), giving, for the case of a single initial dose of N L3 larvae,

$$\int_0^\infty x_1(t)\,dt = \frac{N}{\kappa_1}, \int_0^\infty x_2(t)\,dt = N\frac{\lambda_1}{\kappa_1\kappa_2}, \ldots, \int_0^\infty x_5(t)\,dt = N\frac{\lambda_1 \cdots \lambda_4}{\kappa_1 \cdots \kappa_5}.$$

Now the total number of worms entering the kth stage, $k > 1$, is

$$\lambda_{k-1}\int_0^\infty x_{k-1}(t)\,dt,$$

so the total number of worms ultimately entering the kth stage is $N(\lambda_1 \cdots \lambda_{k-1})/(\kappa_1 \cdots \kappa_{k-1})$. Thus λ_k/κ_k is the fraction of the total number of larvae entering the kth stage that survive to enter the $(k+1)$th stage, $k = 1, \ldots, 4$. The present model can be made stochastic, in which case λ_k/κ_k becomes the probability that an individual larva survives the kth stage.

The parasite population may be studied experimentally by giving the sheep a dose of N L3 larvae at time $t = 0$. The initial solution vector $\mathbf{x}(0)$ of (2) is then given by $x_1(0) = N$, $x_i(0) = 0$, $i = 2, \ldots, 5$, and the complete solution of (2) is

$$\mathbf{x}(t) = e^{\mathbf{A}t}\mathbf{x}(0) \tag{3}$$

$$= \mathbf{T}\,\mathrm{diag}[\exp(-\kappa_1 t), \ldots, \exp(-\kappa_5 t)]\mathbf{T}^{-1}\mathbf{x}(0) \tag{3a}$$

for some nonsingular T, provided the κ_i are distinct, Bellman [5].

If several doses are administered at intervals, or if there is continuous ingestion of L3 larvae at a known rate, the solution of (2) is

$$\mathbf{x}(t) = e^{\mathbf{A}t}\left(\mathbf{x}(0) + \int_0^t e^{-\mathbf{A}s}\,d\mathbf{W}(s)\right). \tag{4}$$

$\mathbf{W}(s)$ is the cumulative vector input function, given, in this case, by

$$\mathbf{W}'(s) = (W(s), 0, 0, 0, 0)$$

where $W(s)$ is the number of L3 larvae ingested in the time interval $(0, s]$. An explicit expression for the matrix exponential $e^{\mathbf{A}t}$ appearing in expressions (3), (4) is given in Lemma 1 of Section 7.

We could include another stage $x_0(t)$, the number of parasites that have died up to the time t, in order to fully describe the parasite population. Equation (2) would become

$$\dot{\mathbf{x}}^* = \mathbf{K}\mathbf{x}^* \tag{5}$$

where

$$\mathbf{K} = \begin{bmatrix} 0 & \boldsymbol{\mu}' \\ 0 & \mathbf{A} \end{bmatrix}, \quad \boldsymbol{\mu}' = (\mu_1, \ldots, \mu_5), \quad \text{and} \quad \mathbf{x}^* = \begin{bmatrix} x_0 \\ \mathbf{x} \end{bmatrix}.$$

The solution of (5) for the case of a single initial dose is $\mathbf{x}^*(t) = e^{\mathbf{K}t}\mathbf{x}^*(0)$,

where $\mathbf{x}^{*\prime}(0) = (0, \mathbf{x}'(0))$. However, since K is singular, the solution of (5) is easily obtained from (3) by means of the identity

$$x_0(t) = x_0(0) + \sum_{i=1}^{5} (x_i(0) - x_i(t)) + W(t).$$

The case where sexual differentiation occurs may be dealt with either by enlarging the system of equations (1) or, more conveniently, by considering two systems of equations similar to (2), one system for females or potential females, and one for males or potential males. Thus in the single initial dose case, the N L3 larvae dosed may be considered as N_0 potential females and N_1 potential males. Then the two systems may be written

$$\begin{aligned} \dot{\mathbf{x}}_0 &= \mathbf{A}_0\mathbf{x}_0 \quad \text{(females)}, \\ \dot{\mathbf{x}}_1 &= \mathbf{A}_1\mathbf{x}_1 \quad \text{(males)}, \end{aligned} \tag{6}$$

each equation with the appropriate initial vector. In (6) A_0, A_1 are matrices similar to A but with λ_i replaced by λ_{0i}, λ_{1i}, respectively, and so on. Equations (6) may also be written

$$\dot{\mathbf{y}} = \mathbf{Y}\mathbf{y} \tag{6a}$$

where

$$\mathbf{y} = \begin{bmatrix} \mathbf{x}_0 \\ \mathbf{x}_1 \end{bmatrix}, \qquad \mathbf{Y} = \mathbf{A}_0 \oplus \mathbf{A}_1 = \begin{bmatrix} \mathbf{A}_0 & \mathbf{0} \\ \mathbf{0} & \mathbf{A}_1 \end{bmatrix},$$

in which case

$$\mathbf{y} = \exp(\mathbf{Y}t)\mathbf{y}(0) = [\exp(\mathbf{A}_0 t) \oplus \exp(\mathbf{A}_1 t)]\mathbf{y}(0).$$

Since the L3 and early L4 stages are not sexually differentiated, we may set $\lambda_{01} = \lambda_{11}$, $\mu_{01} = \mu_{11}$, $\mu_{02} = \mu_{12}$; and since the male stages corresponding to the nonegg-laying and egg-laying adults cannot be differentiated, we can take the number of adult males to be $x_{14} + x_{15}$. Thus from the solutions to the two systems of (6) we can construct a solution vector $\mathbf{u}$ of length 7 of the combined male and female system:

$u_1 = x_{01} + x_{11}$,	number of L3 larvae,
$u_2 = x_{02} + x_{12}$,	number of early L4 larvae,
$u_3 = x_{03}$,	number of late L4 female larvae,
$u_4 = x_{04}$,	number of immature female adults,
$u_5 = x_{05}$,	number of mature or egg-laying female adults,
$u_6 = x_{13}$,	number of late L4 male larvae,
$u_7 = x_{14} + x_{15}$,	number of male adults.

The components of $\mathbf{u}$ correspond to the observable divisions of the parasite population.

3. EGG OUTPUT

A female larva that has been mature for t days is assumed to lay eggs at a rate $\varepsilon(t)$ eggs per day. We now find an expression for $\varepsilon(t)$ in terms of the observable function $E(t)$, the total number of eggs produced by the population up to time t.

At time t consider the worms in the fifth, or egg-producing stage in the age interval $(\tau, \tau + d\tau)$. These are the worms that entered the fifth stage in the time interval $(t - \tau - d\tau, t - \tau)$ and that survived till time t in the fifth stage. The number of worms entering is $\lambda_{04}x_{04}(t - \tau)\,d\tau$, and the number surviving is $\lambda_{04}x_{04}(t - \tau)\exp(-\mu_{05}\tau)\,d\tau$. Each worm in this age interval produces eggs at rate $\varepsilon(\tau)$, so the rate at which eggs are produced by mature worms in age interval $(\tau, \tau + d\tau)$ at time t is $\lambda_{04}\varepsilon(\tau)x_{04}(t - \tau)\exp(-\mu_{05}\tau)\,d\tau$. Integrating over all possible ages, we get for the rate of egg production for the whole population at time t

$$\dot{E}(t) = \lambda_{04}\int_0^t \varepsilon(\tau)x_{04}(t - \tau)\exp(-\mu_{05}\tau)\,d\tau.$$

Let

$$\varepsilon'(\tau) = (0, 0, 0, \varepsilon(\tau)) \quad \text{and} \quad z'(t) = (x_{01}(t), \ldots, x_{04}(t)).$$

Then the foregoing expression may be written

$$E^{(1)}(t) = \lambda_{04}\int_0^t \varepsilon'(\tau)z(t - \tau)\exp(-\mu_{05}\tau)\,d\tau$$

It may easily be seen that the first five derivatives of $E(t)$ are

$$E^{(k+1)}(t) = \lambda_{04}\int_0^t \varepsilon'(\tau)z^{(k)}(t - \tau)\exp(-\mu_{05}\tau)\,d\tau, \qquad k = 0, 1, 2, 3,$$

and

$$E^{(5)}(t) = \lambda_{04}\int_0^t \varepsilon'(\tau)z^{(4)}(t - \tau)\exp(-\mu_{05}\tau)\,d\tau + \lambda_{04}\varepsilon'(t)z^{(3)}(0)\exp(-\mu_{05}t).$$

But $z^{(k)}(t - \tau) = C^k z(t - \tau)$, where C is the submatrix of A_0 obtained by deleting the fifth row and the fifth column. Also, if the characteristic polynomial of C is $\sum_{k=0}^{4} p_k\lambda^k$, by the Cayley–Hamilton theorem,

$$\sum_{k=0}^{4} p_k C^k = 0.$$

Thus, since p_4 is 1

$$\sum_{k=0}^{4} p_k E^{(k+1)}(t) = \lambda_{04}\varepsilon'(t)\mathbf{z}^{(3)}(0)\exp(-\mu_{05}t)$$
$$= N_0\lambda_{01}\lambda_{02}\lambda_{03}\lambda_{04}\varepsilon(t)\exp(-\mu_{05}t),$$

using $\mathbf{z}^{(3)}(0) = \mathbf{C}^3\mathbf{z}(0)$.
That is,

$$\varepsilon(t) = \exp(\mu_{05}t)\left(N_0\prod_{i=1}^{4}\lambda_{0i}\right)^{-1}\sum_{k=0}^{4} p_k E^{(k+1)}(t).$$

Since the characteristic polynomial of $\mathbf{C}$ is $\Pi_{i=1}^{4}(\lambda + \kappa_{0i})$, the p_k are simply symmetrical polynomials in the κ_{0i}, $i = 1, \ldots, 4$.

4. THE CYCLE ON THE PASTURE

Assume that each egg takes a fixed time b_0 to reach the pasture, and let $\tau = t - b_0$, and let $e(t) = \dot{E}(t)$. Introducing the function $f(t)$, which is the proportion of eggs surviving as L3 larvae after a time t on the pasture, we find that the total number of live L3 larvae on the pasture is given by

$$L(\tau) = \int_0^{\tau} e(\tau - x)f(x)\,dx. \tag{7}$$

Note that if $L(\tau)$ and $e(t)$ are given, then (7) can be regarded as an integral equation of the *Faltung* type determining f. In fact,

$$f^*(s) = \frac{L^*(s)}{e^*(s)} \tag{8}$$

where

$$\theta^*(s) = \int_0^{\infty} e^{-sx}\theta(x)\,dx.$$

Equation (8) has practical implications for the situation where L and e can be observed with relative ease; however, this approach will not be pursued further here.

The function $f(t)$ is of some biological interest in its own right. A meaningful functional form for f has been derived and is presented elsewhere [6].

It is also possible to regard the parasites on the pasture in a way similar to that in which the parasites in the sheep have been dealt with. We can

consider the eggs and larval stages and form a model similar to that in Section 2 and $f(t)$ can then be obtained from the solution of the system of differential equations. $E(t)$ corresponds to the $W(t)$ of Section 2. This approach is pursued in the next section where, after some further simplifying assumptions, the whole life cycle is modeled.

The differential equation model for the parasite on the pasture includes all the larval stages. By the use of several survival functions, the more general approach can also be extended to several stages. Consider the extension of Eqs. (7) and (8) to k larval stages. Let $f_i(t)$ be the fraction of eggs dropped onto the pasture at time zero that are in the ith larval stage at time t. Then, if $\mathbf{f}'(t) = (f_1(t), \ldots, f_k(t))$,

$$\mathbf{L}(\tau) = \int_0^t e(\tau - x)\mathbf{f}(x)\, dx \tag{9}$$

where $\mathbf{L}'(t) = (L_1(t), \ldots, L_k(t))$ and $L_i(t)$ is the number of the ith-stage larvae on the pasture at time t. In obvious notation, (8) becomes

$$\mathbf{f}^*(s) = [e^*(s)]^{-1}\mathbf{L}^*(s). \tag{10}$$

5. THE LIFE CYCLE

Suppose we restrict attention now to the female system, and in addition to the five stages in the sheep, we consider four stages on the pasture, with assumptions similar to those concerning stages in the sheep. These stages represent eggs on the pasture and larval stages L1, L2, and L3 on the pasture.

The sheep-pasture system can then be described by the differential equations

$$\dot{\mathbf{x}} = \mathbf{B}\mathbf{x} \tag{11}$$

where $\mathbf{B}$ is a 9×9 matrix with $B_{i,i} = -\kappa_i$, $i = 1, \ldots, 9$, $B_{i+1,i} = \lambda_i$, $i = 1, \ldots, 8$, $B_{1,9} = \lambda_9$, with all other $B_{i,j} = 0$, $B_{i,j}$ denoting the element of $\mathbf{B}$ in the ith row and jth column.

Here it is assumed that the egg production rate is independent of adult age, and delay in eggs reaching the pasture is ignored, so that the rate at which eggs reach the pasture is proportional to the number of adult worms in the sheep. The constant of proportionality is λ_5. Similarly it is assumed that the rate of ingestion of L3 larvae is proportional to the number of L3 larvae on the pasture, the constant of proportionality being λ_9. The constants μ_i, $i = 1, \ldots, 9$, are defined as before with $\kappa_i = \lambda_i + \mu_i$, $i \neq 5$, $\kappa_5 = \mu_5$.

The solution of (11) may be written

$$\mathbf{x}(t) = \mathrm{T}\,\mathrm{diag}[\exp(\rho_1 t), \ldots, \exp(\rho_9 t)]\mathrm{T}^{-1}\mathbf{x}(0)$$

where the ρ_i are the characteristic roots of B assumed distinct, and T is a nonsingular matrix. From this it may be seen that if the characteristic roots ρ_i all have their real parts negative, then each element of $\mathbf{x}(t)$ tends to zero as $t \to \infty$. On the other hand there will be divergence as $t \to \infty$ if there is a characteristic root with positive real part.

The determinant of B is $\lambda_1\lambda_2 \cdots \lambda_9 - \kappa_1 \cdots \kappa_9$, and so the characteristic equation of B is

$$\prod_{i=1}^{9} (\theta + \kappa_i) - \prod_{i=1}^{9} \lambda_i = 0. \tag{12}$$

We may write Eq. (12) in the form

$$\phi(\theta) - a = 0 \tag{13}$$

where

$$\phi(\theta) = \prod_{i=1}^{9} (\theta + \kappa_i) \quad \text{and} \quad a = \prod_{i=1}^{9} \lambda_i.$$

Then it is easily seen that $\phi(\theta)$ is monotonic increasing in the interval $(-\kappa, \infty)$ where $\kappa = \min \kappa_i$. It can be seen that the largest real root of $\phi(\theta) - a = 0$ for $a > 0$ is a monotone increasing function of a and can be obtained continuously from the root $-\kappa$ of $\phi(\theta) = 0$ by continuously varying a. If it is supposed that $-\kappa$ is a simple root of $\phi(\theta) = 0$, which will certainly be so if all the κ_i are distinct, then it can be shown that this largest real root of $\phi(\theta) - a = 0$ for $a > 0$ is also the root with largest real part [9]. Thus when $a < \phi(0)$, that is,

$$\prod_{i=1}^{9} \lambda_i < \prod_{i=1}^{9} \kappa_i,$$

the system (11) is stable since the largest real root is negative. When

$$\prod_{i=1}^{9} \lambda_i = \prod_{i=1}^{9} \kappa_i,$$

Eq. (12) has a zero root, while the remaining roots have negative real parts. In this case, as $t \to \infty$, $\mathbf{x}(t)$ tends to the steady-state solution satisfying $\mathrm{B}\mathbf{x} = \mathbf{0}$. Finally, if

$$\prod_{i=1}^{9} \lambda_i > \prod_{i=1}^{9} \kappa_i,$$

the equation has a positive root, and the system (11) is unstable; that is, the solution diverges to infinity. Thus the largest real root of (12) determines the asymptotic behavior of the system (11). Note that the stability

condition

$$\prod_{i=1}^{9} \lambda_i < \prod_{i=1}^{9} \kappa_i$$

is symmetrical with respect to each stage, and that we already know $\lambda_i \leqslant \kappa_i$, $i \neq 5$.

Perturbation Solution of Equations

Suppose we set $B = A + \lambda_9 D$, where A is a 9×9 version of the A in Section 2 and D is a matrix all of whose elements are zero except that $D_{19} = 1$. Then the solution matrix e^{Bt} can be found in terms of the known matrix e^{At} as an expansion in powers of λ_9. This expansion is [7]

$$\exp(Bt) = \exp(At) + \lambda_9 \int_0^t \exp[A(t-s)]D \exp(As)\, ds$$

$$+ \lambda_9^2 \int_0^t \exp[A(t-s)]D \left\{ \int_0^s \exp[A(s-u)]D \exp(Au)\, du \right\} ds + \cdots.$$

For t fairly small the first-order term should be adequate, as λ_9 will be small. If $X_{i,j}(t)$ is the i, j element of $X(t) = e^{At}$, then the coefficient of λ_9 is a matrix W where

$$W_{i,j} = \int_0^t X_{i1}(t-s) X_{9j}(s)\, ds.$$

This may be explicitly evaluated from the expression for e^{At} using Lemma 2 of Section 7.

6. IMMUNE REACTION

In response to antigenic information transmitted to it from the parasite, the host exhibits an immune reaction to the parasite's presence. To model this, we suppose that the sheep possesses an information dam into which information flows at different rates from the larvae in different stages. Let us say that information flows into this dam at a rate α_j from a jth-stage female or α_{j+5} from a jth-stage male, $j = 1, \ldots, 5$. Also, to account for the loss of immunity during nonexposure, let the information in the dam leak away at a rate ρI proportional to its quantity I. We then have

$$\dot{I} + \rho I = \sum_{i=1}^{10} \alpha_i y_i = \alpha' \mathbf{y} \tag{14}$$

where $\alpha' = (\alpha_1, \ldots, \alpha_{10})$ and y is as defined in Section 2. The immune reaction is assumed to take effect at a time b_1 after I exceeds a threshold level Q, and to lose effect at a time b_2 after I falls below Q. The effect of the immune reaction is to alter the parameters in the matrix Y. Broadly, it is expected that the λ_{0j}, λ_{1j} will be reduced and the μ_{0j}, μ_{1j} increased. In particular, λ_{01}, λ_{14} should be very small while the immune reaction is operating. In this way, the adult population is kept constant while there is a buildup of larvae in the preceding stage. The observable effect of the immune reaction is a sharp leveling of the egg output rate.

If I_0 is the quantity of information in the dam at $t = 0$, then the solution of (14) is

$$I(t) = \left(I_0 + \int_0^t e^{\rho u}\alpha' \mathbf{y}(u)\, du\right) e^{-\rho t}$$

$$= \left(I_0 + \alpha' \int_0^t e^{\rho u}\mathbf{y}(u)\, du\right) e^{-\rho t}. \tag{15}$$

But

$$\int_0^t e^{\rho u}\mathbf{y}(u)\, du = \int_0^t e^{\rho u}\mathbf{Y}^{-1}\dot{\mathbf{y}}(u)\, du$$

$$= \mathbf{Y}^{-1}\int_0^t e^{\rho u}\dot{\mathbf{y}}(u)\, du,$$

which, on integration by parts, becomes

$$\int_0^t e^{\rho u}\mathbf{y}(u)\, du = \mathbf{Y}^{-1}\left[e^{\rho t}\mathbf{y}(t) - \mathbf{y}(0) - \rho\int_0^t e^{\rho u}\mathbf{y}(u)\, du\right];$$

that is,

$$(\mathbf{I} + \rho\mathbf{Y}^{-1})\int_0^t e^{\rho u}\mathbf{y}(u)\, du = \mathbf{Y}^{-1}(e^{\rho t}\mathbf{y}(t) - \mathbf{y}(0)),$$

$$\int_0^t e^{\rho u}\mathbf{y}(u)\, du = (\mathbf{Y} + \rho\mathbf{I})^{-1}(e^{\rho t}\mathbf{y}(t) - \mathbf{y}(0)).$$

Therefore, from (15),

$$I(t) = e^{-\rho t}I_0 + e^{-\rho t}\alpha'(\mathrm{Y} + \rho\mathrm{I})^{-1}(e^{\rho t}\mathrm{y}(t) - \mathrm{y}(0))$$
$$= \alpha'(\mathrm{Y} + \rho\mathrm{I})^{-1}\mathrm{y}(t) + e^{-\rho t}[I_0 - \alpha'(\mathrm{Y} + \rho\mathrm{I})^{-1}\mathrm{y}(0)].$$

In order to estimate the α and ρ we must employ a sequential dosing scheme. Suppose that hosts are dosed with M first-stage larvae at intervals of length Δ, and that initially each host's dam is empty. Let the information present in the dam at time t be $J(t)$. Then

$$J(t) = \frac{M}{N}\left(I(t) + I(t-\Delta) + \cdots + I\left(t - \Delta\left[\frac{t}{\Delta}\right]\right)\right).$$

where $[x]$ denotes the integer part of x.

When the egg output rate is observed to level off at time $t_0 + b_1$, let the parasites be removed and the hosts allowed to rest for various periods of time. Then $J(t_0) = Q$ and after n days rest, the host has a quantity $J(t_0 + b_1)e^{-\rho n}$ of information in the dam. So, if the dosing scheme is recommenced at this time and the egg output rate levels off after further time $t_1 + b_1$, we have

$$Q = J(t_0) = J(t_0 + b_1)\exp[-\rho(n + t_1)] + J(t_1). \tag{16}$$

Here we suppose that $b_2 < n + t_1$, so that the immunity mechanism is triggered again by the second infection. Since we have twelve parameters to estimate in this way, we must have twelve different rest periods. However, it might be desirable to set the α_i corresponding to indistinguishable stages equal. Thus we might set $\alpha_1 = \alpha_6$, $\alpha_2 = \alpha_7$, $\alpha_9 = \alpha_{10}$. Then nine independent equations like (16) estimate α, ρ, and b_1. The foregoing scheme may be modified to estimate b_2 if b_2 is large, but cannot find b_2 if b_2 is small. This is because it takes several days for a second infection to mature.

It is likely that the condition

$$\prod_{i=1}^{9}\lambda_i < \prod_{i=1}^{9}\kappa_i$$

of Section 5 will hold when the immune reaction is operative and fail otherwise. Thus, we envisage the evolution of the host–parasite relationship as a process of the parameters gradually varying until the immune reaction and climatic effects on the pasture stages determine an approximate non-trivial steady-state situation for the complete life cycle.

7. APPENDIX

Inverse of Matrix

The inverse of the 5×5 matrix A required in Section 2 is

$$\mathbf{A}^{-1} = \begin{bmatrix} -\frac{1}{\kappa_1} & & & & \\ -\frac{\lambda_1}{\kappa_1\kappa_2} & -\frac{1}{\kappa_2} & & & \\ -\frac{\lambda_1\lambda_2}{\kappa_1\kappa_2\kappa_3} & -\frac{\lambda_2}{\kappa_2\kappa_3} & -\frac{1}{\kappa_3} & & \\ -\frac{\lambda_1\lambda_2\lambda_3}{\kappa_1\kappa_2\kappa_3\kappa_4} & -\frac{\lambda_2\lambda_3}{\kappa_2\kappa_3\kappa_4} & -\frac{\lambda_3}{\kappa_3\kappa_4} & -\frac{1}{\kappa_4} & \\ -\frac{\lambda_1\lambda_2\lambda_3\lambda_4}{\kappa_1\kappa_2\kappa_3\kappa_4\kappa_5} & -\frac{\lambda_2\lambda_3\lambda_4}{\kappa_2\kappa_3\kappa_4\kappa_5} & -\frac{\lambda_3\lambda_4}{\kappa_3\kappa_4\kappa_5} & -\frac{\lambda_4}{\kappa_4\kappa_5} & -\frac{1}{\kappa_5} \end{bmatrix}.$$

The same expression gives the inverse $(\mathbf{A} + \rho\mathbf{I})^{-1}$ on replacing κ_i by $\kappa_i - \rho$. This in turn gives $(\mathbf{Y} + \rho\mathbf{I})^{-1}$ required in Section 6, since

$$(\mathbf{Y} + \rho\mathbf{I}_{10})^{-1} = (\mathbf{A}_0 + \rho\mathbf{I}_5)^{-1} \oplus (\mathbf{A}_1 + \rho\mathbf{I}_5)^{-1}.$$

Solution of System of Differential Equations

LEMMA 1. *Let* $\mathbf{X}(t)$ *be a matrix whose elements are functions of* t, *satisfying*

$$\frac{d}{dt}\mathbf{X}(t) = \mathbf{A}\mathbf{X}(t), \qquad \mathbf{X}(0) = \mathbf{I},$$

where $\mathbf{I}$ *is the* $n \times n$ *identity matrix, and* $\mathbf{A}$ *is an* $n \times n$ *matrix with elements*

$$A_{i,i} = -\kappa_i, \qquad i = 1, \ldots, n,$$
$$A_{i+1,i} = \lambda_i, \qquad i = 1, \ldots, n-1,$$

and all other

$$A_{i,j} = 0.$$

Then

$$X_{j,l}(t) = \left(\prod_{i=l}^{j-1} \lambda_i\right) \cdot \sum_{i=l}^{j} \left[\frac{\exp(-\kappa_i t)}{\prod_{l \leqslant \nu \leqslant j, \nu \neq i}(\kappa_\nu - \kappa_i)}\right] \tag{17}$$

provided that no two κ_i *are equal.*

(Note that $\Sigma_{i\in\Lambda} f_i = 0$ and $\Pi_{i\in\Lambda} f_i = 1$ if Λ is empty, so that (17) is valid for $l \geqslant j$.)

Proof. $\mathfrak{L}\{X(t)\} = \mathfrak{L}\{AX(t)\}$ where $\mathfrak{L}$ denotes the Laplace transform, $\mathfrak{L}\{f(t)\} = \int_0^\infty f(t)e^{-\theta t}\,dt$. Also

$$\mathfrak{L}\{\dot{X}(t)\} = \theta\mathfrak{L}\{X(t)\} - I;$$

therefore

$$(A - \theta I)\mathfrak{L}\{X(t)\} = -I,$$

so

$$\mathfrak{L}\{X(t)\} = -(A - \theta I)^{-1}.$$

Thus

$$\mathfrak{L}\{X(t)_{j,l}\} = \frac{\prod_{i=l}^{j-1} \lambda_i}{\prod_{i=l}^{j}(\theta + \kappa_i)},$$

from the foregoing expression for A^{-1}. Expanding the right-hand side in partial fractions (assuming the κ_i are distinct), let

$$\left[\prod_{i=l}^{j}(\theta + \kappa_i)\right]^{-1} = \sum_{i=l}^{j} \frac{a_i}{(\theta + \kappa_i)}.$$

Then it is easily seen that

$$a_i = \left[\prod_{\substack{\nu=l \\ \nu \neq i}}^{j} (\kappa_\nu - \kappa_i)\right]^{-1}.$$

Since $\mathfrak{L}^{-1}[(\theta + \kappa_i)^{-1}] = \exp(-\kappa_i t)$, we find that

$$X(t)_{j,l} = \prod_{i=l}^{j-1} \lambda_i \cdot \sum_{i=l}^{j} \frac{\exp(-\kappa_i t)}{\prod_{\substack{\nu=l \\ \nu \neq i}}^{j}(\kappa_\nu - \kappa_i)}.$$

The solution may be also verified by direct substitution.

Another Representation of the Solution

For an arbitrary function $f(x)$ the nth-order divided difference [8] of $f(x)$ with respect to the distinct numbers $x_0, \ldots, x_n$ is defined by

$$f[x_0, \ldots, x_n] = \sum_{i=0}^{n} \frac{f(x_i)}{\prod_{\substack{j=0 \\ j \neq i}}^{n}(x_i - x_j)}, \qquad n = 1, 2, \ldots,$$

$$f[x_0] = f(x_0).$$

A divided difference is a symmetrical function of its arguments; permutation of the arguments leaves its value unaltered. It may be shown [8,

page 250] that if $f(x)$ has a continuous nth derivative, then

$$f[x_0, \ldots, x_n] = \int_0^1 dt_1 \int_0^{t_1} dt_2 \cdots \int_0^{t_{n-1}} dt_n \times f^{(n)}(t_n[x_n - x_{n-1}] + \cdots + t_1[x_1 - x_0] + x_0) \quad (18)$$

provided $n \geqslant 1$. If we set $f_t(\kappa) = e^{\kappa t}$, Eq. (17) may be written

$$X_{j,l}(t) = \left\{\prod_{i=l}^{j-1} \lambda_i\right\} \cdot f_t[-\kappa_l, \ldots, -\kappa_j], \qquad j \geqslant l. \quad (19)$$

The nth derivative of $f_t(\kappa)$ is

$$f_t^{(n)}(\kappa) = t^n e^{\kappa t}$$

so using (18), an expression for $X_{j,l}(t)$ may be obtained as a multiple integral. This may be shown to be the same as the solution obtained by successive integration of the differential equations, or the solution in convolution form obtained by inverting the Laplace transform of $X_{j,l}$ that appears in the proof of Lemma 1 as a product of factors. The integral form of the solution is valid when the κ_i are not all distinct; Eq. (18) completes the definition of a divided difference by continuity.

Recurrence Relationship for Solution

An nth-order divided difference satisfies the recurrence relation

$$f[x_0, \ldots, x_n] = \frac{f[x_0, \ldots, x_{n-1}] - f[x_1, \ldots, x_n]}{x_0 - x_n}$$

provided $n \geqslant 1$ and the x_i are distinct. Thus an nth-order divided difference may be conveniently calculated in tabular form (see [8, page 249]). With κ's as arguments, and $f_t(\kappa) = e^{\kappa t}$ as the function, after multiplication by the appropriate λ's, such a table yields all the elements of the lower triangular matrix e^{At}. When the κ_i are not distinct, we may use the fact that the foregoing recurrence relation is valid if $f^{(n)}(x)$ is continuous and $x_0 \neq x_i$, $i = 1, \ldots, n$, even though the x_i, $i = 1, \ldots, n$, are not all istdinct. Note also that a kth-order divided difference with all of its $k + 1$ arguments equal is given by

$$f[x, \ldots, x] = \frac{f^{(k)}(x)}{k!}.$$

Thus the tabular form of calculation may be modified to give a divided difference with some arguments equal if the equal arguments are permuted so that they are adjacent to one another. Permutation of the arguments in

calculating (19), however, means that all the elements of e^{At} will not be obtained from the one table, as they are when the κ_i are distinct. Note that if the κ_i are not distinct, the diagonal matrix of (3a) in Section 2 must be replaced by a Jordan canonical matrix. Also note the general identity $\exp[A(t+s)] = \exp(At) \cdot \exp(As)$, which may simplify calculations in some cases.

LEMMA 2

$$f_s[-\theta_1, \ldots, -\theta_n] = \int_0^s f_{s-x}[-\theta_1, \ldots, -\theta_k] f_x[-\theta_{k+1}, \ldots, -\theta_n]\, dx.$$

Proof. This result follows immediately from the identity

$$\left[\prod_{i=1}^{n} (\theta + \kappa_i)\right]^{-1} = \left[\prod_{i=1}^{k} (\theta + \kappa_i)\right]^{-1} \cdot \left[\prod_{i=k+1}^{n} (\theta + \kappa_i)\right]^{-1}$$

using Laplace transforms (cf. proof of Lemma 1).

Lemma 2 may be used to evaluate the integrals

$$W_{i,j} = \int_0^t X_{i1}(t-s) X_{9j}(s)\, ds$$

appearing as elements in the matrix coefficient of λ_9 in the perturbation expansion of e^{Bt} (Section 5). It also gives the elements of the matrix coefficients of the higher powers of λ_9 in this expansion. Each element may be expressed as a product of a divided difference of $e^{\kappa t}$ with some arguments repeated and a product of λ's.

LEMMA 3. *The solution of equations*

$$\dot{\mathbf{X}}(t) = \mathrm{B}\mathbf{X}(t), \qquad \mathbf{X}(0) = \mathbf{I}$$

where B *is as in Section* 5 *is given by*

$$X_{j,l}(t) = c_{j,l} \sum_{i=1}^{9} \frac{\phi_{j,l}(\rho_i)}{f'(\rho_i)} \exp(\rho_i t)$$

where

$$c_{j,l} = \prod_{i=l}^{j-1} \lambda_i \quad \text{and} \quad \phi_{j,l}(0) = \prod_{\substack{i<l \\ i>j}} (\kappa_i + 0) \qquad \text{when } j > l,$$

$$c_{j,l} = 1 \quad \text{and} \quad \phi_{j,l}(0) = \prod_{i \neq j} (\kappa_i + 0) \qquad \text{when } j = l,$$

$$c_{j,l} = \prod_{\substack{i \geq l \\ i<j}} \lambda_i \quad \text{and} \quad \phi_{j,l}(0) = \prod_{i=j+1}^{l-1} (\kappa_i + 0) \qquad \text{when } l > j,$$

and

$$f(\theta) = \prod_{i=1}^{9} (\kappa_i + \theta);$$

and ρ_i, $i = 1, \ldots, 9$, *are the characteristic roots (assumed distinct) of* B, *that is, the roots of* $f(\theta) - \Pi_{i=1}^{9} \lambda_i = 0$.

Proof. Set $\mathrm{B} = \Lambda - \mathrm{K}$ where

$$\Lambda = \begin{bmatrix} 0 & \cdots\cdots & \lambda_9 \\ \lambda_1 & & \\ & \ddots & \\ & & \\ & & \\ & \lambda_8 & 0 \end{bmatrix} \quad \text{and} \quad \mathrm{K} = \begin{bmatrix} \kappa_1 & & \\ & \ddots & \\ & & \\ & & \\ & & \kappa_9 \end{bmatrix}.$$

Then $\mathcal{L}\{\mathrm{X}(t)\} = -(\mathrm{B} - \theta\mathrm{I})^{-1}$ (cf. Lemma 1). We calculate B^{-1}.

$$\mathrm{B} = \Lambda - \mathrm{K} = (\Lambda\mathrm{K}^{-1} - \mathrm{I})\mathrm{K} = -(\mathrm{I} - \Lambda\mathrm{K}^{-1})\mathrm{K},$$

so

$$\mathrm{B}^{-1} = -\mathrm{K}^{-1}(\mathrm{I} - \Lambda\mathrm{K}^{-1})^{-1}. \text{ Set } \mathrm{M} = \Lambda\mathrm{K}^{-1}. \text{ Then}$$

$$\mathrm{M} = \begin{bmatrix} 0 & 0 & & & m_9 \\ m_1 & 0 & & & 0 \\ 0 & m_2 & \ddots & & \vdots \\ & \ddots & \ddots & \ddots & \vdots \\ & & \ddots & \ddots & \vdots \\ & & & \ddots & \vdots \\ 0 & & 0 & m_8 & 0 \end{bmatrix}$$

where $m_i = \lambda_i/\kappa_i$, $i = 1, \ldots, 9$, and $\mathrm{B}^{-1} = -\mathrm{K}^{-1}(\mathrm{I} + \mathrm{M} + \mathrm{M}^2 + \cdots)$, provided the series converges. It does if $m = m_1 \cdots m_9 < 1$ for in that situation the characteristic equation of M is $m_1 \cdots m_9 - \theta^9 = 0$, so that by the Cayley–Hamilton theorem $\mathrm{M}^9 = m\mathrm{I}$. Suppose $m < 1$; that is,

$$\prod_{i=1}^{9} \lambda_i < \prod_{i=1}^{9} \kappa_i;$$

then

$$\mathbf{B}^{-1} = -\mathbf{K}^{-1}[(1 + m + m^2 + \cdots)\mathbf{I} + (1 + m + m^2 + \cdots)\mathbf{M} + \cdots + (1 + m + m^2 + \cdots)\mathbf{M}^s]$$
$$= -\frac{1}{1-m}\mathbf{K}^{-1}\mathbf{N} \qquad \text{where } \mathbf{N} = \mathbf{I} + \mathbf{M} + \cdots + \mathbf{M}^s.$$

It is easily found that

$$N_{j,l} = \begin{cases} \prod_{i=l}^{j-1} m_i & \text{if } j > l, \\ 1 & \text{if } j = l, \\ \dfrac{m}{\prod_{i=j}^{l-1} m_i} & \text{if } l > j; \end{cases}$$

while

$$B^{-1}_{j,l} = -[(1-m)\kappa_j]^{-1} N_{j,l}.$$

Clearly the same holds for $m > 1$ by algebraic identity.

If we replace κ_i by $\kappa_i + \theta$ in this expression, we immediately obtain

$$\mathfrak{L}\{\mathbf{X}(t)_{j,l}\} = \begin{cases} \prod_{i=l}^{j-1} \lambda_i \cdot \dfrac{\prod_{i<l,\, i>j}(\kappa_i + \theta)}{[\prod_{i=1}^{9}(\kappa_i + \theta) - \prod_{i=1}^{9}\lambda_i]}, & j > l, \\ \dfrac{\prod_{i \neq j}(\kappa_i + \theta)}{[\prod_{i=1}^{9}(\kappa_i + \theta) - \prod_{i=1}^{9}\lambda_i]}, & j = l, \\ \prod_{\substack{i \geqslant l \\ i < j}} \lambda_i \cdot \dfrac{\prod_{i=j+1}^{l-1}(\kappa_i + \theta)}{[\prod_{i=1}^{9}(\kappa_i + \theta) - \prod_{i=1}^{9}\lambda_i]}, & l > j. \end{cases}$$

This yields the stated result upon expanding the right-hand side in partial fractions and taking the inverse Laplace transform.

REFERENCES

1 J. K. Dineen, A. D. Donald, B. M. Wagland, and J. Offner, The dynamics of the host-parasite relationship, III: The response of sheep to primary infection with *Haemonchus contortus*, *Parasitology* 55(1965), 515–525.

2 J. K. Dineen and B. M. Wagland, The dynamics of the host-parasite relationship, IV: The response of sheep to graded and to repeated infection with *Haemonchus contortus*, *Parasitology* 56(1966), 639–650.

3 J. K. Dineen and B. M. Wagland, The dynamics of the host-parasite relationship, V: Evidence for immunological exhaustion in sheep experimentally infected with *Haemonchus contortus*, *Parasitology* 56(1966), 665–677.

4 B. M. Wagland and J. K. Dineen, The dynamics of the host-parasite relationship, VI: Regeneration of the immune response in sheep infected with *Haemonchus contortus*, *Parasitology* 57(1967), 59–65.
5 R. Bellman, *Stability theory of differential equations*. McGraw-Hill, New York, 1953.
6 G. M. Tallis and A. D. Donald, Further models for the distribution on pasture of infective larvae of the strongyloid nematode parasites of sheep. *Math. Bioscience*. 7(1970), 179–190.
7 R. Bellman, *Perturbation techniques in mathematics, physics and engineering*. Holt, Rinehart and Winston, New York, 1964.
8 E. Isaacson and H. B. Keller, *Analysis of numerical methods*. Wiley, New York, 1966.
9 G. M. Tallis and G. Gordon, A note on the roots of the polynomial equation $f(x) = a$ with reference to stability, II. Submitted for publication, 1969.

A (b)[7]

Some Stochastic Extensions to a Deterministic Treatment of Sheep Parasite Cycles

G. M. TALLIS
Division of Mathematical Statistics, C.S.I.R.O., Newtown, Australia

Communicated by K. E. F. Watt

ABSTRACT

This article provides a stochastic version of a deterministic model for the life cycle of gastrointestinal nematode parasites of sheep. This cycle is conveniently divided into a part that involves the sheep and a part that involves the pasture. Each of these sections of the cycle is described by a compartmental model, and the two parts of the model are then integrated into a single stochastic model. The asymptotic behavior of the final model is found to be different from that of its deterministic analogue.

INTRODUCTION

This article provides elementary stochastic extensions to some of the results in a paper by Gordon *et al.* [1] giving a deterministic, mathematical treatment of the life cycle of certain internal parasites of sheep. The notation of [1] is followed closely and the description of the problem is not repeated here. Numerous equations in [1] are referred to, and the convention [1, (j)] to represent the jth equation of that paper is used. It is intended that [1] be read in conjunction with this article.

The life cycle of the female parasite inside the sheep is modeled below. Thus, a stochastic analogue of [1, (2)] is given. Enlargement of the treatment to cope with two sexes of the parasite can be achieved in the same way as in [1, (6), (6a)]. An analogue to the complete sheep–pasture cycle structure, [1, (11)], is discussed briefly. However, no attention is given to models for the immune reaction, since this topic is somewhat tangential and, in any case, warrants separate consideration.

Most of the main results of interest can be obtained as a special case of the general n-compartment stochastic model. It is convenient to present this theory first and to subsequently particularize it to the case in point.

Mathematical Biosciences 8 (1970), 131–135

THE n-COMPARTMENT STOCHASTIC MODEL

Consider a system with n compartments $\Pi_1, \ldots, \Pi_n$ and let $X(t)$ be the random vector giving the number of individuals in each compartment at time t. The components of $X(t)$ are written as $x_i(t)$, $i = 1, 2, \ldots, n$, and $\sum_{i=1}^{n} x_i(0) = N$. Assume the usual linear transition rates as follows. The probability that a member of Π_i moves to Π_j in the time interval Δt is $\beta_{ij} x_i(t)\,\Delta t + o(\Delta t)$, $\beta_{ii} \equiv 0$. The probability of two changes of state is $o(\Delta t)$. If $p(\mathrm{x}, t)$ is the probability that $X(t) = \mathrm{x}$, then

$$p(\mathrm{x}, t + \Delta t) = \Bigg[\sum_{i=1}^{n}\sum_{j=1}^{n}\beta_{ij}(x_i + 1)\,\Delta t\, E_i E_j^{-1} + \left(1 - \sum_{i=1}^{n}\sum_{j=1}^{n}\beta_{ij} x_i \Delta t\right)\Bigg] p(\mathrm{x}, t) + o(\Delta t) \quad (1)$$

where $E_i E_j^{-1} f(\mathrm{x}) = f(x_1, \ldots, x_i + 1, \ldots, x_j - 1, \ldots, x_n)$. Clearly (1) leads to the set of differential equations

$$p'(\mathrm{x}, t) = \sum_{i=1}^{n}\sum_{j=1}^{n}\beta_{ij}[(x_i + 1)E_i E_j^{-1} - x_i]p(\mathrm{x}, t). \quad (2)$$

Now define

$$P(\mathrm{s}, t) = \sum_{\text{all } x_i} p(\mathrm{x}, t)\prod_{i=1}^{n} s_i^{x_i}$$

by letting $p(\mathrm{x}, t) = 0$ if $\mathrm{x} \notin \{\mathrm{x} \mid x_i \geqslant 0, \sum_{i=1}^{n} x_i = N\}$. If both sides of (2) are multiplied by $\prod_{i=1}^{n} s_i^{x_i}$ and summed, it can be verified that $P(\mathrm{s}, t)$ satisfies

$$\frac{\partial P(\mathrm{s}, t)}{\partial t} = \sum_{i=1}^{n}\sum_{j=1}^{n}\beta_{ij}(s_j - s_i)\frac{\partial P(\mathrm{s}, t)}{\partial s_i}. \quad (3)$$

Let the matrix K have elements $k_{ij} = \beta_{ji}$, $i \neq j$, $k_{ii} = -\sum_{j=1}^{n}\beta_{ij}$ and set $\mathrm{Y}(t) = e^{\mathrm{K}t}$, where $\mathrm{Y}(t)$ has columns $\mathrm{Y}_j(t)$, $\mathrm{Y}_j'(t) = (y_1^{(j)}(t), \ldots, y_n^{(j)}(t))$. Then the solution to (2) subject to the initial condition $P(\mathrm{s}, 0) = \prod_{i=1}^{n} s_i^{x_i(0)}$ is

$$P(\mathrm{s}, t) = \prod_{j=1}^{n}\left[\sum_{i=1}^{n} y_i^{(j)}(t)s_i\right]^{x_j(0)}. \quad (4)$$

To see this, notice first that $\mathrm{Y}(0) = \mathrm{I}_n$ and hence the initial condition is satisfied. Moreover, it is sufficient to verify that a typical factor of (4) satisfies (3). Dropping the superscript, we have

$$\sum_{i=1}^{n} y_i'(t)s_i = \sum_{i,j}\beta_{ij}(s_j - s_i)y_i(t)$$

and equating coefficients of s_i gives

$$y_i'(t) = -\sum_{j=1}^{n} \beta_{ij} y_i(t) + \sum_{j=1}^{n} \beta_{ji} y_j(t), \qquad i = 1, 2, \ldots, n.$$

That is, $\mathrm{y}(t)$ must satisfy

$$\dot{\mathrm{y}}(t) = \mathrm{K}\mathrm{y}(t),$$

which it does since $\dot{\mathrm{Y}}(t) = \mathrm{K}\mathrm{Y}(t)$.

Each element of $\mathrm{Y}(t)$ is nonnegative since $k_{ij} \geqslant 0$, $i \neq j$, and if $\mathbf{1}$ is a vector of n ones, $\mathbf{1}\dot{\mathrm{Y}}(t) = \mathbf{1}'$. This follows because $\mathbf{1}\dot{\mathrm{Y}}(t) = \mathbf{1}'\mathrm{K}\mathrm{Y}(t) = 0'$ and hence $\mathbf{1}'\mathrm{Y}(t) = \mathrm{C}'$ where C is a vector of constants. But $\mathrm{Y}(0) = \mathrm{I}_n$ and hence $\mathrm{C} = \mathbf{1}$. From this it follows that $y_i^{(j)}(t) \geqslant 0$ and $\sum_{i=1}^{n} y_i^{(j)}(t) = 1$ for $t \geqslant 0$.

It is interesting to notice that (4) represents the convolution of n multinomial distributions. Hence $\mu(t) = E\{\mathrm{X}(t)\} = \mathrm{Y}(t)\mathrm{x}(0)$ and $E\{(\mathrm{X}(t) - \mu(t))(\mathrm{X}(t) - \mu(t))'\} = \mathrm{V}(t)$ has elements

$$\begin{aligned} v_{ij}(t) &= -\sum_{k=0}^{n} x_k(0) y_i^{(k)}(t) y_j^{(k)}(t), \qquad i \neq j, \\ v_{ii}(t) &= \sum_{k=0}^{n} x_k(0) y_i^{(k)}(t)(1 - y_i^{(k)}(t)). \end{aligned} \tag{5}$$

This covariance matrix completely specifies the second-order properties of the process since

$$\mathrm{C}(s, t) = E\{(\mathrm{X}(t) - \mu(t))(\mathrm{X}(s) - \mu(s))'\} = \mathrm{V}(\min s, t). \tag{6}$$

The waiting time T_j in the jth compartment has the exponential distribution with parameter $k_{jj} = -\sum_{k=1}^{n} \beta_{jk}$. This provides an additional and useful interpretation of the parameters.

THE PARASITE CYCLE IN THE SHEEP

Define the matrix K as in [1, (5)]; then K defines a set of transition rates for a six-compartment model for death, the three larval stages, immature, and mature female worms; see the introduction of [1]. The numbers in the various stages, therefore, have probabilities specified by (4) with $n = 6$ and $\mathrm{Y}(t) = e^{\mathrm{K}t}$. Note that $\mathrm{x}(0)$ is usually of the form $\mathbf{x}'(0) = [0, N, 0, 0, 0, 0]$, and hence

$$P(\mathbf{s}, t) = \left[\sum_{i=1}^{6} y_i^{(2)}(t) s_i\right]^N.$$

The computation of $\mathrm{Y}(t)$ can be facilitated by special results in [1].

Since

$$\mathbf{K} = \begin{bmatrix} 0 & \boldsymbol{\mu}' \\ 0 & \mathbf{A} \end{bmatrix}$$

where A is defined by [1, (2)] and $\boldsymbol{\mu}$ is the vector of death rates, if $L\{\mathbf{Y}(t)\}$ is the Laplace transform of $\mathbf{Y}(t)$, it follows from the relationship $\dot{\mathbf{Y}}(t) = \mathbf{K}\mathbf{Y}(t)$ that

$$\theta L\{\mathbf{Y}(t)\} - \mathbf{I} = \mathbf{K}L\{\mathbf{Y}(t)\}$$

and

$$L\{\mathbf{Y}(t)\} = (\mathbf{I}\theta - \mathbf{K})^{-1} = \begin{bmatrix} \theta^{-1} & \theta^{-1}\boldsymbol{\mu}'(\mathbf{I}\theta - \mathbf{A})^{-1} \\ 0 & (\mathbf{I}\theta - \mathbf{A})^{-1} \end{bmatrix}.$$

The last result shows that in fact

$$\mathbf{Y}(t) = \begin{bmatrix} 1 & \mathbf{z}'(t) \\ 0 & \mathbf{X}(t) \end{bmatrix}$$

where $\mathbf{X}(t)$ is defined in [1], explicit expressions for the elements are given in the Appendix of that paper, and $z_j(t) = 1 - \sum_{i=1}^{5} X_{ij}(t)$.

The mean of the process is $e^{\mathbf{K}t}\mathbf{x}(0)$, which is in agreement with the deterministic theory. Second-order properties can be studied by use of (5).

Waiting times in the five viable stages of the parasite are exponential with parameters κ_i, $i = 1, 2, \ldots, 5$. Of more interest, however, is the distribution function $F(t)$ of the life-span of a female larva entering the sheep in the third larval stage. It turns out that

$$F(t) = 1 - y_{(1)}^{(2)}(t). \tag{7}$$

THE FULL LIFE CYCLE

Equation [1, (11)] defines the linear system of differential equations appropriate for the study of the full life cycle of the parasite. The purpose of this section is to develop a stochastic analogue. Let

$$\mathbf{D} = \begin{bmatrix} -\kappa_6 & 0 & 0 & 0 \\ \lambda_6 & -\kappa_7 & 0 & 0 \\ 0 & \lambda_7 & -\kappa_8 & 0 \\ 0 & 0 & \lambda_8 & -\kappa_9 \end{bmatrix}$$

where the elements of D are as defined for B of [1, (11)], and set $\mathbf{W}(t) = e^{\mathbf{D}t}$.

Recall the form $P(\mathrm{s}, t) = [\sum_{i=1}^{6} y_i^{(2)}(t)s_i]^N$ presented earlier. Then, for a fixed input of N third-stage larvae at $t = 0$, and if mature females lay eggs according to the probability generating function (pgf) $f(s)$,

a stochastic version of [1, (11)] without feedback ($\lambda_9 = 0$) is

$$Q(\mathrm{s}, \theta; t) = \left[\sum_{i=1}^{5} y_i^{(2)}(t)s_i + y_6^{(2)}(t)s_6 f\left(\sum_{i=1}^{4} w_i^{(1)}(t)\theta_i\right)\right]^N. \tag{8}$$

Expressions for the means of the process are

$$E\{X_i(t)\} = \begin{cases} Ny_i^{(2)}(t), & i = 1, 2, \ldots, 6, \\ Ny_6^{(2)}(t)\mu w_{i-6}^{(1)}(t), & i = 7, \ldots, 10, \quad \mu = f'(1). \end{cases}$$

Variances and covariances can be calculated by converting (8) to cumulant generating function form and taking the appropriate mixed derivatives. The resulting expressions are cumbersome and will not be given here.

For the case of a continuous input model to simulate the grazing sheep, it is mathematically easy and biologically reasonable to assume that the input at time t is proportional to $E\{X_{10}(t)\}$, $\lambda y_6^{(2)}(t)w_4^{(1)}(t)$, say, since usually numerous sheep will be grazing the area. This leads to the following modification to (8).

$$R(\mathrm{s}, \theta; t) = \exp\left\{\lambda \int_0^t [Q(\mathrm{s}, \theta; t - x) - 1]y_6^{(2)}(x)w_4^{(1)}(x)\, dx\right\}. \tag{9}$$

It can be verified that the limiting pgf $R(\mathrm{s}, \theta; \infty)$ specifies an "honest" process with finite moments of all orders. This is in marked contrast to the full deterministic model of [1], which "explodes" with unfavorable combinations of the parameters.

REFERENCE

1 G. Gordon, M. O'Callaghan, and G. M. Tallis. A deterministic model for the life cycle of a class of internal parasites of sheep, *Math. Biosci.* (this issue).

(c)

RESEARCH INTO FERTILITY AND MEAT PRODUCTION IN SHEEP

A (c)[1]

Reprinted from the
AUSTRALIAN JOURNAL OF AGRICULTURAL RESEARCH
VOLUME 11, NUMBER 6, PAGES 1017-1025, 1960

SOME ASPECTS OF THE EFFICIENCY OF LARGE-SCALE ARTIFICIAL INSEMINATION OPERATIONS IN SHEEP

By G. M. Tallis and S. S. Y. Young

Reprinted for the
Commonwealth Scientific and Industrial Research Organization
Australia

SOME ASPECTS OF THE EFFICIENCY OF LARGE-SCALE ARTIFICIAL INSEMINATION OPERATIONS IN SHEEP

By G. M. Tallis* and S. S. Y. Young*

[*Manuscript received June* 14, 1960]

Summary

The problems of estimating each day the proportion of ewes coming into oestrus for the first time and the proportion of ewes not returning after the first service have been investigated and discussed in relation to large-scale artificial insemination operations. The results of these theoretical studies were used to examine data from more than 700 Merino ewes which had undergone artificial insemination, and evidence of inefficiency due to an early failure of vasectomized rams was found. From subsequent analyses it was concluded that the methods developed in this paper for estimating conception rates were fairly sensitive and produced satisfactory results.

I. Introduction

When experimental ewes are mated by artificial insemination (A.I.), the usual practice is to introduce vasectomized (teaser) rams into the main flock in the evening and to draft off marked ewes early the following morning. Ewes served by the teasers during the night are presumed to be in oestrus and are inseminated and placed in a separate paddock. From time to time inseminated ewes rejoin the main flock so that those which have not conceived to the first mating will have a chance of re-insemination. This procedure has been discussed in detail by Dun (1956).

The overall efficiency of these matings depends on such things as the personnel and facilities available, the effectiveness of the teaser rams, the soundness of the insemination techniques, and the fertility of the animals. The general failure of the A.I. operation will, of course, be obvious at the conclusion of the mating period. However, it is clearly desirable to have some way of checking on the procedure during the mating period so that faults may be located and, if possible, rectified well before the conclusion of mating operations. With this end in view, two main problems are considered in this paper:

(1) The estimation of the proportion of ewes expected to come in oestrus for the first time on a given day during the mating period.

(2) The day to day estimation of the proportion of ewes expected not to return after the first insemination.

The first of these problems is important because it provides a basis for evaluating the effectiveness of the teasers in picking out all ewes in oestrus. Moreover, the number of ewes to be inseminated daily will also determine the necessary personnel and equipment required.

* Division of Animal Genetics, C.S.I.R.O., McMaster Laboratory, Glebe, N.S.W.

On the other hand, the answer to the second problem will facilitate a daily check on the fertility of animals and the effectiveness of the insemination technique. Advance knowledge of poor conception rates may allow timely alterations to be made to techniques, and to sire composition or the duration of the mating period or both.

II. Methods

Investigation of the two problems mentioned above necessitates an assumption concerning the distribution of the length of the oestrus cycle (l) in ewes. Data collected during the 1958 experimental matings of Merino sheep carried out by the Sheep Breeding Section of the Division of Animal Genetics, C.S.I.R.O., at the Regional Pastoral Laboratory, Armidale, N.S.W., as well as data from the 1956 experimental matings at the National Field Station, "Gilruth Plains", Cunnamulla, Qld., have been examined. From these data l values were plotted on normal probability paper and no significant departure from normality was found. As histograms of l presented by Kelley (1937) also indicate a normal distribution, it will be assumed here that l is in fact normally distributed with mean μ and variance σ^2 $[N(\mu,\sigma^2,l)]$. Combined estimates from the available mating data are $\bar{l} = 17$ days and $\hat{\sigma}^2 = 6{\cdot}25$.

In addition, it will be assumed that, for a particular ewe with cycle length l during a given period t, the onset of oestrus during that period is distributed rectangularly as follows:

$$F(t) = (1/l)\int_0^t \mathrm{d}x \qquad t$$

$$t/l \qquad t<l$$

$$1 \qquad t\geqslant l$$

Problem 1

In order to estimate the expected proportion of a number (n) of ewes coming in oestrus on the ith day of the mating period, we are led formally to compute $E(1/l)_i$ as

$$E(1/l)_i = p_i = \int_i^\infty \frac{N(\mu\sigma^2 l)}{l}\,\mathrm{d}l.$$

This integral may be approximated by the use of published tables of areas of the normal distribution.

The cumulative proportion of ewes which are expected to come in heat by the ith day is

$$P_i = \sum_{j=1}^{i} p_j.$$

As it may be shown that

$$\lim_{i\to\infty} P_i = 1,$$

the p_i may be thought of as probabilities. It is reasonable therefore to assume that

$$\frac{\sqrt{n}(\bar{p}_i - p_i)}{\sqrt{[p_i(1-p_i)]}}$$

is asymptotically $N(0,1)$, and $\bar{p}_i$ (the estimate of p_i) is expected to lie in the interval

$$p_i \pm t_\alpha \sqrt{\left(\frac{p_i(1-p_i)}{n}\right)},$$

where t_α is the normal standard deviate corresponding to the α probability level.

In field applications, however, when teaser rams are first introduced into a flock of ewes, the proportion of ewes served on the first day is usually larger than expected. Experience has shown that this proportion is approximately twice as large as the proportion expected on the first day. This is probably because the fresh teaser rams serve the ewes that come in oestrus on day 1 as well as some of those

TABLE 1

VALUES OF p'_i, P'_i, AND A_i*

Day	p'_i (%)	P'_i (%)	A_i	Day	p'_i (%)	P'_i (%)	A_i
1	12·0	12·0	0	17	2·2	97·2	0·500
2–8	6·0	18·0–54·0	0	18	1·4	98·5	0·655
9	6·0	60·0	0·001	19	0·8	99·3	0·788
10	6·0	66·0	0·003	20	0·4	99·6	0·885
11	5·9	72·0	0·008	21	0·2	99·8	0·945
12	5·7	77·7	0·023	22	0·1	99·8	0·977
13	5·4	83·0	0·055	23	0·0	99·9	0·992
14	4·8	87·8	0·115	24	0·0	99·9	0·997
15	4·0	91·9	0·212	25	0·0	100·0	0·999
16	3·1	95·0	0·345	26	0·0	100·0	1·000

* p'_i = expected proportion of ewes in heat for the first time on day i.

P'_i = expected cumulative proportion of ewes which have come in heat for the first time by day i.

A_i = area under a normal curve with mean 17 days and standard deviation 2·5 days.

ewes which came on heat on previous days. To overcome this we define p_0 as the expected proportion associated with the day before the commencement of insemination operations, so that p'_1, the actual proportion expected at day 1, is p_0+p_1. It follows therefore that

$$p'_2 = p_2, \qquad p'_i = p_i, \qquad \text{and } P'_i = \sum_{j=1}^{i} p'_j.$$

The values of p'_i and P'_i, based on $N(17,\ 6{\cdot}25,\ l)$, have been calculated and are shown in Table 1.

Problem 2

Solution of the second problem depends largely on assumptions regarding the joint distribution of the first and second cycle lengths, l_1 and l_2. Therefore, data from matings at Armidale in 1958 and 1959 were analysed, and the pooled correlation

coefficient between l_1 and l_2 was found to be $-0{\cdot}11 \pm 0{\cdot}11$. As mating at Armidale continued for only 40 days, those ewes with l_1 and l_2 values of 20 days or over do not appear in the data. The effect of this is to bias zero or negative correlations downwards, so that the actual value of ρ is probably closer to zero than $-0{\cdot}11$. Moreover, since the means and variances of l_1 and l_2 were nearly equal in these data, it will be assumed in the subsequent development that l_1 and l_2 are independently and normally distributed with the same mean and variance.

On the above assumptions, the number of ewes returning for second service can be considered. If we write $n_0+n_1, n_2, \ldots, n_{i-1}$ for ewes brought to first service on day $1, 2, \ldots, i-1$, then, with no conception at first service, no ewes are expected to return for second service on D_1. On D_2, $[n_0 P(1<l<2)+n_1 P(0<l<1)]$ are expected to return, and so on as below:

D_2	D_3	.....	D_i
$n_0 P(1<l<2)$	$n_0 P(2<l<3)$		$n_0 P(i-1<l<i)$
$n_1 P(0<l<1)$	$n_1 P(1<l<2)$		$n_1 P(i-2<l<i-1)$
	$n_2 P(0<l<1)$		$n_2 P(i-3<l<i-2)$
			.....
			$n_{i-1} P(0<l<1)$

where $P(1<l<2)$, for instance, is the probability of a particular ewe having a cycle length of 2 days. Putting $P(i-k-1<l<i-k) = a_{i-k}$, we now define the number of ewes m_i expected to return on D_i, as

$$m_i = \sum_{k=0}^{i-1} n_k a_{i-k},$$

and the cumulative number, M_i, is

$$M_i = \sum_{j=2}^{i} m_j \simeq \sum_{k=0}^{i-1} n_k A_{i-k},$$

where

$$A_{i-k} = \int_{-\infty}^{i-k} N(\mu,\sigma^2,l)\, dl.$$

The values for A_{i-k} with $\mu = 17$ and $\sigma^2 = 6{\cdot}25$ are given in Table 1.

On the assumption that all ewes coming to service on the first occasion would again come to service in the absence of conception,

$$\operatorname{Lim}_{i\to\infty} M_i = n.$$

However, if in fact a proportion $1-s = c$ have conceived as the result of the first service, the expected numbers of returns will be sm_i and sM_i.

Unfortunately the Maximum Likelihood estimate of s, using all the available information, is tedious to obtain; and therefore for field application a more convenient if slightly less efficient estimator is recommended.

This estimator is

$$\hat{s} = N_t/M_t,$$

with variance

$$\text{Var } (\hat{s}) = s(1-s\bar{A}_t)/M_t,$$

where

$$\bar{A}_t = \sum_{k=0}^{t-1} (A_{t-k}n_k)/M_t,$$

and N_t is the cumulative number of ewes returning for second service by D_t. Since $c = 1-s$, it is obvious that the estimator of c is

$$\hat{c} = 1-\hat{s},$$

and that $\hat{c}$ has the same variance as $\hat{s}$.

If the numbers realized at the first service are the same as expected, i.e. np_0, np_1, . . . , np_{i-1}, then the proportion expected to return on D_t, with no conception at first service, is

$$q_t = \sum_{k=0}^{t-1} p_k a_{t-k},$$

and the cumulative proportion to a sufficient approximation is

$$Q_t \simeq \sum_{k=0}^{t-1} p_k A_{t-k}.$$

If we denote the observed, cumulative proportion returning for second service by D_t as R_t, then an estimator of s is

$$\hat{s} = R_t/Q_t.$$

In field operation it is often desirable to know the approximate time when the A.I. operation can be terminated. In this case $\hat{c}$ can also be useful if the desired total conception rate, C (for first and second services), is set in advance. This total conception rate can be calculated approximately as

$$\hat{C}_{t+v} = (n\hat{c}_t+\hat{s}_t M_{t+v}\hat{c}')/n, \qquad (v = 1, 2, \ldots, 13 \text{ and } i \geqslant 25)$$

where $\hat{c}_t$ is the estimate of c on D_t and $\hat{c}'$ is the estimate of the conception rate at second service. The value of $\hat{c}'$ is usually lower than that of $\hat{c}$, and in the present data $\hat{c}' = \frac{1}{2}\hat{c}$ approximately, from the field records of conception subsequent to second service. Because M_{t+v} can be calculated v days ahead of D_t (since A_1 to A_v are only negligible) it is possible to predict the approximate value of C on D_{t+v}. The above formula is useful in the absence of consistent daily variations in conception rates.

III. Field Application

In order to check the overall efficiency of the 1958 artificial insemination operation at Armidale, as well as to illustrate the use of Table 1, analyses of first and second services were carried out. Before computing the number of ewes which are expected to come on heat at any particular day, however, it is necessary to

estimate the "effective number" of ewes in the flock. Such an estimate is required because there is usually a small proportion of ewes, d, in any flock which fail to have any signs of oestrus detected. As d probably differs from flock to flock, and from year to year, it can only be estimated from past experience for any particular flock and year. Once the order of d is known, the effective number of ewes is $n(1-d)$, where n is the initial number of ewes.

At Armidale, two groups of ewes (A and B) were inseminated in 1958. Group A started on day 1 of the mating period and group B started 9 days later. The artificial insemination techniques used were similar to those described by Dun (1956), and $n(1-d)$ was 430 and 285 for groups A and B respectively. However,

TABLE 2

ANALYSIS OF FIRST SERVICE RECORDS FOR TWO GROUPS OF EWES

Artificial insemination data, Armidale, 1958. Effective total for A, 430. Effective total for B, 285

Day	Group A		Group B		Day	Group A		Group B		Day	Group A		Group B	
	Obs. No.*	Exp. No.*	Obs. No.	Exp. No.		Obs. No.	Exp. No.	Obs. No.	Exp. No.		Obs. No.	Exp. No.	Obs. No.	Exp. No.
3	21	26	18	17	13	19	23	12	15	23	6	0	1	0
4	13	26	15	17	14	22	21	11	14	24	9	0	3	0
5	24	26	14	17	15	29	17	14	11	25	9	0	2	0
6	19	26	17	17	16	14	13	13	9	26	8	0	2	0
7	17	26	18	17	17	13	9	9	6	27	3	0	0	0
8	37	26	14	17	18	13	6	8	4	28	3	0	0	0
9	30	26	14	17	19	10	3	1	2	29	1	0	0	0
10	26	26	11	17	20	11	2	3	1	30	1	0	1	0
11	13	25	15	17	21	6	1	3	1	31	0	0		
12	18	25	13	16	22	8	0	1	0	32	1	0		

* Obs. No., observed number. Exp. No., expected number.

during the first 9 days, 168 ewes from group A were inseminated with diluted semen, and the conception rate was subsequently calculated to be 55 per cent. for these ewes. During this period 19 ewes were inseminated with undiluted semen, and undiluted semen was used for all inseminations subsequent to day 10.

For the first seven days the vasectomized rams were caught, greased, and "bagged" prior to introduction into the ewe flock at 5 p.m. The bagging operation was precautionary and consisted of strapping hessian bags to the bellies of the rams to prevent copulation. As a smaller number of ewes than expected were inseminated daily during this period, this practice was discontinued from day 8 onwards.

On day 1 teasers were allowed to run with some of the A ewes and the rest of the A ewes were added to the first group on day 2. Similarly, some B ewes were introduced into the experimental flock on day 10 and the balance were put in on day 11. Unfortunately, no records were taken of the numbers of ewes introduced

on these four days and it will be assumed in subsequent calculations that groups A and B were all introduced on day 1 and day 10 respectively.

For this reason, expected values have been calculated from day 3 onwards and the slight negative bias resulting from this assumption is disregarded. The observed and the expected number of ewes coming in heat daily for both groups (A and B) are shown in Table 2.

For data of group A the agreement between observed and expected numbers is poor. Two features are apparent: firstly, the fit for the first week or so is bad, and secondly, there are an unexpectedly large number of ewes with cycle lengths greater

TABLE 3

ESTIMATED AND OBSERVED CONCEPTION RATES (%) AT FIRST SERVICE FOR TWO GROUPS OF EWES

Day	Group A		Group B		Day	Group A		Group B		Day	Group A	
	Est.*	Obs.*	Est.	Obs.		Est.	Obs.	Est.	Obs.		Est.	Obs.
16	46	58	69	58	24	58	58	60	62	32	64	61
17	53	60	63	61	25	62	59	63	64	33	62	61
18	52	60	48	59	26	62	59	65	65	34	63	62
19	55	55	57	57	27	63	60	63	65	35	63	62
20	64	53	54	58	28	62	59	65	67	36	64	63
21	55	55	57	58	29	63	59	66	69	37	64	64
22	58	54	61	61	30	62	60	66	71	38	64	64
23	57	57	62	61	31	63	60					

* Est., estimated. Obs., observed.

than 21 days. Both these features are probably due to the poor performance of teasers during the first few days of the mating period. On these days, some ewes may have shown signs of oestrus but were not marked by teasers, and these ewes subsequently showed oestrus again later on. Hence the apparent l values of more than 21 days.

On the other hand, the agreement of data from group B with expectation is reasonably good. Thus, the conclusion is drawn that after day 9, when bagging was discontinued, the teasers were working efficiently.

Realized and estimated daily conception rates have been computed for groups A and B. The realized conception rates were calculated from the field data subsequent to the insemination operation, as it was then possible to identify all the ewes which conceived as a result of first service at a given day. The estimated conception rates were calculated from the equations of M_l and $\hat{s}$. As a numerical example, consider the estimation of conception rate at D_{20} for group A. As some of the group A ewes were put in with the teasers on D_1 and the rest on D_2 and the number of ewes served at the end of D_2 was 26, and since n_0, n_1, and n_2 were unknown in this case, the 26 ewes were partitioned as $n_0 = 8$, $n_1 = 9$, $n_2 = 9$. Thus using Tables 1 and 2,

$$\begin{aligned} M_{20} = &(8\times0\cdot885)+(9\times0\cdot788)+(9\times0\cdot655)+\\ &(21\times0\cdot500)+(13\times0\cdot345)+(24\times0\cdot212)+\\ &(19\times0\cdot115)+(17\times0\cdot055)+(37\times0\cdot023)+\\ &(30\times0\cdot008)+(26\times0\cdot003)+(13\times0\cdot001)\\ = &44\cdot44. \end{aligned}$$

From the records $N_{20} = 16$, so that $\hat{s} = 16/44\cdot44 = 36$ per cent., and $\hat{c} = 1-\hat{s} = 64$ per cent. These results are given in Table 3.

In both groups a positive linear trend was apparent. In the case of group A this is to be expected because of the initial low conception rate due to the use of diluted semen and because of the use of an inexperienced operator at the commencement of the operation. The trend in group B is not so marked and could be due to the improvement in the operator's techniques.

As only a very small proportion of ewes have cycle lengths of 14 days or less, the estimated number of ewes returning for second service at the ith day can only be used to calculate the conception rate of the group of ewes which had been inseminated approximately by D_{i-14}. For this reason in Table 3 the estimated conception rate at D_i was compared with the observed conception rate of ewes inseminated by D_{i-14}. This is an arbitary procedure as results on D_{i-13} or D_{i-15} may also be suitable bases for comparison. It is worth noting that the conception rate at D_{i-14} can be used as an indication of the rate at D_i when there is no reason to suspect the existence of a trend, but not otherwise.

It can be seen that the agreement between the estimated and the observed conception rates is good. Large discrepancies, especially prior to D_{20}, may be due to sampling errors, or the arbitrariness of choosing the rates on D_{i-14} as basis for comparison, or both. In general the results show that this method of estimating s is sufficiently sensitive, for practical purposes, after 20 days of insemination when the proportion of ewes returning becomes large.

IV. Discussion

Apart from teaser inefficiency, experience has shown that a sudden change in climatic conditions, such as a severe drop of temperature, may result in a decrease of numbers of ewes in heat for a period of several days. This period of reduced numbers is usually followed by another period of several days when larger than expected numbers of ewes in heat are observed. The additional ewes found in the later period tend to compensate for a smaller number served during the period of stress. In these cases, it seems that severe changes in weather lengthens the oestrus cycles of some ewes, but this effect is clearly distinguishable from the type of teaser effect found in group A. This sensitivity of oestrus lengths to changes in climatic conditions may partly explain the low repeatability of oestrus lengths as estimated in Section II.

The techniques described above are also useful when two or more conception rates based on different mating periods are to be compared. For instance, suppose

two groups of ewes, 1 and 2, have been mated for 30 and 40 days respectively and 30 per cent. of the ewes in group 1 and 35 per cent. of the ewes in group 2 returned for second service. Actual conception rates to first service are not necessarily 70 and 65 per cent. in this case, and in order to compare the two groups $\hat{c}$ must be calculated according to $1-(N_t/M_t)$ as shown earlier. However, if no records of n_k are available approximate values of $\hat{c}$ may be calculated as follows:

Day 30	Day 40
$R = 0{\cdot}30,\ Q = 0{\cdot}789,$	$R = 0{\cdot}35,\ Q = 0{\cdot}998$
$\hat{s} = R/Q = 0{\cdot}38,$	$\hat{s} = R/Q = 0{\cdot}35$
and $\hat{c} = 1-\hat{s} = 0{\cdot}62.$	and $\hat{c} = 1-\hat{s} = 0{\cdot}65$

Thus the actual conception rates are approximately 62 and 65 per cent.

Finally it must be stressed that Table 1 has been constructed on the assumption that l_1 and l_2 are independently and normally distributed, with a mean of 17 days and a standard deviation of 2·5 days. In flocks where these conditions are not fulfilled, different tables must be computed by similar methods to those outlined in Section II.

V. Acknowledgments

The authors wish to thank the late Mr. J. F. Barrett, Executive Officer at the Regional Pastoral Laboratory, Armidale, and Dr. A. A. Dunlop of the Division of Animal Genetics for their cooperation. Acknowledgment is also made of the work of Miss E. Smith and Mr. R. Butler for technical assistance.

VI. References

Dun, R. B. (1956).—Artificial insemination of sheep. *Wool Tech.* 3: 39–42.

Kelley, R. B. (1937).—Studies in fertility in sheep. Bull. Coun. Sci. Industr. Res. Aust. No. 112.

A (c)[2]

THE EFFECTS OF LENGTH OF OESTRUS AND NUMBER OF INSEMINATIONS ON THE FERTILITY AND TWINNING RATE OF THE MERINO EWE

By A. A. DUNLOP* and G. M. TALLIS†

[*Manuscript received August* 29, 1963]

Summary

In a flock of 512 Merino ewes all were inseminated on the first day of oestrus. On the next day half of those still in oestrus were re-inseminated, as were half of those no longer in oestrus. Based on numbers of lambs born subsequently, estimates were made of the separate and joint effects of number of times inseminated and number of days in oestrus.

A second insemination significantly increased the proportion of ewes bearing twins by 4·7%. Other effects were not significant, but suggested that ewes with longer oestrous periods produced slightly more lambs, and that ewes in oestrus on a second day benefited most from re-insemination.

I. INTRODUCTION

There have been many investigations, mainly in eastern Europe, of the effects on lamb production of more than one insemination in a single oestrous period. The rationale of this work appears to fall under three main headings. First, an appreciable fraction of multiple ovulations in sheep are asynchronous (Lysov and Stojanovskaja 1937; Polovceva, Okulicev, and Judovic 1938); so that the ova shed by a ewe may become available for fertilization at different times during a single oestrous period. Second, an increase in the number of lambs born following a second insemination may occur should sperm from an initial insemination at the commencement of oestrus fail to survive when ovulation takes place towards the end of a lengthy heat period. Glembockii and Vasilijev (1944) have produced some indirect evidence favouring this view and they consider that such delayed ovulations frequently tend to be multiple. Finally a direct effect of increased semen dose rate may be operating (Koger 1951). Published work on the effects of multiple insemination has recently been summarized by Salamon (1962). The general situation appears to be that multiple (usually double) insemination commonly gives an increase in number of lambs born. Both fewer dry ewes and more multiple births can bring about this increase, the relative importance of these two changes varying from experiment to experiment, with the latter perhaps being somewhat more important. Variation in results of experiments of this type may well be due to breed differences, to differences in technique (times from commencement of oestrus to the first insemination and between inseminations), and to time of year.

*Division of Animal Genetics, CSIRO, McMaster Laboratory, Glebe, N.S.W.
†Division of Mathematical Statistics, CSIRO, McMaster Laboratory, Glebe, N.S.W.

There have been few reports of work on the effects of multiple insemination in the Australian Merino. Keast and Morley (1949) found no advantage (in fact, a slight disadvantage) in daily insemination of ewes remaining in oestrus for more than 24 hr as compared to similar ewes inseminated once only. However, their numbers of animals were small. Sinclair (1957), on the other hand, found that a second mating, some 6 hr after the first, produced a significant increase of some 13% in the proportion of ewes conceiving. The effect on the number of lambs born was not reported. In the results obtained by Salamon and Robinson (1962), the following percentages of ewes lambed: 60·9% of those inseminated once, 70·3% of those inseminated twice at 8 hr intervals, and 62·5% of those inseminated twice at 24 hr intervals. The differences were not statistically significant. Once again, effects on the number of lambs born were not examined. Thus the relative importance of increases in multiple births and decreases in dry ewes cannot be assessed from existing Australian data.

In contrast to the multiplicity of experiments to evaluate plural insemination, there has been little investigation of a possible relationship between the length of time a ewe stays in oestrus and the number of lambs she will bear. Most workers who have considered this factor at all have apparently assumed that ewes remaining in oestrus for lengthy periods are more likely to benefit from multiple inseminations and have acted accordingly, so that length of oestrus and number of inseminations are confounded in the ensuing lambing data (Kirillov 1938; Glembockii and Vasilijev 1944; Keast and Morley 1949; Lopyrin and Donskaja 1959). In one of these publications (Glembockii and Vasilijev 1944), a change of breed of ram from initial to supplementary inseminations permitted the identification of the lambs resulting from each.

There appear to be no published investigations which permit the estimation of the separate and joint effects on lambing performance of number of inseminations and of length of oestrus. Lysov and Stojanvskaja (1937) approached this most closely when they recorded the heat status of ewes at the time of re-insemination on a second day, but did not obtain this information in the case of ewes inseminated once only. The present paper reports an experiment in which these variables are examined in this way in the Australian Merino.

II. Materials and Methods

(*a*) *Experimental Procedure*

In the course of sheep-breeding experiments at the CSIRO Pastoral Research Laboratory, Armidale, N.S.W., some 700 to 800 Merino ewes ranging from 1½ to 8½ years of age are inseminated annually with semen from 40 to 60 rams. Semen collected by electro-ejaculation is used without dilution or storage. Ewes in oestrus are detected by vasectomized rams whose briskets are smeared with pigmented grease. These rams run with the ewes overnight from 5 p.m. until 6 a.m., and insemination takes place from 8.30 a.m. to approximately 10.30 a.m. In the insemination season commencing on May 8, 1961, ewes which were inseminated for the first time on May 9 (and likewise on succeeding days) were re-teased with rested teasers between 6 and 7 a.m. on the following morning. Thus the mob of ewes initially inseminated on any

one day was, the following morning, divided into two groups, those which were still in oestrus on the second day (i.e. 24–48 hr after the commencement of oestrus) and those which were not. Each of these groups was randomly halved, one half being re-inseminated in the normal course of the day's insemination and the other not re-inseminated. The technique of insemination described by Morrant and Dun (1960) was used. The entire ejaculate from a ram was equally divided among those ewes allotted to him which required insemination on any one day. Thus semen dose rates were relatively high, averaging 0·34 c.c. The effects of semen dose rate and other semen traits have not been considered in the present data. The final second inseminations were made on May 30, and the numbers of offspring born to these ewes were recorded in due course.

TABLE 1

LAMBING PERFORMANCE OF EWES IN RELATION TO OESTRUS LENGTH AND NUMBER OF INSEMINATIONS

No. of Inseminations (i)	Days in Oestrus (j): 1	2
1	*P_{11} = 12 Q_{11} = 189 R_{11} = 66 N_{11} = 267	P_{12} = 1 Q_{12} = 27 R_{12} = 10 N_{12} = 38
2	P_{21} = 18 Q_{21} = 185 R_{21} = 63 N_{21} = 266	P_{22} = 4 Q_{22} = 28 R_{22} = 9 N_{22} = 41

*P_{ij}, number of ewes bearing twins. Q_{ij}, number of ewes bearing singles. R_{ij}, number of ewes not lambing to inseminations at this oestrus.

III. Statistical Methods and Results

The results of the experiment described above are summarized in Table 1. It will be noticed that there are four statistically independent groups, each with three discrete classes. It will be assumed that the outcome in these groups is adequately described by a trinomial statistical model, and that the likelihood of the particular result in the i, jth group is given by

$$L_{ij} = \frac{N_{ij}!}{P_{ij}!\ Q_{ij}!\ R_{ij}!} \cdot p_{ij}^{P_{ij}}\, q_{ij}^{Q_{ij}}\, r_{ij}^{R_{ij}}. \qquad (1)$$

In (1), p_{ij}, q_{ij}, and r_{ij} are the probabilities that a ewe inseminated i times and observed to be in heat on j days, ($i, j = 1, 2$) will give birth to twin, single, and no lambs respectively. Because of the independence of the four sets of data, the combined likelihood for the experiment is given by:

$$L = \prod_{ij} L_{ij}. \qquad (2)$$

It is now possible to postulate the following parameter models (Table 2) for the probabilities associated with Table 1.

In the above models of Table 2, α, a, and s represent effects due to the number of inseminations on twin births, single births, and dry ewes respectively; β, b, and t are equivalent effects attributable to the number of days that a ewe is observed in oestrus; while γ, c, and u may be interpreted as interaction terms. (The analogy between these models and the usual model for the two-way analysis of variance with interaction is obvious.)

TABLE 2

PARAMETER MODELS

Number of Inseminations	Days in Oestrus 1	Days in Oestrus 2
1	$p_{11} = p^* + \alpha + \beta + \gamma$ $q_{11} = q + a + b + c$ $r_{11} = r + s + t + u$	$p_{12} = p + \alpha - \beta - \gamma$ $q_{12} = q + a - b - c$ $r_{12} = r + s - t - u$
2	$p_{21} = p - \alpha + \beta - \gamma$ $q_{21} = q - a + b - c$ $r_{21} = r - s + t - u$	$p_{22} = p - \alpha - \beta + \gamma$ $q_{22} = q - a - b + c$ $r_{22} = r - s - t + u$

*These parameters are subject to the restrictions that: $p + q + r = 1$; $\alpha + a + s = 0$; $\beta + b + t = 0$; $\gamma + c + u = 0$.

Now $\log L$ is a function of the eight unknown parameters p, α, β, γ, q, a, b, c, and maximization in the standard manner shows that:

$$\hat{p} = \sum_{ij} \hat{p}_{ij}/4 \qquad \hat{q} = \sum_{ij} \hat{q}_{ij}/4$$

$$\hat{\alpha} = (\hat{p}_{11} + \hat{p}_{12} - \hat{p}_{21} - \hat{p}_{22})/4 \qquad \hat{a} = (\hat{q}_{11} + \hat{q}_{12} - \hat{q}_{21} - \hat{q}_{22})/4$$

$$\hat{\beta} = (\hat{p}_{11} + \hat{p}_{21} - \hat{p}_{12} - \hat{p}_{22})/4 \qquad \hat{b} = (\hat{q}_{11} + \hat{q}_{21} - \hat{q}_{12} - \hat{q}_{22})/4$$

$$\hat{\gamma} = \hat{p}_{11} - \hat{p} - \hat{\alpha} - \hat{\beta} \qquad \hat{c} = \hat{q}_{11} - \hat{q} - \hat{a} - \hat{b}$$

The variances of the four estimates involving the $\hat{p}_{ij}$ are the same, and in fact equal $(16)^{-1} \sum_{ij} p_{ij}(1-p_{ij})N_{ij}^{-1}$. A similar remark applies to the estimates involving $\hat{q}_{ij}$, and in this case the variances equal $(16)^{-1} \sum_{ij} q_{ij}(1-q_{ij})N_{ij}^{-1}$. The estimates and their standard errors are given in Table 3.

The various combinations of the estimates of Table 3 are:

$\hat{p}_{11} = 0{\cdot}0450$	$\hat{p}_{12} = 0{\cdot}0264$	$\hat{p}_{21} = 0{\cdot}0677$	$\hat{p}_{22} = 0{\cdot}0976$
$\hat{q}_{11} = 0{\cdot}7079$	$\hat{q}_{12} = 0{\cdot}7105$	$\hat{q}_{21} = 0{\cdot}6955$	$\hat{q}_{22} = 0{\cdot}6829$
$\hat{r}_{11} = 0{\cdot}2471$	$\hat{r}_{12} = 0{\cdot}2631$	$\hat{r}_{21} = 0{\cdot}2368$	$\hat{r}_{22} = 0{\cdot}2195$

It is hardly surprising that significantly different estimates were obtained for p, q, and r, which indicate the overall proportions of ewes bearing 2, 1, or 0 lambs respectively ($p + q + r = 1$).

The estimates of the remaining parameters, which are in terms of deviations from an overall mean ($\alpha + a + s = 0$ and so on), are of much more interest. Of the two sets of main effects, the values for number of inseminations are greater than those for duration of oestrus, the former being never less than 1%. Since the effect of two inseminations is equal to that of one insemination but opposite in sign, as defined in Table 2, the difference between ewes with one or two inseminations is thus never less than 2%. The percentage of ewes bearing twins was higher by 4·7 with two inseminations than one ($\hat{a} = -0{\cdot}0235$), while the percentage bearing singles was 2% less ($\hat{a} = +0{\cdot}0100$), and the percentage of dry ewes 2·7% less ($\hat{s} = +0{\cdot}0135$). The effect for twin births was significant at the 5% level, from a one-tailed test based on the assumption that a second insemination would increase the proportion of twins. The effects for single births and dry ewes were not significant individually, but it is worth recording that the significant increase in twinning was accompanied by a decrease in the proportion of dry ewes as well as in the proportion of ewes bearing singles.

TABLE 3

ESTIMATES OF PARAMETERS AND STANDARD ERRORS

Birth type fractions	$\hat{p}^* = 0{\cdot}0591 \pm 0{\cdot}0142$,	$\hat{q}^* = 0{\cdot}6992 \pm 0{\cdot}0277$,	$\hat{r}^* = 0{\cdot}2417 \pm 0{\cdot}0258$
Insemination frequency effects	$\alpha^* = -0{\cdot}0235 \pm 0{\cdot}0142$,	$\hat{a} = 0{\cdot}0100 \pm 0{\cdot}0277$,	$\hat{s} = 0{\cdot}0135 \pm 0{\cdot}0258$
Oestrus length effects	$\hat{\beta} = -0{\cdot}0028 \pm 0{\cdot}0142$,	$\hat{b} = 0{\cdot}0025 \pm 0{\cdot}0277$,	$\hat{t} = 0{\cdot}0003 \pm 0{\cdot}0258$
Interactions	$\hat{\gamma} = 0{\cdot}0121 \pm 0{\cdot}0142$,	$\hat{c} = -0{\cdot}0038 \pm 0{\cdot}0277$,	$\hat{u} = -0{\cdot}0083 \pm 0{\cdot}0258$

The estimated direct effects on lambing status (0, 1, or 2 lambs) of whether a ewe is in oestrus on one day or on two days are not significant and are very small (0·5% and less); their signs are nevertheless in the anticipated directions, i.e. two days in oestrus increases both ewes in lamb and twin births. Although the interaction terms γ, c, and u are not significant, those affecting the proportions of dry ewes and of twins are large enough to be of some interest. They suggest that ewes on heat one day and inseminated once and ewes on heat for two days and inseminated twice will have both more twins and fewer dry ewes than would be expected from the average effects of the main classifications concerned. Unlike combinations of days in oestrus and times inseminated, on the other hand, seem likely to have fewer twins and more dry ewes than expected.

IV. DISCUSSION

The present results on the effect of double as opposed to single insemination, while well short of significance in some cases, are in general agreement with the usually positive results obtained by other workers, but they do not shed any light on which of the several possible mechanisms discussed earlier are involved.

It is hardly surprising that it has not been possible to demonstrate any direct relation between length of oestrus and lambing performance when one considers the error component in the measurement of the former trait: ewes when first drafted may have a range of almost 24 hr in the time since first coming into oestrus. In spite of the absence of any marked main effect of length of oestrus, the estimated values of the interaction terms favour the implicit assumptions of those writers who have considered that ewes exhibiting lengthy heat periods are likely to benefit from multiple inseminations. While it is not possible to compare our results with those obtained in Karakul sheep by Lysov and Stojanovskaja (1937) in these terms, a comparison can be made if our two subclasses of ewes inseminated once are pooled. Having a lower conception rate in their control group (47·8%), they made a major advance of some 11% in the proportion of ewes conceiving as a result of double insemination, an increase which was present in only very minor degree in our data. Their basic proportion of twins, on the other hand, was somewhat higher (7·4% *v.* 4·3%) and their increases in the proportion of twins were naturally somewhat greater on passing from ewes inseminated once to ewes on heat one day but inseminated twice and from these to ewes on heat on two days and inseminated twice (4·9 and 5·1% *v.* 2·5 and 2·9%). No manipulation of inseminations will increase the proportion of twins in the absence of multiple ovulations. However, the benefits estimated in our data as accruing from double insemination of ewes still on season on a second day are so small that the work of re-teasing and re-insemination would be poorly repaid. This is particularly so when one considers the small fraction of ewes — about 0·13 — which exhibited oestrus on the second day under our conditions.

The gains from wholesale double insemination are probably real and of somewhat greater size (an increase of approximately 4·4% of lambs born in the data of Table 1). At the practical level, the net gain for the additional work involved is rather small. In this regard it should be noted that under some circumstances there could be an offsetting reduction in conception due to lowered semen dose rate. This was not the case here, as an examination of our records has shown that ewes inseminated once received almost exactly half the volume of semen received by ewes inseminated twice. While the mechanism of the gains made here through double insemination is not clear, these are large enough to suggest that more searching investigations, where time of onset of oestrus is known much more precisely, and where variations are made in times to the first and between the first and the second inseminations, would give results of both physiological and practical interest.

V. Acknowledgments

Our thanks are due to Messrs. B. Gream, R. W. Moore, and E. K. Yates for assistance in the field.

VI. References

Glembockii, J., and Vasilijev, G. (1944).—Sovkhoz. Proizvod. No. 10/11: 40.
Keast, J. C., and Morley, F. H. W. (1949).—*Aust. Vet. J.* 25: 281.
Kirillov, V. (1938).—*Probl. Anim. Husb., Moscow.* 7: 39.
Koger, M. (1951).—Bull. N. Mex. Agric. Exp. Sta. No. 366.

LOPYRIN, A. I., and DONSKAJA, V. I. (1959).—*Ovchevodstvo* **5**: 18.
LYSOV, A. M., and STOJANOVSKAJA, V. I. (1937).—*Probl. Anim. Husb., Moscow.* **6**: 16.
MORRANT, A. J., and DUN, R. B. (1960).—*Aust. Vet. J.* **36**: 1.
POLOVCEVA, V. V., OKULICEV, G. A., and JUDOVIC, S. S. (1938).—*Probl. Anim. Husb., Moscow* **7**: 55.
SALAMON, S. (1962).—M.Sc.Agr. Thesis, Univ. of Sydney.
SALAMON, S., and ROBINSON, T. J. (1962).—*Aust. J. Agric. Res.* **13**: 52.
SINCLAIR, A. N. (1957).—*Aust. Vet. J.* **33**: 88.

A (c)[3]

THE RELATION OF SEMEN AND VAGINAL MUCUS TRAITS TO FERTILITY IN THE AUSTRALIAN MERINO

By A. A. DUNLOP,* G. M. TALLIS,† G. H. BROWN,† and B. D. GREAM‡

[*Manuscript received July* 7, 1971]

Abstract

Data relating semen traits and vaginal mucus scores in 515 ram years and 4190 first inseminations to subsequent lambing performance of ewes were analysed by a range of statistical methods. The results showed there to be an important curvilinear effect of mucus score at insemination time and linear effects of scores for motility and consistency of semen in the ejaculates used. These effects operated both on fertility and on fecundity. There appeared also to be smaller linear effects of volume of semen inseminated per ewe and estimates of the proportion of abnormal spermatozoa in ejaculates collected before the insemination season on fecundity and probably on fertility. The effects of age of ewe on fecundity to a single insemination are in contrast with the usual effects of age on lambs born under natural mating. In our material two-tooth ewes produced 5–6% more lambs from a single insemination than did older ewes. It is suggested that, while older ewes are more sexually active and shed increased numbers of ova, the reproductive tract becomes a less favourable environment for initiation and completion of pregnancy with increasing age.

I. INTRODUCTION

While much has been published on methods of measuring attributes of ram semen and on their interrelationships, less is known of the influence which semen traits may have on observed fertility. This situation no doubt reflects the facts that, while it is relatively cheap and easy to obtain semen samples for laboratory study, to arrange parallel matings or inseminations and to observe the subsequent lambings over worth-while numbers of ewes is neither cheap nor easy. The subject has been summarized in a review by Emmens and Robinson (1962). An early exception to the general lack of information relating semen traits to fertility in the sheep is the work of Wiggins, Terrill, and Emik (1953), who observed pre-mating semen characteristics and subsequent lambings over a lengthy period with large numbers of sheep of the Rambouillet and related breeds. The only traits both appreciably and significantly correlated with fertility were percentage of normal sperm ($r = 0{\cdot}43$) and, to a lesser extent, the related traits, percentage of abnormal heads and percentage of live normal sperm. More recently workers at the same location (Hulet and Ercanbrack 1962), using a fairly extensive range of semen traits, have observed real and appreciable correlations of a number of these with fertility (motility score $-0{\cdot}60$, pH $-0{\cdot}66$, percentage abnormal

* Division of Animal Genetics, CSIRO, P.O. Box 90, Epping, N.S.W. 2121.

† Division of Mathematical Statistics, CSIRO, King Street, Newtown, N.S.W. 2042.

‡ Division of Animal Genetics, CSIRO, Pastoral Research Laboratory, Private Bag, P.O., Armidale, N.S.W. 2350.

necks −0·66, concentration +0·56, percentage live normal +0·70, percentage abnormal −0·66) while indexes of fertility using some of these attributes have been correlated with fertility (up to $r = 0·74$) in subsequent use. Later work (Hulet, Foote, and Blackwell 1965) showed that a number of these semen characters also had real, though somewhat smaller, relationships with fecundity.

Mating by artificial insemination in sheep-breeding experiments at the CSIRO Pastoral Research Laboratory, Armidale, N.S.W., commenced in 1957 and continued to 1965. This work offered the opportunity of making observations of semen traits and their relation to subsequent fertility. While the final selection of sires actually used was influenced to some extent by the semen picture before mating, particularly in grossly abnormal samples, this was usually a minor factor in comparison with merit in the traits under selection. It was felt therefore that there would be sufficient variation remaining in semen traits to allow an evaluation of their importance in determining reproductive performance. This paper reports the results of such an investigation in inseminations from 1957 to 1965.

II. Materials and Methods

In each of the years 400–800 ewes of mixed ages came forward for mating by artificial insemination, while 50–60 rams were used each season and less than 20% of these were older than 2½ years. Approximately 10 days before the commencement of the mating season in early May, semen was collected by electro-ejaculation from preselected sires and reserves, and was scored for consistency from a range of eight grades similar to those of Gunn, Sanders, and Granger (1942). A small drop of semen was placed on a cover slip on a warm stage and scored subjectively under low power into one of 10 grades of motility. In both scores a grade of 1 was at the most desirable end of the range. The percentage of abnormal sperm was estimated subjectively from a smear stained with haematoxylin and eosin. On the basis of these three scores each chosen sire was either confirmed or, occasionally, replaced by the reserve sire.

The routine of insemination over 30 days was essentially that described by Morrant and Dun (1960). Gradings for motility and consistency as described above were made on a routine basis on every ejaculate, the volume of semen used for each ewe also being recorded. Because of the relatively narrow ewe/ram ratio, fresh undiluted semen was used and the volume of semen per ewe was on the average high at 0·44 ml. The range, however, was appreciable, varying from 0·05 to 2·20 ml. This arose because the ewes to be mated to each sire were predetermined, and the total ejaculate was used irrespective of the number of ewes in a sire group which were in oestrus on any particular day.

Over a considerable period, increasingly more detailed work has indicated that fertility changes accompany the progressive changes in consistency and volume of vaginal mucus during the oestrous period (Kardymovic, Marsakova, and Pavljucuk 1934; Keast and Morley 1949; Morrant and Dun 1960; Restall 1961). Subjective gradings of vaginal mucus were made on all ewes at the time of insemination. The grades used were: 1, clear; 2, clear and copious; 3, cloudy and copious; 4, cloudy; 5, creamy; 6, cheesy.

The traits initially chosen for analysis were those which could be quickly observed on a routine basis during a field insemination programme without appreciably delaying operations. These were volume of semen per ewe, consistency and motility of ejaculate, and mucus score of the ewe. To these was added the estimated percentage of abnormal sperms in the semen smear made before the insemination season. Insufficient labour was available to observe this character on ejaculates during the insemination programme.

Throughout the present analyses, fertility and fecundity, were based on the results of first inseminations. Whether fertility (the presence or absence of a completed pregnancy) or fecundity (the number of offspring born, zero or a larger number) was used in a particular analysis, will be seen to depend on the statistical method in use in each case. As there was naturally a proportion of returns, some ewes were inseminated twice and a small number three times. Criteria were therefore needed to decide which insemination resulted in a birth in such cases of plural insemination. A gestation period of $\geqslant 140$ to $\leqslant 160$ days was deemed acceptable. Thus a single insemination resulting in a birth within this range of gestation lengths was accepted. Where two insemination dates were recorded and only one putative gestation length was within this range, this was accepted. In the rare cases where neither was within the range none was accepted. Where three inseminations had been recorded and only one fell within the above gestation range this was accepted. Where more than one or none fell within the acceptable range none was accepted.

In considering the results of these inseminations, fertility was used in the absolute sense. Thus an inseminated ewe surviving to lambing time must fall into one of two classes. Either the ewe was dry, or she bore at least one lamb. The results from a small number of ewes which were inseminated twice in one oestrus (Dunlop and Tallis 1964) were not considered. Preliminary observation of the relation of the five traits to fertility was carried out graphically and two traits, volume of semen per ewe and percentage of abnormal sperms, were initially deleted as not having a striking relation to fertility. The graphs relating these respective traits to fertility are presented in Figure 1. Here the data on motility, consistency, mucus score, and volume from all nine years are pooled without regard to years, and points depending on fewer than 100 observations are not shown. In the case of the percentage of abnormal sperms the data from 1961 have been excluded, as a consistent technical error in smear preparation in that year resulted in a gross inflation of the proportion of sperms of abnormal appearance. When data were distributed over all the estimated percentages of abnormal sperms the graph was so variable, because of the very small ram numbers represented by individual points, that any trend was difficult to observe. The data were therefore grouped for analysis into the five classes denoted in Figure 1 on the basis of the proportion of abnormal sperms. Each class represents more than 50 ram years and more than 500 ewe years.

The records finally analysed by regression methods contained 4190 first inseminations from 515 rams on a within-year basis. Some rams and many ewes were of course represented in more than one year.

As shown in Figure 1, the effects of consistency and motility appeared to be mainly linear, while, in the case of mucus score, fertility was at a maximum at a mean score probably closer to 2 than 3. To facilitate analysis and to ensure that there were

useful numbers in as many cells as possible in the three-way classification (mucus × consistency × motility) the final three classes in mucus score, the final four classes in consistency, and the final five classes in motility were compressed into a single class in each case. Thus the data finally analysed contained four grades for mucus score, five grades for consistency, and six grades for motility. Weighted mean values on the respective coordinates were computed in the revised classifications. Following this procedure cells containing fewer than four items of data were deleted.

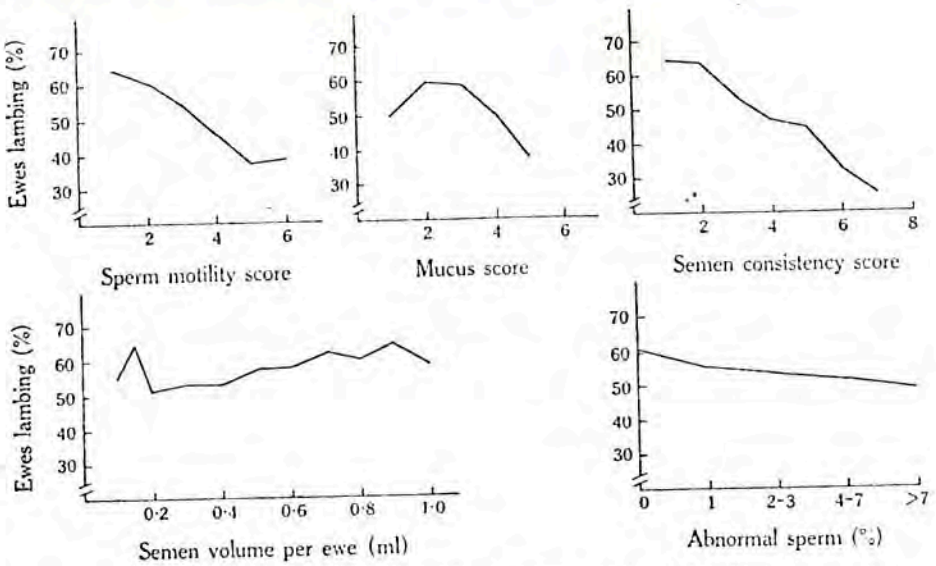


Fig. 1.—Covariation of sperm and mucus traits with fertility.

The models on the basis of which the data were analysed were:

(1) $$P_{ijk} = q + ax_i + bx_i^2 + cy_j + dz_k + e_{ijk}.$$

Here P_{ijk} is the proportion lambing of those ewes which fall in the ith mucus class, the jth consistency class, and the kth motility class; q is a constant; a, b, c, and d are regression coefficients; x_i, y_j, and z_k are mean gradings in the three traits, in the above order; and e_{ijk} is a random error term with an expectation of zero and a variance of $P_{ijk}(1 - P_{ijk})/n_{ijk}$, where n_{ijk} is the number of ewes in the subclass defined by the subscripts.

(2) A model identical with (1) except that e_{ijk} is normally distributed, with a mean of zero and a variance of σ^2.

Since the P_{ijk} may be interpreted as probabilities, model 1 should be the appropriate choice. However, the data were observed over a number of years and the question of between-year homogeneity arises. Techniques for the testing of between-year homogeneity for model 1 have not been developed, hence the need for introducing model 2; i.e. the modification of the error term in model 2 should be regarded as an approximate method for examining between-year differences in regression.

The possibility of improving the analysis by using the arcsine transformation has also been investigated. This, however, did not increase the fraction of controlled variation and the results are not given here. The probable explanation is that most of the P_{ijk} are not near to the extremes of the possible range (zero or one) and little has been gained by the variance-stabilizing transformation.

The parameters q, a, b, c, and d were estimated by the method of Tallis (1964) where a least squares procedure is indicated. The estimates are identical under either model. The relative importance of the several traits was assessed by the squares of the successive multiple correlation coefficients (R^2), as one or more parameters were deleted from the model.

In more detail, the data were initially analysed within individual years under model 1, giving a series of estimates of regression coefficients. Under the simplifying assumptions of model 2 the sets of regression coefficients were tested for heterogeneity between years, while year effects on fertility were also tested for significance.

The regression coefficients were then re-estimated under model 1, the data being pooled on a within-year basis. The relative importance of the several traits on an average basis was estimated by the R^2 technique, using the pooled analysis, while the standard partial regression coefficients from the same analysis gave an alternative solution to the same problem.

Subsequently a re-examination of the data contributing to Figure 1 suggested that both semen volume and percentage of abnormal sperms might have real, though relatively small, effects on fertility. The inclusion of two further variables would have rendered the number of items per cell so small that the methods just described could not have been used. An alternative method of estimating simultaneously the effects of the five variables together with age of ewe was sought. The data were analysed under the following model:

$$y_{ijklmno} = \mu + a_i + b_j + c_k + d_l + e_m + f_n + g_{ijklmno}.$$

Here $y_{ijklmno}$ is the number (0, 1, or 2) of lambs born to the oth ewe in the subclass defined by the subscripts i–n,

μ is the mean of the population being sampled,

a_i is the effect of the ith mucus class, $i = 1 \ldots 4$ and $a_4 = 0$,

b_j is the effect of the jth consistency class, $j = 1 \ldots 5$ and $b_5 = 0$,

c_k is the effect of the kth motility class, $k = 1 \ldots 6$ and $c_6 = 0$,

d_l is the effect of the lth volume class, $l = 1 \ldots 5$ and $d_5 = 0$,

e_m is the effect of the mth class of abnormality percentages, $m = 1 \ldots 5$ and $e_4 = 0$,

f_n is the effect of the nth age of dam class, $n = 1 \ldots 3$ and $f_3 = 0$,

$g_{ijklmno}$ is a deviation of the oth observation in the subclass defined by the subscripts $i \ldots n$ from the sum of μ and the six effects defined by these subscripts. The $g_{ijklmno}$ have an expectation of zero.

It should be noted that the concentration of classes in mucus, consistency, and motility was as already described for the regression analyses. The classes for volume, percentage of abnormalities, and age of dam were:

Class No.	*Volume*	*Abnormalities*	*Age*
1	>0·6 c.c.	0	2 years
2	0·45–0·6 c.c.	1%	3 years
3	0·35–0·4 c.c.	2–3%	>3 years
4	0·25–0·3 c.c.	4–7%	
5	≤0·2 c.c.	>7%	

As the data were non-orthogonal, the parameters were estimated by solving the pertinent least squares equations. The data from each year were analysed separately by analysis of variance, which permitted approximate tests of significance of each of the sets of main effects. With the aim of assessing the relative importance of the six classifications in controlling fecundity, the sometimes unrealistic assumption was made that the classes in each classification constituted a randomly drawn set. Components of variance were then estimated by Henderson's method 3 (1953).

III. Results and Discussion

(*a*) *Regression Analysis*

The partial regression coefficients of fertility on the four traits within individual years are presented in Table 1 together with means of the subclass values in fertility

TABLE 1

PARTIAL REGRESSION COEFFICIENTS OF FERTILITY ON MUCUS SCORE (1), MUCUS SCORE^2 (2), SEMEN CONSISTENCY (3), AND SPERM MOTILITY (4) WITHIN YEARS

Year	Fertility	Partial regression coefficients			
		Trait 1	Trait 2	Trait 3	Trait 4
1957	0·648	+0·491	−0·089	−0·035	−0·011
1958	0·637	+0·217	−0·041	−0·079	−0·016
1959	0·364	+0·178	−0·039	−0·032	−0·050
1960	0·661	+0·384	−0·072	−0·004	−0·013
1961	0·632	+0·255	−0·053	−0·016	−0·022
1962	0·561	+0·231	−0·042	−0·045	−0·006
1963	0·510	+0·274	−0·051	+0·013	−0·042
1964	0·533	+0·524	−0·097	−0·036	−0·058
1965	0·391	−0·008	−0·012	−0·015	−0·096

(proportion of ewes lambing to first insemination) in each year. While there is appreciable variation in the size of coefficients from year to year, there are very few disagreements in sign, the directions in general agreeing with those implicit in Figure 1. The numerical value of the linear effect of mucus score is also consistently larger than the remaining coefficients. Much of the variation in size and sign of coefficients from year to year is no doubt due to the limited numbers within individual years.

The homogeneity test under model 2 for pooling within years by the method given by Rao (1952) consists of tests of two null hypotheses. These are (*a*) that regression equations in different years are equivalent, and (*b*) that the regression equations are equivalent except for the intercepts on the dependent axis. These tests are shown in Table 2.

TABLE 2

TESTS OF REGRESSION HYPOTHESES

Source	DF	MS	*F*
Residuals due to deviations from hypothesis (*a*)	40	9·916	2·762***
Residuals due to deviations from hypothesis (*b*)	32	4·032	1·123 ($P > 0·05$)
Residuals due to separate regressions	268	3·591	

*** $P < 0·001$.

Clearly the first hypothesis is not supported while the second is. Thus, as there is no evidence that the sets of regression coefficients differ between years, and as there are very highly significant differences between intercepts, it is implicit that there are very real mean differences between years in fertility. In view of the large differences in the annual estimates of fertility in Table 1, this is not surprising.

TABLE 3

PARTIAL REGRESSION COEFFICIENTS OF POOLED FERTILITY DATA

Independent variable:	1 Mucus score (linear)	2 Mucus score (quadratic)	3 Semen consistency	4 Sperm motility
Coefficient	0·291***	−0·057***	−0·024**	−0·031***
Standard error	0·046	0·008	0·008	0·007
Standardized coefficient	1·517	−1·714	−0·162	−0·223

** $P < 0·01$. *** $P < 0·001$.

These tests having shown that the rather large differences in fertility of Table 1 were real, but that the various sets of within-year regression coefficients could be considered homogeneous, it seemed reasonable to pool the data on a within-year basis and to re-estimate the regressions as average effects of the several traits on fertility. These estimates of the partial regression coefficients are shown in Table 3, together with their standard errors as estimated under model 1.

All the coefficients are highly significant, indicating real effects of these traits on fertility, with mucus scores apparently being of considerably greater importance than semen traits if the standard partial regression coefficients are regarded as measures of relative importance. The reality and nature of these effects are as might have been expected from published work to which reference has already been made.

The squares of the simple and multiple correlations relating the four variables to fertility on an intra-year basis are shown in Table 4.

TABLE 4

SQUARES OF SIMPLE AND MULTIPLE CORRELATIONS WITH FERTILITY

Trait(s)	Correlation²	Trait(s)	Correlation²
1 (mucus score)	0·028	1, 3, 4	0·130
2 (mucus score²)	0·049	1, 4	0·114
3 (semen consistency)	0·066	2, 3	0·118
4 (sperm motility)	0·087	2, 3, 4	0·152
1, 2	0·119	2, 4	0·135
1, 2, 3	0·195	3, 4	0·102
1, 2, 4	0·212	1, 2, 3, 4	0·231
1, 3	0·096		

Clearly, as these values are so low, there is considerable variation in fertility not controlled by the attributes considered here. This is not surprising when one considers that no account was taken of age differences in the ewes in analysis to this point, or of differences among the observers who made the subjective gradings over the course of the work. A further contribution to error must be the coarse expression of fertility (zero or one) in individual ewe years.

(*b*) *Analyses of Variance of Fecundity*

The main effects for the six traits, as estimated by the least squares procedure and averaged over the eight years, are shown in Table 5 together with their standard

TABLE 5

ESTIMATES OF EFFECTS OF SIX TRAITS ON FECUNDITY TO ONE INSEMINATION AND STANDARD ERRORS (IN PARENTHESIS)

Parameter No.:	1	2	3	4	5	6
Trait						
Mucus score	0·09 (0·03)	0·23 (0·03)	0·21 (0·03)	0		
Semen consistency	0·15 (0·05)	0·04 (0·04)	0·04 (0·04)	0 (0·04)	0	
Sperm motility	0·28 (0·05)	0·26 (0·04)	0·23 (0·05)	0·13 (0·05)	0·06 (0·05)	0
Semen volume	0·05 (0·03)	0·07 (0·03)	0·02 (0·03)	0 (0·03)	0	
Abnormal sperm (%)	0·04 (0·03)	0·03 (0·03)	−0·03 (0·03)	0	−0·05 (0·03)	
Age of ewe	0·05 (0·03)	−0·01 (0·02)	0			

errors. The results of the tests of significance of the six traits in each of the eight years are summarized in Table 6. These tests should of course be viewed conservatively in view of the distribution problems involved. In producing the averages of Table 5 individual annual estimates were weighted by the inverse of their standard errors. It should be noted that the effects here are on fecundity rather than on fertility as used in the regression analyses. A scale of 0, 1, 2, . . . was deemed more appropriate

to analysis of variance than a 0, 1 scale. There were, however, no triplets and only 180 pairs of twins resulting from the 4250 first inseminations analysed here. It is therefore not surprising that these estimates present a picture in good general agreement with Figure 1 and, where applicable, with the regression analyses. The overall effects of volume of semen and percentage of abnormalities are smaller than those of mucus, consistency, and motility. The generally high semen dose rate, and the fact that abnormalities were estimated from a single ejaculate removed in time from those used for insemination, are consistent with these modest effects.

TABLE 6

COMPONENTS OF VARIANCE EXPRESSED AS PERCENTAGES AND SIGNIFICANCE OF MAIN EFFECTS

Year:	1957	1958	1959	1960	1962	1963	1964	1965	Mean
Source									
Error	84·3	92·1	93·6	91·5	90·3	96·2	92·9	83·7	90·6
Age of ewe	2·5*	1·3	—	0·1	1·0	0·5	0·1	0·9	0·8
Abnormal sperm (%)	1·4	—	0·0	—	0·4	—	—	3·1***	0·6
Semen volume	1·4	0·9	0·1	0·1	—	0·2	—	—	0·3
Sperm motility	—	—	2·3*	0·3	3·3**	2·6*	0·5	11·2***	2·5
Semen consistency	4·5*	1·5	2·3**	0·6	1·1	—	—	0·5	1·3
Mucus score	6·0**	4·2**	1·6**	7·3***	4·1***	0·6	6·6***	0·7	3·9

*$P<0·05$. **$P<0·01$. ***$P<0·001$.

The effects of age of dam are of considerable interest in that they contrast sharply with earlier data from the same stocks under natural mating (de Haas and Dunlop 1969) and with data on other Australian Merinos (Turner and Dolling 1965). Here we have fecundity per ewe having a first artificial insemination and present at lambing as opposed to fecundity per ewe present at lambing under natural service conditions and irrespective of whether 0, 1, or more oestrous periods resulted in services. In our data fecundity, as here defined, is higher in two-tooth than in mature ewes. The estimate at this age in Table 5 is just significant if an additional decimal place is considered.

The present data are probably a net result of several influences. Thus 2% of these two-tooth ewes produced twins as a result of the first insemination while the equivalent figure for older ewes was 5%. This trend agrees with the proportions of multiple births produced by these ages in complete lambings (de Haas and Dunlop 1969). It also agrees with previously unpublished data on the proportions of multiple ovulations in 68 two-tooth ewes and 306 older Merino ewes of these stocks ranging from $2\frac{1}{2}$ to $9\frac{1}{2}$ years of age. On laparotomy the fractions of multiple ovulations were found to be 7·4 and 13·4% respectively. The direction of this age difference agrees with that in ovulation rate reported by Giles (1969) in Bungaree Merinos. Secondly, in our material 40% of two-tooth ewes failed to lamb to a first insemination compared with 46% of older ewes. Finally, the fractions of ewes for which an oestrous period was not recorded and which were therefore not included in these analyses were 23% for two-tooths and 5% for older ewes. The make-up of this total is interesting as two drought years were included. The percentages of apparently non-cycling ewes were:

	two-tooth	*older ewes*
drought years	50	18
non-drought years	6	2

Most, though not quite all, of the differences between two-tooth and older ewes come from the drought years when the younger sheep were probably restricted in development. This age difference is in the same direction as that recorded in Merinos by Mullaney (1966), though his fractions of inactive ewes were much lower than ours.

It appears that older ewes, when inseminated once, may be slightly more efficient in producing multiple births by reason of higher ovulation rates, but that with increasing age they became somewhat less efficient in producing a lamb at all. One may speculate that ageing through wear and tear, through infections, and possibly through cumulative effects of plant oestrogens might be expected to bring this about. On the other hand the greater proportion of two-tooth ewes not having a first insemination may be due to (*a*) fewer two-tooth ewes ovulating presumably through immaturity, or (*b*) lesser libido and experience of two-tooth ewes.

The fairly extensive data of de Haas and Dunlop (1969), which came from the same stocks of sheep in earlier years and at three locations, have been re-analysed to estimate age differences in the number of lambs born per ewe to first service under natural mating. The figures were: two-tooth, 0·73; four-tooth, 0·84; older than four-tooth, 0·93. These are to be compared with the age effects expressed as deviations in Table 5, i.e. 0·05, −0·01, and 0·0. Clearly the trend is reversed and the age effects we have found under artificial insemination do not apply under conditions of natural service. This reversal was present in data from each of the locations prior to pooling to produce the first set of figures above. The only explanation which seems acceptable is that under unrestricted natural service the more experienced older ewes may be more successful competitors for male favours, so that these animals receive several services each, spread in varying degrees over the oestrous period. Lambourne (1956) has presented data suggesting that this may be so in the New Zealand Romney. His data also suggest that ewes remain in oestrus longer with increasing age. Some Australian workers (Sinclair 1957; Salamon and Robinson 1962; Dunlop and Tallis 1964), as well as a number of Russian authors summarized by Salamon (1962), have shown that plural services or inseminations increase either the number of ewes lambing or the number of lambs born or both.

In order to judge from the regression analyses, the relative importance of each of the variables, it is informative to look at the degree of statistical control (in terms of R^2 where R is either the multiple correlation or the simple correlation depending on the number of independent variables under consideration) exhibited jointly and separately by various combinations. For example, the control by variables 1 and 2 jointly (denoted by $R^2_{(1,2)}$), summed with $R^2_{(3,4)}$ is 0·221 and is not very different from $R^2_{(1,2,3,4)} = 0{\cdot}231$. This is to be expected on *a priori* grounds, because ewe and ram characteristics are independent. The situation within ewe and ram traits is somewhat different. The fraction of control in variation due to characteristics 1 and 2 jointly ($R^2_{(1,2)} = 0{\cdot}119$) is considerably in excess of that controlled by 1 or 2 separately ($R^2_{(1)} = 0{\cdot}028$, $R^2_{(2)} = 0{\cdot}049$). This is not surprising for it is evident from Figure 1 that there is a pronounced curvilinear effect of mucus score on fertility. However, the excess of joint control of variation over individual control is much less marked in traits 3 and 4 ($R^2_{(3,4)} = 0{\cdot}102$, $R^2_{(3)} = 0{\cdot}087$, $R^2_{(4)} = 0{\cdot}066$) so that there is less to be gained by considering them jointly. The simple correlations of traits 1 and 2 and of 3 and 4 are both positive (0·98 and 0·51 respectively). The correlation between 1 and 2

must be strongly positive as 2 is the square of 1 which is always positive. The correlation between consistency (3) and motility (4) is also expected to be positive because in less concentrated samples of semen the swirling wave motion characteristic of the numerically lower scores of motility is absent, irrespective of the degree of activity of individual sperm. In spite of the fact that both of these simple correlations are positive they have a different effect on the joint control of variation when taken pairwise or separately. For ewe characteristics 1 and 2 the curvilinear relationship with fertility has boosted the importance of their joint effect whilst for ram characteristics 3 and 4 (both approximately linearly related to fertility) their mutual correlation has rendered the joint effect only slightly more important than their separate effects.

While one may look at the linear and quadratic contributions of mucus score separately, it seems more meaningful to compare the total control of fertility by mucus score with that by consistency or by motility. Such a comparison leads to the conclusion that mucus score has somewhat more control of fertility than either consistency or motility of the semen, though not markedly more than their joint effects.

Light may also be thrown on the same problem by consideration of the components of variance from the analysis of variance of the data classified on six criteria. The components estimated for individual years and expressed as percentages also appear in Table 6 together with the simple means of the annual percentages. Where the estimate of a variance component was negative it was assumed to be zero and this is denoted by a dash. The summary of tests of significance of main effects and the components of variance as expressed in Table 6 support conclusions already drawn from the regression analyses, as well as the suggestions from the earlier graphical presentation of the raw bulk data. Thus mucus score was by far the most consistently significant source of variation and the variance component due to this cause was considerably the largest of the six classifications considered. The effects of motility were significant in four years out of eight and the effects (Table 5) were at least as large in some cases as those of mucus score. This trait ranked second in the size of variance component. Consistency appeared to be the next most important trait, being significant in only two years, and having smaller effects and a variance component averaging only about half as large as that due to motility.

Of the three remaining traits, age and percentage of abnormalities were each significant only in a single year while the effects of semen volume did not have a significant effect on fecundity in any individual year. Each of these traits was responsible for less than 1% of the variation.

In the three traits mucus score, motility, and consistency, which were analysed both by regression and analysis of variance, it is of interest to draw together the various measures by which one might consider their relative importance. This is done in Table 7.

By every criterion but one the descending order of importance is mucus, motility, consistency. The exception is the size of effects from the analysis of variance. An examination of Figure 1 and Table 5 will make it clear that this contradiction is more apparent than real. It will be noted that in solving the least squares equations, the terminal effect has been set equal to zero (except in the case of percentage abnormalities where the effects are in any case small). This has meant that in the linear cases the largest effect is the sum of differences of like sign between successive adjacent effects.

Had a linearizing transformation been applied to the mucus data the preceding statement could have applied to this trait also. It would then in all probability have stood first in size of effects. The present effect of score 1 can be looked on as the sum of differences of unlike sign.

TABLE 7

MEASURES OF IMPORTANCE OF THREE TRAITS

Trait: / Measure	Mucus			Sperm motility	Semen consistency
	Linear	Joint	Quadratic		
Standard partial regression coefficient	1·52		−1·71	−0·22	−0·16
r^2 or R^2		0·12		0·09	0·07
Order of size of effects		2		1	3
Times significant		6/8		4/8	2/8
Mean component percentage		3·9		2·5	1·3

In passing it seems worth while recording that the control of variation in fertility within individual years by mucus score (linear + quadratic) may be positively related to fertility and that this is more marked when control by mucus score is expressed as a fraction of the total variation controlled ($R^2_{(1,2)}/R^2_{(1,2,3,4)}$). There is naturally an inverse relation between fertility and control of variation by consistency and motility. If these relationships were confirmed in considerable quantities of independent data, and if it were shown to apply to fecundity as well as to fertility, there could be practical implications for the sheep husbandman who is able before mating to predict successfully the overall outcome of the following lambing.

In considering any possible applications of the present work, it is likely that, while the normal features of good husbandry of sires, such as avoiding nutritional deficiencies, heat stress, and infections, will result in some benefits through improved semen quality, the greatest gains will accrue from efforts to ensure that insemination takes place at the optimum stage of the oestrous cycle as measured by the condition of the vaginal mucus. One may reasonably guess that this will apply to individually controlled natural service (hand service) as well as to artificial insemination. Unfortunately, however, to thus optimize mucus score at insemination will incur a considerable labour cost in determining the time of onset of oestrus and in inseminating at intervals through 24 hour days. The prospective user will have to make his own cost benefit calculations. A cheaper and possibly less effective alternative or addition is to undertake multiple insemination. We have seen that there is evidence of its benefits. Certainly the cheapest alternative, unless there are overwhelming considerations to the contrary, is to revert to natural service, when the multiple services that usually take place will ensure that one of these at least is reasonably close to the optimum stage of the oestrous cycle and that one profits by any additional benefits of multiple services, however these are mediated.

IV. Acknowledgments

It is a pleasure to record our appreciation of the major assistance of Mr. E. K. Yates among the many helpers in the field, and of the contributions made in analyses of data by Misses E. Smith and E. McKay.

V. References

DUNLOP, A. A., and TALLIS, G. M. (1964).—The effects of length of oestrus and number of inseminations on the fertility and twinning rate of the Merino ewe. *Aust. J. agric. Res.* **15**, 282–8.

EMMENS, C. W., and ROBINSON, T. J. (1962).—*In* "The Semen of Animals and Artificial Insemination", ed. J. P. Maule, pp. 205–51. Tech. Bull. Commonw. Bur. Anim. Breed. Genet. No. 15.

GILES, J. R. (1969).—Fertilisation and embryo mortality in Bungaree Merinos mated in autumn. *Aust. J. exp. Agric. Anim. Husb.* **9**, 377–80.

GUNN, R. M. C., SANDERS, R. N., and GRANGER, W. (1942).—Studies in fertility in sheep. 2. Seminal changes affecting fertility in rams. Coun. scient. ind. Res. Aust. Bull. No. 148.

DE HAAS, H. J., and DUNLOP, A. A. (1969).—The effects of some variables on the components of reproduction rate in the Merino. *Aust. J. agric. Res.* 20, 549–59.

HENDERSON, C. R. (1953).—Estimation of variance and covariance components. *Biometrics* **9**, 226–52.

HULET, C. V., and ERCANBRACK, S. K. (1962).—A fertility index for rams. *J. Anim. Sci.* **21**, 489–93.

HULET, C. V., FOOTE, W. C., and BLACKWELL, R. L. (1965).—Relationship of semen quality and fertility in the ram to fecundity in the ewe. *J. Reprod. Fert.* **9**, 311–15.

KARDYMOVIC, M., MARSAKOVA, A., and PAVLJUCUK, V. (1934).—Insemination of sheep at different times during oestrus. *Problemÿ Zivot.* **5**, 110–15.

KEAST, J. C., and MORLEY, F. H. W. (1949).—Some observations on artificial insemination of sheep. *Aust. vet. J.* **25**, 281–7.

LAMBOURNE, L. J. (1956).—Mating behavour. Proc. Ruakura Fmrs' Conf. Week, pp. 6–20.

MORRANT, A. J., and DUN, R. B. (1960).—Artificial insemination of sheep. II. Techniques and equipment used at Trangie Agricultural Experiment Station. *Aust. vet. J.* **36**, 1–7.

MULLANEY, P. D. (1966).—Prenatal losses in sheep in western Victoria. *Proc. Aust. Soc. Anim. Prod.* **6**, 56–9.

RAO, C. R. (1952).—"Advanced Statistical Methods in Biometric Research." (John Wiley & Sons: New York.)

RESTALL, B. J. (1961).—Proc. Conf. Artificial Breeding of Sheep in Australia held at Univ. of New South Wales, August 1961, pp. 67–75.

SALAMON, S. (1962).—M.Sc. Thesis, Univ. of Sydney.

SALAMON, S., and ROBINSON, T. J. (1962).—Studies on the artificial insemination of Merino sheep. I. The effects of frequency and season of insemination, age of the ewe, rams, and milk diluents on lambing performance. *Aust. J. agric. Res.* **13**, 52–68.

SINCLAIR, A. N. (1957).—The effect of variation of time of mating, mating frequency, and semen dose rate on conception in Merino sheep. *Aust. vet. J.* **33**, 88–91.

TALLIS, G. M. (1964).—The use of models in the analysis of some classes of contingency tables. *Biometrics* **20**, 832–9.

TURNER, HELEN N., and DOLLING, C. H. S. (1965).—Vital statistics for an experimental flock of Merino sheep. II. The influence of age on reproductive performance. *Aust. J. agric. Res.* **16**, 699–712.

WIGGINS, E. L., TERRILL, C. E., and EMIK, L. O. (1953).—Relationships between libido, semen characteristics and fertility in range rams. *J. Anim. Sci.* **12**, 684–96.

A (c)[4]

THE RELATIONSHIP BETWEEN LIVE MEASUREMENTS AND EDIBLE MEAT IN MERINO WETHERS

By G. M. TALLIS,* HELEN NEWTON TURNER,† and G. H. BROWN†

[*Manuscript received August* 29, 1963]

Summary

Seventy-five Merino wethers of a medium Peppin strain were slaughtered at 7 months of age after a series of live measurements had been made, weights then being taken of carcass and of edible meat after boning out. Fat content was insufficient for trimming, and bone weight was obtained by difference.

Weight of edible meat was highly correlated (0·95) with liveweight before slaughter, and the inclusion of any other measurement in a multiple correlation analysis failed to raise this value. Variation in bone weight contributed only 25% of the variation in carcass weight, and the ratio of meat to bone was positively correlated (0·54) with liveweight before slaughter.

If total amount of edible meat is accepted as the criterion for meat production, liveweight before slaughter was a satisfactory predictor for these sheep. It is suggested that simplification of criteria along these lines is desirable to aid in the selection of sheep for meat production, though more work is required on sheep of other ages and other breeds.

I. INTRODUCTION

In the past, relatively few investigations of the relationships between live measurements and carcass characteristics have been carried out for sheep. Most of the work so far reported has been exclusively concerned with the study of factors influencing growth, fertility, wool production, and efficiency (Taneja 1955; Cassard *et al.* 1956) on the one hand, and various measurements and scores associated with carcass grading on the other (Robinson, Binet, and Doig 1956; Kemp, Bull, and Bear 1953). However, any experiment which is connected with the production of meat necessarily demands the joint consideration of both aspects.

In order to devise procedures for selecting sheep for mutton production, it is essential first of all to define mutton type. Once this step is made, the next requirement is to establish suitable tools which can be used in a selection programme and can be applied to exert selection pressure for the defined type. The above-mentioned problem of definition is not an easy one, since it will depend largely on market and other commercial factors, the relative importance of each not being easily disentangled.

For the purpose of the present work, however, and as a first approximation, we classify those sheep which at a specified age yield the maximum weight of edible meat as being the most desirable mutton type. Admittedly, this definition is defective in several respects as it assumes, among other things, that the distribution and quality of the meat on the carcass is unimportant. There is no doubt that these features should be examined more closely, but in order to make initial progress the simplified definition of desirable type is subsequently assumed here.

*Division of Mathematical Statistics, CSIRO, McMaster Laboratory, Glebe, N.S.W.
†Division of Animal Genetics, CSIRO, North Ryde, N.S.W.

In brief, therefore, the aim of the present investigation is to find some live body measurements of sheep which are at the same time repeatable and useful predictors of total edible meat. This work falls naturally into two sections and is reported below.

TABLE 1

ANALYSIS OF REPEATED MEASUREMENTS ON THE SAME SHEEP

Character	Components of Variance*					Mean	Coeff. of Variation	Repeatability*	
	$\hat{\sigma}_e^2$	$\hat{\sigma}_s^2$	$\hat{\sigma}_t^2$	$\hat{\sigma}_{st}^2$	$\hat{\sigma}^2$	$\hat{\mu}$	$\hat{V}$	$\hat{r}_1$	$\hat{r}_2$
Wither to pinbone	0·300	0·499	0·127	0·084	1·010	23·70	4·2	0·49	0·58
Width of hips	0·002	0·089	0·000	0·002	0·093	6·09	5·0	0·96	0·97
Depth of chest	0·059	0·124	0·004	0·032	0·219	11·49	4·1	0·57	0·66
Elbow to coronet	0·012	0·231	0·015	0·005	0·262	15·60	3·3	0·88	0·90

* Components of variance: $\hat{\sigma}_e^2$, between repeated measurements at the same time on the same sheep.
$\hat{\sigma}_s^2$, between sheep.
$\hat{\sigma}_t^2$, between times on the same sheep.
$\hat{\sigma}_{st}^2$, sheep × times interaction.
$\hat{\sigma}^2 = \hat{\sigma}_e^2 + \hat{\sigma}_s^2 + \hat{\sigma}_t^2 + \hat{\sigma}_{st}^2$.

Repeatability: $r_1 = \sigma_s^2/\sigma^2$.
$r_2 = \sigma_s^2/(\sigma^2 - \frac{1}{2}\sigma_e^2)$.

II. EXPERIMENTAL

(*a*) *Repeatability*

Prior to the main experiment, four body dimensions were measured on 40 live Merino wethers at the CSIRO field station at Armidale, N.S.W. The dimensions were wither to pinbone, width at hips, depth of chest, and elbow to coronet, all being measured in accordance with the standards set out by Turner *et al.* (1953). The sheep were caught and all characters measured twice by a single observer. Each animal was subsequently re-caught and another two sets of measurements were taken by the same observer. Thus there was a total of four measurements for every character on each wether, and suitable partitioning of the total variance for the four characters was subsequently carried out by the standard analysis of variance routine.

The estimated components of variance, means, coefficients of variation, and repeatability of measurements by a single observer are given in Table 1. The total variance, σ^2, is defined as the sum of the error, sheep, times, and sheep × times components, $\sigma^2 = \sigma_e^2 + \sigma_s^2 + \sigma_t^2 + \sigma_{st}^2$, obtained from the analysis of variance table, and the coefficient of variation and the estimate of repeatability are given by $V = 100\sigma/\mu$ and $r_1 = \sigma_s^2/\sigma^2$ respectively.

TABLE 2

CORRELATION COEFFICIENTS, MEANS, AND STANDARD DEVIATIONS

For definitions see text

Variable	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}	x_{11}	x_{12}	x_{13}	x_{14}	x_{15}	Mean $\hat{\mu}_t$	S.D. $\hat{\sigma}_t$
x_1	1	0·99	0·62	0·48	0·90	0·96	0·27	0·79	0·81	0·84	0·83	0·69	0·92	0·53	0·87	26·72	4·54
x_2		1	0·69	0·45	0·89	0·95	0·29	0·77	0·81	0·82	0·81	0·67	0·92	0·53	0·86	18·64	3·42
x_3			1	0·11	0·50	0·54	0·29	0·39	0·52	0·44	0·46	0·28	0·56	0·36	0·51	69·60	1·64
x_4				1	0·51	0·56	−0·13	0·43	0·43	0·52	0·56	0·58	0·48	0·19	0·50	9·22	1·63
x_5					1	0·92	0·31	0·88	0·79	0·74	0·78	0·68	0·88	0·40	0·88	43·76	8·53
x_6						1	0·21	0·83	0·81	0·84	0·83	0·73	0·96	0·50	0·93	61·04	8·43
x_7							1	−0·14	0·13	0·12	0·14	0·05	0·23	0·12	0·22	226·00	10·29
x_8								1	0·74	0·68	0·71	0·64	0·79	0·36	0·80	0·35	0·07
x_9									1	0·74	0·72	0·68	0·75	0·39	0·72	21·00	1·02
x_{10}										1	0·80	0·74	0·77	0·72	0·59	5·30	0·34
x_{11}											1	0·82	0·72	0·34	0·72	9·85	0·53
x_{12}												1	0·52	0·06	0·60	13·84	0·63
x_{13}													1	0·61	0·92		
x_{14}														1	0·26		
x_{15}															1		

It will be noticed from Table 1 that the V's are consistently small and range from 3 to 5%. On the other hand, whereas the repeatabilities of width of hips and elbow to coronet measurements are high, those for depth of chest and wither to pinbone are unsatisfactorily low. In the case of the latter two measurements, an inspection of the components of variance reveals that maximum increase in precision is achieved by repeated measurements at a single catching. Since this is also the most economical in terms of time and labour, each character was measured twice in later studies. It is estimated that this procedure increases repeatability to $r_2 = \sigma_g^2/(\sigma^2 - \frac{1}{2}\sigma_e^2)$, recorded in the last column of Table 1.

(*b*) *Live Measurements and the Boned-out Carcass*

In May 1962, 75 Merino wethers approximately 7 months old were weighed and measured at the CSIRO field station at Deniliquin. The sheep were subsequently despatched to Sydney for slaughter at the Sydney Meat Preserving Co., and the carcasses taken to a commercial boning-out works. Here the carcasses were weighed and then stripped of all edible meat, which also was weighed. The weight of the skeleton was obtained as the difference between the carcass weight and the total weight of edible meat. It was originally intended that the fat would be trimmed off and weighed separately, but the amount present was negligible and could not be separated.

From the above measurements and the records of the individual animals the following data were available for each wether:

x_1 = carcass weight (lb),
x_2 = total weight of edible meat in the carcass (lb),
x_3 = percentage of edible meat = $100x_2/x_1$,
x_4 = birth weight (lb),
x_5 = weaning weight (lb),
x_6 = live slaughter weight (lb),
x_7 = age at slaughter (days),
x_8 = average daily gain (lb) = $(x_6 - x_4)/x_7$,
x_9 = length of live animal from wither to pinbone (in.),
x_{10} = width at hips (on live animal — in.),
x_{11} = depth of chest (on live animal — in.),
x_{12} = elbow to coronet (on live animal — in.),
$x_{13} = x_6/x_{12}$,
$x_{14} = x_{10}/x_{12}$,
$x_{15} = x_6/x_{10}$.

The means and standard deviations of characters x_1 to x_{12} are given in Table 2, together with the estimated correlation matrix of all 15 characters.

The ratios x_{13}, x_{14}, and x_{15} were calculated to see whether or not such compounds are of greater predictive value for x_2 and x_3 than the raw measurements individually. Clearly, however, both x_{13} and x_{15} are so closely correlated with x_6 that they can provide little, if any, information additional to that already provided by x_6 alone. On the other hand, x_{14} is not so highly correlated with its components and may therefore provide additional information concerning conformation. Nevertheless, it does not appear to be a valuable index for predicting either x_2 or x_3.

There are numerous interesting relationships which may be inferred from Table 2. Obviously the most important single character for predicting x_2 is x_6, and their correlation coefficient is estimated as 0·95. However, since, by definition, it is the total weight of edible meat at a fixed age which is important, $\rho_{2,6\cdot7}$ was also estimated and was again found to be 0·95. A multiple regression analysis was then carried out with x_2 as the dependent variable and x_4, x_6, x_7, x_8, x_9, x_{10}, x_{11}, x_{12} as independent variables. The multiple correlation coefficient, $\hat{R}$, was found to be 0·96, which indicated that no sensible improvement in prediction power had been achieved by considering additional variables to x_6. The analysis was then re-run keeping age x_7 fixed, and the results again showed that $\hat{R}$ was not appreciably greater than $\hat{\rho}_{2,6\cdot7}$.

There are too many other relationships to be discussed in detail here. However, attention is drawn to the rather high correlation between x_4 and x_2 which suggests, perhaps, that early selection for x_2 is a possibility. It is also encouraging, although hardly surprising, to notice that x_2 is highly correlated with average daily gains, x_8. Thus those animals which are most satisfactory from the producer's viewpoint may also be the most satisfactory for the buyer.

Since the variances of carcass weight and weight of edible meat are 20·33 and 11·71 and the covariance between the two is 15·35, the variance of the weight of skeleton is found to be 1·34. This emphasizes the relatively small contribution of skeleton to the variance of carcass weights (about 25%) and, therefore, differences in such weights may be chiefly attributable to differences in meatiness.

As a final point of interest, the correlation between x_6 and the ratio of edible meat to bone was calculated and was found to be 0·54. This relationship, again, is not surprising from developmental considerations, but it further suggests that selection carried out as suggested above would tend to increase the proportion of edible meat available in the carcass.

III. Conclusions and Discussion

The main conclusion which may be drawn from this work is quite clear. If one accepts the definition of carcass quality given in the introduction, then the results here reported strongly suggest that selection on body weight, preferably age-corrected, will usually pick those sheep which have the greatest weight of edible meat at a given age, without the aid of any other live measurement. As the heritability of body weight, including weaning weight, is high (Young, Turner, and Dolling 1960; Young *et al.* 1964), there should be response to selection for it; and, although no genetic correlations between body weight and amount of edible meat are available, the very high phenotypic correlation makes it reasonable to assume that there would be a strong correlated response. Moreover, body weight is the only measurement required, since the inclusion of measurements on additional characters will probably not appreciably affect the selection pressure exerted on total weight of edible meat. In addition, if criteria other than total weight of edible meat are considered, then ratios of body measurements involving body weight probably will not provide much additional information to body weight alone. Therefore, in this case it may be ratios such as x_{14}, x_{10}/x_{11}, x_{11}/x_{12}, and so on which are important.

However, this experiment was conducted on only a moderately large sample of Merino wethers, at a single age, and the results require confirmation, particularly at other ages and with other breeds. Young (personal communication) has obtained estimates of over 0·7 for the repeatability of body weight for Merino rams weighed at 6, 12, and 18 months, but the bone–muscle–fat relationships at 12 months of age or more still require investigation.

A next step in generalizing somewhat the restrictive definition of carcass quality may be to regard those carcasses which are worth most to the butcher as being the most desirable. Thus the carcass may be thought of as cut into retail joints, each joint weighed, and the total value of the carcass calculated from these weights and some average price per pound. However, the coarser the grouping of retail cuts, the more hope there is of finding practical selection tools, and initially it may be satisfactory only to consider, say, the relative weights of the front to the back of the carcass, where separation is at the 13th rib. It is worth noticing that the oversimplified definition given in Section I assigns equal value to the meat of all cuts and is, therefore, a special case of the new definition.

In any case, whatever steps are taken to broaden the investigation, it seems essential to keep the objectives as clear and simple as possible if positive results are to be obtained. Because of the general confusion with regard to the definition of carcass and meat qualities, it may be best to oversimplify the issue initially and to introduce complications when (and if) they arise.

Finally, other interesting aspects of the problem of producing meat are to be found in the interrelations of food efficiency, daily gain, and the growth of an optimum carcass. One might guess that the fastest-gaining animals are generally the most efficient (Knapp and Baker 1944) and may also produce optimum carcasses.

No matter what the answers may be, these questions must be cleared up before satisfactory tools can be developed for the selection of mutton sheep. Furthermore, since it is clear that a suitable definition of carcass quality depends on the type of market for which the animals are intended, several definitions may be necessary. For instance, the definition for the grower of prime lambs will be different from that of the grower selling older sheep for slaughter. However, only when matters of definition and selection criteria have been settled will it be possible to investigate the situation genetically with a view to developing selection indexes and predictors of genetic progress under selection.

IV. References

Cassard, D. W., Gregory, P. W., Weir, W. C., and Wilson, J. F. (1956).—Environmental factors affecting body dimensions in yearling Hampshire ewes. *J. Anim. Sci.* 15: 922–9.

Kemp, J. D., Bull, S., and Bear, H. W. (1953).—The economy and nutritive value of different cuts of lamb of different grades. *J. Anim. Sci.* 12: 338–46.

Knapp, B., jr., and Baker, A. L. (1944).—Correlation between rate and efficiency of gain in steers. *J. Anim. Sci.* 3: 219–23.

Robinson, T. J., Binet, F. E., and Doig, A. G. (1956).—Fat lamb studies in Victoria. I. An assessment of the relative value of various external measurements for differentiating between various grades of export lamb carcasses. *Aust. J. Agric. Res.* 7: 345–65.

Taneja, G. C. (1955).—Mutton qualities in Australian Merino sheep. *Aust. J. Agric. Res.* 6: 882–90.

TURNER, HELEN NEWTON, HAYMAN, R. H., RICHES, J. H., ROBERTS, N. F., and WILSON, L. T. (1953).—Physical definition of sheep and their fleece. CSIRO Aust., Div. Anim. Hlth. Prod. Rep. No. 4 (Ser. S.W.-2).

YOUNG, S. S. Y., BROWN, G. H., TURNER, HELEN NEWTON, and DOLLING, C. H. S. (1964).—Genetic and phenotypic parameters for body weight and greasy fleece weight at weaning in Australian Merino sheep (in preparation).

YOUNG, S. S. Y., TURNER, HELEN NEWTON, and DOLLING, C. H. S. (1960).—Comparison of estimates of repeatability and heritability for some production traits in Merino rams and ewes. II. Heritability. *Aust. J. Agric. Res.* 11: 604–17.

(d)

THEORETICAL DEVELOPMENTS PERTINENT TO SECTIONS A(b) AND A(c)

Commonwealth of Australia
COMMONWEALTH SCIENTIFIC AND INDUSTRIAL RESEARCH ORGANIZATION
Reprinted from "Biometrics"
Vol. 20, No. 4 Pages No. 832-839 December 1964

THE USE OF MODELS IN THE ANALYSIS OF SOME CLASSES OF CONTINGENCY TABLES

G. M. Tallis[1]

Division of Mathematical Statistics, C.S.I.R.O., McMaster Laboratory, Glebe, N.S.W., Australia

INTRODUCTION

In some types of biological experiments, it frequently happens that the results take the form of a k-dimensional table, the cells giving the outcomes of independent multinomial trials. This would be the case, for instance, in an experiment to determine the viability of a particular organism in different environments. The experimenter may be interested in testing the effects of r different media and c temperature conditions on the survival rate of his test organism, at a particular age. For this purpose the experimenter may set up rc tubes in such a way that i, jth tube contains the ith medium, is held at the jth temperature level and holds a random sample of N_{ij} of his organisms. After the elapse of a predetermined period of time, he counts the number of organisms surviving in each tube and he is then interested in analysing the effects of temperature and media on the proportion of survivors. His data may be arranged in an $r \times c$ table recording the proportion of survivors, and this example is obtained as a special case from the more general situation by using $k = 2$ and using binomial trials.

It is the purpose of this paper to suggest a method of analysing such data, but the development is mainly in terms of two-dimensional tables with independent binomial trials. Since extensions to more involved situations are easily effected, more complicated cases will receive no further attention here. The same methods of analysis may also be used in somewhat different circumstances and the necessary modifications for dealing with $r \times c$ contingency tables in the same way are indicated as an example.

METHODS

It will subsequently be assumed that we have an $r \times c$ table of independent binomial trials with parameters p_{ij} and N_{ij} and outcomes

[1]Now at Department of Biostatistics, Johns Hopkins University, Baltimore, Md., U.S.A.

n_{ij}, $i = 1, 2, \cdots, r$; $j = 1, 2, \cdots c$. Suppose now we write the parameter model

$$p_{ij} = p + \alpha_i + \beta_j + \gamma_{ij}, \tag{1}$$

then it is required to estimate the $(r + 1)(c + 1)$ unknown parameters p, α_i, β_j and γ_{ij}. The analogy between (1) and the standard two-way analysis of variance model is obvious and we use the same identifiability constraints:

$$\begin{aligned} h_q &\equiv \sum_j \gamma_{qj} = 0, \qquad 1 \leq q \leq r; \\ h_{r+s} &\equiv \sum_i \gamma_{is} = 0, \qquad 1 \leq s \leq c - 1; \\ h_{r+c} &\equiv \sum_i \alpha_i = 0; \\ h_{r+c+1} &\equiv \sum_j \beta_j = 0. \end{aligned}$$

We now write the full Likelihood equation, L,

$$L = C \prod_{ij} p_{ij}^{n_{ij}} q_{ij}^{N_{ij}-n_{ij}} \tag{2}$$

where C is a constant which does not depend on any of the unknown parameters. It is found by standard procedures that L is maximised when

$$n_{ij}/N_{ij} = \hat{p}_{ij} = \hat{p} + \hat{\alpha}_i + \hat{\beta}_j + \hat{\gamma}_{ij}, \quad i = 1, 2, \cdots r, \quad j = 1, 2, \cdots c.$$

By using the constraint relations $\sum \hat{\alpha}_i = \sum \hat{\beta}_j = \sum_i \hat{\gamma}_{ij} = \sum_j \hat{\gamma}_{ij} = 0$, it follows that

$$\begin{aligned} \hat{p} &= \sum_{ij} \hat{p}_{ij}/rc, \qquad \hat{p} + \hat{\beta}_j = \sum_{i=1}^{r} \hat{p}_{ij}/r \\ \hat{p} + \hat{\alpha}_i &= \sum_{j=1}^{c} \hat{p}_{ij}/c, \qquad \hat{\gamma}_{ij} = \hat{p}_{ij} - \hat{p} - \hat{\alpha}_i - \hat{\beta}_j . \end{aligned} \tag{3}$$

The large sample variances of these estimates, which are obviously unbiased, can be calculated directly, but a more systematic approach is to compute the elements of the information matrix, I, and to invert it. If we use the notation $I(\theta)$ and $I(\theta, \phi)$ to represent the information and co-information associated with the parameters θ and ϕ, then

$$\begin{aligned} I(p) &= \sum_{ij} N_{ij}(p_{ij}q_{ij})^{-1} \\ I(\alpha_i) &= \sum_j N_{ij}(p_{ij}q_{ij})^{-1} = I(\alpha_i, p) \\ I(\beta_j) &= \sum_i N_{ij}(p_{ij}q_{ij})^{-1} = I(\beta_j, p) \\ I(\gamma_{ij}) &= N_{ij}(p_{ij}q_{ij})^{-1} = I(\alpha_i, \beta_j) = I(\gamma_{ij}, p) = I(\alpha_i, \gamma_{ij}) = I(\beta_j, \gamma_{ij}) \end{aligned} \tag{4}$$

and all other elements of I are zero.

It is convenient at this stage to relabel the parameters as follows: $\alpha_i = \theta_i$ for $1 \leq i \leq r$, $\beta_j = \theta_{r+j}$ for $1 \leq j \leq c$, $\gamma_{st} = \theta_{r+rc+t}$ for $1 \leq s \leq r$ and $1 \leq t \leq c$, and $p = \theta_{(r+1)(c+1)}$. Now the constraint equations can be written $h_i(\theta) = 0$, $1 \leq i \leq r + c + 1$ and, since I is singular, another matrix HH′ where $\mathrm{H} = (h_{ij})$, $h_{ij} = \partial h_i(\theta)/\partial\theta_j$, must be added (Aitchison and Silvey [1960]). The new matrix $\mathrm{I} + \mathrm{HH}'$ is non-singular and of rank $(r + 1)(c + 1)$ and, if we let

$$\begin{bmatrix} \mathrm{I} + \mathrm{HH}' & -\mathrm{H} \\ -\mathrm{H}' & 0 \end{bmatrix}^{-1} = \begin{bmatrix} \mathrm{U} & \mathrm{V} \\ \mathrm{V} & \mathrm{W} \end{bmatrix},$$

then U is an estimate of the covariance matrix of the estimated parameters. In this case, of course, H consists of entries of 0's and 1's and can readily be written down by inspection.

As an example let $r = c = 2$. Then the relevant restrictions are specified by

$$h_1 \equiv \alpha_1 + \alpha_2 = h_2 \equiv \beta_1 + \beta_2 = h_3 \equiv \gamma_{11} + \gamma_{12} = h_4 \equiv \gamma_{21} + \gamma_{22} = h_5 \equiv \gamma_{11} + \gamma_{21} = 0$$

and

$$\mathrm{H} = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & 0 \\ 0 & 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix}, \qquad \mathrm{HH}' = \begin{bmatrix} 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 2 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 2 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}.$$

However, in this case the problem is best handled by building the constraints directly into the models for the probabilities. Thus

$$\begin{aligned} p_{11} &= p + \alpha + \beta + \gamma, & p_{12} &= p + \alpha - \beta - \gamma, \\ p_{21} &= p - \alpha + \beta - \gamma, & p_{22} &= p - \alpha - \beta + \gamma, \end{aligned} \tag{5}$$

and there are effectively four parameters, which are estimated by

$$\hat{p} = \sum_{i,j}^{2} \hat{p}_{ij}/4, \qquad \hat{\alpha} = (\hat{p}_{11} + \hat{p}_{12} - \hat{p}_{21} - \hat{p}_{22})/4, \tag{6a}$$

$$\hat{\beta} = (\hat{p}_{11} + \hat{p}_{21} - \hat{p}_{12} - \hat{p}_{22})/4, \quad \hat{\gamma} = \hat{p}_{11} - \hat{p} - \hat{\alpha} - \hat{\beta}. \tag{6b}$$

The variances of the estimates are easily obtained directly and it is found that they are all equal to $(16)^{-1} \sum_{ij} N_{ij}^{-1} p_{ij} q_{ij}$.

Missing cells

It may happen that no results are available for several cells of the $r \times c$ table. This situation, although inconvenient, can be dealt with, and the procedure will be illustrated for one missing cell.

If the (m, n)th cell is missing, there are still $r + c + 1$ equations, but the parameter γ_{mn} cannot be estimated. The analysis remains the same except that γ_{mn} is deleted from the parameter set and the constraint equations are modified by letting $\gamma_{mn} = 0$.

We have

$$\hat{p}_{ij} = \hat{p} + \hat{\alpha}_i + \hat{\beta}_j + \hat{\gamma}_{ij} , \qquad i, j \neq m, n,$$

and

$$\sum_{i,j \neq m,n} \hat{p}_{ij} = (rc - 1)\hat{p} - \hat{\alpha}_m - \hat{\beta}_n = \hat{p}'_{..}$$

$$\sum_{i \neq n} \hat{p}_{mj} = (c - 1)(\hat{p} + \hat{\alpha}_m) - \hat{\beta}_n = \hat{p}'_{m.}$$

$$\sum_{i \neq m} \hat{p}_{in} = (r - 1)(\hat{p} + \hat{\beta}_n) - \hat{\alpha}_m = \hat{p}'_{.n} .$$

Setting

$$\hat{d}_{m.} = \hat{p}'_{m.}/(c - 1) - \hat{p}'_{..}/(rc - 1)$$

$$= +\hat{\alpha}_m rc/(rc - 1) - \hat{\beta}_n c(r - 1)/(c - 1)(rc - 1)$$

$$\hat{d}_{.n} = \hat{p}'_{.n}/(r - 1) - \hat{p}'_{..}/(rc - 1)$$

$$= -\hat{\alpha}_m (c - 1)r/(r - 1)(rc - 1) + \hat{\beta}_n rc/(rc - 1)$$

it is found that

$$\hat{\alpha}_m = \hat{d}_{m.} + \hat{d}_{.n}(r - 1)/(c - 1)r$$

$$\hat{\beta}_n = \hat{d}_{.n} + \hat{d}_{m.}(c - 1)/(r - 1)c.$$

The covariance matrix of the estimates is obtained as previously. However, the rank of $I + HH'$ is now $rc + r + c$. If there are several missing cells, it may be best to carry out a standard iterative solution of the equations

$$\hat{p}_{ij} = \hat{p} + \hat{\alpha}_i + \hat{\beta}_j + \hat{\gamma}_{ij} .$$

First guesses of the true M. L. estimates may be obtained from equations (3) suitably modified for number deficiencies. The iteration proceeds according to Aitchison and Silvey (*loc. cit.*).

MODIFICATIONS AND EXTENSIONS

The extension of these methods to tables of any dimension can be effected in an obvious way. When there are no missing cells the M. L. estimates are easily obtained, although the number of parameters rapidly increases with k. Hypotheses with regard to subsets of the parameter set $\{\theta_i\}$ are tested by introducing additional equations of constraint: again reference is made to Aitchison and Silvey, where the procedure is thoroughly discussed in terms of the Wald test and the Lagrange-multiplier test criteria.

As further illustration, model (1) can also be used in the analysis of $r \times c$ contingency tables where the margins are not held fixed. In this instance the likelihood function takes the multinomial form

$$L = C \prod_{ij} p_{ij}^{n_{ij}} \tag{8}$$

where $\sum_{ij} n_{ij} = N$ and $\sum_{ij} p_{ij} = 1$. It is seen that $p = 1/rc$, always, and therefore does not need to be estimated. We have therefore $r + c + rc$ parameters and $r + c + 1$ restrictions which leaves a total of $rc - 1$ independent parameters to be estimated. The M. L. estimates are of the same form as (3) with $\hat{p}_{ij} = n_{ij}/N$. The information matrix can be constructed from the relations,

$$I(\alpha_i) = N \sum_{j=1}^{c} p_{ij}^{-1}, \qquad I(\beta_j) = N \sum_{i=1}^{r} p_{ij}^{-1}, \tag{9}$$

$$I(\alpha_i, \beta_j) = I(\beta_j, \gamma_{ij}) = I(\alpha_i, \gamma_{ij}) = I(\gamma_{ij}) = N p_{ij}^{-1},$$

the remaining terms being zero. The adjustment matrix **H** is again easily written down by inspection and the same theory as in the previous section applied.

The use of these techniques in any problem must be made with caution and the applicability of the model ascertained. It is not true, for instance, that if all $\gamma_{ij} = 0$ then there is independence in the table. As a matter of fact, if there is statistical independence, each γ_{ij} must satisfy the relation $\gamma_{ij} = rc\alpha_i\beta_j$. Thus, independence can be checked by these means but this particular hypothesis is probably best tested by the standard chi-square procedure.

Returning now to the case where the cells are statistically independent, the situation may arise that a model of the form

$$\hat{p}_{ij} = \beta_0 + \beta_1 i + \beta_2 i^2 + \beta_3 j + \beta_4 j^2 + \epsilon_{ij},$$

$$E(\epsilon_{ij}) = 0, \qquad V(\epsilon_{ij}) = p_{ij}(1 - p_{ij})/n_{ij}$$

may be more appropriate than the interaction model. Maximum likelihood estimates of the parameters may again be obtained, but unfortunately the work is long and tedious. An unweighted least squares solution, however, is easily derived as follows. Write

$$\mathbf{A} = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 \\ 1 & 2 & 4 & 1 & 1 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & r & r^2 & 1 & 1 \\ 1 & 1 & 1 & 2 & 4 \\ 1 & 2 & 4 & 2 & 4 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & r & r^2 & c & c^2 \end{bmatrix} \quad \hat{\mathbf{p}} = \begin{bmatrix} \hat{p}_{11} \\ \hat{p}_{21} \\ \vdots \\ \hat{p}_{r1} \\ \hat{p}_{12} \\ \hat{p}_{22} \\ \vdots \\ \hat{p}_{rc} \end{bmatrix} \quad \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_4 \end{bmatrix};$$

then it is required to minimise $(\hat{\mathbf{p}} - \mathbf{A}\boldsymbol{\beta})'(\hat{\mathbf{p}} - \mathbf{A}\boldsymbol{\beta})$ with respect to $\boldsymbol{\beta}$. Thus, if $\mathbf{S} = \mathbf{A}'\mathbf{A}$,

$$\hat{\boldsymbol{\beta}} = \mathbf{S}^{-1}\mathbf{A}'\hat{\mathbf{p}}, \quad E(\hat{\boldsymbol{\beta}}) = \mathbf{S}^{-1}\mathbf{A}'E(\hat{\mathbf{p}}) = \mathbf{S}^{-1}\mathbf{S}\boldsymbol{\beta} = \boldsymbol{\beta}$$

and the variance-covariance matrix for $\hat{\boldsymbol{\beta}}$ is

$$E\{(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})'\} = \mathbf{S}^{-1}\mathbf{A}'E\{(\hat{\mathbf{p}} - \mathbf{p})(\hat{\mathbf{p}} - \mathbf{p})'\}\mathbf{A}\mathbf{S}^{-1} = \mathbf{S}^{-1}\mathbf{A}'\mathbf{D}\mathbf{A}\mathbf{S}^{-1}$$

where

$$\mathbf{D} = \text{diag}\,\{p_{11}(1 - p_{11})/n_{11}\,,\, p_{21}(1 - p_{21})/n_{21}\,,\, \cdots\,,\, p_{rc}(1 - p_{rc})/n_{rc}\}\,.$$

Obviously this approach can be extended to cover more general situations.

EXAMPLE

In order to illustrate these methods with a numerical example, we consider the data of Dunlop and Tallis [1963]. In this study, breeding ewes were classified according to whether they remained in oestrus one or two days. Roughly, half of each group was then singly inseminated and the other half doubly inseminated, on consecutive days. The aim of the experiment was to estimate the effects of the number of inseminations and the number of days in oestrus on the lambing performance of the ewes. From the subsequent lambing records Table 1 was compiled. The appropriate parameter models are listed in Table 2. In these models, α, a, and s represent effects due to the number of inseminations on twin births, single births, and dry ewes

respectively; β, b, and t are equivalent effects attributable to the number of days that a ewe is observed in oestrus; while γ, c, and u may be be interpreted as interaction terms.

TABLE 1

LAMBING PERFORMANCE OF EWES IN RELATION TO OESTRUS LENGTH AND NUMBER OF INSEMINATIONS

No. of Inseminations (i)	Days in Oestrus (j) 1	2
1	*P_{11} = 12 Q_{11} = 189 R_{11} = 66 N_{11} = 267	P_{12} = 1 Q_{12} = 27 R_{12} = 10 N_{12} = 38
2	P_{21} = 18 Q_{21} = 185 R_{21} = 63 N_{21} = 266	P_{22} = 4 Q_{22} = 28 R_{22} = 9 N_{22} = 41

*P_{ij}, number of ewes bearing twins. Q_{ij}, number of ewes bearing singles. R_{ij}, number of ewes not lambing to inseminations at this oestrus.

TABLE 2

PARAMETER MODELS

Number of Inseminations	Days in Oestrus 1	2
1	$p_{11} = p^* + \alpha + \beta + \gamma$ $q_{11} = q + a + b + c$ $r_{11} = r + s + t + u$	$p_{12} = p + \alpha - \beta - \gamma$ $q_{12} = q + a - b - c$ $r_{12} = r + s - t - u$
2	$p_{21} = p - \alpha + \beta - \gamma$ $q_{21} = q - a + b - c$ $r_{21} = r - s + t - u$	$p_{22} = p - \alpha - \beta + \gamma$ $q_{22} = q - a - b + c$ $r_{22} = r - s - t + u$

*These parameters are subject to the restrictions that: $p + q + r = 1$; $\alpha + a + s = 0$; $\beta + b + t = 0$; $\gamma + c + u = 0$.

By the methods of this paper the M. L. estimates are found to be

$$\hat{p} = \sum_{ij} \hat{p}_{ij}/4 \qquad \hat{q} = \sum_{ij} \hat{q}_{ij}/4$$

$$\hat{\alpha} = (\hat{p}_{11} + \hat{p}_{12} - \hat{p}_{21} - \hat{p}_{22})/4 \qquad \hat{a} = (\hat{q}_{11} + \hat{q}_{12} - \hat{q}_{21} - \hat{q}_{22})/4$$

$$\hat{\beta} = (\hat{p}_{11} + \hat{p}_{21} - \hat{p}_{12} - \hat{p}_{22})/4 \qquad \hat{b} = (\hat{q}_{11} + \hat{q}_{21} - \hat{q}_{12} - \hat{q}_{22})/4$$

$$\hat{\gamma} = \hat{p}_{11} - \hat{p} - \hat{\alpha} - \hat{\beta} \qquad \hat{c} = \hat{q}_{11} - \hat{q} - \hat{a} - \hat{b}.$$

The variance of the four estimates involving the $\hat{p}_{ij}$ are the same, and in fact equal $(16)^{-1} \sum_{ij} p_{ij}(1 - p_{ij})N_{ij}^{-1}$. A similar remark applies to the estimates involving $\hat{q}_{ij}$, and in this case the variances equal $(16)^{-1} \sum_{ij} q_{ij}(1 - q_{ij})N_{ij}^{-1}$. The estimates and their standard errors are given in Table 3.

TABLE 3

ESTIMATES OF PARAMETERS AND STANDARD ERRORS

Birth type fractions	$\hat{p} = 0.0591 \pm 0.0142$,	$\hat{q} = 0.6992 \pm 0.0277$,	$\hat{r} = 0.2417 \pm 0.0258$
Insemination frequency effects	$\hat{\alpha} = -0.0235 \pm 0.0142$,	$\hat{a} = 0.0100 \pm 0.0277$,	$\hat{s} = 0.0135 \pm 0.0258$
Oestrus length effects	$\hat{\beta} = -0.0028 \pm 0.0142$,	$\hat{b} = 0.0025 \pm 0.0277$,	$\hat{t} = 0.0003 \pm 0.0258$
Interactions	$\hat{\gamma} = 0.0121 \pm 0.0142$,	$\hat{c} = -0.0038 \pm 0.0277$,	$\hat{u} = -0.0083 \pm 0.0258$

For a detailed discussion of the biological implications of these results, as well as for a fuller description of the experiment, the reader is referred to the original paper.

REFERENCES

Aitchison, J. and Silvey, S. D. [1960]. Maximum-likelihood estimation procedures and associated tests of significance. *J. Roy. Statist. Soc. B. 22*, 154–71.

Dunlop, A. A. and Tallis, G. M. [1964]. The effects of length of oestrus and number of inseminations on the fertility and twinning rate of the Merino ewe. *Aust. J. Agric. Res. 15*, 282–8.

A (d)[2]

TECHNOMETRICS VOL. 10, NO. 1 FEBRUARY 1968

The Use of Fractional Moments for Estimating the Parameters of a Mixed Exponential Distribution*

G. M. TALLIS[1] AND R. LIGHT

The Johns Hopkins University

In this paper the use of fractional moments for estimation purposes is discussed. These ideas are illustrated by means of the mixed exponential distribution.

The estimation of the three parameters of the above distribution by the method of moments and by maximum likelihood is investigated numerically in detail. As anticipated, the efficiency of the former method can be greatly increased by using approximately optimal combinations of moments. It is found that the moment method requires only a small amount of calculation when compared with the maximum likelihood method, although charts are presented to greatly ease the computational burden of the latter method.

I. INTRODUCTION

Of all the procedures of estimating parameters, the method of moments is perhaps the oldest. In many cases it leads to tractable operations where other methods become computationally complicated and it is mainly for this reason that moment estimators are used at all today.

The attitude taken in this paper is that, although this method as usually applied may be inefficient for a particular problem [see e.g. Fisher (1922)], by resorting to fractional moments (or fractional absolute moments) the efficiency may be appreciably increased. This idea is discussed in relation to the problem of estimating the three parameters of a mixture of two exponential distributions.

The latter problem has been investigated in detail by Rider (1961) using the first three power moments. We show here by extensive numerical work that, for some combinations of the parameter values, these three moments provide extremely inefficient estimates of the parameters. However, by a suitable choice of fractional moments, this position can be greatly improved.

Weiner (1962) has studied this problem from the point of view of obtaining maximum likelihood (m.l.) estimates, and it is clear from his results that the amount of computational work involved is enormous. As a by-product of some of our calculations we have been able to construct charts to reduce this labour.

Received September 1966; Revised May 1967.

* Paper No. BU-133 of the Biometrics Department, and Paper No. PB522 of the Plant Breeding Department, Cornell University.

[1] Now at C. S. I. R. O., Sydney, Australia.

Nevertheless, a high speed computer is still required to complete the work whereas moment estimates can be obtained on a desk calculator.

In the following section we set out some of the properties and procedures pertinent to estimation by the method of moments. A lemma is proved establishing a sufficient condition for moment estimators and m.l. estimators to be the same. From investigations in later sections, it appears that, in some cases, the method of moments must be carefully applied if large losses of information are to be avoided.

A search of the literature has not revealed any papers which have investigated the efficacy of using fractional moments to increase the efficiency of estimation. In fact, fractional moments appear to have been scarcely used for any purposes at all.

An interesting preliminary example to emphasize the points to be made subsequently concerns the estimation of the parameter γ of the frequency function $[\Gamma(\gamma)]^{-1}x^{\gamma-1}e^{-x}$, $\gamma > 0$. It is found that, by drawing a random sample, $x_1, x_2, \cdots, x_n$, and computing $S_\alpha = \sum_{i=1}^{n} x_i^\alpha/n$, $\alpha > 0$, a moment estimate of γ, $\hat{\gamma}$, is obtained from the equation

$$S_\alpha = \Gamma(\hat{\gamma} + \alpha)/\Gamma(\hat{\gamma}).$$

Moreover,

$$V(\hat{\gamma}) \simeq \frac{\Gamma(\gamma + 2\alpha)\Gamma(\gamma) - \Gamma^2(\gamma + \alpha)}{n\Gamma^2(\gamma + \alpha)[\psi(\gamma + \alpha) - \psi(\gamma)]^2}$$

where $\psi(x)$ is the digamma function. It can be shown that $\lim_{\alpha \to 0} V(\hat{\gamma}) = [n\psi'(\gamma)]^{-1}$ which is the variance for the maximum likelihood estimator of γ and this stresses the possibility of increasing the efficiency of the method of moments by considering fractional powers of X.

II. Methods

Consider the distribution function $F(x, \theta)$ with unknown vector of parameters $\theta' = (\theta_1, \theta_2, \cdots, \theta_k)$. Let the vector S be defined by $S' = (S_{\alpha_1}, S_{\alpha_2}, \cdots, S_{\alpha_k})$ where $S_{\alpha_i} = \sum_{j=1}^{n} |x_j|^{\alpha_i}/n$, $\{x_j\}$ is a random sample of size n taken from $F(x, \theta)$ and the α_i are positive real numbers.

We define the moment estimate of θ by the matrix equation

$$\mathbf{m}(\tilde{\theta}) = \mathbf{S} \tag{1}$$

where $\mathbf{m}(\theta) = (m_{\alpha_1}(\theta), \cdots, m_{\alpha_k}(\theta))$ and

$$m_{\alpha_i}(\theta) = \int_{-\infty}^{\infty} |x|^{\alpha_i}\, dF(x, \theta).$$

It will be assumed that the α_i are linearly ordered and that $E\{|x|^{2\alpha_k}\}$ exists, thus ensuring finite variances for the S_{α_i}. Moreover, the functions $m_{\alpha_i}(\theta)$ will be presumed to possess continuous partial derivations of the first order.

Now let θ_0 be a first guess at $\tilde{\theta}$, then neglecting second order terms, the first correction terms, δ, are obtained from

$$\Delta = \mathbf{S} - \mathbf{m}(\theta_0) = \mathbf{H}(\theta_0)\delta \tag{2}$$

where $H(\theta) = [h_{ij}(\theta)]$ and $h_{ij}(\theta) = \partial m_i(\theta)/\partial\theta_j$. Thus if $|H(\theta_0)| \neq 0$,

$$\delta = H(\tilde{\theta}_0)^{-1} \Delta. \tag{3}$$

In order to calculate the large sample variance matrix for $\tilde{\theta}$, we note that

$$dm(\tilde{\theta}) = H(\theta)\, d\tilde{\theta} = dS$$

and

$$H(\theta)V(\tilde{\theta})H(\theta)' = V(S)$$

whence

$$V(\tilde{\theta}) = H(\theta)^{-1}V(S)H(\theta)'^{-1} \tag{4}$$

where $V(S) = n^{-1}[m_{\alpha_i+\alpha_j}(\theta) - m_{\alpha_i}(\theta)m_{\alpha_j}(\theta)]$ is the covariance matrix for S. The efficiency of the estimator $\tilde{\theta}$ can be measured by the ratio $|V(\hat{\theta})|/|V(\tilde{\theta})|$, where $V(\hat{\theta})$ is the covariance matrix of the maximum likelihood, m.l., estimator of θ, $\hat{\theta}$ [Cramér (1946), page 494].

We now note some properties of the estimator $\tilde{\theta}$. Firstly, the equation $m(\theta) = m$ has a unique solution in some neighbourhood of m and θ if $|H(\theta)| \neq 0$. This follows from the classical inversion theorem of analysis. Since each component of S tends in probability to the corresponding component of $m(\theta)$, it is easily verified that S tends to $m(\theta)$ in probability i.e. $\lim_{n\to\infty} P(||S - m(\theta)|| < \epsilon) = 1$ for arbitrary $\epsilon > 0$. Thus, for sufficiently large n we are guaranteed a unique solution to (1) with probability arbitrarily near one, provided only that $|H(\theta)| \neq 0$.

It also follows easily from the above results and from the continuity of the inverse transformation that $\tilde{\theta}$ is consistent. Moreover, under rather general conditions on the functions $m_i(\theta)$. Cramér (1946 page 366) points out that $\tilde{\theta}$ is asymptotically normally distributed with mean vector θ and covariance $V(\tilde{\theta})$ given by (4).

Since the moment estimators obtained in this paper will be compared with maximum likelihood, m.l., estimators for efficiency, it is interesting to notice one result connecting the two types of estimation procedures. The following lemma is illustrative and the regularity conditions stated in the reference cited are assumed to hold here.

Lemma

If an unbiased estimator, $T(x)$, of some strictly monotone, differentiable function of θ, $\tau(\theta)$, satisfying the Cramér–Rao lower bound exists for θ in some interval, then there exists a moment estimator of θ, $\tilde{\theta}$, such that $\tilde{\theta} = \hat{\theta}$, where $\hat{\theta}$ is the m.l. estimator of θ.

Proof

Since $T(x)$ satisfies the Cramér–Rao lower bound and $E(T) = \tau(\theta)$, we know that [see Rao (1965, page 264)]

$$T = \lambda(\theta)f'/f + \tau(\theta)$$

and hence

$$f'/f = T/\lambda(\theta) - \tau(\theta)/\lambda(\theta), \qquad \lambda(\theta) \neq 0 \quad \text{for all} \quad \theta.$$

Now let $D^{-1}(1/\lambda(\theta)) = A(\theta)$ and $D^{-1}(\tau(\theta)/\lambda(\theta)) = -B(\theta)$, then

$$f(x, \theta) = \exp\{A(\theta)T(x) + B(\theta) + C(x)\}.$$

If a random sample of size n is taken from a population with frequency function $f(x, \theta)$ we have that

$$\log L(\theta) = A(\theta) \sum_{i=1}^{n} T(x_i) + nB(\theta) + D(x)$$

and

$$\frac{d \log L(\hat{\theta})}{d\theta} = \bar{T}/\lambda(\hat{\theta}) - \tau(\hat{\theta})/\lambda(\hat{\theta}) = 0, \qquad \bar{T} = \sum_{i=1}^{n} T(x_i)/n,$$

whence $\hat{\theta} = \tau^{-1}(\bar{T})$, which is precisely the moment estimator of θ using the function $T(x)$.

As an example, since in the case of a normal distribution of known variance X is an unbiased estimate of the mean, μ, attaining the minimum variance bound, it follows from the above result that $\bar{x}$ is both the moment and m.l. estimator of μ. In fact, since a necessary and sufficient condition that the Cramér–Rao lower bound be attained by an estimator $T(x)$ of some function $\tau(\theta)$, $E(T) = \tau$, is that $f(x, \theta)$ be of exponential form, there exists a moment estimator which is equal to the m.l. estimator for the parameters of most of the usual one parameter distributions. Undoubtedly stronger and more useful theorems can be established, but we leave the issue at this point.

III. An Application

In our case we consider the mixture of two exponential distributions and

$$dF(x, \theta) = f(x, \theta)\,dx = (\theta_3\theta_1 e^{-\theta_1 x} + \theta_4\theta_2 e^{-\theta_2 x})\,dx, \qquad \theta_4 = 1 - \theta_3$$

whence it is found that

$$m_{\alpha_i} = E\{X^{\alpha_i}\} = \Gamma(\alpha_i + 1)(\theta_3\theta_1^{-\alpha_i} + \theta_4\theta_2^{-\alpha_i}).$$

Now $\mathrm{H}(\theta) = [h_{ij}(\theta)]$, $h_{ij}(\theta) = \partial m_{\alpha_i}/\partial\theta_j$, and letting $\theta_2 = k\theta_1$ after differentiation,

$$\mathbf{H}(\theta) = \mathbf{A}(\theta)\mathrm{H}^*(\theta)\mathrm{B}(\theta)\mathbf{C}(\theta)$$

where $\mathbf{A}(\theta) = \mathrm{diag}\,(\theta_1^{-\alpha_1}, \theta_1^{-\alpha_2}, \theta_1^{-\alpha_3})$, $\mathrm{B}(\theta) = \mathrm{diag}\,(\theta_3, \theta_4, 1)$, $\mathbf{C}(\theta) = \mathrm{diag}\,(\theta_1^{-1}, \theta_1^{-1}, 1)$ and

$$\mathbf{H}^*(\theta) = \begin{bmatrix} -\Gamma(\alpha_1 + 1)\alpha_1 & -\Gamma(\alpha_1 + 1)\alpha_1 k^{-(\alpha_1+1)} & \Gamma(\alpha_1 + 1)(1 - k^{-\alpha_1}) \\ -\Gamma(\alpha_2 + 1)\alpha_2 & -\Gamma(\alpha_2 + 1)\alpha_2 k^{-(\alpha_2+1)} & \Gamma(\alpha_2 + 1)(1 - k^{-\alpha_2}) \\ -\Gamma(\alpha_3 + 1)\alpha_3 & -\Gamma(\alpha_3 + 1)\alpha_3 k^{-(\alpha_3+1)} & \Gamma(\alpha_3 + 1)(1 - k^{-\alpha_3}) \end{bmatrix}.$$

It can also be verified that

$$\mathrm{V}(\mathrm{S}) = \mathrm{A}(0)\mathrm{V}^*(\mathrm{S})\mathrm{A}(0) = n^{-1}[E\{x^{\alpha_i+\alpha_j}\} - E\{x^{\alpha_i}\}E\{x^{\alpha_j}\}]$$

where

$$n\mathrm{V}^*(\mathrm{S}) = [\Gamma(\alpha_i + \alpha_j + 1)(\theta_3 + \theta_4 k^{-(\alpha_i+\alpha_j)}) - \Gamma(\alpha_i + 1)\Gamma(\alpha_j + 1)(\theta_3 + \theta_4 k^{-\alpha_i})(\theta_3 + \theta_4 k^{-\alpha_j})].$$

Thus we have that

$$|\mathrm{V}(\tilde{\theta})| = |\mathrm{H}(0)|^{-2}\,|\mathrm{V}(\mathrm{S})| = \theta_1^4\theta_3^{-2}\theta_4^{-2}\,|\mathrm{H}^*(0)|^{-2}\,|\mathrm{V}^*(\mathrm{S})| \tag{6}$$

and we notice that $\mathrm{H}^*(0)$ and $\mathrm{V}^*(\mathrm{S})$ do not depend on θ_1.

Now the maximum likelihood (m.l.) function is given by

$$L(\theta) = \prod_{i=1}^{n} f(x_i\,,\,\theta) = \prod_{i=1}^{n} \{\theta_3\theta_1 e^{-\theta_1 x_i} + \theta_4\theta_2 e^{-\theta_2 x_i}\} \tag{7}$$

and in order to maximize (7) with respect to θ we calculate the likelihood equations

$$\left.\begin{aligned}
\frac{\partial \log L(\hat{\theta})}{\partial \theta_1} &= \sum_{i=1}^{n} [f(x_i\,,\,\hat{\theta})]^{-1}[\hat{\theta}_3(1 - \hat{\theta}_1 x_i)e^{-\hat{\theta}_1 x_i}] = 0\\
\frac{\partial \log L(\hat{\theta})}{\partial \theta_2} &= \sum_{i=1}^{n} [f(x_i\,,\,\hat{\theta})]^{-1}[\hat{\theta}_4(1 - \hat{\theta}_2 x_i)e^{-\hat{\theta}_2 x_i}] = 0\\
\frac{\partial \log L(\hat{\theta})}{\partial \theta_3} &= \sum_{i=1}^{n} [f(x_i\,,\,\hat{\theta})]^{-1}[\hat{\theta}_1 e^{-\hat{\theta}_1 x_i} - \hat{\theta}_2 e^{-\hat{\theta}_2 x_i}] = 0
\end{aligned}\right\} \tag{8}$$

The information matrix $\mathrm{I}(0)$ can be written in the form

$$\mathrm{I}(0) = n\mathrm{C}(0)\mathrm{B}(0)\mathrm{I}^*(0)\mathrm{B}(0)\mathrm{C}(0),$$

where the elements of $\mathrm{I}^*(0)$ do not depend on θ_1, and are defined by

$$\begin{aligned}
I_{11}^*(0) &= \int_0^\infty [g(x)]^{-1}e^{-2x}(1 - x)^2\,dx\\
I_{12}^*(0) &= \int_0^\infty [g(x)]^{-1}e^{-x(1+k)}(1 - x)(1 - kx)\,dx\\
I_{22}^*(0) &= \int_0^\infty [g(x)]^{-1}e^{-2kx}(1 - kx)^2\,dx\\
I_{13}^*(0) &= \int_0^\infty [g(x)]^{-1}e^{-2x}(1 - ke^{-(k-1)x})(1 - x)\,dx\\
I_{23}^*(0) &= \int_0^\infty [g(x)]^{-1}e^{-(k+1)x}(1 - ke^{-(k-1)x})(1 - kx)\,dx\\
I_{33}^*(0) &= \int_0^\infty [g(x)]^{-1}e^{-2x}(1 - ke^{-(k-1)x})^2\,dx
\end{aligned} \tag{9}$$

where $g(x) = \theta_3 e^{-x} + \theta_4 k e^{-kx}$.

We notice now that the efficiency of the method of moments, $|V(\tilde{0})|^{-1}\,|I(0)|^{-1}$, can be written as $|H^*(0)|^2\,|V^*(S)|^{-1}\,|I^*(0)|^{-1}$ and is therefore independent of θ_1. The efficiency is thus a function of θ_3 and k and we use the notation $E(k, \theta_3)$ to emphasize this.

The Tables

By use of the formulae developed above, tables 1 and 2 have been constructed. Firstly, for fixed k and θ_3, $E_1(k, \theta_3)$ has been calculated using $\alpha_1 = 1$, $\alpha_2 = 2$ and $\alpha_3 = 3$. The integrals $I_{ij}^*(0)$ were evaluated numerically using the fifteen point Laguerre integration technique. It can be seen from Table 1 that the efficiency falls off alarmingly with increasing k and θ_3.

A search was subsequently made to find more satisfactory combinations of the α_i using increments of .25 and keeping α_1 fixed at 1. The results of this work are given in Table 2 where the combinations giving the greatest efficiency are tabulated. A second efficiency figure, $E_2(k, \theta_3)$, was then calculated and the results listed in Table 1. It is clear from these figures that the use of carefully selected moment combination can considerably increase the efficiency

TABLE I

The Efficiencies $E_1(k, \theta_3)$ and $E_2(k, \theta_3)$ for various values of k and θ_3

		k					
θ_3		1.5	2	3	4	5	10
.1	$E_1(k, \theta_3)$	.942	.737	.407	.196	.155	.041
	$E_2(k, \theta_3)$	.950	.829	.711	.589	.561	.501
.2	$E_1(k, \theta_3)$	.912	.643	.311	.174	.110	.022
	$E_2(k, \theta_3)$	.946	.842	.726	.640	.621	.441
.3	$E_1(k, \theta_3)$	.868	.574	.257	.138	.083	.015
	$E_2(k, \theta_3)$	.946	.858	.727	.671	.624	.502
.4	$E_1(k, \theta_3)$	.824	.533	.220	.112	.065	.011
	$E_2(k, \theta_3)$	.948	.862	.734	.681	.614	.509
.5	$E_1(k, \theta_3)$	.784	.475	.190	.093	.052	.008
	$E_2(k, \theta_3)$	.950	.872	.756	.677	.625	.507
.6	$E_1(k, \theta_3)$	.747	.436	.165	.078	.042	.006
	$E_2(k, \theta_3)$	.950	.880	.768	.676	.642	.512
.7	$E_1(k, \theta_3)$	.714	.402	.146	.066	.035	.005
	$E_2(k, \theta_3)$	.954	.880	.772	.697	.655	.528
.8	$E_1(k, \theta_3)$	.682	.371	.127	.055	.028	.004
	$E_2(k, \theta_3)$	.957	.876	.771	.718	.664	.535
.9	$E_1(k, \theta_3)$	.652	.343	.111	.046	.022	.002
	$E_2(k, \theta_3)$	.956	.884	.778	.734	.665	.522

of the estimation procedure for large k and θ_3. If $k < 1.5$ and $\theta_3 < .3$ there is little difference between E_1 and E_2.

Because of the large amount of computation involved a full search for optimal α_i was not conducted. If all three moments were allowed to vary, perhaps even greater gains in efficiency could be achieved. Nevertheless, the tables can serve as a guide and should prevent drastic loss of information as a result of choosing inefficient moment combinations. The required sample moments are readily calculated with the aid of a table of square roots.

The Solution of the Equations

The first task is to obtain an initial guess of θ, θ_0 say, from the data. Once θ_0 is to hand, an approximately optimal set of moments, α_i, can be obtained from Table 2 and then the $S_{\alpha i} = \sum_{j=1}^{n} x_j^{\alpha_i}/n$ calculated. It is then possible to form the vector $\Delta_0 = \mathrm{S} - \mathrm{m}(\theta_0)$ and to find $\delta_0 = \mathrm{H}(\theta_0)^{-1}\Delta_0$. If the com-

TABLE 2

Suitable combinations of α_2 and α_3 ($\alpha_1 = 1$) for various values of k and θ_3

		k					
θ_3		1.5	2	3	4	5	10*
.1	α_2	2.25	2.00	1.50	1.50	.75	.75
	α_3	2.75	2.25	1.75	1.75	1.50	1.25
.2	α_2	2.00	1.75	1.25	.75	.75	.75
	α_3	2.50	2.00	1.50	1.50	1.25	1.25
.3	α_2	2.00	1.50	1.25	.75	.75	.50
	α_3	2.25	1.75	1.50	1.25	1.25	.75
.4	α_2	1.75	1.25	.75	.75	.50	.50
	α_3	2.00	1.75	1.50	1.25	1.25	.75
.5	α_2	1.75	1.25	.75	.75	.50	.50
	α_3	2.00	1.50	1.25	1.25	.75	.75
.6	α_2	1.50	1.25	.75	.50	.50	.25
	α_3	2.00	1.50	1.25	1.25	.75	.50
.7	α_2	1.50	1.25	.75	.50	.50	.25
	α_3	1.75	1.50	1.25	.75	.75	.50
.8	α_2	1.50	1.25	.50	.50	.50	.25
	α_3	1.75	1.50	1.25	.75	.75	.50
.9	α_2	1.50	.75	.50	.50	.25	.25
	α_3	1.75	1.50	1.25	.75	.75	.50

* For $k > 10$ and all θ_3, $\alpha_2 = .25$ and $\alpha_3 = .50$ are satisfactory.

ponents of θ_0 are written θ_{0i}, $i = 1, 2, 3$, and $k_0 = \theta_{02}/\theta_{01}$,

$$H(\theta_0) = A(\theta_0)H^*(\theta_0)B(\theta_0)C(\theta_0)$$

Thus the inversion of $H(\theta_0)$ reduces essentially to the inversion of the simple 3×3 matrix $H^*(\theta_0)$.

The new estimate $\theta_1 = \theta_0 + \delta_0$ can now be used for another iteration and it may be found that, provided k_0 is not greatly different to k_1, $H^*(\theta_0)$ can be used in place of $H^*(\theta_1)$ to save one matrix inversion. However, the last iteration should use an updated estimate of H^* to ensure that δ is sufficiently close to zero. Once $\bar{\theta}$ is calculated, $V^*(S)$ can be obtained and

$$\hat{V}(\bar{\theta}) = C(\bar{\theta})^{-1}B(\bar{\theta})^{-1}H^*(\bar{\theta})^{-1}V^*(S)H^*(\bar{\theta})'^{-1}B(\bar{\theta})^{-1}C(\bar{\theta})^{-1}.$$

If $\bar{\theta}$ changes appreciably from θ_0, it may be desirable to select another set of moments in order to maximize the efficiency. However, if the first guesses are of the correct order or magnitude, this will probably not be necessary.

The solution of the m.l. equations (S) can be carried out according to standard methods outlined in Weiner (1962). The work is long and tedious and the calculations require a high speed computer for their completion.

Alternatively, use can be made of charts 1 to 6 which are by-products of Tables 1 and 2. We notice that $V(\hat{\theta}) = I(\theta)^{-1}$ can be written as

$$V(\hat{\theta}) = n^{-1}C(\theta)^{-1}B(\theta)^{-1}V^*(\theta)B(\theta)^{-1}C(\theta)^{-1}$$

where $V^*(\theta) = I^*(\theta)^{-1}$, the elements of $I^*(\theta)$ being defined by (9). The elements of $V^*(\theta)$ are graphed in the six charts for various values of θ_3 and k.

In order to make use of this information to solve (S), an initial value, θ_0, is substituted into the left hand side of (S) to calculate Δ_0. Using θ_0 and the charts, $V(\theta_0)$ is calculated and then a new estimate of $\hat{\theta}$, θ_1 obtained from the formula $\theta_1 = \theta_0 + \delta_0$, where $\delta_0 = V(\theta_0)\ \Delta_0$. A few iterations should lead to good estimates of $\hat{\theta}$ and $V(\hat{\theta})$.

Although the use of charts greatly reduces the labour of calculating $\hat{\theta}$, the moment procedure is less work. Certainly, the latter method could also be greatly shortened by the construction of charts similar to those for the m.l. solution. This program has not been carried out.

Censoring

Suppose that the data are censored in such a way that we have full information on all $X \leq T$ and we know only that $X > T$ for other values of X. We define a new variable

$$Y = \begin{matrix} X, & X \leq T \\ T, & X > T \end{matrix}.$$

and

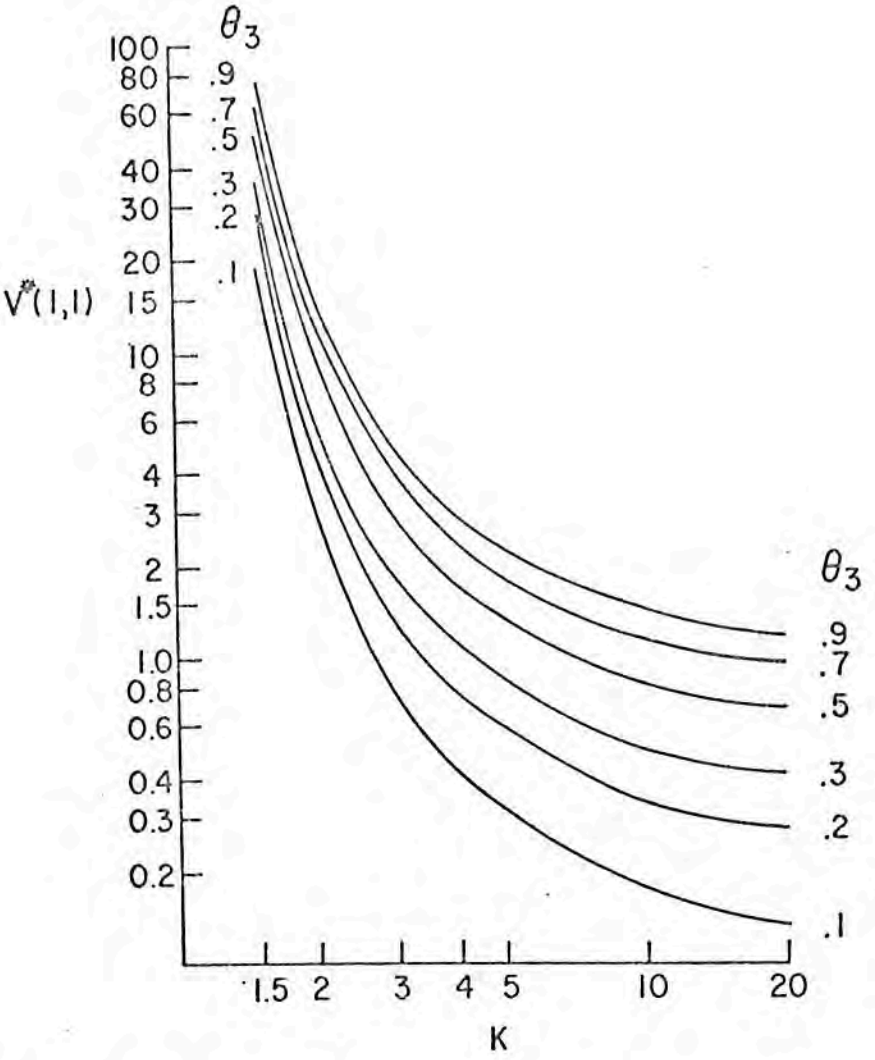
$$E\{Y^\alpha\} = \theta_3\theta_1^{-\alpha}I(\alpha + 1, \theta_1 T) + \theta_4\theta_2^{-\alpha}I(\alpha + 1, \theta_2 T) + T^\alpha[\theta_3 e^{-\theta_1 T} + \theta_4 e^{-\theta_2 T}]$$

where

$$I(\alpha + 1, \theta_1 T) = \int_0^{T\theta_1} t^\alpha e^{-t}\, dt.$$

CHART I

$V^*(1, 1)$ for various values of k and θ_3



It is found that in this case

$$\mathbf{H}(\theta)' = [\mathrm{H}_1'(\theta), \mathrm{H}_2'(\theta), \mathrm{H}_3'(\theta)]$$

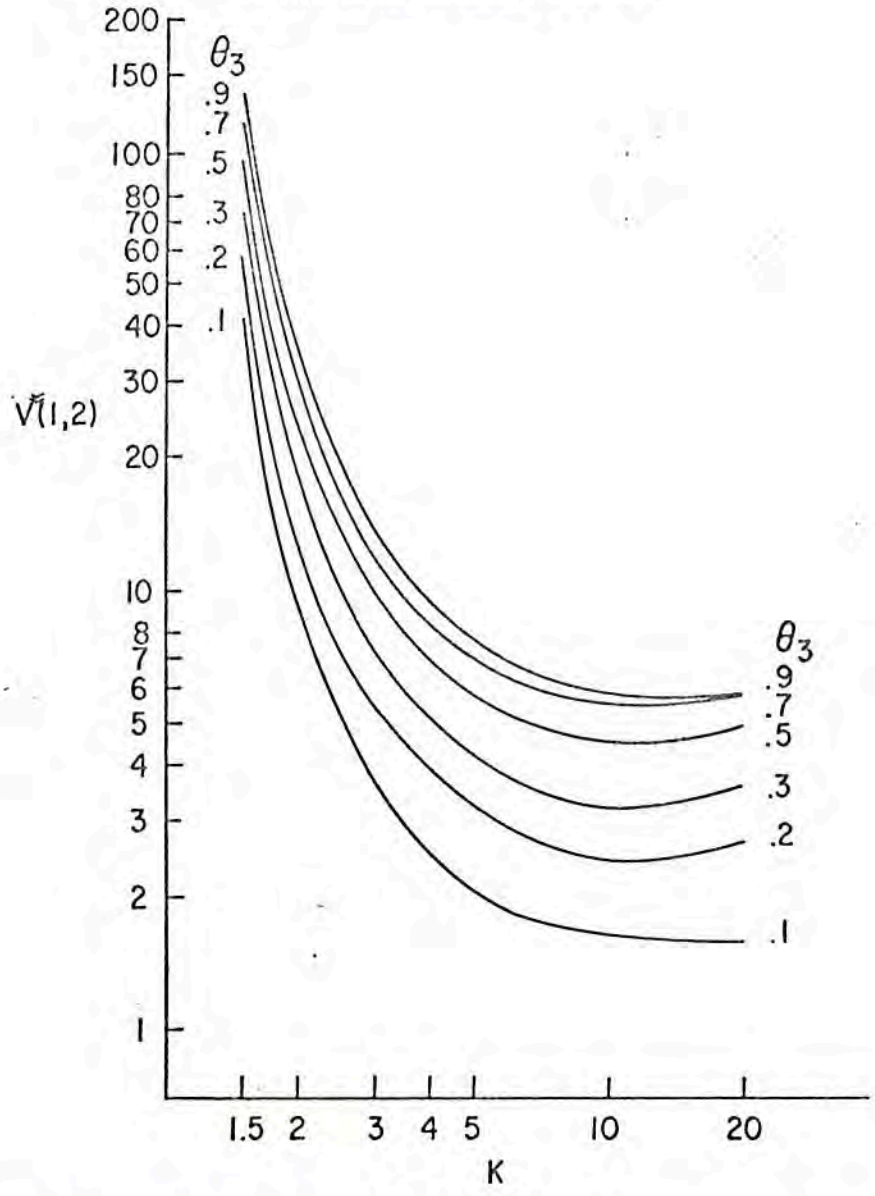
where

$$\mathbf{H}_i'(\theta) = [-I(\alpha_i + 1, \theta_1 T)\theta_3\alpha_i\theta_1^{-(\alpha_i+1)}, -I(\alpha_i + 1, \theta_2 T)\theta_4\alpha_i\theta_2^{-(\alpha_i+1)},$$
$$\theta_1^{-\alpha_i}I(\alpha_i + 1, \theta_1 T) - \theta_2^{-\alpha_i}I(\alpha_i + 1, \theta_2 T) + T^{\alpha_i}(e^{-\theta_1 T} - e^{-\theta_2 T})]$$

The task of computing $\tilde{\theta}$ and $\hat{V}(\tilde{\theta})$ can now be carried out as indicated above using

$$S_{\alpha_i} = n^{-1}\left\{\sum_{i=1}^{t} x_i^{\alpha_i} + (n - t)T^{\alpha_i}\right\},$$

CHART II
$V^*(1, 2)$ for various values of k and θ_3



where t is the number of x's in the sample less than T. The m.l. procedure must also be suitably modified and the charts are no longer of any use.

IV. Numerical Example

In order to compare the two methods of estimation a numerical example was tried. A sample of 196 observations was drawn from a population having a mixed exponential distribution with parameters $\theta_1 = .1$, $\theta_2 = 1$ and $\theta_3 = .5$.

This sample was constructed with the aid of a table of random, exponentially distributed numbers.

Using the parameter values for θ_0 , for the m.l. solution Δ_0' turned out to be [137.67, −3.54, −2.59]. The matrix $V^*(\theta_0)$ was then obtained from the charts and $V(\theta_0) = n^{-1}C(\theta_0)^{-1}B(\theta_0)^{-1}V^*(\theta_0)^{-1}B(\theta_0)^{-1}C(\theta_0)^{-1}$ calculated. It was

CHART III

$V^*(1, S)$ for various values of k and θ_3

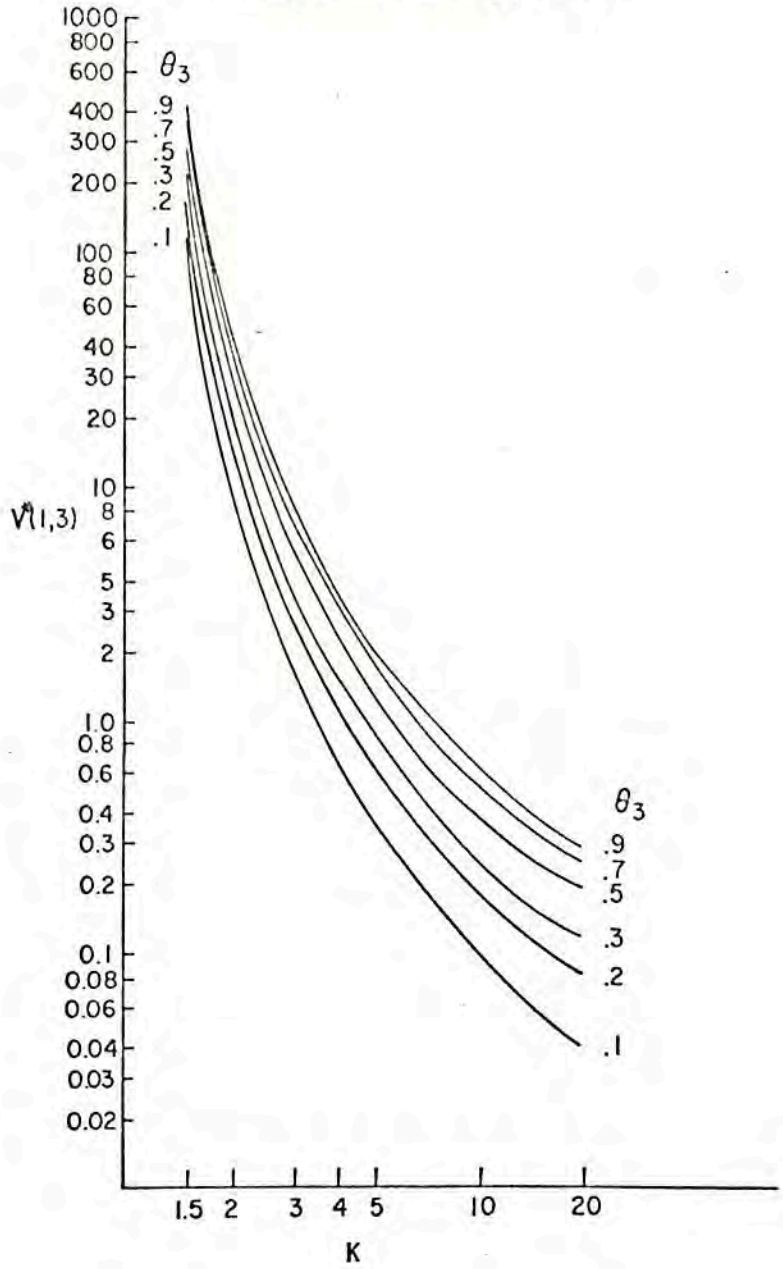
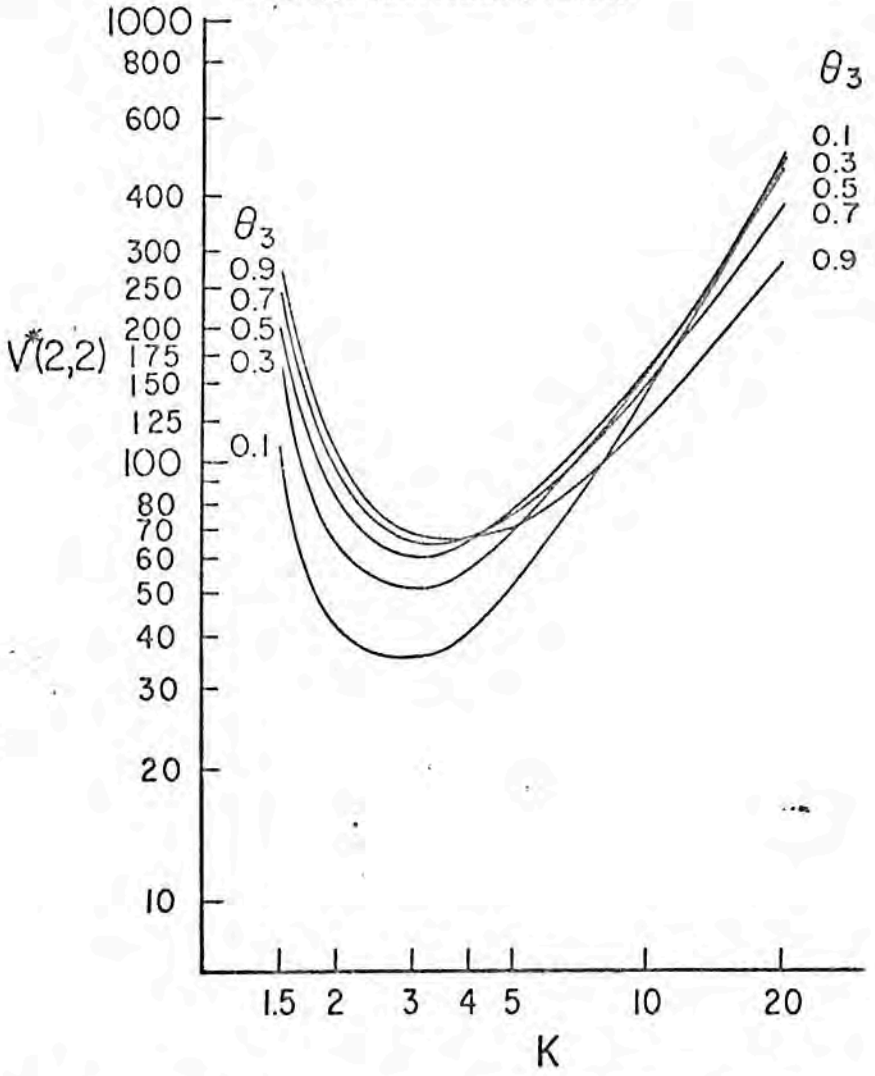


CHART IV

$V^*(2, 2)$ for various values of k and θ_3



found that $\delta_0' = [-0.26, -.264, -.082]$, and after six iterations $\hat{\theta}_1 = .077$, $\hat{\theta}_2 = .806$ and $\hat{\theta}_3 = .418$ with covariance matrix

$$\mathbf{V}(\hat{\theta}) = \begin{bmatrix} .00011 & .00052 & .00026 \\ .00052 & .01600 & .00381 \\ .00026 & .00381 & .00297 \end{bmatrix}.$$

In the case of the moment estimation procedure, using $\alpha_1 = .5$, $\alpha_2 = .75$ and $\alpha_3 = 1.00$, $S_{\alpha_1} = 1.9056$, $S_{\alpha_2} = 3.2648$ and $S_{\alpha_3} = 6.1908$ for the sample. Again, using the parameter values as the initial guess, $\Delta_0' = (-.06125, -.22116, -.69076)$ and $\delta_0' = \Delta_0' \mathrm{H}(\theta_0)^{-1/} = (.03039, .21281, .10346)$.

After four iterations $\Delta_5 = 0$ to five decimal places and $\bar{\theta}_1 = .078$, $\bar{\theta}_2 = .905$

and $\bar{\theta}_3 = .435$. The final covariance matrix was

$$V(\bar{\theta}) = \begin{bmatrix} .00014 & .00112 & .00049 \\ .00112 & .03503 & .00846 \\ .00049 & .00846 & .00471 \end{bmatrix}.$$

For the particular example chosen the moment estimates are all closer to the true values than are the m.l. estimates, although the overall efficiency of the former method is only about 50%. The rate of convergence of the moment iteration procedure seems to be considerably higher than that for the m.l. method.

CHART V

$V^*(2, S)$ for various values of k and θ_3

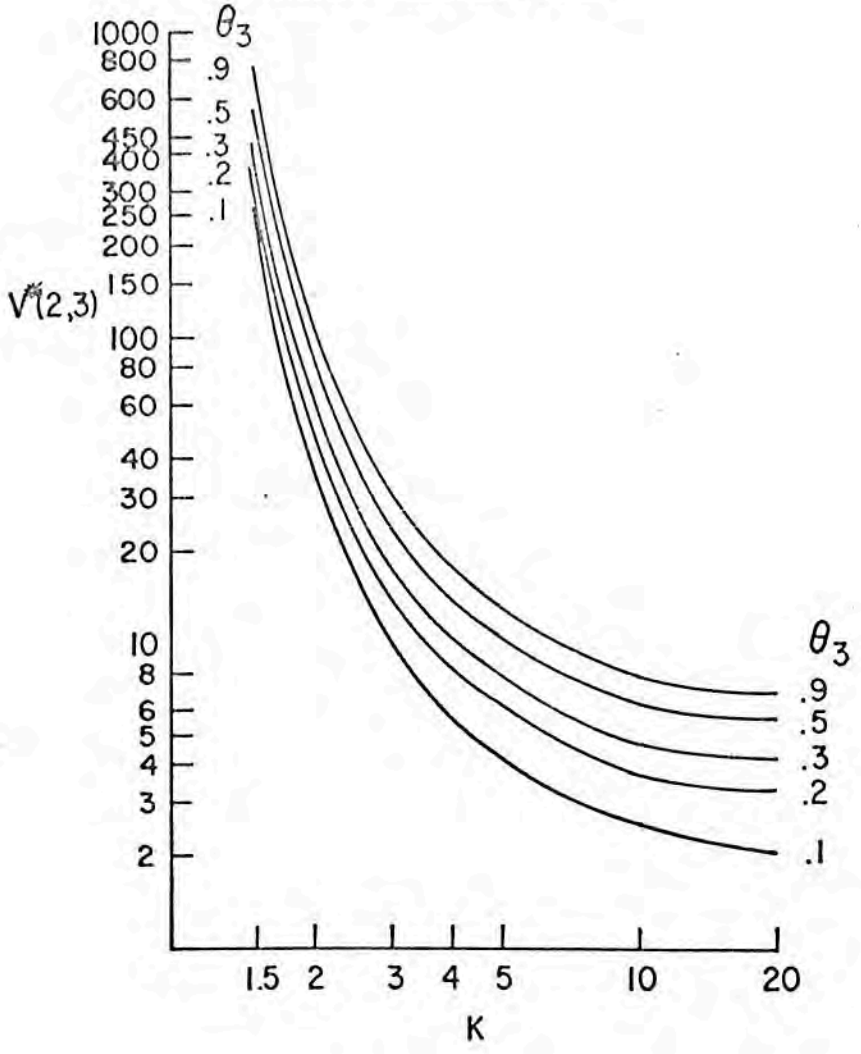
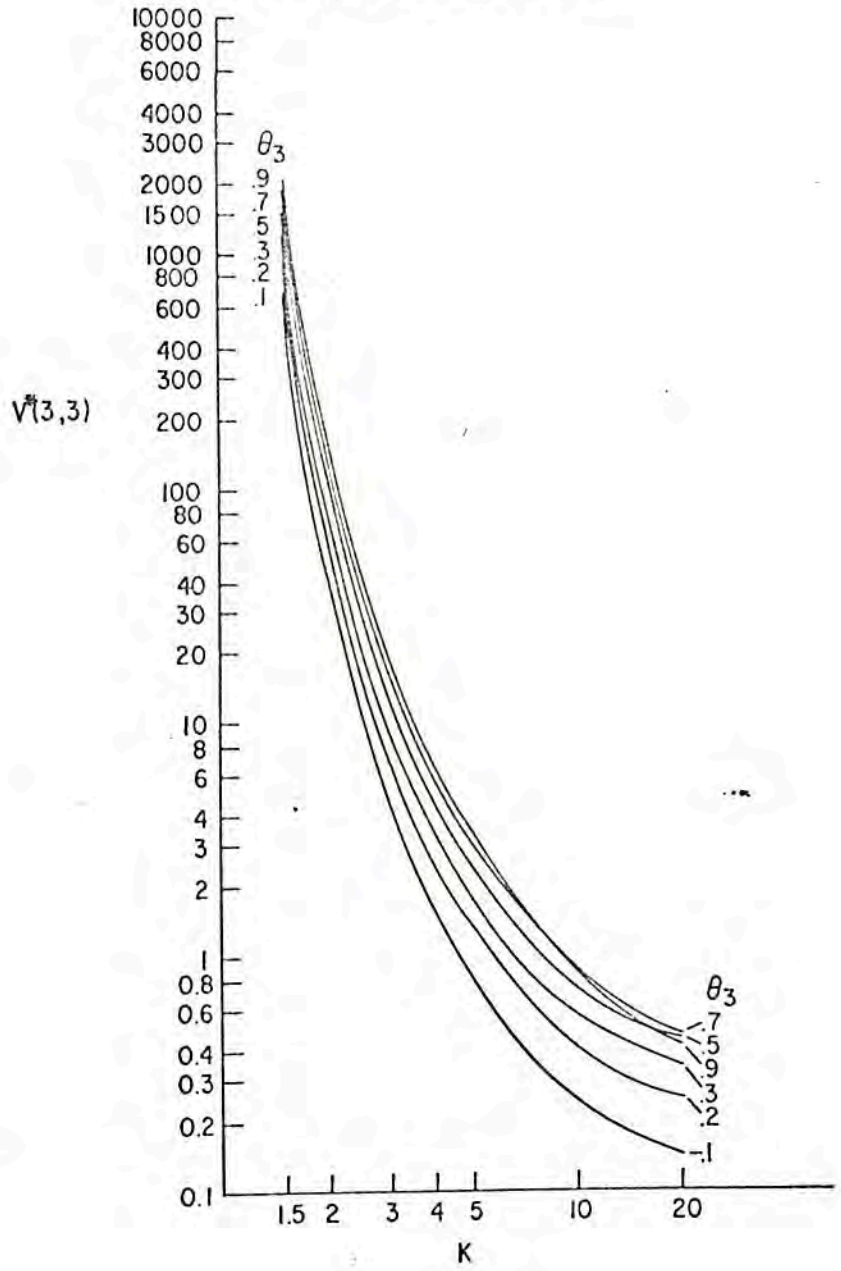


CHART VI

$V^*(3, 3)$ for various values of k and θ_3



ACKNOWLEDGMENT

The authors wish to thank the referee for his helpful comments.

References

1. Cramér, H., 1946. *Mathematical Methods of Statistics*. Princeton University Press.
2. Fisher, R. U., 1922. On the mathematical foundations of theoretical statistics. *Philos. Trans. Roy. Soc. A222*, 309–368.
3. Rao, C. R., 1965. *Linear Statistical Inference and its Applications*. John Wiley and Sons Inc. New York.
4. Rider, P., 1961. The method of moments applied to a mixture of two exponential distributions. *Ann. Math. Stat. 32*, 143–147.
5. Weiner, S., 1962. Samples from mixed-exponential populations. Mimeo. paper, ARINC Research Corp., Washington, D. C.

A (d)[3]

J. Appl. Prob. 6, 389–398 (1969)
Printed in Israel

THE IDENTIFIABILITY OF MIXTURES OF DISTRIBUTIONS

G. M. TALLIS, *C.S.I.R.O., New South Wales* 2042

1. Introduction

This paper considers aspects of the following problem. Let $F(x,\theta)$ be a distribution function, d.f., in x for all θ and a Borel measurable function of θ. Define the mixture (Robbins (1948)),

$$F(x) = \int_{-\infty}^{\infty} F(x,\theta)d\Phi(\theta) \tag{1}$$

where Φ is a d.f., then it is of interest to determine conditions under which $F(x)$ and $F(x,\theta)$ uniquely determine Φ. If there is only one Φ satisfying (1), F is said to be an identifiable mixture. Usually a consistency assumption is used whereby it is presumed that there exists at least one solution to (1).

The above definition of identifiability will be extended somewhat in Sections 2 and 3 below.

2. Countably infinite mixtures

The following form of (1) will now be discussed. Let

$$F(x) = \sum_{i=1}^{\infty} \beta_i F_i(x), \text{ a.e., } \sum_{1}^{\infty} |\beta_i| < \infty, \quad \sum_{1}^{\infty} \beta_i = 1 \tag{2}$$

where for all i $F_i(x)$ is a d.f., then F will be called a countably infinite mixture. At this stage no restriction of the form $\beta_i \geqq 0$ is invoked. The mixture (2) is identifiable if a unique set of β_i satisfies the equation. By assumption, there exists at least one such set.

It can be assumed at the outset that $F_i(-1) = 0$ and $F_i(1) = 1$ for all i since this condition can always be achieved, if necessary, by a suitable transformation without altering the problem.

A necessary and sufficient condition for identifiability in the case when the set $\{F_i\}_1^n$ is finite has been established by Teicher (1963). It turns out that this condition is equivalent to $\{F_i\}_1^n$ being a linearly independent set.

The infinite set $\{F_i\}_1^\infty$ will be said to be

Received in revised form 12 November 1968.

Definition 1: linearly independent if every finite subset is linearly independent;

Definition 2: strongly independent if $\sum_1^\infty a_i F_i(x) \equiv 0$ implies $a_i \equiv 0$ for $\sum_1^\infty |a_i| < \infty$;

Definition 3: mean square independent if, for $\sum_1^\infty |a_i| < \infty$,

$$\lim_{n\to\infty} \int_{-1}^{1} \left\{ \sum_1^n a_i F_i(x) \right\}^2 dx = 0$$

implies $a_i \equiv 0$.

Note that strong independence is equivalent to identifiability.

Theorem 1. A necessary condition that (2) is identifiable is that $\{F_i\}_1^\infty$ is linearly independent.

Proof. By a suitable renumbering if necessary, suppose that $F_k(x) = \sum_{i=1}^{k-1} a_i F_i(x)$ then

$$F(x) = \sum_{i=1}^{k-1} (\beta_i + a_i\beta_k) F_i(x) + \sum_{j=k+1}^{\infty} \beta_j F_j(x).$$

Set $\beta_i' = \beta_i + \varepsilon a_i$, $i = 1,2,\cdots,k-1$, $\beta_k' = \beta_k - \varepsilon$ and $\beta_i' = \beta_i$ for $i > k$ and notice that $\sum_1^{k-1} a_i = 1$. For arbitrary ε, $\sum_1^k \beta_i = \sum_1^k \beta_i'$ and it is clear that the set of solutions has cardinal number c.

The condition of linear independence is not sufficient for identifiability. For example, take any strongly independent set $\{F_i\}_1^\infty$ and form the new set $\{G_i\}_1^\infty$ where $G_{i+1} = F_i$ and $G_1 = \sum_1^\infty \beta_i F_i$, $\beta_i > 0$, $\sum_1^\infty \beta_i = 1$. The set $\{G_i\}_1^\infty$ is linearly independent but not strongly independent.

Some notation will now be introduced. By Theorem 1 it can be assumed that $\{F_i\}_1^\infty$ is linearly independent and hence by the Gram-Schmidt orthogonalisation process there exists an associated set of orthonormal functions $\{\phi_i\}_1^\infty$. Let $\int_{-1}^1 \phi_i(x) F_j(x) dx = k_{ij}$ and define the infinite matrix $K = [k_{ij}]$, then K is an upper semimatrix. Subsequently the arbitrary, complete set of orthonormal functions $\{\psi_i\}_1^\infty$ will be used in conjunction with the infinite matrix $A = [a_{ij}]$, $a_{ij} = \int_{-1}^1 \psi_i(x) F_j(x) dx$ and the relationship between A and K established.

Below, l^1 and l^2 refer to the sets of all infinite sequences of real numbers which are respectively, absolutely and square summable. The notation L^2 is an abbreviation of $L^2(-1,1)$, the set of all Lebesgue square integrable functions defined on the interval $[-1,1]$.

The domain of K will be restricted to l^1 and for any set $\{F_i\}_1^\infty$ the set $\gamma \subset L^2$ of interest is defined by

$$\gamma = \{f;\ f \in L^2,\ \lim_{n\to\infty} \int_{-1}^{1} \left[f(x) - \sum_{i=1}^{n} a_i F_i(x) \right]^2 dx = 0,\ a \in l^1\}.$$

It can be verified that $\{\phi_i\}_1^\infty$ also spans γ and, in fact, for suitable $b \in l^2$,

$$\lim_{n\to\infty} \int_{-1}^{1} \left[f(x) - \sum_{i=1}^{n} b_i \phi_i(x) \right]^2 dx = 0 \text{ for all } f \in \gamma.$$

Some of the equations which appear later may fail to hold on a Lebesgue set of measure zero. However, since this possibility does not affect the argument, repetitious use of the a.e. notation will not be made.

Theorem 2. Every solution to (2) is a solution to $K\beta = \alpha$ where $\alpha' = [\alpha_1, \alpha_2, \cdots]$, $\alpha_i = \int_{-1}^{1} \phi_i(x) F(x) dx$, and conversely. Moreover (2) is identifiable if and only if K^{-1} exists.

Proof. Firstly, every solution of (2) satisfies $K\beta = \alpha$ since

$$\int_{-1}^{1} \phi_i(x) F(x) dx = \alpha_i = \sum_{1}^{\infty} k_{ij} \beta_j$$

for all i, by the dominated convergence theorem. On the other hand, for every solution to $K\beta = \alpha$

$$\int_{-1}^{1} \phi_i(x) \left[F(x) - \sum_{1}^{\infty} \beta_j F_j(x) \right] dx = 0$$

for all i. Since the term in square brackets is a linear combination of the $F_j \in \gamma$, it is a linear combination of the ϕ_j with all Fourier coefficients zero. This implies that $F(x) = \sum_1^\infty \beta_j F_j(x)$ and the set $\{\beta_i\}_1^\infty$ is a solution to (2). The last part of the theorem is now obvious.

Now for any $\{w_i\}_1^\infty$, $\sum_1^\infty w_j F_j(x) = \sum_1^\infty y_j \phi_j(x)$ for suitable choice of $\{y_i\}_1^\infty$. In fact if both sides of the above equation are multiplied by ϕ_i and integrated, the equation $Kw = y$ results. Repeat the above process using F_i and define the matrix

$$D = [d_{ij}], \; d_{ij} = \int_{-1}^{1} F_i(x) F_j(x) dx,$$

then

$$Dw = K'y = K'(Kw) = (K'K)w.$$

Since the above equation holds for all w, $D = K'K$. Note that the associativity relation $K'(Kw) = (K'K)w$ can be justified since K is an upper semi-matrix.

Furthermore, $Kw = 0$ implies $K'(Kw) = (K'K)w = 0$. On the other hand, $0 = (K'K)w = K'(Kw)$ implies that $w'K'(Kw) = 0$ and $Kw = 0$. Thus K^{-1} exists if and only if D^{-1} exists.

Corollary 1. A necessary and sufficient condition that (2) is identifiable is that D^{-1} exists.

If $\{\psi_i\}_1^\infty$ is a complete set of orthonormal functions with respect to Lebesgue measure on $[-1,1]$ then, for suitable $\{z_i\}_1^\infty$,

$$\sum_1^\infty w_i F_i(x) = \sum_1^\infty y_i \phi_i(x) = \sum_1^\infty z_i \psi_i(x).$$

From the above equation the following relationships can be deduced

$$Kw = y = Bz$$

$$Aw = B'y = z$$

where $B = [b_{ij}]$, $b_{ij} = \int_{-1}^{1} \phi_i(x)\psi_j(x)dx$. Thus for all $\{w_i\}_1^\infty$ $Kw = B(Aw)$, $Aw = B'(Kw)$ and therefore $K = BA$ and $A = B'K$. Similarly for all $\{y_i\}_1^\infty$ $y = B'(By)$ which implies that $B'B = I$, i.e., B is orthogonal. Finally $Kw = y = 0$ implies $B'y = Aw = 0$ and $Aw = z = 0$ implies $Bz = Kw = 0$ and therefore K^{-1} exists if and only if A^{-1} exists.

Corollary 2. If $\{\psi_i\}_1^\infty$ is a complete set of orthonormal functions on $[-1,1]$ then, in the notation used above, $A = B'K$, $K = BA$, $B'B = I$ and a necessary and sufficient condition for (2) to be identifiable is that A^{-1} exists.

In summary, (2) is identifiable if and only if K^{-1} exists, if and only if D^{-1} exists, if and only if A^{-1} exists.

Theorem 3. The set $\{F_i\}_1^\infty$ is strongly independent if and only if it is mean square independent.

Proof. Let $\{F_i\}_1^\infty$ be mean square independent and suppose that $\sum_1^\infty a_i F_i(x) = 0$ then

$$\lim_{n\to\infty} \int_{-1}^{1} \left\{ \sum_1^n a_i F_i(x) \right\}^2 dx = \int_{-1}^{1} \left\{ \sum_1^\infty a_i F_i(x) \right\}^2 dx = 0$$

and $a_i \equiv 0$. The interchange of the order of summation and integration can be justified by the dominated convergence theorem.

If, on the other hand, $\{F_i\}_1^\infty$ is strongly independent, the same calculation shows that $\sum_1^\infty a_i F_i(x) = 0$ a.e. implies $a_i \equiv 0$.

Although the above results have been presented for d.f.s the same arguments may hold if $F_i(x)$ is replaced by its derivative, if it exists, or by an associated transform such as the characteristic function. Ingenuity is required to select the optimum form and the best method of attacking specific mixtures.

The conditions of (2) can be changed somewhat without altering the results. If it is assumed that there is one set $\{\beta_i\}_1^\infty$ which satisfies (2) and which has the properties $\beta_i > 0$ for all i and $\sum_1^\infty \beta_i = 1$, and if $F(x)$ is called identifiable when this is the only set with these properties, then the above theorems still apply. In fact if $\{\beta_i^1\}_1^\infty$ is another solution to (2) then $\{t\beta_i + (1-t)\beta_i^1\}_1^\infty$, is also a solu-

tion for $0 \leqq t \leqq 1$. Thus any neighbourhood of β contains other solutions with the same properties as $\{\beta_i\}_1^\infty$.

Some sufficient conditions for K^{-1} to exist

Rewrite $Kx = y$ in the form $x = (I - K)x + y$ where $I = [\delta_{ij}]$; then, setting $C = I - K$, if $\sum_{i,j} |c_{ij}| < \infty$, the determinant of C, Δ, exists. If $\Delta \neq 0$ then there is a unique solution to $Kx = y$ (Kantorovich and Krylov (1958)).

If $\sum_{j=1}^\infty |c_{ij}| < 1$, the system is called regular. The result which is most appropriate to the present problem concerning regular systems is the following theorem.

Theorem 4. A regular system can have no more than one solution tending to zero, i.e., such that $\lim_{i\to\infty} \beta_i = 0$.

A simple proof of this theorem, which will be used later, can be found in Kantorovich and Krylov.

A third condition on C will now be established. Define the norm of C as $\|C\|^2 = \sum_{i,j} c_{ij}^2$ and suppose $\|C\|^2 < \infty$, then $\|Cx\| \leqq \|C\| \, \|x\|$ for $x \in l^2$. To see this let $y_i(n) = \sum_{j=1}^n c_{ij}x_j$ then, by the Cauchy-Schwartz inequality

$$|y_i(n)| \leqq \sum_{j=1}^n |c_{ij}x_j| \leqq \left(\sum_{j=1}^n c_{ij}^2\right)^{1/2} \left(\sum_{j=1}^n x_j^2\right)^{1/2},$$

and hence

$$\sum_{i=1}^m y_i^2(n) \leqq \sum_{i=1}^m \left(\sum_{j=1}^n c_{ij}^2\right) \left(\sum_{j=1}^n x_j^2\right).$$

Upon letting n and then m tend to infinity the required result is obtained.

Theorem 5. A sufficient condition for $Kx = y$ to have a unique solution is that $\|I - K\| < 1$.

Proof. Let $C = I - K$ then, since $\|Cx_1 - Cx_2\| = \|C(x_1 - x_2)\| < \|x_1 - x_2\|$ for $x_1, x_2 \in l^2$, C is a contraction operator in l^2 and has a unique fixed point x_0, say. Thus $Cx_0 = x_0$, $Kx_0 = 0$ and $x_0 = 0$ is the only solution to the homogeneous equation $Kx = 0$. Suppose $Kx_1 = Kx_2 = y$, then $K(x_1 - x_2) = 0$ and $x_1 = x_2$. The same arguments apply equally to the matrices D and A.

Three examples will now be discussed. In these cases it is convenient to work with the frequency functions.

Example 1. Consider the class of frequency functions $\{f_i(x)\}_1^\infty$ defined by

$$f_i(x) = \begin{cases} 2^i & 1 - 2^{-(i-1)} \leqq x \leqq 1 - 2^{-i}, \\ 0 & \text{otherwise.} \end{cases}$$

Let $d_{ij} = \int_{-\infty}^{\infty} f_i(x) f_j(x) dx$, with a slight abuse of notation, then the matrix D is obviously diagonal. Thus $Dx = 0$ has the unique solution $x = 0$ and hence D^{-1} exists. The mixture $f(x) = \sum_1^\infty \beta_j f_j(x)$ is identifiable.

Example 2. In this case define

$$f_i(x) = \begin{cases} 1 & \dfrac{i-1}{2} \leqq x \leqq \dfrac{i+1}{2}, \\ 0 & \text{otherwise.} \end{cases}$$

It is clear that D is now a band matrix with elements $\frac{1}{2}$, 1, $\frac{1}{2}$ in each row. The equation $Dx = 0$ leads to the difference equation

$$\tfrac{1}{2}x_{i-1} + x_i + \tfrac{1}{2}x_{i+1} = 0, \qquad i = 1, 2, \cdots$$

defining $x_0 = 0$. If $g(s) = \sum_1^\infty x_i s^i$ and $\sum_1^\infty |x_i| < \infty$, then $g(0) = 0$ and

$$g(s)\left(1 + \frac{s}{2} + \frac{1}{2s}\right) = \tfrac{1}{2}x_1, \qquad 0 < s \leqq 1.$$

If $x_1 \neq 0$, the above equation contradicts the absolute convergence of the series $\{x_i\}_1^\infty$ and hence $g(s) \equiv 0$ for $0 \leqq s \leqq 1$ implying $x_i \equiv 0$. Again D^{-1} exists and any mixture involving the f_i is identifiable.

Before discussing the third example the following useful fact is noted. Let Λ be an infinite diagonal matrix with elements $\lambda_i \neq 0$, then if $(\Lambda D \Lambda)^{-1}$ exists so does D^{-1}. For suppose D^{-1} does not exist then for some $x \neq 0$, $Dx = 0 = (\Lambda D \Lambda)\Lambda^{-1}x$ and $(\Lambda D \Lambda)$ has no inverse since $\Lambda^{-1}x \neq 0$. Thus to test the strong independence of $\{F_i\}_1^\infty$ it may be more convenient to test $\{\lambda_i F_i\}_1^\infty$ for suitable λ_i.

Example 3. Set

$$f_i(x) = \begin{cases} \theta^i & 0 \leqq x \leqq \theta^{-i} \\ 0 & \text{otherwise.} \end{cases} \qquad \theta^{1/2} = \gamma > 3$$

It turns out that, in this example, D is not an easy form to test the required inverse property. Consider, therefore, the related sequence $\{g_i\}_1^\infty$, $g_i = \theta^{-i/2} f_i$. For $\{g_i\}_1^\infty$

$$D = \begin{bmatrix} 1 & \gamma^{-1} & \gamma^{-2} & \cdots \\ \gamma^{-1} & 1 & \gamma^{-1} & \cdots \\ \gamma^{-2} & \gamma^{-1} & 1 & \cdots \\ \vdots & \vdots & \vdots & \end{bmatrix},$$

and, since $\sum_{\substack{j=1 \\ j \neq i}}^\infty |d_{ij}| < 2 \sum_{j=1}^\infty \gamma^{-} < 1$ an application of Theorem 4 shows that D^{-1} exists implying that $\{g_i\}_1^\infty$ and also $\{f_i\}_1^\infty$ are strongly independent.

3. The general case

Attention will now be given to the general mixture (1). As was done in Section (2) a somewhat broader attitude will be taken and Φ will not necessarily be a d.f. It will be assumed that, by suitable transformation, (1) can be written as

$$F(x) = \int_{-1}^{1} F(x,\theta)d\Phi(\theta), \tag{3}$$

with $F(-1) = 0$ and $F(1)=1$. In order to make further progress it seems necessary to impose conditions on $F(x,\theta)$. For convenience it will be assumed that $T(x,\theta) = \partial F(x,\theta)/\partial\theta$ is continuous in θ, and square integrable over $[-1,1] \times [-1,1]$.

Integrating (3) by parts gives

$$F(x) = F(x,1) - \int_{-1}^{1} T(x,\theta)\,\Phi(\theta)d\theta,$$

which can be put in the form

$$L(x) = \int_{-1}^{1} T(x,\theta)\Phi(\theta)d\theta, \tag{4}$$

where $L(x) = F(x,1) - F(x)$. It is easiest to discuss identifiability in terms of (4). Thus, (3) will be said to be identifiable if there is a unique square integrable solution to (4).

Define the symmetric kernel

$$K(x,y) = \int_{-1}^{1} T(x,z)T(y,z)dz$$

with eigenvalues and eigenfunctions λ_i and ϕ_i respectively, i.e., $\int_{-1}^{1}\phi_i(x)\,K(x,y)dx = \lambda_i\phi_i(x)$. An application of the Hilbert-Schmidt theorem for unsymmetric kernels, Tricomi ((1957), page 150) gives the following theorem.

Theorem 6. A necessary and sufficient condition that (3) is an identifiable mixture is that $\{\phi_i\}_1^\infty$ is a complete set, i.e., zero is not in the spectrum of K.

In fact, more can be said about (3) than is indicated by the last theorem. If one removes the requirement of consistency mentioned in the introduction, then (4) does not necessarily possess a solution. Let $\pi_i = \int_{-1}^{1}\phi_i(x)L(x)dx$, $i = 1,2,\cdots$, then, according to the standard theory of Fredholm integral equations of the first kind,

(a) if $\sum_{i=1}^{n} \lambda_i^{-2}\pi_i^2$ does not converge, there does not exist a solution to (4) in L^2;

(b) if $\sum_{i=1}^{\infty} \lambda_i^{-2}\pi_i^2 < \infty$ and $\{\phi_i\}_1^{\infty}$ is not complete in L^2, the set of solutions to (4) has cardinal number c;

(c) if $\sum_{i=1}^{\infty} \lambda_i^{-2}\pi_i^2 < \infty$ and $\{\phi_i\}_1^{\infty}$ is complete in L^2 then (4) (and (3)) has a unique solution.

Clearly (c) is the interesting case and Φ can be estimated by

$$\Phi_n(\theta) = \sum_{i=1}^{n} \pi_i \lambda_i^{-1} \phi_i(\theta).$$

The rapidity of the convergence of $\sum_{i=1}^{n} \pi_i \lambda_i^{-1}$ determines the accuracy of the mean-square approximation.

The above expression can be used to estimate the moments $\mu_j = \int_{-1}^{1} \theta^j d\Phi(\theta)$ which uniquely determine Φ. Define

$$\mu_j(n) = \int_{-1}^{1} [\tfrac{1}{2} - j\theta^{j-1}]\Phi_n(\theta)d\theta,$$

then, by the Cauchy-Schwartz inequality

$$\begin{aligned} |\mu_j - \mu_j(n)| &\leqq \int_{-1}^{1} |\tfrac{1}{2} - j\theta^{j-1}| \, |\Phi_n(\theta) - \Phi(\theta)| \, d\theta \\ &\leqq \left(\int_{-1}^{1} (\tfrac{1}{2} - j\theta^{j-1})^2 d\theta \right)^{1/2} \left(\int_{-1}^{1} [\Phi_n(\theta) - \Phi(\theta)]^2 d\theta \right)^{1/2} \\ &\leqq (2)^{1/2}(\tfrac{1}{4} - j^2/(2j-1))^{1/2}\varepsilon, \end{aligned}$$

where n is chosen so that

$$\sum_{i=n+1}^{\infty} \lambda_i^{-2}\pi_i^2 < \varepsilon^2.$$

A parallel theory can be developed for the countably infinite case. If D is assumed to satisfy the requirement that $\sum_{i,j} d_{ij}^2 < \infty$ then, defining ϕ_i and λ_i as the eigenvectors and eigenvalues of D, i.e., $D\phi_i = \lambda_i\phi_i$, and setting $L_i = \int_{-1}^{1} F_i(x)F(x)dx$ and $\pi_i = \phi_i' L$, similar statements to (a), (b) and (c) above hold.
Explicitly

(a′) if $\sum_{i=1}^{n} \lambda_i^{-2}\pi_i^2$ does not converge, there does not exist a solution to $\boldsymbol{L} = D\boldsymbol{\beta}$ in l^2;

(b′) if $\sum_{i=1}^{\infty} \lambda_i^{-2}\pi_i^2 < \infty$ and $\{\phi_i\}_1^{\infty}$ is not complete in l^2, the set of solutions to $L = D\beta$ has cardinal number c;

(c′) if $\sum_{i=1}^{\infty} \lambda_i^{-2}\pi_i^2 < \infty$ and $\{\phi_i\}_1^{\infty}$ is complete, there is a unique l^2 solution to the equation.

In the case of (c′), $\beta = \sum_{i=1}^{\infty} \pi_i\lambda_i^{-1}\phi_i$. If $\beta_n = \sum_{i=1}^{n} \pi_i\lambda_i^{-1}\phi_i$ is taken as an approximation to β, then $\| \beta_n - \beta \| = \sum_{n+1}^{\infty} \pi_i^2\lambda_i^{-2}$ which can be made arbitrarily small by taking n large.

The proof of these results flows from the fact that D is a completely continuous operator.

In general, the application of Theorem 6 is difficult. However, it does allow for the satisfactory and immediate discussion of some cases. Consider the probability generating function $[1 + p(s-1)]^n$ for the binomial distribution with parameters p and n. Let

$$P(s) = \int_0^1 [1 + p(s-1)]^n d\Phi(p),$$

then $P(\cdot)$ does not specify an identifiable mixture. To see this apply Theorem 6 and note that

$$n^2(x-1)(y-1)\int_0^1 [1 + p(x-1)]^{n-1}[1 + p(y-1)]^{n-1}dp$$

is not a closed kernel since non-trivial, $L^2(0,1)$ functions orthogonal to it are readily found.

It is well known that any function belonging to $L^2(-1,1)$ can be approximated in the mean-square by a polynomial. Explicitly, if $f \in L^2(-1,1)$, then

$$\lim_{n\to\infty} \int_{-1}^{1} \left[f(x) - \sum_{i=0}^{n} \alpha_i x^i \right]^2 dx = 0$$

for suitable choice of α_i and

$$\int_{-1}^{1} K(x,y)f(y)dy = \sum_{i=0}^{\infty} \alpha_i g_i(x),$$

where $g_i(x) = \int_{-1}^{1} y^i K(x,y)dy$.

Now if the system $\{g_i\}_0^{\infty}$ is strongly independent then, if $\sum_{i=0}^{\infty} \alpha_i g_i(x) \equiv 0$, $\alpha_i \equiv 0$, $f(x) \equiv 0$ and zero cannot be an eigenvalue of the kernel K. Thus the techniques of the previous sections may be useful in the application of Theorem 6.

Acknowledgement

The author is grateful to a number of readers of an earlier version of this paper for suggestions and comments. However, he particularly appreciates the help and encouragement of Dr. W. Davis in the later stages of the work.

References

KANTOROVICH, L. V. AND KRYLOV, V. I. (1959) *Approximate Methods of Higher Analysis*. Translated by C. D. Benster. Noordhoff, Groningen.

ROBBINS, H. (1948) Mixture of distributions. *Ann. Math. Statist.* **19**, 360–369.

TEICHER, H. (1963) Identifiability of finite mixtures. *Ann. Math. Statist.* 34, 1265–1269.

TRICOMI, F. G. (1957) *Integral Equations*. Interscience Publishers Inc., New York.

NOTE ON IDENTIFIABILITY OF MIXTURES OF DISTRIBUTIONS

G.M. TALLIS

(Unpublished)

I Introduction

The problem of identifiability of mixtures concerns the transformation

$$F(x) = \int_{\Omega} F(x,\theta)\,dG(\theta) \qquad (1)$$

where $F(x,\theta)$ is a distribution function, d.f., for all $\theta\in\Omega$ and G is a d.f. defined on Ω. Standard measurability conditions imposed on $F(x,\theta)$ ensure that the integral makes sense. It is easy to see that F is a d.f. and it is called a mixture. The family $F(x,\theta)$, $\theta\in\Omega$, is referred to as the kernel of the mixture and G is the mixing d.f. The mixture F is said to be identifiable iff there is a unique G leading to F.

The main task is to find conditions to impose on the kernel which will guarantee identifiability. This note proposes a necessary and sufficient condition which is developed from two basic properties of a metric. The method, which requires a minimum of mathematical structure is applied to retrieve some known results as special cases.

II Results

Let M be a set with a function ρ defined on $M\times M$ such that

(1) $\rho : M\times M \to E_1$

(2) $\rho(x,x) = 0 \quad \forall x\in M$

(3) $\rho(x,y) > 0 \quad \forall x\neq y\in M.$

Now consider a non-constant function $f : \mathcal{D}\to M$ with range $f(\mathcal{D}) = R\subset M$.

Definition

The norm of f, $\|f\|$, is defined by

$$\|f\| = \inf_{\alpha} \{\rho(f(x),f(y)) \leq \alpha\rho(x,y) \ \forall x,y\in\mathcal{D}\}$$

Note that since f is non-constant, $\|f\| > 0$. Moreover, if f^{-1} exists

$$\|f^{-1}\| = \inf_{\alpha} \{\rho(f^{-1}(u),f^{-1}(v)) \leq \alpha\rho(u,v) \ \forall u,v\in\mathcal{R}\}.$$

Theorem

f^{-1} exists and $\|f^{-1}\| < \infty$ iff $\rho(f(x),f(y)) \geq \alpha\rho(x,y)$ $\forall x,y\in\mathcal{D}$ for some $\alpha>0$.

Proof

Sufficiency

If $f(x) = f(y)$ then

$$0 = \rho(f(x), f(y)) \geq \alpha\rho(x,y)$$

which implies that $\rho(x,y) = 0$, that $x=y$ and that f^{-1} exists. Moreover,

$$\rho(f(x),f(y)) \geq \alpha\rho(f^{-1}(f(x),f^{-1}(f(y)))$$

or

$$\rho(f^{-1}(u),f^{-1}(v)) \leq \alpha^{-1}\rho(u,v) \ \forall u,v\in\mathcal{R}$$

and hence $\|f^{-1}\| \leq \alpha^{-1} < \infty$.

Necessity

If f^{-1} exists and $\|f^{-1}\| < \infty$, then

$$\rho(x,y) = \rho(f^{-1}(f(x)),f^{-1}(f(y))) \leq \|f^{-1}\| \ \rho(f(x),f(y))$$

and

$$\rho(f(x),f(y)) \geq \alpha\rho(x,y)$$

for $\alpha = \|f^{-1}\|^{-1}$ and $\forall x,y\in\mathcal{D}$.

To retrieve a standard result needed later, let M be a linear space with inner product $p(x,y)$. Then, if ρ is the metric induced by the norm $[p(x,x)]^{\frac{1}{2}} = \|x\|$ i.e. $\rho(x,y) = \|x-y\|$, and if T is a linear transformation of M onto $\mathcal{R}=M$, the above theorem shows that T^{-1} exists and $\|T^{-1}\| < \infty$ iff

$\|Tx\| \geq \alpha\|x\|$ $\forall x \in M$, $\alpha > 0$, (see e.g. Taylor (1958), page 86).

III Application

The above theorem is now applied to the mixture problem. Let M be the set of all d.f.'s, and without loss of generality assume that for $F \in M$, $F(0) = 0$, $F(1) = 1$ and $\Omega = [0,1]$. Put

$$\rho(F,G) = \int_0^1 [F(x) - G(x)]^2 \, dx,$$

then ρ satisfies the conditions of the theorem since $\rho(F,G) = 0$ implies $F = G$.

The function f that maps $\mathcal{D} = M$ to M is now defined by

$$F(x) = \int_0^1 F(x,\theta) dG(\theta) \tag{2}$$

With the above definition of ρ, then,

Corollary

F is an identifiable mixture iff

$$\int_0^1 [\int_0^1 F(x,\theta)(dG_1(\theta) - dG_2(\theta))]^2 dx \geq \alpha \int_0^1 [G_1(\theta) - G_2(\theta)]^2 \, d\theta \tag{3}$$

for some $\alpha > 0$.

This result will now be used to look at two special cases.

Case 1 (Tallis (1969), Theorem 6)

Suppose it is possible to integrate (2) by parts to get

$$L(x) = \int_0^1 T(x,\theta) G(\theta) d\theta$$

for suitable T and L. Applying (3) the condition becomes

$$\int_0^1 [\int_0^1 [T(x,\theta)(G_1(\theta) - G_2(\theta))] d\theta]^2 dx \geq \alpha \int_0^1 [G_1(\theta) - G_2(\theta)]^2 d\theta$$

and putting $G_1(\theta) - G_2(\theta) = \chi(\theta)$ and

$$K(\theta,\phi) = \int_0^1 T(x,\theta) T(x,\phi) dx$$

we get

$$\int_0^1 \int_0^1 \chi(\theta) K(\theta,\phi) \chi(\phi) d\theta d\phi \geq \alpha \int_0^1 \chi^2(\theta) d\theta \tag{4}$$

Regarding K as a symmetric, linear operator defined on $\mathcal{L}^2[0,1]$ and using the norm $\int_0^1 f^2(x) dx$, (4) can be written as

$$\|K\chi\|^2 \geq \alpha \|\chi\|^2.$$

From the discussion following Theorem 1 this holds for all $\chi \in \mathcal{L}^2[0,1]$ iff the inverse of K exists i.e. zero is not in the spectrum of K and the associated set of eigenfunctions is complete. (See also Tricomi (1957), page 124).

Case 2 (Tallis (1969), Corollary 1)

Let $\{\theta_j\}$ be a countable sequence, $\theta_j < \theta_k$, $j < k$, and put $\mathcal{D}$ equal to the subset of M whose members are absolutely continuous with respect to μ, where μ is the counting measure defined on $\{\theta_j\}$. Now,

$$F(x) = \sum_{i=1}^{\infty} \beta_i F_i(x), \quad \beta_i \geq 0, \quad \sum_{i=1}^{\infty} \beta_i = 1 \qquad (5)$$

where $F_i(x) = F(x,\theta_i)$. The criterion for identifiability becomes

$$\int_0^1 [\sum_{i=1}^{\infty} \gamma_i F_i(x)]^2 \, dx \geq \alpha \int_0^1 [G_1(\theta) - G_2(\theta)]^2 \, d\theta$$

or

$$\underset{\sim}{\gamma}' K \underset{\sim}{\gamma} \geq \alpha \int_0^1 [G_1(\theta) - G_2(\theta)]^2 \, d\theta, \quad k_{ij} = \int_0^1 F_i(x) F_j(x) \, dx.$$

Thus a necessary condition for (5) to be identifiable is that $\underset{\sim}{\gamma}' K \underset{\sim}{\gamma} > 0$, $\underset{\sim}{\gamma} \neq 0$, $\sum_{i=1}^{\infty} \gamma_i^2 < \infty$ and this is true iff K has an inverse i.e. zero is not in the spectrum of K and the associated eigenvectors span ℓ^2.

The latter condition is also sufficient for identifiability, for then there exists a constant α^* such that

$$\underset{\sim}{\gamma}' K \underset{\sim}{\gamma} \geq \alpha^* \| \underset{\sim}{\gamma} \|^2 \geq \alpha^* \int_0^1 [G_1(\theta) - G_2(\theta)]^2 \, d\theta.$$

In conclusion notice that metrics other than $\int_0^1 [F(x) - G(x)]^2 \, dx$ can be used. The fundamental set M can also be adjusted to suit particular problems. For instance $\rho(F,G) = \sup_x |F(x) - G(x)|$ is a possibility for M as used above, although for other metrics and different assumptions it may be convenient to take M as consisting of equivalence classes with respect to a measure. These situations will not be pursued.

REFERENCES

Tallis, G.M. (1969). The identifiability of mixtures of distribution. J. Appl. Prob. 6 : 389-398.

Taylor, A.E. (1958). An Introduction to Functional Analysis. John Wiley & Sons, New York.

Tricomi, F.G. (1957). Integral Equations. Interscience Publishers Inc., New York.

A(d)[5]

J. Appl. Prob. 6, 74–87 (1969)
Printed in Israel

GENERAL MODELS FOR *r*-MOLECULAR REACTIONS

G. M. TALLIS AND
R. T. LESLIE, *C.S.I.R.O., Newtown, N.S.W.**

1. Introduction

In the present paper we consider the r-molecular reversible reaction $rA \rightleftharpoons B$ from several viewpoints. The deterministic theory for integral reaction orders is considered first and is subsequently extended to cover the case of fractional order reactions. Stochastic models are then proposed, the analyses being carried through by spectral methods and, in the case of first order reactions, the first passage time problem is also examined. Finally, we use a diffusion theory approach to the problem to obtain results which are valid for a large number of molecules.

The formulation of stochastic models for chemical reaction kinetics has received considerable attention recently and several cases have been treated in some detail (see McQuarrie (1967)). However, in two important respects, these models tend to fall short of adequacy. Firstly, some of the results are applicable only to solutions containing small numbers of molecules. Secondly, the reaction rates are calculated on the assumption that, in a small period of time, the probability of reaction between a pair of molecules is the same for all molecule pairs.

Models proposed, for example, by Ishida (1964) and Darvey, Ninham and Staff (1966) take the probability of collisions for the case $rA \rightleftharpoons B$ as being proportional to the number of combinations r at a time of the molecules in the whole space occupied by the reactants.

In deriving the differential equations for the stochastic processes, the limiting process as $\Delta t \to 0$ reduces to microscopic dimensions the region in which a single molecule can contact other molecules. This is of course due to the finiteness of the velocity of the particles, a feature which is emphasised more in liquids than in gases. A convincing argument is therefore required to justify the use of the standard combinatorial collision rates.

It seems highly probable that such an argument will not be forthcoming since it is well known that the order of a chemical reaction is not necessarily the same as its molecularity, (see Laidler (1950) Chapter 1). For these reasons in the sub-

Received in revised form 10 June 1968.
* The authors' exact address is C.S.I.R.O., Alpha House, 60 King St., Newtown, N.S.W. 2042.

sequent development combinatorial collision rates are not explicitly used. Instead, the rates of the forward and backward reactions, λ and μ, are allowed to be arbitrary functions of the state of the system.

In order to develop some feel for the problem it is instructive to pursue the following elementary argument. Let the total volume V (taken as unity) of solvent contain a molecules of A at time t and consider the reaction $2A \rightleftharpoons B$. Suppose that V is decomposed into K smaller non-overlapping cells of size $\Delta V = 1/K$. Then, on the assumption of independent occupancy of the cells and multinomial distribution of molecules between cells, the parameters would be a and $1/K = \Delta V = p_i$ $(i = 1, \cdots, K)$.

It is essential to fix the elementary time interval Δt. Given Δt, let ΔV be so chosen that it is the largest volume for which in time Δt all pairs of molecules within ΔV have equal chance of collision. The magnitude of ΔV is then taken to be dependent only on the mean molecular velocity (determined in turn by the temperature), and not on the concentration of A.

Let X_i be the number of A-molecules in the ith cell. Then conditionally on $X_i = x$ the collision rate (probability of a collision in time Δt in cell i) could fairly be taken as $\lambda(\Delta V)\binom{x}{2}\Delta t$. The intensity λ has been written as a function of ΔV since, for fixed Δt, $\lambda(\Delta V)$ will probably be a monotone decreasing function of ΔV.

We seek an expression for the elementary probability that in time Δt the state of the system will change from a to $a-1$. This may be found from the unconditional expected values of the rates for a single cell appropriately combined over all cells. The marginal distribution of $\{X_i\}$ is binomial with $\Pr\{X_i = x\} = \binom{a}{x}(\Delta V)^x(1-\Delta V)^{a-x}$ and, applying this to the conditional rate for cell i, the unconditional rate becomes $\frac{1}{2}a(a-1)\lambda(\Delta V)(\Delta V)^2\Delta t$, denoting the probability (to order $O(\Delta t)^2$) that there is a collision in ΔV in time Δt. If Y_i is the number of collisions in cell i we then have $\Pr\{Y_i = 1\} = \frac{1}{2}a(a-1)\lambda(\Delta V)(\Delta V)^2\Delta t + O(\Delta t)^2$, and since the probability of collisions in two cells simultaneously is $O(\Delta t)^2$, we can write

$$\Pr\left\{\sum_{i=1}^{K} Y_i = 1\right\} = \tfrac{1}{2}a(a-1)\lambda(\Delta V)\Delta V\Delta t + O(\Delta t)^2 \tag{1}$$

and

$$\Pr\left\{\sum_{i=1}^{K} Y_i = 0\right\} = 1 - \Pr\left\{\sum_{i=1}^{K} Y_i = 1\right\}.$$

Hence finally the probability of the system changing from the state represented by a to that represented by $(a-1)$ is (1). It is intuitive however that this probability should depend only on a, representable by $\alpha(a)\Delta t$, whence

$$\alpha(a) = \tfrac{1}{2}a(a-1)\Delta V\lambda(\Delta V)$$

or

$$\lambda(\Delta V) = f(a)/\Delta V \quad \text{for suitable choice of } f.$$

We now have a slightly different way of looking at the problem. It is clear that $\alpha(a)$ is proportional to $a(a-1)$ if and only if $\lambda(\Delta V) = C/\Delta V$, i.e., $f(a) = C$. This implies that the probability $\lambda(\Delta V)\Delta V \Delta t$ that two molecules collide and react should be independent of the concentration of A. When the order of the reaction is the same as the molecularity then $\alpha(a) = O(a^2)$ and $f(a)$ may well be closely approximated by a constant. For reactions of higher molecularity than 2 the factor $\binom{a}{r}$ appears with $\lambda(\Delta V)$, and as there are no known reactions with order greater than 2, we cannot accept combinatorial collision rates.

There still remains the problem of taking the limit $\Delta t \to 0$ as $\lambda(\Delta V)$ and ΔV are dependent on Δt. A possible approach is to accept the associated differential equation as an approximation to the difference equation for finite Δt and assume that the solution of the differential equation is close to the solution of the difference equation.

2. Deterministic theory

We consider the r molecular reaction $r\mathrm{A} \rightleftharpoons \mathrm{B}$ of order n and let $x(t)$ be the proportion of the total concentration C of A molecules in state A at time t. Then we can write the following, in general, non-linear differential equation:

$$\frac{dx(t)}{dt} = \mu[1 - x(t)] - \lambda x(t)^n, \tag{1}$$

where μ is the rate of breakdown of B to A and λ is C^{n-1} times the rate of the forward reaction (see Laidler (1950)).

The explicit solution of (1) for arbitrary n poses difficulties. However, when $\mu = 0$, i.e., the reaction is irreversible, the general solution is, using $x(0) = 1$,

$$x(t) = [\lambda t(n-1) + 1]^{-1/(n-1)} \tag{2}$$

while for $n = 1$, (1) integrates to

$$x(t) = \mu/(\lambda + \mu) + [\lambda/(\lambda + \mu)]e^{-t(\mu+\lambda)}. \tag{3}$$

The equilibrium concentration x_∞ is obtained from (1) by letting t tend to infinity. Thus

$$\lim_{t\to\infty} \frac{dx(t)}{dt} = \mu[1 - x_\infty] - \lambda x_\infty^n = 0$$

and x_∞ is given by the solution of the equation

$$\frac{x_\infty^n}{1 - x_\infty} = \mu/\lambda. \tag{4}$$

We now solve (1) for the case $n = 2$ i.e., the Riccati equation

$$x' = -\lambda x^2 - \mu x + \mu.$$

Firstly note that x_∞ is a particular solution and hence the general solution is $x(t) = x_\infty + [v(t)]^{-1}$ where v is the general solution of the equation

$$v' - (2\lambda x_\infty + \mu)v - \lambda = 0.$$

Thus finally,

$$x(t) = x_\infty + \frac{(2\lambda x_\infty + \mu)}{\lambda}\left[\left(\frac{2\lambda x_\infty + \mu}{\lambda(1 - x_\infty)} + 1\right)e^{(2\lambda x_\infty + \mu)t} - 1\right]^{-1}. \tag{5}$$

For $n \geqq 3$ and integral valued we can attempt the following series solution. Let $x(t) = \sum_{j=0}^{\infty} a_j t^j$ and define the sequence $\{a_j^{(n)}\}$ as the n-fold convolution of the series $\{a_j\}$ with itself, then from (1)

$$\sum_{j=0}^{\infty}(j+1)a_{j+1}t^j + \mu\sum_{j=0}^{\infty}a_j t^j + \lambda\sum_{j=0}^{\infty}a_j^{(n)}t^j = \mu$$

and, equating coefficients of t^j on both sides, $a_0 = 1$, $a_1 = -\lambda$ and

$$a_j = -\frac{1}{j}[\mu a_{j-1} + \lambda a_{j-1}^{(n)}], \quad j > 1. \tag{6}$$

The coefficients can therefore be computed recursively from (6).

An inspection of (3) and (5) suggests that the transformation $y = (x - x_\infty)^{-1} = \sum a_j t^j$ will produce a rapidly converging series. Under this, (1) becomes

$$-y^{-2}y' + \mu(x_\infty + y^{-1}) + \lambda(x_\infty + y^{-1})^n = \mu. \tag{7}$$

If both sides of (7) are multiplied by y^n and the results $y^n[\mu x_\infty - \mu + \lambda x_\infty^n] = 0$, $-y^{n-2}y' = -(n-1)^{-1}\frac{d}{dt}y^{n-1}$ used, we obtain

$$\frac{d}{dt}y^{n-1} = (n-1)\left[\lambda\sum_{k=0}^{n-1}\binom{n}{k}x_\infty^k y^k + \mu y^{n-1}\right]$$

from which we obtain the recursion

$$a_{j+1}^{(n-1)} = \frac{(n-1)}{j+1}\left[\lambda\sum_{k=0}^{n-1}\binom{n}{k}x_\infty^k a_j^{(k)} + \mu a_j^{(n-1)}\right]. \tag{8}$$

Since $y(0) = (1 - x_\infty)^{-1} = a_0$ and

$$a_j^{(n-1)} = \sum a_{i_1}a_{i_2}a_{i_{n-1}} + (n-1)a_0^{n-2}a_j,$$

where the summation is taken over all values of i_k such that $\sum_k i_k = j$ but $i_k \neq j$ all k, it is clear that a_j can also be obtained recursively from (8).

Notice now that $a_j^{(n-1)} = \sum_{k=0}^{j} a_k a_{j-k}^{(n-2)} > a_j^{(n-2)}$ since all terms in the series are positive and $a_0 = (1 - x_\infty)^{-1}$. Thus we have

$$a_j^{(n-1)} < \frac{(n-1)}{j}[\lambda(1+x_\infty)^n + \mu]a_{j-1}^{n-1} < \{(n-1)[\lambda(1+x_\infty)^n + \mu]\}^j a_0^{n-1}/j!$$

and hence the series $\sum a_j t^j$ converges more rapidly than

$$(1 - x_\infty)^{-(n-1)} \exp\{(n-1)[\lambda(1+x_\infty)^n + \mu]t\}.$$

We next consider the case when n is not necessarily an integer. Let $x(t)^n$ have the series expansion $\sum c_j t^j$, then the problem is to define the coefficients c_j in terms of the a_j. Define $\log x(t) = v(t) = \sum b_j t^j$, then $\log x(t)^n = nv(t) = \sum n b_j t^j$ and by identifying coefficients of t^j

$$b_0 = 0 \quad b_2 = a_2 - a_1^2 \quad b_4 = a_4 - 4a_3a_1 - 3a_2^2 + 12a_2a_1^2 - 6a_1^4$$

$$b_1 = a_1 \quad b_3 = a_3 - 3a_2a_1 + 2a_1^3 \quad b_5 = a_5 - 5a_4a_1 - 10a_3a_2 + 20a_3a_1^2$$

$$+ 30a_2^2a_1 - 60a_2a_1^3 + 24a_1^5$$

and

$$c_0 = 1 \quad c_1 = nb_1 \quad c_2 = nb_2 + n^2b_1^2 \quad c_3 = nb_3 + 3n^2b_2b_1 + n^3b_1^3$$

$$c_4 = nb_4 + 4n^2b_3b_1 + 3n^2b_2^2 + 6n^3b_2b_1^2 + n^4b_1^4$$

$$c_5 = nb_5 + 5n^2b_4b_1 + 10n^2b_3b_2 + 10n^3b_3b_1^2 + 15n^3b_2^2b_1 + 10n^4b_2b_1^3 + n^5b_1^5 .$$

Thus the c_j can be expressed in terms of the a_j by substituting for b_j in the above expressions. If more than five terms are required, they may be obtained from the list given in Kendall and Stuart ((1958), pages 69–71), by applying the same procedure as above. Clearly the recursion

$$a_j = -\frac{1}{j}[\mu a_{j-1} + \lambda c_{j-1}], \quad j > 1$$

can now be carried out using $a_0 = 1$ and $a_1 = -\lambda$.

In some cases the constants λ, μ and n are unknown and must be determined by experiment. Suppose $x(t)$ is estimated at the time points $t_1, t_2, \cdots, t_N$ and that $\max(t_{i+1} - t_i)$ is not too large. Then, writing $x(t_i) = x_i$, we can fit a polynomial of degree $(N-1)$, $\sum_{j=0}^{N-1} a_j t^j$, to pass through the points (t_i, x_i), $i = 1, 2, \cdots, N$. Thus

$$\alpha = T^{-1}x$$

where $\alpha' = (\alpha_0, \alpha_1, \cdots, \alpha_{N-1})$, $x' = (x_1, x_2, \cdots, x_N)$ and T is the $(N \times N)$ Vandermonde matrix $[t_i^{j-1}]$. The parameters λ, μ, and n may now be calculated by the principle of least squares by setting

$$F(\lambda, \mu, n) = \sum_{i=1}^{N} \left[\sum_{j=0}^{N-1} \alpha_j j t_i^{j-1} - \mu(1 - x_i) + \lambda x_i^n\right]^2$$

and carrying out the required minimisation by standard methods.

3. Discrete stochastic theory

(a) *Distribution*

We are still concerned with the reaction $rA \underset{\mu}{\overset{\lambda}{\rightleftharpoons}} B$ and initially it is assumed that the reaction is of the first order. Suppose there is a total of rN molecules of A then if λ is the rate at which r molecules of A combine to form one molecule of B and X is the number of B molecules,

$$p_x(t+\Delta t) = p_x(t)(1-\mu x\Delta t - \lambda\overline{N-x}\Delta t) + p_{x-1}(t)\lambda\overline{N-x+1}\Delta t + p_{x+1}(t)\mu\overline{x+1}\Delta t.$$

Putting $p_{-1}(t) \equiv p_{N+1}(t) \equiv 0$, we obtain a birth and death type differential equation

$$(1) \qquad p_x'(t) = -p_x(t)(\mu x + \lambda\overline{N-x}) + p_{x-1}(t)\lambda\overline{N-x+1} + p_{x+1}(t)\mu\overline{x+1}$$

which holds for $0 \leqq x \leqq N$. Multiplying both sides of (1) by s^x and summing from $x = -1$ to $x = N+1$ it is found that the probability generating function for $p_x(t)$, $P(s,t)$, satisfies the following Lagrange partial differential equation:

$$(2) \qquad \frac{\partial P(s,t)}{\partial t} = N\lambda(s-1)P(s,t) + [\mu - s(\mu-\lambda) - \lambda s^2]\ \frac{\partial P(s,t)}{\partial s}.$$

From (2) we obtain the equivalent system of equations

$$\frac{dt}{-1} = \frac{ds}{(1-s)(\mu+\lambda s)} = \frac{dP}{-N\lambda(s-1)P},$$

which give the independent solutions

$$(1-s)(\mu+\lambda s)^{-1}e^{-t(\mu+\lambda)} = C_1$$

$$P(\mu+\lambda s)^{-N} = C_2,$$

and the general solution is

$$P = [\mu+\lambda s]^N f\left(\frac{1-s}{\mu+\lambda s}e^{-t(\mu+\lambda)}\right),$$

where f is an arbitrary function.

If there are m B units at $t=0$, $0 \leqq m \leqq N$, the boundary condition is $P(s,0) = s^m$ and hence

$$f\left(\frac{1-s}{\mu+\lambda s}\right) = s^m/(\mu+\lambda s)^N$$

and

$$f(x) = (\mu+\lambda)^{-N}[1-\mu x]^m[1+\lambda x]^{N-m}.$$

After some algebra we obtain finally

$$P(s,t) = (\mu+\lambda)^{-N}[\mu(1-e^{-t(\mu+\lambda)}) + s(\lambda+\mu e^{-t(\mu+\lambda)})]^m \times [\mu+\lambda e^{-t(\mu+\lambda)} + \lambda s(1-e^{-t(\mu+\lambda)})]^{N-m}. \tag{3}$$

The steady state distribution is obviously binomial with parameters $p = \lambda/(\mu+\lambda)$ and N while for $m = 0$ the general solution is also binomial with parameters $p = \lambda(1-e^{-t(\mu+\lambda)})/(\lambda+\mu)$ and N. For $m \neq 0, N$, (3) is the convolution of two binomial distributions and hence the mean and variance of X is easily obtained.

Since $X = X_1 + X_2$, $E(X) = E(X_1) + E(X_2)$ and $V(X) = V(X_1) + V(X_2)$, we have

$$E(X) = (\mu+\lambda)^{-1} m(\lambda+\mu e^{-t(\mu+\lambda)}) + (N-m)(\mu+\lambda)^{-1}\lambda(1-e^{-t(\mu+\lambda)})$$

$$V(X) = m(\mu+\lambda)^{-2}(\lambda+\mu e^{-t(\mu+\lambda)})\mu(1-e^{-t(\mu+\lambda)}) + (N-m)(\mu+\lambda)^{-2}(\mu+\lambda e^{-t(\mu+\lambda)})\lambda(1-e^{-t(\mu+\lambda)}). \tag{4}$$

Putting $m = 0$ in the formula for $E(X)$ we obtain the desired agreement with (3) of 2, since in the deterministic case it is the concentration of A which is followed.

The case $n = 2$ has been solved by Ishida (1964) for irreversible reactions only. The difficulties of obtaining explicit solutions for the general situations are emphasised by this work and we therefore turn to other methods.

Again we assume that there are N r A units and let $\lambda(N-x)$ be the rate at which B units are formed given there are x units of B and $\mu(x)$ be the rate of breakdown of B to A. Then setting $p_{-1}(t) \equiv p_{N+1}(t) \equiv 0$ we have the differentio-difference equation

$$p_x'(t) = -p_x(t)\{\lambda(N-x) + \mu(x)\} + p_{x-1}(t)\lambda(N-x+1) + p_{x+1}(t)\mu(x+1) \tag{5}$$

which holds for $0 \leqq x \leqq N$. Such equations have been studied, for example, by Moran (1963), where further references will be found. Now let

$$A = \begin{bmatrix} -\lambda(N) & \mu(1) & 0 & 0 & \cdots & 0 & 0 \\ \lambda(N) & -\lambda(N-1)+\mu(1) & \mu(2) & 0 & \cdots & 0 & 0 \\ \cdot & \cdot & \cdot & \cdot & & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdots & \cdot & \cdot \\ 0 & 0 & 0 & 0 & \cdots & \lambda(1) & -\mu(N) \end{bmatrix},$$

then we can write the system (5) in the vector form

$$p'(t) = Ap(t) \tag{6}$$

with the initial condition $p(0)$. The solution to (6) is therefore

$$p(t) = e^{At}p(0) \tag{7}$$

where $e^{At} \equiv \sum_{j=0}^{\infty} A^j t^j / j!$. However, (7) is not in a useful form for large values of t and it seems preferable to adopt a spectral approach.

Suppose A has $N+1$ distinct characteristic roots $\theta_0, \theta_1, \cdots, \theta_N$, $\theta_i \geqq \theta_j$ for $i \geqq j$, then, as can easily be verified directly, A is singular since $p'_N(t) = -\sum_{j=0}^{N-1} p'_j(t)$ and hence $\theta_N = 0$. Let $\Omega = \operatorname{diag}(\theta_0, \theta_1, \cdots, \theta_N)$ and H be defined by $\Omega = H^{-1}AH$, the transformation $y(t) = H^{-1}p(t)$ then reduces (6) to

$$y'(t) = H^{-1}AHy(t) = \Omega y(t). \tag{8}$$

We solve (8) component-wise and write $y_j(t) = e^{\theta_j t}c_j$. Now letting $E(t) = \operatorname{diag}(e^{\theta_0 t}, e^{\theta_1 t}, \cdots, 1)$ we have $y(t) = E(t)c$ and $p(t) = HE(t)c$. But $p(0) = Hc$ and hence finally

$$\begin{aligned} p(t) &= HE(t)H^{-1}p(0) \\ &= \sum_{j=0}^{N} e^{\theta_j t} h_j h_j^{-} p(0), \end{aligned} \tag{9}$$

where h_j is the jth column of H and h_j^{-} is the jth row of H^{-1}. The limiting form of (9) is

$$p(\infty) = h_N h_N^{-} p(0). \tag{10}$$

However, the stationary distribution can be obtained explicitly from (5). In fact, setting $p'_x(t) = 0$ and writing $p_x(\infty) = p_x$, we obtain the equation

$$p_x[\mu(x) + \lambda(N-x)] = p_{x-1}\lambda(N-x+1) + p_{x+1}\mu(x+1)$$

which leads to the recursion relation

$$\lambda(N-x)p_x = \mu(x+1)p_{x+1}$$

whence

$$p_x = \prod_{j=0}^{x-1} \frac{\lambda(N-j)}{\mu(j+1)} p_0 \tag{11}$$

and p_0 is chosen so that $\sum_{x=0}^{N} p_x = 1$.

For example, when $r = n = 2$ the appropriate terms for λ and μ are $\lambda(N-j) = \lambda\overline{2N-j}\ \overline{2N-2j-1}/2$ and $\mu(j+1) = \mu\overline{j+1}$ where λ and μ are constants. Applying (11) it is found that

$$p_x = p_0(\lambda/2\mu)^x [2N]_{2x}/x!$$

In order to verify that $p(\infty)$ exists we notice that since A is a band matrix there exists a non-singular diagonal matrix D such that $B = D^{-1}AD$ is a symmetric band matrix (Bodewig (1959), page 259). Moreover,

$$|A - \theta I| = |B - \theta I|$$

and hence all the θ_i are real.

However, e^{At} has positive elements if and only if $a_{ij} \geqq 0$, for $i \neq j$, (Bellman (1960), page 172), and hence $p_x(t) \geqq 0$ for all x. It is not difficult to verify now that, in view of the constraint $\sum p_x(t) = 1$, $0_i \leqq 0$ for all i which implies the existence of the limiting distributions.

(b) *First passage times*

It may be of interest in the case of the discrete stochastic models discussed above to know the distribution of the first passage time through, say, $N - a$. That is, starting with N molecules in state A, $x = N$, what can be said with regard to the time required until there are a molecules in state B for the first time from the start of the experiment?

It appears to be simplest to work with the backward equations making $N - a$ an absorbing barrier. Thus,

$$(12)\quad p_y'(t) = -[\lambda y + \mu(N-y)]p_y(t) + \lambda y p_{y-1}(t) + \mu(N-y)p_{y+1}(t), N-a < y \leqq N$$

$$p_{N-a}(t) \equiv 1,$$

where $p_y(t)$ stands for $p_{y,N-a}(t)$. Now let

$$p_y^* = \int_0^\infty e^{-st} p_y(t)\, dt,$$

then (12) can be transformed into the second order, homogeneous difference equation

$$(13)\quad \mu(N-y)p_{y+1}^* + \lambda(y)p_{y-1}^* - [\lambda(y) + \mu(N-y) + s]\, p_y^* = 0, N-a+1 \leqq y \leqq N$$

$$p_{N-a}^* = 1/s.$$

Equation (13) can be solved in principle by the method of Laplace (Jordan (1960)), for polynomial $\lambda(i)$, $\mu(i)$, but here we consider explicitly only the linear case $\lambda(i) = \lambda i$, $\mu(i) = \mu i$, with λ and μ constant. It should be noted that at the upper limit, N, of the range of values of y for which the recurrence relation holds, the coefficient $\mu(N-y)$ of one of the extreme terms, p_{y+1}^*, vanishes. This implies that the system is effectively of first order, for which a single boundary condition suffices to make the problem determinate; this is supplied as $sp_{N-a}^* = 1$. We proceed as though the equation is second order, and expect to find that one of the solutions is singular or will not satisfy the boundary condition.

The general solution is then of the form

$$p_y^* = c_1 \int_0^{t_1} t^{y-1} v\, dt + c_2 \int_0^{t_2} t^{y-1} v\, dt,$$

where t_1 and t_2 are the roots of

$$-\mu t^2 + (\mu - \lambda)t + \lambda = 0$$

and

$$\log v = D^{-1}\left\{\frac{t^2\mu(N+1) - t(s+\mu N) + \lambda}{t(1-t)(\mu t+\lambda)}\right\}.$$

Hence $t_1 = 1$, $t_2 = -\lambda/u$ and the initial condition is $p^*_{N-a} = 1/s$. However, the root $-\lambda/\mu$ leads to a divergent integral and hence we set $c_2 = 0$.

Since $p_N(t)$ is the probability that the first passage time T does not exceed t, we require the Laplace transform $Q(s)$ of $p'_N(t)$ to define the density function of T. By using the initial condition it can be shown that

$$Q(s) = \frac{\int_0^1 t^N(1-t)^{s/(\lambda+\mu)-1}(\mu t+\lambda)^{-(N+1+s)}dt}{\int_0^1 t^{N-a}(1-t)^{s/(\lambda+\mu)-1}(\mu t+\lambda)^{-(N+1+s)}dt}. \tag{14}$$

It may be noted that while the integrals in (14) are defined only for $s > 0$, their ratio tends to 1 as $s \to 0$; this follows on integration by parts and cancellation of the factor $(\lambda+\mu)/s$.

The mean first passage time is then found to be

$$E(T) = Q'(0) = \lim_{s\to 0} s^{-1}(1-Q(s)) \tag{15}$$

$$= (\lambda+\mu)^N \int_0^1 t^{N-a}(1-t^a)(1-t)^{-1}(\mu t+\lambda)^{-(N+1)}dt.$$

4. Diffusion theory

(a) *Distribution*

The discrete stochastic results discussed above seem to be of limited value. In practice the number of molecules will be extremely large and it appears that it is the asymptotic distribution for large N that is relevant. This distribution cannot be readily studied unless tractable, explicit solutions are obtained. For arbitrary A, the exact behaviour of the latent roots as N gets large seems to be more or less unpredictable (see Lederman and Reuter (1952)), and therefore another approach to the problem is required.

We assume here that the concentration of A, at time t, $X(t)$ undergoes a diffusion process. Specifically let

$$\theta(x) = \mu\exp\{-K_1\Delta^2 x\}(\Delta t)^{\frac{1}{2}}(1-x) + \alpha(x)/2$$

$$\phi(x) = \lambda\exp\{-K_2\Delta^2 x\}(\Delta t)^{\frac{1}{2}}x^n + \alpha(x)/2 \tag{1}$$

be the probabilities that $X(t)$ moves from x to $x+\Delta x$ and x to $x-\Delta x$ respectively during time $(\Delta t)^{\frac{1}{2}}$. The probability of no change is $1-\theta-\phi$. In (1) μ, λ, K_1 and K_2 are constants while $\alpha(x)$ is a small disturbance.

It is now found that the instantaneous mean is

$$\lim_{\Delta x,\Delta t\to 0}\{\theta(x)-\phi(x)\}\Delta x/\Delta t = \mu(1-x)-\lambda x^n,\ \Delta^2 x = \Delta t \tag{2}$$

and the instantaneous variance is

$$\lim_{\Delta x,\Delta t\to 0}\{\theta(x)+\phi(x)-[\theta(x)-\phi(x)]^2\}\Delta^2 x/\Delta t = \alpha(x). \tag{3}$$

Now by definition the instantaneous mean is

$$\beta(x) = \lim_{\Delta t\to 0} E[X(t+\Delta t)-X(t)\mid X(t)=x]/\Delta t$$

and it is clear that this expression reduces to the differential equation for the deterministic process when $X(t)$ is not a random function.

From (2) and (3) we can write down the appropriate time-homogeneous differential equation for $p(x,t)$, the frequency function for $X(t)$, as

$$\frac{\partial}{\partial x}\left\{\frac{\partial}{\partial x}\alpha(x)\,p(x,t)/2-\beta(x)\,p(x,t)\right\} = \frac{\partial}{\partial t}p(x,t) \tag{4}$$

with boundary condition

$$\left[\frac{\partial}{\partial x}\alpha(x)\,p(x,t)/2-\beta(x)\,p(x,t)\right]_{x=0}^{x=1} = 0. \tag{5}$$

The unconstrained stationary distribution is obtainable from (4) immediately. Let $F(x) = 2\int^x\{\beta(w)/\alpha(w)\}\,dw$ be such that

$$\lim_{\alpha(x)\to 0}[\alpha(x)]^{-1}\exp\{F(x)-F(x_\infty)\} = 0,\quad x\neq x_\infty,$$

then the limiting distribution $t\to\infty$ is

$$\Pi(x) = C[\alpha(x)]^{-1}\exp\{F(x)-F(x_\infty)\},\ C^{-1} = \int_0^1[\alpha(x)]^{-1}\exp\{F(x)-F(x_\infty)\}dx \tag{6}$$

and it has the property that

$$\lim_{\alpha(x)\to 0}\Pi(x) = \delta(x-x_\infty)$$

in agreement with the deterministic theory.

For example, if $\alpha(x)\equiv\sigma^2$ then (6) becomes

$$\Pi(x) = C\sigma^{-2}\exp\{-x\sigma^{-2}[-\mu(2-x)+2\lambda x^n/(n+1)] - \sigma^{-2}[2n\mu-(n-1)\mu x_\infty]\,x_\infty/(n+1)\}$$

and for $n=1$ we obtain the more familiar form

$$\Pi(x) = C\exp\left\{-\frac{(\mu+\lambda)}{\sigma^2}[x-\mu/(\mu+\lambda)]^2\right\},$$

where $C = (\Pi\sigma^2)^{-\frac{1}{2}}(\mu+\lambda)^{\frac{1}{2}}\{\Phi(k_1) - \Phi(k_2)\}^{-1}$, $\Phi(x)$ is the distribution function of a standard normal variate,

$$k_1 = \lambda 2^{\frac{1}{2}}/(\mu+\lambda)^{\frac{1}{2}}\sigma \quad \text{and} \quad k_2 = -\mu 2^{\frac{1}{2}}/(\mu+\lambda)^{\frac{1}{2}}\sigma.$$

In general, (4) can be solved approximately subject to (5) by the methods of Keilson (1964).

For the case $n = 1$ we attempt a general solution. This is facilitated by the change of variable $y = 1 - x$ and the appropriate differential equation becomes

$$\frac{\sigma^2}{2}\frac{\partial^2 p}{\partial y^2} - \frac{\partial}{\partial y}\{(\lambda(1-y) - \mu y)p\} = \frac{\partial p}{\partial t} \tag{7}$$

with initial condition $p(y,0) = \delta(y)$ and boundary condition $\int_0^1 p(y,t)dy = 1$. Now let

$$\phi(\theta,t) = \int_{-\infty}^{\infty} e^{-\theta y} p(y,t)\,dy,$$

then (7) transforms to the simple Lagrange form

$$\frac{\partial\phi}{\partial t} = \frac{\sigma^2}{2}\theta^2\phi - \theta\left[(\lambda+\mu)\frac{\partial\phi}{\partial\theta} + \lambda\phi\right] \tag{8}$$

with initial condition $\phi(\theta,0) = 1$. The equivalent system of equations for (8) is

$$\frac{dt}{1} = \frac{d\theta}{\theta(\lambda+\mu)} = \frac{-d\phi}{\phi\theta(\lambda - \sigma^2\theta/2)}$$

which leads to the solutions

$$\theta e^{-(\lambda+\mu)t} = C_1, \quad \exp\{(\lambda\theta - \sigma^2\theta^2/4)/(\lambda+\mu)\}\phi = C_2.$$

Thus

$$\phi(\theta,t) = \exp\{(-\lambda\theta + \sigma^2\theta^2/4)/(\lambda+\mu)\} f(\theta e^{-(\lambda+\mu)t})$$

where f is an arbitrary function. Using the initial condition it is found that

$$\phi(\theta,t) = \exp\{-\lambda\theta(1 - e^{-t(\lambda+\mu)})/(\lambda+\mu) + \theta^2\sigma^2(1 - e^{-2t(\lambda+\mu)})4(\lambda+\mu)\}. \tag{9}$$

It is clear from (9) that the unrestricted process has a normal distribution with mean $\mu(t) = \lambda(1 - e^{-t(\lambda+\mu)})/(\lambda+\mu)$ and variance $\sigma^2(t) = \sigma^2(1 - e^{-2t(\lambda+\mu)})/2(\lambda+\mu)$. In order to impose the boundary condition we notice that, for small σ^2, $\int_0^1 p(y,t)dy$ is essentially equal to $\int_0^\infty p(y,t)dy$. Therefore it should be sufficient to impose the boundary condition

$$\frac{\sigma^2}{2}\frac{\partial p(0,t)}{\partial y} - \beta(0)\,p(0,t) = 0. \tag{10}$$

Consider the expression

$$p^*(y,t) = f(y,t) + g(t) \int_{-\infty}^{-\varepsilon} a(u) f(y-u,t)\,du, \quad \varepsilon > 0, \tag{11}$$

where $f(y,t)$ is the normal frequency function with mean and variance given above and $g(t)$ and $a(u)$ are suitable functions. Equation (11) satisfies the initial condition $p^*(y,0) = \delta(y)$ and will approximately satisfy (7) if $g'(t) \doteq 0$. Now

$$p^*(0,t) = f(0,t) + g(t) \int_{-\infty}^{-\varepsilon} a(u) f(-u,t)\,du$$

and

$$\frac{\partial p^*(0,t)}{\partial y} = u(t) f(0,t)/\sigma^2(t) - g(t) \int_{-\infty}^{-\varepsilon} a(u) \frac{\partial}{\partial u} f(-u,t)\,du \tag{12}$$

$$= u(t) f(0,t)/\sigma^2(t) - g(t)\, a(-\varepsilon) f(-\varepsilon,t) + g(t) \int_{-\infty}^{-\varepsilon} a'(u) f(-u,t)\,du.$$

Notice that (12) holds for arbitrary $\varepsilon > 0$ and, from (9), $u(t)/\sigma^2(t) = 2\lambda/\sigma^2(1+e^{-t(\lambda+u)})$. Then, from (10) and (11) it follows that we can obtain suitable g and a from the formulae $\lambda(1+e^{-t(\lambda+u)})^{-1} - g(t)\,a(0)\sigma^2/2 - \lambda = 0$ and

$$\int_{-\infty}^{-\varepsilon} [a'(u) - 2\lambda/\sigma^2 a(u)] f(-u,t)\,du = 0.$$

Thus

$$g(t) = (1+e^{t(\lambda+u)})^{-1} \text{ and } a(u) = -2\lambda/\sigma^2 e^{2\lambda u/\sigma^2}$$

and

$$p^*(y,t) = f(y,t) - (1+e^{t(\lambda+u)})^{-1} 2\lambda\sigma^{-2} \int_{-\infty}^{-0} e^{2\lambda u/\sigma^2} f(y-u,t)\,du. \tag{13}$$

Since $g'(t)$ tends to zero exponentially, p^* should be a good approximation to p for moderate values of t. For small σ^2, as would be expected in this case, we have the approximation $p(y,t) \doteq (1+e^{-t(\lambda+u)})^{-1} f(y,t)$, $0 \leqq y \leqq 1$.

If more accurate approximations to p are required for small values of t, set $g(t) \equiv 1$ and use (11) to obtain solutions to (7) for various small intervals of t. If this is done, each solution satisfies (7) exactly and the boundary condition is approximately satisfied.

(b) *Mean first passage times*

Supposing the reaction starts at $y = a$, $(y = 1-x)$, and it is required to know $E\{T(b)\}$, where $T(b)$ is the first passage time through b, $a < b \leqq 1$. Then, if $m(a) = E\{T(b)\}$, we have the following differential equation to solve:

$$\frac{\sigma^2}{2}\frac{d^2m}{da^2} + \{\lambda(1-a)^n - \mu a\}\frac{dm}{da} = -1 \tag{14}$$

subject to the boundary conditions $dm(0)/da = 0$ and $m(b) = 0$ (Cox and Miller (1965)).

The solution to (14) is found to be

$$m(a) = \frac{2}{\sigma^2}\int_a^b \exp\left\{\frac{\sigma^2}{2}\left[\frac{\lambda(1-t)^{n+1}}{n+1} + \frac{\mu t^2}{2}\right]\right\}\int_0^t \exp\left\{-\frac{\sigma^2}{2}\left[\frac{\lambda(1-x)^{n+1}}{n+1} + \frac{ux^2}{2}\right]\right\}dx\,dt.$$

References

BELLMAN, R. (1960) *Introduction to Matrix Analysis*. McGraw-Hill, New York.

BODEWIG, E. (1959) *Matrix Calculus*. North-Holland, Amsterdam.

COX, D. R. AND MILLER, H. D. (1965) *The Theory of Stochastic Processes*. John Wiley & Sons Inc., New York.

DARVEY, I. G., NINHAM, B. W. AND STAFF, P. J. (1966) Stochastic models for second order chemical reaction kinetics. The equilibrium state. *J. Chem. Phys.* 45, 2145–2155.

ISHIDA, K. (1964) Stochastic model for bimolecular reaction. *J. Chem. Phys.* 41, 2472–2478.

JORDAN, K. (1960) *Calculus of Finite Differences*. Chelsea, New York.

KEILSON, J. (1964) A review of transient behavior in regular diffusion and birth-death processes. *J. Appl. Prob.* 1, 247–266.

KENDALL, M. G. AND STUART, A. (1958) *The Advanced Theory of Statistics*. Charles Griffin & Co. Ltd., London.

LAIDLER, K. L. (1950) *Chemical Kinetics*. McGraw-Hill, New York.

LEDERMANN, W. AND REUTER, G. E. H. (1953) Spectral theory for the differential equations of simple birth and death processes. *Phil. Trans.* A 246, 321–369.

MCQUARRIE, D. A. (1967) Stochastic approach to chemical kinetics. *J. Appl. Prob.* 4, 413–478.

MORAN, P. A. P. (1963) Some general results on random walks, with genetic applications. *J. Aust. Math. Soc.* 3, 468–479.

A (d)[6]

Reprinted from
BIOMETRICS
THE BIOMETRIC SOCIETY, Vol. 23, No. 1, March 1970

ESTIMATING THE DISTRIBUTION OF SPHERICAL AND ELLIPTICAL BODIES IN CONGLOMERATES FROM PLANE SECTIONS

G. M. TALLIS

Division of Mathematical Statistics, C. S. I. R. O., 60 King St., Newtown, N.S.W., 2042, Australia

SUMMARY

Problems associated with the estimation of the distribution of a certain material, B, in a conglomerate, C, are discussed. The well-known and general result that the proportion of B in C can be estimated without bias from a random plane cut of C is presented briefly. Assumptions with regard to the shape of the deposits of B are then made, thus allowing more information to be obtained from two-dimensional analysis.

Initially the bodies of B are taken to be spherical and some extensions and modifications to Wicksell's classical results in this field are suggested. In particular, the Holmes bias of thin section analysis is considered and this problem is dealt with in general for spherical bodies.

Subsequently, the shapes of the deposits are considered to be ellipsoidal with a special orientation pattern. Some explicit results are obtained with little effort by methods analogous to those introduced in the spherical case.

Some of the formulae are illustrated by means of two short, numerical examples.

1. INTRODUCTION

Aspects of the following problem are considered in this paper. A certain material, which will be called A, contains bodies consisting of a second type of material, B. The conglomerate of A and B will be referred to as C. It is required to estimate the proportion of C which is occupied by B from observations made on plane sections or thin slices.

If certain assumptions are made with regard to the shapes of the bodies, more information can be obtained. In some cases it is feasible to estimate from sections the actual distribution function of the volumes of the deposits of B. This is true, for instance, in the classical cases where the bodies are spherical or elliptical with special orientation patterns.

This type of situation arises in various branches of science. In petrography, modal analysis can be undertaken by studying thin sections. The ratio of the total area of the section to the area occupied by the particular rock type B provides an estimate of the concentration of B in the rock mass. Apparently, these ideas are far from new although they have only recently been put onto a satisfactory statistical footing (Chayes [1956]).

Again, in pathology for instance, it is sometimes important to obtain an estimate of the proportion of a particular mass of tissue occupied by certain

glands or cell types, Thin sections of the tissue can be prepared and photographed and precisely the same procedures of petrography applied. The proportion of the total area of the photograph occupied by the glands can be ascertained to obtain the required estimate.

This paper arose from studies in wool growth which concerned the distribution of wax glands in the skin. These glands are roughly ellipsoidal and have one axis more or less normal to the skin surface. Trephine samples of the skin are taken from which vertical and transverse sections are cut and photographed. After enlargement of the photographs, sections of the wax glands can be identified and measured. An estimate of the distribution of the glands in various breeds of sheep could provide some insight as to their function in wool growth.

Basic mathematical papers in this area were written by Wicksell [1925; 1926]. He related the distribution of the diameter of spheres embedded in a mass of material to the distribution of the diameter of circles formed by randomly cutting the mass with a plane. These circles on the face of the cut were measured to obtain the necessary information from which the distribution of the diameter (or volume) of the spheres was ultimately inferred. Wicksell also obtained similar results for ellipsoidal bodies although the general case where the ellipsoids have random orientation proved intractable. There is a summary of these findings in Kendall and Moran [1963].

Some of the procedures which Wicksell suggests require complicated calculations. It is our purpose to simplify these methods where possible by allowing several plane sections to be taken and by introducing parametric models. The relationship between the moments of circle radii obtained in plane sections to the moments of the radii of spheres in the mass is obtained by a direct argument. This allows Wicksell's results to be obtained quickly and also provides an approach for more complicated situations. The treatment is more general than that of the original papers.

It may happen that, for various reasons, bodies greater or smaller than a certain size cannot be measured satisfactorily on the plane sections. This was the case in the studies described by Wicksell where bodies less than a certain size could not be properly identified. Although Wicksell neglected this source of bias, truncation type procedures are required. These are developed below.

There is an interesting source of bias related to thin section analysis known as the Holmes effect. If the material B is opaque and is set in transparent material A, the amount of B will be systematically overestimated if thin sections are used to obtain area measurements in transmitted light. This is because the apparent area of a particular body will be that of the maximum cross-sectional area of this body in the section. What is actually required, of course, is the area at the surface of the section.

If $2k$ is the width of the section, it is almost obvious without calculation that for a sphere of radius r this bias results in an overestimate of $\pi r^2 k/(r + k)$ for the average of cross-sectional area (see Figure 3). For a fuller discussion of this topic see Chayes [1956]. The Holmes effect is examined in this paper

from the point of view of Wicksell's fundamental integral equation relating distribution of radii of spheres in the sampled mass to the distribution of circle radii in plane sections.

2. GENERAL RESULT RELATING AREA TO VOLUME

It was pointed out in the Introduction that, in order to obtain an unbiased estimate of the proportion of the total volume of the conglomerate C occupied by the material B, it is sufficient to measure this proportion on the face of a plane cut of C. This result is sufficiently important and easy to establish that the proof will be outlined here for completeness.

Consider a unit volume of conglomerate C consisting of two types of material, A and B. For $0 \leq x \leq 1$ let $a(x)$ be the area occupied by B in the vertical plane located at x (see Figure 1). Clearly the total volume of B in the unit volume of C is

$$V_B = \int_0^1 a(x)\, dx = E\{a(x)\}$$

if the frequency function of the cutting plane is uniform on the unit interval.

Now define

$$\sigma_B^2 = \int_0^1 a^2(x)\, dx - V_B^2$$

and let n random planes intersect C at x_i with associated areas $a_i = a(x_i)$, $i = 1, 2, \cdots, n$. If $\bar{a} = \sum_1^n a_i/n$ and $s^2 = \sum_1^n (a_i - \bar{a})^2/(n-1)$ then $E\{\bar{a}\} = V_B$ and $E\{s^2\} = \sigma_B^2$. Although the distribution of $\bar{a}$ is not known in general, $v = (\bar{a} - V_B)/s$ will be asymptotically distributed as a standardised normal variate for $\sigma_B^2 > 0$.

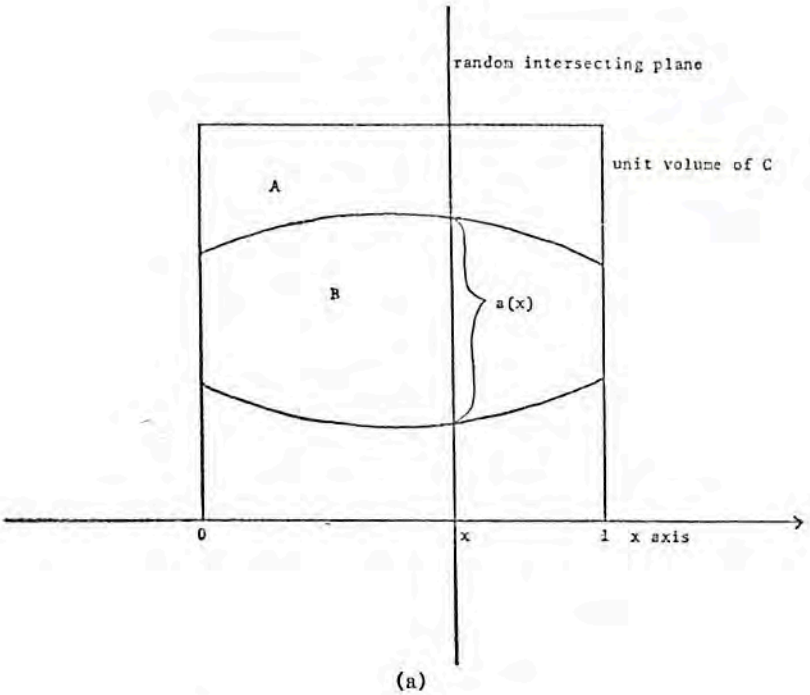
The parameter σ_B^2 has intuitive appeal in describing the heterogeneity of the distribution of B throughout C. For example, if B lies in horizontal strata, σ_B^2 will be small and great precision is expected from two-dimensional analyses.

Referring back to the problem of wax glands in the skin of sheep, the procedure is very simple when only an estimate of the proportion of the total tissue occupied by the glands is required. From several, n say, photographs of skin sections the proportion of area associated with the glands to the total area of the photographs, a_i, can be obtained. The statistic s^2 is then calculated and v used to find a suitable confidence interval for V_B.

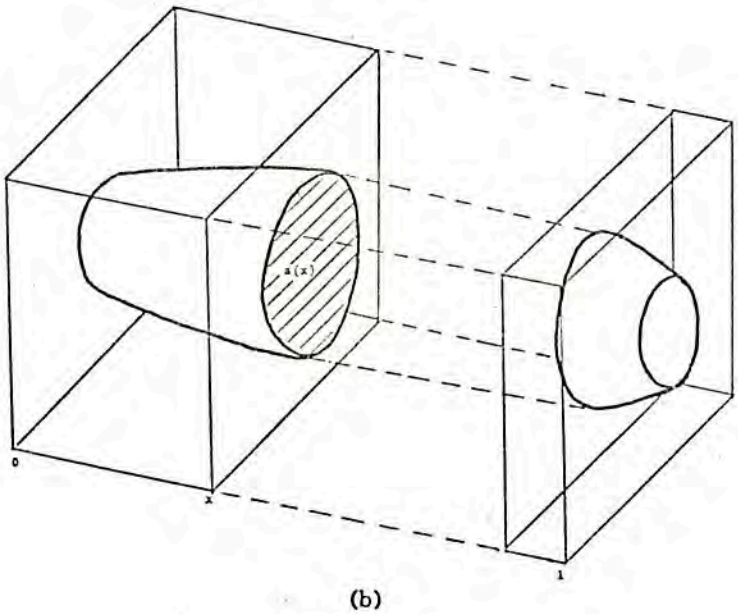
3. SPHERICAL BODIES

(a) *General*

It is useful to build notation and technique by considering first the case where the bodies of material B are spherical. Our approach short-cuts some of Wicksell's development and allows for extensions. The work is carried out in terms of the radius of the spheres, r, and some of the Wicksell notation has been changed as a concession to modern convention. Let



(a)



(b)

FIGURE 1

PLANE SAMPLING IN THREE DIMENSIONS

λ = the average number of sphere centres per unit volume of conglomerate C,

$G(r)$ = the distribution function of sphere radii, r, of material B, with power moments μ_i and derivative $g(r)$ (when it exists),

$F(r)$ = the distribution function of sphere radii which intersect a random plane through C, with power moments m_i and derivative $f(r)$ (when it exists),

$\Phi(y)$ = the distribution function of the radius of circles, y, formed on a random plane through C by intersection with the spheres, with power moments η_i and derivative $\phi(y)$ (when it exists).

It will be assumed that all the moments of G exist and, in fact, that $m_G(t) = \int_0^\infty e^{tr}\, dG(r)$ exists for some $t > 0$. This is a sufficient condition for the moment sequence μ_i to uniquely determine G. It can be shown that existence of $m_G(t)$ implies the existence of $m_F(t)$ which in turn implies the existence of $m_\Phi(t)$.

The first task is to establish the relationship between G and F. Consider a small interval I; then the average number of spheres with radius $r \in I$ which intersect a random plane is approximately $2\lambda r \int_I dG(r)$ per unit area of the plane. Hence the relationship

$$dF(r) = r\, dG(r)/r_0\,, \qquad r_0 = \int_0^\infty r\, dG(r) = \mu_1\,, \tag{3.1a}$$

and if G is absolutely continuous

$$f(r) = r\, g(r)/r_0\,, \qquad r_0 = \int_0^\infty r\, g(r)\, dr = \mu_1\,. \tag{3.1b}$$

If y is the radius of a circle on the intersecting plane with associated sphere of radius r, and if w is the distance from the centre of the sphere to the slice, then $y^i = (r^2 - w^2)^{\frac{1}{2}i}$ (see Figure 2). It will be assumed that the position of the cut has density $dw/2r$ so that

$$E\{y^i \mid r\} = \int_{-r}^{r} (r^2 - w^2)^{\frac{1}{2}i}\, dw/2r = r^i C_i^{-1},$$

where

$$C_i^{-1} = \int_0^{\frac{1}{2}\pi} (\cos\theta)^{i+1}\, d\theta = \frac{(2m-1)!!\,\pi}{(2m)!!\,2}\,, \qquad j = 2m-1$$

$$= \frac{(2m)!!}{(2m+1)!!}\,, \qquad j = 2m$$

and $n!! = n(n-2)(n-4)\cdots$, $1!! = 0!! = (-1)!! = 1$. Thus the unconditional expectation of $C_i y^i$ is $C_i\eta_i = m_i = \mu_{i+1}/\mu_1$, $j \geq 1$. Since

$$\int_0^\infty r^{-1}\, dF(r) = r_0^{-1} \int_0^\infty dG(r) = r_0^{-1}, \tag{3.2}$$

the relationship $2\eta_{-1}/\pi = m_{-1} = r_0^{-1} = \mu_1^{-1}$ holds.

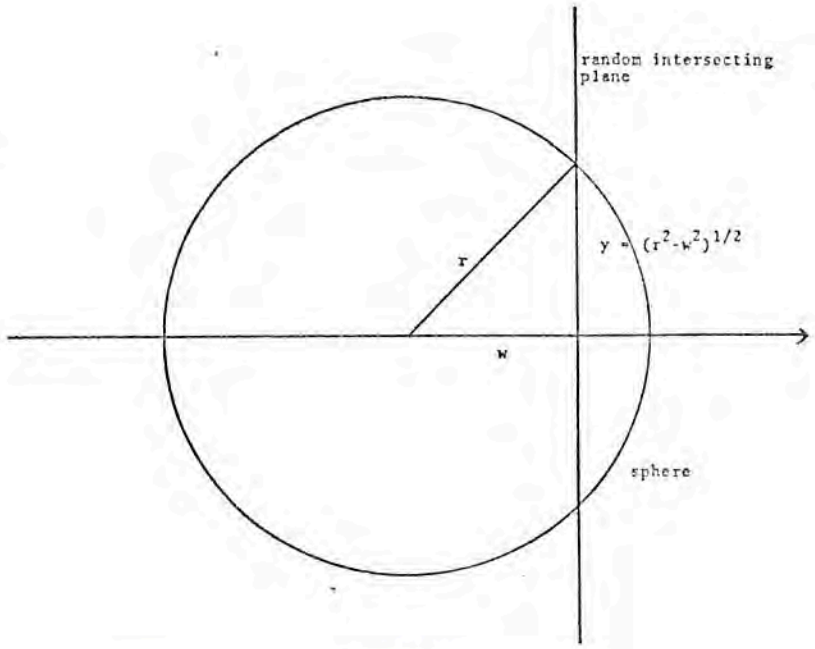


FIGURE 2
THE RANDOM INTERSECTION OF A PLANE WITH A SPHERE

If n circles are measured on random plane sections of the material, then an unbiased estimate of m_j , $j \geq -1$, is given by $\bar{m}_j = C_j \sum_{i=1}^{n} y_i^j/n$. An estimate of μ_j is $\bar{\mu}_j = \bar{m}_{j-1}/\bar{m}_{-1}$. Although it is possible to obtain simple expressions for the variance of $\bar{m}_j$, $j > 0$, the variance of $\bar{m}_{-1}$ does not exist. However, $\bar{\mu}_j$ is certainly a consistent estimator of μ_j .

The following equation defines Φ in terms of F:

$$\Phi(y) = \int_0^{\frac{1}{2}\pi} \cos\theta \, F(y/\cos\theta) \, d\theta. \tag{3.3a}$$

To verify this, note that the right hand side of (3.3a) is a distribution function and

$$\eta_j = j \int_0^{\infty} y^{j-1}[1 - \Phi(y)] \, dy = j \int_0^{\infty} y^{j-1} \int_0^{\frac{1}{2}\pi} [1 - F(y/\cos\theta)] \cos\theta \, d\theta \, dy$$

$$= \int_0^{\frac{1}{2}\pi} \cos\theta^{j+1} \, d\theta \int_0^{\infty} jx^{j-1}[1 - F(x)] \, dx = C_j^{-1} m_j .$$

The interchange of the order of integration is justified by Tonelli's theorem.

Let $y = x \cos\theta$; then (3.3a) can be written as

$$\frac{\Phi(y)}{y^2} = \int_y^{\infty} \frac{F(x)}{x^2(x^2 - y^2)^{\frac{1}{2}}} \, dx.$$

Set $y^2 = u$ and $x^2 = v$ and write $\Phi(\sqrt{u})/u = \alpha(u)$ and $F(\sqrt{v})/(2v^{3/2}) = \beta(v)$; then the equation becomes

$$\alpha(u) = \int_u^\infty \frac{\beta(v)}{(v-u)^{\frac{1}{2}}}\,dv.$$

Now

$$\int_w^\infty \frac{\alpha(u)}{(u-w)^{\frac{1}{2}}}\,du = \int_w^\infty \left\{\int_u^\infty \frac{\beta(v)}{(v-u)^{\frac{1}{2}}}\,dv\right\}\frac{du}{(u-w)^{\frac{1}{2}}}$$

and it can be verified that the integral on the left hand side exists for $w > 0$. Again, by Tonelli's theorem the order of integration can be reversed on the right hand side to give

$$\int_w^\infty \beta(v)\left\{\int_w^v (v-u)^{-\frac{1}{2}}(u-w)^{-\frac{1}{2}}\,du\right\}dv.$$

The inner integral reduces to $\int_0^1 t^{-\frac{1}{2}}(1-t)^{-\frac{1}{2}}\,dt = \pi$ with the substitution $u = w + (v-w)t$. Thus

$$\int_w^\infty \beta(v)\,dv = \pi^{-1}\int_w^\infty \frac{\alpha(u)}{(u-w)^{\frac{1}{2}}}\,du$$

and

$$\int_w^\infty \frac{F(y)}{y^2}\,dy = 2\pi^{-1}\int_w^\infty \frac{\Phi(y)}{y(y^2-w^2)^{\frac{1}{2}}}\,dy, \qquad w > 0.$$

Hence for $x > 0$

$$F(x) = -2\pi^{-1}x^2 D_x \int_x^\infty \frac{\Phi(y)}{y(y^2-x^2)^{\frac{1}{2}}}\,dy, \text{ i.e.,} \tag{3.4}$$

where $D_x k(x)$ is the derivative of k. Since $F(0) = 0$, (3.4) uniquely determines F, and G is recovered from the relationship

$$G(r) = r_0\int_0^r x^{-1}\,dF(x), \qquad r_0^{-1} = \int_0^\infty r^{-1}\,dF(r).$$

When the frequency functions ϕ, f, and g exist (3.3a) becomes

$$\begin{aligned}\phi(y) &= \int_0^{\frac{1}{2}\pi} f(y/\cos\theta)\,d\theta \\ &= yr_0^{-1}\int_y^\infty \frac{g(r)\,dr}{(r^2-y^2)^{\frac{1}{2}}}\end{aligned} \tag{3.3b}$$

which is Wicksell's fundamental integral equation. By the same manipulation as above the solution to (3.3b) can be written as

$$G(r) = 1 - \int_r^\infty \phi(y)(y^2-r^2)^{-\frac{1}{2}}\,dy \Big/ \int_0^\infty y^{-1}\phi(y)\,dy. \tag{3.5}$$

From now on we assume that the required frequency functions exist and that the necessary calculations can be justified by, if necessary, imposing mild restrictions on the problem such as the assumption that the radii are bounded. Although, more mathematical satisfaction is achieved by working

with the distribution functions, the frequency function form is the one which will be most acceptable in practice.

It is often convenient to parametrise the problem from the start although the estimated moments could be used to fit some general curve, one of the Pearson class, for example. For illustration suppose that, for a particular problem, it is reasonable to assume that $g(r)$ is of the form

$$g(r) = [\Gamma(\gamma)]^{-1}e^{-r}r^{\gamma-1}, \qquad \gamma > 0;$$

then g is estimated once an estimate of γ is available. In this case

$$f(r) = [\Gamma(\gamma + 1)]^{-1}e^{-r}r^{\gamma},$$

$m_1 = \gamma + 1$, $\hat{\gamma} = \bar{m}_1 - 1$ and $\text{var}\,\hat{\gamma} = C_1^2(\eta_2 - \eta_1^2)/n$.

Equation (3.3b) can be used to check the adequacy of the parametrisation. Once $f(r)$ is estimated, an estimate of $\phi(y)$ can be calculated from (3.3b) and compared with the observed frequency function. If the agreement is poor a more general parametric model can be tried. For example, if the one-parameter gamma distribution used above produces unsatisfactory results, the two-parameter gamma distribution can perhaps be fitted with a happier outcome.

There remains the problem of estimating λ. The average number of spheres which are cut by a random plane per unit area of the plane is $2\lambda r_0$. If A is the total area of the intersection and n is the number of spheres cut by the plane, then the equation $\hat{\lambda} = n/A2\bar{r}_0$ gives an estimate of λ, where $\bar{r}_0$ is an estimate of r_0. If $\bar{r}_0$ is calculated from the harmonic mean of the observed radii of circles in the plane, i.e. $\bar{r}_0^{-1} = \bar{m}_{-1} = 2\bar{\eta}_{-1}/\pi$, the behaviour of $\hat{\lambda}$ may not be stable since $\bar{r}_0^{-1}$ has infinite variance. More satisfaction can be achieved from parametrised versions. For instance, in the case considered above where $g(r)$ has a one-parameter gamma distribution, $m_{-1} = \gamma^{-1} = r_0^{-1}$, and hence a suitable estimator is $\bar{r}_0^{-1} = (\bar{m}_1 - 1)^{-1}$.

Some idea of the variance of $\hat{\lambda}$ can be obtained by assuming that n is distributed approximately as a Poisson variate with parameter $2A\lambda r_0$, and that n and $\bar{m}_1$ are uncorrelated. Under these assumptions, since

$$\text{var}\,(\bar{m}_1) = \text{var}\,(\hat{\gamma}) = C_1^2(\eta_2 - \eta_1^2)/n,$$

the large sample variance of $\hat{\lambda}$ can be calculated by the delta method. In cases where λ and r_0 are small the approximation should be satisfactory and if A is large $\hat{\lambda}$ should be stable. The routine algebra will not be carried out here.

Of course, a direct estimate of λ can be obtained by serial section. The actual number, n, of sphere centres in the volume of material sampled, v, can be counted and n should be approximately Poisson-distributed with parameter λv.

(b) *Truncation*

It can happen that circles with radii greater than R, say, are ill defined or difficult to measure. In these cases it is desirable to have a truncation pro-

cedure to fall back on. It is found that if the problem is parametrised, $g(r)$ can be estimated from observations made on plane sections using only circles with $y \leq R$.

Consider the truncated frequency function $\tau(y) = \phi(y)/\Phi(R)$, $\Phi(R) = \int_0^R \phi(x)\,dx$. Then, writing

$$\eta_i(R) = \int_0^R y^i \tau(y)\,dy,$$

the following relationship can be verified

$$\begin{aligned}\Phi(R)\eta_i(R) &= \int_0^R y^i \phi(y)\,dy = \int_0^R y^{i+1} \int_y^\infty f(r) r^{-1}(r^2 - y^2)^{-\frac{1}{2}}\,dr\,dy \\ &= k_i(0) m_i - \int_R^\infty r^i f(r) k_i(R/r)\,dr,\end{aligned} \tag{3.6}$$

where

$$k_i(\theta) = \int_\theta^1 v^{i+1}(1 - v^2)^{-\frac{1}{2}}\,dv, \qquad k_i(0) = C_i^{-1}.$$

As an example of the use of (3.6) substitute the gamma form of $f(r)$ used above and set $j = 0$. Thus, writing Φ as a function of γ and R, we have

$$\begin{aligned}\Phi(R, \gamma) &= 1 - \int_R^\infty [\Gamma(\gamma + 1)]^{-1} e^{-r} r^\gamma (1 - R^2/r^2)\,dr \\ &= \Gamma(R, \gamma) + R^2 \Gamma(\gamma - 1)[\Gamma(\gamma + 1)]^{-1}[1 - \Gamma(R, \gamma - 2)],\end{aligned}$$

where

$$\Gamma(x, \theta) = \int_\theta^x [\Gamma(\theta)]^{-1} e^{-y} y^{\theta - 1}\,dy.$$

If from a sample of plane sections it is found that, of the total number of circles, n_0 have radii less than R, then

$$\Phi(R, \hat{\gamma}) = n_0/n \tag{3.7}$$

defines an estimator for γ. As $\Phi(R, \gamma)$, a decreasing function of γ, is readily plotted as a function of γ from existing tables of the incomplete gamma function, a solution to (3.7) can be found with relative ease. The large sample variance of $\hat{\gamma}$ can be calculated by standard methods although its form would not be pleasing since it involves $\partial\Phi/\partial\gamma$.

(c) *The Holmes effect*

So far the analyses have been entirely in terms of measurements made on the face of a plane intersecting the conglomerate C. If thin sections are used and the spherical bodies B are opaque, the Holmes bias mentioned in the Introduction can be expected.

Consider a sphere of radius r and a random thin slice of thickness $2k$ (Figure 3). Define the sets A_i as $A_1 = [-r - k, -k)$, $A_2 = [-k, k)$, $A_3 =$

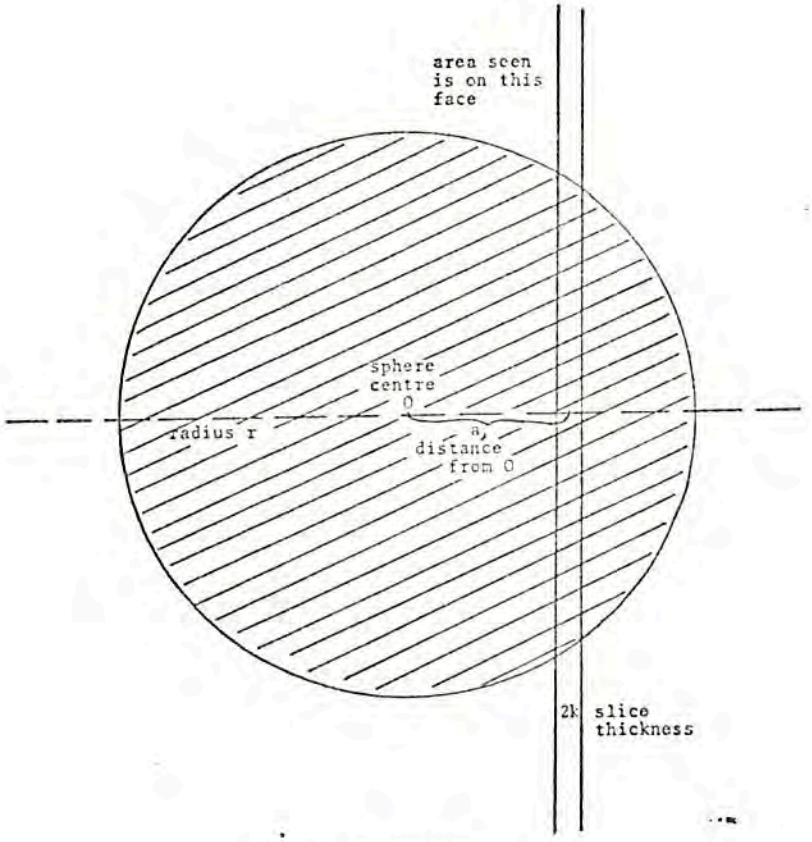


FIGURE 3
ILLUSTRATION OF THE HOLMES EFFECT

$[k, r + k]$, and if a is the position of the centre of the slice then, under the assumption that the sphere is sliced at random,

$$\Pr\{a \in A_1\} = \Pr\{a \in A_3\} = r/2(r + k), \qquad \Pr\{a \in A_2\} = k/(r + k).$$

Further

$$E\{y^t \mid r, a \in A_1\} = E\{y^t \mid r, a \in A_3\} = C_i^{-1} r^t, \qquad E\{y^t \mid r, a \in A_2\} = r^t,$$

and hence

$$E\{y^t \mid r\} = C_i^{-1} r^{t+1}/(r + k) + r^t k/(r + k). \tag{3.8}$$

Let $f_k(r) = \dfrac{(r+k)\,g(r)}{k+\mu_1}$, then $\eta_j = E[y^j] = \int_0^\infty E[y^j \mid r]\, f_k(r)\, dr$

$$= \frac{C_j^{-1}\,\mu_{j+1}}{k+\mu_1} + \frac{k\,\mu_j}{k+\mu_1} \tag{3.9}$$

$$\phi(y) = \left[y \int_0^{\frac{1}{2}\pi} g(y/\cos\theta)[\cos\theta]^{-1}\, d\theta + kg(y) \right] \Big/ (\lambda + \mu_1) \tag{3.10}$$
$$= y \int_y^\infty \beta(x)(x^2 - y^2)^{-\frac{1}{2}}\, dx + k\beta(y), \quad \beta = g/(\lambda + \mu_1).$$

Equation (3.10) transforms to

$$\tilde{g}(v) = \int_v^\infty h(u)\sqrt{v}(u - v)^{-\frac{1}{2}}\, du + kh(v), \tag{3.11}$$

where $\tilde{g}(v) = \phi(\sqrt{v})/2\sqrt{v}$ and $h(u) = \beta(\sqrt{u})/2\sqrt{u}$, by setting $x^2 = u$ and $y^2 = v$.

The solution to (3.11) will now be briefly outlined. Let $K(v, u) = \sqrt{v}(u - v)^{-\frac{1}{2}}$ and write (3.11) as

$$\tilde{g} = Kh + kh, \qquad \text{where} \quad Kh = \int_v^\infty K(v, u)h(u)\, du. \tag{3.12}$$

Now, the equation $Kh_0 = \tilde{g}$ can be solved. Explicitly

$$h_0(u) = -D_v\pi^{-1} \int_u^\infty \tilde{g}(v)[v(v - u)]^{-\frac{1}{2}}\, dv$$

$$= -\pi^{-1} \int_0^1 \tilde{g}'(u/t)t^{-2}(1 - t)^{-\frac{1}{2}}\, dt$$

after the substitution $u = tv$. For notation, the inversion will be written as $h_0 = K^{-1}\tilde{g}$.

A second approximation to h is $h_1 = K^{-1}\tilde{g} - kK^{-1}(K^{-1}\tilde{g})$ and h_1 satisfies (3.12) to $O(k^2)$ which would be satisfactory for most practical purposes. It has been assumed, of course, that $\tilde{g}$ is such that the required mathematical operations can be justified.

A more rigorous discussion of (3.12) is possible. By composition (3.12) can be transformed into the equivalent equation

$$(K_2 - k^2)h = (K - k)\tilde{g}, \qquad K_2h = K(Kh). \tag{3.13}$$

It can be verified that because of the boundedness of r, K_2 is a bounded and continuous kernel. As such the method of iteration appropriate to Volterra integral equations of the second kind is applicable and a unique solution is guaranteed; see, e.g. Tricomi [1957].

4. ELLIPSOIDS OF REVOLUTION

The next simplest case to consider is the one where the bodies are ellipsoids of revolution with the z-axis in the North-South position. The relevant equation is

$$(x^2 + y^2)/\alpha^2 + z^2/\gamma^2 = 1. \tag{4.1}$$

Before proceeding we display the notation. Let

λ = the average number of ellipsoid centres per unit volume of conglomerate C,

$g(\alpha, \gamma)$ = the frequency function of spheroid diameters, α, and major axis, γ, of material B with power moments μ_{ij} ,

$f(\alpha, \gamma)$ = the frequency function of diameters and major axes of spheroids which intersect a random plane through C, with power moments m_{ij} ,

$\phi(y, z)$ = the frequency function of minor, y, and major, z, axes of ellipses formed on a random plane through C by intersection with the spheroids, with power moments η_{ij} .

The average number of ellipsoids with diameter and major axis in the set $[\alpha, \alpha + d\alpha] \times [\gamma, \gamma + d\gamma]$ which are intersected by a random plane parallel to the z-axis is $2\lambda\alpha g(\alpha, \gamma)\, d\alpha\, d\gamma$ per unit area of the plane. Hence

$$f(\alpha, \gamma) = \alpha g(\alpha, \gamma)/\alpha_0 , \qquad \alpha_0 = \mu_{1,0} = \int_0^\infty \int_0^\infty \alpha g(\alpha, \gamma)\, d\alpha\, d\gamma. \tag{4.2}$$

If y and z are the minor and major axes of the ellipse formed by the plane cutting the ellipsoid at w, then

$$y = \alpha(1 - w^2/\alpha^2)^{\frac{1}{2}}, \qquad z = \gamma(1 - w^2/\alpha^2)^{\frac{1}{2}},$$

and if the cut has density $dw/2\alpha$,

$$\begin{aligned} E\{y^i z^j \mid \alpha, \gamma\} &= \alpha^i \gamma^j \int_{-\alpha}^{\alpha} (1 - w^2/\alpha^2)^{\frac{1}{2}(i+j)}\, dw/2\alpha \\ &= \alpha^i \gamma^j C_{i+j}^{-1} . \end{aligned}$$

Clearly, $C_{i+j}\eta_{ij} = E\{C_{i+j} y^i z^j\} = m_{ij} = \mu_{i+1,j}/\alpha_0$, and the estimation of m_{ij} and μ_{ij} proceeds in the same way as for the spherical case.

The best approach in this more complicated situation is to establish an integral equation, similar to (3.3b), relating g and ϕ. From the above relationship between η_{ij} and m_{ij}

$$\phi(y, z) = \int_0^{\frac{1}{2}\pi} f(y/\cos\theta, z/\cos\theta)(\cos\theta)^{-1}\, d\theta, \tag{4.3}$$

which can be written as

$$\phi(s, st) = \int_0^{\frac{1}{2}\pi} f(s/\cos\theta, st/\cos\theta)(\cos\theta)^{-1}\, d\theta$$

by setting $y = s$, $z = ts$. Now let $s = v\cos\theta$, then

$$\begin{aligned} \phi(s, st) &= \int_s^\infty f(v, vt)(v^2 - s^2)^{-\frac{1}{2}}\, dv \\ &= \alpha_0^{-1} \int_s^\infty v g(v, vt)(v^2 - s^2)^{-\frac{1}{2}}\, dv. \end{aligned} \tag{4.4}$$

By the methods of section 3

$$\alpha_0^{-1}\int_v^\infty w g(w, wl)\,dw = 2\pi^{-1}\int_v^\infty w\phi(w, wl)(w^2 - y^2)^{-\frac{1}{2}}\,dw$$

and

$$g(y, z) = \frac{-D_v 2\pi^{-1}\int_v^\infty w\phi(w, wl)(w^2 - y^2)^{-\frac{1}{2}}\,dw\Big|_{l=v/y}}{\int_0^\infty\int_0^\infty y^{-1}\phi(y, z)\,dy\,dz}. \tag{4.5}$$

The problem could be tackled from a parametric viewpoint by assigning a suitable family of distributions to $g(\alpha, \gamma)$ and estimating the parameters by the method of moments as for the spherical case. Since no new principle is involved, this approach will not be developed further.

The volume, V, of an ellipsoid of revolution is $V = 4\pi\alpha^2\gamma/3$ and hence $E\{V^i\} = \nu_i = 4\pi\mu_{2i,i}/3 = 4\pi\alpha_0 m_{2i-1,i}/3$. From a sample of size n of ellipses measured on the face of random planes intersecting the material it is possible to estimate ν_i by

$$\bar{\nu}_i = 2\pi^2 C_{3i-1}\sum_{k=1}^{n} y_k^{2i-1}z_k^i\left[3\sum_{k=1}^{n} y_k^{-1}\right]^{-1}. \tag{4.6}$$

Unfortunately, $\bar{\nu}_i$ is only asymptotically unbiased and, for reasons discussed in section 3, it has infinite variance. However, useful information can be obtained with little effort since if $\bar{\nu}_1$ is known and an estimate of V_B, $\bar{a}$ say, (section 2), is also available, then λ can be estimated by $\hat{\lambda} = \bar{a}\bar{\nu}_1^{-1}$; see Example 2 below.

5. GENERAL ELLIPSOIDS

In this section the basic assumption that the bodies are ellipsoids oriented in the North-South position is maintained. However, the ellipsoids are of the general form

$$x^2/\alpha^2 + y^2/\beta^2 + z^2/\gamma^2 = 1, \tag{5.1}$$

and the x and y axes are subject to a random rotation, θ, about the z-axis according to the density $d\theta/2\pi$.

If $g(\alpha, \beta, \gamma)$ is the joint frequency function for α, β, and γ, the principal axes of the bodies in the conglomerate C, then in order to proceed to a reasonably tractable solution it is assumed that g factorises as $g(\alpha, \beta, \gamma) = g_1(\alpha, \beta)g_2(\gamma)$. This assumption leads to great simplification when it is used in conjunction with transverse and vertical sectioning of C. The solution will only be outlined since it can be reduced to the situations discussed in sections 3 and 4 above.

(a) *Transverse sectioning*

It is assumed that both transverse and vertical sectioning of C is possible. The notation and assumptions used in this and the following section are con-

sistent with those of previous sections and need no further elaboration. In analogy with (3.1) and (4.2)

$$f(\alpha, \beta, \gamma) = \gamma g_2(\gamma) g_1(\alpha, \beta)/\gamma_0 , \qquad \gamma_0 = \int_0^\infty \gamma g_2(\gamma)\, d\gamma. \tag{5.2}$$

Suppose the plane intersects a particular ellipsoid with axes α, β, and γ at $z = w$, then the density of w is $dw/2\gamma$. If x and y are the major and minor axes of the associated ellipse in the plane,

$$E\{x^i y^j \mid \alpha, \beta, \gamma\} = \int_{-\gamma}^{\gamma} \alpha^i \beta^j (1 - w^2/\gamma^2)^{\frac{1}{2}(i+j)}\, dw/2\gamma$$
$$= C_{i+j}^{-1} \alpha^i \beta^j .$$

The methods of section 4 can now be used in obtaining $g_1(\alpha, \beta)$.

(b) *Vertical sectioning*

Under a random rotation θ of the x, y axes about the central, vertical axis z the equation of the ellipsoid becomes

$$(x \cos\theta + y \sin\theta)^2/\alpha^2 + (x \sin\theta - y \cos\theta)^2/\beta^2 + z^2/\gamma^2 = 1. \tag{5.3}$$

From (5.3) it is found that

$$z = \gamma(1 - w^2/h^2), \qquad h^2(\alpha, \beta, \gamma) = \alpha^2 \cos^2\theta + \beta^2 \sin^2\theta,$$

on setting $x = w$, the position of the vertical cut. But it can be shown that the diameter of the ellipsoid, i.e. the distance between the two tangent planes parallel to $x = w$ is $2h$ and hence, given α, β, γ, and θ

$$E\{z^i \mid \alpha, \beta, \gamma, \theta\} = \gamma^i C_i^{-1}$$

since the cut has density $dw/2h$. It is easily verified that

$$f(\alpha, \beta, \gamma) = g_1(\alpha, \beta) g_2(\gamma) \bar{h}(\alpha, \beta)/h_0 ,$$

where

$$\bar{h}(\alpha, \beta) = \int_{-\pi}^{\pi} h(\alpha, \beta, \theta)\, d\theta/2\pi$$

$$h_0 = \int_0^\infty \int_0^\infty \bar{h}(\alpha, \beta) g_1(\alpha, \beta)\, d\alpha\, d\beta.$$

The function $g_2(\gamma)$ can be calculated according to the methods of section 3 by noticing that

$$g_2(\gamma) = \int_0^\infty \int_0^\infty f(\alpha, \beta, \gamma)\, d\alpha\, d\beta = f_2(\gamma), \quad \text{say.}$$

6. NUMERICAL EXAMPLES

To illustrate the use of some of the formulae, two numerical examples are given.

Example 1

A large block of Swiss cheese was thinly sliced and one hundred slices were drawn at random with replacement. The maximum and minimum diameters of each hole appearing in the slices were obtained and y, the geometric mean, was calculated. The resulting empirical frequency function of y is given in Table 1. From the latter figures it was required to estimate the distribution function, $G(r)$, of the radii of the spherical airspaces in the cheese.

Formula (3.5) was used to obtain a numerical estimate of $G(r)$ and this is also given in Table 1. Let y_i, $i = 1, 2, \cdots, 18$ represent the class boundaries for Table 1; then $G(r)$ was calculated from the empirical frequency polygon by the formulae

$$\int_0^\infty \phi(y)y^{-1}\,dy \simeq \sum_{i=1}^{18} \frac{(\phi_i y_{i+1} - \phi_{i+1}y_i)}{(y_{i+1} - y_i)} \log\left(\frac{y_{i+1}}{y_i}\right)$$

$$\int_{y_i}^\infty \frac{\phi(y)}{(y^2 - y_i^2)^{\frac{1}{2}}}\,dy \simeq \sum_{i=1}^{18} \left\{\frac{(\phi_{i+1} - \phi_i)}{(y_{i+1} - y_i)}\left[(y_{i+1}^2 - y_i^2)^{\frac{1}{2}} - (y_i^2 - y_i^2)^{\frac{1}{2}}\right]\right.$$
$$\left. + \left(\frac{\phi_i y_{i+1} - \phi_{i+1}y_i}{y_{i+1} - y_i}\right) \log\left[\frac{y_{i+1} + (y_{i+1}^2 - y_i^2)^{\frac{1}{2}}}{y_i + (y_i^2 - y_i^2)^{\frac{1}{2}}}\right]\right\}$$

Some trouble is experienced in the neighbourhood of $r = 0$. The estimate of $G(r)$ becomes slightly negative and this is due to the fact that not every

TABLE 1

EMPIRICAL FREQUENCY FUNCTION AND ESTIMATE OF $G(r)$ FOR AIRSPACE IN SWISS CHEESE

$y_i + \frac{1}{2}$(cms)	ϕ_i	Est $G(r)$
.125	.005	.005
.225	.031	.014
.325	.051	.025
.425	.054	.035
.525	.074	.042
.625	.147	.055
.725	.133	.079
.825	.216	.209
.925	.147	.562
1.025	.060	.799
1.125	.041	.884
1.225	.021	.946
1.325	.006	.976
1.425	.006	.982
1.525	.003	.989
1.625	.003	.992
1.725	.003	.996

$y_i + \frac{1}{2}$ = class centres, ϕ_i = estimate of frequency in class i.

TABLE 2

GLAND VOLUME ANALYSIS IN SKIN TISSUE OF SHEEP

Sheep no.	P	$V \times 10^4$	N
1	.055	13.1	42
2	.069	16.6	42
3	.074	15.2	49
4	.069	30.8	22
5	.066	23.9	28
6	.065	22.8	29
7	.059	38.9	15
8	.036	19.1	19
9	.043	33.5	13
10	.040	22.9	18
11	.061	42.6	14
12	.051	21.8	23
13	.044	22.6	20
14	.029	22.3	13
15	.043	16.6	26
16	.054	28.4	19

P = proportion of gland/(mm)^3 of skin tissue,
V = average gland volume (formula 4.6),
N = approximate number of glands/(mm)^3 skin tissue.

frequency function when used in (3.5) leads to a distribution function. The true ϕ does, but in this case the estimate does not. To overcome this, the last positive value of G estimated by the formula was used to obtain a linear interpolation passing through $G(0) = 0$. Of course, if the problem is parametrised as discussed above, these problems are avoided.

Actually, μ_1 calculated from $G(r)$ is 0.84. When estimated from the formula $(2/\pi)\eta_{-1} = \mu_1^{-1}$ a figure of 1.01 is obtained, giving reasonable agreement.

Example 2

During the study of wool growth described in the Introduction mid-side skin samples of 16 sheep were taken. From photographs of the skin sections, tracings of the glands were made and cut out. These cut-outs were roughly elliptical and the major and minor axes were measured as the maximum and minimum diameters of the particular tracing. The total weight of the tracings was also obtained.

From the above information the proportion of gland occupying a (mm)^2 of the vertical skin slice was calculated and is reported in Table 2. From formula (4.6) the average volume of the glands was estimated, leading ultimately to an estimate of the number of glands per (mm)^3 of skin tissue.

Information of the type presented in Table 2 is being used to assist in the interpretation of patterns of wool growth. Fuller details will be reported elsewhere.

ACKNOWLEDGEMENTS

I wish to thank Mr. Ralph Chapman for bringing some of the problems treated in this paper to my attention. I am also indebted to Mr. M. O'Callaghan for his help with the numerical work of Example 1, and to the referees for valuable criticisms and suggestions.

L'ESTIMATION DE LA DISTRIBUTION DE CORPS SPHERIQUES OU ELLIPTIQUES A L'INTERIEUR DE CONGLOMERATS, A PARTIR DE SECTIONS PLANES

RESUME

Les problèmes de l'estimation de la distribution d'une certaine matière B à l'intérieur d'un conglomérat sont discutés dans ce travail. Le résultat général et bien connu suivant lequel la proportion de B dans C peut être estimée sans biais à partir d'une section plane aléatoire de C est présenté brièvement. Certaines hypothèses concernant la forme des dépôts de B permettent d'obtenir plus d'informations de l'analyse à deux dimensions.

Initialement les fragments de B sont choisis de forme sphérique et on suggère certaines modifications et extensions des résultats classiques de Wicksell dans ce domaine. En particulier le biais de Holmes dans l'analyse des sections planes minces est considéré très généralement pour les corps sphériques.

On considère par la suite des dépôts de forme ellipsoïdale convenablement orientés. Des résultats explicites sont obtenus facilement par analogie avec le cas sphérique.

Quelques formules sont illustrées par deux exemples numériques courts.

REFERENCES

Chayes, F. [1956]. *Petrographic Modal Analysis: an Elementary Statistical Appraisal.* Wiley, New York.

Kendall, M. G. and Moran, P. A. P. [1963]. *Geometrical Probability.* Griffin, London.

Tricomi, F. G. [1957]. *Integral Equations.* Interscience Publishers Inc., New York.

Wicksell, S. D. [1925]. The corpuscle problem. Part I. *Biometrika 17*, 84–99.

Wicksell, S. D. [1926]. The corpuscle problem. Part II. *Biometrika 18*, 151–72.

Received May 1968, *Revised July* 1969

Sampling Methods for Estimating Average Faecal Egg-Count in Animal Populations

By G. M. Tallis and D. Culpin

Division of Mathematical Statistics Technical Paper No. 27

Commonwealth Scientific and Industrial
Research Organization, Australia
Melbourne 1969

SAMPLING METHODS FOR ESTIMATING AVERAGE FAECAL EGG-COUNT IN ANIMAL POPULATIONS

By G. M. Tallis* and D. Culpin*

Summary

This paper presents a sampling procedure for estimating, to a nominated precision, the average faecal worm egg-count of a group of animals. Formulae for calculating the correct number of animals to sample from a fixed population are developed. The treatment is general and allows both within and between animal sampling variability to be taken into account. Specific reference is made to sheep.

I. Introduction

During the study of parasitic diseases of sheep the veterinarian is often faced with the task of periodically estimating the level of infection in various experimental flocks. If the prime interest is in some internal egg-laying worm, the standard practice for estimating the level of infection in a sheep is to collect a faecal sample from the sheep and, from this, to subsequently make an estimate of the egg-count. If the egg-count is assumed to be associated with the level of infection within the sheep, then this level may be measured by such egg-counts. The average egg-count for the flock can then be used as a measure of the general level of infection.

For small flocks it is quite feasible to obtain a faecal sample from each sheep. But for larger flocks it is more economical to estimate the average egg-count by the average egg-count of a random sample drawn from the flock. This procedure naturally entails a concomitant increase in the errors of estimation, but by choice of sample size it is usually possible to ensure that the mean flock egg-count is estimated with a desired precision. It is the main purpose of this paper to develop methods for choosing the sample size.

This problem was originally suggested to us by Dr H. Gordon of the McMaster Laboratory, Division of Animal Health, CSIRO, and it therefore has particular relevance to the egg-counting technique developed by him (Gordon and Whitlock 1939).† However, the theory is of a general nature and should be applicable to other situations and to animals other than sheep.

Those who wish to gain an idea of the methods without an understanding of the assumptions underlying them may pass directly to Section VI, where a summary of the methods and a numerical example are presented.

Sections II–V should be comprehensible to anyone having some familiarity with elementary statistics since the more involved arguments are relegated to the Appendices.

* Division of Mathematical Statistics, CSIRO, 60 King Street, Newtown, N.S.W. 2042.

† Gordon, H. McL., and Whitlock, H. V. (1939).—A new technique for counting nematode eggs in sheep faeces. *J. Coun. scient. ind. Res. Aust.* 12, 50–2.

II. Terminology, General Assumptions, and Relations

The flock size will be denoted by N and the sheep will be numbered $1, 2, \ldots, N$. It will be assumed that a random sample of n sheep has been selected for faecal collection. The sheep in the sample will be numbered $1, 2, \ldots, n$, this numbering being distinct from the numbering of the flock. (When a sheep is referred to by number, the context should indicate which numbering is relevant.)

The average number of eggs per gram of dry faeces from sheep number s will be denoted by λ_s ($s = 1, 2, \ldots, N$). The mean and variance of the set $\lambda_1, \lambda_2, \ldots, \lambda_N$ will be denoted by λ and $\sigma^2_{(\lambda)}$ respectively, and these are defined by

$$\lambda = \sum_{s=1}^{N} \lambda_s / N$$

and

$$\sigma^2_{(\lambda)} = \sum_{s=1}^{N} (\lambda_s - \lambda)^2 / (N-1).$$

It is our purpose to estimate λ. The quantity $\sigma^2_{(\lambda)}$ could also be called the between sheep variance of egg-count.

For the moment it will be assumed that there is a method for obtaining for any sheep s an unbiased estimate $\hat{\lambda}_s$ of λ_s, and that the estimates from different sheep are independent. Let $\sigma^2_s = \mathrm{Var}(\hat{\lambda}_s)$. Then it will be further assumed that there is an unbiased estimate $\hat{\sigma}^2_s$ of σ^2_s. Later on (see Section IV) a way of obtaining these estimates will be considered. Finally, let

$$\sigma^2 = \sum_{s=1}^{N} \sigma^2_s / N.$$

Now from the sample of n sheep we have n estimates $\hat{\lambda}_1, \hat{\lambda}_2, \ldots, \hat{\lambda}_n$ of egg-counts and n estimates $\hat{\sigma}^2_1, \hat{\sigma}^2_2, \ldots, \hat{\sigma}^2_n$ of variances. A natural estimator to choose for the flock average egg-count λ is the sample average

$$\hat{\lambda} = \sum_{i=1}^{n} \hat{\lambda}_i / n. \tag{1}$$

In fact, it can be shown that $\hat{\lambda}$ is an unbiased estimate of λ. Also

$$\mathrm{Var}(\hat{\lambda}) = \frac{\sigma^2}{n} + \left(\frac{1}{n} - \frac{1}{N}\right)\sigma^2_{(\lambda)}. \tag{2}$$

To prove results of this type, the fairly standard methods of the statistics of finite sampling can be used. As an illustration of the methods, $E(\hat{\lambda})$ and $\mathrm{Var}(\hat{\lambda})$ are evaluated in Appendix I. Note that two levels of statistical distribution are involved; (1) the sampling distribution of the random sample of sheep from the flock, and (2) the distribution of the estimate of egg-count, $\hat{\lambda}_s$, for a particular sheep s.

For later reference we shall state here two further results. Let

$$\hat{\sigma}^2 = \sum_{i=1}^{n} \hat{\sigma}^2_i / n \tag{3}$$

and

$$\hat{\sigma}^2_{(\lambda)} = \sum_{i=1}^{n} (\hat{\lambda}_i - \hat{\lambda})^2 / (n-1) - \hat{\sigma}^2. \tag{4}$$

Then $\hat{\sigma}^2$ and $\hat{\sigma}^2_{(\lambda)}$ are unbiased estimates of σ^2 and $\sigma^2_{(\lambda)}$ respectively. (These results are not proved in this paper.)

III. Choice of Sample Size

We now turn to the question of choosing an appropriate sample size. It is required that λ be estimated by $\hat{\lambda}$ with a certain precision. In order to state this requirement more specifically we shall stipulate that we wish $\hat{\lambda}$ to satisfy the condition

$$\text{Prob}\{\,|\,\hat{\lambda}-\lambda\,|\,\leqslant p\lambda\} = 1-\beta. \tag{5}$$

This condition could be expressed alternatively as follows: With probability $1-\beta$ the true value of λ lies in the range $\hat{\lambda}/(1+p)$ to $\hat{\lambda}/(1-p)$. For example, if we choose β equal to 0·05 and p equal to 0·1, the range is $0{\cdot}91\hat{\lambda}$ to $1{\cdot}11\hat{\lambda}$, and λ lies in this interval with a probability of 0·95. The smaller p is, the smaller is the range; the smaller β is, the larger is the probability that λ lies in the range.

Using the Central Limit Theorem it can be assumed that $\hat{\lambda}$ is approximately normally distributed with mean λ and variance $\text{Var}(\hat{\lambda})$. We shall denote the $100(1-\beta)$ two-tailed percentage point of the normal distribution by n_β. We can now write

$$\text{Prob}\left\{\frac{|\,\hat{\lambda}-\lambda\,|}{\sqrt{\{\text{Var}(\hat{\lambda})\}}} \leqslant n_\beta\right\} \simeq 1-\beta.$$

Comparing this with (5), it is seen that

$$p\lambda \simeq n_\beta\sqrt{\{\text{Var}(\hat{\lambda})\}}. \tag{6}$$

From equations (2) and (6), we find that

$$n \simeq N\left\{\frac{1+\sigma^2/\sigma^2_{(\lambda)}}{1+Np^2/(n^2_\beta c^2)}\right\}, \tag{7}$$

where $c = \sigma_{(\lambda)}/\lambda$ and is subsequently referred to as the coefficient of variation.

We have thus found an expression for the sample size n. Of course, λ, σ^2, and $\sigma^2_{(\lambda)}$ are unknown, and hence suitable estimates must first be obtained in order to use equation (7). This problem will be discussed in Section V.

IV. Estimating the Egg-count of a Sheep

We present a method by which unbiased estimators $\hat{\lambda}_s$ and $\hat{\sigma}^2_s$ of λ_s and σ^2_s respectively may be found.

We must make some further assumptions. Suppose that the number of eggs in a randomly collected gram of dry faeces from sheep s has a Poisson distribution with parameter (and therefore mean) λ_s. If, from sheep s, g_s grams of faeces is sampled and X_s is the number of eggs contained in this sample, then, given g_s, X_s has a Poisson distribution with parameter $\lambda_s g_s$. In practice, the situation will be that approximately the same amount of faeces will be sampled from each sheep. This average or "aimed at" amount, which we shall denote by g, will be known, but the particular amounts sampled from individual sheep will not. We thus assume that g_s has some distribution which does not depend on s, and which has mean g and variance $\sigma^2_{(g)}$. Now

the distribution of X_s will be a mixture of two distributions: the distribution of X_s given g_s, which is Poisson, and the distribution of g_s.

Let

$$\hat{\lambda}_s = X_s/g \tag{8}$$

and

$$\hat{\sigma}_s^2 = (\hat{\lambda}_s^2\sigma_{(g)}^2 + g\hat{\lambda}_s)/(g^2+\sigma_{(g)}^2). \tag{9}$$

Using the assumptions just made, it is shown in Appendix II that $E(\hat{\lambda}_s) = \lambda_s$ and $E(\hat{\sigma}_s^2) = \sigma_s^2$.

We have therefore produced estimators $\hat{\lambda}_s$ and σ_s^2 whose existence was required in Section II.

V. Use of the Formula for Sample Size

In Section III we obtained a formula, (7), for the sample size n required for a certain precision of estimation of the flock average egg-count λ by the sample average egg-count $\hat{\lambda}$. The use of (7) depends on our having knowledge of λ, σ^2, and $\sigma_{(\lambda)}^2$.

If we have prior estimates $\bar{\lambda}$, $\bar{\sigma}^2$, and $\bar{\sigma}_{(\lambda)}^2$ of λ, σ^2, and $\sigma_{(\lambda)}^2$, we may write, in analogy with (7), the formula

$$\bar{n} = N\left\{\frac{1+\bar{\sigma}^2/\bar{\sigma}_{(\lambda)}^2}{1+Np^2/(n_\beta^2\bar{c}^2)}\right\}, \tag{10}$$

where $\bar{c} = \bar{\sigma}_{(\lambda)}/\bar{\lambda}$. We shall henceforth use $\bar{n}$ as an estimate of the n given by (7).

In order to obtain the prior estimates just referred to, we shall suppose that we have a preliminary sample of m sheep drawn from the flock, and that $\hat{\lambda}_s$ and $\hat{\sigma}_s^2$ are estimated as in Section IV. Then, if we let $c_g = \sigma_{(g)}/g$, from equation (3)

$$\bar{\sigma}^2 = (c_g^2 \sum_{i=1}^{m} \hat{\lambda}_i^2/m + \bar{\lambda}/g)/(1+c_g^2),$$

and $\bar{\lambda}$ and $\bar{\sigma}_{(\lambda)}^2$ are defined (analogously with (1) and (4)) by

$$\bar{\lambda} = \sum_{i=1}^{m} \hat{\lambda}_i/m$$

and

$$\bar{\sigma}_{(\lambda)}^2 = \sum_{i=1}^{m} (\hat{\lambda}_i - \bar{\lambda})^2/(m-1) - \bar{\sigma}^2.$$

Eliminating $\sum_{i=1}^{m} \hat{\lambda}_i^2$ between the equations for $\bar{\sigma}^2$ and $\bar{\sigma}_{(\lambda)}^2$, we obtain

$$\bar{n} = N\left\{\frac{\bar{c}^2 + 1/(\bar{\lambda}g) + c_g^2(1+\bar{c}^2)}{(\bar{c}^2 + Np^2/n_\beta^2)(1+c_g^2/m)}\right\}. \tag{11}$$

If it happens that $\bar{\sigma}_{(\lambda)}^2$, and thus $\bar{c}^2$, are negative, it is suggested that in (11) $\bar{c}^2$ be set to zero. This will give

$$\bar{n} = n_\beta^2\{1/(\bar{\lambda}g) + c_g^2\}/\{p^2(1+c_g^2/m)\}. \tag{11a}$$

As special experimentation would be necessary to determine $\sigma^2_{(g)}$, for convenience it can be assumed to be zero, and we accordingly write

$$\overline{n} = N\left(\frac{\overline{c}^2+1/(\overline{\lambda}g)}{\overline{c}^2+Np^2/n^2_{\overline{\beta}}}\right), \tag{12}$$

or, if $\overline{c}$ has to be set to zero,

$$\overline{n} = n^2_{\overline{\beta}}/(\overline{\lambda}gp^2). \tag{12a}$$

The estimate $\overline{n}$ given by (12) (or (12a)) will mostly underestimate n. If it does, the larger $\sigma^2_{(g)}$ is, the greater will this underestimation be. If some estimate of $\sigma^2_{(g)}$ has been obtained, then the $\overline{n}$ given by (11) (or (11a)) should be used. In practice this means that variation in the weights of the samples of faeces should be kept as small as possible, and if it cannot be kept small some estimation of its magnitude should be obtained and made use of.

Having chosen a preliminary sample of m sheep from the flock from which to calculate $\overline{n}$, when $m \geqslant \overline{n}$ the value of $\overline{\lambda}$ calculated from this preliminary sample can be used as the final estimate of λ. Otherwise, if $m < \overline{n}$, it would be necessary to sample a further $\overline{n}-m$ sheep from the flock and to calculate $\hat{\lambda}$ from the combined sample.

It is realized that it may not be convenient to resample a flock of sheep at a later date. To avoid this inconvenience, an alternative and less accurate approach is to use equation (12) (or (12a)) (or (11) (or (11a))) for calculating $\overline{n}$, employing any pertinent estimates $\overline{\lambda}$ and $\overline{c}$. [It is because λ is more easily estimated than $\sigma^2_{(\lambda)}$, and c is generally less variable with time and locality than $\sigma^2_{(\lambda)}$, that $\overline{n}$ was defined in formulae (11) and (12) in terms of $\overline{\lambda}$ and $\overline{c}$ rather than in terms of $\overline{\lambda}$ and $\overline{\sigma}^2_{(\lambda)}$. Similarly g and c_g were used instead of g and $\sigma^2_{(g)}$.] Since conditions in the flock may vary widely from time to time, even past information on the same flock may be unreliable as an indication of present conditions. However, an educated guess can be better than nothing, especially if conservative figures are used, that is, figures which give rise to a conservative estimate $\overline{n}$ of n. From this viewpoint, an estimate of λ is conservative if it is small; so the smallest value of $\overline{\lambda}$ consistent with existing knowledge of the flock should be used. In formula (12), if $\overline{\lambda} > n^2_{\overline{\beta}}/(Np^2g)$ then a conservative value of $\overline{c}$ would be the largest likely value, while if $\overline{\lambda} < n^2_{\overline{\beta}}/(Np^2g)$ a conservative $\overline{c}$ would be the smallest value that might occur.

VI. Summary and Numerical Example

In this section a summary is given of the method of estimating sample size which has evolved from previous sections. It should not be necessary to read Sections II–V in order to obtain a working understanding of the method.

Suppose there is a flock of $N = 500$ sheep and we want to estimate the average egg-count. To do this we propose to sample n sheep from the flock and use the average egg-count of the sample, which we denote by $\hat{\lambda}$, as an estimate of the average egg-count of the flock. A faecal sample of size g grams is taken from each sheep in the sample and the egg-count per gram estimated from it. We shall take in our example $g = 1/60$, which is the amount of dry faeces from which the counts are made.

The problem is to find n. The value of n naturally depends on our having certain prior information about the flock; this can be obtained from a preliminary sample taken from the flock or from other pertinent data.

Firstly, we shall deal with the case when a preliminary sample of, say, $m = 50$ sheep has been taken from the flock. From this sample the average egg-count ($\overline{\lambda}$) and the between sheep variance for egg-count ($\overline{\sigma}^2_{(\lambda)}$) are calculated. The formula for $\overline{\sigma}^2_{(\lambda)}$ is

$$\overline{\sigma}^2_{(\lambda)} = \sum_{i=1}^{m} (\hat{\lambda}_i - \overline{\lambda})^2/(m-1) - \overline{\lambda}/g,$$

where $\hat{\lambda}_1, \hat{\lambda}_2, \ldots, \hat{\lambda}_m$ are the egg-counts from the preliminary sample. The two estimates $\hat{\lambda}$ and $\hat{\sigma}^2_{(\lambda)}$ may, for example, have values 1500 and 1200^2 respectively. The coefficient of variation is $\overline{c} = 1200/1500 = 0{\cdot}8$. An estimate of the sample size n is given by

$$N\left\{\frac{\overline{c}^2 + 1/(\overline{\lambda}g)}{\overline{c}^2 + Np^2/n_\beta^2}\right\} = 500\left\{\frac{0{\cdot}8^2 + 1/\{1500 \times \frac{1}{60}\}}{0{\cdot}8^2 + 500 \times 0{\cdot}1^2/1{\cdot}96^2}\right\} = 175.$$

It is therefore necessary to sample an additional $175 - 50 = 125$ sheep from the flock. From this combined sample of 175 the average egg-count $\hat{\lambda}$ is calculated.

The value of $1{\cdot}96$ chosen for n_β in the above formula is the 95% point of the normal distribution. The choice of the constant p as $0{\cdot}1$ enables us to say that with 95% probability the average egg-count for the flock lies between the two values

$$\frac{\hat{\lambda}}{1+p} = 0{\cdot}91\hat{\lambda} \text{ and } \frac{\hat{\lambda}}{1-p} = 1{\cdot}11\hat{\lambda}.$$

Secondly, if from past data we can obtain estimates of the average egg-count and the coefficient of variation, these can be used to estimate sample size in the same way as was done with the values obtained from a preliminary sample. However, in general these estimates will not be as reliable, for the condition of a flock may vary considerably with time. To counter this, a conservative estimate of n will be obtained if a small rather than a large value of $\overline{\lambda}$ is used. When $\overline{\lambda} > n_\beta^2/(Np^2g)$, $\overline{c}$ should be chosen large and when $\overline{\lambda} < n_\beta^2/(Np^2g)$, $\overline{c}$ should be small.

In the above example, suppose that 1500 and $0{\cdot}8$ were values of $\overline{\lambda}$ and $\overline{c}$ respectively obtained from past data on the flock. If it were thought that there had been some increase in infection since these values were found, it might still be better to take $\overline{\lambda}$ equal to 1500 instead of some larger value. As $n_\beta^2/(Np^2g) = 46$, which is less than $\overline{\lambda}$, a conservative estimate of $\overline{c}$ would perhaps be 1. Past observation of the behaviour of $\overline{\lambda}$ and $\overline{c}$ would be valuable in decisions of this type.

If it is found that $\overline{\sigma}^2_{(\lambda)}$, and thus $\overline{c}^2$, are negative, then $\overline{c}^2$ should be made zero and the estimate of sample size given above should be replaced by the estimate $n_\beta^2/(\overline{\lambda}gp^2)$.

Finally, it should be noted that the formula given for the sample size assumes that the amount of dry faeces from which egg-counts are obtained is constant and known. If these conditions are not satisfied the variability should be estimated (as σ_s), and the more general approach developed in this paper must be used.

VII. Acknowledgment

The authors wish to thank Dr H. Gordon for suggesting the problem and for helpful discussion on matters of application.

Appendix I

Evaluation of $E(\hat{\lambda})$ and $\mathrm{Var}(\hat{\lambda})$

In distinction from the procedure adopted in the text, we shall number the elements of a random sample $j_1, j_2, \ldots, j_n$. We shall denote expectations and variances with respect to the sampling distribution by E_j and Var_j respectively, and expectations with respect to the distribution of egg-counts within a sheep by E_λ.

Using well-known results of the theory of finite sampling,

$$E_j\{\sum_{i=1}^{n}\lambda_{j_i}/n\} = \lambda$$

and

$$\mathrm{Var}_j\{\sum_{i=1}^{n}\lambda_{j_i}/n\} = (1/n-1/N)\sigma_\omega^2.$$

Then

$$E(\hat{\lambda}) = E_j[E_\lambda\{\sum_{i=1}^{n}\hat{\lambda}_{j_i}/n \mid j\}] = E_j\{\sum_{i=1}^{n}\lambda_{j_i}/n\} = \lambda$$

and

$$\mathrm{Var}(\hat{\lambda}) = E\{(\hat{\lambda}-\lambda)^2\} = E[\{\sum_{i=1}^{n}(\hat{\lambda}_{j_i}-\lambda_{j_i})/n+(\sum_{i=1}^{n}\lambda_{j_i}/n-\lambda)\}^2]$$

$$= E_j[E_\lambda\{\sum_{i=1}^{n}(\hat{\lambda}_{j_i}-\lambda_{j_i})^2/n^2 \mid j\}]+E_j[\{\sum_{i=1}^{n}\lambda_{j_i}/n-\lambda\}^2]$$

(since the two cross-product terms vanish, one of them vanishing because the estimates $\hat{\lambda}_s$ are independent)

$$= E_j\{\sum_{i=1}^{n}\sigma_{j_i}^2/n^2\}+\mathrm{Var}_j\{\sum_{i=1}^{n}\lambda_{j_i}/n\}$$

$$= \frac{\sigma^2}{n}+\left(\frac{1}{n}-\frac{1}{N}\right)\sigma_\omega^2.$$

Appendix II

Proof that $E(\hat{\lambda}_s) = \lambda_s$ and $E(\hat{\sigma}_s^2) = \sigma_s^2$

Let the probability generating function (p.g.f.) of the conditional random variable $X_s \mid g_s$ be denoted by $P_s(t \mid g_s)$. Then, $X_s \mid g_s$ being Poisson with parameter $\lambda_s g_s$,

$$P_s(t \mid g_s) = \exp\{\lambda_s g_s(t-1)\}.$$

Let the distribution function of g_s be Φ and the moment generating function be m. Then, if $P_s(t)$ is the p.g.f. of the unconditional random variable X_s, it follows that

$$P_s(t) = \int_0^\infty P_s(t \mid g_s)\mathrm{d}\Phi(g_s) = \int_0^\infty \exp\{\lambda_s g_s(t-1)\}\mathrm{d}\Phi(g_s) = m\{\lambda_s(t-1)\}.$$

It is immediately clear that $p^{(j)}(1) = \lambda_s^j m^{(j)}(0)$, where $f^{(j)} = \mathrm{d}^j f/\mathrm{d}x^j$; and hence $\mu_{s[j]} = \lambda_s^j \mu_j(g)$, where $\mu_{s[j]}$ is the jth factorial moment of X_s and $\mu_j(g)$ is the jth power moment of Φ. In particular $E(X_s) = \lambda_s g$ and $\mathrm{Var}(X_s) = \lambda_s^2 \sigma_{(g)}^2 + \lambda_s g$.

If $\hat{\lambda}_s = X_s/g$, then $E(\hat{\lambda}_s) = \lambda_s$ and $\sigma_s^2 = \mathrm{Var}(\hat{\lambda}_s) = \lambda_s^2 \sigma_{(g)}^2/g^2 + \lambda_s/g$. If $\hat{\sigma}_s^2$ is defined as in the main text, it follows easily that $E(\hat{\sigma}_s^2) = \sigma_s^2$.

A(d)[8]

SIAM J. APPL. MATH.
Vol. 21, No. 1, July 1971

A NOTE ON THE ROOTS OF THE POLYNOMIAL EQUATION $f(x) = a$ WITH REFERENCE TO STABILITY*

G. M. TALLIS† AND G. GORDON‡

1. Introduction. We are concerned here with the behavior of the roots of the equation $f(x) = a$, where f is a polynomial and a is a parameter.

A polynomial equation $f(x) = 0$ will be called stable if the real part of each root is negative. The first result provides a sufficient condition for $f(x) = a$ to be stable when $f(x) = 0$ is stable. The second theorem considers the case where $f(x) = 0$ has a simple real root greater than the real part of all the other roots. For a certain class of polynomials, it gives a sufficient condition for $f(x) = a$ to have the same property.

These results are applied to a cyclic compartmental model specified by a system of first order, linear differential equations with constant coefficients.

2. Results.

THEOREM 1. *Let* $f(x) = \prod_{i=1}^{n}(x - \alpha_i)$; *then*:

(a) *if* $|a| < \prod_{i=1}^{n}|\mathrm{Re}\,\alpha_i|$, *the equation* $f(x) - a = 0$ *has no roots on the imaginary axis*;

(b) *if* $|a| < \prod_{i=1}^{n}|\mathrm{Re}\,\alpha_i|$, *and the equation* $f(x) = 0$ *is stable, then the equation* $f(x) - a = 0$ *is also stable.*

Proof. (a) Suppose $f(x) - a = 0$ has a root on the imaginary axis. Then $f(ik) = a$ for some real k and

$$|f(ik)| = |a|. \tag{1}$$

But

$$|f(ik)| = \prod_{i=1}^{n}|ik - \alpha_i| \geqq \prod_{i=1}^{n}|\mathrm{Re}\,\alpha_i|.$$

So, if $\prod_{i=1}^{n}|\mathrm{Re}\,\alpha_i| > |a|$, equation (1) cannot be satisfied.

(b) Let the real parts of the roots of $f(x) - a = 0$ be denoted by $\beta_i(a)$, $i = 1, \cdots, n$. Each $\beta_i(a)$ is a continuous function of a (see [2]) and $\beta_i(0) < 0$, $i = 1, \cdots, n$. Suppose $f(x) - a_1 = 0$ has a root with positive real part for some a_1 such that

$$|a_1| < \prod_{i=1}^{n}|\mathrm{Re}\,\alpha_i| = \prod_{i=1}^{n}|\beta_i(0)|.$$

Then "by continuity" there will be an a_2, with $0 < |a_2| < |a_1|$, for which the equation $f(x) - a_2 = 0$ has a root with real part zero. But $|a_2| < \prod_{i=1}^{n}|\mathrm{Re}\,\alpha_i|$,

* Received by the editors September 18, 1969, and in revised form September 9, 1970.

† Department of Statistics, University of Adelaide, Adelaide, 5001, South Australia.

‡ Division of Mathematical Statistics, OSIRO, Newton, New South Wales, 2042, Australia.

contradicting (a). Hence the roots of $f(x) - a = 0$ have negative real parts if $|a| < \prod_{i=1}^{n} |\mathrm{Re}\ \alpha_i|$.

Note that if $f(x)$ has real coefficients, the complex roots of $f(x) = 0$ occur in conjugate pairs. In this case the real parts of the complex roots may be obtained from the coefficients of x in the quadratic terms of the factorization of $f(x)$. If all the α_i are real and less than zero, and $f(x) = \sum_{j=0}^{n} c_j x^j$ with $c_n = 1$, then $f(x) - a = 0$ is stable if $|a| < c_0 = (-1)^n \prod_{i=1}^{n} \alpha_i$.

The polynomials considered subsequently will have real coefficients. Thus, let $f(x) = \sum_{j=0}^{n} c_j x^j$, c_j real, $c_n = 1$. Suppose $f(x)$ has m pairs of conjugate complex roots and $n - 2m$ real roots, so that

$$f(x) = \prod_{i=1}^{n} (x - \alpha_i),$$

where $\alpha_i = x_i + iy_i$ and $\alpha_{2j} = \bar{\alpha}_{2j-1}$, $1 \leqq j \leqq m$; i.e., $y_{2j} = -y_{2j-1}$, $1 \leqq j \leqq m$, and $y_i = 0$, $i > 2m$. The roots of $f(x) - a = 0$, a real, will be denoted by $\alpha_i(a)$, $\alpha_i(0) = \alpha_i$, $i = 1, \cdots, n$.

THEOREM 2. *Suppose that*:

(i) *the equation $f(x) = 0$ has a simple real root, α_n say, such that $\alpha_n > \mathrm{Re}\ \alpha_i$, $i < n$, and hence $\alpha_n > \gamma$, where γ is the largest real root of $f'(x) = 0$;*

(ii) *$A - x_j \geqq |y_j|$, $j = 1, \cdots, 2m$, for some A, $\alpha_n \geqq A > \gamma$; and*

(iii) *$a \geqq f(A)$.*

Then the equation $f(x) = a$ also has a simple real root $\alpha_n(a)$ such that $\alpha_n(a) > \mathrm{Re}\ \alpha_i(a)$, $i < n$.

Proof. We verify first that $\alpha_n > \gamma$ as stated in (i). By the Gauss–Lucas lemma, the zeros of $f'(x) = 0$ lie in the smallest convex polygon in the complex plane containing the roots of $f(x) = 0$ (see [2]). If $\alpha_n < \gamma$, this result contradicts the assumption in (i) that $\alpha_n > \mathrm{Re}\ \alpha_i$, while if $\alpha_n = \gamma$, there is a contradiction with the assumption that α_n is simple. Thus $\alpha_n > \gamma$ as required.

Since $c_n = 1$, $f'(x) > 0$ for $x > \gamma$, and so $f(x)$ is monotonic increasing for $x > \gamma$. Thus if $a > f(\gamma)$, the equation $f(x) - a = 0$ will have a real, simple root $\alpha_n(a)$, which is a monotonic increasing function of a. Also, since $\alpha_n \geqq A > \gamma$, by the monotonicity of $f(x)$, it follows that $f(A) \leqq 0$. Now, for $1 \leqq j \leqq n$, let $\alpha_j(a) = x_j(a) + iy_j(a)$, $x_j(0) = x_j$, $y_j(0) = y_j$. Both $x_j(a)$ and $y_j(a)$ are continuous functions of a. Suppose for some $i < n$ and some $a \in [f(A), \infty)$, $x_i(a) \geqq \alpha_n(a)$. Then, since $x_i(0) < \alpha_n(0)$ and $0 \in [f(A), \infty)$, by continuity, there exists an $a_1 \in [f(A), \infty)$, $0 < a_1 \leqq a$ if $a > 0$, $0 > a_1 \geqq a$ if $a < 0$, such that $x_i(a_1) = \alpha_n(a_1) = \alpha$, say; i.e., $\alpha_i(a_1) = \alpha + ik$, where $k = y_i(a_1)$.

Thus $a_1 = f(\alpha) = f(\alpha + ik)$, so that

$$|f(\alpha)| = |f(\alpha + ik)|.$$

Now

$$|f(\alpha)| = \prod_{j=1}^{n} (\alpha - \alpha_j)$$

$$= \prod_{j=1}^{n} [(\alpha - x_j)^2 + y_j^2]^{1/2},$$

while

$$|f(\alpha + ik)| = \left| \prod_{j=1}^{n} (\alpha + ik - \alpha_j) \right|$$
$$= \prod_{j=1}^{n} |\alpha - x_j + i(k - y_j)|$$
$$= \prod_{j=1}^{n} [(\alpha - x_j)^2 + (k - y_j)^2]^{1/2}.$$

Thus

$$\prod_{j=1}^{n} [(\alpha - x_j)^2 + y_j^2]^{1/2} = \prod_{j=1}^{n} [(\alpha - x_j)^2 + (k - y_j)^2]^{1/2}. \tag{2}$$

Let

$$R_1 = \prod_{j=1}^{2m} \frac{[(\alpha - x_j)^2 + (k - y_j)^2]^{1/2}}{[(\alpha - x_j)^2 + y_j^2]^{1/2}},$$
$$R_2 = \prod_{j=2m+1}^{n} \frac{[(\alpha - x_j)^2 + k^2]^{1/2}}{[(\alpha - x_j)^2]^{1/2}}$$

and $R = R_1 R_2$. Clearly $R_2 \geqq 1$, with equality if and only if $k = 0$. On the other hand, R_1 can be rewritten as

$$R_1 = \prod_{r=1}^{m} \frac{[(\alpha - x_{2r})^2 + (k - y_{2r})^2]^{1/2}[(\alpha - x_{2r})^2 + (k + y_{2r})^2]^{1/2}}{(\alpha - x_{2r})^2 + y_{2r}^2}.$$

Set $d = \alpha - x$; then the square of a typical factor is of the form

$$\frac{[d^2 + y^2 + k^2 - 2ky][d^2 + y^2 + k^2 + 2ky]}{[d^2 + y^2]^2}$$
$$= \frac{[d^2 + y^2 + k^2]^2 - 4k^2y^2}{[d^2 + y^2]^2}$$
$$= \frac{[d^2 + y^2]^2 + 2k^2[d^2 + y^2] + k^4 - 4k^2y^2}{[d^2 + y^2]^2}$$
$$= 1 + \frac{k^2[k^2 + 2(d^2 - y^2)]}{[d^2 + y^2]^2}.$$

Since $a_1 \geqq f(A)$, by the monotonicity of $f(x)$, $\alpha_n(a_1) \geqq A$. Thus

$$d = \alpha_n(a_1) - x \geqq A - x \geqq |y|,$$

so that

$$d^2 \geqq y^2.$$

It follows that $k^2 + 2(d^2 - y^2) > 0$ for all $k \neq 0$, and so $R > 1$ for all $k \neq 0$. By (2), $R = 1$, so we conclude that $k = 0$. But this implies that $\alpha_i(a_1) = \alpha_n(a_1)$, i.e., that there is a double real root of $f(x) - a_1 = 0$. This implies that $\alpha_n(a_1)$

is a real root of $f'(x) = 0$, contradicting the choice of γ as the largest real root of $f'(x) = 0$, since $\alpha_n(a_1) \geqq A > \gamma$.

COROLLARY. *Suppose that*:

(i) $f(x) = 0$ *has real roots, and the largest root* α_n *is simple*;

(ii) $a > f(\gamma)$, *where* γ *is the largest real root of* $f'(x) = 0$.

Then $f(x) = a$ *also has a simple real root,* $\alpha_n(a)$, *such that* $\alpha_n(a) > \operatorname{Re} \alpha_i(a)$, $i < n$.

Proof. Since $m = 0$, $R = R_2 > 1$ unless $k = 0$, and contradiction follows as before.

When $\alpha_n < 0$, that is when $f(x) = 0$ has real roots and is stable, the above corollary may be used instead of Theorem 1(b). The conditions on a given by the two results are different, however, and neither includes the other.

3. Application. During a mathematical investigation of the life cycle of an intestinal parasite of sheep [1] the linear system of differential equations $d\mathbf{x}/dt = B\mathbf{x}$, where

$$B = \begin{bmatrix} -\kappa_1 & & & & & & & & \lambda_9 \\ \lambda_1 & -\kappa_2 & & & & & & & \\ & \lambda_2 & -\kappa_3 & & & & & & \\ & & \cdot & \cdot & & & & & \\ & & & \cdot & \cdot & & & & \\ & & & & & & \lambda_8 & -\kappa_9 \end{bmatrix}$$

was used to determine the number of parasites in each of 9 stages of the life cycle at time t after an initial dose. The parameter λ_i determines the rate at which parasites transfer from the ith stage to the $(i + 1)$th, except for the 5th stage where λ_5 is the egg production rate of a female parasite. The parameter κ_i determines the decay of parasites in the ith stage, due to death and transfer to the next stage. Thus $\lambda_i > 0$, $\kappa_i > 0$, $\lambda_i \leqq \kappa_i$, $i \neq 5$.

It is important to investigate the stability of the above system. The characteristic equation associated with B is

$$\prod_{i=1}^{9} (x + \kappa_i) - \prod_{i=1}^{9} \lambda_i = 0,$$

which is of the form $f(x) - a = 0$, with all roots of $f(x) = 0$ real and negative, and with

$$a = \prod_{i=1}^{9} \lambda_i > 0.$$

Moreover, $f(x) = \prod_{i=1}^{9} (x + \kappa_i)$ is a monotonic increasing function of x for $x \in (-\kappa, \infty)$, where $\kappa = \min \kappa_i$.

Thus (i) if $\prod_{i=1}^{9} \lambda_i > \prod_{i=1}^{9} \kappa_i$, there is a real root greater than zero and the system is unstable;

(ii) if $\prod_{i=1}^{9} \lambda_i < \prod_{i=1}^{9} \kappa_i$, by Theorem 1(b) the system is stable.

The corollary to Theorem 2 shows that when $a > 0$ the root of $f(x) - a = 0$ with the greatest real part is the largest real root, $\kappa(a)$. Since $\kappa(a)$ is a monotonic

increasing function of a, and $\kappa(a) = 0$ if $a = \prod_{i=1}^{9} \kappa_i$, Theorem 2 establishes the stability conditions (i) and (ii) independently of the result of Theorem 1. Since the root with the greatest real part is indicated when either (i) or (ii) holds, it also provides information useful for finding the asymptotic form of the solution to the system.

When (iii) $\prod_{i=1}^{9} \lambda_i = \prod_{i=1}^{9} \kappa_i$, the characteristic equation associated with B has a zero root, but no root with positive real part. By considering the diagonal form of the matrix B and the corresponding expression for e^{Bt}, it can be seen that the solution of the system of differential equations is bounded. That is, each element of $\mathbf{x}(t)$ for any t is less than a constant depending only on the parameters and the initial conditions. These bounds should not be confused with the number of parasites tolerated by the sheep. Even if (ii) or (iii) holds it would be possible for the numbers of parasites indicated by the solution $\mathbf{x}(t)$ to exceed the biological limits.

Note that the stability conclusions reached above may be generalized immediately to a cyclic compartmental model with an arbitrary number of compartments and leakage of flow ($\lambda_i < \kappa_i$), conservation of flow ($\lambda_i = \kappa_i$), or generation of flow ($\lambda_i > \kappa_i$) at any stage. This is because the form of the characteristic equation depends on the pattern of the matrix B, and the stability arguments above are independent of the degree of the characteristic polynomial.

REFERENCES

[1] G. GORDON, M. O'CALLAGHAN AND G. M. TALLIS, *A deterministic model for the life-cycle of a class of internal parasites of sheep*, Math. Biosci., 8 (1970), pp. 209–226.

[2] M. MARDEN, *The Geometry of the Zeros of a Polynomial in a Complex Variable*, Mathematical Surveys, No. 3, American Mathematical Society, New York, 1949.

A(d)[9]

A RELATIONSHIP BETWEEN MOMENT ESTIMATORS AND MAXIMUM LIKELIHOOD ESTIMATORS

G.M. Tallis and M. Hudson

(Unpublished)

The one parameter exponential family

$$f(x,\theta) = \exp\{\alpha(\theta)\, T(x) + \beta(\theta) + B(x)\},\ \theta \in \Omega, \quad (1)$$

where f is a density function with respect to a measure μ, features widely in the statistical literature. There are well known results which connect sufficient statistics, complete sufficient statistics and minimum variance bound estimators with (1).

In this note it is shown, under two types of regularity conditions, that a necessary and sufficient condition for there to exist a maximum likelihood estimator (m.l.e.), $\hat{\theta}_n$, for θ which is equal to a moment estimator (m.e.), $\bar{\theta}_n$, is that the parent density be of the form (1).

Results

(A) Under Standard Regularity Conditions

Four definitions, D1(a) - D4(a) are followed by Theorem 1(a).

D1(a) The Class $\mathcal{R}(\theta)$

Let $f(x,\theta)$ be a density function with respect to a σ-finite measure μ, defined on a subset R of the real line and depending on a parameter $\theta \in \Omega$. Then $f \in \mathcal{R}(\theta)$ if

(i) f satisfies the conditions necessary for establishing the Cramér-Rao inequality;

(ii) f satisfies the conditions necessary for deriving the asymptotic properties of m.l.e.'s (see Rao (1965), pages 263 and 299).

D2(a) The Class $\mathcal{E}(\theta)$

A density function $f \in \mathcal{R}(\theta)$ also belongs to the class $\mathcal{E}(\theta)$ if for $\theta \in \Omega$

(i) $f(x,\theta) = \exp\{\alpha(\theta)\, T(x) + \beta(\theta) + B(x)\}$;

(ii) $E[T(X)] = \tau(\theta) = -\beta'(\theta)/\alpha'(\theta)$;

(iii) $\tau'(\theta)$ exists for $\theta \in \Omega$ and $\tau'(\theta) \neq 0$;

(iv) $0 < V[T(X)|\theta] < \infty$, $T(R) = \tau(\Omega)$.

D3(a) A M.E. of θ.

A m.e. of θ will be said to exist if for $f \in \mathcal{R}(\theta)$

(i) there exist functions $T(X)$ and $\tau(\theta)$ such that $E[T(X)] = \tau(\theta)$, $\theta \in \Omega$;

(ii) $\tau'(\theta)$ exists for $\theta \in \Omega$ and $\tau'(\theta) \neq 0$;

(iii) $0 < V[T(X)|\theta] < \infty$, $T(R) = \tau(\Omega)$.

Then, for a random sample of size n, a m.e. for θ is defined to be a consistent root $\bar{\theta}_n$ of

$$\tau(\bar{\theta}_n) = \bar{T}_n = \Sigma\, T(X_i)/n$$

and the set of consistent roots will be written $\{\bar{\theta}_n\}$.

D4(a) A m.l.e. of θ.

A m.l.e. of θ, $\hat{\theta}_n$, for $f \in \mathcal{R}(\theta)$, is defined as a consistent root of the equation

$$\sum_{i=1}^{n} \frac{\partial}{\partial\theta} \ln f(X_i, \hat{\theta}_n) = 0.$$

The set of consistent roots will be written $\{\hat{\theta}_n\}$.

Theorem 1(a)

For $f \in \mathcal{R}(\theta)$, $f \in \mathcal{E}(\theta)$ iff there exists $\bar{\theta}_n \in \{\bar{\theta}_n\}$ and $\hat{\theta}_n \in \{\hat{\theta}_n\}$ such that $\bar{\theta}_n = \hat{\theta}_n$.

Proof

If $f \in \mathcal{E}(\theta)$, then the likelihood function is

$$L(\underset{\sim}{x},\theta) = \exp\{\alpha(\theta) \sum_1^n T(x_i) + n\beta(\theta) + \sum_1^n B(x_i)\}$$

and

$$\frac{\partial}{\partial\theta} \ln L(\underset{\sim}{x},\hat{\theta}) = \alpha'(\hat{\theta}_n) \sum_1^n T(x_i) + n\beta'(\hat{\theta}_n) = 0$$

i.e. $$\bar{T}_n = -\beta'(\hat{\theta}_n)/\alpha'(\hat{\theta}_n) = \tau(\hat{\theta}_n) = \tau(\bar{\theta}_n)$$

and $$\{\hat{\theta}_n\} = \{\bar{\theta}_n\}.$$

Now suppose $\bar{\theta}_n = \hat{\theta}_n$ for $\bar{\theta}_n \in \{\bar{\theta}_n\}$ and $\hat{\theta}_n \in \{\hat{\theta}_n\}$. Then there exist functions $T(x)$ and $\tau(\theta)$ such that $E[T(X)] = \tau(\theta)$, $\tau'(\theta)$ exists for $\theta \in \Omega$ and $\tau'(\theta) \neq 0$. Thus

$$\tau(\hat{\theta}_n) = \tau(\bar{\theta}_n) = \bar{T}_n.$$

By the Central Limit Theorem

$$\sqrt{n}(\bar{T}_n - \tau(\theta)) \xrightarrow{L} N(0, V[T])$$

and since $f \in \mathcal{R}(\theta)$ and $\hat{\theta}_n$ is a consistent root of the likelihood equation

$$\sqrt{n}(\hat{\theta}_n - \theta) \xrightarrow{L} N(0, I^{-1}(\theta)), \quad I(\theta) = E[(\frac{\partial}{\partial\theta} \ln f(X,\theta))^2]$$

by standard theory. It follows from Rao (loc. cit.), page 320, that

$$\sqrt{n}(\tau(\hat{\theta}_n) - \tau(\theta)) = \sqrt{n}(\bar{T}_n - \tau(\theta)) \xrightarrow{L} N(0,[\tau'(\theta)]^2 I^{-1}(\theta))$$

whence $V[T] = [\tau'(\theta)]^2 I^{-1}(\theta)$. Thus $V[T]$ achieves the minimum variance bound. Again from Rao, page 264, this implies that

$$f(x,\theta) = \exp\{\alpha(\theta)\, T(x) + \beta(\theta) + B(x)\}$$

where $E[T(X)] = \tau(\theta) = -\beta'(\theta)/\alpha'(\theta)$.

This result emphasises that a moment estimator is not even a candidate for asymptotic efficiency outside $\mathcal{E}(\theta)$.

Under Special Regularity Conditions

In this section we replace D1(a)-D4(a) by D1(b)-D4(b) and Theorem 1(a) by Theorem 1(b).

D1(b) The Class $\mathcal{R}^x(\theta)$

Let $f(x,\theta)$ be a density function with respect to a σ-finite measure μ, defined on R and depending on a parameter $\theta \in \Omega$. Then $f \in \mathcal{R}^x(\theta)$ if $\partial f(x,\theta)/\partial\theta$ exists for $\theta \in \Omega$ and f is continuous in x.

D2(b) The Class $\mathcal{E}^x(\theta)$

The density function $f \in \mathcal{E}^x(\theta) \subset \mathcal{R}^x(\theta)$ if

(i) = D2(a)(i)

(ii) = D2(a)(ii)

(iii) $T(R) = \tau(\Omega)$ and there exists a partition of R and Ω by intervals, I_{x_j} and I_{θ_j}, $j=1,2,\ldots k$, such that for all j

(1) $T(x)$ is one to one on I_{x_j};

(2) $\tau(\theta)$ is one to one on I_{θ_j};

(3) $T(I_{x_j}) = \tau(I_{\theta_j})$.

D3(b) A M.E. Set for θ.

A m.e. set is said to exist for θ for $f \in \mathcal{R}^X(\theta)$ if

(i) = D3(a)(i)

(ii) = D2(b)(iii).

Then for a random sample of size n, a m.e. set for θ is defined by

$$\{\bar{\theta}_n\}^* = \{\theta; \tau(\theta) = \bar{T}_n\}, \quad \bar{T}_n = \sum_{i=1}^{n} T(X_i)/n.$$

D4(b) A M.L.E. Set for θ.

A m.l.e. set for θ, is said to exist for $f \in \mathcal{R}^X(\theta)$ if $L(\underset{\sim}{x},\theta) = \prod_{i=1}^{n} f(x_i,\theta)$, $n=1,2,\ldots$ has stationary values which are found as roots of the equation

$$\sum_{i=1}^{n} \frac{\partial}{\partial\theta} \ln f(x_i,\theta) = \sum_{i=1}^{n} g(x_i,\theta) = 0;$$

we put

$$\{\hat{\theta}_n\}^* = \{\theta; \sum_{i=1}^{n} g(x_i,\theta) = 0\}.$$

Theorem 1(b)

For $f \in \mathcal{R}^X(\theta)$, $f \in \mathcal{E}^X(\theta)$ iff there exist $\bar{\theta}_n \in \{\bar{\theta}_n\}^*$ and $\hat{\theta}_n \in \{\hat{\theta}_n\}^*$ such that $\bar{\theta}_n = \hat{\theta}_n$.

Proof

If $f \in \mathcal{E}^X(\theta)$ then the likelihood function is

$$L(\underset{\sim}{x},\theta) = \exp\{\alpha(\theta) \sum_{1}^{n} T(x_i) + n\beta(\theta) + \sum_{1}^{n} B(x_i)\}$$

and equating $\partial \ln L(\underset{\sim}{x},\theta)/\partial\theta$ to zero gives

$$\tau(\theta) = \bar{T}_n$$

and clearly $\{\bar{\theta}_n\}^* = \{\hat{\theta}_n\}^*$.

For $x \in I_{x_j}$ and $\theta \in I_{\theta_j}$ let $t = T_j(x)$ and $\omega = \tau_j(\theta)$, then $f(x,\theta) = f(T_j^{-1}(t), \tau_j^{-1}(\omega)) = \phi(t,\omega)$, say, where T_j and τ_j are the restrictions of T and τ to I_{x_j} and I_{θ_j} respectively. Moreover, since $\bar{\theta}_n = \hat{\theta}_n$ for $\bar{\theta}_n \in \{\bar{\theta}_n\}^*$ and $\hat{\theta}_n \in \{\bar{\theta}_n\}^*$, $\bar{\theta}_n = \hat{\theta}_n = \tau_j^{-1}(\bar{T}_n)$ where $\bar{T}_n = \sum_{i=1}^{n} T(x_i)/n$, $x_i \in I_{x_j}$ for all i. It follows that, by the 1:1 property of τ_j, $\hat{\omega}_n = \tau_j(\hat{\theta}_n) = \bar{T}_n$. Thus

$$0 = \sum_{i=1}^{n} \frac{\partial}{\partial\omega} \ln \phi(t_i,\hat{\omega}_n) = \sum_{i=1}^{n} k(t_i,\bar{t}_n), \text{ say,} \qquad (2)$$

holds for t_i in $I_t = T(I_x)$, dropping the subscript j.

The functional equation (2) is of the form

$$\sum_{i=1}^{m+1} h(y_i,z) = 0; \quad \sum_{i=1}^{m+1} y_i = (m+1)z; \quad y_i, z \in I_t$$

which reduces to

$$\sum_{i=1}^{m+1} p(u_i,z) = 0 \quad \sum_{i=1}^{m+1} u_i = 0, \quad u_i = y_i - z,$$

where the u_i belong to some interval I_u containing zero. Note that $u_{m+1} = -\sum_{i=1}^{m} u_i$ and $p(0,z) = 0$ since

for $u_i \equiv 0$ $(m+1)p(0,z) = 0$. Set $u_2 = \ldots = u_m = 0$ so that $u_{m+1} = -u_1$ and $p(u,z) = -p(-u,z)$. Thus

$$p(u_1,z) + \ldots + p(u_m,z) = -p(-\sum_1^m u_i,z) = p(\sum_1^m u_i,z) \qquad (3)$$

and, putting $m = 2$, we have the Cauchy functional equation

$$p(u_1,z) + p(u_2,z) = p(u_1 + u_2,z),$$

which has the unique continuous solution, Aczél (1966) page 31,

$$p(u,z) = c(z)u$$

for some constant, $c(z)$, and for u in a subinterval of I_u. The relation extends to all of I_u by (3). Thus

$$h(y,z) = c(z)\ (y-z),\ y \in I_t,\ z \in I_t$$

and

$$\frac{\partial \ln \phi(t,\omega)}{\partial \omega}\Big|_{\omega=\bar{t}_n} = k(t,\bar{t}_n) = c(\bar{t}_n)\ (t-\bar{t}_n)$$

holds for all $\bar{t}_n \in I_t$. Clearly, therefore,

$$\frac{\partial}{\partial \omega} \ln \phi(t,\omega) = c(\omega)\ (t-\omega),\ \omega \in I_t$$

and $\phi(t,\omega)$ is of the form (1) for $(t,\omega) \in I_t \times I_t$.

Back transformation to x and θ leaves the form unchanged and $f(x,\theta)$ is of exponential form for $x \in I_x$ and $\theta \in \tau^{-1}(I_t)$.

By applying this analysis to all I_{x_j}, I_{θ_j} pairs, the result follows by continuity and the assumption that $T(R) = \tau(\Omega)$.

REFERENCES

ACZEL, J. (1966) Lectures on Functional Equations and their Applications.
Academic Press, New York.

RAO, C.R. (1965) Linear Statistical Inference and its Applications.
John Wiley and Sons, New York.

A(d)[10]

Commonwealth of Australia
COMMONWEALTH SCIENTIFIC AND INDUSTRIAL RESEARCH ORGANIZATION

Proc. Camb. Phil. Soc. (1971), 69, 309
PCPS 69-31
Printed in Great Britain

A note on sufficient statistics and the exponential family

By G. M. TALLIS
Division of Mathematical Statistics, C.S.I.R.O.,
60 King Street, Newtown, N.S.W. 2042, Australia

(*Received* 14 *February* 1970)

1. *Introduction.* The relationship between sufficient statistics and the exponential family was first investigated by Pitman (8) and Koopman (7). The treatments assumed a density function $f(x;\theta)$ which was at least differentiable with respect to both arguments.

More recently the problem has been taken up again in an attempt to remove unnecessary restrictions. Dynkin (6) developed a theorem, restated by Brown (3), which required that $f(x;\theta)$ be continuously differentiable with respect to x on an interval. Subsequently Brown(3) proved a number of more general results under very mild measure theoretic restrictions on the family of densities and the statistic. Unfortunately, the proofs are long and involved. Other work in this area has been reported by Denny(4), (5).

It is the purpose of this note to prove, by elementary argument, a result that holds under very weak conditions. It turns out that Theorem 1 of this paper is nearly equivalent to Theorem 8·1′ of Brown. The main difference concerns the restriction placed on $f(x;\theta)$. Brown assumes that for all θ of interest and, for A any Lebesgue measurable set, f satisfies

$$\int_A f(x;\theta)\,d\mu = 0 \Leftrightarrow \int_A d\mu = 0,$$

where μ is Lebesgue measure. I assume the continuity of f at a certain point x_0 for all θ and this leads to a considerable shortening and simplification of the proof.

The same ideas are used to suggest an extension to the case where θ and x are vector valued. The purpose of this work is to provide a result of some generality which gives insight into underlying relationships without requiring a long and difficult proof.

2. *The main result.* The most important theorem will be established first. Theorem 2 is of a similar nature and its proof is only outlined.

Definition 1. A density function $f(x;\theta)$ with respect to a measure μ will be said to be of the one-parameter, continuous exponential type if

$$f(x;\theta) = \exp\{\alpha(\theta)\,A(x) + \beta(\theta) + B(x)\}, \tag{1}$$

where $A(x)$ and $B(x)$ are continuous functions of x in the interval R, $A(x)$ is strictly monotone in an interval contained in R, $\alpha(\theta)$ and $\beta(\theta)$ are bounded for $\theta \in \Omega$.

Definition 2. A statistic $t_n(x_1, \ldots, x_n)$, which will be written as $t_n(\mathbf{x})$, will be said to be additively sufficient for $\theta \in \Omega$ if

(a) $L(\mathbf{x};\theta) = \prod_{i=1}^{n} f(x_i;\theta) = g_n(t_n;\theta)\,h_n(\mathbf{x})$,

(b) $t_n(\mathbf{x}) = \sum_{i=1}^{n} k(x_i)$,

(c) $k(x)$ is continuous for $x \in R$, $k(x_0) = 0$ for some $x_0 \in R$ and $k(x)$ is strictly monotone in a neighbourhood of x_0.

It is assumed that $f(x;\theta) \neq 0$ for $x \in R$ and $\theta \in \Omega$.

LEMMA 1. *It can be assumed without loss of generality that* $h_n(\mathbf{x})$ *is of the form* $\prod_{i=1}^{n} h_1(x_i)$ *and* $g_1(t_1;\theta)$ *is continuous at* $t_1 = 0$.

Proof. Suppose $L(\mathbf{x};\theta) = g_n(t_n;\theta)\,h_n(\mathbf{x})$, then for any other $\theta' \in \Omega$

$$L(\mathbf{x};\theta') = g_n(t_n;\theta')\,h_n(\mathbf{x})$$

and hence $$L(\mathbf{x};\theta) = [g_n(t_n;\theta)/g_n(t_n;\theta')]\,L(\mathbf{x};\theta') = g_n^*(t_n;\theta)\,h_n^*(\mathbf{x}),$$

say, which proves the first part of the lemma. The second observation follows from the equation $g_1^*(t_1;\theta) = f(x;\theta)/f(x;\theta')$, the continuity of f and Definition 2(c).

THEOREM 1. *If* $f(x;\theta)$ *is continuous for* $x \in R$ and $f(x;\theta) \neq 0$ *for* $x \in R$ *and* $\theta \in \Omega$, *then an additively sufficient statistic for* θ *exists if and only if* $f(x;\theta)$ *is of the one-parameter, continuous exponential form.*

Proof. Let $f(x;\theta)$ be of the form (1). By assumption an $x_0 \in R$ exists such that $A(x)$ is strictly monotone in a neighbourhood of x_0. Set $k(x) = A(x) - A(x_0)$, then

$$L(\mathbf{x};\theta) = \exp\left\{\alpha(\theta)\sum_{i=1}^{n} k(x_i) + n\gamma(\theta)\right\}\exp\left\{\sum_{i=1}^{n} B(x_i)\right\},$$

where $\gamma(\theta) = \beta(\theta) + \alpha(\theta)\,A(x_0)$, and an additively sufficient statistic for θ exists.

On the other hand note that

$$f(x;\theta) = g_1(k(x);\theta)\,h_1(x)$$

and
$$\begin{aligned} f(x_1;\theta)f(x_2;\theta) &= g_2(k(x_1)+k(x_2);\theta)\,h_1(x_1)\,h_1(x_2) \\ &= g_1(k(x_1);\theta)\,g_1(k(x_2);\theta)\,h_1(x_1)\,h_1(x_2) \end{aligned}$$

by Definition 2 and Lemma 1. Thus

$$g_2(k(x_1)+k(x_2);\theta) = g_1(k(x_1);\theta)\,g_1(k(x_2);\theta)$$

and setting $x_2 = x_0$ and $g_1(0;\theta) = [C(\theta)]^{-1}$

$$g_2(k(x_1);\theta) = [C(\theta)]^{-1}\,g_1(k(x_1);\theta)$$

whence $$g_1(k(x_1)+k(x_2);\theta) = C(\theta)\,g_1(k(x_1);\theta)\,g_1(k(x_2);\theta). \tag{2}$$

Upon taking logarithms of both sides of (2) the functional equation takes the form

$$\Pi(u+v;\theta) = \Pi(u;\theta) + \Pi(v;\theta) - \gamma(\theta), \tag{3}$$

where u and v belong to some interval I containing zero. Subtract $\gamma(\theta)$ from both sides of (3) and let $\phi(u;\theta) = \Pi(u;\theta) - \gamma(\theta)$, then (3) becomes the standard Cauchy equation

$$\phi(u+v;\theta) = \phi(u;\theta) + \phi(v;\theta). \tag{4}$$

The function $\phi(u;\theta)$ is continuous at $u = 0$ by Lemma 1 and (4) shows that ϕ is continuous everywhere in I by letting v tend to zero. The unique continuous solution to (4) is Aczél ((1), p. 31)

$$\phi(u;\theta) = \alpha(\theta)\, u.$$

Finally
$$f(x;\theta) = \exp\{\alpha(\theta)\, k(x) + \gamma(\theta) + B(x)\}$$

which is of the one-parameter continuous exponential form.

A local result similar to that of Brown, Theorem 8·1′, is obtained by assuming that f is only continuous at x_0 and $f(x_0;\theta) \neq 0$.

3. *An extension.* In this section the Definitions 1 and 2 and Theorem 1 will be recast to take care of the case where θ and x are q and r dimensional vectors respectively.

Definition 3. A density function $f(\mathrm{x};\theta)$ with respect to a measure μ will be said to be of the continuous exponential form if

$$f(\mathrm{x};\theta) = \exp\{\alpha'(\theta)\, \mathrm{A}(\mathrm{x}) + \beta(\theta) + B(\mathrm{x})\}, \tag{5}$$

where $\mathrm{A}'(\mathrm{x}) = [A_1(\mathrm{x}), A_2(\mathrm{x}), \ldots, A_s(\mathrm{x})]$, for all i, $A_i(\mathrm{x})$ and $B(\mathrm{x})$ are continuous functions of x in a closed region R of Euclidean r-space, $\alpha'(\theta) = [\alpha_1(\theta), \ldots, \alpha_s(\theta)]$ and $\alpha_i(\theta)$ and $\beta(\theta)$ are bounded for $\theta \in \Omega$.

Definition 4. A statistic t_n based on a random sample of size n from $f(\mathrm{x};\theta)$ will be said to be additively sufficient for $\theta \in \Omega$ if

(*a*) $L(\mathrm{x}_1, \ldots, \mathrm{x}_n;\theta) = \prod_{i=1}^{n} f(\mathrm{x}_i;\theta) = g_n(\mathrm{t}_n;\theta)\, h_n(\mathrm{x}_1, \ldots, \mathrm{x}_n)$;

(*b*) $\mathrm{t}_n = \phi_n\left[\sum_{i=1}^{n} \mathrm{k}(\mathrm{x}_i)\right]$, where ϕ_n is a continuous vector valued function and $\mathrm{k}'(\mathrm{x}) = [k_1(\mathrm{x}), \ldots, k_s(\mathrm{x})]$;

(*c*) the $k_i(\mathrm{x})$ are continuous for $\mathrm{x} \in R$, a closed region, and $\mathrm{k}(\mathrm{x}_0) = 0$ for some $\mathrm{x}_0 \in R$;

(*d*) for $n = 1$, $f(\mathrm{x};\theta) = g_1(\mathrm{t}_1;\theta)\, h_1(\mathrm{x})$ and $g_1(\mathrm{t}_1;\theta)$ is continuous at $\mathrm{t}_1 = \mathrm{t}(x_0)$ for $\theta \in \Omega$.

It will be assumed that $f(\mathrm{x};\theta) \neq 0$ for $\mathrm{x} \in R$ and $\theta \in \Omega$. Moreover, for notational convenience, any function of $\mathrm{x}_1, \ldots, \mathrm{x}_n$, $p(\mathrm{x}_1, \ldots, \mathrm{x}_n)$ say, will be written as $p(\mathrm{y})$.

Lemma 2. *It can be assumed without loss of generality that* $h_n(\mathrm{y})$ *is of the form*

$$\prod_{i=1}^{n} h_1(\mathrm{x}_i).$$

The proof of this is the same as for Lemma 1, and the general result can now be stated.

Theorem 2. *If* $f(\mathrm{x};\theta)$ *is a density function with respect to* μ *and* $f(\mathrm{x};\theta) \neq 0$ *for* $\mathrm{x} \in R$, $\theta \in \Omega$, *then an additively sufficient statistic for* $\theta \in \Omega$ *exists if and only if* $f(\mathrm{x};\theta)$ *is of continuous exponential form.*

Proof. The proof will only be outlined since it parallels the proof of Theorem 1.

Let $f(\mathrm{x},\theta)$ be of form (5) and set

$$k_i(\mathrm{x}) = A_i(\mathrm{x}) - A_i(\mathrm{x}_0),$$

then $$L(\mathbf{y};\theta) = \exp\left\{\alpha'(\theta)\sum_{i=1}^{n} \mathrm{k}(\mathrm{x}_i) + n\gamma(\theta)\right\} \exp\left\{\sum_{i=1}^{n} B_i(\mathrm{x}_i)\right\}$$

where $\gamma(\theta) = \beta(\theta) + \alpha'(\theta)\,\mathrm{A}(\mathrm{x}_0)$. Hence, an additively sufficient statistic for θ is $t_n(\mathbf{y}) = \sum_{i=1}^{n} \mathrm{k}(\mathrm{x}_i)$.

Again
$$\begin{aligned} f(\mathrm{x};\theta) &= g_1(\phi_1[k(\mathrm{x})];\,\theta)\,h_1(\mathrm{x}) \\ &= G_1(k(\mathrm{x});\theta)\,h_1(\mathrm{x}) \end{aligned}$$

say, and
$$\begin{aligned} f(\mathrm{x}_1;\theta)f(\mathrm{x}_2;\theta) &= g_2(\phi_2[\mathrm{k}(\mathrm{x}_1)+\mathrm{k}(\mathrm{x}_2)];\,\theta)\,h_1(\mathrm{x}_1)\,h_1(\mathrm{x}_2) \\ &= G_2(\mathrm{k}(\mathrm{x}_1)+\mathrm{k}(\mathrm{x}_2);\,\theta)\,h_1(\mathrm{x}_1)\,h_1(\mathrm{x}_2) \\ &= G_1(\mathrm{k}(\mathrm{x}_1);\theta)\,G_1(\mathrm{k}(\mathrm{x}_2);\theta)\,h_1(\mathrm{x}_1)\,h_1(\mathrm{x}_2). \end{aligned}$$

The same steps as for Theorem 1 lead to the functional equation

$$\Pi(\mathbf{u}+\mathbf{v};\theta) = \Pi(\mathbf{u};\theta) + \Pi(\mathbf{v};\theta) - \gamma(\theta) \tag{6}$$

where $\mathbf{u}$ and $\mathbf{v}$ belong to some closed neighbourhood N of $\mathbf{0}$. This leads to the multivariate Cauchy equation

$$\phi(\mathbf{u}+\mathbf{v};\theta) = \phi(\mathbf{u};\theta) + \phi(\mathbf{v};\theta) \tag{7}$$

which is continuous at $\mathbf{u} = \mathbf{0}$ and hence everywhere in N. The unique continuous solution to (7) is

$$\phi(\mathbf{u};\theta) = \alpha'(\theta)\,\mathbf{u}$$

(Aczél (1), p. 215) and finally

$$f(\mathrm{x};\theta) = \exp\{\alpha'(\theta)\,\mathrm{k}(\mathrm{x}) + \gamma(\theta) + B(\mathrm{x})\}.$$

4. *Removal of continuity restrictions on $f(x;\theta)$.* The continuity restrictions which have been placed on $f(x;\theta)$ can, if necessary, be lifted in various ways. For instance, in the case of the local result mentioned at the end of section 2, it is sufficient to require that $f(x;\theta)$ be measurable and greater than zero in a neighbourhood of x_0 for all $\theta \in \Omega$.

The proof will be outlined. First, k and its inverse are both measurable functions on a neighbourhood of x_0 and, because of the condition placed on $f(x;\theta)$, this implies the measurability of g_1. Identical argument to that used in Theorem 1 leads to equation (4), where ϕ is measurable on an interval I. A result of Banach (2) can then be used to infer that $f(x,\theta)$ has the continuous exponential form near x_0.

Alternatively, for $f(x;\theta) > 0$ a.e. let $r(x;\theta) = f(x;\theta)/f(x;\theta')$ and suppose $r(x_0;\theta) \neq 0$ and that $r(x;\theta)$ is a.e. bounded in a neighbourhood of x_0 for all $\theta \in \Omega$. Then it is readily verified that g_1 is integrable on an interval I containing zero. The argument leading to (2) then holds a.e. and the technique described in Aczél ((1) p. 190) can be used to give the result.

Other modifications are possible. In particular the assumption of continuity of $g_1(\mathrm{t}_1;\theta)$ in Definition 4(d) can be removed when $q = r = s$ by imposing suitable conditions on $\mathrm{k}(\mathrm{x})$. However, most of the other generalizations seem artificial and they will not be pursued here.

I wish to thank Mr David Culpin and a referee for their suggestions and assistance with this work.

REFERENCES

(1) Aczél, J. *Lectures on functional equations and their applications* (New York; Academic Press, 1966).
(2) Banach, S. Sur l'equation fonctionelle $f(x+y) = f(x)+f(y)$. *Fund. Math.* **1** (1920), 123–124.
(3) Brown, L. Sufficient statistics in the case of independent random variables. *Ann. Math. Statist.* **35** (1964), 1456–1474.
(4) Denny, J. L. Sufficient conditions for a family of probabilities to be exponential. *Proc. Nat. Acad. Sci. U.S.A.* **57** (1967), 1184–1187.
(5) Denny, J. L. Note on a theorem of Dynkin on the dimension of sufficient statistics. *Ann. Math. Statist.* **40** (1969), 1474–1476.
(6) Dynkin, E. B. Necessary and sufficient statistics for a family of probability distributions. *Select. Transl. Math. Statist. Prob.* **1** (1961), 23–41.
(7) Koopman, B. O. On distributions admitting a sufficient statistic. *Trans. Amer. Math. Soc.* **39** (1936), 399.
(8) Pitman, E. J. G. Sufficient statistics and intrinsic accuracy. *Proc. Cambridge Philos. Soc.* **32** (1936), 567.

B

RESEARCH INTO SOME SPECIFIC PROBLEMS IN MEDICINE WITH SPECIAL REFERENCE TO BREAST CANCER

B[1]

Aspects of the Reliability of a Urinary 17-Hydroxycorticosteroid Assay

GORDON SARFATY AND MICHAEL TALLIS

Endocrine Research Unit, Cancer Institute, Melbourne, Victoria; and Division of Mathematical Statistics, C.S.I.R.O., Newtown, N.S.W., Australia

ABSTRACT. The presence of unsuspected random, nonphysiological fluctuations may in certain circumstances reduce the ability of an assay procedure to define differences between individuals or groups of subjects, particularly if the procedure has been initially designed for a different purpose. In the example of urinary 17-hydroxycorticosteroids, the assay can be regarded as implicitly designed to distinguish normals from extremes of hyper- and hypoendocrine function. When the assay is used to define possible differences in a different population, for example, women with breast cancer, previously acceptable criteria may not be adequate for such a purpose. To define sources of variation which could affect the assay's utility for this purpose, measurement was made of the urinary 17-hydroxycorticosteroid excretion for 3 consecutive days in 44 healthy women. Besides assessing the usual chemical criteria of the method reliability and the known sources of variation due to age, height and weight, the magnitude of variation due to differences associated with single 24-hr estimate between and within women was measured. This revealed that approximately 34% of the variance was due to differences within women as compared with 63% due to differences between women. It was concluded that a single 24-hr estimate provides a poor estimate of an individual's true 17-hydroxycorticosteroid excretion. An acceptably reliable estimate of an individual's true mean value, *i.e.*, reduction of within subject variance to approximately 10%, can only be obtained by urine collection for 5 consecutive days. (*J Clin Endocr* 31: 52, 1970)

QUANTITATIVE techniques for the measurements of steroids are often designed with the assumption that they are capable of distinguishing individuals with hyper- and hypoendocrine function from those of a third population, regarded, at least in respect of endocrine disorder, as normal. The index of a measurement's utility for this purpose is generally based on whether or not the assay conforms to standards of chemically determined criteria in respect of accuracy (*i.e.*, without bias), specificity and repeatability (usually equated with precision). In respect of a single 24-hr excretion estimate, these criteria identify *in vitro* variation of the steroid estimated, but ignore the possibility of nonphysiological fluctuations or statistical randomness of the characteristic being measured. Without an awareness of the presence and limits of this potential source of error there could be a significant reduction in the efficiency of detecting real differences between individuals as well as groups.

The breast cancer population can be used as an example of the need to consider undefined statistical randomness in assay reliability. In a group of women with this disease (1) urinary 17-OHCS differed little from the range expected in normal subjects (2–4). An apparent conclusion is that women with breast cancer have no overt abnormality of adrenal function in respect of these metabolites. However, when using the assay to classify breast cancer responders and nonresponders to adrenalectomy, small differences were found between the subgroups (1).

The implication that these two breast cancer populations have differences in adrenal function that can be used to classify individual response, and that the differences are small compared with those of normal, needs to be considered in respect of the ability of the 17-OHCS assay to measure any individual's true value.

The breast cancer group is a specific

Received December 24, 1969.

example of the general problem; it therefore was considered important to reappraise assay reliability in a group of healthy women. This has been done by designing urinary 17-OHCS assays to define *in vitro* aspects of reliability for comparison with other studies, as well as to define a possible random source of subject variability.

Urinary 17-OHCS assays were arranged to obtain the following information:

1. the relative performance of sodium borohydride and sodium periodate used to obtain the final chromogen for the Zimmermann reaction;
2. the usefulness of correcting derived urinary values for losses in the method by use of an external standard;
3. whether concomitant measurement of creatinine is of aid in determining the completeness of the 24-hr urine collection;
4. the magnitude of sources of variability in the assay due to differences between subjects, within subjects, and the error due to assay technique;
5. the influence of the subject's age, height and weight on the 24-hr excretion value.

Results indicated that the component of variance due to differences between women was approximately twice as large as that due to day-to-day differences within women. The latter was sufficiently large to emphasize the importance of analyzing more than a single day's urine sample if a satisfactory estimate is to be made of an individual's true value.

Materials and 17-OHCS Assay Methods

Subjects. The subjects studied were apparently healthy women who agreed to participate in the study. They were obtained through voluntary workers' organizations attached to hospitals and represent a middle range socioeconomic class. To be reasonably certain of their apparent health, each person completed a brief questionnaire. Forty-four were considered to be reasonably healthy and were taking neither oral contraceptives nor other medication at the time of the study. Each person carried out her usual daily activities during the study period. The numbers of women in each decade were: 10–20 yr, 1; 20–30 yr, 8; 30–40 yr, 9; 40–50 yr, 8; 50–60 yr, 13; and 60–70 yr, 5 subjects, respectively.

Urine collections. Three consecutive 24-hr urine collections were obtained from each woman. Urine, collected into polyethylene containers without preservative, was kept in a portable cooler, maintained at approximately 10 C. On receipt in the laboratory, the volume was measured and aliquots taken for immediate creatinine determination. The remainder of the specimen was frozen and maintained at −10 C until assayed for 17-OHCS.

Urinary 17-hydroxycorticosteroids. The method of assay was essentially that described by Wilson and Lipsett (3), the only alteration being in the Zimmermann reaction. This reaction was modified because of difficulties in maintaining the stability of ethanolic KOH in a non-air-conditioned laboratory during hot weather.

Tetramethylammonium hydroxide, 25% w/v (BDH, laboratory reagent grade) was found to be a satisfactory substitute for ethanolic KOH. It had the advantage in being storable at room temperature without deterioration. Using this alkali, the chromogenic ratio of dehydroepiandrosterone[1] (DHA) to 11β-hydroxyetiocholanolone averaged 1.26 (n = 8), comparing favorably the reported value of 1.35 (3) for ethanolic KOH.

The following 17-OHCS assays were made in duplicate: each day of the 3 consecutive days' urine collections; the second of the 3-day collection to which the model steroids DHA and tetrahydrocortisol (THF) were added to measure recovery; distilled water samples in the same urine batch to which the same model steroids were added, in the same amounts, to obtain a comparison of recoveries between urine and water. Amounts of model steroids used ranged from 10 to 100 μg, approximating urine levels of between 3 and 30 mg/24 hr. The actual amount added in any given batch was selected from a table of random numbers.

In 5 subjects, the recovery of tetrahydrocortisol (THF) was measured in each of the 3-days' collections. This latter series of assays was also made in duplicate.

Urinary creatinine. Creatinine was measured with a Technicon AutoAnalyzer in each 24-hr urine sample to determine its usefulness as a means of checking the adequacy of the 24-hr volume.

Results and Statistical Analyses

Urinary 17-hydroxycorticosteroids. Fig. 1 shows a histogram of 17-OHCS values

[1] Trivial names are used for the following steroids: 3α,11β-dihydroxy-5β-androstan-17-one = 11β-hydroxyetiocholanolone; 3α,11β,17β,21-tetrahydroxy-5β-pregnan-20-one = tetrahydrocortisol or THF.

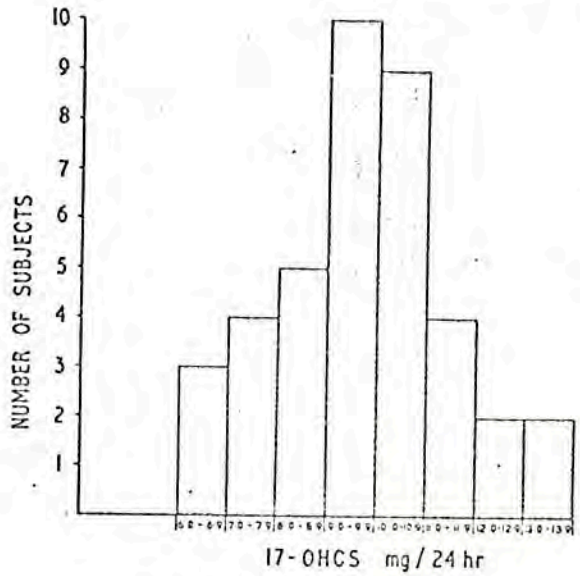


FIG. 1. Urinary values in 44 healthy woman.

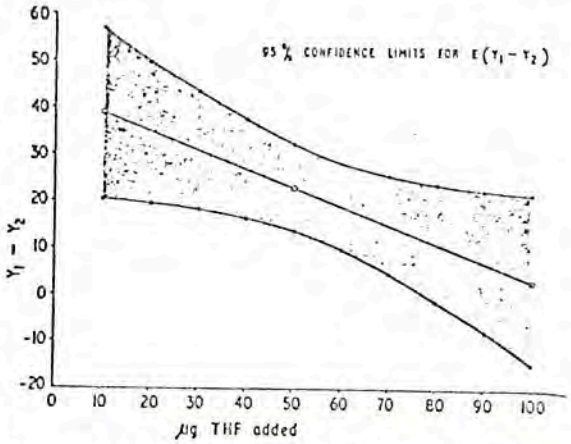


FIG. 2. Estimated regression of recovery of THF from urine (Y_1) and water (Y_2).

constructed from the average of three daily values found in the 44 women. The mean value for all the women was 9.86 mg 24 hr and the range 6.0–13.9.

Recoveries of tetrahydrocortisol and dehydroepiandrosterone. Table 1 shows the values at each recovery level and mean results obtained when THF and DHA were added to aliquots of water and to equal volumes of each of the three successive days' urine collections obtained from 15 of the subjects.

TABLE 1. Per cent recovery of steroids added to urines of different subjects

Amount added (μg)	Tetrahydrocortisol		Dehydroepiandrosterone	
	Urine	Water	Urine	Water
10	73	110	10	0
20	69	97	10	0
20	65	98	5	0
30	73	97	0	0
40	41	113	10	0
50	92	97	6	0
60	80	89	0	0
60	77	99	5	0
60	80	86	3	0
70	80	106	0	1
70	81	92	7	0
70	72	86	4	0
80	88	101	6	0
90	95	102	4	0
100	85	90	3	0
Mean 55.3	76.8	97.5	4.9	0

An inspection of Table 1 reveals a number of pertinent details. First, there was essentially zero recovery of DHA from water, while the recovery of this steroid from urine showed an apparent negative regression of recovery with amount added. This regression was calculated and the slope did not significantly differ from zero. Thus, the average recovery difference of 4.9, together with the calculated confidence interval (2.96, 6.77), seems to provide a satisfactory description of the significantly greater recovery of DHA from urine than from water over a wide range of concentrations.

The recovery of THF, however, presents a different picture. The expected recovery from urine increases significantly with concentration and the differences between the estimated regression equations (for urine Y_1, and water, Y_2) was 42.4–0.39 x where x is the amount added in mg. This line and the 95% confidence bands are illustrated in Fig. 2, the bands being calculated by methods described by Miller (5).

Sources and magnitude of variability in urinary 17-OHCS assay values. In Table 2 are presented complete data for urine volume,

TABLE 2. Data used to determine source and magnitude of variability in 17-OHCS and usefulness of creatinine

Subject	Daily urine (vol ml)	Creatinine (g/24 hr)	17-Hydroxycorticosteroids (mg/24 hr)	
			R	R2
A	940	1.54	7.3	7.3
	1750	1.28	8.5	9.0
	1500	1.20	8.8	7.9
B	1000	1.04	7.1	7.1
	1300	0.91	7.6	8.4
	670	0.89	9.3	8.7
C	820	1.12	9.0	9.0
	880	1.39	11.4	11.7
	700	1.10	8.6	8.8
E	1570	1.10	10.2	9.7
	1415	1.20	13.3	12.8
	1290	0.90	8.4	8.6
F	2025	1.50	9.8	9.2
	1775	1.30	10.9	10.3
	2500	1.40	10.6	11.4
G	900	0.90	13.4	13.4
	550	1.24	8.7	9.6
	455	1.03	8.0	7.7
H	1160	0.93	9.4	7.5
	840	0.69	6.3	6.3
	900	0.88	7.0	6.7
J	600	0.78	5.1	5.2
	1000	1.08	7.1	7.1
	1650	1.25	9.1	9.8
M	1190	1.62	8.9	8.5
	1800	1.71	8.7	8.3
	1630	1.52	9.5	9.5
N	800	1.17	10.4	9.9
	1560	1.37	9.1	10.1
	1030	1.04	9.4	9.4
O	1050	2.48	6.8	7.8
	870	2.04	6.9	7.3
	780	1.85	7.1	7.1
P	740	1.09	5.8	6.2
	720	1.15	7.4	7.4
	815	1.18	6.6	6.8
Q	1260	1.41	13.1	13.1
	1510	1.52	13.7	13.1
	1290	1.73	16.7	15.5
R	2400	0.89	6.2	7.0
	2290	0.92	6.7	6.7
	1610	0.97	7.7	7.7
S	1675	—	10.8	10.8
	1150	—	11.9	11.2
	1590	—	9.9	9.9
T	1510	0.79	5.9	6.4
	1560	0.64	5.1	5.6
	1500	0.90	7.8	7.3
U	1380	0.88	9.0	9.0
	1300	0.86	10.2	9.2
	1100	1.00	11.0	10.6
W	780	1.09	7.1	7.6
	1270	1.18	9.8	10.6
	1010	1.10	8.5	9.2
Y	1180	—	10.3	10.3
	1680	—	9.8	9.8
	1140	—	9.6	9.9

creatinine and 17-OHCS obtained from 19 of the subjects. These data were used to construct analysis of variance tables for 17-OHCS and for creatinine.

Referring to the analysis of variance in Table 3 and on equating the estimated mean squares with their expectations and solving the resulting equations, it is found that

$$\hat{\sigma}_e^2 = 0.1581, \hat{\sigma}_d^2 = (3.4854 - 0.1581)/2 = 1.6636, \text{ and}$$

$$\hat{\sigma}_s^2 = (22.0105 - 3.4854)/6 = 3.0875.$$

Thus, the total variance of a single observation is the sum of the three components, measurement error, between day variations and subject variability. Clearly, the technique error is negligible, but the day-to-day variance contributes appreciably to the total variance. This contribution can be reduced according to the formula

$$\hat{\sigma}_{(n)}^2 = 3.0875 + 1.8217/n$$

where n is the number of days on which measurements are made. This reduction in variance is illustrated in Fig. 3.

An estimate of the repeatability (intraclass correlation) of the daily 17-OHCS measurements is defined by

TABLE 3. Analysis of variance of 17-OHCS

Source	Degrees of freedom	Sums of squares	Estimated squares	Expected mean squares
Between subjects (s)	18	396.1882	22.0105	$\sigma_e^2+2\sigma_d^2+6\sigma_s^2$
Between days within subjects (d)	38	132.4467	3.4854	$\sigma_e^2+2\sigma_d^2$
Error (e)	57	9.0100	0.1581	σ_e^2

$$\hat{\rho}_{17\text{-OHCS}} = \hat{\sigma}_s^2/(\hat{\sigma}_s^2 + \hat{\sigma}_d^2 + \hat{\sigma}_e^2)$$
$$= 3.0875/4.9092 = 0.6289.$$

Usefulness of urinary creatinine measurement as compared with 17-OHCS values. The analysis of variance of the creatinine result is given in Table 4. Proceeding as for the 17-OHCS analysis of variance, the estimates of creatinine variance components are

$$\hat{\sigma}_d^2 = 0.0232 \quad \text{and} \quad \hat{\sigma}_s^2 = 0.1063.$$

Similarly, the estimate of the intraclass correlation is:

$$\hat{\rho}_{\text{creatinine}} = \hat{\sigma}_s^2/(\hat{\sigma}_s^2 + \hat{\sigma}_d^2) = 0.1063/0.1295 = 0.82.$$

This estimate of day-to-day repeatability for creatinine is higher than that for 17-OHCS, and hence creatinine appears to be more reliable in detecting errors in urine collection. However, the difference between the two values may not be sufficiently great to warrant special emphasis being placed on this estimation when measuring consecutive urine samples for total 17-OHCS.

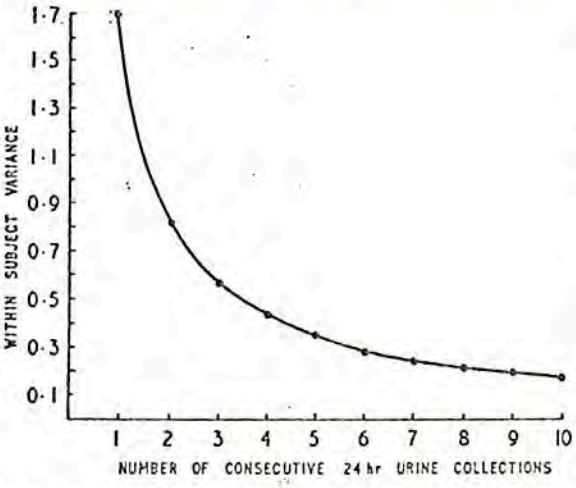


FIG. 3. Reduction of within-subject variance by increasing number of days taken for urine collection.

Correction for method losses. Using the data presented in Table 5, the intraclass correlations for the loss uncorrected and the loss corrected 17-OHCS values were estimated. Although the sample is small, the uncorrected and corrected intraclass correlations were 0.74 and 0.13, respectively, indicating a marked loss of repeatability associated with the correction procedure.

Urinary 17-OHCS, age, height, weight and surface area. In order to establish the relative importance of relationships between 17-OHCS and body weight, height, surface area and age, a multiple regression with 17-OHCS as the dependant variable and the rest as the independant variables was carried out.

From an inspection of the estimated regression coefficients, together with their standard errors, it appeared that age and surface area were important contributors to the total variance of 17-OHCS values and that height and weight were insignificant. The analysis summarized in Table 6 bears this out. It is clear from the results in Table 6 that age is the most important variable, followed by surface area. If these two variables are known, there appears to be no advantage, from a prediction viewpoint, in using either height or weight.

The regression equation using age and surface area was:

$$\text{17-OHCS} = 5.8717 - 0.0883A + 4.8537\,SA$$

and the respective standard errors of the estimated coefficients were 3.0281, 0.0188 and 1.8804. Inspection of the regression

TABLE 4. Analysis of variance of creatinine

Source	Degrees of freedom	Sums of squares	Estimated mean squares	Expected mean squares
Between subjects	16	5.4743	0.3421	$\sigma_d^2+3\sigma_s^2$
Error	34	0.7873	0.0232	σ_d^2

equation also reveals that increasing age is associated with a fall of 17-OHCS values and that increasing surface area is accompanied by increasing values of 17-OHCS.

Discussion

When comparing results obtained from an assay designed to provide an adequate description of individuals in a group, it is important to consider how variables other than those contributed to by the chemical analysis may affect the appraisal.

For example, while emphasis is often placed on detecting, estimating and correcting results for chemical errors due to lack of quantitation, specificity or sensitivity, less consideration is given to incorporating corrections which may arise from well-known variables such as age, sex, body weight and height. It is only when gross evidence of individual variability is obtained, exampled by diurnal variation of plasma 17-OHCS or menstrual fluctuation in estrogens, that attention becomes focused on biologic factors as a potential source of error in assay reliability. Thus, an estimate of plasma 17-OHCS when assessing normal adrenal function can be in error by 100% if diurnal variation is ignored (6). Likewise, an estimation of estrogen content in a single 24-hour urine collection can be in gross error if attempting to define normal cyclic ovarian activity (7).

In the case of urinary 17-OHCS, the assay appears to be chemically reliable when distinguishing the extremes of adrenal function, that is, hyper and hypo from normal individuals (2, 3).

However, in the context of attempting to define differences between women with breast cancer whose urinary 17-OHCS values approximate those of apparently healthy women, usual chemical reliability criteria may be insufficient and this prompted the present reappraisal.

The fact that the mean value reported here derived from the average of a three-consecutive-day analysis corresponded closely with that reported by Wilson and Lipsett (3) (mean 9.86 mg/24 hr *vs.* 9.4 mg/24 hr reported) suggested that no gross discrepancy exists in the performance of the same method in the two different laboratories and also implies that our additional findings are generally valid. This is further supported by the narrower range of our values, 6.0 – 13.9 mg/24 hr *vs.* 5.0 – 18.0 mg/24 hr reported by Wilson and Lipsett, since a three-day average should provide a better estimate than a single day's excretion.

TABLE 5. Urinary 17-OHCS corrected for method losses

Subject	Day	17-OHCS mg/24 hr		% Recovery THF
		Uncorrected	Corrected	
1	1	10.3	13.2	79
	2	9.8	13.0	75
	3	9.8	12.6	78
2	1	13.2	15.0	88
	2	—	—	—
	3	14.8	19.8	75
3	1	7.4	12.8	58
	2	10.2	16.6	61
	3	8.9	15.8	56
4	1	7.0	10.8	65
	2	9.3	14.1	66
	3	9.4	13.4	70
5	1	9.0	11.2	79
	2	9.7	13.6	71
	3	10.8	18.6	58
Intraclass correlation (r)		0.74	0.13	

TABLE 6. Analysis of variance of regression of 17-OHCS *vs.* age, height, body weight and surface area

Source	Degrees of freedom	Sum of squares	Mean squares	F ratio
A	1	48.1867	48.1867	19.2 ($p < .001$)
SA/A	1	17.8039	17.8039	6.95 ($p < .05$)
H, W/A, SA	2	9.3724	4.6862	1.83
Total regression	4	75.3630	18.8407	7.36 ($p < .01$)
Error	36	92.1772	2.5605	

A = age, SA = surface area; H = height; W = weight.

As it was not possible to check the reduction/oxidation characteristics by conventional tracer isotopic techniques, this aspect of reliability was examined by comparing recoveries from urine and water at different levels of added model steroids sufficient to approximate those levels likely to be found in urine. Reagent reliability was revealed by the consistent 100% reduction of DHA and virtually 100% reduction and oxidation of THF when these compounds were added to water. Recoveries from urine being significantly lower than those from water is not surprising and presumably reflects both an inhibition of reduction and oxidation by substances in urine. As all urines tested negatively for glucose prior to analysis, this compound could not be responsible for disturbing the periodate oxidation. An unusual finding, however, was the increasing recovery of THF with increasing amounts added (Table 1; Fig. 2), suggesting whatever inhibition is present is less effective at high THF levels. Although the same trend was present for DHA, the regression was not significant, presumably due to the small differences in the recovery values. The differences in recoveries with increasing amounts added cannot be due to pipetting errors in adding the model compounds as the same pipette was used for additions to both water and urine.

An attempt to improve the assay by correcting for losses measured by recoveries of the external standard was not successful, as this led to a marked loss of repeatability, presumably due to the introduction of a further error involved in the assay of the recovery urine sample.

Because the basic validity of assay quantitation depends on the 24-hour urine volume, an independent means of assessing the correctness of volume by concomitant measurement of another solute, *e.g.*, creatinine, could yield more reliable data. Cramer *et al.* (8) found that fluctuations in urine volume were of the same order as the fluctuations in creatinine and considered no advantage is gained from expressing solutes as a function of creatinine as compared with 24-hour urine volume.

It was also of interest in regard to 24-hour urine volumes to enquire whether perfectly reproducible daily urine volumes could improve the assay. The correlation between urine volume and 17-OHCS was calculated on a within-subject basis and found to be $\hat{\rho} = 0.27$. This indicates that, at best, if urine volume could be perfectly controlled, the within-subject variance of 1.7 would be reduced by $(0.27)^2 \times 1.7 = 0.12$, to 1.58. The difference is insignificant.

Consideration of biologic variations as a source of error revealed age and surface area to be two important parameters contributing to variation in the assay results. The negative regression of 17-OHCS with increasing age was reported by other observers (9) using a comparable (10) assay method. Although Bulbrook *et al.* (11) and Tanner *et al.* (12), using similar procedures, found a positive regression of 17-OHCS with body weight, the present study reveals

surface area to be a more important variable than weight alone.

In the suggestion that urinary 17-OHCS steroid metabolites can fluctuate over long periods of time (9), the contribution of shifts in method reliability has not been clear. The analysis of variance of the 17-OHCS reported here demonstrated the methodology error to be minimal, approximately 3%, and this includes urine collection errors. The within-subject variance, however, was of an unsuspected magnitude, contributing to approximately 34% of the variability in single 24-hour 17-OHCS value. From this evidence it can be seen that a 24-hour value provides a poor estimate of an individual's true 17-OHCS excretion when required for the purposes of attempting to distinguish healthy women from those with breast cancer or, for that matter, any disease in which urinary 17-OHCS values approximate those found in health.

In general, the 24-hour excretion of 17-OHCS by women unassociated with the additional effects of disease changes as a result of a number of influences. First, there is a long-term trend reflecting physiological changes with age. Second, there is a random day-to-day component which may be due to factors such as short-term fluctuations in body weight, in rest-activity cycles, and in day-to-day pyschological interactions with the environment. By taking an average of several days' urine this component is smoothed out to reveal a clearer picture of the true, nonrandom part. In the data reported here, this random source of variability was relatively large.

Other individual, nonrandom cyclic components may exist but these have not been demonstrated in this short-term study. If they are present, it is clearly desirable that measurements be made at the same part of a cycle for each individual. Otherwise, the within-subject variance is increased still further, decreasing reliability and making the classification of individuals into groups more difficult.

Acknowledgments

The cooperation of the Women's Auxillaries of Prince Henry's Hospital, the Peter MacCallum Clinic and their individual volunteers, the assistance of Miss Greta Petersen, and the laboratory facilities provided by Professor Bryan Hudson, are acknowledged.

References

1. Bulbrook, R. D., F. C. Greenwood, and J. L. Hayward, *Lancet* 1: 1154, 1960.
2. Few, J. D., *J Endocr* 22: 31, 1961.
3. Wilson, H., and M. B. Lipsett, *Anal Biochem* 5: 217, 1963.
4. Metcalf, M. G., *J Endocr* 26: 415, 1963.
5. Miller, R. G., Simultaneous Statistical Inference, McGraw-Hill, New York, 1966.
6. Brown, H., E. Engbert, S. Wallach, and E. L. Simmons, *J Clin Endocr* 17: 1191, 1957.
7. Brown, J. B., *J Obstet Gynaec Brit Emp* 66: 795, 1959.
8. Cramer, K., H. Cramer, and S. Selander, *Clin Chim Acta* 55: 331, 1967.
9. Borth, R., A. Linder, and A. Riondel, *Acta Endocr (Kobenhavn)* 25: 33, 1957.
10. Thomas, B. S., *J Clin Endocr* 25: 710, 1965.
11. Bulbrook, R. D., *In* Hayward, J. L., and R. D. Bulbrook (eds.), Clinical Evaluation of Breast Cancer, Academic Press, London, 1966, p. 77.
12. Tanner, J. M., M. J. R. Healey, R. H. Whitehouse, and A. C. Edgson, *J Endocr* 19: 87, 1959.

B [2]

Biometrika (1969), 56, 3, *p.* 505
Printed in Great Britain

Note on a calibration problem

By G. M. TALLIS
C.S.I.R.O., Newtown, New South Wales

Summary

The problem of obtaining a satisfactory estimate of a variable X from another variable Y, where X and Y have joint frequency function $\phi(x, y; \theta)$, is considered under the restriction that only Y and $Y|X = x$ can be observed. This raises the question as to whether or not ϕ can be determined from the frequency functions of Y and $Y|X = x$. It is found that the latter problem is related to the theory of the identifiability of mixtures of distributions. The information associated with Y and the bivariate viewpoint are used for a fresh approach to a problem of calibration. The standard technique is examined in the light of the new approach.

1. The problem

In this note an aspect of the following problem is considered. Individuals from a fixed population are taken with a view to obtaining information with regard to a variable X which is associated with each member of the population. Unfortunately, X cannot be observed directly and, instead, an associated variable Y is measured. It is assumed that approximately: (*a*) by repeated sampling, the marginal distribution of Y can be determined; and (*b*) the conditional distribution of $Y|x$ can be found by suitable experimentation.

A typical example of this situation concerns the measurement of certain chemical compounds, X, in the blood or urine of patients. The amount of compound cannot be directly observed but, instead, after suitable preparation, a reading is made on a machine, Y, which has been standardized against known amounts of the substance being measured. From the machine reading and the standard curve it is required to estimate X. For the purposes of this note this type of situation will be referred to as the calibration problem.

The theoretical question of deciding when a satisfactory prediction of X can be made given the marginal distribution of Y and the conditional distribution of $Y|x$ is discussed below. In applications, however, the technique requires that the population which is being sampled remains fixed. In practice this assumption may be hard to justify and a frequent check on the distributions of Y and $Y|x$ may be necessary. Nevertheless, this is hardly an indictment of the method since the standard approach to this problem provides an estimator for Y which itself must be checked repeatedly in many cases. When the correlation between X and Y is low, the old method can be very inefficient.

Parametric forms of the relevant frequency functions will subsequently be assumed. In fact, the joint frequency function for the random vector (X, Y) will be written as

$$\phi(x, y; \theta) = h(y; \delta) f(x|y; \pi) = k(x; \rho)\, g(y|x; \gamma).$$

In the context of the previous paragraphs, it is assumed that $h(y; \delta)$ and $g(y|x; \gamma)$ are known, i.e. δ and γ are known. If in the conditional frequency function $f(x|y; \pi)$, π is some function of δ and γ, $\pi = \psi(\delta, \gamma)$, the calibration problem will be said to be determinate. This simply

means that π and hence f is determined once δ and γ are known. Since

$$f(x|y;\pi) = k(x;\rho)\,g(y|x;\gamma)/h(y;\delta),$$

it is seen that the calibration problem is determinate if ρ is some function of γ and δ.

Estimation proceeds in two stages. First, from a random sample from the population under study, δ is estimated, by $\hat{\delta}$ say, from n observations made on Y. Secondly, a calibration experiment is run using fixed and known values of $X = x$, and a further m Y readings are obtained. From the pairs (x_i, y_i) $(i = 1, \ldots, m)$, γ is estimated, by $\hat{\gamma}$ say, leading to an estimator for π, $\hat{\pi} = \psi(\hat{\delta}, \hat{\gamma})$.

Thus the difference between this and the classical approach to the problem is that additional information is introduced in connexion with the distribution of Y. However, even if h and g are known, this does not necessarily guarantee that the conditional distribution of $X|y$ can be obtained. Only if the problem is 'determinate' will a solution be possible in this way.

The frequency function $f(x|y;\pi)$ provides the maximum information with regard to X for an observed y. Moreover, if $r(y) = E\{X|y\}$, then $r(y)$ estimates X with minimum mean square error and hence provides a suitable estimator.

2. Example

Consider the case where X and Y are jointly distributed in the bivariate normal distribution $N(\mu_x, \mu_y, \sigma_x^2, \sigma_y^2, \sigma_{xy})$; then the calibration problem is determinate. In this instance $\delta' = (\mu_y, \sigma_y^2)$ and $\gamma' = (\alpha_1, \beta_1, \sigma_1^2)$, since Y is distributed as $N(\mu_y, \sigma_y^2)$ and $Y|X = x$ is distributed as $N(\alpha_1 + \beta_1 x, \sigma_1^2)$. It is found after some elementary algebra that $\pi' = (\alpha_2, \beta_2, \sigma_2^2)$, where

$$\alpha_2 = \beta_1^{-1}(\mu_y \sigma_1^2/\sigma_y^2 - \alpha_1),$$
$$\beta_2 = (\sigma_y^2 - \sigma_1^2)\,(\beta_1 \sigma_y^2)^{-1},$$
$$\sigma_2^2 = \sigma_1^2(\sigma_y^2 - \sigma_1^2)\,(\beta_1^2 \sigma_y^2)^{-1}.$$

Thus the frequency function for $X|Y = y$ can be found from a knowledge of h and g. Estimates of α_2, β_2 and σ_2^2 are obtained by replacing α_1, β_1, σ_1^2, μ_y and σ_y^2 by their usual estimates. The properties of $\hat{\pi}$ will not be discussed here since they digress from the main theme of this note.

3. Determinancy as a special case of the identifiability of mixtures

It is interesting to look at the problem in a slightly different way. Since

$$h(y) = \int_{-\infty}^{\infty} g(y|x)\,k(x)\,dx \tag{1}$$

and h and g are assumed known, the determinacy of the calibration problem is equivalent to finding a unique solution to the above Fredholm integral equation of the first kind. This is a special case of the problem of the identifiability of mixtures of distributions.

An immediate and useful result can be obtained from the above observation. If $g(y|x)$ is $N(\alpha + \beta x, \sigma^2)$ then the calibration problem is determinate, since by taking the Fourier transform of both sides of (1) and rearranging,

$$k^*(\theta) = h^*(\theta/\beta)\exp(-i\theta\alpha/\beta + \theta^2\sigma^2/\beta^2). \tag{2}$$

In (2), k^* and h^* are the Fourier transforms of k and h respectively. Since k is uniquely determined by k^*, the result follows. The example of § 2 is a special case of this proposition.

Examples of nondeterminacy are easily constructed. Let

$$h(y) = \begin{cases} \frac{1}{2} & (y=0), \\ 0 & (y \neq 0, 1), \\ \frac{1}{2} & (y=1), \end{cases} \qquad g(y|x) = \begin{cases} x & (y=0), \\ 0 & (y \neq 0, 1), \\ 1-x & (y=1), \end{cases}$$

then

$$f(x|y) = \begin{cases} 2x\,k(x) & (y=0), \\ 0 & (y \neq 0, 1), \\ 2(1-x)\,k(x) & (y=1). \end{cases}$$

Clearly $k(x)$ is determined only up to the first moment, i.e.

$$\int_0^1 xk(x)\,dx = \tfrac{1}{2}.$$

4. Discussion

The results of § 3 can be used to discuss certain practical situations. For instance, in the context of the example of § 1, it may be necessary to estimate the particular chemical compound in male and female patients. In this case two distinct populations may be involved. Now, if it is reasonable to assume that the characteristics of the measuring apparatus are described by $g(y|x;\gamma) = N(\alpha+\beta x, \sigma^2)$ and that the performance does not change during the course of the experiment, then, in the representation $\phi(x, y;\theta) = k(x;\rho)\,(y|x;\gamma)$, ρ is some function of γ and δ. This follows from the determinacy of the calibration problem. Thus, the only way in which a change in the joint frequency function of X and Y can be achieved is through a change in $h(y;\delta)$, the frequency function for Y.

Under the above conditions, therefore, in order to obtain a satisfactory estimate of $f(x|y;\pi)$ for both the male and female populations, only a single estimate of α and β from the machine is required. However, $h(y;\delta)$ must be estimated separately for the two populations so that appropriate estimates of $f(x|y;\pi)$ can be obtained.

It is interesting to examine a special case of the calibration problem discussed by Graybill (1961, p. 125) where part of the sampling is carried out in a population with $X = x_0$ fixed but unknown. Let $k(x) = \delta(x_0)$, where $\delta(.)$ is the Dirac function. Then

$$h(y) = \int_{-\infty}^{\infty} \delta(x_0)\,g(y|x)\,dx = g(y|x_0).$$

Clearly, x_0 can be considered as one of the parameters to be estimated in the conditional frequency function g. In fact, here $\delta' = [\gamma', x_0]$ and δ can be estimated from formulae given by Graybill. Therefore, this particular situation can be handled without any preliminary calibration experiment.

One current method of approaching the calibration problem in the case that $E(Y|x) = r(x)$ is linear amounts to equating $r(x) = \alpha_1 + \beta_1 x$ to an observed $Y = y$ and solving the resulting equation for x (Bennett & Franklin, 1954, p. 231). This leads to the estimator

$$\hat{x}_0 = (y-\alpha_1)/\beta_1 = \mu_x + (y-\mu_y)/\beta_1,$$

where α_1, β_1, μ_x and μ_y are replaced by suitable estimates. However, the estimator for X proposed here is

$$\hat{x} = \mu_x + (y-\mu_y)\,(\beta_1\sigma_x^2/\sigma_y^2)$$

and $\beta_1^{-1} = \beta_1\sigma_x^2/\sigma_y^2$ if and only if $|\rho| = 1$. Therefore, the old procedure amounts to using,

or using an estimate of, the wrong regression line and could lead for $|\rho| \ll 1$ to a serious loss of efficiency in terms of mean square error. The results of a simulation study emphasizing this point will be reported elsewhere.

I am indebted to Professor E. J. Williams for his comments and especially for pointing out the equivalence of determinacy and identifiability. I am also grateful to Mr H. Weiler for a number of interesting discussions.

REFERENCES

BENNETT, C. A. & FRANKLIN, N. L. (1954). *Statistical Analysis in Chemistry and the Chemical Industry*. New York: Wiley.

GRAYBILL, A. F. (1961). *An Introduction to Linear Statistical Models*. New York: McGraw-Hill.

[*Received December* 1968. *Revised May* 1969]

Metrika

Internationale Zeitschrift für theoretische und angewandte Statistik
International Journal for Theoretical and Applied Statistics

Sonderdruck aus:
Volumen 15
1970
Fasc. 1/2

Physica-Verlag · Wien - Würzburg

Some Extensions of Discriminant Function Analysis

By G. M. TALLIS, Newtown [1])

I. Introduction

The concept of discriminant function analysis was first introduced by R. A. FISHER as a tool to assist in the taxonomic classification of several species of Iris. This technique has undergone considerable extension and general discussions on the topic are given in RAO, ANDERSON and elsewhere.

It is the purpose of this paper to extend the ideas of discriminant analysis to a stochastic process $X(t)$. The large sample dispersion matrix of the estimated discriminant function is developed and is used to obtain revised classification regions and errors of misclassification. The modified analysis which includes the errors of estimation is illustrated by means of Fisher's classical example cited above. It is found that the consideration of sampling errors may be important in any discriminant analysis.

The extensions discussed here were initiated by investigations into human mammary carcinoma. Numerous workers have found a relationship between the levels of certain hormones excreted in the urine of women in advanced stages of breast cancer and their responsiveness to endocrine ablation. In fact, by using two of these hormone levels in a discriminant analysis it has been possible to identify a group of women with lowered responsiveness to this sort of treatment. It was felt that some accuracy of the analysis was lost by not taking account of the marked regression of these variables on age. Moreover, from a statistical viewpoint, this important problem does not seem to have received due attention and this paper attempts to correct this deficiency.

II. Results

We consider the n-dimensional vector $X(t)$ with multivariate normal distribution function $N(\mu(t), V)$. It is assumed that $X(t)$ belongs to one of two populations, Π_1 or Π_2, where the members of Π_i follow the normal distribution

[1]) G. M. TALLIS, Division of Mathematical Statistics, C.S.I.R.O., 60 King St., Newtown, N.S.W., 2042.

function $N(\mu_i(t), V)$ $i = 1,2$. The problem is, given a particular vector $x(t)$, how do we best assign $x(t)$ to either Π_1 or Π_2?

This is the well known problem of discriminant analysis with the modification of considering X as a stochastic process i.e. a function of time t. The relevant theory is discussed in ANDERSON and we list some of the results here for completeness and to establish notation.

Let $C(i|j)$ be the cost of assigning $x(t)$ to Π_i when it actually comes from Π_j, $i \neq j$. Moreover, if R_1 and R_2 are regions of n-space E_n,

$$R_1 \cup R_2 = E_n,$$
$$R_1 \cap R_2 = \Phi,$$

such that if $x(t) \in R_i$, we assign $x(t)$ to Π_i and if Π_i has prior probability p_i, $p_1 + p_2 = 1$, then the expected cost of misclassification is

$$E(C) = C(2|1)p_1 \int_{R_2} \mathrm{d}N(\mu_1(t), V) + C(1|2)p_2 \int_{R_1} \mathrm{d}N(\mu_2(t), V). \tag{1}$$

Using the above notation we have the following theorem the proof of which may be found in ANDERSON.

Theorem

The regions of classification, R_1 and R_2, which minimise $E(C)$ for fixed t are specified by

$$R_1: [X(t) - \tfrac{1}{2}(\mu_1(t) + \mu_2(t))]' V^{-1} [\mu_1(t) - \mu_2(t)] \geqslant \log k$$
$$R_2: [X(t) - \tfrac{1}{2}(\mu_1(t) + \mu_2(t))]' V^{-1} [\mu_1(t) - \mu_2(t)] < \log k$$

where $k = p_2 C(1|2)[p_1 C(2|1)]^{-1}$.

Of course, the constant k is usually unknown and is therefore put equal to one. Thus, if the expression on the left hand side is negative, $x(t)$ is assigned to Π_2 and if it is positive, $x(t)$ is classified as belonging to Π_1. If $C(1|2) = C(2|1)$ then the above scheme minimises the expected proportion of misclassification.

It turns out subsequently that it is convenient to work with

$$D(t) = X'(t) V^{-1} (\mu_1(t) - \mu_2(t)) = X'(t) l(t) \tag{2}$$

which will be called the discriminant function. Notice that $D(t)$ is normally distributed with mean $\beta_i(t) = l'(t)\mu_i(t)$ and variance $\sigma_1^2(t) = \sigma_2^2(t) = l'(t) V l(t)$. Using the minimax principle, we now find a $b(t)$ such that

$$C(2|1)p_1 \int_{b(t)}^{\infty} \mathrm{d}N(\beta_1(t), \sigma^2(t)) = C(1|2)p_2 \int_{-\infty}^{b(t)} \mathrm{d}N(\beta_2(t), \sigma^2(t))$$

and

$$R_1: x'(t) l(t) \geqslant b(t)$$
$$R_2: x'(t) l(t) < b(t).$$

In case $C(2|1)p_1 = C(1|2)p_2$, $b(t) = \frac{1}{2}(\beta_1(t) + \beta_2(t))$ and we have the situation of the discussion following the theorem.

Unfortunately, in practice neither V nor $\mu_i(t)$ are known and they must be estimated from suitable data. In order to make further progress we must assume some parametric form for $\mu_i(t) = [\mu_i^{(1)}(t), \ldots, \mu_i^{(n)}(t)]$. Let $\Gamma' = [1, t, \ldots, t^{k-1}]$, then we will assume that $\mu_i(t) = B_i \Gamma$, where B_i is an $(n \times k)$ matrix of unknown constants. If S, $\hat{B}_1$ and $\hat{B}_2$ are estimates of V, B_1 and B_2 respectively, then the estimated index value is

$$\hat{D}(t) = X'(t) S^{-1} \hat{d}(t) \tag{3}$$

where $\hat{d}(t) = (\hat{B}_1 - \hat{B}_2)\Gamma$.

Now suppose a random sample of n_i vectors are taken from Π_i and is denoted by $Y_i = [y_i^{(1)}, \ldots, y_i^{(n_i)}]$ where $y_i^{(j)}$ is an $(n \times 1)$ vector and hence Y_i is $(n \times n_i)$. Then an unbiased estimate of B_i is given by

$$\hat{B}_i = C_i A_i^{-1}$$

where

$$A_i = T_i T_i', \quad C_i = Y_i T_i', \quad T_i = [\Gamma_i^{(1)}, \ldots, \Gamma_i^{(n_i)}].$$

The obtain a pooled and unbiased estimate of V set

$$S = \sum_{i=1}^{2} (Y_i Y_i' - \hat{B}_i A_i \hat{B}_i')/(n_1 + n_2 - 2k).$$

It is well known that the matrices $\hat{B}_1$, $\hat{B}_2$ and S are independently distributed. In fact, $(n_1 + n_2 - 2k)S$ has a WISHART distribution with parameters $(n_1 + n_2 - k)$ and V and if $\hat{B}_i$ is rolled out into a vector of dimension $(nk \times 1)$, $\hat{\beta}_i' = [\beta_i^{(1)'}, \ldots \beta_i^{(n)'}]$, then $\hat{\beta}_i$ is distributed as $N(\beta_i, V \otimes A_i^{-1})$.

Before proceeding we notice that (3) can be written in somewhat different form. Let $Z(t) = I_n \otimes \Gamma'$, then (3) becomes

$$\hat{D}(t) = X'(t) S^{-1} Z(t)(\hat{\beta}_1 - \hat{\beta}_2).$$

This representation of $\hat{D}(t)$ facilitates the following derivation.

Write $S = V + \Delta V$ and $\hat{d}(t) = d(t) + \Delta d(t)$ where $E(\Delta V) = 0$ and $E(\Delta d(t)) = 0$, then

$$\begin{aligned} \hat{\mathfrak{L}}(t) &= (V + \Delta V)^{-1}(d(t) + \Delta d(t)) \\ &= \sum_{n=0}^{\infty} (V^{-1} \Delta V)^n V^{-1}(d(t) + \Delta d(t)). \end{aligned}$$

Neglecting terms of $0(N^{-3/2})$, $N = n_1 + n_2$, we have

$$E(\hat{\mathfrak{L}}(t)) = V^{-1} d(t) + E(V^{-1} \Delta V V^{-1} \Delta V V^{-1} d(t)).$$

In order to evaluate the expectation on the right of the above equation $\Delta V V^{-1} \Delta V = A$, say, must be written out as follows:

$$A = [a_{ij}] = \left[\sum_s \sum_r \Delta v_{ir} v_{rs}^{-1} \Delta v_{sj} \right]$$

and

$$E[A] = \left[\sum_s \sum_r v_{rs}^{-1} \operatorname{cov}(\hat{v}_{ir}, \hat{v}_{sj}) \right]$$

where v_{rs}^{-1} is the r,s^{th} element of V^{-1}. If we substitute

$$\text{cov}\,(\hat{v}_{ir}, \hat{v}_{sj}) = (N-2k)^{-1}(v_{is}v_{rj} + v_{ij}v_{rs})$$

in the above expression and carry out some simplification we obtain

$$E(A) = (N-2k)^{-1}(1+n)V$$

and finally

$$E\{\hat{\mathfrak{L}}(t)\} = [1 + (N-2k)^{-1}(1+n)]\mathfrak{L}(t). \tag{5}$$

From (5) it is clear that the bias in $\hat{\mathfrak{L}}(t)$ would be small for N reasonably large and probably not worth correcting for. Moreover, to the order of accuracy required by the subsequent expansions, $E\{\hat{\mathfrak{L}}(t)\}$ can be replaced by $\mathfrak{L}(t)$ without altering the final results.

Now

$$S\mathfrak{L}(t) = \hat{d}(t) = (V + \Delta V)(\mathfrak{L}(t) + \Delta\mathfrak{L}(t)) = (d(t) + \Delta d(t))$$

whence

$$\Delta\mathfrak{L}(t) = V^{-1}(\Delta d(t) - \Delta V\mathfrak{L}(t))$$

neglecting the term $\Delta V\Delta\mathfrak{L}(t)$ which is $0(N^{-1})$.

Therefore

$$\begin{aligned} L(t) &= [\text{cov}\,(\mathfrak{L}_i(t), \mathfrak{L}_j(t))] = E[\Delta\mathfrak{L}(t)\Delta\mathfrak{L}'(t)] \\ &= V^{-1}[E(\Delta d(t)\Delta d'(t)) + E(\Delta V\mathfrak{L}(t)\mathfrak{L}'(t)\Delta V)]V^{-1} \end{aligned}$$

the remaining terms dropping out because ΔV and $\Delta d(t)$ are independently distributed. Since $\hat{d}(t)$ is distributed as $N(d(t), \Gamma'(A_1^{-1} + A_2^{-1})\Gamma V)$, the first term of L is $\Gamma'(A_1^{-1} + A_2^{-1})\Gamma V^{-1}$.

In order to evaluate the second term write

$$C(t) = \mathfrak{L}(t)\mathfrak{L}'(t), \text{ and}$$
$$M(t) = \Delta VC(t)\Delta V,$$

then

$$M(t) = [m_{qr}(t)] = \left[\sum_s\sum_t C_{st}(t)\Delta v_{tr}\Delta v_{qs}\right]$$

and

$$E[M(t)] = \left[\sum_s\sum_t C_{st}(t)\,\text{cov}\,(\hat{v}_{tr}, \hat{v}_{qs})\right] = (N-2k)^{-1}[VC(t)V + V\ \text{trace}\ VC(t)]$$

after substituting for $\text{cov}\,(\hat{v}_{tr}, \hat{v}_{qs})$ and simplifying. Thus

$$L(t) = \Gamma'(A_1^{-1} + A_2^{-1})\Gamma V^{-1} + (N-2k)^{-1}\,V^{-1}[d(t)d'(t)V^{-1} + Id'(t)V^{-1}d(t)].$$

We now calculate $\text{var}\,(\hat{D}(t))$ after noticing that

$$E\{(\hat{D}(t))\} = E[\hat{\mathfrak{L}}'(t)X(t)] = \mathfrak{L}'(t)\mu_i(t),$$

to $O(N^{-1})$, if $X(t) \in \Pi_i$. By definition

$$\text{Var}\,(\hat{D}(t) = E[\hat{\mathfrak{L}}'(t)X(t) - \mathfrak{L}'(t)\mu_i(t)]^2$$

and, writing

$$\hat{\mathfrak{L}}'(t)X(t) = (\mathfrak{L}(t) + \Delta\mathfrak{L}(t))'(\mu_i(t) + \Delta X(t))$$

we obtain finally

$$\mathrm{Var}(\hat{D}(t)) = \mu_i'(t) L(t) \mu_i(t) + \ell'(t) V \ell(t) + E[\Delta \ell'(t) \Delta X(t) \Delta X'(t) \Delta \ell(t)].$$

The expectation in the above equation is clearly equal to trace $L(t)V$ and the final expression is

$$\begin{aligned} \mathrm{Var}(\hat{D}(t)) = {} & d'(t) V^{-1} d(t)[1 + (n+1)/(N-2k)] \\ & + n\Gamma'(A_1^{-1} + A_2^{-1})\Gamma + \mu_i'(t) L(t) \mu_i(t). \end{aligned} \tag{6}$$

The expression for var $(\hat{D}(t))$ can be compared with

$$\sigma^2(t) = \ell'(t) V \ell(t) = d'(t) V^{-1} d(t).$$

It is clear that the additional terms of $O(N^{-1})$ may be appreciable and that var $(\hat{D}(t))$ may be considerably greater than σ^2. Moreover, var $(\hat{D}(t))$ depends on Π_i, σ_i^2 say, and hence the value of $b(t)$ should be computed from the equation

$$C(2|1) p_1 \int_{b(t)}^{\infty} dN(\beta_1(t), \sigma_1^2(t)) = C(1|2) p_2 \int_{-\infty}^{b(t)} dN(\beta_2(t), \sigma_2^2(t)).$$

In particular when $C(2|1)p_1 = C(1|2)p_2$

$$b(t) = (\sigma_2(t)\beta_1(t) + \sigma_1(t)\beta_2(t))/(\sigma_1(t) + \sigma_2(t)).$$

In order to decide initially whether or not Π_1, and Π_2 are distinct populations, it is important to test the null hypothesis $\beta_1 - \beta_2 = 0$. Let

$$T^2 = (\hat{\beta}_1 - \hat{\beta}_2)' [S^{-1} \otimes (A_1^{-1} + A_2^{-1})^{-1}](\hat{\beta}_1 - \hat{\beta}_2), \tag{7}$$

then, under the null hypothesis, T^2 has a central T^2 distribution with parameters, nk and $(N - 2k - nk + 1)$.

The above results simplify in the classical case when $k = 1$, i.e. when the Π_i distribution is $(N(\mu_i, V)$. It is found in these circumstances that

$$\hat{\ell} = S^{-1} d, \; d = (\bar{x}_1 - \bar{x}_2),$$

$$S = (N-2)^{-1} \left[\sum_{j=1}^{n_1} (x_{1j} - \bar{x}_1)(x_{1j} - \bar{x}_1)' + \sum_{j=1}^{n_2} (x_{2j} - \bar{x}_2)(x_{2j} - \bar{x}_2)' \right]$$

where the notation is obvious. The expectation $E\hat{\ell})$ can be found from (5) by setting $k = 1$ and suppressing the index t, while L takes the form

$$L = (N-2)^{-1} V^{-1} [dd' V^{-1} + I d' V^{-1} d] + \frac{N}{n_1 n_2} V^{-1}.$$

The above formula simplifies still further if $n_1 = n_2 = m$ to

$$L \simeq V^{-1} (2m)^{-1} (4I + dd' V^{-1} + I d' V^{-1} d].$$

The variance of $\hat{D}$ turns out to be

$$\mathrm{Var}(\hat{D}) = d' V^{-1} d[1 + (n+1)/(N-2)] + nN/n_1 n_2 + \mu_i' L \mu_i$$

and the test of the null hypothesis $\mu_1 - \mu_2 = 0$ can be made with the statistic

$$T^2 = \frac{n_1 n_2}{N}(\bar{x}_1 - \bar{x}_2)' S^{-1}(\bar{x}_1 - \bar{x}_2)$$

which has the central T^2 distribution with parameters n and $(N-n-1)$.

Example

In order to illustrate the above results the Iris data of Fisher as recorded in RAO [page 248] will be used. Thus, $n = 4$ and $k = 1$ and it is found that $\hat{\xi}' = [-3.0528, -18.0230, 21.7662, 30.8442]$ and the estimated mean of the discriminant function in Π_1 and Π_2 are 65.5785 and -37.6550 respectively.

Under the standard analysis of FISHER the common standard deviation of $\hat{D}$ is 10.1604. The boundary value b assuming $C(2|1)p_1 = C(1|2)p_2$ is 13.9618 and the probability of misclassification is

$$Pr\{\hat{D} > 13.9618 | \hat{D} \sim N(-37.655, 103.2335)\} = 10^{-7} \times 1.88 .$$

By the methods suggested in this paper, $\hat{\xi}$, of course, remains unaltered but the variance of $\hat{D}$ in Π_1 and Π_2 are 372.2685 and 312.3221 respectively. The revised boundary value is 11.6974 and the revised probability of misclassification is

$$Pr(\hat{D} > 11.6974 | \hat{D} \sim N(-37.655, 313.2331)\} = 10^{-3} \times 2.61 .$$

Although in this example the probability of misclassification is still small under the revised analysis, it is considerably larger than that given by the standard analysis. It appears, therefore, that to obtain even rough approximations to probabilities of misclassification, at least all terms of $0(N^{-1})$ should be included in the analysis.

A Fortran programme has been written to cover the general analysis reported above. Aspects of this programme, further extensions and numerical examples based on extensive medical data will be reported elsewhere.

III. Acknowledgement

I wish to thank Mr. George BROWN for checking the formulae for $k = 1$. D. Shaw simplified two of the results and performed the computations of the example. I am also indebted to Dr. Gordon SARFATY for raising some of the questions considered here.

References

ANDERSON, T. W.: Introduction to Multivariate Statistical Analysis, New York, 1958.

FISHER, R. A.: The use of multiple measurements in taxonomic problems. Ann. Eugen. 7, 179—188, 1936.

RAO, C. R.: Advanced Statistical Methods in Biometric Research. New York, 1952.

B[4]

Discriminant Analysis By Computer

By G. M. Tallis*, D. E. Shaw*, J. Williams* and G. Sarfaty#

1 INTRODUCTION

In a recent paper, Tallis (1969) [4], Fisher's discriminant function was re-examined from two points of view. Firstly, extensions were made to allow the random n-vector X characterizing individuals to be a stochastic process, X(t). Specifically, X(t) was assumed to be distributed multinormally with mean vector $\underset{\sim}{\mu}_i(t)$ and co-variance matrix V [i.e. $X(t) \sim N(\underset{\sim}{\mu}_i(t), V)$], if X(t) belongs to population Π_i, i = 1,2. A polynomial form for $\underset{\sim}{\mu}_i(t)$ was achieved by setting $\underset{\sim}{\mu}_i(t) = B_i\Gamma$, where B_i is a matrix of coefficients and $\Gamma' = [1,t, \ldots , t^{k-1}]$. This approach allows, for instance, discriminant analysis to be applied where the members of the two populations are of different ages, the mean thus being a function of the individual's age. In this case, of course, each age has its own special discriminant function. The standard analysis is retrieved by setting k = 1.

The second extension was to consider the effect of sampling errors on the estimated discriminant function and on associated errors of misclassification. The large sample covariance matrix, L(t), for the estimated coefficients of the discriminant, $\underset{\sim}{l}(t)$, was obtained and it was shown that, in the case of Fisher's famous Iris example, it appeared to be important to consider the sampling variance of $\underset{\sim}{l}(t)$,when calculating the probabilities of misclassification.

The full analysis of $k > 1$ is computationally strenuous. Therefore, a FORTRAN programme, DISCRIM, has been written to provide an output of all useful steps in the modified discriminant analysis. This paper discusses DISCRIM and indicates further extensions to the work in [4].

11 INDIVIDUAL PROBABILITIES

A brief sketch of results additional to those discussed in [4] will now be presented. Full details are omitted.

Let $X(t) \sim N(\underset{\sim}{\mu}_i(t),V)$ if X(t) belongs to Π_i, i = 1,2, and suppose the proportion of Π_i in the whole population is p_i ($p_1 + p_2 = 1$). Then if $f_i(x,t)$ is the frequency function corresponding to $N(\underset{\sim}{\mu}_i(t),V)$, the frequency function for the total population is given by the mixture

$$f(x,t) = p_1 f_1(x,t) + p_2 f_2(x,t).$$

The conditional probability that an individual belongs to Π_1 given X(t) = x is $p_1 f_1(x,t)/f(xt)$. If the algebra is carried out this becomes

$$[1 + \exp \{ \ln(p_2/p_1) + D(t) - A(t) \}]^{-1} = P_1(D,t) \text{ say.}$$

where

$$D(t) = X'(t) \ \underset{\sim}{l}(:),$$

$$\underset{\sim}{l}(t) = V^{-1}[\underset{\sim}{\mu}_2(t) - \underset{\sim}{\mu}_1(t)]$$

and

$$A(t) = \tfrac{1}{2}[\underset{\sim}{\mu}_2(t) + \underset{\sim}{\mu}_1(t)]'V^{-1}[\underset{\sim}{\mu}_2(t) - \underset{\sim}{\mu}_1(t)].$$

Now

$$\text{Var } \{ D(t) \} = \sigma^2(t) = \underset{\sim}{l}'(t)\underset{\sim}{V}^{-1}\underset{\sim}{l}(t)$$

and we introduce

$$D'(t) = [D(t) - A(t)]/\sigma(t)$$

which will be called the standardised discriminant. The graphs of

$$P_1(D',t) = [1 + \exp \{ \ln(p_2/p_1) + \sigma(t)D'(t) \}]^{-1}$$

against D' (t) for various values of t give a comparison of the performance of the discriminant function for different values of the time parameter.

The above analysis is for the case where the parameters of $f_i(x,t)$, i = 1,2, are assumed to be known. When estimates are used the procedure of [4] is appropriate. In this case it can reasonably be assumed that the estimate of D(t), $\hat{D}(t)$, obtained from [4] equation (3) is approximately normally distributed in Π_i with mean and variance estimable from formulae in [4]. As for D(t),

$$\hat{D}'(t) = [\hat{D}(t) - \hat{A}(t)]/\hat{\sigma}(t)$$

is a more or less standardised version of $\hat{D}(t)$. The expressions $\hat{A}(t)$ and $\hat{\sigma}(t)$ indicate estimates of A(t) and $\sigma(t)$ based on a sample.

In Π_i, the frequency function for $\hat{D}'(t)$, $h_i(\hat{D}',t)$ say, can be obtained from that of $\hat{D}(t)$. Then, the conditional probability analogous to $P_1(D',t)$ is

$$Q_1(\hat{D}',t) = p_1 h_1(\hat{D}',t) / [p_1 h_1(\hat{D}',t) + p_2 h_2(\hat{D}',t)].$$

The graph of $Q_1(\hat{D}',t)$ against $\hat{D}'$ can be compared for various values of t. Moreover, it can also be compared with graphs of the estimates of $P_1(D',t)$, $\hat{P}_1(\hat{D}',t)$, obtained by replacing the population parameters by their sample estimates. The graphs of $\hat{P}_1(\hat{D}',t)$ can be regarded as indicative of the limiting efficiency which can be achieved by increasing the sample size. To a sufficient degree of approximation, $\hat{P}_1(\hat{D}',t)$, is uniformly greater than $Q_1(\hat{D}',t)$ for $\hat{D}' > 0$ and uniformly less than $\hat{D}' < 0$. This reflects the

* *Division of Mathematical Statistics, C.S.I.R.O., 60 King St., Newtown, N.S.W., 2042.* # *Endocrine Research Unit, Cancer Institute, 278 William St., Melbourne, Vic., 3000.*

increase in uncertainty, in terms of the probability of misclassification given $\hat{D}'$, as a result of sampling variability in the parameter estimates.

These points are all illustrated in the example.

III PROGRAMS

A. Discrim

A program, DISCRIM, has been written in FORTRAN 1V for the CDC 3200, to perform the operations outlined in [4] and above. In writing this program, an attempt was made to satisfy several requirements dictated by its possible use.

(i) DISCRIM has been written in such a way that the major part of it should be transferable, without significant programming effort, to another small or medium-sized computer. Any section which depends on an optional computer facility such as an on-line graph plotter may be deleted without affecting the remainder of the program.

(ii) DISCRIM will handle, within the core-storage limitations of a small or medium-sized computer, problems involving large amounts of data. The number of attributes recorded for each individual, and the order of the poly nomial to be fitted, are limited by the available storage. Values of 20 for each of these numbers would not overtax the storage on the CDC 3200. The number of individuals that can be processed in one run is unlimited.

(iii) By using a free-format input routine*, which counts the number of attributes for each individual and the number of individuals in each group, the data preparation requirements for DISCRIM have been made very simple. Each individual is represented by one data card, and the cards to be processed must be arranged in two groups corresponding to Π_1 and Π_2. In addition, two parameter cards are necessary, specifying the order of polynomial to be fitted, the output options chosen and the ages at which the process is to be studied.
Results may be updated by adding the cards for incoming individuals to the existing data pack and re-running DISCRIM.

(iv) The output from DISCRIM is in two parts as discussed below.

(a) The line-printer output.

Firstly, the data for all the individuals may be listed; this option is controlled by the first parameter card. Then DISCRIM prints the number of attributes, the number in each group, and the order of the polynomial fitted. This is followed by

$\hat{B}_1$, $\hat{B}_2$, S — as defined in [4].

R_1, R_2 — the residual sums of squares and products matrix for each group.

Then for each age, t, specified by the second parameter card, DISCRIM prints

$\hat{\mu}_1(t)$, $\hat{\mu}_2(t)$ — the estimated means, in each group, of $\underline{X}(t)$,

$\hat{\underline{l}}(t)$, the vector of estimated coefficients for the discriminant,

$\underline{L}(t)$ the large sample variance-covariance matrix of $\hat{\underline{l}}(t)$

$E_1[\hat{D}(t)]$, $Var_1[\hat{D}(t)]$, $E_2[\hat{D}(t)]$, $Var_2[\hat{D}(t)]$ — where E_i and Var_i refer to the expectation and variance in group i,

$\hat{A}(t)$ as defined in Section II above,

$\hat{\underline{l}}^*(t) = \hat{\underline{l}}(t)/\hat{\sigma}(t)$,

$\hat{A}^*(t) = \hat{A}(t)/\hat{\sigma}(t)$.

(b) The graphical output.

The output to the on-line graph plotter may be suppressed by adjusting the first parameter card. If called for it may take one of two forms. The first form shows the graphs of $Q_1(\hat{D}',t)$ for all specified t. For the second form, each t occupies a separate graph, and curves for $Q_1(\hat{D}',t)$ and $P_1(\hat{D}',t)$ are shown.

B Disprob

A second program, DISPROB, has been written to calculate $Q_1(\hat{D}',t)$ and $\hat{P}_1(\hat{D}',t)$ for individuals. The data for DISPROB consists of certain of the results from DISCRIM, and data for the individuals for which the probabilities are required. Again, free-format input makes the preparation of data for DISPROB very simple.

IV EXAMPLE

In order to illustrate the output of DISCRIM, medical data kindly made available by Dr. R.D. Bulbrook of the Imperial Cancer Research Fund, London was used. These data relate tumour growth response and non-response to adrenalectomy of patients with advanced breast cancer to their urinary excretion of two steroid hormone metabolites, 17-hydroxycorticosteroids (X_1) and aetiocholanolone (X_2). The use of discriminant analysis in this context was described by Bulbrook, Greenwood and Hayward (1960) [2].

The statistical problem was to construct a two-variable discriminant function, based on X_1 and X_2, to separate responders and non-responders. Previous workers have noted marked linear regression of the variables with age and hence the analysis in [4], with k = 2, seemed to be appropriate. In Bulbrook's random sample of women undergoing the operation 47 responded and 117 failed to respond.

The analysis was run for k = 1 and for k = 2. In the latter case values t = 25(5)65 were used. Table 1 shows the results for k = 1, and Table 2 the results for k = 2, t = 25. Tables 1 and 2 are identical with the lay-out of results from

* The free-format input routine was developed by D. Culpin for the Computer Library of the Division of Mathematical Statistics.

TABLE 1

NO. OF ATTRIBUTES = 2
NO. IN 1ST GROUP = 47
NO. IN 2ND GROUP = 117
DEGREE OF POLYNOMIAL = 0

EST. OF B1
9.29149
971.0

EST. OF B2
9.52393
650.359

POOLED EST. OF V
1.50304×10^{1} 6.63102×10^{2}
6.63102×10^{2} 2.23882×10^{5}

RESIDUAL SSP MATRIX FOR GROUP 1
5.71177×10^{2} 4.72768×10^{4}
4.72768×10^{4} 1.33044×10^{7}

RESIDUAL SSP MATRIX FOR GROUP 2
1.86375×10^{3} 6.01457×10^{4}
6.01457×10^{4} 2.29646×10^{7}

MU1
9.29149 971.0

MU2
9.52393 650.359

COEFFICIENTS
-9.04705×10^{-2} 1.70014×10^{-3}

MATRIX L(T)
2.60045×10^{-3} -8.50192×10^{-6}
-8.50192×10^{-6} 1.89033×10^{-7}

GROUP 1 MEAN = 0.810234 VARIANCE = 0.885617
GROUP 2 MEAN = 0.244069 VARIANCE = 0.846805
A(T) = 0.527151

COEFFICIENTS FOR STANDARDISED DISCRIMINANT
−0.120236 0.002260

A FOR STANDARDISED DISCRIMINANT = 0.700590

TABLE 2

NO. OF ATTRIBUTES = 2
NO. IN 1ST GROUP = 47
NO. IN 2ND GROUP = 117
DEGREE OF POLYNOMIAL = 1

EST. OF B1
8.75616 0.0109968
1667.58 −14.3092

EST. OF B2
15.8199 −0.126070
801.948 −3.03542

POOLED EST. OF V
1.42622×10^{1} 6.52720×10^{2}
6.52720×10^{2} 2.20569×10^{5}

RESIDUAL SSP MATRIX FOR GROUP 1
5.70651×10^{2} 4.79604×10^{4}
4.79604×10^{4} 1.24149×10^{7}

RESIDUAL SSP MATRIX FOR GROUP 2
1.71129×10^{3} 5.64749×10^{4}
5.64749×10^{4} 2.28762×10^{7}

AGE 25

MU1
9.03108 1309.85

MU2
12.6681 726.063

COEFFICIENTS
−0.435067 0.00393422

MATRIX L(T)
2.12955×10^{-2} -7.02157×10^{-5}
-7.02157×10^{-5} 1.39722×10^{-6}

GROUP 1 MEAN = 1.224122 VARIANCE = 6.872249
GROUP 2 MEAN = −2.654999 VARIANCE = 7.261795

A(T) = −0.715438

COEFFICIENTS FOR STANDARDISED DISCRIMINANT
−0.220987 0.001998

A FOR STANDARDISED DISCRIMINANT = −0.363250

DISCRIM; for $k \geqslant 2$, the section of output headed "AGE 25" would be repeated for each age, t, specified.

Since

$$\hat{D}'(t) = [\hat{D}(t) - \hat{A}(t)] / \hat{\sigma}(t),$$

where

$$\hat{D}(t) = \underset{\sim}{x}'\hat{\underset{\sim}{l}}(t),$$

revised values of $\hat{l}_i(t)$, $\hat{l}_i^*(t)$, and of $\hat{A}(t)$, $\hat{A}^*(t)$, allow $\hat{D}'(t)$ to be calculated directly from any observational vector $\underset{\sim}{x}$. In the output from DISCRIM, the $\hat{l}_i^*(t)$ are given under "COEFFICIENTS FOR STANDARDISED DISCRIMINANT", and

$A^*(t)$ under "A FOR STANDARDISED DISCRIMINANT".

E.g., when k = 2, t = 25,

$$\hat{\sigma}(25) = \sqrt{E_1[\hat{D}(25)] - E_2[\hat{D}(25)]}$$

$$= \sqrt{1.224122 + 2.654999}$$

$$= 1.969548.$$

Therefore,

$$\hat{l}_1^*(25) = \frac{-0.435067}{1.969548} = -0.220897$$

$$\hat{l}_2^*(25) = \frac{0.00393422}{1.969548} = 0.001998$$

$$\hat{A}^*(25) = \frac{-0.715438}{1.969548} = -0.363250$$

and hence

$$\hat{D}'(25) = -0.220897\ X_1 + 0.001998\ X_2 - 0.363250.$$

The residual sums of squares and products matrices are included in the output in order that tests may be made of the significance of including the highest order term in t in the fitted polynomial. If we denote by $R_{i,k}$ the residual sums of squares and products matrix for group i when a $(k-1)^{th}$ order polynomial is fitted, then the likelihood ratio criterion for testing the null hypothesis that the term in t^{k-1} has no effect in group i is

$$\lambda = \frac{|R_{i,k}|^{n_i/2}}{|R_{i,k-1}|^{n_i/2}}$$

where n_i is the number in the i^{th} group, Anderson [1]. For testing, one usually uses $U = \lambda^{2/n_i}$, where

$U \sim U_{n,1,n_i-k}$. For n = 2

$$(U_{2,1,n_i-k})^{-1/2}(1 - [U_{2,1,n_i-k}]^{1/2})\,(n_i-k-1) = F_{2,2(n_i-k-1)}.$$

TABLE 3

	t = 25		t = 45		t = 65	
$\hat{\underset{\sim}{\mu}}_1(t)$	9.03108	1309.85	9.25101	1023.67	9.47095	737.487
$\hat{\underset{\sim}{\mu}}_2(t)$	12.6681	726.063	10.1467	665.354	7.62535	604.646
$\hat{\underset{\sim}{l}}(t)$	−0.435067	0.00393422	−0.158636	0.00209395	−0.117796	0.000253675
$\underset{\sim}{L}(t)$	2.12955×10^{-2}		3.48754×10^{-3}		9.52175×10^{-3}	
	-7.02157×10^{-5}	1.39722×10^{-6}	-1.19312×10^{-5}	2.4274×10^{-7}	-2.77339×10^{-5}	6.10478×10^{-7}
$E_1[\hat{D}(t)]$	1.224122		0.675970		1.302721	
$V_1[\hat{D}(t)]$	6.872249		1.306953		1.284046	
$E_2[\hat{D}(t)]$	−2.654999		−0.216418		1.051618	
$V_2[\hat{D}(t)]$	7.261795		1.285518		1.006449	
A(t)	−0.715438		0.229776		1.177169	
$\hat{\underset{\sim}{l}}^*(t)$	−0.220897	0.001998	−0.167928	0.002217	0.235074	0.000506
$\hat{A}^*(t)$	−0.363250		0.243236		2.349164	

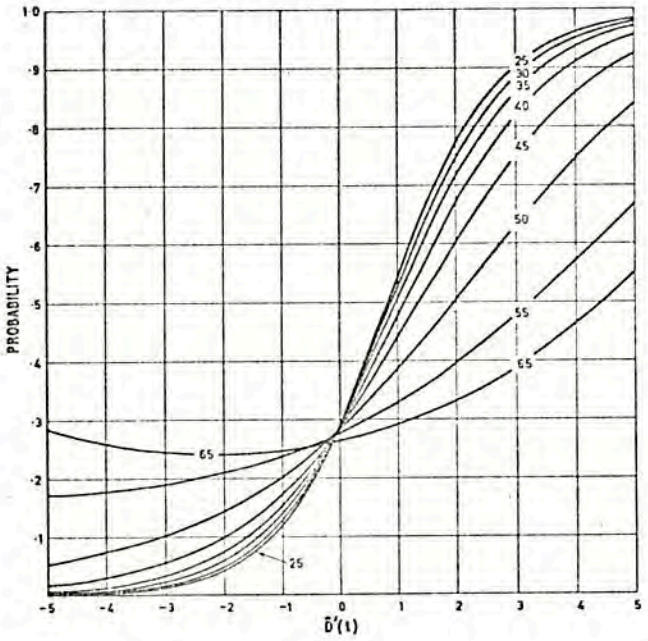


Figure 1.

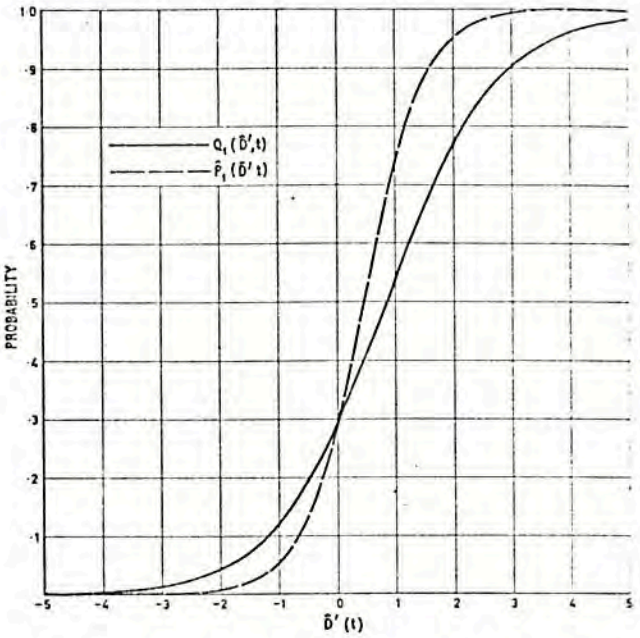


Figure 2.

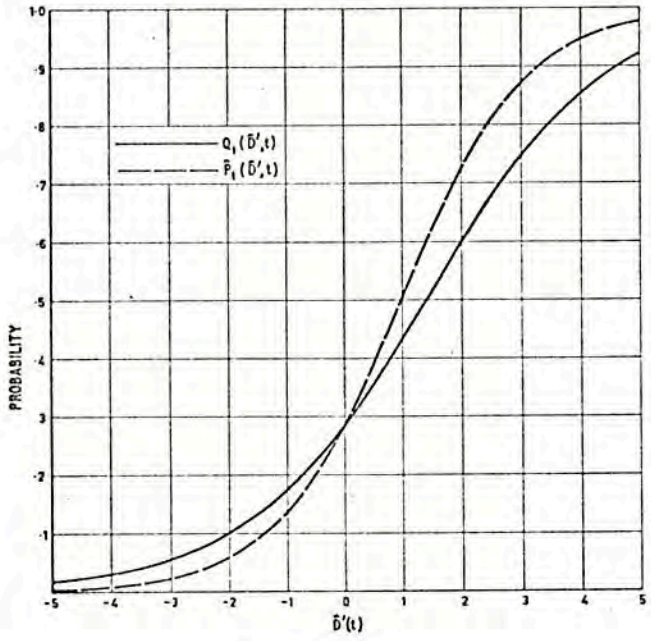


Figure 3

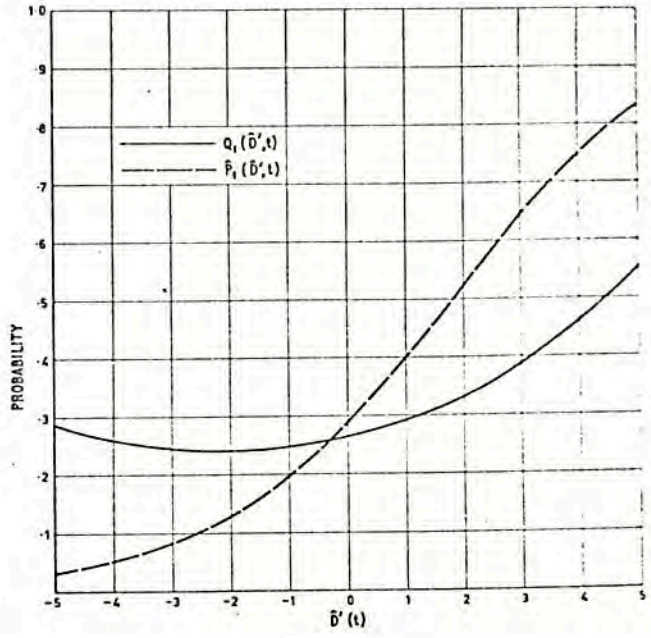


Figure 4

For $n > 2$, an approximation must be made, Hill and Davis [3], to obtain significance points for $U_{n,1,n_i-k}$. If one or both of the groups indicate that the term under consideration is significant, then it would seem reasonable to retain it in both groups. For the data in our example

$|R_{1,1}| = 5.36407 \times 10^9$ $|R_{1,2}| = 4.78438 \times 10^9$

$|R_{2,1}| = 3.91828 \times 10^{10}$ $|R_{2,2}| = 3.59584 \times 10^{10}$

Therefore

for group 1, $U = 0.892$, $\Pr\{U_{2,1,45} > U\} = 0.081$

for group 2, $U = 0.918$, $\Pr\{U_{2,1,115} > U\} = 0.007$.

A similar test for $k = 3$ was not significant and hence only the linear component appears to be important.

In Table 3 are presented, for purposes of comparison between ages, the results for $k = 2$ when $t = 25$, 45 and 65.

Note that the confidence intervals, shown below, for the $l_i(t)$ are much shorter at t = 45 than at t = 25 or t = 65.

t = 25

95% confidence limits for $l_1(25)$ are −0.721, −0.149
for $l_2(25)$ are 0.00162, 0.00625

t = 45

95% confidence limits for $l_1(45)$ are −0.274, −0.0429
for $l_2(45)$ are 0.00113, 0.00306

t = 65

95% confidence limits for $l_1(65)$ are −0.309, 0.0735
for $l_2(65)$ are −0.00128, 0.00179

(We have used confidence limits of the form $\hat{l}_i \pm 1.96\sqrt{\text{var } l_i}$ as a large-sample approximation). The difference in precision noted above in estimation of the $l_i(t)$ is to be expected in that the mean age of the patients providing the data is 49.58 and relatively few of the patients have ages close to 25 or 65.

Figures 1 — 4 illustrate the graphical output from DISCRIM. Figure 1 shows the curves of $Q_1(\hat{D}',t)$ against the standardised discriminant $\hat{D}'(t)$, for t = 25(5)55, 65. The curve for t = 60 is omitted because it is practically the same as the curve for t = 65. These curves illustrate that discrimination becomes less reliable as age increases.

Figures 2, 3 and 4 show the curves of $Q_1(\hat{D}',t)$ and $\hat{P}_1(\hat{D}',t)$ for t = 25, 45 and 65 respectively. As mentioned earlier, the probability $Q_1(\hat{D}',t)$ is less than $\hat{P}_1(\hat{D}',t)$ for $\hat{D}'(t) > 0$ (approximately) and greater for $\hat{D}'(t) < 0$ (approximately), indicating an increase in uncertainty as a result of sampling variability. Also, it might be noted that the probability curve for t = 45 lies closer to its limiting curve than is the case for t = 25 or t = 65. This again reflects a scarcity of data in the region of t = 25 and of t = 65.

ACKNOWLEDGEMENT

The authors wish to acknowledge Dr. Bulbrook's kindness in making the data available for these analyses.

REFERENCES

[1] ANDERSON, T.W. (1958): *Introduction to Multivariate Statistical Analysis.* (page 187). *(John Wiley & Sons: New York).*

[2] BULBROOK, R.D. GREENWOOD, F.C. AND HAYWARD J.L. (1960): Selection of breast cancer patients for adrenalectomy or hypophysectomy by determination of urinary 17-hydroxycortico-steroids and aetiocholanolone. *Lancet,i:* 1154.

[3] HILL, G.W. AND DAVIS, A.W. (1968): Generalized asymptotic expansions of Cornish-Fisher type. *Ann.Math.Statist., 39.* 1264.

[4] TALLIS, G.M. (1969): Some extensions of discriminant function analysis. *Metrika, XV* (in press).

B[5]

Reprinted from THE LANCET, *October* 3, 1970, *pp.* 685-687

PROBABILITY OF A WOMAN WITH ADVANCED BREAST CANCER RESPONDING TO ADRENALECTOMY OR HYPOPHYSECTOMY

GORDON SARFATY
Endocrine Research Unit, Cancer Institute, Melbourne, Victoria, Australia
MICHAEL TALLIS
Division of Mathematical Statistics, C.S.I.R.O., Newtown, New South Wales, Australia

Summary Data on urinary steroids in women with breast cancer have been used to produce probability curves allowing the chance of success with adrenalectomy or hypophysectomy to be computed for individual patients. Graphs have been drawn for probability of success against size of discriminant (11-deoxy-17-ketosteroids+17-hydroxycorticosteroids) in 164 cases of breast cancer. In older patients this approach could be made more useful by substituting time to ablation for 17-hydroxycorticosteroids. This probability method may provide the clinician with a useful tool in assessing an individual patient's likely response to endocrine surgery.

Introduction

THE search for variables [1] to predict the response of women with advanced breast cancer to adrenalectomy or hypophysectomy has been triggered off by the low remission-rate achieved by ablative surgery. Quantitative indices, such as the latent period and urinary androgenic steroid metabolites, provide valuable descriptions of group behaviour but are not much help in individual cases because of overlap between responding and non-responding groups. Discriminant function analysis [2] of response-related urinary steroid metabolites [3] achieves greater precision of classification between responding and non-responding groups especially when combined with the latent period.[4]

Besides the problem of group overlap, classification of a patient simply as a responder (positive discriminant) or non-responder (negative discriminant) only gives the probability of response of the group (i.e., about 30%

probability of success for non-responders and 70% for responders), and results in an unnecessary loss of information in respect of an individual's potential to respond to ablative surgery.

The difficulty can be avoided if the size as well as the sign of the discriminant is used. We have done this with preoperative measurements of urinary total 11-deoxy-17-ketosteroids (11-deoxy-17-K.S.), urinary 17-hydroxycorticosteroids (17-OHC.S.), and the time to ablation (latent period, or free interval), to calculate discriminant values.

Material and Methods

Dr. R. D. Bulbrook and Mr. J. L. Hayward of the Imperial Cancer Research Fund, London, kindly made available data on 164 of their patients with advanced breast cancer who underwent either total bilateral adrenalectomy or hypophysectomy. Patients having a regression after ablation were classified as responders, the remainder as non-responders.[5]

Urinary steroid metabolites were measured by the methods of Bulbrook et al.[3] Urinary total 11-deoxy-17-K.S. (i.e., ætiocholanolone, androsterone, and dehydroepiandrosterone) were found to be equivalent to using ætiocholanolone alone. Of the clinical data the interval from when the patient was first seen, or when they had had a mastectomy, to the time of ablation was used in preference to the conventional "free" interval since it was not available for all patients. We subsequently found that both intervals gave the same probabilities of remission.

We used an extension of the Fisher[6] discriminant analysis which treats responding and non-responding groups as samples of the true population and provides an improved account of discriminant analysis by considering errors made by this sampling.[7] The consideration of errors also allowed a comparison to be made between estimates of the population probabilities (referred to in the text as the estimate) and those obtained if the sample was regarded as the true population (referred to as the limit).

Fig. 1—Probability of remission after bilateral adrenalectomy or hypophysectomy in advanced breast cancer using a discriminant calculated from urinary steroid metabolite values.

(a) Effect of population sampling, showing estimate curve (E) of sample and limit curve (L) for true population.

(b) Probability of remission at different ages of ablation.

Fig. 2.—Limited potential for improving probabilities for older patients.

Fig. 3—Effect of time to ablation (or free interval) on the probability of remission.

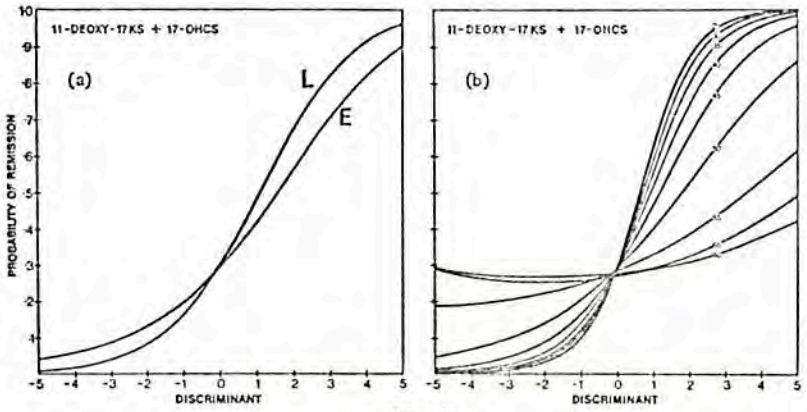


Fig. 1.

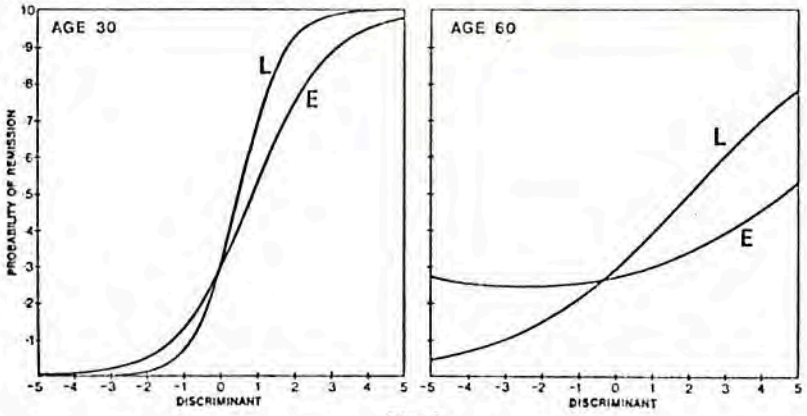


Fig. 2.

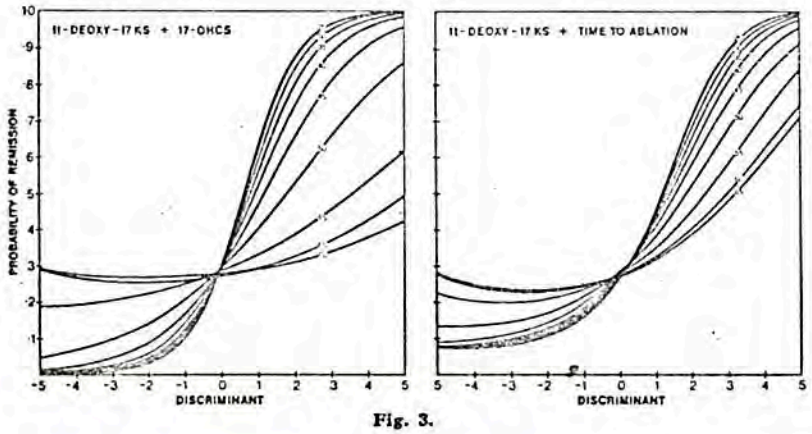


Fig. 3.

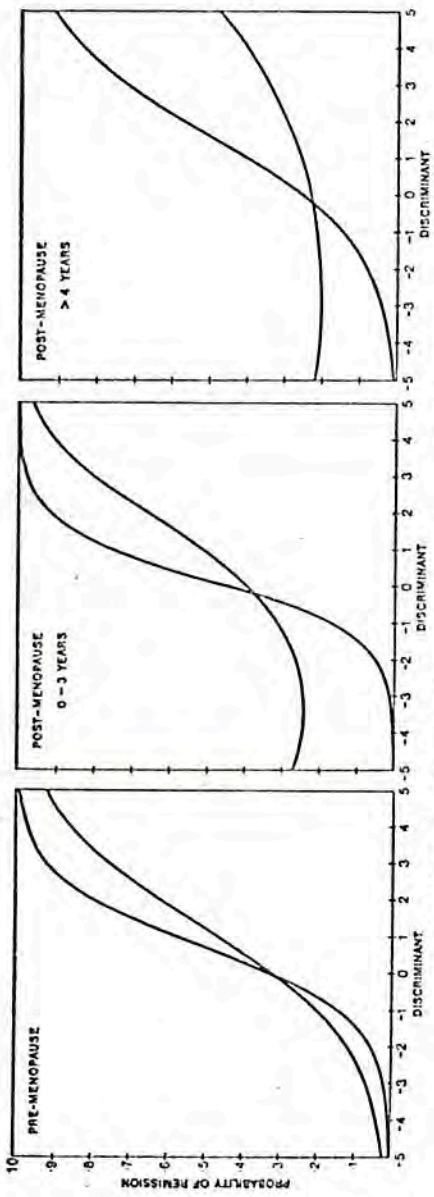


Fig. 4—Role of the patient's menstrual status at the time of ablation.

Individual probabilities of response were generated by a programme written in Fortran for the CDC 3200 computer.[8] The output was plotted to yield the probability of remission as a function of the discriminant. Figs. 1–4 were redrawn from the computer plots.

Results and Discussion

Population Sampling and Age at Ablation

The effect of population sampling is seen in fig. 1a. When errors associated with the discriminant function statistics are neglected the limit curve is obtained. The difference between the limit (L) and estimated (E) curve is not great, revealing the absence of a major sampling effect on estimated probabilities. For example, a discriminant value of 2 read on the estimate curve results in a probability of remission of approximately 55%, whereas the limit that can be obtained from the true population is approximately 65%. Conversely, a discriminant of −2 gives an estimate probability of 15% and a limit probability of 8%.

The patient's age at ablation is an important variable when determining response (fig. 1b). Urinary steroids seem to be of greatest value in determining response when measured in women below the age of fifty.

What potential is there for improvement in the older age-group with the use of these steroid metabolites? This can be determined by comparing the estimate and the limit probability curves at ages 30 and 60 (fig. 2). With a discriminant of 1, at age thirty gives a limit probability about 16% higher than the estimate, whereas the same discriminant at age sixty can be expected to be improved by only 8%. It seems that in older women this urinary steroid discriminant is neither useful nor potentially useful in detecting probabilities of remission.

Time to Ablation (or Free Interval)

When the latent period of the disease is substituted for urinary 17-OHc.s. as a discriminant variable, we get a substantial improvement in the likelihood of remission in older patients (fig. 3), and a slightly diminished probability of response in the age-groups around thirty years. We do not know why slower tumour growth, as indicated by the latent period, has a more important bearing on the outcome of surgery than does the urinary steroid pattern in older women.

Menopausal Status

Although the cyclic-functioning ovary is a control mechanism of tumour growth in some patients, it does not necessarily imply that the menopausal status of a woman is a determinant of remission to major ablation. Some workers regard the menopausal state as important, others disagree.[1] Fig. 4 shows the estimate and limit curves for three menstrual groupings; premenopause, up to three years after the menopause,

and more than four years postmenopausal.

The curves for premenopausal and postmenopausal patients are very similar to the age-probability curves in fig. 2, and the age forty to forty-five curve in fig. 1b. These findings suggest that the menstrual status may be an important determinant of probability of response only in so far as it reflects the patient's age.

Clinical application of discriminant analysis has not resulted in the hoped-for improvement in the selection of women for major endocrine surgery.[9] The reason for this is the inability of the discriminant to identify clearly individual responders or non-responders.

The probability approach to the use of discriminants for determining the likelihood of patient response can provide the clinician with more useful information when assessing the role of surgery in the individual woman. Whether the objective measurement of probability will be of greater value than clinical judgement of the likelihood of response can only be decided by a prospective study.

Requests for reprints should be addressed to G. S.

REFERENCES

1. Fairgrieve, J. *Surgery Gynec. Obstet.* 1965, 120, 371.
2. Kendall, M. G., Stuart, A. Advanced Theory of Statistics; vol. III. New York, 1966.
3. Bulbrook, R. D., Greenwood, F. C., Hayward, J. L. *Lancet*, 1960, i, 1154.
4. Wilson, R. E., Crocker, D. N., Fairgrieve, J., Bartholomay, A. F., Emerson, K., Moore, F. D. *J. Am. med. Ass.* 1967, 199, 474.
5. Sarfaty, G. A. *Med. J. Aust.* 1969, i, 398.
6. Fisher, R. A. *Ann. Eugen.* 1936, 7, 179.
7. Tallis, G. M. *Metrika* (in the press).
8. Tallis, G. M., Shaw, D. E., Williams, J., Sarfaty, G. A. *Aust. Comp. J.* 1970, 2, 3.
9. Atkins, H. J. B., Bulbrook, R. D., Falconer, M. A., Hayward, J. L., MacLean, K. S., Schurr, P. H. *Lancet*, 1968, ii, 1255.

B [6]

BASIC RESULTS OF A STUDY OF BILATERAL ADRENALECTOMY FOR ADVANCED BREAST CANCER

URINARY STEROIDS AND RELATED DATA IN 148 PATIENTS

GORDON SARFATY,* PAULA PITT* AND MICHAEL TALLIS†

Cancer Institute, Melbourne, and University of Adelaide

Med. J. Aust., 1973, 2: 877-881.

Over a five-year interval 148 women who had advanced breast cancer were treated by bilateral adrenalectomy. A detailed preoperative study was made of their urinary 17-ketosteroid metabolites of plasma androgens. In addition, a variety of clinical variables and tumour characteristics which might discriminate between responses to the treatment were examined. Significant differences were found between remitting and non-remitting groups for steroidal and clinical measurements. Related studies suggest that traditional discriminant analysis may not be the best means of utilizing the results.

IN 1967 a study of adrenalectomy in advanced breast cancer was commenced at the Cancer Institute, Victoria. Its aim was to research the potential for a clinically useful prediction of response to this form of therapy. A preliminary account of the experimental design and initial clinical results was presented in this Journal in 1969 (Sarfaty). We now present a detailed summary of the basic data for 148 adrenalectomies during the five-year period 1967 to 1972. These results include observations of clinical features of the patient's illness, tumour characteristics and the excretion of urinary steroid metabolites.

Studies of factors that might be useful in the prediction of response to endocrine ablation are neither recent nor novel. In 1900, Boyd reported on the clinical aspects of remission in 54 women with advanced breast cancer treated by oophorectomy. The subsequent introduction of treatment by major endocrine ablation, adrenalectomy and hypophysectomy, and development of the quantitative analyses of urinary steroids, encouraged further work on the preoperative classification of response (Bulbrook *et alii*, 1960).

Although clinical variables, steroid metabolites and tumour receptor proteins (reviewed in Fairgrieve, 1965; Forrest and Kunkler, 1968; Hayward, 1970; Breast Cancer Task Force, 1972) can all be used preoperatively to classify women as remitters, no successful clinical application to patient selection has been reported. Because of this it seemed that a reappraisal of the general and specific aspects of the response-prediction problem could be worthwhile. Our work has been especially concerned with the analysis of urinary neutral 17-ketosteroids, with the concept of remission, and with the biometrical problems of applying the results.

We have raised the question of the importance of these matters previously (Sarfaty, 1969; Sarfaty and Tallis, 1970*a* and *b*). The discussion here will be extended and a proposal made for considering the response to adrenalectomy as a continuous process or variable. By this definition some point of response occurs at which it becomes possible to classify a patient as a remitter or non-remitter to the ablation. This approach should avoid the difficulties of analyses which assume discrete categories of remission and non-remission.

THE PATIENTS

The 148 women studied were either post-menopausal or had had a previous oophorectomy. If a patient had had a previous hysterectomy, gonadotrophin and ovarian hormone assays were used to decide if there was cyclic ovarian activity. At adrenalectomy all women had disseminated cancer; 51% had been treated by systemic drug therapy and 30% by a previous oophorectomy. Oophorectomy was not always independent of drug therapy.

Patients were selected for adrenalectomy when this was considered to be the best means of palliating continuing tumour growth. Techniques of patient assessment and other aspects of the protocols have been described previously (Sarfaty, 1969).

A general account of the outcome, in respect of operative mortality, response and survival is summarized in the following list.

October, 1967, to November, 1972

Total operations	148
Deaths at operation	4
Mortality rate	2·9%
Unclassifiable responses	10
Total confirmed deaths[1]	103

Classifiable Patients (134)

Remissions	42 (31·3%)
Remitters still in remission	15 (11·2%)
Remitters still living	28 (20·8%)
No remission	92 (68·7%)
Non-remitters still living	15 (11·2%)

A basic purpose of the work was to find variables which would significantly classify the differences between remitting and

* Endocrine Research Unit.

† Department of Statistics, University of Adelaide.

Address for reprints: Dr Gordon Sarfaty, Head, Endocrine Research Unit, Cancer Institute, 278 William Street, Melbourne, Vic. 3000.

[1] By inspection of the death certificate issued by the Registrar of Births, Deaths and Marriages, Victoria.

TABLE 1
Comparison of Clinical Data for Remitting and Non-Remitting Patients

Clinical Data	Remitters			Non-Remitters			
	Number of Patients	Mean	SD	Number of Patients	Mean	S.D.	
Number of pregnancies	42	2·4	1·5	92	2·4	1·6	n.s.
Age at first pregnancy (years)	37	26·8	6·1	78	25·6	5·7	n.s.
Age at clinical diagnosis (years)	42	49·0	9·0	92	50·0	9·3	n.s.
Age at adrenalectomy (years)	42	51·3	8·2	92	53·6	8·7	n.s.
Delay in reporting symptoms (months)	42	14·5	45·8	92	3·7	9·0	n.s.
Latent ("free") interval (months)	31	43·5	39·3	67	32·6	37·9	n.s.
Diagnosis to adrenalectomy (months)	42	63·4	56·9	92	42·7	45·9	s
Adrenalectomy to death (months)	14	15·7	9·1	77	7·7	6·7	—
Diagnosis to death (months)	14	79·6	73·1	77	50·4	48·2	—
Surface area (m^2)	41	1·67	0·14	89	1·61	0·15	s

SD = Standard deviation.
s = Significance difference ($P < 0·05$)
n.s. = Difference not significant
— = Not analysed
} See text.

non-remitting groups. We therefore collected data that might reasonably assist that aim. As some restrictions have to be placed on this type of data gathering, we recognize the possibility of important omissions. There was, for example, no opportunity of examining steroid receptor proteins in the tumours from our patients.

When all data were gathered and the patients divided into two groups according to their clinical response, results between the two groups were tabulated. Statistical significance was then measured from the analysis of raw and logarithmically transformed data. Where necessary, due allowance was made for discrepancies in variance between the two groups.

Data gathered from the clinical assessment of the patients are recorded in Tables 1 and 2. Of the 34 observations, only two showed significant differences between remitters and non-remitters. One of these was surface area of the women. The other was the time interval between diagnosis and adrenalectomy.

This period includes the latent or "free" interval which, although longer in remitters, did not reach statistical significance. It should be noted that a zero value for latent period was given to patients who did not have a mastectomy before their adrenalectomy.

The fact of surface area being significantly different should not be surprising, as steroidal hormone production and excretion are often a function of body size.

Survival intervals, that is adrenalectomy to death, and diagnosis to death, have not yet been analysed. They cannot be examined by standard methods as some women in the group are still living. This bias, coupled with the fact that new patients are coming in at random times, introduces a complicated truncation effect which must be corrected for in the estimation of mean survival. Special methods have been developed to deal with this problem and a full analysis of survival will be presented separately.

The menopausal condition of patients is frequently discussed as a potential means of differentiating the responses to adrenalectomy. We could not show a significant difference in adrenalectomy remission rates between women who were premenopausal and those who were postmenopausal at the onset of their cancer. Nor was there any significant difference in remissions between those groups and patients who went through the menopause during the interval between the original diagnosis of their disease and their adrenalectomy.

Clinical stage at presentation, previous drug therapy and site of metastatic lesions do not significantly differentiate the responders. Patients with visceral lesions are said to remit less often than those with lesions at other sites (Fairgrieve, 1965). In the sample of patients reported here (Table 1) no significant differences were found on comparing the remission rates for visceral lesions with lesions in non-visceral sites.

TABLE 2
Comparisons of Clinical Data for Remitting (42) and Non-Remitting (92) Patients

Clinical Groupings	Remitters	Non-Remitters	Percentage of Total Patients
Mastectomy:			
No mastectomy	8	22	22
Unilateral mastectomy	31	65	72
Bilateral mastectomy	3	5	6
Menopausal state			
Premenopausal	23	42	49
Postmenopausal	14	44	43
Premenopausal at diagnosis and postmenopausal at adrenalectomy	6	5	8
Premenopausal at diagnosis and postoophorectomy at adrenalectomy	16	36	39
Oophorectomy:			
Prophylactic oophorectomy	4	13	13
Therapeutic oophorectomy	12	26	26
Drugs:			
Any drug within 30 days of adrenalectomy	7	15	16
Corticosteroids within 12 months of adrenalectomy	0	4	3
Site of lesions:			
Skeleton	27	60	65
Skin	27	54	60
Lymph nodes	26	64	67
Breast	11	27	28
Pulmonary	13	29	31
Liver	0	6	5
Cerebral	1	1	2
Abdominal	0	4	3
Haemopoietic and bone marrow	2	3	4
Clinical stage of disease at presentation:			
Stage 1	4	14	13
Stage 2	15	41	42
Stage 3	3	11	10
Stage 4	4	11	11

It should be noted that some of the above comparisons are based on small differences within subsets of the data. Examples are the number of patients having prophylactic oophorectomy, lesion sites such as liver, brain and abdomen, and bilateral mastectomy. Conclusions where the numbers are small should be regarded as tentative.

THE TUMOUR

In 78 cases it was possible to trace the tumour specimen and to reexamine the histology. The reexamination was carried out by Dr R. Motteram, who classified and graded the cancers

according to the International Histological Classification of Tumours, No. 2 (Scarff and Torloni, 1968). Frequency of types was as follows:

Type	Frequency
Ductal	66 (85%)
Lobular	8 (10%)
Medullary	3 (4%)
Apocrine	1 (1%)

Remission rates were compared by tumour type, ductal grade, fibrosis, and the presence of lymphocytes. Because of small numbers available in subsets, maximum use has been made of data by pooling groups, i.e. ductal versus non-ductal, low-grade ductal versus higher grades, and so on. The results are shown in Table 3.

TABLE 3
Remission Rates and Tumour Features

Tumour Features	Remissions	No Remission	Remission Rate	
Type:				
Ductal	16	30	0·35	n.s.
Non-ductal	3	7	0·30	
Ductal grade:				
Low grade (I)	10	17	0·37	n.s.
Higher grades (II and III)	6	13	0·40	
Fibrosis grades:				
None to minimal (− and +)	9	22	0·29	n.s.
Moderate to maximum (+ + and + + +)	10	15	0·40	
Lymphocytes:				
Lymphocytes absent (−)	17	24	0·41	$P<0·10$
Lymphocytes present (+ to + + +)	2	13	0·13	

n.s. = not significantly different ($P>0·10$).

In agreement with a number of other studies (reviewed by Hayward, 1970), no significant differences have been revealed for any of the groups.

However, a point of interest arises in connexion with the presence of lymphocytes in the tumour. While the difficulties of assessing tumour lymphocytes are recognized, the remission rate is lower in the women whose tumours were associated with lymphocytes. This difference was significant at the 10% level. If the relationship is valid, there could be an inverse role of involvement by the immune system in tumour response to ablation. A larger sample of patients is needed to confirm these results.

THE URINARY STEROIDS

The principal circulating androgenic steroids in plasma and their precursors are known to be dehydroepiandrosterone, dehydroepiandrosterone sulphate, androstenedione and testosterone (Vande Wiele *et alii*, 1963). Testosterone and androstenedione are interconvertible, while the products of androstenedione metabolism are androsterone (A) and etiocholanolone (E). The latter two steroids are excreted in the urine as glucuronide and sulphate conjugates. Plasma dehydroepiandrosterone (D) and its sulphate are also interconvertible. Dehydroepiandrosterone is excreted in the urine as the glucuronide, while the sulphate conjugate is excreted directly. The glucuronide and sulphate conjugates of A, E and D comprise the 11-deoxy fraction of the total urinary neutral 17-ketosteroids.

Previously reported studies of androgen metabolites have confined their observations to compounds measured as combined glucuronide and sulphate conjugates. Thus, when differences between remission and non-remission have been found it could not be decided whether they were due to relative changes within the conjugates or to changes in the actual metabolite itself.

We measured separately in the glucuronide and sulphate conjugate fractions of urine, the total neutral 17-ketosteroids (17-KS), the total 11-deoxy-17-ketosteroids, as well as the individual compounds dehydroepiandrosterone, androsterone and etiocholanolone.

Methods used were modifications of standard procedures involving spectrophotometry of the Zimmermann reacting material in groups of the steroids, and gas-liquid chromatography of trimethylsilylethers for the individual constituents of the 11-deoxy groups (Zimmermann, 1935; Kirschner and Lipsett, 1963).

Results are shown in Table 4. Statistical tests reveal significant differences for six of the ten types of compounds measured, again both on raw and transformed data. We cannot satisfactorily explain why the same compound(s) conjugated as glucuronide or sulphate vary in significance between remitters and non-remitters.

GENERAL DISCUSSION

Bilateral adrenalectomy as palliative treatment for advanced breast cancer has a number of advantages over other forms of systemic therapy. Remission rates and time in remission are higher for adrenalectomy than for either hormonal or cytotoxic therapy (Atkins *et alii*, 1966; Byron, 1967; Dao and Nemoto, 1965; Sarfaty, Pitt and Tallis, unpublished observations). On recovery from ablative surgery, the patient is not chronically disturbed by the hazards associated with cytotoxic drugs and pharmacological doses of hormones.

The disadvantages of adrenalectomy are the relatively low remission rate, the operative mortality, and the need for a major surgical procedure in patients whose remaining life expectancy is limited. At the time adrenalectomy is contemplated, the

TABLE 4
Comparison of Urinary Steroids (µg/24 hrs) for Remitting and Non-Remitting Patients

Urinary Steroids	Remitters			Non-Remitters			
	Number of Patients	Mean	SD	Number of Patients	Mean	SD	
Glucuronide fraction:							
Total 17-ketosteroids	40	3,510	1,220	86	2,790	1,340	s
Total 11-deoxy-17-ketosteroids	40	2,059	1,193	85	1,329	947	s
Androsterone	40	671	510	86	548	710	n.s.
Etiocholanolone	40	987	506	87	716	541	s
Dehydroepiandrosterone	37	23	45	78	110	575	n.s.
Sulphate fraction:							
Total 17-ketosteroids	37	902	530	76	1,626	750	n.s.
Total 11-deoxy-17-ketosteroids	40	493	342	85	276	295	s
Androsterone	40	172	172	87	86	121	s
Etiocholanolone	40	167	149	87	107	120	s
Dehydroepiandrosterone	37	148	183	79	103	148	n.s.

SD = Standard deviation.
s = Significant difference ($P<0·05$)
n.s. = Difference not significant
} See text.

expectation from breast cancer life tables is about three years (Tallis *et alii*, 1973) and in our present group was about 12 months (Table 1). Seventy-eight per cent of the women have already had previous major surgery, either a unilateral or bilateral mastectomy.

It is also common for patients at this stage of the disease to have been treated with radiotherapy, hormones or cytotoxic drugs, either serially or in combination. A further disadvantage of major ablation is the postoperative need of continually supervised corticosteroid replacement therapy.

In this position it would be ideal to have a clear-cut pronouncement on the expectancy of remission, and therefore avoid unnecessary treatment for patients who will not benefit. However, even with intensive studies no major clinical application has resulted. The ten-year research of steroid-based discriminants described by Atkins and his colleagues (1968) concluded with: "hopes of a single, all-embracing test that would differentiate clearly between responsive and unresponsive patients have not been realized."

What are the reasons for this failure? One possibility is that the parameters cannot distinguish between two populations. We have minimized this in regard to steroid excretion by estimating these compounds in five-day urine collections (Sarfaty and Tallis, 1970*a*).

Another reason may be that remission and non-remission categories are unrealistic when cancer growth is considered as a biological phenomenon. If this is true discrete categories may be artefacts of the methods used for assessment of patients. As a corollary, applying discriminants to data analysis may not be the optimum approach to segregating the tumour responses to ablation.

When it occurs, regression of the cancer takes place over a variable time. Sometimes the process is rapid and dramatic, sometimes it is slow and can only be decided in retrospect. In our own cases it has taken up to eight months, with repeated lesion measurement and skeletal surveys, to be reasonably certain of response classification. On other occasions it appeared that growth became static and neither progressed nor regressed. In other patients growth appears to continue, but at a slower rate than previously. Rarely, the response seems bizarre with tumour progressing at one site and regressing at another. Less objectively, patients sometimes insist that their condition has improved, but no objective basis can be made out for this.

These apparent realities of the tumour-host relationship are difficult to reconcile with a dichotomy of clinical response. This may be partly due to the imprecise techniques available for clinically measuring tumour growth and to the need for their frequent, critical application. Another factor may be the strong natural desire to demonstrate a "cure" with present cancer therapies. This can lead to a misunderstanding of what is achieved by endocrine ablative therapy.

With clinical tools, node, skin, bone and visceral lesions may be undetectable and indistinguishable from zero growth rate. But we know that recurrence is inevitable. The resulting difficulties of classifying the outcome of ablation prompted Bulbrook *et alii* (1960) to use a third, intermediate category of response in his discriminant studies. A three-category classification confounds the analysis which requires ablation response to be a dichotomy.

Discriminant functions have been calculated in this context because the problem was conveniently thought of in terms of a mixed population. While discriminant analysis (Fisher, 1936) is a classical and useful technique for assigning an unclassified member of a mixed population to one or other component, its application to the situation in breast cancer may be inappropriate.

We consider that ablation response should not be regarded as a mixture, but as a homogeneous continuous phenomenon. In this context discriminant analysis or modifications of it (Tallis, 1970; Tallis *et alii*, 1970) is not the desirable statistical technique to apply to the tumour growth response following ablation or any form of anticancer therapy.

A more reasonable model might be to consider tumour growth as a continuous variable and the classification of remission as a threshold phenomenon. Thus when a certain point of diminished tumour growth is reached following adrenalectomy clinical remission is considered present.

This clinically-defined point of tumour growth rate will result in an artificial cleavage of the population to provide a group of patients who are said to be remitters by presently used criteria.

If results of the observation we have presented are considered in this way, suitable statistical treatment can recognize remitting and non-remitting groups and quantify the probability of remission.

The net result may seem no different from previous aims of finding remission probabilities (Sarfaty and Tallis, 1970*b*). The fact is, however, that the new analysis will not be based on a possibly erroneous assumption that tumour growth rate in breast cancer is a mixed population phenomenon. Treating tumour growth as a continuous variable should lead to greater accuracy and can provide further opportunities for development in this field.

In summary, it is clear from the present work that it is possible to detect statistically significant differences between remission and non-remission groups as currently defined. These differences are present in respect of host parameters, that is, the time interval of known disease, surface area and urinary steroid metabolites. There are no significant differences between the groups in respect of the tumour characteristics dealt with here.

Definition of significant differences between the two groups is no guarantee of useful clinical application. Response to adrenalectomy may best be regarded as a continually graded change in tumour growth, preserving the terms remission and non-remission as convenient expressions for present clinical practice.

Meanwhile, application of the present results to defining a patient's probability of remission is proceeding. If carefully applied, the technique is likely to improve patient selection in what to date has been a matter of an intuitive judgement, often difficult for both patient and surgeon.

ACKNOWLEDGEMENTS

These studies were commenced in the Department of Medicine, Monash University, at Prince Henry's Hospital. Then Professor Bryan Hudson made facilities available in that department while Dr T. H. Ackland and Dr N. Johnson operated on patients at the Royal Melbourne Hospital. At the Peter MacCallum Clinic the late Dr W. P. Holman and Dr W. Moon assisted with patient selection and assessment. Later, Professor R. Bennett and Dr D. Chan also aided the selection of patients, and surgery was additionally carried out by the former at St Vincent's Hospital. The frequent radiological examinations required have been continuously reported on by Dr John Martin and Dr Barry Drake. The Outpatient Department and Visiting Nursing Service of the Peter MacCallum Clinic have willingly provided help for our rigorous standards of patient assessment. In the Endocrine Research Unit laboratory, Miss G. Peterson, Mr O. Sautner, Mr G. Joannou and Mrs H. Gardner maintained a high standard of technological proficiency. We acknowledge with pleasure the contributions of all and Dr R. D. Wright for continual encouragement.

REFERENCES

ATKINS, H. J. B., FALCONER, M. A., HAYWARD, J. L., *et alii* (1966), The timing of adrenalectomy and of hypophysectomy in the treatment of advanced breast cancer, *Lancet*, 1: 827.

ATKINS, H. J. B., BULBROOK, R. D., FALCONER, M. A., *et alii* (1968), Ten years' experience of steroid assays in the management of breast cancer, *Lancet*, 2: 1255.

BREAST CANCER TASK FORCE of the National Cancer Institute (1972), *Workshop on Breast Cancer*, Bethesda, Maryland.

BULBROOK, R. D., GREENWOOD, F. C., and HAYWARD, J. L. (1960), Selection of breast cancer patients for adrenalectomy or hypophysectomy by determination of urinary 17-hydroxy-corticosteroids and etiocholanolone, *Lancet*, 1: 1154.

BULBROOK, R. D. (1970), Prognostic value of steroid assays in human breast cancer, in *Advances in Steroid Biochemistry and Pharmacology*, edited by Briggs, H. M., Academic Press, London, 1: 387.

BYRON, R. J., jr (1967), Randomised adrenalectomies in advanced breast cancer, in *Major Endocrine Surgery for the Treatment of Cancer of the Breast in Advanced Stages*, edited by Dargent and Romien, p. 253, quoted by Hayward, J. L., 1970, *op. cit.*

DAO, T. L., and NEMOTO, T. (1965), An evaluation of adrenalectomy and androgen in disseminated mammary carcinoma, *Surg. Gynec. Obstet.*, 121: 1257.

FAIRGRIEVE, J. (1965), Selective criteria for surgical removal of the endocrine glands in advanced breast cancer, *Surg. Gynec. Obstet.*, 120: 371.

FISHER, R. A. (1936), The use of multiple measurements in taxonomic problems, *Ann. Eugen. (Lond.)*, 8: 376.

FORREST, A. P. M., and KUNKLER, P. B. (1968), editors, *Prognostic Factors in Breast Cancer*, Livingstone, Edinburgh.

HAYWARD, J. L., and BULBROOK, R. D. (1968), Urinary steroids and prognosis in breast cancer, in *Prognostic Factors in Breast Cancer*, edited by Forrest, A. P. M., and Kunkler, P. B., Livingstone, Edinburgh; 383.

HAYWARD, J. L. (1970), *Hormones and Human Breast Cancer*, Heinemann Medical Books, London.

KIRSCHNER, M. A., and LIPSETT, M. B. (1963), Gas-liquid chromatography in quantitative analysis of urinary 11-deoxy-17-ketosteroids, *J. clin. Endocr.*, 23: 255.

SARFATY, G. (1969), Preliminary account of a study concerned with individual patient response to adrenalectomy in advanced breast cancer, MED. J. AUST., 1: 398.

SARFATY, G., and TALLIS, M. (1970*a*), Aspects of the reliability of a urinary 17-hydroxysteroid assay, *J. clin. Endocr.*, 31: 52.

SARFATY, G., and TALLIS, M. (1970*b*), Probability of a woman with advanced breast cancer responding to adrenalectomy or hypophysectomy, *Lancet*, 2: 685.

SCARFF, R. W., and TORLONI, H. (1968), Histological typing of breast tumours, in *International Histological Classification of Tumours, No. 2*, WHO, Geneva: 19.

TALLIS, G. M. (1970), Some extensions of discriminant function analysis, *Metrika*, 15: 86.

TALLIS, G. M., SHAW, D. E., WILLIAMS, J., and SARFATY, G. (1970), Discriminant analysis by computer, *Aust. Comput. J.*, 2: 3.

TALLIS, G. M., SARFATY, G., and LEPPARD, P. (1973), *The Use of a Probability Model for the Construction of Age Specific Life Tables for Women with Breast Cancer*, University of Adelaide and Cancer Institute, Melbourne.

VANDE WIELE, R. L., MACDONALD, P. C., GURPIDE, E., and LIEBERMAN, S. (1963), Studies on the secretion and interconversion of the androgens, *Recent Progr. Hormone Res.*, 19: 275.

ZIMMERMANN, W. (1935), *Eine Farbreaktion der Sexualhormone und ihre Anwendung zur quantitativen colorimetrischen Bestimmung, Hoppe-Seylers Z. physiol. Chem.*, 233: 257.

AUSTRALASIAN MEDICAL PUBLISHING COMPANY LIMITED
71-79 Arundel Street, Glebe, N.S.W., 2037

B [7]

COMPUTERS AND BIOMEDICAL RESEARCH 8, 1–7 (1975)

A General Classification Model with Specific Application to Response to Adrenalectomy in Women with Breast Cancer

G. M. TALLIS AND P. LEPPARD

Department of Statistics, University of Adelaide

AND

G. SARFATY

Endocrine Research Unit, Cancer Institute, Melbourne

Received October 25, 1973

INTRODUCTION

We consider the general situation of a mixture of two populations Π_1 and Π_2. If X is a random vector of measurements associated with each individual of the population, a model for the density of X is specified by

$$f(\mathrm{x}) = pf_1(\mathrm{x}) + qf_2(\mathrm{x}), p + q = 1, \tag{1}$$

where $f_i(\mathrm{x})$ is the density of X in Π_i, $i = 1, 2$, and p is the proportion of individuals in Π_1.

A classical statistical problem is to assign individuals to either Π_1 or Π_2 on the basis of the observed vector $\mathrm{X} = \mathrm{x}$. Another way of looking at this is to calculate the conditional probability that the individual belongs to Π_1, say, given $\mathrm{X} = \mathrm{x}$. This probability is given by

$$P(\mathrm{x}) = Pr\{\Pi_1|\mathrm{x}\} = \frac{pf_1(\mathrm{x})}{pf_1(\mathrm{x}) + qf_2(\mathrm{x})}. \tag{2}$$

Under the usual assumptions that X is distributed as a multinormal distribution with mean vector μ and covariance Σ, $N(\mu, \Sigma)$, $P(\mathrm{x})$ reduces to

$$P(\mathrm{x}) = \left(1 + \frac{q}{p}\exp\{2\mathrm{d}'\mathrm{x} - \mu_2'\Sigma^{-1}\mu_2 + \mu_1'\Sigma^{-1}\mu_1\}\right)^{-1}, \tag{3}$$

where $\mathrm{d} = \Sigma^{-1}(\mu_2 - \mu_1)$ is the vector of discriminant coefficients. For an analogous but alternative approach, see Cornfield and Truett (*1*).

The purpose of this note is to propose a different model which should have wide applications. The results will be developed and discussed using the specific example of clinical response to adrenalectomy in women with breast cancer.

The Model

We are concerned with predicting the clinical assessment of remission and nonremission in women suffering breast cancer following adrenalectomy. In this context Π_1 is the population of all women with the disease who remit and Π_2 is the nonremitting population. The vector $\mathbf{X}$ consists of certain measurements made on each woman and may include age, surface area, certain time intervals related to the disease, and some chemical assessment of the internal hormonal environment. The statistical exercise is to estimate the chances of remission prior to the operation.

It seems reasonable to postulate that response to adrenalectomy is continuous rather than discrete as required by the mixture $f(\mathbf{x})$ described above (see Sarfaty *et al.* (*4*)). In fact, let X_0 be this response, X_0 being closely associated with the rate of tumour growth. We assume that X_0 and $\mathbf{X}$ have a density function $g(x_0, \mathbf{x})$.

Now the clinical condition of remission is regarded as occurring if and only if $X_0 > a$, for some fixed constant a on the response scale. Thus, nonresponse corresponds to $X_0 \leqslant a$. If this is so, the classification remission and nonremission introduces a truncation on X_0, the truncation being effective in $(n+1)$ dimensions. Thus the remitting population has density function

$$h(x_0, \mathbf{x}) = g(x_0, \mathbf{x})/\alpha(a), \tag{4}$$

where

$$\alpha(a) = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \int_{a}^{\infty} g(x_0, \mathbf{x}) dx_0 \, d\mathbf{x}.$$

What is required is the conditional density of $X_0|\mathbf{X} = \mathbf{x}$,

$$\phi(x_0|\mathbf{x}) = g(x_0, \mathbf{x})/k(\mathbf{x}), \tag{5}$$

where $k(\mathbf{x})$ is the joint density function of $\mathbf{X}$. If $P(\mathbf{x}) = Pr\{\text{Remission}|\mathbf{x}\}$, then

$$P(\mathbf{x}) = \int_{a}^{\infty} \phi(x_0|\mathbf{x}) dx_0. \tag{6}$$

In order to apply the above model, we assume, with no loss in generality, that X_0 is scaled so that $E[X_0] = 0$ and $\text{var}[X_0] = 1$. Moreover, suppose $g(x_0, \mathbf{x}) = N(\nu, \Sigma)$, where

$$\nu = \begin{bmatrix} 0 \\ \mu \end{bmatrix}, \quad \Sigma = \begin{bmatrix} 1 & \sigma' \\ \sigma & V \end{bmatrix},$$

$$\mu = E[\mathbf{X}], \ \text{var}[\mathbf{X}] = V, \ \text{and} \ \text{cov}[X_0, \mathbf{X}] = \sigma.$$

Then by standard distribution theory,

$$\phi(x_0|\mathbf{x}) = N(\sigma' V^{-1}(\mathbf{x} - \mu), 1 - R^2), \tag{7}$$

where $R^2 = \sigma' V^{-1} \sigma$.

If

$$\Phi(a) = \int_{-\infty}^{a} \frac{e^{-t^2/2} dt}{\sqrt{2\pi}},$$

then finally,

$$P(\mathrm{x}) = 1 - \Phi\left(\frac{a - \sigma' V^{-1}(\mathrm{x} - \mu)}{\sqrt{1 - R^2}}\right). \tag{8}$$

Put $W = \sigma' V^{-1}(\mathrm{X} - \mu)$; then $\mathrm{var}[W] = R^2$, $E[W] = 0$, and

$$P(\mathrm{x}) = Q(u, R) = 1 - \Phi\left(\frac{a - Ru}{\sqrt{1 - R^2}}\right)$$

for $U = W/R$. The function $Q(u, R)$ can be plotted as a function of u for $-3 \leqslant u \leqslant 3$. Clearly,

$$\lim_{R \to 1} Q(u, R) = \begin{cases} 1, & u > a \\ \frac{1}{2}, & u = a \\ 0, & u < a \end{cases}$$

which specifies the optimum probability curve. Thus, in general, the larger R, the more desirable the probability plot, since by suitably scaling X_0 it can be assumed that $R \geqslant 0$.

By constructing $Q(u, R)$ for various subvectors of X, the plots of Q can be compared and the effect of omitting certain variables from the analysis assessed. This technique is illustrated in the next section.

From first principles, or by specializing (3) of Tallis (*5*), the expected value of X_i when $X_0 > a$ is

$$E_R[X_i] = \mu_i + \sigma_i \rho_{0i} \phi(a)/\alpha(a) \quad i = 1, \ldots, n, \tag{9}$$

and when $X_0 \leqslant a$,

$$E_{NR}[X_i] = \mu_i - \sigma_i \rho_{0i} \phi(a)/[1 - \alpha(a)] \quad i = 1, \ldots, n. \tag{10}$$

Thus the effect of the truncation is to produce two different vectors of means $E_R[\mathrm{X}]$ and $E_{NR}[\mathrm{X}]$ as in the mixture model.

We note that part of the above development is essentially that of Hannan and Tate (*2*) who investigated some extensions of the classical biserial correlation model. Equations (9) and (10) are given in vector notation in the latter paper, but applications of the general model to personalised probabilities are not discussed there. These general procedures may extend usefully to situations where there are more than two classes, but this matter will not be pursued further here.

Example

The following measurements were made by the Endocrine Research Unit, Peter McCallum Clinic, on 120 women with breast cancer who had an adrenalectomy (Sarfaty *et al.* (*4*)):

Y_1 = Age at clinical diagnosis
Y_2 = Surface area

Y_3 = 17-Ketosteroid glucuronide
Y_4 = 11-Deoxy-17-ketosteroid glucuronide
Y_5 = 11-Deoxy-17-ketosteroid sulphate
Y_6 = Androsterone glucuronide
Y_7 = Etiocholanolone glucuronide
Y_8 = Androsterone sulphate
Y_9 = Etiocholanolone sulphate
Y_{10} = Time delay before reporting
Y_{11} = Reporting to adrenalectomy

The means and standard deviations of the Y_i for women in the remitting and non-remitting groups are given in Table 1.

TABLE 1

MEANS AND STANDARD DEVIATIONS FOR THE Y_i

	Remitters ($n_1 = 39$) Mean	SD	Nonremitters ($n_2 = 81$) Mean	SD
Y_1	49.45	8.97	50.85	8.74
Y_2	1.67	0.14	1.61	0.15
Y_3	3.56	1.20	2.68	1.28
Y_4	2083.72	1196.71	1320.44	944.58
Y_5	495.05	346.84	283.94	298.86
Y_6	684.31	510.56	538.14	722.24
Y_7	993.56	510.79	694.64	517.37
Y_8	174.49	174.69	85.21	121.92
Y_9	162.80	148.03	108.65	119.06
Y_{10}	1.31	4.03	0.26	0.55
Y_{11}	5.41	4.91	3.68	4.02

The proportion of women in the remitting group, 0.325, provides an estimate of a, $\hat{a}$, as the solution of

$$\int_{\hat{a}}^{\infty} \phi(t)dt = 0.325.$$

Thus $\hat{a} = 0.454$.

In order to apply normal distribution theory, it was assumed that $X_i = \Phi^{-1}[F_i(Y)]$, $i = 1, \ldots, 11$, have a multivariate normal distribution with mean 0 and covariance matrix equal to the correlation matrix R. The functions $F_i(\cdot)$ are the respective marginal distribution functions of the Y_i, and Φ^{-1} is the inverse function of the standard normal probability integral Φ. This model is an extension and a new application of a model reported by Moran (3). Under this type of transformation, scale and location are standardized while the relationship between the variables Y_i is preserved. This is what is required for the present problem.

In order to apply this theory, it was convenient to trivially modify the original Y_i to avoid the possibility of ties. This was done by first scaling the Y_i to eliminate decimal quantities and then adding a five-digit random number on (0, 1) as a new decimal. Thus if the scaled Y_i was 1132, then the modified value might be 1132.14527. This effectively removes the possibility of ties and simplifies the application of the theory.

For each modified Y_i, a sample distribution function is constructed, $F_i^n(y)$, where n is the sample size. Thus, we are concerned on each margin with the ordered sample $Y_{i(1)}, Y_{i(2)}, \ldots, Y_{i(n)}$, which defines $F_i^n(y)$ uniquely. The required transformation model is then approximated by mapping $Y_{i(j)}$ to u_j, where u_j is the $(j/n \times 100)$th percentile of Φ. Such a form of approximation is necessary since the individual $F_i(y)$'s are unknown.

This procedure was followed for each element of the 120 vectors of observations $\mathbf{Y}_i$ and produced 120 new vectors which approximated $\mathbf{X}_i$. Each component had an identical mean and variance of 0.0333 and 1.0658, respectively. The deviations from the theoretical values of 0 and 1 are due to the approximation used in transforming $Y_{i(120)}$, but they diminish with an increase in sample size. The correlations between $X_j, j = 1, \ldots, 11$, were calculated by standard techniques and are given in Table 2.

TABLE 2

CORRELATION TABLE FOR THE X_i

	X_0	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}	X_{11}
ρ'	1	−.09	.24	.39	.44	.36	.30	.39	.39	.30	.17	.27
		1	−.15	−.24	−.25	−.12	−.10	−.21	−.15	−.14	−.08	−.47
			1	.20	.17	.16	.09	.23	.13	.14	.05	−.10
				1	.77	.46	.56	.82	.49	.56	.13	.13
					1	.53	.64	.79	.49	.42	−.11	.06
C						1	.40	.51	.69	.62	−.05	−.02
							1	.67	.50	.36	−.10	−.03
								1	.53	.59	−.05	.10
									1	.71	−.12	.04
										1	−.07	.10
											1	−.21
												1

In order to estimate $\rho_{0i}, i = 1, \ldots, 11$, use was made of formulas (9) and (10) which when solved for ρ_{0i} and substituting observed values of $\hat{\mu}_i = 0.033$, $E_R[X_i] = \bar{x}_{iR}$, $E_{NR}[X_i] = \bar{x}_{iNR}$ (see Table 3), $\hat{\sigma} = 1.0658$, and $\hat{a} = 0.454$ yields

$$\hat{\rho}_{0i} = \frac{0.325\bar{x}_{iR} - 0.675\bar{x}_{iNR} + 0.35 \times 0.033}{2 \times 1.0658 \times \phi(\hat{a})}.$$

These values are recorded on the top line of Table 2 also.

TABLE 3

SAMPLE MEANS OF THE X_i FOR REMITTERS AND NONREMITTERS

	Remitters	Nonremitters
X_1	−0.075	0.086
X_2	0.305	−0.097
X_3	0.475	−0.179
X_4	0.533	−0.207
X_5	0.440	−0.162
X_6	0.374	−0.131
X_7	0.475	−0.179
X_8	0.479	−0.181
X_9	0.375	−0.131
X_{10}	0.229	−0.061
X_{11}	0.346	−0.117

The estimate of the multiple correlation coefficient R was $\hat{R} = 0.702$, and using this parameter estimate, $Q_1(u, \hat{R})$ was tabulated against u (Table 4).

TABLE 4

PROBABILITY VALUES FOR TWO GROUPINGS OF THE Y_i

u	Q_1	Q_2
−3.0	0.00016	0.00023
−2.5	0.00096	0.00126
−2.0	0.00453	0.00551
−1.5	0.01714	0.01953
−1.0	0.05223	0.05648
−0.5	0.12915	0.13425
0.0	0.26196	0.26498
0.5	0.44267	0.44055
1.0	0.63643	0.62889
1.5	0.80012	0.79028
2.0	0.90909	0.90076
2.5	0.96625	0.96116
3.0	0.98987	0.98755

The variables Y_5, Y_7, Y_9 were deleted from the analysis and $Q(u, \hat{R})$ was recalculated, Q_2 say, with $\hat{R} = 0.691$. These variables are difficult and strenuous to obtain and it would be desirable not to have to measure them on the pre-adrenalectomy patients provided the probability of response is not drastically altered. The function

$Q_2(u, \hat{R})$ is also tabulated (Table 4), and it can be seen that the overall effect of omitting Y_5, Y_7, Y_9 is negligible.

Discussion

Although the model for $P(x)$ as specified by Eq. (8) has been developed in terms of a specific biological situation, it is clear that the results have a wide potential use. Whenever a population is dichotomized by means of a truncation as hypothesized for X_0, (8) may provide a suitable means of calculating conditional probabilities for the occurrence or nonoccurrence of a particular event. In general, then, if the event A is associated with $a < X_0$ and the complimentary event $\bar{A}$ with $X_0 \geq a$, (8) gives the conditional probability of A given $X = x$.

Specific applications to response and nonresponse to treatment in any trials where response is assessed on a go, no-go basis follows by direct analogy. Whenever a continuous response coupled with a threshold principle specifies a reasonable model for the process, these results provide a more attractive analysis than the classical mixed population approach of discriminant analysis. The fact is that in many cases, two populations do not exist and one is invariably involved with truncation in a single population.

ACKNOWLEDGMENT

The authors thank the referee for comments which helped to improve the presentation of this paper.

REFERENCES

1. Cornfield, J., Truett, J., and Kannel, W. A multivariate analysis of the risk of coronary heart disease in Framingham. *J. Chron. Dis.* 20, 511 (1967).
2. Hannan, J. F., and Tate, R. F. Estimation of the parameters for a multivariate normal distribution when one variable is dichotomized. *Biometrika* 52, 664 (1965).
3. Moran, P. A. P. Testing for correlation between non-negative variates. *Biometrika*, 54 (3, 4), 385 (1967).
4. Sarfaty, G., Pitt, P., and Tallis, G. M. Basic results of a study of bilateral adrenalectomy for advanced breast cancer—Urinary steroid related data in 148 patients. *Med. J. Aust.* 2 (19), 877 (1973).
5. Tallis, G. M. The moment generating function of the truncated multi-normal distribution. *J. Royal Statist. Soc. B* 23, 223 (1961).

B[8]

ON THE OPTIMAL ALLOCATION OF CLINICAL TREATMENTS

G.M. Tallis, P. Leppard and G. Sarfaty

University of Adelaide
and
Endocrine Research Unit, Cancer Institute, Melbourne

To appear in "Mathematical Biosciences"

ABSTRACT

The optimal allocation of medical treatment when the probability of response is a known function of measurements made prior to the treatment is discussed. The treatment of women with advanced breast cancer is used as an example.

1. INTRODUCTION

In this note we consider the problem of the optimal allocation of treatment for a specific disease when the probability of response is some known function of measurements made on the patient prior to treatment. Let there be k eligible treatments, $T_1, T_2 \ldots T_k$, and suppose a vector $\underset{\sim}{X}$ of measurements is made on each patient leading to probabilities of successful treatment using T_i of $P_i(\underset{\sim}{x})$, $i=1,2,\ldots,k$. We require an optimal procedure for deciding when to use T_i for any given patient.

It is intuitive that T_i should be used when $P_i(\underset{\sim}{x}) \geq P_j(\underset{\sim}{x})$, $j=1,2,\ldots k$. If $P_i(\underset{\sim}{x}) = P_j(\underset{\sim}{x})$ for some $i \neq j$, it is evidently immaterial whether T_i or T_j is used and the decision of which to apply should be made on other grounds.

We show that the above procedure is optimal in the sense that it maximises the expected probability of treatment response. Some extensions of these ideas are suggested and the specific application of the results to the treatment of women with advanced breast cancer discussed.

It will be assumed that $\underset{\sim}{X}$ has a distribution function $\Phi(\underset{\sim}{x})$. In practice, Φ does not have to be known for the procedures to be applied.

2. METHODS

We need the following standard result stated by Rao (1965, p377).

Lemma

Let $f_i(\underset{\sim}{x})$, $i=1,2,\ldots,k$, $f_i(\underset{\sim}{x}) \neq f_j(\underset{\sim}{x})$ $i \neq j$ *be* μ *measurable functions mapping* E_n *to* E_1 *and such that* $\int_{E_n} |f_i(\underset{\sim}{x})| d\mu < \infty$ *for all* i. *If* $A_1, A_2, \ldots, A_k$ *is a partition of* E_n, A_i μ *measurable for all* i, *then the integral* $\sum_{i=1}^{k} \int_{A_i} f_i(\underset{\sim}{x}) d\mu$ *is a maximum for* $A_i = R_i$ *where* $\underset{\sim}{x} \in R_i$ *implies that* $f_i(\underset{\sim}{x}) \geq f_j(\underset{\sim}{x})$ *for all* j *and* $R_1, R_2, \ldots, R_k$ *is a partition of* E_n.

Proof

$$\sum_{i=1}^{k} \int_{R_i} f_i(\underset{\sim}{x}) d\mu = \sum_{i=1}^{k} \sum_{j=1}^{k} \int_{R_i \cap A_j} f_i(\underset{\sim}{x}) d\mu$$

$$\geq \sum_{i=1}^{k} \sum_{j=1}^{k} \int_{R_i \cap A_j} f_j(\underset{\sim}{x}) d\mu$$

$$= \sum_{j=1}^{k} \sum_{i=1}^{k} \int_{R_i \cap A_j} f_j(\underset{\sim}{x}) d\mu = \sum_{j=1}^{k} \int_{A_j} f_j(\underset{\sim}{x}) d\mu.$$

We now apply the above result to the problem of defining optimal patient selection criteria. If $\underset{\sim}{X}$ is n-dimensional, we need regions in n-space, A_i, such that when $\underset{\sim}{x} \in A_i, T_i$ is used as the treatment for the particular patient. It is required to choose A_i so that $\&(A_1, \ldots, A_k)$ is a maximum, where

$$\&(A_1, \ldots, A_k) = \sum_{i=1}^{k} \int_{A_i} P_i(\underset{\sim}{x}) d\Phi(\underset{\sim}{x}).$$

Using the lemma, let $f_i(\underset{\sim}{x}) = P_i(\underset{\sim}{x})$ and $d\mu = d\Phi$; $A_i = R_i$ where $\underset{\sim}{x} \in R_i$ implies that $P_i(\underset{\sim}{x}) \geq P_j(\underset{\sim}{x})$ for all j and $R_1, \ldots, R_k$ is a partition of E_n. The proportion of patients to get T_i is α_i, where

$$\alpha_i = \int_{R_i} d\Phi(\underset{\sim}{x}).$$

A special case of interest is when $k=2$ and $P_2(\cdot)$ is independent of $\underset{\sim}{x}$. Now $R_1 = \{\underset{\sim}{x} : P_1(\underset{\sim}{x}) \geq P_2\}$ and

$$\mathcal{E}(R_1, R_2) = (1-\alpha_1)P_2 + \int_{R_1} P_1(\underset{\sim}{x})\, d\Phi(\underset{\sim}{x}).$$

3. EXTENSIONS

There are obvious extensions of the above ideas. For instance, if the disease is potentially lethal, let L be the survival time following treatment. It can be anticipated that the density function for L will depend on $\underset{\sim}{x}$ and $T_i, g_i(\ell;\underset{\sim}{x})$ say. In practice, after T_i is given, further treatments may be applied. Thus g_i summarises the overall effect of giving T_i first, followed by additional treatment, on the survival pattern.

Put $\Pi_i(\underset{\sim}{x}) = \int_0^\infty \ell . g_i(\ell;\underset{\sim}{x})\, d\ell$, $i=1,2,\ldots,k$. Then $E[L]$ is a maximum, where $E[L] = \sum_{i=1}^{k} \int_{R_i} \Pi_i(\underset{\sim}{x})\, d\Phi(\underset{\sim}{x})$, when $R_i = \{\underset{\sim}{x} : \Pi_i(\underset{\sim}{x}) \geq \Pi_j(\underset{\sim}{x})\ \forall_j\}$.

4. APPLICATION

As a particular example we cite the treatment of women with advanced breast cancer by adrenalectomy, T_1 and radiotherapy, T_2. Previous work of Tallis, Sarfaty and Leppard (1974) has investigated the relationship between the results of adrenalectomy and measurements, $\underset{\sim}{X}$, made prior to the operation. These measurements, for example, include the age of the woman when reporting with breast cancer and the level of various steroid concentrations. The probability of clinical remission as a result of adrenalectomy was estimated as a function of these variables. As a consequence of this modelling, a score $U = u$ is assigned to each woman based on her particular vector of measurements, where $U \sim N(0,1)$, in the population of all women with breast cancer. The probability of remitting as a result of T_1 given $U = u$ is then

$$P_1(u) = 1-\Phi\{(a-\rho u)/(1-\rho^2)^{\frac{1}{2}}\},\quad \Phi'(y) = \phi(y) = \exp\{-y^2/2\}/\sqrt{2\Pi},$$

where a and ρ are parameters intrinsic to $\underset{\sim}{X}$ and remission to T_1.

Estimates of a and ρ are .466 and .7121 respectively, (cited reference).

Remission to T_2 is not known as a function of $\underset{\sim}{X}$ but is about .25 on a group basis. Now the equation

$$P_1(u) = 1 - \phi\left(\frac{.466-.7121u}{\sqrt{1-.7121^2}}\right) = .25$$

is satisfied for $u \approx 0$, and hence all patients with $U > 0$ have T_1 and those with $U \leq 0$ have T_2. The overall performance of the procedure is measured by the average remission rate

$$\&(R_1, R_2) = .25 \times .5 + .5 \int_0^\infty P_1(u)\phi(u)du = .40,$$

since $a_1 = \int_0^\infty \phi(u)du = .5$ and $\int_0^\infty P_1(u)\phi(u)du = .54$ by numerical integration. This compares with the overall remission to adrenalectomy, T_1, of .32.

BIBLIOGRAPHY

Rao, C.R. (1965). Linear Statistical Inference and Its Applications. New York: Wiley.

Tallis, G.M., Sarfaty, G., Leppard, P. (1974). A general classification model with specific application to response to adrenalectomy in women with breast cancer. Computers & Biomed. Res. 7 (5).

B[9]

THE USE OF A PROBABILITY MODEL FOR THE CONSTRUCTION OF AGE SPECIFIC LIFE TABLES FOR WOMEN WITH BREAST CANCER

BY

G.M. TALLIS, G. SARFATY, P. LEPPARD

JUNE 1973

UNIVERSITY OF ADELAIDE

AND

CANCER INSTITUTE, MELBOURNE

With the co-operation of the Anti-Cancer Council of Victoria

THE USE OF A PROBABILITY MODEL FOR THE CONSTRUCTION OF AGE SPECIFIC LIFE TABLES FOR WOMEN WITH BREAST CANCER.

G.M. TALLIS,[+] G. SARFATY,[*] P. LEPPARD [+]

[+] DEPARTMENT OF STATISTICS, UNIVERSITY OF ADELAIDE, ADELAIDE

[*] ENDOCRINE RESEARCH UNIT, CANCER INSTITUTE, MELBOURNE.

FOREWORD

These tables were produced to assist with the study of breast cancer by the Endocrine Research Unit, Cancer Institute, Melbourne. The technology required for their calculation was developed by the Department of Statistics, University of Adelaide, and the necessary data were supplied by the Central Cancer Registry of the Anti-Cancer Council of Victoria.

CONTENTS

APPENDICES

I Introduction

Breast cancer is the commonest malignancy of women in middle-life and accounts for about ⅓ of all cancers of women (Ross, 1969). The mortality rate in Australia, as elsewhere in the world, is approximately 240 per 1,000,000 (Commonwealth Bureau of Census & Statistics) and has remained unchanged for the last 20 years.

Since the incidence of this disease has not altered appreciably over the years, this appears to show that the various treatments adopted have been ineffective, despite advances in surgical and other techniques. Research indicates that growth of breast cancers is closely related to the quantitative secretions of the steroid synthesising endocrine glands.

As the development, growth and senescence of the endocrine glands is age related it is important to know what age structure exists in the mortality due to breast cancer. It is also desirable to know if tumour types, disease stage or treatment affect age specific morbidity.

To develop this type of information requires initial access to a large body of data on breast cancer in women. This was obtained from the Central Cancer Registry, Melbourne, [Rankin, 1971], through the co-operation of Drs. N. Gray and

R. Rankin of the Anti-Cancer Council of Victoria.

A project was therefore set up with aims to:

1. Construct a full set of life tables from registry records 1946 to 1970;
2. Disassociate the contribution of the disease to mortality from that of other causes so that the relationship of morbidity to age could be studied.
3. Apply the technique to an assessment of tumour histology, stage and treatment.

The subsequent discussion will be primarily concerned with the technical problems of achieving these three objectives.

II THE DATA

(a) The Layout

The form of recording relevant information on each registrant has been standardized by the Registry and is reported elsewhere Rankin (1971). The pertinent available data from the point of view of constructing the life tables were;

1. the year of registration;
2. the age at registration (see Table I)
3. the number of years surviving registration; women living longer than 15 years after registration are simply recorded as having survived this period;
4. the estimated time delay, subsequently referred to as lag, between first observing the disease and registration (not available for all registrants).

TABLE I

Age Distribution at Registration

Age	<30	30-39	40-49	50-59	60-69	70-79	>80
Frequency	62	599	1774	1988	2280	1746	647

Further information allowed the data to be subdivided according to various factors into subsets but a certain basic layout of all subsets of the data was followed consistently throughout the analyses. All women who had been observed for

a period of fifteen years or more were classified according to the number of years surviving registration. This process was repeated for all women who had been observed for fourteen years, and so on. The following table summarises the procedure.

TABLE II

General Data Layout

Years observed	Total	No. dying in xth year after registration x=1	2	.	.	.	.	.	14	15
1	N_1	$n_{1,1}$								
2	N_2	$n_{2,1}$	$n_{2,2}$							
.	.	.	.							
.	.	.	.							
.	.	.	.							
14	N_{14}	$n_{14,1}$	$n_{14,2}$	.	.	.			$n_{14,14}$	
15	N_{15}	$n_{15,1}$	$n_{15,2}$	.	.	.			$n_{15,14}$	$n_{15,15}$

It should also be noted that

$$\sum_{j=1}^{i} n_{ij} + s_i = N_i \qquad i = 1,\ldots 15$$

where s_i = number surviving period i. A numerical example of Table II is given in Section III (Table V).

Women who were lost to the Registry due to incomplete follow up were entered in the row corresponding to the number of years they survived before their record was terminated.

The data were then corrected for lag as outlined in the next section and tables analogous to Table II constructed. Hence the period of observation effectively dates from when the disease was first discovered by the patient.

Note that the procedure of the previous paragraph requires the assumption that the data are stationary from the point of view of mortality. This assumption was tested and found to be reasonable, see Section IV (a).

Let $P' = (P_1, P_2, \ldots, P_{15})$, $\bar{p}_{ij} = n_{ij}/N_i$, $M_j = \sum_{i=j}^{15} N_i$

and set

$$\bar{P}_j = \sum_{i=j}^{15} w_{ji}\bar{p}_{ij}, \quad w_{ji} = N_i/M_j.$$

Then, putting $\bar{\underset{\sim}{P}}' = (\bar{P}_1, \bar{P}_2, \ldots, \bar{P}_{15})$

$$E[\bar{\underset{\sim}{P}}] = \underset{\sim}{P}$$

where $E[\ ]$ is the expectation operator. Obviously, P_j is the probability that any woman, satisfying the requirements of the subset, will die in the jth year after first observing the disease, $1 - \sum_{j=1}^{15} P_j$ is the probability of surviving 15 years after the year of first observation and $N = \sum_{i=1}^{15} N_i$ is the total number of women in the subset.

Write $\bar{\underset{\sim}{p}}_i' = (\bar{p}_{i1}, \bar{p}_{i2}, \ldots, \bar{p}_{ii})$ where, of course,

$$E[\bar{\underset{\sim}{p}}_i'] = \underset{\sim}{P}_i' = (P_1, P_2, \ldots, P_i)$$

and the covariance matrix of $\bar{\underset{\sim}{p}}_i$, $V[\bar{\underset{\sim}{p}}_i] = V_i$ is given by

$$N_i V_i = \begin{bmatrix} P_1(1-P_1) & -P_1P_2 & \cdots & -P_1P_1 \\ -P_2P_1 & P_2(1-P_2) & \cdots & -P_2P_1 \\ \cdot & \cdot & & \\ -P_1P_1 & -P_1P_2 & \cdots & P_1(1-P_1) \end{bmatrix}.$$

Now let $\bar{\underset{\sim}{p}} = (\bar{\underset{\sim}{p}}_1', \bar{\underset{\sim}{p}}_2', \ldots, \bar{\underset{\sim}{p}}_{15}')$ and

$$W = \begin{bmatrix} w_{1,1} & w_{1,2} & 0 & w_{1,3} & 0 & 0 & w_{1,4} & 0 & \cdots \\ 0 & 0 & w_{2,2} & 0 & w_{2,3} & 0 & 0 & w_{2,4} & \cdots \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \end{bmatrix}$$

then $\bar{\underset{\sim}{P}} = W\,\bar{\underset{\sim}{p}}$, $E[\bar{\underset{\sim}{P}}] = \underset{\sim}{P}$ and

$$V[\bar{\underset{\sim}{P}}] = W\ V[\bar{\underset{\sim}{p}}]W'$$

where $$V[\bar{\underset{\sim}{p}}] = V_1 \oplus V_2 \oplus \ldots \oplus V_{15}.$$

If W is replaced by another row stochastic matrix an unbiased estimator of $\underset{\sim}{P}$ will result. But W has been chosen to minimise trace $V[\bar{\underset{\sim}{P}}]$ within the class of linear unbiased estimators. This follows from the observation that the ith element on the diagonal of $V[\bar{\underset{\sim}{P}}]$ is

$$V[\sum_{j=1}^{15} w_{1j}\bar{P}_{j1}] = [P_1(1-P_1)]\underset{\sim}{w}_1'\mathcal{N}_1\underset{\sim}{w}_1$$

where $\underset{\sim}{w}_1' = (w_{1,1}, w_{1,1+1}, \ldots, w_{1,15})$ is to be found and $\mathcal{N}_1^{-1} = \text{diag}(N_1, N_{1+1}, \ldots, N_{15})$. The minimum of $\underset{\sim}{w}_1'\mathcal{N}_1\underset{\sim}{w}_1$ subject to $\underset{\sim}{1}'\underset{\sim}{w}_1 = 1$ is obtained by the usual Lagrange methods. Thus

$$2^{-1}\frac{\partial}{\partial \underset{\sim}{w}_1}[\underset{\sim}{w}_1'\mathcal{N}_1\underset{\sim}{w}_1 + 2\lambda\underset{\sim}{w}_1'\underset{\sim}{1}] = \mathcal{N}_1\underset{\sim}{w}_1 + \lambda\underset{\sim}{1}$$

and equating this result to $\underset{\sim}{0}$ and solving shows that $\underset{\sim}{w}_1 = \lambda\mathcal{N}_1^{-1}\underset{\sim}{1}$ where λ is chosen so that $\underset{\sim}{1}'\underset{\sim}{w}_1 = 1$. Clearly, $\underset{\sim}{w}_1' = (N_1, N_{1+1}, \ldots, N_{15}) \,/\, \sum_{j=1}^{15} N_j$.

(b) Time Delay to Registration

Actuarial calculations are pertinent if dated from the age of first observing the disease rather than from the age of registration, and a method was devised to correct for this lag. Unfortunately accurate estimates of this time delay are not in general available, but Registry records provided the following frequency distribution.

TABLE III

Distribution of Delay to Registration

Lag (in years)	0-.5	.5-1	1-2	2-5	>5	no information
Frequency	4392	1155	594	922	455	1578

The density function

$$f(x) = \begin{cases} k\,\dfrac{\alpha\, e^{-\alpha x}}{1-e^{-\alpha}} & 0 \leq x \leq 1 \\ (1-k)\beta\, e^{-\beta(x-1)} & x > 1 \end{cases}$$

where $0 < k < 1$ and $\alpha, \beta > 0$.

was fitted to the data and the maximum likelihood estimates of the parameters were

$$\hat{k} = .738, \quad \hat{\alpha} = 2.664, \quad \hat{\beta} = .365.$$

The model describes the data extremely well as can be seen by comparing the observed probabilities with those estimated from the fitted model. The estimated mean lag is 1.36 years.

TABLE IV

Observed and Estimated Lag Probabilities

Lag (years)	0-.5	.5-1	1-2	2-5	>5
observed proportion	.584	.154	.079	.122	.061
model probability	.584	.154	.080	.121	.061

This estimated density, together with Registry values for individual lags, ages, and length of survival after registration were used to calculate modified layout tables (see Section II (a))dated from time of observation. Full details are omitted.

III The Model

With the objectives detailed in the Introduction in mind, a suitable model describing the distribution of length of life for women with breast cancer had to be found. A competing risk type model was set up and can be rationalised in the following way.

It was assumed that there are two competing causes of death for women with breast cancer; the disease itself, and other causes. The model assumes that these two risks act independently in the following way;

$$F(x+\Delta x) = F(x)[1-\phi_c(x)\Delta x][1-\phi_o(a+x)\Delta x] + O(\Delta x) \qquad (1)$$

where X is the length of survival of a woman first observing breast cancer at age a, and ϕ_c, ϕ_o are the forces of mortality due to the disease and other causes respectively. Equation (1) leads to the differential equation

$$F'(x) = -F(x)[\phi_c(x) + \phi_o(a+x)]$$

with solution

$$F(x) = 1 - \exp\left\{-\int_0^x [\phi_c(t) + \phi_o(a+t)]dt\right\} \qquad (2)$$

The function $F(x) = \Pr\{X \leq x\}$ specifies a distribution function for X in terms of the two components of the overall force of mortality function $\phi(x) = \phi_c(x) + \phi_o(a+x)$, where

the origin is taken as the time of first observing the disease.

The function ϕ_o was obtained from the Australian Life Tables (1960-1962) for women, after a trivial correction for the prevalence of breast cancer had been applied. A function of the form $\phi_o(t) = \gamma e^{\delta t}$ was fitted to the corrected tabular values and the parameters γ and δ had numerical values 3.607×10^{-5} and 9.261×10^{-2} respectively. The actual fit was excellent as can be verified by comparing the life expectancies calculated by numerical quadrature from

$$F_o(x) = 1 - \exp\left\{ - \int_0^x \phi_o(t)\,dt \right\}$$

with those of the life table. In most cases the two figures agree to within .20 of a year.

The parametric form of ϕ_c, $\phi_c(x,\underset{\sim}{\theta})$ say, was chosen to suit the emperical force of mortality function. The latter was calculated from the formula

$$\bar{\phi}_c(j-\tfrac{1}{2}) = \bar{P}_j/[1-(\bar{F}_{j-1}+\bar{F}_j)/2] - \phi_o(\bar{a}+j-\tfrac{1}{2})$$

where $\bar{F}_o = 0$, $\bar{F}_j = \sum_{i=1}^{j} \bar{P}_i$ and $\bar{a}$ is the average age of the group. The term $(\bar{F}_{j-1}+\bar{F}_j)/2$ estimates $F(j-\frac{1}{2})$ and $\bar{P}_j$ is an estimate of $f(j-\frac{1}{2}) = F'(j-\frac{1}{2})$, by the Mean Value Theorem. Thus $\bar{\phi}_c(j-\frac{1}{2})$ estimates $f(j-\frac{1}{2})/(1-F(j-\frac{1}{2}))$ as required. The estimates $\bar{P}_j$ were calculated on the combined data of 9096 women where $\bar{a} = 58.86$.

TABLE V

Layout for the Combined Group

Years Observed	Total	No. dying in the xth year after Registration														
		x=1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1	246	41														
2	372	50	59													
3	429	59	53	40												
4	447	41	67	50	36											
5	461	51	63	50	27	19										
6	538	58	67	45	53	36	17									
7	495	57	56	45	39	31	28	16								
8	523	49	60	66	37	31	29	24	13							
9	544	47	67	55	44	45	18	22	17	5						
10	491	56	47	50	43	29	23	23	20	12	12					
11	456	47	60	40	42	30	24	28	21	13	11	10				
12	483	58	58	54	41	28	36	30	17	15	15	8	12			
13	438	46	63	41	32	40	29	25	20	16	9	9	8	8		
14	409	63	57	37	33	27	19	24	14	25	13	8	4	8	6	
15	2764	333	364	279	233	206	178	142	123	94	76	90	87	60	54	46
$\bar{P}$		.116	.129	.101	.082	.069	.056	.051	.040	.032	.027	.028	.027	.021	.019	.017
$\bar{\phi}_c$		.114	.147	.131	.115	.103	.104	.088	.074	.065	.072	.079	.061	.057	.049	

The general form of $\phi_c(x,\underset{\sim}{\theta})$ fitted to all groups of data was

$$\phi_c(x,\underset{\sim}{\theta}) = \begin{cases} \theta_0 & 0 \leqslant x \leqslant 1 \\ e^{-\theta_1-\theta_2 x} & x > 1 \end{cases}$$

IV Fitting the Model

(a) Estimation of $\underset{\sim}{\theta}$.

Use can be made of the $\phi_c(x)$ of the previous section to estimate $\underset{\sim}{\theta}$. Thus, for instance, $\ln \bar{\phi}_c(\frac{1}{2})$ estimates θ_0, θ_0^0 say, and $-\ln \bar{\phi}_c(i-\frac{1}{2})$ estimates $\theta_1 + \theta_2.(i-\frac{1}{2})$ for $i \geqslant 2$. The parameters θ_1 and θ_2 can be fitted by least squares using $-\ln \bar{\phi}_c(i-\frac{1}{2})$, $i = 2,3,\ldots 15$ to give estimates θ_1^0 and θ_2^0.

Once $\underset{\sim}{\theta}^0$ is to hand, $F(x,\underset{\sim}{\theta}^0)$ can be calculated and also $P_j(\underset{\sim}{\theta}^0)$, $j = 1,2,\ldots 15$, where

$$P_j(\underset{\sim}{\theta}^0) = F(j,\underset{\sim}{\theta}^0) - F(j-1,\underset{\sim}{\theta}^0).$$

The statistic

$$A(\underset{\sim}{\theta}^0, V^{-1}[\underset{\sim}{\bar{P}}]) = [\underset{\sim}{P}(\underset{\sim}{\theta}^0)-\underset{\sim}{\bar{P}}]' V^{-1}[\underset{\sim}{\bar{P}}][\underset{\sim}{P}(\underset{\sim}{\theta}^0)-\underset{\sim}{\bar{P}}]$$

is a measure of agreement between $\underset{\sim}{P}(\underset{\sim}{\theta}^0)$ and $\underset{\sim}{\bar{P}}$. In fact it is shown in Section VI that

$$\inf_{\underset{\sim}{\theta}} A(\underset{\sim}{\theta}, V^{-1}[\underset{\sim}{\bar{P}}]) = A(\underset{\sim}{\theta}^*, V^{-1}[\underset{\sim}{\bar{P}}])$$

is asymptotically distributed as a chi-square variate with 12 degrees of freedom.

The vector $\underset{\sim}{\theta}^{o}$ provides a starting value from which $\underset{\sim}{\theta}^{*}$ can be found by means of a minimisation program, Nelder and Mead (1965). The matrix $V[\overline{\underset{\sim}{P}}]$, although unknown, is estimated by using the consistent estimates $\overline{\underset{\sim}{P}}$ of $\underset{\sim}{P}(\underset{\sim}{\theta})$ without affecting the distribution theory.

It is found that $\underset{\sim}{\theta}^{o}$ provides a satisfactory starting vector for all minimisations in various sub-groups of the data, whether or not these are formed by age, tumour histology, stage of disease at reporting or the type of treatment given. For this particular work, the rate of convergence of the minimisation procedure is rapid.

The stationarity of the data was examined by comparing the pre-1960 registrants with those who registered after 1960. Models of the above form were fitted to each group and it was found that there were no important differences in the respective parameters, suggesting that there has been no significant change in the mortality pattern of the disease with time. This conforms with observations elsewhere.

(b) The Relationship of Age to $\underset{\sim}{\theta}$.

If the complete set of data is sorted into subgroups according to age, a, it is found that $\underset{\sim}{\theta}^{*}$ varies quite appreciably with a. Hence we put

$$\underset{\sim}{\theta}(a) = B\,\underset{\sim}{a}$$

where

$$B = \begin{bmatrix} \beta_{00}, \beta_{01}, \ldots \beta_{0r} \\ \beta_{10}, \beta_{11}, \ldots \beta_{1r} \\ \beta_{20}, \beta_{21}, \ldots \beta_{2r} \end{bmatrix}, \quad \underset{\sim}{a}' = (1, a, a^2, \ldots a^r)$$

for suitable choice of B and r.

The following procedure has been adopted to estimate B from the data for a given r. The registrants are linearly ordered according to age at reporting and then roughly the 2,000 youngest women form the first age group. Second and subsequent groups are formed by dropping out approximately the 500 youngest members of the previous group and adding approximately the next 500 older women, and so on. The data are worked through in this way until the oldest possible group of 2,000 is formed. The resulting groups, say k in number, with average ages at reporting a_i, $i = 1,2,\ldots k$ are then separately subjected to the estimation procedure as described in the previous section.

At each age $\bar{a}_i$ the estimate $\underset{\sim}{\theta}_i^*$ and its covariance matrix $V[\underset{\sim}{\theta}_i^*] = \Sigma_{ii}$ are obtained. From the results of Section VI

$$\Sigma_{ii} = (Q' V^{-1}[\underset{\sim}{\bar{P}}_i] Q)^{-1}$$

where $q_{st} = \partial P_s(\underset{\sim}{\theta})/\partial\theta_t$ and $V[\underset{\sim}{\bar{P}}_i]$ is the

covariance matrix of $\bar{\underset{\sim}{P}}_i$, $\bar{\underset{\sim}{P}}_i$ being calculated from the appropriate lay-out of the ith age-group.

Now $\underset{\sim}{\theta}(a) = B\underset{\sim}{a}$ can be written in the form

$$\underset{\sim}{\theta}(a) = A\underset{\sim}{\beta}$$

where $A = (I \otimes \underset{\sim}{a}')$ and $\underset{\sim}{\beta}' = (\beta_{00},\ldots\beta_{0r},\beta_{10},\ldots\beta_{1r},\beta_{20},\ldots\beta$

Let

$$\underset{\sim}{\theta}_i^* = A_i\underset{\sim}{\beta} + \underset{\sim}{\varepsilon}_i \quad E[\underset{\sim}{\varepsilon}_i] = \underset{\sim}{0}, \quad V[\underset{\sim}{\varepsilon}_i] = \Sigma_{ii}$$

then it is required to estimate $\underset{\sim}{\beta}$ from the $\underset{\sim}{\theta}_i^*$ for the various ages $\bar{a}_i$.

If $Q(\underset{\sim}{\beta}) = \sum_{i=1}^{k} (\underset{\sim}{\theta}_i^* - A_i\underset{\sim}{\beta})' \Sigma_{ii}^{-1} (\underset{\sim}{\theta}_i^* - A_i\underset{\sim}{\beta})$

then Q can be minimised with respect to $\underset{\sim}{\beta}$ to give the estimate

$$\hat{\underset{\sim}{\beta}}_0 = \left(\sum_{i=1}^{k} A_i'\Sigma_i^{-1}A_i\right)^{-1} \sum_{i=1}^{k} A_i'\Sigma_i^{-1}\underset{\sim}{\theta}_i^*.$$

Because of data overlap, the $\underset{\sim}{\theta}_i^*$ are not all independent and $\hat{\underset{\sim}{\beta}}_0$ is not an efficient estimator, although it is unbiassed. The following table gives the age boundaries and the numbers in the various age groups for calculating the $\underset{\sim}{\theta}_i^*$.

TABLE VI

Age Boundaries	Numbers
1-46	2078
38-48	1909
42-51	1997
46-54	1807
49-58	1981
53-61	1851
56-64	1975
60-67	1834
63-71	1988
66-74	1843
69-79	1806
71-100	1931

Let

$$X = \begin{bmatrix} A_1 \\ A_2 \\ \vdots \\ A_k \end{bmatrix} \qquad \underset{\sim}{\hat{\theta}} = \begin{bmatrix} \theta_1^* \\ \theta_2^* \\ \vdots \\ \theta_k^* \end{bmatrix}$$

and put

$$\Sigma_o = (\underset{\sim}{\hat{\theta}}-X\underset{\sim}{\hat{\beta}}_o)(\underset{\sim}{\hat{\theta}}-X\underset{\sim}{\hat{\beta}}_o)' = \begin{bmatrix} \Sigma_{11} & \cdots & \Sigma_{1k} \\ \vdots & & \\ \Sigma_{k1} & \cdots & \Sigma_{kk} \end{bmatrix}.$$

The above table shows that $\underset{\sim}{\theta}_i^*$ is independent of $\underset{\sim}{\theta}_j^*$ for $|j-i| > 2$. Thus Σ_o can be simplified since all but the diagonal and first and second off-diagonal blocks are 0.

Put $D_i = \text{diag}(\sigma_{i0}, \sigma_{i1}, \sigma_{i2})$, where σ_{ij} is the jth element of the diagonal of Σ_{ii}, and define $R_{ij} = D_i^{-\frac{1}{2}} \Sigma_{ij} D_j^{-\frac{1}{2}}$ for $j = i+1, i+2$ and let $\bar{R}_1 = \sum_{i=1}^{k-1} R_{i,i+1}/(k-1)$ and $\bar{R}_2 = \sum_{i=1}^{k-2} R_{i,i+2}/(k-2)$. If $\bar{\Sigma}_{ii+1} = D_i^{\frac{1}{2}} \bar{R}_1 D_{i+1}^{\frac{1}{2}}$ and $\bar{\Sigma}_{ii+2} = D_i^{\frac{1}{2}} \bar{R}_2 D_{i+2}^{\frac{1}{2}}$, then construct $\bar{\Sigma}$, say, using $\bar{\Sigma}_{ij} = 0$ for $|i-j| > 2$, $\bar{\Sigma}_{ii} = \Sigma_{ii}$ and $\bar{\Sigma}_{ii+1}$, $\bar{\Sigma}_{ii+2}$ as above. A new estimate $\hat{\beta}_1$ is obtained by minimising

$$Q(\underset{\sim}{\beta}) = (\underset{\sim}{\hat{\theta}}-X\underset{\sim}{\beta})'\bar{\Sigma}^{-1}(\underset{\sim}{\hat{\theta}}-X\underset{\sim}{\beta})$$

with respect to $\underset{\sim}{\beta}$. The whole process is iterated until no significant change in $\underset{\sim}{\hat{\beta}}_i$ is evident. The procedure converges rapidly in two or three iterations to $\underset{\sim}{\hat{\beta}}$ say.

The assumption on which the above technique rests is that the correlation structure between $\underset{\sim}{\theta}_i^*$, $\underset{\sim}{\theta}_{i+1}^*$ and $\underset{\sim}{\theta}_{i+2}^*$ is the same for all i. This is a standard assumption in time series analysis.

If $\hat{\Sigma}$ is the estimate of Σ corresponding to $\hat{\beta}$

$$\hat{\underset{\sim}{\theta}}(a) = A\hat{\underset{\sim}{\beta}}$$

and

$$V[\hat{\underset{\sim}{\theta}}(a)] = A[X\hat{\Sigma}^{-1}X']^{-1}A' .$$

Thus finally, the appropriate estimate of $F(x,\underset{\sim}{\theta})$ for a woman reporting breast cancer at age a is $F(x,\hat{\underset{\sim}{\theta}}(a))$.

The appropriate degree of the polynomials, r, to use can be determined by comparing the sum of the residuals $A(\underset{\sim}{\theta}_1^*, V^{-1}[\underset{\sim}{P}_1])$ to the sum of the residuals obtained by using $\hat{\underset{\sim}{\theta}}(a)$ in place of the $\underset{\sim}{\theta}_1^*$. It was found in the current data that with r=4 these two sums were 166.99 and 179.91 respectively, showing satisfactory agreement. Smaller values of r produced unsatisfactorily wide discrepancies between the sums.

Extrapolation at both ends of the age range is necessary in order to obtain results for women ranging from 35 to 85 years. Because of the polynomial approximation $\hat{\underset{\sim}{\theta}}(a)$, and because the sizes of the two terminal estimation groups are of the order of 2,000, with average ages greater than 40 and less than 80, respectively, it was necessary to examine the younger and older age structures more carefully. Two separate analyses were carried out by the above procedures using group sizes of 500 instead of 2000 and shifting the position of estimation by approximately 250 instead of 500 as

described earlier. In this way, essentially the first two groups and the last two groups of Table I were reanalysed.

Thus, polynomial approximations $\underset{\sim}{\vartheta}_1(a)$ valid for $35 \leqslant a \leqslant 43$, $\underset{\sim}{\vartheta}_2(a)$ valid for $43 < a < 75$ and $\underset{\sim}{\vartheta}_3(a)$ valid for $75 \leqslant a \leqslant 85$ were arrived at where $\underset{\sim}{\vartheta}_i(a)$, $i=1,3$ arose from the analysis of the end groups and $\underset{\sim}{\vartheta}_2(a)$ was obtained from the large group analysis.

$35 \leqslant a \leqslant 43$

$$\vartheta_0(a) = 1.17 - 5.24 \times 10^{-2}a + 6.24 \times 10^{-4}a^2$$

$$\vartheta_1(a) = 4.16 - 1.13 \times 10^{-1}a + 1.39 \times 10^{-3}a^2$$

$$\vartheta_2(a) = 3.17 \times 10^{-2} + 2.62 \times 10^{-3}a - 2.29 \times 10^{-5}a^2$$

$43 < a < 75$

$$\vartheta_0(a) = 7.29 - 5.32 \times 10^{-1}a + 1.44 \times 10^{-2}a^2 - 1.70 \times 10^{-4}a^3 + 7.34 \times 10^{-7}a^4$$

$$\vartheta_1(a) = -22.11 + 1.95a - 5.72 \times 10^{-2}a^2 + 7.22 \times 10^{-4}a^3 - 3.32 \times 10^{-6}a^4$$

$$\vartheta_2(a) = 4.04 \times 10^{-3} - 3.49 \times 10^{-2}a + 2.04 \times 10^{-3}a^2 - 3.63 \times 10^{-5}a^3 + 2.06 \times 10^{-7}a^4$$

$75 \leqslant a \leqslant 85$

$$\vartheta_0(a) = -3.39 + 8.54 \times 10^{-2}a - 5.16 \times 10^{-4}a^2$$

$$\vartheta_1(a) = -15.00 + 4.45 \times 10^{-1}a - 2.93 \times 10^{-3}a^2$$

$$\vartheta_2(a) = 4.96 - 1.22 \times 10^{-1}a + 7.53 \times 10^{-4}a^2$$

(c) Testing the Model

In order to test the adequacy of the fitting procedure the combined group of women was broken into 6 non-intersecting subsets with average ages 38.9, 49.0, 57.1, 64.6, 71.8 and 81.3. Each group was used to estimate a value of $\underset{\sim}{\theta}^*$ and the associated residuals were summed giving a value of 82.25. This value, being the sum of 6 independent chi-square variates, is itself distributed as a chi-square on 72 degrees of freedom, and is not significant (χ^2_{72} (5%) = 92.8). The corresponding total residual using $\hat{\underset{\sim}{\theta}}(a)$ in place of $\underset{\sim}{\theta}^*$ for each group is 91.4.

As a further comparison, the combined data were analysed using both Cutler's method (1958) for constructing life tables and the procedure outlined previously. The results of the two methods, denoted A and B respectively, are shown in Table VII.

The agreement between the two methods for estimating the proportions of survivors is remarkable. However, the parametric approach has the advantage in that the functional form of $F(x)$ is determined allowing expectations to be calculated. These are unobtainable from Method A. Moreover, the partitioning of the force of mortality into the required components ϕ_c and ϕ_o, where ϕ_c is a function of the age of first observing the disease, can be negotiated using Method B. Method A allows no such flexibility and, in fact, this procedure cannot be used to construct life tables of the type reported here

TABLE VII

Comparison of Methods A and B for Estimating Pr{live≥x}

x	A	B
1	.886	.884
2	.759	.756
3	.660	.654
4	.580	.571
5	.513	.503
6	.458	.447
7	.409	.399
8	.369	.359
9	.336	.324
10	.308	.294
11	.279	.268
12	.249	.245
13	.225	.224
14	.203	.205
15	.184	.189

V The Tables

A (a) Main Life Tables

For each age, a, of first observing the disease a set of life tables is constructed using $F(x,\underset{\sim}{\theta}(a))$. The second column gives $P_i(\underset{\sim}{\theta}(a)) = F(i,\underset{\sim}{\theta}(a))-F(i-1,\underset{\sim}{\theta}(a))$, an estimate of the probability of dying in the period $(a+i-1,a+i)$ where $i=1,2,\ldots,91-a$.

The fourth column records the conditional probability, Q, of dying in the ith year after reporting, i.e. at age a+i, given the woman has survived i years. Thus,

$$Q_i(\underset{\sim}{\theta}(a)) = P_i(\underset{\sim}{\theta}(a))/(1 - \sum_{j=0}^{i-1} P_j(\underset{\sim}{\theta}(a))),\ P_0(\underset{\sim}{\theta}(a)) = 0.$$

Column six gives the normal life expectancy for a woman who has survived to age A, while column seven gives the comparable figures for women observing breast cancer at age a. In fact,

$$E[\overbrace{BC \mid A=a+i}] = \int_0^\infty [1-F(x+i,\underset{\sim}{\theta}(a))]dx/[1-F(i,\underset{\sim}{\theta}(a))]$$

provides the required estimate.

Columns three, five and eight give the standard errors of the corresponding estimate appearing in the previous column. These figures give some idea of the precision of the tables and they are calculated according to the methods described in Section A (b).

A (b) Precision of Estimation in A(a)

It was shown in Section IV that $\underset{\sim}{\hat{\theta}}(a)$ has covariance matrix

$$V[\underset{\sim}{\hat{\theta}}(a)] = A\, V[\underset{\sim}{\hat{\beta}}]A', \quad A = I \otimes \underset{\sim}{a}'$$

and since all the quantities in the Main Life Tables depend on $F(x,\underset{\sim}{\hat{\theta}}(a))$, it is therefore possible to calculate standard errors for them. The procedures are outlined below.

Put $F(x,\underset{\sim}{\theta}(a)) = F(x,\underset{\sim}{\theta},a)$, then

$$F(x,\underset{\sim}{\theta},a) = \begin{cases} 1 - \exp\{B(x,\underset{\sim}{\theta},a)\} & 0 \leqslant x < 1 \\ 1 - \exp\{A(x,\underset{\sim}{\theta},a)\} & x \geqslant 1 \end{cases}$$

where

$$B(x,\underset{\sim}{\theta},a) = -\theta_0 x - \int_a^{a+x} \phi_0(t)dt$$

$$A(x,\underset{\sim}{\theta},a) = -\theta_0 - \theta_2^{-1}[e^{-\theta_1-\theta_2} - e^{-\theta_1-\theta_2 x}] - \int_a^{a+x} \phi_0(t)dt.$$

The fact that $\underset{\sim}{\theta}$ is a function of a is suppressed.

Then

$$\frac{\partial F}{\partial \theta_0}(x,\underset{\sim}{\theta},a) = \begin{cases} x \exp\{B(x,\underset{\sim}{\theta},a)\} & 0 \leqslant x < 1 \\ \exp\{A(x,\underset{\sim}{\theta},a)\} & x \geqslant 1 \end{cases}$$

$$\frac{\partial F(x,\underset{\sim}{\theta},a)}{\partial \theta_1} = \begin{cases} 0 & 0 \leqslant x < 1 \\ -\exp\{A(x,\underset{\sim}{\theta},a)\}\ C(x,\underset{\sim}{\theta}) & x \geqslant 1 \end{cases}$$

$$\frac{\partial F(x,\underset{\sim}{\theta},a)}{\partial \theta_2} = \begin{cases} 0 & 0 \leqslant x < 1 \\ -\exp\{A(x,\underset{\sim}{\theta},a)\}\theta_2^{-1}[C(x,\underset{\sim}{\theta})-D(x,\underset{\sim}{\theta})] & x \geqslant 1 \end{cases}$$

where

$$C(x,\underset{\sim}{\theta}) = \theta_2^{-1}[e^{-\theta_1-\theta_2}-e^{-\theta_1-\theta_2 x}]$$

$$D(x,\underset{\sim}{\theta}) = [xe^{-\theta_1-\theta_2 x}-e^{-\theta_1-\theta_2}]\text{.}$$

Now let $G = (g_{ij})$, $g_{ij} = \partial F(i,\underset{\sim}{\theta},a)/\partial\theta_j$ and

$$M = \begin{bmatrix} 1 & 0 & \ldots & 0 & 0 \\ -1 & 1 & 0\ldots & 0 & 0 \\ \vdots & & & & \\ 0 & 0 & \ldots & -1 & 1 \end{bmatrix}\text{.}$$

If $Q = (q_{ij})$, $q_{ij} = \partial P_i(\underset{\sim}{\theta})/\partial\theta_j$, then $Q = MG$.
Thus for each of the groups used in constructing the age relationship for $\underset{\sim}{\theta}(a)$

$$V[\underset{\sim}{\theta}^*] = [Q'V^{-1}[\underset{\sim}{\bar{P}}]Q]^{-1} = \Sigma$$

where $V[\underset{\sim}{\bar{P}}]$ is the covariance matrix of the $\bar{P}_j$ for the particular age group concerned. Moreover

$$V[\underset{\sim}{P}(\underset{\sim}{\theta}^{*})] = Q\,\Sigma\,Q'$$

and for the estimator $\underset{\sim}{\hat{\theta}}(a)$,

$$V[\underset{\sim}{P}(\underset{\sim}{\hat{\theta}}(a))] = Q\,A\,V\,[\underset{\sim}{\hat{\beta}}]A'\,Q'.$$

The covariance matrix of the $Q_i(\underset{\sim}{\hat{\theta}}(a))$ is built up from $V[\underset{\sim}{P}(\underset{\sim}{\hat{\theta}}(a))]$ by means of the following expression;

$$\mathrm{Cov}[Q_i(\underset{\sim}{\hat{\theta}}(a)),\ Q_j(\underset{\sim}{\hat{\theta}}(a))] = Q_iQ_j\left\{\frac{\mathrm{Cov}(P_i,P_j)}{P_iP_j} + \sum_{\ell=1}^{j-1}\frac{\mathrm{Cov}(P_i,P_\ell)}{P_i(1-\pi_j)}\right.$$

$$\left. + \sum_{k=1}^{i-1}\frac{\mathrm{Cov}(P_j,P_k)}{P_j(1-\pi_i)} + \sum_{k=1}^{i-1}\sum_{\ell=1}^{j-1}\frac{\mathrm{Cov}(P_\ell,P_k)}{(1-\pi_i)(1-\pi_j)}\right\}$$

where $\pi_i = \sum_{s=0}^{i-1} P_s$ and $\pi_j = \sum_{t=0}^{j-1} P_t$.

Now let $E[X|m,\underset{\sim}{\theta},a]$ be the conditional life expectancy of a woman first observing breast cancer at age a and surviving to $a+m$.

For $m = 0$ we have

$$E[X|0,\underset{\sim}{\theta},a] = \int_0^{\infty}[1 - F(x,\underset{\sim}{\theta},a)]dx$$

and

$$\frac{\partial}{\partial\theta_o}E[X|0,\underset{\sim}{\theta},a] = -\int_0^1 x\,\exp\{B(x)\}dx - \int_1^{\infty}\exp\{A(x)\}dx$$

$$\frac{\partial}{\partial \theta_1} E[X|0,\underset{\sim}{\theta},a] = \int_1^\infty \exp\{A(x)\}C(x)dx$$

$$\frac{\partial}{\partial \theta_2} E[X|0,\underset{\sim}{\theta},a] = \theta_2^{-1} \int_1^\infty \exp\{A(x)\}[C(x)-D(x)]dx$$

For $m > 0$

$$E[X|m,\underset{\sim}{\theta},a] = \int_0^\infty \exp\{A(m+x,\underset{\sim}{\theta},a)-A(m,\underset{\sim}{\theta},a)\}dx$$

$$\frac{\partial}{\partial \theta_0} E[X|m,\underset{\sim}{\theta},a] = 0$$

$$\frac{\partial}{\partial \theta_1} E[X|m,\underset{\sim}{\theta},a] = \int_0^\infty \exp\{A(m+x,\underset{\sim}{\theta},a) - A(m,\underset{\sim}{\theta},a)\} \times [C(m+x,\underset{\sim}{\theta}) - C(m,\underset{\sim}{\theta})]dx$$

$$\frac{\partial}{\partial \theta_2} E[X|m,\underset{\sim}{\theta},a] = \theta_2^{-1} \int_0^\infty \exp\{A(m+x,\underset{\sim}{\theta},a) - A(m,\underset{\sim}{\theta},a)\} \times [C(m+x,\underset{\sim}{\theta}) - D(m+x,\underset{\sim}{\theta}) - C(m,\underset{\sim}{\theta}) + D(m,\underset{\sim}{\theta})]dx.$$

Finally

$$V\{E[X|m,\underset{\sim}{\hat{\theta}},a]\} = \left(\frac{\partial}{\partial \underset{\sim}{\theta}} E[X|m,\underset{\sim}{\theta},a]\right)' V(\underset{\sim}{\hat{\theta}}(a))\left(\frac{\partial}{\partial \underset{\sim}{\theta}} E[X|m,\underset{\sim}{\theta},a]\right)$$

B (a) Tables by Stage, Histology and Stage × Histology

Traditionally Breast Cancer is divided into four distinct stages of the disease, (S_1, S_2, S_3, S_4) and three degree types of tumour malignancy: well differentiated, moderate and anaplastic (H_1, H_2, H_3). It is well known that stage and histology significantly affect survival.

In order to quantify this observation, the Registry data were sorted by histology into three groups corresponding to H_i, $i = 1,2,3$ and by stage into four groups corresponding to S_j, $j = 1,2,3,4$. Parameter vectors $\underset{\sim}{\theta}^i$, $i = 1,2,3$, for histology and ${}^j\underset{\sim}{\theta}$, $j=1,2,3,4$, for stage were calculated according to the methods previously described. The average ages $\bar{a}^i$, ${}^j\bar{a}$ in the seven groups varied (see Table VIII) and it was decided to correct each set of parameters to the age of 60. This was accomplished according to the formulae

$$\underset{\sim}{\theta}^i(60) = \underset{\sim}{\theta}^i + \underset{\sim}{\partial}(60) - \underset{\sim}{\partial}(\bar{a}^i) \qquad i = 1,2,3$$

$${}^j\underset{\sim}{\theta}(60) = {}^j\underset{\sim}{\theta} + \underset{\sim}{\partial}(60) - \underset{\sim}{\partial}({}^j\bar{a}) \qquad j = 1,2,3,4$$

Once the $\underset{\sim}{\theta}^i(60)$ and ${}^j\underset{\sim}{\theta}(60)$ were known, suitable estimates of $V[\underset{\sim}{\theta}^i(60)]$ were calculated from the formula

$$V[\underset{\sim}{\theta}^i(60)] = V(\underset{\sim}{\theta}^i) + V[\underset{\sim}{\partial}(60) - \underset{\sim}{\partial}(\bar{a}^i)]$$

It follows from Section IV (b) that

$$V[\underset{\sim}{\hat{\theta}}(60)-\underset{\sim}{\hat{\theta}}(\bar{a}^1)] = (A_1-A_2)V(\underset{\sim}{\hat{\beta}})(A_1-A_2)'$$

where

$$A_1 = I \otimes (1,60,60^2,60^3)$$
$$A_2 = I \otimes (1,\bar{a}^1,(\bar{a}^1)^2,(\bar{a}^1)^3).$$

Similarly $V[^j\underset{\sim}{\theta}(60)]$ can be calculated.

TABLE VIII

Parameters Used in the Calculation of the Histology and Stage Life Tables.

	H_1	H_2	H_3	S_1	S_2	S_3	S_4
N	441	932	2764	2001	1821	837	713
$\bar{a}$	59.2	58.5	58.0	61.9	60.6	63.4	62.3
$\theta_0(\bar{a})$	.056	.086	.105	.038	.087	.142	.361
$\theta_1(\bar{a})$	2.320	1.970	1.870	2.700	1.780	1.360	.963
$\theta_2(\bar{a})$	.076	.084	.054	.028	.101	.061	.012
$\theta_0(60)$	.056	.086	.105	.040	.087	.144	.362
$\theta_1(60)$	2.340	2.000	1.910	2.670	1.780	1.310	.940
$\theta_2(60)$	.071	.076	.045	.034	.101	.070	.017

Life Tables for a=60 have been calculated based on the estimates $\underset{\sim}{\theta}^1(60)$ and $^j\underset{\sim}{\theta}(60)$. These give a direct comparison of the effect of histology and stage on life expectancy and this effect can be measured against the comparable table in the Main Life Tables.

By using estimates $\theta_k^1(a) = \frac{\vartheta_k(a)}{\vartheta_k(60)} \cdot \theta_k^1(60)$ and ${}^j\theta_k(a) = \frac{\vartheta_k(a)}{\vartheta_k(60)} \cdot {}^j\theta_k(60)$ $k = 0,1,2$ it is possible to calculate the survival expectations for reporting ages ranging from 35 to 85 as detailed in the Histology and Stage Life Tables.

The implicit assumption underlying these estimates is that the ratio of the parameters at age a to the parameters values at 60 for the various stages and histologies is the same as $\vartheta_k(a)/\vartheta_k(60)$ $k = 0,1,2$. This is probably reasonable and, if accepted, by-passes a full and expensive age analysis on the seven sub-groups in turn.

The effect on expectation of the two classifications is so marked that it is imperative to examine the joint situation. That is, life tables for age 60 and life expectations from 35 to 85 are required for women in S_j and with histology type H_i $j = 1,2,3,4$ and $i = 1,2,3$. Let ${}^j\underset{\sim}{\theta}^1$ be the parameter vector which is appropriate for S_j and H_i registrants. Then, if

$$ {}^j\theta_k^1(60) = \frac{\theta_k^1(60)}{\vartheta_k(60)} \cdot \frac{{}^j\theta_k(60)}{\vartheta_k(60)} $$

${}^j\underset{\sim}{\theta}^1(60)$ can be constructed from known vectors $\underset{\sim}{\theta}^1(60)$, ${}^j\underset{\sim}{\theta}(60)$ and $\underset{\sim}{\vartheta}(60)$. Stage × Histology Life Tables were constructed on this basis.

By using identical methods to tnose already discussed, survival expectations for the same range of reporting ages were calculated for the various stage × histology classifications.

In order to test the above multiplicative assumption, the data were classified according to stage and histology type, yielding a two way table, see Table IX.

TABLE IX

Frequencies and Residuals for Stage × Histology Life Tables

	S_1	S_2	S_3	S_4
H_1	177 21.4	59 13.9	36 19.6	18 10.8
H_2	272 17.9	252 8.5	97 20.9	49 13.1
H_3	560 18.5	637 14.4	196 9.2	176 50.5

Unfortunately, due to missing information, many cells had small numbers. Nevertheless it was possible to calculate $A({}^{j}\underset{\sim}{\varrho}^{1}(60), V^{-1}[\bar{\underset{\sim}{P}}])$ for all i and j. In general, the residuals were satisfactorily low. The one large value for cell (3,4) results from the estimated $P_k({}^{4}\underset{\sim}{\varrho}^{3})$ being too high. In view of the fact that different data sets were used to calculate the ${}^{j}\underset{\sim}{\varrho}^{1}(60)$ and to test the validity of their construction, the evidence strongly

suggests that the synthesised vectors for S_j and H_i are entirely satisfactory. This conclusion is reinforced if one abandons classifications involving S_4 women. The fact is that, once in S_4, histology plays a more or less insignificant role. For this classification the order of the differences are small and cannot be accurately established by the present methods. Moreover, from a pragmatic viewpoint, it is quite unnecessary to attempt such a resolution. For these reasons, only H_i, S_j classifications for $i,j = 1,2,3$ were used in the Stage × Histology Life Tables. For any histology, the appropriate table for S_4 women is the S_4 table.

B (b) Precision of Estimation of Expectations in B(a)

The method for obtaining $V\{E[X|m,\underset{\sim}{\theta},a]\}$ for stage, histology and stage × histology that was derived in Section V A(b) i.e.

$$V\{E[X|m,\underset{\sim}{\hat{\theta}},a]\} = \left(\frac{\partial}{\partial \underset{\sim}{\theta}} E[X|m,\underset{\sim}{\theta},a]\right)' V(\underset{\sim}{\hat{\theta}}(a))\left(\frac{\partial}{\partial \underset{\sim}{\theta}} E[X|m,\underset{\sim}{\theta},a]\right)$$

can be used here with slight modification.

For $\underset{\sim}{\hat{\theta}}(a) \equiv {}^{j}\underset{\sim}{\theta}^{i}(a)$, approximately

$$\mathrm{Cov}[{}^{j}\theta_k{}^{1}(a),\ {}^{j}\theta_\ell{}^{1}(a)]$$

$$= {}^{j}\theta_k{}^{1}(a)\,{}^{j}\theta_\ell{}^{1}(a)\Bigg\{ \frac{\mathrm{Cov}[\theta_k{}^{1}(60),\theta_\ell{}^{1}(60)]}{\theta_k{}^{1}(60)\theta_\ell{}^{1}(60)} + \frac{\mathrm{Cov}[{}^{j}\theta_k(60),{}^{j}\theta_\ell(60)]}{{}^{j}\theta_k(60)\,{}^{j}\theta_\ell(60)}$$

$$+ \frac{4\ \mathrm{Cov}[\theta_k(60),\theta_\ell(60)]}{\theta_k(60)\theta_\ell(60)} + \frac{\mathrm{Cov}[\theta_k(a),\theta_\ell(a)]}{\theta_k(a)\theta_\ell(a)}$$

$$- \frac{2\ \mathrm{Cov}[\theta_k(60),\theta_\ell(a)]}{\theta_k(60)\theta_\ell(a)} - \frac{2\ \mathrm{Cov}[\theta_k(a),\theta_\ell(60)]}{\theta_k(a)\theta_\ell(60)} \Bigg\}$$

where $\underset{\sim}{\theta}(a) = A\,\underset{\sim}{\hat{\beta}}, \qquad A = I \otimes \underset{\sim}{a}'$

$\underset{\sim}{\theta}(60) = A_{60}\underset{\sim}{\hat{\beta}}$

$\mathrm{Cov}[\underset{\sim}{\theta}(a),\underset{\sim}{\theta}(60)] = A\ V(\underset{\sim}{\hat{\beta}})A_{60}'$.

For $\underset{\sim}{\hat{\theta}}(a) \equiv \underset{\sim}{\theta}^{1}(a)$,

$$\mathrm{Cov}[\theta_k{}^{1}(a),\ \theta_\ell{}^{1}(a)]$$

$$= \theta_k{}^{1}(a)\theta_\ell{}^{1}(a)\Bigg\{\frac{\mathrm{Cov}[\theta_k{}^{1}(60),\theta_\ell{}^{1}(60)]}{\theta_k{}^{1}(60)\theta_\ell{}^{1}(60)} + \frac{\mathrm{Cov}[\theta_k(a),\theta_\ell(a)]}{\theta_k(a)\theta_\ell(a)}$$

$$+ \frac{\mathrm{Cov}[\theta_k(60),\theta_\ell(60)]}{\theta_k(60)\theta_\ell(60)} - \frac{\mathrm{Cov}[\theta_k(a),\theta_\ell(60)]}{\theta_k(a)\theta_\ell(60)} - \frac{\mathrm{Cov}[\theta_k(60),\theta_\ell(a)]}{\theta_k(60)\theta_\ell(a)} \Bigg\}$$

For $\underset{\sim}{\hat{\theta}}(a) \equiv {}^{j}\underset{\sim}{\theta}(a)$, $\underset{\sim}{\theta}^{1}(a)$ is replaced in the above expression by ${}^{j}\underset{\sim}{\theta}(a)$.

C. Tables by Treatment, Histology × Treatment, and Stage × Treatment

Three main types of treatment of Breast Cancer are commonly recognised; radiotherapy, surgery and a combination of both (T_1, T_2, T_3). Using identical methods to those of B(a) life tables were constructed at age 60 for T_1, T_2 and T_3 and the various treatment × histology and treatment × stage groupings. In general the fitting procedure works extremely well, as can be seen from Table X.

TABLE X

Frequencies and Residuals for treatment × stage and treatment × histology life tables

	S_1	S_2	S_3	H_1	H_2	H_3
T_1	980	639	162	172	356	742
	24.7	20.5	32.3	8.3	18.2	18.8
T_2	100	188	309	45	81	176
	23.2	12.0	61.7	19.1	13.8	5.0
T_3	734	801	270	151	361	1074
	18.5	26.9	6.5	11.03	9.3	18.3

For the same reasons as discussed earlier, $T_i \times S_4$ $i = 1,2,3$ was not investigated. Analogously, for any treatment the appropriate table for S_4 women is the S_4 table.

D. Q Tables for Stage, Histology and Stage × Histology

Using estimates of $\underset{\sim}{\theta}^{1}(a), {}^{j}\underset{\sim}{\theta}(a), {}^{j}\underset{\sim}{\theta}^{1}(a)$ already obtained, extensive tables of $Q_x(\underset{\sim}{\theta}(a)) = P_x(\underset{\sim}{\theta}(a))/(1-\sum_{s=0}^{x-1} P_s(\underset{\sim}{\theta}(a))$ were calculated for a=35,(1),85 and for the various stage, histology and stage × histology classifications. These tables of the conditional probability $Q_x(\underset{\sim}{\theta}(a))$ of dying in (x,x+1), given that the disease was first observed at age a and the woman has survived a period of x years, are of actuarial interest. The tables are currently available but are not included in this volume.

E. Graphs of ϕ_c

Graphs of the force of mortality function due to breast cancer, ϕ_c ,were plotted for the stage and histology classifications for observation age 60.Plots were also obtained of $\phi_c(\underset{\sim}{\theta}(a))$ for a=35,60 and 85.

This set of graphs provides a visual impression of the effects of age, stage and histology on the morbidity of breast cancer.

VI The Asymptotic Properties of $\underset{\sim}{\theta}^*$ and $A(\underset{\sim}{\theta}^*, V^{-1}[\overline{\underset{\sim}{P}}])$.

The large sample properties of the statistics $\underset{\sim}{\theta}^*$ and $A(\underset{\sim}{\theta}^*, V^{-1}[\overline{\underset{\sim}{P}}])$ introduced in Section IV will be discussed as a special case of some general results which are of interest in their own right. Let $\overline{\underset{\sim}{X}}' = (\overline{X}_1, \overline{X}_2, \ldots, \overline{X}_k)$ be a second order random vector, where $\overline{X}_i$, the ith component, is calculated from a random sample of size n_i. Further, suppose $E[\overline{\underset{\sim}{X}}] = \underset{\sim}{\mu}(\underset{\sim}{\theta})$, where $\underset{\sim}{\theta}$ is a $q \times 1$ vector of unknown parameters belonging to Ω, and that $\overline{\underset{\sim}{X}}$ tends in law to a normal distribution with mean $\underset{\sim}{\mu}(\underset{\sim}{\theta})$ and non-singular covariance matrix ~~$V(\underset{\sim}{n})$~~ V/n, ~~$\underset{\sim}{n}' = (n_1, n_2, \ldots, n_k)$~~, ~~$n_i = n\alpha_i, \alpha_i > 0$~~ fixed.

We consider the estimator $\underset{\sim}{\theta}^*$ of $\underset{\sim}{\theta}$ defined by

$$A(\underset{\sim}{\theta}^*, S) = \inf_{\underset{\sim}{\theta} \in \Omega} (\overline{\underset{\sim}{X}} - \underset{\sim}{\mu}(\underset{\sim}{\theta}))' S (\overline{\underset{\sim}{X}} - \underset{\sim}{\mu}(\underset{\sim}{\theta})) \qquad \dagger$$

where S is symmetric and non-singular. Thus, $\underset{\sim}{\theta}$ minimises the quadratic form $A(\underset{\sim}{\theta}, S)$ and $A(\underset{\sim}{\theta}^*, S)$ is some measure of the agreement between $\overline{\underset{\sim}{X}}$ and $\underset{\sim}{\mu}(\underset{\sim}{\theta}^*)$.

The following theorem shows that $A(\underset{\sim}{\theta}^*, S)$ is asymptotically distributed as a chi-square variable with $(k-q)$ degrees of freedom for suitable choice of S. As a by-product, the large sample distribution of $\underset{\sim}{\theta}^*$ is also obtained.

† Ω is assumed to satisfy the Implicit Function Theorem ensuring $\underset{\sim}{\theta}_n^*$ exists, is consistent and $\sqrt{n}(\underset{\sim}{\theta}_n^* - \underset{\sim}{\theta})$ tends to a normal distribution.

Theorem

Let $\underset{\sim}{\mu}(\underset{\sim}{\theta})$ have continuous first and second partial derivatives for $\underset{\sim}{\theta} \in \Omega$ and define

(1) $Q = [q_{st}]$, $q_{st} = \partial \mu_s(\underset{\sim}{\theta})/\partial \theta_t$, rank $Q \underset{\theta}{\equiv} q$

(2) $A = Q(Q'SQ)^{-1}Q'S$

(3) $\Sigma = (I-A)V(\underset{\sim}{n})\ (I-A')$.

Then, if Σ^g is a generalised inverse of Σ i.e. $\Sigma\Sigma^g\Sigma = \Sigma$, $A(\underset{\sim}{\theta}^*, \Sigma^g)$ is asymptotically distributed as $\chi^2_{(k-q)}$.

Proof

The proof of the above result will be outlined by means of the notation $\underset{\sim}{\overset{P}{}}$, where for a sequence of random variables U_n, $U_n \underset{\sim}{\overset{P}{}} V$ means that U_n converges in probability to V i.e. for all $\varepsilon>0$, $\lim_{n\to\infty} \Pr\{|U_n-V|>\varepsilon\} = 0$. Final justification of the calculations are omitted since they are available elsewhere, see e.g. Rao 1965 Chapter 6.

Now set

$$\phi_i(\underset{\sim}{\overline{X}}, \underset{\sim}{\theta}) = \tfrac{1}{2} \frac{\partial}{\partial \theta_i} A(\underset{\sim}{\theta}, S)$$

$$= -\underset{\sim}{\overline{X}}' S\, \underset{\sim}{\mu}_i(\underset{\sim}{\theta}) + \underset{\sim}{\mu}'(\underset{\sim}{\theta}) S \underset{\sim}{\mu}_i(\underset{\sim}{\theta})$$

where $\underset{\sim}{\mu}_i(\underset{\sim}{\theta}) = \frac{\partial}{\partial \theta_i} \underset{\sim}{\mu}(\underset{\sim}{\theta})$. Put $\underset{\sim}{\theta} = \underset{\sim}{\theta}^*$, then

$$\phi_i(\underset{\sim}{\overline{X}}, \underset{\sim}{\theta}^*) = 0 \quad i = 1,2,\ldots,q.$$

Write this in vector notation

$$\underset{\sim}{\phi}(\underset{\sim}{\bar{X}}, \underset{\sim}{\theta}^*) = \underset{\sim}{0}$$

and observe that $\underset{\sim}{\phi}(\underset{\sim}{\mu}(\underset{\sim}{\theta}), \underset{\sim}{\theta}) = \underset{\sim}{0}$.

By the Mean Value Theorem it can be shown that

$$\sqrt{n}\,\phi_i(\underset{\sim}{\bar{X}}, \underset{\sim}{\theta}^*) = 0 \overset{P}{\underset{\sim}{}} \phi_i(\underset{\sim}{\mu}(\underset{\sim}{\theta}), \underset{\sim}{\theta}) + \sqrt{n} \sum_{j=1}^{k} \frac{\partial}{\partial \bar{X}_j} \phi_i(\underset{\sim}{\mu}(\underset{\sim}{\theta}), \underset{\sim}{\theta})(\bar{X}_j - \mu_j(\underset{\sim}{\theta}))$$

$$+ \sqrt{n} \sum_{r=1}^{q} \frac{\partial}{\partial \theta_r} \phi_i(\underset{\sim}{\mu}(\underset{\sim}{\theta}), \underset{\sim}{\theta})(\theta_r^* - \theta_r).$$

But $\frac{\partial}{\partial \bar{X}_j} \phi_i(\underset{\sim}{\mu}(\underset{\sim}{\theta}), \underset{\sim}{\theta}) =$ -jth element of $S\,\underset{\sim}{\mu}_i(\underset{\sim}{\theta})$

and

$$\frac{\partial}{\partial \theta_r} \phi_i(\underset{\sim}{\mu}(\underset{\sim}{\theta}), \underset{\sim}{\theta}) = \underset{\sim}{\mu}_r'(\underset{\sim}{\theta}) S\, \underset{\sim}{\mu}_i(\underset{\sim}{\theta})$$

so that

$$\underset{\sim}{0} = \underset{\sim}{\phi}(\underset{\sim}{\bar{X}}, \underset{\sim}{\theta}^*) \overset{P}{\underset{\sim}{}} -\sqrt{n}\,Q'S(\underset{\sim}{\bar{X}} - \underset{\sim}{\mu}(\underset{\sim}{\theta})) + \sqrt{n}\,Q'SQ(\underset{\sim}{\theta}^* - \underset{\sim}{\theta})$$

and $\sqrt{n}(\underset{\sim}{\theta}^* - \underset{\sim}{\theta}) \overset{P}{\underset{\sim}{}} \sqrt{n}(Q'SQ)^{-1}Q'S(\underset{\sim}{\bar{X}} - \underset{\sim}{\mu}(\underset{\sim}{\theta}))$.

Now consider

$$\sqrt{n}\left(\underset{\sim}{\bar{X}} - \underset{\sim}{\mu}(\underset{\sim}{\theta}^*)\right) \overset{P}{\underset{\sim}{}} \sqrt{n}\left(\underset{\sim}{\bar{X}} - \underset{\sim}{\mu}(\underset{\sim}{\theta})\right) - \sqrt{n}Q(\underset{\sim}{\theta}^* - \underset{\sim}{\theta})$$

$$\overset{P}{\underset{\sim}{}} \sqrt{n}[I - Q(Q'SQ)^{-1}Q'S](\underset{\sim}{\bar{X}} - \underset{\sim}{\mu}(\underset{\sim}{\theta}))$$

$$= \sqrt{n}[I - A](\underset{\sim}{\bar{X}} - \underset{\sim}{\mu}(\underset{\sim}{\theta})),$$

then $\underset{\sim}{\bar{X}} - \underset{\sim}{\mu}(\underset{\sim}{\theta}^*)$ is asymptotically distributed as $N(\underset{\sim}{0}, \Sigma)$,

$\Sigma = (I-A)V(\underset{\sim}{n})(I-A')$. We note the following facts;

(a) rank $\Sigma = \rho(\Sigma)$ = trace $(I-A) = k-q$ since $\rho(V(\underset{\sim}{n})) = k$;

(b) $A(\underset{\sim}{\theta}^{\#},\Sigma^{g})$ is asymptotically distributed as $\chi^2_{(k-q)}$ since, from known results on quadratic forms (Searle 1971, page 69), $\underset{\sim}{Z}'B\underset{\sim}{Z}$ is distributed as $\chi^2_{(d)}$ if

$$VBVBV = VBV, \text{ trace } BV=d,$$

where $\underset{\sim}{Z}$ is distributed as $N(\underset{\sim}{0},V)$. Choosing $B = V^{g}$, $VV^{g}VV^{g}V = VV^{g}V$ and trace $V^{g}V = \rho(V) = \rho(V^{g})$, the result follows from (a).

Lemma 1

$[\Sigma+Q(Q'S^{-1}Q)^{-1}Q']^{-1}$ is a generalised inverse of Σ.

Proof

Notice that $\Sigma SQ = 0$ since $(I-A')SQ = 0$ and also $\rho(Q'S^{\frac{1}{2}}) = q$ and hence

$$\rho\begin{bmatrix} \Sigma^{\frac{1}{2}}S^{\frac{1}{2}} \\ Q' \quad S^{\frac{1}{2}} \end{bmatrix} = k.$$

Also, putting $E^{\frac{1}{2}} = (Q'S^{-1}Q)^{-\frac{1}{2}}$

$$\rho\begin{bmatrix} I & 0 \\ 0 & E^{\frac{1}{2}} \end{bmatrix}\begin{bmatrix} \Sigma^{\frac{1}{2}}S^{\frac{1}{2}} \\ Q' \quad S^{\frac{1}{2}} \end{bmatrix}\begin{bmatrix} I & 0 \\ 0 & S^{-1} \end{bmatrix}S^{-\frac{1}{2}} = \rho\begin{bmatrix} \Sigma^{\frac{1}{2}} \\ E^{\frac{1}{2}} \quad Q' \end{bmatrix} = k$$

since the rank of a matrix is not altered by multiplying it by a non-singular matrix. Hence

$$\rho\left[(\Sigma^{\frac{1}{2}'}, Q E^{\frac{1}{2}'}) \begin{pmatrix} \Sigma^{\frac{1}{2}} \\ E^{\frac{1}{2}} Q' \end{pmatrix} \right] = \rho[\Sigma + Q(Q'S^{-1}Q)^{-1}Q'] = k$$

and by Rao (loc.cit. page 30) $[\Sigma+Q(Q'S^{-1}Q)^{-1}Q']^{-1}$ is a generalised inverse of Σ.

Corollary 1

$$A(\underset{\sim}{\theta}^{*}, V^{-1}(\underset{\sim}{x})) \overset{P}{\sim} \chi^2_{(k-q)}.$$

Proof

Put $S = V^{-1}(\underset{\sim}{n})$ in the theorem, then

$$\Sigma = (I-A)V(\underset{\sim}{x})(I-A') = V(\underset{\sim}{n}) - AV(\underset{\sim}{x}) - V(\underset{\sim}{x})A' + AV(\underset{\sim}{x})A'$$

$$= V(\underset{\sim}{x}) - Q(Q'V(\underset{\sim}{x})Q)^{-1}Q'$$

and hence

$$\Sigma + Q(Q'V(\underset{\sim}{x})Q)^{-1}Q' = V(\underset{\sim}{x})$$

and

$$(\underset{\sim}{\bar{X}} - \underset{\sim}{\mu}(\underset{\sim}{\theta}^{*}))' V^{-1}(\underset{\sim}{x})(\underset{\sim}{\bar{X}} - \underset{\sim}{\mu}(\underset{\sim}{\theta}^{*})) \overset{P}{\sim} \chi^2_{(k-q)}.$$

Corollary 2

In the notation of previous sections,

$$(\underset{\sim}{\bar{P}} - \underset{\sim}{P}(\underset{\sim}{\theta}^{*}))' V^{-1}(\underset{\sim}{\bar{P}})(\underset{\sim}{\bar{P}} - \underset{\sim}{P}(\underset{\sim}{\theta}^{*})) \overset{P}{\sim} \chi^2_{(k-q)}.$$

Proof

This is a direct application of the Theorem, Lemma 1 and Corollary 1.

Corollary 3

For $S = V^{-1}(\Pi)$, $\underset{\sim}{\theta}^* \overset{P}{\sim} N(\underset{\sim}{\theta}, (Q'V^{-1}(\Pi)Q)^{-1})$

Proof

In the proof of the Theorem it was shown that

$$\sqrt{n}(\underset{\sim}{\theta}^* - \underset{\sim}{\theta}) \overset{P}{\sim} \sqrt{n}(Q'SQ)^{-1}Q'S(\overline{\underset{\sim}{X}} - \underset{\sim}{\mu}(\underset{\sim}{\theta}))$$

and the result follows by putting $S = V^{-1}(\Pi)$.

REFERENCES

Cutler, S.J. and Ederer, F. (1958)

Maximum utilization of the life table method in analyzing survival.

J. Chron. Dis., December, p699.

Nelder, J.A. and Mead, R. (1965)

A simplex method for function minimization.

The Computer Journal, 7, p308.

Rao, C.R. (1965)

Linear Statistical Inference and its Applications.

(Wiley , New York)

Ross, W.L. (1969)

The magnitude of the breast cancer problem in the U.S.A.

Cancer, 24, p1106.

Rankin, R.W. (1971)

The Central Cancer Registry,Melbourne 1940-1970.

Med. J. of Aust., 1, p750.

Searle, S.R. (1971)

Linear Models. (Wiley , New York)

Commonwealth Bureau of Census and Statistics (1972)

Personal communication.

APPENDIX A

MAIN LIFE TABLES

The first value of A in each of the following tables is the age of first observation of breast cancer.

A	P	SE(P)	Q	SE(Q)	E(N)	E(BC)	SE(E(BC))
35	.09731	.00892	.09731	.00892	41.48	14.62	.811
36	.10952	.00638	.12132	.00697	40.56	15.14	.889
37	.08815	.00375	.11113	.00546	39.62	16.17	1.025
38	.07176	.00232	.10178	.00441	38.69	17.14	1.155
39	.05904	.00175	.09322	.00386	37.75	18.03	1.270
40	.04904	.00167	.08540	.00373	36.81	18.83	1.367
41	.04110	.00172	.07826	.00389	35.88	19.54	1.442
42	.03473	.00177	.07175	.00418	34.95	20.16	1.495
43	.02958	.00177	.06582	.00450	34.03	20.68	1.526
44	.02537	.00174	.06045	.00479	33.11	21.11	1.536
45	.02192	.00169	.05558	.00503	32.19	21.44	1.528
46	.01906	.00162	.05117	.00522	31.28	21.67	1.502
47	.01668	.00153	.04721	.00534	30.38	21.81	1.463
48	.01470	.00144	.04365	.00540	29.49	21.87	1.412
49	.01303	.00135	.04046	.00542	28.60	21.84	1.351
50	.01163	.00126	.03763	.00540	27.72	21.74	1.283
51	.01045	.00117	.03513	.00534	26.84	21.57	1.210
52	.00945	.00109	.03294	.00524	25.98	21.34	1.133
53	.00861	.00100	.03103	.00512	25.12	21.05	1.055
54	.00791	.00092	.02941	.00498	24.28	20.71	.977
55	.00732	.00084	.02804	.00483	23.44	20.32	.900
56	.00683	.00076	.02693	.00466	22.61	19.89	.825
57	.00643	.00069	.02606	.00448	21.80	19.43	.752
58	.00611	.00061	.02543	.00429	20.99	18.94	.683
59	.00586	.00054	.02502	.00411	20.20	18.42	.617
60	.00567	.00048	.02484	.00392	19.42	17.88	.556
61	.00554	.00042	.02489	.00372	18.65	17.32	.498
62	.00546	.00036	.02516	.00354	17.89	16.75	.445
63	.00543	.00031	.02566	.00335	17.15	16.17	.396
64	.00544	.00028	.02639	.00317	16.42	15.58	.351
65	.00550	.00026	.02736	.00299	15.71	14.99	.311
66	.00558	.00027	.02858	.00282	15.01	14.40	.273
67	.00570	.00030	.03006	.00265	14.33	13.81	.240
68	.00586	.00035	.03181	.00249	13.66	13.22	.210
69	.00605	.00040	.03384	.00234	13.01	12.64	.183
70	.00623	.00047	.03618	.00219	12.38	12.06	.159
71	.00644	.00053	.03884	.00205	11.77	11.50	.138
72	.00667	.00060	.04184	.00191	11.17	10.94	.119
73	.00691	.00067	.04521	.00178	10.59	10.40	.102
74	.00715	.00074	.04897	.00166	10.02	9.87	.088
75	.00738	.00081	.05316	.00155	9.48	9.35	.075
76	.00760	.00087	.05781	.00144	8.96	8.85	.064
77	.00779	.00093	.06295	.00134	8.45	8.36	.054
78	.00796	.00098	.06862	.00124	7.96	7.88	.046
79	.00809	.00103	.07487	.00115	7.49	7.43	.039
80	.00817	.00107	.08174	.00106	7.04	6.99	.033
81	.00819	.00109	.08928	.00098	6.61	6.57	.027
82	.00815	.00111	.09753	.00090	6.20	6.16	.023
83	.00804	.00111	.10656	.00083	5.80	5.77	.019
84	.00785	.00110	.11643	.00076	5.43	5.40	.016
85	.00757	.00107	.12718	.00070	5.07	5.05	.013
86	.00722	.00104	.13890	.00064	4.73	4.71	.011
87	.00679	.00098	.15162	.00059	4.41	4.39	.009
88	.00628	.00092	.16544	.00054	4.10	4.09	.007
89	.00572	.00084	.18040	.00049	3.81	3.80	.006
90	.00511	.00076	.19658	.00044	3.54	3.53	.005

A	P	SE(P)	Q	SE(Q)	E(N)	E(BC)	SE(E(BC))
36	.09006	.00782	.09006	.00782	40.56	14.43	.725
37	.11179	.00582	.12285	.00631	39.62	14.81	.789
38	.08976	.00339	.11245	.00493	38.69	15.82	.912
39	.07291	.00210	.10293	.00399	37.75	16.77	1.028
40	.05987	.00161	.09421	.00351	36.81	17.64	1.130
41	.04965	.00155	.08625	.00343	35.88	18.43	1.216
42	.04155	.00160	.07900	.00361	34.95	19.12	1.282
43	.03507	.00164	.07239	.00389	34.03	19.72	1.327
44	.02983	.00164	.06639	.00419	33.11	20.22	1.353
45	.02557	.00161	.06096	.00446	32.19	20.62	1.360
46	.02208	.00155	.05604	.00468	31.28	20.93	1.351
47	.01919	.00148	.05160	.00484	30.38	21.14	1.326
48	.01679	.00140	.04761	.00495	29.49	21.27	1.289
49	.01479	.00132	.04404	.00500	28.60	21.31	1.241
50	.01312	.00123	.04086	.00501	27.72	21.27	1.185
51	.01172	.00115	.03804	.00498	26.84	21.15	1.123
52	.01053	.00106	.03556	.00492	25.98	20.97	1.057
53	.00954	.00098	.03340	.00482	25.12	20.72	.988
54	.00871	.00090	.03154	.00471	24.28	20.42	.918
55	.00801	.00082	.02996	.00457	23.44	20.07	.848
56	.00743	.00075	.02865	.00442	22.61	19.68	.779
57	.00696	.00067	.02761	.00426	21.80	19.24	.712
58	.00657	.00060	.02682	.00409	20.99	18.77	.648
59	.00627	.00054	.02627	.00392	20.20	18.28	.587
60	.00603	.00047	.02597	.00374	19.42	17.76	.529
61	.00586	.00041	.02590	.00357	18.65	17.22	.475
62	.00574	.00035	.02607	.00339	17.89	16.66	.425
63	.00568	.00031	.02647	.00321	17.15	16.09	.378
64	.00567	.00027	.02712	.00304	16.42	15.52	.336
65	.00569	.00025	.02802	.00287	15.71	14.94	.297
66	.00576	.00025	.02917	.00271	15.01	14.35	.262
67	.00587	.00027	.03059	.00255	14.33	13.77	.230
68	.00600	.00031	.03228	.00239	13.66	13.19	.201
69	.00615	.00036	.03426	.00224	13.01	12.61	.175
70	.00635	.00042	.03656	.00210	12.38	12.04	.152
71	.00656	.00048	.03917	.00197	11.77	11.48	.132
72	.00678	.00054	.04214	.00184	11.17	10.93	.114
73	.00701	.00061	.04547	.00171	10.59	10.38	.098
74	.00724	.00067	.04921	.00160	10.02	9.86	.084
75	.00745	.00073	.05337	.00149	9.48	9.34	.072
76	.00768	.00079	.05800	.00138	8.96	8.84	.061
77	.00787	.00085	.06312	.00128	8.45	8.35	.052
78	.00803	.00090	.06877	.00119	7.96	7.88	.044
79	.00816	.00094	.07500	.00110	7.49	7.42	.037
80	.00824	.00098	.08186	.00102	7.04	6.99	.031
81	.00826	.00100	.08938	.00094	6.61	6.56	.026
82	.00821	.00102	.09763	.00086	6.20	6.16	.022
83	.00810	.00102	.10665	.00080	5.80	5.77	.018
84	.00790	.00101	.11650	.00073	5.43	5.40	.015
85	.00762	.00099	.12725	.00067	5.07	5.05	.013
86	.00727	.00095	.13895	.00061	4.73	4.71	.010
87	.00683	.00090	.15168	.00056	4.41	4.39	.009
88	.00632	.00084	.16548	.00051	4.10	4.09	.007
89	.00575	.00077	.18044	.00047	3.81	3.80	.006
90	.00514	.00070	.19661	.00042	3.54	3.53	.005

A	P	SE(P)	Q	SE(Q)	E(N)	E(BC)	SE(E(BC))
37	.08391	.00735	.08391	.00735	39.62	14.26	.704
38	.11369	.00573	.12410	.00617	38.69	14.52	.761
39	.09109	.00332	.11352	.00481	37.75	15.51	.881
40	.07386	.00205	.10384	.00390	36.81	16.43	.993
41	.06055	.00160	.09499	.00346	35.88	17.28	1.093
42	.05015	.00156	.08693	.00342	34.95	18.05	1.174
43	.04192	.00162	.07958	.00363	34.03	18.72	1.237
44	.03534	.00165	.07290	.00393	33.11	19.30	1.279
45	.03004	.00165	.06684	.00423	32.19	19.77	1.302
46	.02574	.00161	.06136	.00450	31.28	20.16	1.307
47	.02221	.00155	.05641	.00472	30.38	20.44	1.295
48	.01930	.00148	.05195	.00487	29.49	20.63	1.269
49	.01689	.00139	.04796	.00497	28.60	20.74	1.231
50	.01489	.00131	.04439	.00502	27.72	20.76	1.183
51	.01321	.00122	.04122	.00502	26.84	20.70	1.128
52	.01181	.00113	.03843	.00499	25.98	20.57	1.066
53	.01063	.00104	.03598	.00491	25.12	20.37	1.001
54	.00964	.00096	.03386	.00482	24.28	20.11	.933
55	.00882	.00088	.03205	.00469	23.44	19.80	.865
56	.00813	.00080	.03054	.00455	22.61	19.44	.797
57	.00757	.00072	.02930	.00440	21.80	19.04	.731
58	.00710	.00065	.02834	.00423	20.99	18.60	.666
59	.00673	.00058	.02764	.00406	20.20	18.13	.605
60	.00644	.00051	.02719	.00388	19.42	17.63	.546
61	.00622	.00044	.02700	.00370	18.65	17.10	.491
62	.00606	.00038	.02705	.00352	17.89	16.57	.440
63	.00597	.00033	.02736	.00334	17.15	16.01	.392
64	.00592	.00028	.02792	.00316	16.42	15.45	.348
65	.00592	.00026	.02873	.00299	15.71	14.88	.308
66	.00597	.00025	.02981	.00282	15.01	14.30	.272
67	.00605	.00026	.03116	.00266	14.33	13.73	.239
68	.00617	.00030	.03279	.00250	13.66	13.15	.209
69	.00632	.00035	.03472	.00234	13.01	12.58	.182
70	.00650	.00040	.03697	.00219	12.38	12.02	.159
71	.00669	.00047	.03954	.00205	11.77	11.46	.137
72	.00690	.00053	.04247	.00192	11.17	10.91	.119
73	.00712	.00060	.04577	.00179	10.59	10.37	.102
74	.00735	.00066	.04947	.00167	10.02	9.84	.088
75	.00757	.00072	.05361	.00155	9.48	9.33	.075
76	.00778	.00078	.05820	.00144	8.96	8.83	.064
77	.00797	.00084	.06330	.00134	8.45	8.34	.054
78	.00813	.00089	.06894	.00124	7.96	7.87	.046
79	.00825	.00093	.07515	.00115	7.49	7.42	.039
80	.00832	.00097	.08199	.00106	7.04	6.98	.033
81	.00834	.00100	.08950	.00098	6.61	6.56	.027
82	.00829	.00101	.09773	.00090	6.20	6.16	.023
83	.00817	.00102	.10674	.00083	5.80	5.77	.019
84	.00797	.00101	.11658	.00076	5.43	5.40	.016
85	.00769	.00099	.12732	.00070	5.07	5.05	.013
86	.00733	.00095	.13901	.00064	4.73	4.71	.011
87	.00689	.00090	.15173	.00058	4.41	4.39	.009
88	.00637	.00084	.16553	.00053	4.10	4.09	.007
89	.00580	.00077	.18048	.00048	3.81	3.80	.006
90	.00518	.00070	.19665	.00044	3.54	3.53	.005

A	P	SE(P)	Q	SE(Q)	E(N)	E(BC)	SE(E(BC))
38	.07888	.00730	.07888	.00730	38.69	14.08	.709
39	.11519	.00592	.12505	.00635	37.75	14.25	.764
40	.09214	.00341	.11433	.00494	36.81	15.22	.884
41	.07461	.00211	.10452	.00400	35.88	16.12	.997
42	.06109	.00166	.09557	.00357	34.95	16.95	1.096
43	.05053	.00163	.08741	.00356	34.03	17.69	1.178
44	.04220	.00170	.07999	.00379	33.11	18.34	1.239
45	.03556	.00173	.07326	.00411	32.19	18.89	1.280
46	.03021	.00172	.06716	.00443	31.28	19.35	1.301
47	.02587	.00168	.06165	.00471	30.38	19.70	1.304
48	.02232	.00161	.05669	.00493	29.49	19.97	1.290
49	.01940	.00153	.05223	.00509	28.60	20.14	1.262
50	.01698	.00144	.04825	.00519	27.72	20.22	1.221
51	.01498	.00135	.04470	.00523	26.84	20.22	1.171
52	.01330	.00125	.04156	.00523	25.98	20.14	1.114
53	.01190	.00116	.03880	.00518	25.12	20.00	1.051
54	.01073	.00107	.03640	.00510	24.28	19.78	.985
55	.00975	.00098	.03433	.00499	23.44	19.51	.916
56	.00894	.00089	.03259	.00486	22.61	19.19	.847
57	.00827	.00081	.03115	.00471	21.80	18.82	.779
58	.00771	.00073	.03000	.00454	20.99	18.41	.712
59	.00727	.00065	.02913	.00437	20.20	17.96	.648
60	.00691	.00057	.02853	.00418	19.42	17.48	.586
61	.00663	.00050	.02820	.00400	18.65	16.98	.528
62	.00643	.00043	.02813	.00381	17.89	16.46	.474
63	.00629	.00037	.02833	.00362	17.15	15.92	.423
64	.00621	.00032	.02878	.00343	16.42	15.37	.376
65	.00619	.00028	.02951	.00324	15.71	14.81	.333
66	.00621	.00026	.03050	.00306	15.01	14.25	.294
67	.00627	.00027	.03178	.00288	14.33	13.68	.259
68	.00637	.00030	.03335	.00271	13.66	13.11	.226
69	.00650	.00035	.03522	.00254	13.01	12.55	.198
70	.00666	.00041	.03741	.00239	12.38	11.99	.172
71	.00685	.00047	.03994	.00223	11.77	11.44	.149
72	.00705	.00054	.04282	.00209	11.17	10.89	.129
73	.00726	.00061	.04608	.00195	10.59	10.35	.111
74	.00748	.00067	.04975	.00181	10.02	9.83	.095
75	.00769	.00074	.05386	.00169	9.48	9.32	.081
76	.00790	.00081	.05843	.00157	8.96	8.82	.069
77	.00808	.00086	.06350	.00145	8.45	8.34	.059
78	.00824	.00092	.06912	.00135	7.96	7.87	.050
79	.00835	.00097	.07531	.00125	7.49	7.42	.042
80	.00843	.00100	.08213	.00115	7.04	6.98	.035
81	.00844	.00103	.08962	.00106	6.61	6.56	.030
82	.00839	.00105	.09784	.00098	6.20	6.15	.025
83	.00826	.00105	.10684	.00090	5.80	5.77	.021
84	.00806	.00105	.11667	.00082	5.43	5.40	.017
85	.00777	.00102	.12740	.00076	5.07	5.05	.014
86	.00740	.00099	.13908	.00069	4.73	4.71	.012
87	.00696	.00094	.15179	.00063	4.41	4.39	.010
88	.00644	.00088	.16558	.00058	4.10	4.09	.008
89	.00586	.00080	.18053	.00052	3.81	3.80	.006
90	.00523	.00072	.19669	.00047	3.54	3.53	.005

A	P	SE(P)	Q	SE(Q)	E(N)	E(BC)	SE(E(BC))
39	.07499	.00743	.07499	.00743	37.75	13.92	.715
40	.11628	.00619	.12571	.00661	36.81	14.01	.767
41	.09290	.00355	.11487	.00513	35.88	14.96	.887
42	.07514	.00219	.10496	.00416	34.95	15.84	1.001
43	.06146	.00173	.09593	.00372	34.03	16.64	1.100
44	.05081	.00171	.08771	.00371	33.11	17.36	1.180
45	.04241	.00178	.08025	.00396	32.19	17.98	1.240
46	.03571	.00181	.07348	.00430	31.28	18.50	1.280
47	.03033	.00180	.06736	.00463	30.38	18.93	1.299
48	.02597	.00175	.06185	.00492	29.49	19.27	1.299
49	.02241	.00168	.05689	.00514	28.60	19.50	1.283
50	.01949	.00159	.05244	.00530	27.72	19.65	1.253
51	.01707	.00150	.04848	.00540	26.84	19.71	1.211
52	.01507	.00140	.04497	.00544	25.98	19.69	1.159
53	.01340	.00130	.04187	.00543	25.12	19.60	1.100
54	.01201	.00120	.03916	.00538	24.28	19.43	1.035
55	.01085	.00110	.03682	.00529	23.44	19.20	.968
56	.00988	.00100	.03482	.00517	22.61	18.92	.898
57	.00908	.00091	.03316	.00503	21.80	18.58	.829
58	.00842	.00082	.03180	.00487	20.99	18.20	.760
59	.00788	.00073	.03075	.00469	20.20	17.78	.693
60	.00745	.00065	.02998	.00451	19.42	17.33	.629
61	.00711	.00057	.02950	.00431	18.65	16.85	.568
62	.00685	.00050	.02930	.00411	17.89	16.35	.510
63	.00667	.00043	.02938	.00392	17.15	15.83	.456
64	.00655	.00037	.02972	.00372	16.42	15.29	.406
65	.00649	.00032	.03035	.00352	15.71	14.74	.361
66	.00648	.00029	.03126	.00332	15.01	14.19	.319
67	.00652	.00029	.03246	.00313	14.33	13.63	.280
68	.00660	.00031	.03395	.00295	13.66	13.07	.246
69	.00671	.00035	.03576	.00277	13.01	12.51	.215
70	.00686	.00041	.03789	.00260	12.38	11.96	.187
71	.00703	.00047	.04037	.00243	11.77	11.41	.162
72	.00722	.00054	.04321	.00227	11.17	10.87	.140
73	.00742	.00062	.04643	.00212	10.59	10.34	.120
74	.00763	.00069	.05006	.00198	10.02	9.82	.103
75	.00784	.00076	.05413	.00184	9.48	9.31	.088
76	.00804	.00083	.05867	.00171	8.96	8.81	.075
77	.00822	.00089	.06372	.00158	8.45	8.33	.064
78	.00837	.00095	.06931	.00147	7.96	7.86	.054
79	.00848	.00100	.07548	.00136	7.49	7.41	.046
80	.00855	.00104	.08228	.00125	7.04	6.97	.038
81	.00856	.00107	.08976	.00116	6.61	6.56	.032
82	.00850	.00109	.09796	.00106	6.20	6.15	.027
83	.00837	.00109	.10694	.00098	5.80	5.77	.022
84	.00816	.00108	.11676	.00090	5.43	5.40	.019
85	.00787	.00106	.12748	.00082	5.07	5.04	.015
86	.00750	.00103	.13916	.00075	4.73	4.71	.013
87	.00704	.00098	.15185	.00069	4.41	4.39	.010
88	.00652	.00091	.16564	.00063	4.10	4.09	.009
89	.00593	.00084	.18058	.00057	3.81	3.80	.007
90	.00529	.00075	.19673	.00052	3.54	3.53	.006

A	P	SE(P)	Q	SE(Q)	E(N)	E(BC)	SE(E(BC))
40	.07225	.00754	.07225	.00754	36.81	13.72	.707
41	.11696	.00637	.12606	.00679	35.88	13.80	.755
42	.09335	.00365	.11514	.00527	34.95	14.73	.873
43	.07545	.00225	.10517	.00427	34.03	15.58	.984
44	.06169	.00178	.09608	.00381	33.11	16.35	1.081
45	.05097	.00176	.08783	.00381	32.19	17.04	1.159
46	.04253	.00182	.08034	.00406	31.28	17.63	1.217
47	.03581	.00186	.07357	.00440	30.38	18.13	1.254
48	.03042	.00184	.06745	.00474	29.49	18.53	1.271
49	.02605	.00179	.06194	.00503	28.60	18.84	1.269
50	.02249	.00172	.05701	.00525	27.72	19.05	1.251
51	.01957	.00163	.05259	.00541	26.84	19.18	1.219
52	.01715	.00153	.04867	.00550	25.98	19.21	1.176
53	.01516	.00142	.04520	.00554	25.12	19.17	1.123
54	.01350	.00132	.04216	.00552	24.28	19.05	1.064
55	.01212	.00121	.03952	.00547	23.44	18.87	.999
56	.01097	.00111	.03725	.00537	22.61	18.63	.932
57	.01002	.00101	.03534	.00525	21.80	18.33	.863
58	.00924	.00091	.03376	.00510	20.99	17.98	.794
59	.00859	.00082	.03251	.00493	20.20	17.59	.727
60	.00807	.00073	.03156	.00475	19.42	17.17	.661
61	.00766	.00064	.03092	.00456	18.65	16.71	.598
62	.00734	.00056	.03057	.00436	17.89	16.23	.539
63	.00710	.00049	.03052	.00415	17.15	15.72	.483
64	.00693	.00042	.03075	.00395	16.42	15.20	.431
65	.00684	.00036	.03127	.00374	15.71	14.67	.383
66	.00679	.00032	.03208	.00354	15.01	14.13	.339
67	.00680	.00030	.03319	.00334	14.33	13.58	.298
68	.00686	.00031	.03461	.00315	13.66	13.03	.262
69	.00695	.00035	.03635	.00296	13.01	12.48	.229
70	.00708	.00040	.03842	.00278	12.38	11.93	.199
71	.00724	.00047	.04084	.00260	11.77	11.38	.173
72	.00742	.00054	.04363	.00243	11.17	10.85	.149
73	.00761	.00061	.04680	.00227	10.59	10.32	.129
74	.00781	.00069	.05040	.00211	10.02	9.80	.110
75	.00801	.00076	.05443	.00197	9.48	9.30	.094
76	.00820	.00083	.05894	.00183	8.96	8.80	.081
77	.00838	.00090	.06396	.00170	8.45	8.32	.068
78	.00852	.00096	.06952	.00157	7.96	7.86	.058
79	.00863	.00101	.07567	.00145	7.49	7.40	.049
80	.00869	.00105	.08245	.00134	7.04	6.97	.041
81	.00870	.00108	.08990	.00124	6.61	6.55	.034
82	.00864	.00110	.09809	.00114	6.20	6.15	.029
83	.00850	.00111	.10706	.00105	5.80	5.76	.024
84	.00829	.00110	.11686	.00096	5.43	5.39	.020
85	.00799	.00108	.12757	.00088	5.07	5.04	.016
86	.00761	.00105	.13924	.00081	4.73	4.71	.014
87	.00715	.00099	.15192	.00073	4.41	4.39	.011
88	.00661	.00093	.16570	.00067	4.10	4.09	.009
89	.00601	.00085	.18063	.00061	3.81	3.80	.007
90	.00537	.00077	.19678	.00055	3.54	3.53	.006

A	P	SE(P)	Q	SE(Q)	E(N)	E(BC)	SE(E(BC)) 50
41	.07068	.00751	.07068	.00751	35.88	13.57	.680
42	.11721	.00641	.12612	.00682	34.95	13.62	.725
43	.09351	.00366	.11514	.00528	34.03	14.51	.837
44	.07555	.00226	.10514	.00428	33.11	15.34	.943
45	.06175	.00178	.09603	.00382	32.19	16.09	1.034
46	.05102	.00175	.08777	.00380	31.28	16.74	1.108
47	.04257	.00182	.08028	.00404	30.38	17.31	1.161
48	.03585	.00185	.07352	.00438	29.49	17.78	1.195
49	.03046	.00184	.06742	.00471	28.60	18.15	1.209
50	.02610	.00179	.06195	.00499	27.72	18.42	1.205
51	.02255	.00171	.05705	.00521	26.84	18.61	1.186
52	.01964	.00162	.05268	.00537	25.98	18.71	1.154
53	.01724	.00152	.04882	.00546	25.12	18.72	1.111
54	.01525	.00141	.04541	.00549	24.28	18.66	1.059
55	.01361	.00130	.04245	.00547	23.44	18.52	1.000
56	.01224	.00120	.03988	.00541	22.61	18.32	.938
57	.01111	.00109	.03770	.00531	21.80	18.06	.872
58	.01018	.00099	.03589	.00518	20.99	17.75	.806
59	.00941	.00090	.03441	.00503	20.20	17.39	.740
60	.00879	.00080	.03328	.00486	19.42	16.99	.675
61	.00829	.00071	.03246	.00468	18.65	16.56	.613
62	.00789	.00062	.03195	.00448	17.89	16.10	.553
63	.00759	.00054	.03175	.00428	17.15	15.61	.497
64	.00737	.00046	.03185	.00408	16.42	15.11	.444
65	.00723	.00040	.03226	.00387	15.71	14.59	.395
66	.00715	.00035	.03297	.00367	15.01	14.06	.350
67	.00713	.00032	.03398	.00347	14.33	13.52	.309
68	.00716	.00032	.03532	.00327	13.66	12.98	.271
69	.00723	.00034	.03698	.00308	13.01	12.44	.237
70	.00734	.00039	.03899	.00289	12.38	11.89	.207
71	.00748	.00045	.04135	.00271	11.77	11.36	.180
72	.00764	.00052	.04408	.00253	11.17	10.82	.155
73	.00783	.00059	.04721	.00237	10.59	10.30	.134
74	.00802	.00067	.05076	.00221	10.02	9.79	.115
75	.00821	.00074	.05475	.00205	9.48	9.28	.098
76	.00839	.00081	.05923	.00191	8.96	8.79	.084
77	.00856	.00088	.06421	.00177	8.45	8.31	.071
78	.00870	.00094	.06975	.00164	7.96	7.85	.060
79	.00881	.00099	.07587	.00152	7.49	7.40	.051
80	.00886	.00104	.08263	.00140	7.04	6.96	.043
81	.00886	.00107	.09006	.00129	6.61	6.55	.036
82	.00879	.00109	.09823	.00119	6.20	6.15	.030
83	.00865	.00110	.10718	.00109	5.80	5.76	.025
84	.00843	.00109	.11698	.00100	5.43	5.39	.021
85	.00813	.00107	.12767	.00092	5.07	5.04	.017
86	.00774	.00104	.13932	.00084	4.73	4.71	.014
87	.00726	.00099	.15200	.00077	4.41	4.39	.012
88	.00672	.00092	.16577	.00070	4.10	4.09	.009
89	.00611	.00085	.18069	.00063	3.81	3.80	.008
90	.00545	.00076	.19683	.00057	3.54	3.53	.006

A	P	SE(P)	Q	SE(Q)	E(N)	E(BC)	SE(E(BC))
42	.07029	.00730	.07029	.00730	34.95	13.42	.637
43	.11704	.00625	.12588	.00665	34.03	13.46	.678
44	.09337	.00358	.11489	.00516	33.11	14.32	.781
45	.07544	.00220	.10487	.00417	32.19	15.12	.878
46	.06167	.00173	.09578	.00371	31.28	15.84	.962
47	.05096	.00170	.08753	.00368	30.38	16.46	1.029
48	.04254	.00176	.08007	.00391	29.49	16.99	1.077
49	.03584	.00180	.07334	.00423	28.60	17.43	1.106
50	.03047	.00178	.06729	.00455	27.72	17.77	1.118
51	.02613	.00173	.06187	.00482	26.84	18.02	1.112
52	.02260	.00166	.05703	.00503	25.98	18.18	1.093
53	.01970	.00156	.05273	.00517	25.12	18.25	1.061
54	.01732	.00147	.04893	.00525	24.28	18.23	1.019
55	.01535	.00136	.04561	.00528	23.44	18.15	.969
56	.01373	.00126	.04273	.00526	22.61	17.99	.913
57	.01238	.00115	.04026	.00519	21.80	17.77	.854
58	.01127	.00105	.03818	.00510	20.99	17.50	.793
59	.01036	.00095	.03648	.00497	20.20	17.17	.731
60	.00961	.00085	.03513	.00482	19.42	16.80	.669
61	.00901	.00076	.03412	.00465	18.65	16.40	.609
62	.00853	.00067	.03344	.00447	17.89	15.96	.552
63	.00815	.00058	.03308	.00429	17.15	15.49	.497
64	.00787	.00050	.03305	.00409	16.42	15.01	.445
65	.00768	.00043	.03333	.00389	15.71	14.50	.397
66	.00756	.00037	.03393	.00369	15.01	13.98	.352
67	.00750	.00033	.03484	.00350	14.33	13.46	.311
68	.00749	.00032	.03609	.00330	13.66	12.93	.274
69	.00754	.00033	.03767	.00311	13.01	12.39	.240
70	.00763	.00037	.03960	.00292	12.38	11.86	.209
71	.00775	.00042	.04189	.00274	11.77	11.33	.182
72	.00790	.00049	.04457	.00257	11.17	10.80	.157
73	.00807	.00056	.04764	.00240	10.59	10.28	.136
74	.00825	.00063	.05115	.00224	10.02	9.77	.117
75	.00843	.00070	.05510	.00208	9.48	9.27	.100
76	.00861	.00077	.05954	.00194	8.96	8.78	.085
77	.00877	.00084	.06449	.00180	8.45	8.30	.072
78	.00890	.00090	.06999	.00167	7.96	7.84	.061
79	.00900	.00095	.07609	.00154	7.49	7.39	.052
80	.00905	.00099	.08282	.00142	7.04	6.96	.043
81	.00905	.00103	.09024	.00131	6.61	6.54	.036
82	.00897	.00105	.09838	.00121	6.20	6.14	.030
83	.00883	.00106	.10732	.00111	5.80	5.76	.025
84	.00860	.00106	.11709	.00102	5.43	5.39	.021
85	.00828	.00104	.12777	.00093	5.07	5.04	.017
86	.00788	.00100	.13941	.00085	4.73	4.70	.014
87	.00740	.00096	.15208	.00078	4.41	4.39	.012
88	.00684	.00089	.16584	.00071	4.10	4.09	.010
89	.00622	.00082	.18075	.00064	3.81	3.80	.008
90	.00555	.00074	.19689	.00058	3.54	3.53	.006

A	P	SE(P)	Q	SE(Q)	E(N)	E(BC)	SE(E(BC))
43	.07107	.00694	.07107	.00694	34.03	13.33	.582
44	.11645	.00594	.12535	.00633	33.11	13.31	.620
45	.09293	.00340	.11437	.00490	32.19	14.15	.712
46	.07511	.00209	.10439	.00396	31.28	14.92	.799
47	.06143	.00164	.09532	.00352	30.38	15.60	.874
48	.05079	.00161	.08712	.00349	29.49	16.19	.934
49	.04242	.00167	.07971	.00370	28.60	16.69	.976
50	.03578	.00170	.07304	.00399	27.72	17.10	1.001
51	.03044	.00169	.06706	.00429	26.84	17.41	1.009
52	.02614	.00164	.06170	.00454	25.98	17.62	1.002
53	.02263	.00157	.05694	.00473	25.12	17.75	.982
54	.01976	.00148	.05273	.00487	24.28	17.79	.951
55	.01741	.00138	.04902	.00494	23.44	17.76	.912
56	.01546	.00128	.04580	.00496	22.61	17.65	.865
57	.01386	.00118	.04302	.00494	21.80	17.47	.814
58	.01254	.00108	.04066	.00487	20.99	17.23	.759
59	.01145	.00098	.03870	.00478	20.20	16.94	.703
60	.01056	.00089	.03712	.00466	19.42	16.60	.647
61	.00983	.00079	.03591	.00451	18.65	16.22	.591
62	.00925	.00070	.03505	.00435	17.89	15.81	.536
63	.00879	.00062	.03453	.00418	17.15	15.37	.484
64	.00844	.00053	.03434	.00400	16.42	14.90	.435
65	.00819	.00046	.03448	.00382	15.71	14.41	.389
66	.00802	.00040	.03496	.00363	15.01	13.91	.346
67	.00791	.00035	.03577	.00344	14.33	13.39	.306
68	.00788	.00032	.03692	.00325	13.66	12.87	.269
69	.00789	.00032	.03841	.00307	13.01	12.34	.236
70	.00795	.00034	.04026	.00289	12.38	11.82	.207
71	.00806	.00039	.04249	.00271	11.77	11.29	.180
72	.00819	.00045	.04510	.00254	11.17	10.77	.156
73	.00834	.00051	.04812	.00238	10.59	10.26	.134
74	.00851	.00058	.05157	.00222	10.02	9.75	.116
75	.00868	.00065	.05547	.00207	9.48	9.25	.099
76	.00885	.00072	.05987	.00192	8.96	8.77	.084
77	.00900	.00078	.06478	.00179	8.45	8.29	.072
78	.00913	.00084	.07026	.00166	7.96	7.83	.061
79	.00922	.00089	.07632	.00153	7.49	7.39	.051
80	.00927	.00093	.08303	.00142	7.04	6.95	.043
81	.00926	.00097	.09042	.00131	6.61	6.54	.036
82	.00918	.00099	.09855	.00120	6.20	6.14	.030
83	.00902	.00100	.10746	.00111	5.80	5.76	.025
84	.00878	.00100	.11722	.00101	5.43	5.39	.021
85	.00846	.00098	.12789	.00093	5.07	5.04	.017
86	.00805	.00095	.13951	.00085	4.73	4.70	.014
87	.00755	.00090	.15217	.00078	4.41	4.39	.012
88	.00698	.00085	.16592	.00071	4.10	4.08	.010
89	.00635	.00078	.18082	.00064	3.81	3.80	.008
90	.00566	.00070	.19695	.00058	3.54	3.53	.006

A	P	SE(P)	Q	SE(Q)	E(N)	E(BC)	SE(E(BC))
44	.07681	.00387	.07681	.00387	33.11	13.38	.322
45	.11010	.00309	.11926	.00331	32.19	13.45	.345
46	.08879	.00180	.10920	.00257	31.28	14.21	.395
47	.07246	.00113	.10004	.00208	30.38	14.89	.441
48	.05978	.00091	.09171	.00187	29.49	15.49	.480
49	.04983	.00090	.08416	.00189	28.60	16.01	.511
50	.04193	.00093	.07734	.00202	27.72	16.44	.532
51	.03561	.00095	.07118	.00220	26.84	16.77	.545
52	.03051	.00095	.06566	.00237	25.98	17.02	.548
53	.02636	.00093	.06072	.00252	25.12	17.18	.544
54	.02297	.00089	.05632	.00263	24.28	17.26	.533
55	.02018	.00084	.05244	.00271	23.44	17.26	.516
56	.01788	.00079	.04903	.00276	22.61	17.19	.494
57	.01598	.00073	.04608	.00278	21.80	17.05	.469
58	.01441	.00067	.04355	.00278	20.99	16.85	.441
59	.01310	.00062	.04142	.00275	20.20	16.60	.412
60	.01203	.00056	.03968	.00270	19.42	16.29	.382
61	.01116	.00050	.03830	.00264	18.65	15.95	.351
62	.01045	.00045	.03729	.00257	17.89	15.56	.321
63	.00988	.00040	.03662	.00248	17.15	15.15	.292
64	.00943	.00035	.03630	.00239	16.42	14.70	.264
65	.00909	.00030	.03631	.00230	15.71	14.24	.237
66	.00885	.00025	.03666	.00220	15.01	13.76	.212
67	.00868	.00022	.03736	.00210	14.33	13.26	.189
68	.00859	.00019	.03839	.00199	13.66	12.76	.167
69	.00856	.00018	.03978	.00189	13.01	12.24	.148
70	.00858	.00019	.04153	.00179	12.38	11.73	.130
71	.00865	.00021	.04366	.00169	11.77	11.22	.113
72	.00875	.00024	.04618	.00159	11.17	10.71	.099
73	.00887	.00028	.04912	.00150	10.59	10.20	.086
74	.00902	.00032	.05249	.00140	10.02	9.70	.074
75	.00917	.00036	.05633	.00131	9.48	9.21	.064
76	.00931	.00040	.06066	.00123	8.96	8.73	.055
77	.00945	.00044	.06551	.00115	8.45	8.26	.047
78	.00956	.00048	.07092	.00107	7.96	7.81	.040
79	.00964	.00051	.07694	.00099	7.49	7.36	.034
80	.00966	.00054	.08359	.00092	7.04	6.94	.028
81	.00963	.00056	.09094	.00085	6.61	6.52	.024
82	.00954	.00057	.09902	.00079	6.20	6.13	.020
83	.00936	.00058	.10789	.00073	5.80	5.75	.017
84	.00910	.00058	.11762	.00067	5.43	5.38	.014
85	.00876	.00057	.12824	.00062	5.07	5.03	.012
86	.00833	.00055	.13984	.00057	4.73	4.70	.010
87	.00781	.00053	.15247	.00052	4.41	4.38	.008
88	.00721	.00050	.16619	.00047	4.10	4.08	.006
89	.00655	.00046	.18106	.00043	3.81	3.80	.005
90	.00584	.00041	.19716	.00039	3.54	3.53	.004

A	P	SE(P)	Q	SE(Q)	E(N)	E(BC)	SE(E(BC))
45	.07828	.00395	.07828	.00395	32.19	13.16	.313
46	.11182	.00320	.12132	.00343	31.28	13.23	.335
47	.08970	.00185	.11076	.00265	30.38	13.99	.383
48	.07286	.00116	.10117	.00215	29.49	14.67	.428
49	.05987	.00093	.09249	.00193	28.60	15.27	.466
50	.04973	.00092	.08465	.00194	27.72	15.78	.495
51	.04172	.00095	.07759	.00208	26.84	16.19	.514
52	.03535	.00097	.07126	.00225	25.98	16.51	.525
53	.03022	.00096	.06560	.00242	25.12	16.74	.526
54	.02607	.00093	.06057	.00256	24.28	16.89	.520
55	.02270	.00089	.05613	.00267	23.44	16.94	.507
56	.01993	.00084	.05223	.00274	22.61	16.92	.489
57	.01766	.00078	.04883	.00278	21.80	16.83	.466
58	.01580	.00073	.04592	.00279	20.99	16.67	.440
59	.01426	.00067	.04345	.00277	20.20	16.44	.412
60	.01300	.00061	.04141	.00273	19.42	16.17	.383
61	.01197	.00055	.03978	.00268	18.65	15.85	.353
62	.01114	.00049	.03854	.00261	17.89	15.48	.323
63	.01047	.00043	.03767	.00252	17.15	15.08	.293
64	.00994	.00038	.03718	.00243	16.42	14.65	.265
65	.00954	.00033	.03704	.00233	15.71	14.20	.238
66	.00924	.00028	.03727	.00223	15.01	13.73	.213
67	.00903	.00024	.03785	.00213	14.33	13.24	.189
68	.00891	.00021	.03879	.00202	13.66	12.74	.168
69	.00885	.00019	.04009	.00191	13.01	12.23	.148
70	.00885	.00019	.04178	.00181	12.38	11.72	.129
71	.00890	.00021	.04385	.00170	11.77	11.21	.113
72	.00899	.00024	.04632	.00160	11.17	10.71	.098
73	.00911	.00027	.04922	.00150	10.59	10.20	.085
74	.00925	.00031	.05255	.00140	10.02	9.70	.073
75	.00940	.00035	.05636	.00131	9.48	9.21	.063
76	.00955	.00040	.06067	.00122	8.96	8.73	.054
77	.00968	.00044	.06550	.00114	8.45	8.27	.046
78	.00980	.00047	.07090	.00106	7.96	7.81	.039
79	.00987	.00050	.07690	.00098	7.49	7.37	.033
80	.00990	.00053	.08354	.00091	7.04	6.94	.028
81	.00987	.00055	.09088	.00084	6.61	6.53	.023
82	.00977	.00057	.09896	.00077	6.20	6.13	.019
83	.00959	.00056	.10783	.00071	5.80	5.75	.016
84	.00933	.00058	.11755	.00065	5.43	5.38	.013
85	.00898	.00057	.12818	.00060	5.07	5.03	.011
86	.00853	.00055	.13977	.00055	4.73	4.70	.009
87	.00800	.00053	.15240	.00050	4.41	4.38	.008
88	.00740	.00049	.16612	.00046	4.10	4.08	.006
89	.00672	.00046	.18100	.00041	3.81	3.80	.005
90	.00599	.00041	.19710	.00038	3.54	3.53	.004

A	P	SE(P)	Q	SE(Q)	E(N)	E(BC)	SE(E(BC))
46	.08057	.00398	.08057	.00398	31.28	12.86	.298
47	.11360	.00326	.12355	.00350	30.38	12.95	.319
48	.09069	.00187	.11254	.00270	29.49	13.71	.365
49	.07335	.00116	.10257	.00218	28.60	14.38	.407
50	.06005	.00093	.09357	.00196	27.72	14.97	.443
51	.04972	.00092	.08547	.00197	26.84	15.47	.470
52	.04161	.00095	.07821	.00210	25.98	15.87	.487
53	.03517	.00097	.07172	.00227	25.12	16.17	.496
54	.03002	.00095	.06595	.00243	24.28	16.39	.496
55	.02587	.00092	.06085	.00257	23.44	16.51	.488
56	.02251	.00088	.05636	.00267	22.61	16.55	.475
57	.01976	.00082	.05245	.00273	21.80	16.51	.456
58	.01752	.00077	.04907	.00276	20.99	16.39	.433
59	.01569	.00071	.04620	.00276	20.20	16.21	.407
60	.01419	.00064	.04380	.00273	19.42	15.98	.379
61	.01296	.00058	.04185	.00268	18.65	15.68	.351
62	.01197	.00052	.04034	.00262	17.89	15.35	.322
63	.01117	.00046	.03923	.00254	17.15	14.97	.293
64	.01054	.00041	.03852	.00245	16.42	14.56	.265
65	.01005	.00035	.03820	.00236	15.71	14.13	.238
66	.00968	.00030	.03825	.00225	15.01	13.67	.213
67	.00941	.00026	.03869	.00215	14.33	13.19	.190
68	.00924	.00023	.03951	.00204	13.66	12.70	.168
69	.00915	.00020	.04071	.00193	13.01	12.20	.148
70	.00912	.00020	.04230	.00182	12.38	11.70	.129
71	.00914	.00021	.04428	.00171	11.77	11.20	.113
72	.00921	.00023	.04669	.00161	11.17	10.69	.098
73	.00931	.00026	.04952	.00151	10.59	10.19	.085
74	.00944	.00030	.05281	.00141	10.02	9.69	.073
75	.00958	.00034	.05657	.00131	9.48	9.21	.063
76	.00972	.00038	.06083	.00122	8.96	8.73	.053
77	.00985	.00042	.06563	.00114	8.45	8.26	.045
78	.00995	.00046	.07100	.00105	7.96	7.81	.039
79	.01002	.00049	.07698	.00098	7.49	7.37	.032
80	.01005	.00052	.08361	.00090	7.04	6.94	.027
81	.01001	.00054	.09093	.00083	6.61	6.53	.023
82	.00991	.00055	.09899	.00076	6.20	6.13	.019
83	.00973	.00056	.10785	.00070	5.80	5.75	.016
84	.00946	.00056	.11756	.00064	5.43	5.38	.013
85	.00910	.00055	.12818	.00059	5.07	5.03	.011
86	.00865	.00054	.13977	.00054	4.73	4.70	.009
87	.00812	.00051	.15239	.00049	4.41	4.38	.007
88	.00750	.00048	.16611	.00044	4.10	4.08	.006
89	.00681	.00045	.18099	.00040	3.81	3.80	.005
90	.00608	.00040	.19709	.00036	3.54	3.53	.004

A	P	SE(P)	Q	SE(Q)	E(N)	E(BC)	SE(E(BC))
47	.08347	.00392	.08347	.00392	30.38	12.53	.278
48	.11538	.00325	.12589	.00351	29.49	12.62	.299
49	.09171	.00185	.11448	.00270	28.60	13.37	.341
50	.07391	.00114	.10418	.00218	27.72	14.04	.380
51	.06031	.00092	.09490	.00196	26.84	14.61	.413
52	.04960	.00091	.08658	.00196	25.98	15.10	.438
53	.04158	.00093	.07915	.00208	25.12	15.48	.454
54	.03509	.00094	.07253	.00225	24.28	15.77	.461
55	.02991	.00093	.06666	.00240	23.44	15.97	.460
56	.02575	.00089	.06149	.00253	22.61	16.07	.451
57	.02239	.00085	.05697	.00262	21.80	16.09	.437
58	.01967	.00079	.05306	.00267	20.99	16.03	.419
59	.01745	.00073	.04971	.00269	20.20	15.91	.396
60	.01564	.00067	.04689	.00269	19.42	15.71	.371
61	.01417	.00061	.04456	.00265	18.65	15.46	.345
62	.01297	.00055	.04271	.00260	17.89	15.16	.317
63	.01201	.00049	.04131	.00253	17.15	14.81	.290
64	.01125	.00043	.04034	.00245	16.42	14.43	.263
65	.01064	.00038	.03979	.00236	15.71	14.02	.237
66	.01018	.00033	.03964	.00226	15.01	13.58	.212
67	.00984	.00028	.03990	.00216	14.33	13.12	.189
68	.00961	.00024	.04056	.00205	13.66	12.64	.168
69	.00946	.00021	.04162	.00194	13.01	12.15	.148
70	.00938	.00020	.04309	.00183	12.38	11.66	.129
71	.00937	.00020	.04497	.00172	11.77	11.16	.113
72	.00941	.00022	.04728	.00162	11.17	10.66	.098
73	.00949	.00025	.05004	.00152	10.59	10.17	.085
74	.00959	.00028	.05325	.00142	10.02	9.68	.073
75	.00971	.00032	.05695	.00132	9.48	9.19	.062
76	.00984	.00036	.06116	.00123	8.96	8.72	.053
77	.00995	.00040	.06591	.00114	8.45	8.25	.045
78	.01005	.00043	.07124	.00105	7.96	7.80	.038
79	.01011	.00047	.07718	.00098	7.49	7.36	.032
80	.01013	.00049	.08378	.00090	7.04	6.93	.027
81	.01009	.00051	.09107	.00083	6.61	6.52	.023
82	.00998	.00053	.09911	.00076	6.20	6.13	.019
83	.00979	.00054	.10795	.00070	5.80	5.75	.016
84	.00952	.00054	.11765	.00064	5.43	5.38	.013
85	.00915	.00053	.12825	.00058	5.07	5.03	.011
86	.00870	.00052	.13983	.00053	4.73	4.70	.009
87	.00816	.00049	.15244	.00048	4.41	4.38	.007
88	.00754	.00046	.16615	.00044	4.10	4.08	.006
89	.00685	.00043	.18102	.00040	3.81	3.80	.005
90	.00611	.00039	.19712	.00036	3.54	3.53	.004

A	P	SE(P)	Q	SE(Q)	E(N)	E(BC)	SE(E(BC))
48	.08680	.00381	.08680	.00381	29.49	12.16	.256
49	.11712	.00318	.12825	.00345	28.60	12.27	.276
50	.09275	.00180	.11651	.00265	27.72	13.00	.315
51	.07451	.00111	.10594	.00214	26.84	13.65	.351
52	.06064	.00089	.09644	.00192	25.98	14.21	.381
53	.04996	.00088	.08794	.00192	25.12	14.68	.404
54	.04164	.00090	.08036	.00205	24.28	15.04	.418
55	.03509	.00091	.07363	.00220	23.44	15.32	.423
56	.02988	.00089	.06769	.00235	22.61	15.50	.422
57	.02571	.00085	.06247	.00247	21.80	15.58	.413
58	.02236	.00081	.05794	.00255	20.99	15.59	.399
59	.01964	.00075	.05404	.00260	20.20	15.52	.381
60	.01744	.00069	.05072	.00261	19.42	15.38	.360
61	.01565	.00063	.04795	.00260	18.65	15.17	.336
62	.01421	.00057	.04571	.00256	17.89	14.91	.311
63	.01304	.00051	.04396	.00251	17.15	14.60	.286
64	.01210	.00045	.04268	.00244	16.42	14.25	.260
65	.01136	.00040	.04185	.00235	15.71	13.87	.235
66	.01078	.00035	.04146	.00226	15.01	13.45	.211
67	.01035	.00030	.04151	.00216	14.33	13.01	.189
68	.01003	.00026	.04197	.00206	13.66	12.55	.167
69	.00981	.00022	.04286	.00195	13.01	12.08	.148
70	.00968	.00020	.04418	.00185	12.38	11.60	.130
71	.00962	.00020	.04593	.00174	11.77	11.11	.113
72	.00962	.00021	.04812	.00163	11.17	10.62	.098
73	.00966	.00023	.05077	.00153	10.59	10.14	.085
74	.00973	.00027	.05390	.00143	10.02	9.65	.073
75	.00982	.00030	.05751	.00133	9.48	9.17	.063
76	.00993	.00034	.06165	.00124	8.96	8.70	.054
77	.01002	.00037	.06635	.00115	8.45	8.24	.046
78	.01010	.00041	.07162	.00106	7.96	7.79	.039
79	.01015	.00044	.07751	.00098	7.49	7.35	.032
80	.01015	.00046	.08406	.00091	7.04	6.93	.027
81	.01010	.00049	.09132	.00083	6.61	6.52	.023
82	.00999	.00050	.09933	.00077	6.20	6.12	.019
83	.00979	.00051	.10814	.00070	5.80	5.74	.016
84	.00951	.00051	.11781	.00064	5.43	5.38	.013
85	.00915	.00050	.12839	.00059	5.07	5.03	.011
86	.00869	.00049	.13995	.00053	4.73	4.70	.009
87	.00815	.00047	.15254	.00048	4.41	4.38	.007
88	.00752	.00044	.16624	.00044	4.10	4.08	.006
89	.00683	.00041	.18110	.00040	3.81	3.80	.005
90	.00609	.00037	.19718	.00036	3.54	3.53	.004

A	P	SE(P)	Q	SE(Q)	E(N)	E(BC)	SE(E(BC))
49	.09040	.00368	.09040	.00368	28.60	11.77	.235
50	.11877	.00308	.13057	.00335	27.72	11.89	.254
51	.09378	.00173	.11859	.00257	26.84	12.60	.289
52	.07515	.00106	.10780	.00208	25.98	13.23	.323
53	.06103	.00085	.09813	.00187	25.12	13.78	.351
54	.05019	.00085	.08949	.00188	24.28	14.22	.371
55	.04178	.00087	.08180	.00200	23.44	14.57	.384
56	.03517	.00087	.07500	.00215	22.61	14.83	.389
57	.02993	.00085	.06901	.00229	21.80	14.99	.387
58	.02575	.00081	.06377	.00240	20.99	15.06	.378
59	.02239	.00077	.05923	.00248	20.20	15.06	.365
60	.01969	.00071	.05535	.00252	19.42	14.97	.348
61	.01750	.00065	.05208	.00254	18.65	14.82	.327
62	.01573	.00059	.04938	.00252	17.89	14.61	.305
63	.01430	.00053	.04722	.00248	17.15	14.34	.282
64	.01315	.00047	.04558	.00243	16.42	14.03	.258
65	.01223	.00042	.04443	.00235	15.71	13.68	.234
66	.01151	.00036	.04375	.00227	15.01	13.29	.211
67	.01095	.00031	.04353	.00218	14.33	12.88	.189
68	.01053	.00027	.04377	.00208	13.66	12.44	.169
69	.01023	.00023	.04446	.00198	13.01	11.99	.149
70	.01002	.00021	.04559	.00187	12.38	11.52	.131
71	.00990	.00020	.04718	.00177	11.77	11.05	.115
72	.00984	.00020	.04923	.00166	11.17	10.57	.100
73	.00984	.00022	.05175	.00156	10.59	10.09	.086
74	.00987	.00025	.05476	.00146	10.02	9.61	.075
75	.00993	.00028	.05828	.00136	9.48	9.14	.064
76	.01000	.00031	.06233	.00127	8.96	8.68	.055
77	.01007	.00035	.06694	.00118	8.45	8.22	.046
78	.01013	.00038	.07214	.00109	7.96	7.77	.039
79	.01016	.00041	.07797	.00101	7.49	7.34	.033
80	.01014	.00044	.08447	.00093	7.04	6.92	.028
81	.01008	.00046	.09168	.00085	6.61	6.51	.023
82	.00995	.00047	.09964	.00078	6.20	6.11	.019
83	.00975	.00048	.10841	.00072	5.80	5.74	.016
84	.00946	.00048	.11805	.00066	5.43	5.37	.013
85	.00909	.00048	.12860	.00060	5.07	5.03	.011
86	.00863	.00046	.14013	.00055	4.73	4.69	.009
87	.00809	.00044	.15270	.00050	4.41	4.38	.007
88	.00747	.00042	.16638	.00045	4.10	4.08	.006
89	.00678	.00039	.18122	.00041	3.81	3.79	.005
90	.00604	.00035	.19728	.00037	3.54	3.53	.004

A	P	SE(P)	Q	SE(Q)	E(N)	E(BC)	SE(E(BC))
50	.09410	.00357	.09410	.00357	27.72	11.37	.217
51	.12030	.00297	.13279	.00324	26.84	11.51	.235
52	.09478	.00166	.12065	.00249	25.98	12.19	.268
53	.07580	.00102	.10972	.00202	25.12	12.80	.299
54	.06146	.00083	.09994	.00183	24.28	13.32	.325
55	.05049	.00082	.09121	.00185	23.44	13.74	.344
56	.04198	.00084	.08345	.00197	22.61	14.07	.356
57	.03532	.00084	.07660	.00212	21.80	14.31	.360
58	.03005	.00082	.07059	.00226	20.99	14.46	.358
59	.02586	.00078	.06535	.00237	20.20	14.52	.350
60	.02250	.00073	.06083	.00244	19.42	14.50	.337
61	.01979	.00067	.05699	.00248	18.65	14.41	.320
62	.01761	.00062	.05377	.00249	17.89	14.25	.301
63	.01585	.00056	.05115	.00247	17.15	14.03	.280
64	.01444	.00050	.04909	.00244	16.42	13.76	.258
65	.01330	.00044	.04756	.00238	15.71	13.44	.236
66	.01240	.00039	.04655	.00231	15.01	13.09	.214
67	.01169	.00033	.04603	.00222	14.33	12.71	.193
68	.01114	.00029	.04600	.00213	13.66	12.30	.172
69	.01073	.00025	.04645	.00203	13.01	11.87	.153
70	.01044	.00022	.04736	.00193	12.38	11.42	.135
71	.01024	.00020	.04876	.00183	11.77	10.96	.118
72	.01011	.00020	.05063	.00172	11.17	10.50	.103
73	.01005	.00021	.05300	.00162	10.59	10.03	.090
74	.01003	.00023	.05587	.00152	10.02	9.56	.077
75	.01005	.00026	.05926	.00142	9.48	9.10	.067
76	.01008	.00029	.06320	.00132	8.96	8.64	.057
77	.01012	.00033	.06772	.00123	8.45	8.19	.049
78	.01014	.00036	.07283	.00114	7.96	7.75	.041
79	.01015	.00039	.07858	.00105	7.49	7.32	.035
80	.01011	.00041	.08501	.00097	7.04	6.90	.029
81	.01003	.00043	.09216	.00090	6.61	6.50	.024
82	.00989	.00045	.10007	.00082	6.20	6.10	.020
83	.00968	.00046	.10879	.00076	5.80	5.73	.017
84	.00938	.00046	.11838	.00069	5.43	5.37	.014
85	.00901	.00046	.12889	.00063	5.07	5.02	.012
86	.00855	.00044	.14039	.00058	4.73	4.69	.010
87	.00800	.00043	.15293	.00052	4.41	4.38	.008
88	.00738	.00040	.16658	.00047	4.10	4.08	.006
89	.00670	.00037	.18139	.00043	3.81	3.79	.005
90	.00597	.00034	.19744	.00039	3.54	3.52	.004

A	P	SE(P)	Q	SE(Q)	E(N)	E(BC)	SE(E(BC))
51	.09780	.00350	.09780	.00350	26.84	10.98	.202
52	.12167	.00289	.13486	.00316	25.96	11.12	.220
53	.09572	.00160	.12264	.00242	25.12	11.78	.252
54	.07646	.00099	.11165	.00197	24.28	12.36	.282
55	.06194	.00081	.10182	.00180	23.44	12.85	.306
56	.05084	.00081	.09305	.00183	22.61	13.25	.324
57	.04226	.00083	.08528	.00197	21.80	13.56	.335
58	.03555	.00083	.07842	.00212	20.99	13.78	.339
59	.03025	.00080	.07241	.00226	20.20	13.91	.337
60	.02604	.00076	.06719	.00237	19.42	13.96	.329
61	.02267	.00071	.06272	.00245	18.65	13.93	.317
62	.01996	.00065	.05893	.00249	17.89	13.83	.301
63	.01779	.00059	.05579	.00249	17.15	13.66	.283
64	.01603	.00053	.05326	.00248	16.42	13.44	.263
65	.01462	.00047	.05130	.00244	15.71	13.17	.242
66	.01349	.00041	.04991	.00238	15.01	12.85	.221
67	.01260	.00036	.04905	.00231	14.33	12.50	.200
68	.01190	.00031	.04870	.00222	13.66	12.12	.179
69	.01136	.00027	.04887	.00213	13.01	11.72	.160
70	.01095	.00023	.04953	.00203	12.38	11.29	.142
71	.01065	.00020	.05070	.00193	11.77	10.86	.125
72	.01044	.00019	.05237	.00183	11.17	10.41	.109
73	.01031	.00020	.05455	.00172	10.59	9.96	.095
74	.01023	.00022	.05726	.00162	10.02	9.50	.082
75	.01019	.00024	.06050	.00151	9.48	9.05	.071
76	.01018	.00028	.06431	.00141	8.96	8.60	.061
77	.01017	.00031	.06870	.00132	8.45	8.16	.052
78	.01017	.00034	.07371	.00122	7.96	7.72	.044
79	.01014	.00037	.07937	.00114	7.49	7.30	.037
80	.01008	.00040	.08571	.00105	7.04	6.88	.032
81	.00998	.00042	.09278	.00097	6.61	6.48	.026
82	.00982	.00043	.10062	.00089	6.20	6.09	.022
83	.00959	.00044	.10928	.00082	5.80	5.72	.018
84	.00928	.00045	.11881	.00075	5.43	5.36	.015
85	.00890	.00044	.12928	.00069	5.07	5.01	.013
86	.00844	.00043	.14073	.00063	4.73	4.68	.010
87	.00789	.00042	.15323	.00057	4.41	4.37	.008
88	.00728	.00039	.16684	.00052	4.10	4.07	.007
89	.00660	.00036	.18162	.00047	3.81	3.79	.006
90	.00588	.00033	.19764	.00042	3.54	3.52	.005

A	P	SE(P)	Q	SE(Q)	E(N)	E(BC)	SE(E(BC))
52	.10136	.00350	.10136	.00350	25.98	10.60	.192
53	.12287	.00285	.13673	.00312	25.12	10.74	.210
54	.09660	.00156	.12452	.00239	24.28	11.36	.241
55	.07712	.00097	.11355	.00195	23.44	11.91	.270
56	.06245	.00081	.10374	.00180	22.61	12.37	.294
57	.05126	.00081	.09499	.00185	21.80	12.75	.311
58	.04260	.00083	.08724	.00199	20.99	13.04	.321
59	.03584	.00083	.08042	.00216	20.20	13.24	.325
60	.03052	.00080	.07445	.00231	19.42	13.35	.323
61	.02629	.00075	.06929	.00242	18.65	13.39	.315
62	.02291	.00070	.06487	.00250	17.89	13.35	.303
63	.02019	.00064	.06116	.00254	17.15	13.24	.288
64	.01801	.00058	.05811	.00255	16.42	13.07	.271
65	.01626	.00052	.05569	.00253	15.71	12.84	.251
66	.01485	.00046	.05387	.00249	15.01	12.57	.231
67	.01373	.00040	.05262	.00244	14.33	12.26	.211
68	.01283	.00034	.05192	.00236	13.66	11.92	.190
69	.01213	.00029	.05176	.00228	13.01	11.54	.171
70	.01158	.00025	.05214	.00218	12.38	11.14	.152
71	.01117	.00022	.05305	.00208	11.77	10.73	.135
72	.01086	.00020	.05448	.00198	11.17	10.30	.119
73	.01064	.00020	.05645	.00187	10.59	9.87	.104
74	.01049	.00021	.05896	.00177	10.02	9.43	.090
75	.01039	.00023	.06203	.00166	9.48	8.99	.078
76	.01031	.00026	.06568	.00156	8.96	8.55	.067
77	.01026	.00030	.06993	.00145	8.45	8.11	.057
78	.01021	.00033	.07481	.00135	7.96	7.69	.049
79	.01014	.00036	.08035	.00126	7.49	7.27	.042
80	.01005	.00039	.08659	.00117	7.04	6.86	.035
81	.00992	.00041	.09356	.00108	6.61	6.46	.030
82	.00974	.00043	.10132	.00099	6.20	6.08	.025
83	.00949	.00044	.10990	.00092	5.80	5.70	.021
84	.00918	.00044	.11937	.00084	5.43	5.35	.017
85	.00879	.00044	.12977	.00077	5.07	5.01	.014
86	.00832	.00043	.14117	.00070	4.73	4.68	.012
87	.00777	.00041	.15362	.00064	4.41	4.37	.010
88	.00716	.00039	.16719	.00058	4.10	4.07	.008
89	.00649	.00036	.18193	.00053	3.81	3.79	.006
90	.00578	.00033	.19791	.00048	3.54	3.52	.005

A	P	SE(P)	Q	SE(Q)	E(N)	E(BC)	SE(E(BC))
53	.10471	.00356	.10471	.00356	25.12	10.23	.185
54	.12387	.00285	.13836	.00314	24.28	10.37	.203
55	.09740	.00156	.12626	.00240	23.44	10.95	.233
56	.07777	.00097	.11539	.00196	22.61	11.47	.262
57	.06300	.00082	.10566	.00182	21.80	11.90	.285
58	.05172	.00083	.09700	.00189	20.99	12.25	.302
59	.04301	.00085	.08932	.00205	20.20	12.51	.313
60	.03621	.00084	.08258	.00223	19.42	12.69	.317
61	.03085	.00081	.07669	.00239	18.65	12.79	.315
62	.02660	.00076	.07161	.00251	17.89	12.81	.307
63	.02321	.00070	.06729	.00259	17.15	12.76	.296
64	.02048	.00064	.06368	.00264	16.42	12.65	.281
65	.01829	.00057	.06075	.00265	15.71	12.48	.264
66	.01653	.00051	.05846	.00264	15.01	12.25	.245
67	.01512	.00045	.05678	.00260	14.33	11.98	.225
68	.01399	.00039	.05569	.00254	13.66	11.67	.205
69	.01309	.00033	.05518	.00247	13.01	11.33	.185
70	.01238	.00028	.05523	.00238	12.38	10.96	.166
71	.01182	.00024	.05584	.00228	11.77	10.58	.148
72	.01140	.00022	.05700	.00218	11.17	10.17	.131
73	.01107	.00020	.05873	.00208	10.59	9.76	.115
74	.01083	.00021	.06102	.00197	10.02	9.33	.101
75	.01065	.00023	.06389	.00186	9.48	8.91	.087
76	.01051	.00026	.06735	.00175	8.96	8.48	.075
77	.01039	.00029	.07144	.00164	8.45	8.06	.065
78	.01029	.00032	.07617	.00153	7.96	7.64	.055
79	.01018	.00035	.08157	.00143	7.49	7.23	.047
80	.01005	.00038	.08768	.00133	7.04	6.83	.040
81	.00989	.00040	.09454	.00123	6.61	6.44	.034
82	.00968	.00042	.10220	.00114	6.20	6.06	.028
83	.00941	.00044	.11069	.00105	5.80	5.69	.024
84	.00908	.00044	.12007	.00097	5.43	5.33	.020
85	.00867	.00044	.13040	.00089	5.07	4.99	.016
86	.00820	.00043	.14173	.00081	4.73	4.67	.014
87	.00765	.00042	.15412	.00074	4.41	4.36	.011
88	.00704	.00039	.16763	.00068	4.10	4.06	.009
89	.00637	.00037	.18232	.00062	3.81	3.78	.007
90	.00567	.00033	.19826	.00056	3.54	3.52	.006

A	P	SE(P)	Q	SE(Q)	E(N)	E(BC)	SE(E(BC))
54	.10777	.00365	.10777	.00365	24.28	9.88	.180
55	.12466	.00290	.13972	.00320	23.44	10.01	.197
56	.09811	.00158	.12782	.00244	22.61	10.56	.227
57	.07841	.00098	.11713	.00200	21.80	11.03	.255
58	.06357	.00084	.10756	.00187	20.99	11.43	.279
59	.05224	.00085	.09904	.00195	20.20	11.75	.296
60	.04348	.00087	.09150	.00213	19.42	11.99	.307
61	.03665	.00086	.08488	.00232	18.65	12.15	.311
62	.03126	.00082	.07911	.00249	17.89	12.23	.309
63	.02698	.00077	.07416	.00262	17.15	12.24	.302
64	.02357	.00071	.06996	.00271	16.42	12.18	.290
65	.02083	.00064	.06648	.00277	15.71	12.06	.276
66	.01863	.00058	.06369	.00279	15.01	11.89	.259
67	.01685	.00051	.06155	.00278	14.33	11.66	.241
68	.01543	.00044	.06003	.00274	13.66	11.40	.221
69	.01428	.00038	.05913	.00268	13.01	11.09	.202
70	.01337	.00033	.05883	.00261	12.38	10.76	.182
71	.01264	.00028	.05911	.00252	11.77	10.40	.163
72	.01207	.00024	.05997	.00242	11.17	10.02	.146
73	.01162	.00022	.06142	.00232	10.59	9.63	.129
74	.01127	.00021	.06346	.00221	10.02	9.23	.113
75	.01099	.00023	.06610	.00209	9.48	8.82	.099
76	.01077	.00025	.06936	.00198	8.96	8.40	.086
77	.01059	.00028	.07325	.00186	8.45	7.99	.074
78	.01042	.00031	.07780	.00175	7.96	7.59	.064
79	.01026	.00035	.08305	.00164	7.49	7.18	.054
80	.01008	.00038	.08902	.00153	7.04	6.79	.046
81	.00988	.00040	.09575	.00142	6.61	6.40	.039
82	.00964	.00043	.10328	.00132	6.20	6.03	.033
83	.00934	.00044	.11167	.00122	5.80	5.67	.028
84	.00899	.00045	.12095	.00113	5.43	5.32	.023
85	.00857	.00045	.13119	.00104	5.07	4.98	.019
86	.00809	.00044	.14243	.00096	4.73	4.66	.016
87	.00753	.00043	.15475	.00088	4.41	4.35	.013
88	.00692	.00040	.16819	.00080	4.10	4.06	.011
89	.00626	.00037	.18282	.00073	3.81	3.78	.009
90	.00556	.00034	.19871	.00066	3.54	3.51	.007

A	P	SE(P)	Q	SE(Q)	E(N)	E(BC)	SE(E(BC))
55	.11048	.00377	.11048	.00377	23.44	9.54	.174
56	.12524	.00296	.14079	.00328	22.61	9.67	.192
57	.09874	.00161	.12919	.00251	21.80	10.18	.221
58	.07904	.00100	.11876	.00205	20.99	10.61	.249
59	.06417	.00085	.10942	.00192	20.20	10.98	.272
60	.05281	.00087	.10110	.00201	19.42	11.27	.289
61	.04402	.00089	.09375	.00220	18.65	11.48	.300
62	.03715	.00088	.08730	.00241	17.89	11.62	.304
63	.03173	.00084	.08170	.00259	17.15	11.68	.303
64	.02743	.00079	.07690	.00274	16.42	11.68	.296
65	.02399	.00072	.07286	.00284	15.71	11.61	.285
66	.02123	.00065	.06955	.00291	15.01	11.48	.271
67	.01901	.00058	.06692	.00293	14.33	11.31	.255
68	.01721	.00051	.06496	.00293	13.66	11.08	.237
69	.01577	.00044	.06364	.00290	13.01	10.82	.218
70	.01461	.00038	.06295	.00284	12.38	10.52	.199
71	.01367	.00032	.06288	.00277	11.77	10.19	.180
72	.01292	.00028	.06342	.00268	11.17	9.84	.161
73	.01232	.00024	.06456	.00258	10.59	9.48	.144
74	.01184	.00023	.06633	.00247	10.02	9.10	.127
75	.01145	.00023	.06871	.00236	9.48	8.71	.112
76	.01113	.00025	.07173	.00224	8.96	8.31	.098
77	.01086	.00028	.07541	.00212	8.45	7.92	.085
78	.01063	.00031	.07977	.00200	7.96	7.52	.073
79	.01040	.00034	.08483	.00188	7.49	7.13	.063
80	.01017	.00037	.09063	.00177	7.04	6.74	.054
81	.00992	.00040	.09721	.00165	6.61	6.37	.046
82	.00963	.00043	.10460	.00154	6.20	6.00	.039
83	.00931	.00044	.11286	.00143	5.80	5.64	.033
84	.00893	.00045	.12203	.00133	5.43	5.30	.027
85	.00849	.00045	.13216	.00123	5.07	4.96	.023
86	.00799	.00045	.14331	.00113	4.73	4.64	.019
87	.00743	.00043	.15553	.00104	4.41	4.34	.016
88	.00681	.00041	.16889	.00095	4.10	4.05	.013
89	.00615	.00038	.18345	.00087	3.81	3.77	.011
90	.00545	.00035	.19927	.00080	3.54	3.50	.009

A	P	SE(P)	Q	SE(Q)	E(N)	E(BC)	SE(E(BC))
56	.11279	.00387	.11279	.00387	22.61	9.24	.169
57	.12561	.00303	.14158	.00336	21.80	9.35	.186
58	.09928	.00164	.13035	.00257	20.99	9.81	.214
59	.07965	.00102	.12025	.00211	20.20	10.21	.241
60	.06480	.00087	.11121	.00197	19.42	10.54	.264
61	.05343	.00089	.10316	.00207	18.65	10.80	.281
62	.04461	.00091	.09605	.00227	17.89	10.98	.291
63	.03771	.00090	.08982	.00249	17.15	11.10	.296
64	.03226	.00086	.08443	.00269	16.42	11.14	.295
65	.02793	.00080	.07983	.00285	15.71	11.13	.289
66	.02446	.00073	.07599	.00296	15.01	11.05	.279
67	.02168	.00066	.07287	.00304	14.33	10.92	.265
68	.01943	.00058	.07045	.00307	13.66	10.74	.249
69	.01761	.00051	.06870	.00307	13.01	10.51	.232
70	.01614	.00044	.06761	.00305	12.38	10.25	.214
71	.01495	.00038	.06716	.00300	11.77	9.96	.195
72	.01399	.00032	.06735	.00293	11.17	9.64	.177
73	.01320	.00028	.06817	.00284	10.59	9.30	.158
74	.01257	.00025	.06963	.00274	10.02	8.95	.141
75	.01205	.00024	.07174	.00264	9.48	8.58	.125
76	.01161	.00025	.07450	.00252	8.96	8.20	.110
77	.01124	.00027	.07794	.00240	8.45	7.82	.096
78	.01092	.00030	.08208	.00228	7.96	7.44	.083
79	.01061	.00034	.08694	.00215	7.49	7.06	.072
80	.01032	.00037	.09255	.00203	7.04	6.69	.062
81	.01001	.00040	.09895	.00191	6.61	6.32	.053
82	.00968	.00042	.10619	.00179	6.20	5.96	.045
83	.00931	.00044	.11430	.00167	5.80	5.61	.038
84	.00890	.00045	.12333	.00155	5.43	5.27	.032
85	.00843	.00046	.13334	.00144	5.07	4.94	.027
86	.00792	.00045	.14437	.00134	4.73	4.63	.022
87	.00734	.00044	.15650	.00123	4.41	4.32	.019
88	.00672	.00042	.16976	.00114	4.10	4.03	.015
89	.00605	.00039	.18423	.00104	3.81	3.76	.013
90	.00536	.00036	.19997	.00095	3.54	3.50	.010

A	P	SE(P)	Q	SE(Q)	E(N)	E(BC)	SE(E(BC))
57	.11468	.00394	.11468	.00394	21.80	8.95	.162
58	.12578	.00307	.14208	.00341	20.99	9.04	.179
59	.09974	.00167	.13131	.00262	20.20	9.46	.205
60	.08024	.00103	.12162	.00215	19.42	9.82	.231
61	.06545	.00068	.11293	.00201	18.65	10.11	.253
62	.05409	.00090	.10521	.00211	17.89	10.34	.269
63	.04526	.00092	.09839	.00232	17.15	10.50	.280
64	.03834	.00091	.09244	.00255	16.42	10.59	.285
65	.03286	.00087	.08730	.00276	15.71	10.62	.284
66	.02850	.00081	.08294	.00293	15.01	10.59	.278
67	.02500	.00074	.07934	.00306	14.33	10.50	.269
68	.02218	.00066	.07646	.00314	13.66	10.37	.256
69	.01990	.00058	.07427	.00319	13.01	10.18	.242
70	.01805	.00051	.07277	.00320	12.38	9.96	.225
71	.01654	.00044	.07193	.00318	11.77	9.70	.208
72	.01531	.00037	.07176	.00314	11.17	9.42	.190
73	.01431	.00032	.07224	.00308	10.59	9.11	.172
74	.01349	.00028	.07339	.00299	10.02	8.78	.154
75	.01281	.00026	.07520	.00290	9.48	8.43	.138
76	.01223	.00025	.07768	.00279	8.96	8.08	.122
77	.01175	.00027	.08086	.00268	8.45	7.72	.107
78	.01132	.00030	.08476	.00256	7.96	7.35	.094
79	.01092	.00033	.08940	.00243	7.49	6.99	.082
80	.01055	.00036	.09480	.00230	7.04	6.63	.070
81	.01017	.00039	.10101	.00218	6.61	6.27	.061
82	.00979	.00042	.10807	.00205	6.20	5.92	.052
83	.00937	.00044	.11602	.00192	5.80	5.57	.044
84	.00892	.00045	.12490	.00180	5.43	5.24	.037
85	.00842	.00046	.13476	.00168	5.07	4.92	.031
86	.00787	.00045	.14567	.00156	4.73	4.61	.026
87	.00728	.00044	.15767	.00145	4.41	4.31	.022
88	.00664	.00042	.17082	.00134	4.10	4.02	.018
89	.00597	.00040	.18519	.00124	3.81	3.75	.015
90	.00528	.00036	.20083	.00114	3.54	3.49	.013

A	P	SE(P)	Q	SE(Q)	E(N)	E(BC)	SE(E(BC))
58	.11614	.00398	.11614	.00398	20.99	8.68	.154
59	.12578	.00309	.14231	.00344	20.20	8.76	.170
60	.10012	.00168	.13207	.00265	19.42	9.14	.195
61	.08082	.00103	.12284	.00217	18.65	9.45	.219
62	.06613	.00088	.11458	.00203	17.89	9.71	.240
63	.05480	.00090	.10723	.00212	17.15	9.90	.255
64	.04597	.00092	.10076	.00234	16.42	10.03	.266
65	.03903	.00091	.09513	.00258	15.71	10.10	.270
66	.03352	.00087	.09029	.00280	15.01	10.12	.270
67	.02912	.00081	.08622	.00298	14.33	10.07	.265
68	.02558	.00073	.08289	.00312	13.66	9.97	.257
69	.02272	.00066	.08028	.00322	13.01	9.83	.245
70	.02040	.00058	.07838	.00328	12.38	9.65	.231
71	.01851	.00050	.07716	.00330	11.77	9.43	.216
72	.01696	.00043	.07662	.00329	11.17	9.17	.199
73	.01569	.00037	.07676	.00326	10.59	8.89	.182
74	.01464	.00031	.07758	.00320	10.02	8.59	.165
75	.01377	.00028	.07908	.00312	9.48	8.27	.149
76	.01303	.00026	.08127	.00303	8.96	7.94	.133
77	.01240	.00027	.08418	.00293	8.45	7.60	.118
78	.01185	.00029	.08782	.00282	7.96	7.25	.104
79	.01135	.00032	.09222	.00270	7.49	6.90	.091
80	.01088	.00035	.09740	.00257	7.04	6.55	.079
81	.01042	.00038	.10341	.00245	6.61	6.21	.068
82	.00997	.00041	.11027	.00232	6.20	5.86	.059
83	.00949	.00043	.11803	.00219	5.80	5.53	.050
84	.00899	.00045	.12674	.00206	5.43	5.20	.043
85	.00845	.00045	.13644	.00193	5.07	4.89	.036
86	.00787	.00045	.14720	.00180	4.73	4.58	.031
87	.00726	.00044	.15907	.00168	4.41	4.29	.026
88	.00660	.00042	.17209	.00156	4.10	4.00	.021
89	.00592	.00040	.18634	.00145	3.81	3.73	.018
90	.00522	.00036	.20187	.00134	3.54	3.48	.015

A	P	SE(P)	Q	SE(Q)	E(N)	E(BC)	SE(E(BC))
59	.11718	.00396	.11718	.00396	20.20	8.44	.145
60	.12564	.00307	.14232	.00342	19.42	8.50	.160
61	.10044	.00168	.13265	.00264	18.65	8.83	.183
62	.08140	.00103	.12394	.00217	17.89	9.11	.205
63	.06683	.00087	.11615	.00202	17.15	9.32	.225
64	.05555	.00089	.10924	.00212	16.42	9.49	.240
65	.04673	.00091	.10316	.00234	15.71	9.59	.249
66	.03977	.00090	.09789	.00258	15.01	9.64	.254
67	.03423	.00086	.09340	.00281	14.33	9.63	.254
68	.02979	.00080	.08966	.00300	13.66	9.57	.250
69	.02621	.00072	.08665	.00316	13.01	9.47	.242
70	.02330	.00065	.08435	.00327	12.38	9.32	.232
71	.02093	.00057	.08276	.00334	11.77	9.13	.219
72	.01899	.00049	.08187	.00337	11.17	8.91	.204
73	.01740	.00042	.08167	.00337	10.59	8.66	.189
74	.01607	.00036	.08216	.00335	10.02	8.39	.173
75	.01497	.00031	.08335	.00330	9.48	8.09	.157
76	.01403	.00028	.08525	.00323	8.96	7.79	.142
77	.01323	.00027	.08788	.00315	8.45	7.46	.127
78	.01253	.00028	.09126	.00305	7.96	7.14	.112
79	.01191	.00031	.09541	.00294	7.49	6.80	.099
80	.01133	.00034	.10035	.00283	7.04	6.47	.087
81	.01078	.00037	.10613	.00270	6.61	6.13	.076
82	.01024	.00039	.11279	.00258	6.20	5.80	.065
83	.00969	.00042	.12035	.00245	5.80	5.48	.056
84	.00913	.00044	.12888	.00232	5.43	5.16	.048
85	.00854	.00044	.13841	.00218	5.07	4.85	.041
86	.00792	.00044	.14900	.00205	4.73	4.55	.035
87	.00727	.00044	.16071	.00192	4.41	4.26	.029
88	.00659	.00042	.17360	.00180	4.10	3.98	.025
89	.00589	.00039	.18772	.00167	3.81	3.72	.021
90	.00518	.00036	.20312	.00155	3.54	3.46	.017

A	P	SE(P)	Q	SE(Q)	E(N)	E(BC)	SE(E(BC))
60	.11782	.00391	.11782	.00391	19.42	8.22	.136
61	.12538	.00302	.14213	.00336	18.65	8.25	.150
62	.10072	.00165	.13308	.00261	17.89	8.54	.170
63	.08197	.00101	.12494	.00215	17.15	8.78	.191
64	.06755	.00086	.11766	.00201	16.42	8.96	.209
65	.05634	.00088	.11122	.00210	15.71	9.09	.223
66	.04753	.00090	.10558	.00232	15.01	9.17	.233
67	.04056	.00089	.10072	.00257	14.33	9.19	.237
68	.03499	.00085	.09661	.00281	13.66	9.17	.237
69	.03050	.00079	.09324	.00301	13.01	9.10	.234
70	.02687	.00071	.09059	.00317	12.38	8.98	.227
71	.02392	.00063	.08866	.00330	11.77	8.83	.217
72	.02149	.00055	.08743	.00338	11.17	8.64	.205
73	.01950	.00047	.08690	.00342	10.59	8.42	.192
74	.01784	.00040	.08708	.00344	10.02	8.17	.178
75	.01645	.00035	.08797	.00343	9.48	7.90	.163
76	.01528	.00030	.08958	.00339	8.96	7.62	.148
77	.01428	.00028	.09193	.00333	8.45	7.32	.134
78	.01340	.00028	.09504	.00325	7.96	7.01	.120
79	.01263	.00030	.09893	.00316	7.49	6.69	.106
80	.01192	.00032	.10364	.00306	7.04	6.38	.094
81	.01125	.00035	.10919	.00294	6.61	6.06	.082
82	.01062	.00038	.11562	.00283	6.20	5.74	.072
83	.00999	.00040	.12298	.00270	5.80	5.42	.062
84	.00935	.00042	.13131	.00257	5.43	5.11	.054
85	.00870	.00043	.14065	.00244	5.07	4.81	.046
86	.00803	.00044	.15107	.00231	4.73	4.52	.039
87	.00734	.00043	.16262	.00217	4.41	4.23	.033
88	.00663	.00041	.17535	.00204	4.10	3.96	.028
89	.00590	.00039	.18932	.00191	3.81	3.70	.024
90	.00517	.00036	.20459	.00178	3.54	3.45	.020

A	P	SE(P)	Q	SE(Q)	E(N)	E(BC)	SE(E(BC))
61	.11809	.00382	.11809	.00382	18.65	8.01	.127
62	.12506	.00294	.14180	.00328	17.89	8.02	.140
63	.10097	.00162	.13341	.00255	17.15	8.27	.159
64	.08255	.00099	.12586	.00211	16.42	8.47	.178
65	.06830	.00084	.11913	.00198	15.71	8.62	.195
66	.05717	.00086	.11319	.00208	15.01	8.72	.208
67	.04838	.00089	.10803	.00231	14.33	8.77	.217
68	.04139	.00088	.10361	.00256	13.66	8.77	.221
69	.03578	.00083	.09993	.00281	13.01	8.73	.222
70	.03125	.00077	.09697	.00302	12.38	8.64	.218
71	.02757	.00069	.09473	.00320	11.77	8.52	.212
72	.02455	.00061	.09319	.00333	11.17	8.36	.203
73	.02207	.00053	.09236	.00342	10.59	8.17	.192
74	.02000	.00046	.09225	.00348	10.02	7.95	.180
75	.01828	.00039	.09285	.00351	9.48	7.70	.167
76	.01682	.00034	.09418	.00350	8.96	7.44	.153
77	.01557	.00030	.09627	.00348	8.45	7.16	.140
78	.01449	.00029	.09911	.00343	7.96	6.88	.126
79	.01353	.00029	.10276	.00336	7.49	6.58	.113
80	.01267	.00031	.10722	.00327	7.04	6.27	.101
81	.01187	.00034	.11254	.00318	6.61	5.97	.089
82	.01112	.00037	.11875	.00307	6.20	5.66	.078
83	.01039	.00039	.12589	.00295	5.80	5.36	.068
84	.00967	.00041	.13402	.00283	5.43	5.06	.059
85	.00894	.00042	.14317	.00270	5.07	4.77	.051
86	.00821	.00043	.15341	.00257	4.73	4.48	.044
87	.00747	.00042	.16478	.00243	4.41	4.20	.037
88	.00671	.00041	.17734	.00230	4.10	3.93	.032
89	.00595	.00039	.19116	.00216	3.81	3.68	.027
90	.00519	.00036	.20628	.00203	3.54	3.43	.023

A	P	SE(P)	Q	SE(Q)	E(N)	E(BC)	SE(E(BC))
62	.11805	.00372	.11805	.00372	17.89	7.83	.120
63	.12471	.00286	.14140	.00319	17.15	7.81	.132
64	.10122	.00158	.13367	.00249	16.42	8.01	.150
65	.08314	.00098	.12674	.00207	15.71	8.18	.168
66	.06908	.00084	.12058	.00196	15.01	8.29	.183
67	.05803	.00086	.11518	.00208	14.33	8.36	.195
68	.04926	.00088	.11051	.00231	13.66	8.39	.204
69	.04226	.00087	.10658	.00258	13.01	8.37	.208
70	.03661	.00083	.10336	.00284	12.38	8.31	.208
71	.03203	.00076	.10085	.00306	11.77	8.21	.206
72	.02829	.00068	.09904	.00325	11.17	8.08	.200
73	.02520	.00060	.09795	.00339	10.59	7.91	.192
74	.02265	.00052	.09757	.00350	10.02	7.72	.182
75	.02051	.00044	.09791	.00357	9.48	7.50	.170
76	.01870	.00037	.09898	.00361	8.96	7.26	.158
77	.01716	.00033	.10081	.00362	8.45	7.00	.145
78	.01583	.00030	.10341	.00360	7.96	6.73	.132
79	.01466	.00029	.10681	.00356	7.49	6.45	.120
80	.01361	.00030	.11104	.00350	7.04	6.17	.107
81	.01266	.00033	.11613	.00342	6.61	5.87	.096
82	.01176	.00036	.12212	.00332	6.20	5.58	.085
83	.01091	.00038	.12905	.00322	5.80	5.29	.074
84	.01009	.00040	.13698	.00311	5.43	5.00	.065
85	.00928	.00042	.14593	.00298	5.07	4.72	.057
86	.00847	.00042	.15598	.00286	4.73	4.44	.049
87	.00766	.00042	.16717	.00272	4.41	4.17	.042
88	.00685	.00041	.17956	.00259	4.10	3.90	.036
89	.00605	.00039	.19321	.00245	3.81	3.65	.030
90	.00526	.00036	.20817	.00231	3.54	3.41	.026

A	P	SE(P)	Q	SE(Q)	E(N)	E(BC)	SE(E(BC))
63	.11776	.00364	.11776	.00364	17.15	7.65	.115
64	.12439	.00280	.14100	.00312	16.42	7.61	.126
65	.10150	.00155	.13393	.00244	15.71	7.78	.143
66	.08377	.00097	.12762	.00204	15.01	7.90	.160
67	.06988	.00084	.12205	.00195	14.33	7.99	.175
68	.05892	.00087	.11720	.00210	13.66	8.03	.187
69	.05017	.00090	.11306	.00236	13.01	8.03	.195
70	.04315	.00088	.10963	.00265	12.38	7.99	.199
71	.03746	.00083	.10689	.00292	11.77	7.92	.199
72	.03282	.00076	.10486	.00316	11.17	7.81	.196
73	.02901	.00068	.10354	.00336	10.59	7.66	.191
74	.02585	.00059	.10293	.00352	10.02	7.49	.183
75	.02321	.00050	.10303	.00364	9.48	7.30	.174
76	.02099	.00043	.10387	.00373	8.96	7.08	.163
77	.01910	.00036	.10547	.00377	8.45	6.84	.151
78	.01747	.00032	.10784	.00379	7.96	6.59	.139
79	.01605	.00030	.11102	.00378	7.49	6.33	.127
80	.01478	.00030	.11502	.00375	7.04	6.05	.115
81	.01363	.00032	.11990	.00369	6.61	5.78	.103
82	.01258	.00035	.12568	.00362	6.20	5.50	.092
83	.01159	.00038	.13240	.00353	5.80	5.22	.081
84	.01064	.00040	.14012	.00343	5.43	4.94	.072
85	.00972	.00042	.14889	.00332	5.07	4.66	.063
86	.00882	.00043	.15875	.00319	4.73	4.39	.055
87	.00793	.00043	.16976	.00306	4.41	4.13	.047
88	.00706	.00042	.18197	.00292	4.10	3.87	.041
89	.00620	.00040	.19545	.00278	3.81	3.62	.035
90	.00537	.00037	.21025	.00264	3.54	3.39	.029

A	P	SE(P)	Q	SE(Q)	E(N)	E(BC)	SE(E(BC))
64	.11732	.00360	.11732	.00360	16.42	7.48	.112
65	.12416	.00278	.14066	.00310	15.71	7.41	.123
66	.10184	.00154	.13426	.00241	15.01	7.55	.140
67	.08443	.00098	.12856	.00204	14.33	7.64	.157
68	.07072	.00087	.12358	.00198	13.66	7.70	.171
69	.05983	.00090	.11928	.00216	13.01	7.72	.182
70	.05110	.00092	.11568	.00246	12.38	7.70	.190
71	.04405	.00090	.11277	.00277	11.77	7.64	.194
72	.03832	.00085	.11056	.00307	11.17	7.55	.194
73	.03361	.00077	.10904	.00333	10.59	7.43	.191
74	.02972	.00068	.10822	.00355	10.02	7.27	.186
75	.02648	.00058	.10813	.00373	9.48	7.10	.179
76	.02376	.00050	.10876	.00387	8.96	6.90	.169
77	.02144	.00042	.11015	.00396	8.45	6.68	.159
78	.01946	.00036	.11231	.00402	7.96	6.45	.148
79	.01773	.00032	.11528	.00405	7.49	6.20	.136
80	.01620	.00031	.11908	.00405	7.04	5.94	.124
81	.01483	.00033	.12376	.00403	6.61	5.68	.112
82	.01358	.00035	.12933	.00398	6.20	5.41	.101
83	.01242	.00038	.13587	.00391	5.80	5.14	.090
84	.01133	.00041	.14339	.00382	5.43	4.87	.080
85	.01029	.00043	.15197	.00371	5.07	4.61	.070
86	.00928	.00044	.16165	.00360	4.73	4.34	.061
87	.00830	.00044	.17248	.00347	4.41	4.09	.053
88	.00735	.00044	.18452	.00333	4.10	3.84	.046
89	.00642	.00042	.19783	.00319	3.81	3.60	.040
90	.00553	.00039	.21247	.00304	3.54	3.36	.034

A	P	SE(P)	Q	SE(Q)	E(N)	E(BC)	SE(E(BC))
65	.11683	.00363	.11683	.00363	15.71	7.32	.110
66	.12407	.00281	.14048	.00313	15.01	7.23	.122
67	.10226	.00155	.13471	.00243	14.33	7.33	.139
68	.08514	.00100	.12962	.00207	13.66	7.40	.155
69	.07158	.00091	.12521	.00205	13.01	7.43	.170
70	.06075	.00095	.12148	.00227	12.38	7.42	.181
71	.05203	.00097	.11842	.00260	11.77	7.38	.188
72	.04495	.00095	.11605	.00295	11.17	7.30	.192
73	.03916	.00088	.11436	.00328	10.59	7.20	.192
74	.03438	.00079	.11337	.00357	10.02	7.07	.189
75	.03041	.00069	.11310	.00382	9.48	6.91	.184
76	.02708	.00059	.11355	.00402	8.96	6.73	.177
77	.02426	.00049	.11476	.00417	8.45	6.52	.167
78	.02184	.00041	.11673	.00429	7.96	6.31	.157
79	.01975	.00036	.11951	.00436	7.49	6.07	.146
80	.01792	.00033	.12313	.00440	7.04	5.83	.134
81	.01628	.00033	.12761	.00441	6.61	5.58	.122
82	.01481	.00036	.13301	.00439	6.20	5.33	.111
83	.01345	.00039	.13935	.00434	5.80	5.07	.100
84	.01219	.00042	.14670	.00427	5.43	4.81	.089
85	.01099	.00045	.15510	.00419	5.07	4.55	.079
86	.00986	.00046	.16460	.00408	4.73	4.30	.069
87	.00877	.00047	.17526	.00396	4.41	4.05	.061
88	.00772	.00046	.18714	.00382	4.10	3.80	.053
89	.00672	.00044	.20028	.00368	3.81	3.57	.046
90	.00576	.00042	.21476	.00352	3.54	3.34	.039

A	P	SE(P)	Q	SE(Q)	E(N)	E(BC)	SE(E(BC))
66	.11640	.00374	.11640	.00374	15.01	7.17	.110
67	.12419	.00289	.14054	.00322	14.33	7.05	.121
68	.10280	.00159	.13537	.00250	13.66	7.12	.139
69	.08592	.00104	.13085	.00214	13.01	7.16	.155
70	.07248	.00096	.12700	.00215	12.38	7.17	.170
71	.06169	.00102	.12382	.00242	11.77	7.14	.181
72	.05295	.00103	.12131	.00279	11.17	7.08	.188
73	.04583	.00100	.11947	.00318	10.59	6.99	.192
74	.03997	.00092	.11833	.00354	10.02	6.87	.192
75	.03511	.00082	.11789	.00386	9.48	6.73	.189
76	.03104	.00071	.11818	.00414	8.96	6.56	.184
77	.02761	.00060	.11922	.00436	8.45	6.38	.176
78	.02469	.00050	.12102	.00454	7.96	6.17	.167
79	.02217	.00042	.12363	.00467	7.49	5.96	.156
80	.01997	.00037	.12707	.00476	7.04	5.73	.145
81	.01802	.00035	.13138	.00481	6.61	5.49	.133
82	.01628	.00037	.13661	.00483	6.20	5.24	.122
83	.01469	.00040	.14278	.00481	5.80	5.00	.110
84	.01323	.00044	.14996	.00477	5.43	4.75	.099
85	.01186	.00047	.15819	.00470	5.07	4.50	.088
86	.01057	.00049	.16753	.00461	4.73	4.25	.078
87	.00935	.00050	.17802	.00450	4.41	4.01	.069
88	.00819	.00049	.18974	.00437	4.10	3.77	.060
89	.00709	.00048	.20273	.00422	3.81	3.54	.052
90	.00606	.00045	.21706	.00407	3.54	3.31	.045

A	P	SE(P)	Q	SE(Q)	E(N)	E(BC)	SE(E(BC))
67	.11617	.00392	.11617	.00392	14.33	7.01	.110
68	.12457	.00302	.14094	.00335	13.66	6.87	.121
69	.10350	.00166	.13631	.00260	13.01	6.92	.139
70	.08678	.00109	.13234	.00224	12.38	6.94	.156
71	.07341	.00103	.12902	.00227	11.77	6.92	.170
72	.06262	.00108	.12636	.00258	11.17	6.87	.181
73	.05385	.00110	.12438	.00299	10.59	6.80	.188
74	.04666	.00106	.12308	.00342	10.02	6.69	.191
75	.04072	.00097	.12249	.00382	9.48	6.56	.191
76	.03577	.00086	.12262	.00418	8.96	6.41	.188
77	.03161	.00073	.12349	.00448	8.45	6.24	.182
78	.02807	.00061	.12513	.00473	7.96	6.05	.175
79	.02504	.00051	.12758	.00493	7.49	5.84	.165
80	.02241	.00043	.13085	.00508	7.04	5.63	.155
81	.02009	.00039	.13500	.00518	6.61	5.40	.144
82	.01803	.00039	.14006	.00525	6.20	5.17	.132
83	.01617	.00041	.14607	.00527	5.80	4.93	.120
84	.01447	.00045	.15309	.00526	5.43	4.69	.109
85	.01290	.00049	.16116	.00522	5.07	4.45	.098
86	.01144	.00051	.17034	.00515	4.73	4.21	.087
87	.01007	.00053	.18068	.00505	4.41	3.97	.077
88	.00878	.00053	.19225	.00493	4.10	3.74	.068
89	.00756	.00051	.20509	.00479	3.81	3.51	.059
90	.00643	.00049	.21927	.00463	3.54	3.29	.052

A	P	SE(P)	Q	SE(Q)	E(N)	E(BC)	SE(E(BC))
68	.11630	.00414	.11630	.00414	13.66	6.86	.110
69	.12529	.00317	.14178	.00352	13.01	6.70	.121
70	.10438	.00174	.13764	.00273	12.38	6.72	.137
71	.08774	.00114	.13415	.00235	11.77	6.72	.154
72	.07437	.00109	.13132	.00240	11.17	6.68	.168
73	.06354	.00115	.12916	.00274	10.59	6.62	.179
74	.05470	.00116	.12770	.00320	10.02	6.53	.186
75	.04743	.00111	.12693	.00366	9.48	6.41	.189
76	.04140	.00101	.12688	.00410	8.96	6.27	.189
77	.03634	.00089	.12758	.00448	8.45	6.11	.185
78	.03207	.00075	.12905	.00481	7.96	5.94	.179
79	.02842	.00062	.13132	.00508	7.49	5.74	.172
80	.02528	.00052	.13443	.00530	7.04	5.54	.162
81	.02253	.00044	.13841	.00547	6.61	5.32	.152
82	.02009	.00041	.14331	.00558	6.20	5.10	.140
83	.01792	.00043	.14916	.00565	5.80	4.87	.129
84	.01595	.00046	.15602	.00568	5.43	4.63	.118
85	.01414	.00050	.16394	.00567	5.07	4.40	.106
86	.01247	.00053	.17296	.00562	4.73	4.17	.095
87	.01092	.00055	.18316	.00555	4.41	3.93	.085
88	.00948	.00056	.19458	.00544	4.10	3.71	.075
89	.00813	.00055	.20729	.00531	3.81	3.48	.066
90	.00688	.00052	.22133	.00516	3.54	3.27	.057

A	P	SE(P)	Q	SE(Q)	E(N)	E(BC)	SE(E(BC))
69	.11696	.00437	.11696	.00437	13.01	6.69	.108
70	.12640	.00332	.14315	.00369	12.38	6.52	.118
71	.10549	.00181	.13942	.00286	11.77	6.52	.134
72	.08879	.00119	.13636	.00246	11.17	6.50	.151
73	.07534	.00114	.13398	.00253	10.59	6.45	.164
74	.06443	.00120	.13229	.00289	10.02	6.37	.175
75	.05549	.00121	.13131	.00337	9.48	6.27	.181
76	.04811	.00115	.13105	.00387	8.96	6.14	.184
77	.04196	.00104	.13154	.00433	8.45	6.00	.183
78	.03679	.00091	.13282	.00474	7.96	5.83	.180
79	.03241	.00076	.13490	.00509	7.49	5.65	.174
80	.02864	.00063	.13782	.00538	7.04	5.45	.166
81	.02538	.00052	.14162	.00561	6.61	5.25	.156
82	.02251	.00046	.14634	.00578	6.20	5.03	.146
83	.01996	.00045	.15202	.00590	5.80	4.81	.135
84	.01767	.00047	.15872	.00598	5.43	4.58	.124
85	.01559	.00050	.16648	.00601	5.07	4.36	.112
86	.01369	.00054	.17535	.00599	4.73	4.13	.101
87	.01194	.00057	.18540	.00594	4.41	3.90	.090
88	.01032	.00058	.19668	.00586	4.10	3.68	.080
89	.00882	.00057	.20925	.00574	3.81	3.46	.071
90	.00743	.00055	.22316	.00560	3.54	3.25	.062

A	P	SE(P)	Q	SE(Q)	E(N)	E(BC)	SE(E(BC))
70	.11833	.00458	.11833	.00458	12.38	6.52	.105
71	.12798	.00345	.14516	.00384	11.77	6.33	.114
72	.10685	.00187	.14177	.00297	11.17	6.32	.130
73	.08995	.00123	.13907	.00256	10.59	6.29	.145
74	.07633	.00118	.13707	.00263	10.02	6.22	.158
75	.06526	.00124	.13580	.00301	9.48	6.13	.168
76	.05618	.00125	.13527	.00352	8.96	6.02	.174
77	.04866	.00118	.13551	.00403	8.45	5.89	.176
78	.04239	.00106	.13653	.00451	7.96	5.73	.175
79	.03709	.00091	.13837	.00494	7.49	5.56	.171
80	.03258	.00076	.14107	.00530	7.04	5.38	.165
81	.02870	.00063	.14466	.00559	6.61	5.18	.157
82	.02531	.00053	.14917	.00583	6.20	4.97	.148
83	.02233	.00048	.15466	.00600	5.80	4.76	.137
84	.01967	.00048	.16118	.00612	5.43	4.54	.127
85	.01728	.00051	.16877	.00619	5.07	4.32	.116
86	.01510	.00054	.17748	.00621	4.73	4.10	.105
87	.01312	.00057	.18737	.00619	4.41	3.88	.094
88	.01129	.00059	.19850	.00613	4.10	3.66	.084
89	.00962	.00059	.21093	.00603	3.81	3.44	.074
90	.00808	.00057	.22471	.00590	3.54	3.23	.065

A	P	SE(P)	Q	SE(Q)	E(N)	E(BC)	SE(E(BC))
71	.12062	.00472	.12062	.00472	11.77	6.34	.100
72	.13009	.00355	.14793	.00395	11.17	6.14	.109
73	.10848	.00190	.14478	.00305	10.59	6.12	.123
74	.09122	.00125	.14236	.00263	10.02	6.07	.138
75	.07732	.00120	.14068	.00271	9.48	6.00	.150
76	.06601	.00127	.13978	.00311	8.96	5.90	.159
77	.05674	.00126	.13966	.00363	8.45	5.78	.164
78	.04906	.00118	.14035	.00416	7.96	5.64	.166
79	.04263	.00105	.14189	.00464	7.49	5.48	.164
80	.03720	.00089	.14430	.00507	7.04	5.31	.160
81	.03257	.00074	.14762	.00543	6.61	5.12	.154
82	.02856	.00061	.15189	.00572	6.20	4.92	.146
83	.02506	.00053	.15715	.00595	5.80	4.72	.137
84	.02197	.00049	.16344	.00612	5.43	4.50	.127
85	.01921	.00050	.17083	.00623	5.07	4.29	.117
86	.01672	.00054	.17936	.00628	4.73	4.07	.106
87	.01447	.00057	.18908	.00629	4.41	3.85	.096
88	.01241	.00059	.20005	.00625	4.10	3.64	.086
89	.01054	.00059	.21233	.00617	3.81	3.43	.076
90	.00884	.00058	.22597	.00605	3.54	3.22	.067

A	P	SE(P)	Q	SE(Q)	E(N)	E(BC)	SE(E(BC))
72	.12405	.00479	.12405	.00479	11.17	6.14	.095
73	.13279	.00360	.15160	.00402	10.59	5.94	.103
74	.11041	.00190	.14857	.00309	10.02	5.91	.116
75	.09259	.00125	.14633	.00267	9.48	5.86	.130
76	.07827	.00122	.14490	.00277	8.96	5.78	.141
77	.06665	.00128	.14429	.00320	8.45	5.67	.149
78	.05713	.00126	.14454	.00373	7.96	5.55	.153
79	.04925	.00117	.14566	.00426	7.49	5.40	.155
80	.04266	.00102	.14769	.00475	7.04	5.24	.153
81	.03709	.00086	.15066	.00517	6.61	5.06	.148
82	.03233	.00071	.15460	.00552	6.20	4.87	.142
83	.02821	.00059	.15957	.00580	5.80	4.68	.134
84	.02460	.00052	.16560	.00601	5.43	4.47	.125
85	.02141	.00051	.17274	.00616	5.07	4.26	.115
86	.01857	.00053	.18104	.00625	4.73	4.05	.105
87	.01600	.00056	.19055	.00628	4.41	3.84	.095
88	.01369	.00058	.20134	.00626	4.10	3.62	.086
89	.01159	.00059	.21345	.00620	3.81	3.42	.076
90	.00969	.00058	.22694	.00610	3.54	3.21	.067

A	P	SE(P)	Q	SE(Q)	E(N)	E(BC)	SE(E(BC))
73	.12883	.00480	.12883	.00480	10.59	5.92	.090
74	.13616	.00363	.15630	.00407	10.02	5.72	.098
75	.11264	.00188	.15326	.00311	9.48	5.69	.110
76	.09403	.00126	.15109	.00271	8.96	5.64	.123
77	.07915	.00125	.14981	.00286	8.45	5.55	.134
78	.06712	.00130	.14943	.00332	7.96	5.45	.141
79	.05730	.00126	.14999	.00388	7.49	5.32	.145
80	.04920	.00115	.15150	.00442	7.04	5.17	.145
81	.04243	.00099	.15400	.00491	6.61	5.00	.143
82	.03672	.00082	.15752	.00532	6.20	4.83	.138
83	.03184	.00067	.16209	.00566	5.80	4.64	.131
84	.02761	.00057	.16777	.00591	5.43	4.44	.123
85	.02391	.00052	.17460	.00610	5.07	4.24	.114
86	.02064	.00052	.18261	.00622	4.73	4.03	.105
87	.01773	.00055	.19188	.00627	4.41	3.82	.095
88	.01512	.00058	.20244	.00627	4.10	3.61	.086
89	.01277	.00059	.21435	.00622	3.81	3.41	.076
90	.01065	.00059	.22767	.00612	3.54	3.20	.068

A	P	SE(P)	Q	SE(Q)	E(N)	E(BC)	SE(E(BC))
74	.13521	.00483	.13521	.00483	10.02	5.69	.086
75	.14027	.00369	.16220	.00416	9.48	5.50	.095
76	.11519	.00187	.15899	.00315	8.96	5.47	.107
77	.09552	.00129	.15676	.00279	8.45	5.41	.120
78	.07991	.00131	.15552	.00303	7.96	5.33	.131
79	.06738	.00135	.15529	.00357	7.49	5.22	.137
80	.05721	.00129	.15609	.00417	7.04	5.09	.141
81	.04885	.00114	.15794	.00474	6.61	4.94	.140
82	.04190	.00096	.16088	.00523	6.20	4.78	.137
83	.03605	.00078	.16494	.00564	5.80	4.60	.131
84	.03106	.00064	.17016	.00595	5.43	4.41	.124
85	.02674	.00055	.17657	.00618	5.07	4.21	.116
86	.02298	.00053	.18423	.00633	4.73	4.01	.107
87	.01965	.00055	.19317	.00641	4.41	3.81	.097
88	.01670	.00058	.20345	.00643	4.10	3.61	.088
89	.01406	.00060	.21511	.00638	3.81	3.40	.078
90	.01171	.00061	.22822	.00629	3.54	3.20	.069

A	P	SE(P)	Q	SE(Q)	E(N)	E(BC)	SE(E(BC))
75	.14769	.00920	.14769	.00920	9.48	5.26	.164
76	.14709	.00726	.17257	.00831	8.96	5.08	.184
77	.12010	.00358	.17029	.00624	8.45	5.04	.217
78	.09885	.00274	.16894	.00586	7.96	4.98	.249
79	.08196	.00291	.16854	.00687	7.49	4.89	.274
80	.06837	.00294	.16911	.00843	7.04	4.79	.290
81	.05734	.00270	.17069	.01005	6.61	4.66	.299
82	.04829	.00230	.17332	.01154	6.20	4.52	.300
83	.04077	.00185	.17703	.01283	5.80	4.37	.295
84	.03447	.00145	.18186	.01391	5.43	4.21	.286
85	.02913	.00117	.18787	.01478	5.07	4.03	.272
86	.02457	.00107	.19509	.01544	4.73	3.85	.255
87	.02064	.00111	.20359	.01592	4.41	3.67	.237
88	.01723	.00120	.21341	.01621	4.10	3.48	.217
89	.01426	.00127	.22460	.01634	3.81	3.29	.197
90	.01168	.00130	.23723	.01633	3.54	3.11	.177

A	P	SE(P)	Q	SE(Q)	E(N)	E(BC)	SE(E(BC))
76	.15845	.00996	.15845	.00996	8.96	4.97	.163
77	.15001	.00773	.17825	.00893	8.45	4.82	.184
78	.12256	.00379	.17722	.00674	7.96	4.76	.217
79	.10075	.00296	.17707	.00644	7.49	4.68	.251
80	.08327	.00316	.17783	.00768	7.04	4.58	.278
81	.06911	.00315	.17953	.00953	6.61	4.47	.296
82	.05755	.00285	.18222	.01146	6.20	4.34	.306
83	.04803	.00239	.18593	.01324	5.80	4.20	.308
84	.04010	.00189	.19072	.01482	5.43	4.05	.305
85	.03346	.00147	.19664	.01617	5.07	3.89	.297
86	.02785	.00122	.20374	.01729	4.73	3.72	.284
87	.02308	.00117	.21206	.01818	4.41	3.55	.268
88	.01901	.00124	.22168	.01885	4.10	3.38	.249
89	.01553	.00133	.23264	.01931	3.81	3.20	.230
90	.01255	.00138	.24501	.01958	3.54	3.03	.210

A	P	SE(P)	Q	SE(Q)	E(N)	E(BC)	SE(E(BC))
77	.16866	.01053	.16866	.01053	8.45	4.71	.157
78	.15360	.00813	.18477	.00950	7.96	4.57	.178
79	.12529	.00394	.18487	.00719	7.49	4.50	.211
80	.10265	.00311	.18582	.00694	7.04	4.41	.243
81	.08441	.00332	.18766	.00836	6.61	4.30	.272
82	.06958	.00328	.19044	.01045	6.20	4.18	.290
83	.05744	.00292	.19420	.01263	5.80	4.05	.301
84	.04743	.00240	.19899	.01468	5.43	3.91	.305
85	.03911	.00188	.20486	.01651	5.07	3.76	.303
86	.03217	.00147	.21188	.01811	4.73	3.61	.296
87	.02633	.00127	.22010	.01945	4.41	3.45	.284
88	.02142	.00127	.22957	.02054	4.10	3.28	.269
89	.01728	.00135	.24036	.02139	3.81	3.12	.252
90	.01379	.00142	.25253	.02202	3.54	2.95	.233

A	P	SE(P)	Q	SE(Q)	E(N)	E(BC)	SE(E(BC))
78	.17837	.01080	.17837	.01080	7.96	4.47	.146
79	.15789	.00837	.19216	.00986	7.49	4.33	.167
80	.12828	.00399	.19326	.00749	7.04	4.25	.197
81	.10453	.00317	.19521	.00726	6.61	4.15	.230
82	.08535	.00338	.19806	.00882	6.20	4.04	.256
83	.06976	.00330	.20185	.01108	5.80	3.92	.275
84	.05700	.00289	.20664	.01344	5.43	3.79	.286
85	.04650	.00234	.21248	.01568	5.07	3.65	.291
86	.03782	.00181	.21944	.01770	4.73	3.50	.290
87	.03061	.00144	.22756	.01947	4.41	3.35	.283
88	.02462	.00130	.23691	.02097	4.10	3.20	.273
89	.01963	.00133	.24755	.02222	3.81	3.04	.259
90	.01549	.00140	.25955	.02322	3.54	2.89	.243

A	P	SE(P)	Q	SE(Q)	E(N)	E(BC)	SE(E(BC))
79	.18765	.01079	.18765	.01079	7.49	4.24	.133
80	.16286	.00842	.20048	.01001	7.04	4.11	.152
81	.13149	.00394	.20246	.00761	6.61	4.02	.180
82	.10634	.00316	.20530	.00743	6.20	3.92	.211
83	.08606	.00336	.20906	.00908	5.80	3.80	.236
84	.06961	.00323	.21380	.01145	5.43	3.68	.253
85	.05621	.00278	.21957	.01393	5.07	3.55	.264
86	.04523	.00221	.22643	.01628	4.73	3.41	.269
87	.03623	.00170	.23443	.01841	4.41	3.27	.268
88	.02883	.00140	.24365	.02029	4.10	3.12	.263
89	.02274	.00131	.25414	.02190	3.81	2.98	.253
90	.01775	.00135	.26596	.02323	3.54	2.83	.241

A	P	SE(P)	Q	SE(Q)	E(N)	E(BC)	SE(E(BC))
80	.19656	.01065	.19656	.01065	7.04	4.03	.120
81	.16855	.00837	.20978	.01005	6.61	3.89	.138
82	.13491	.00383	.21250	.00763	6.20	3.80	.163
83	.10806	.00313	.21612	.00753	5.80	3.70	.192
84	.08650	.00332	.22071	.00929	5.43	3.58	.215
85	.06912	.00313	.22632	.01177	5.07	3.46	.231
86	.05506	.00263	.23301	.01434	4.73	3.33	.242
87	.04365	.00206	.24083	.01677	4.41	3.20	.247
88	.03438	.00159	.24986	.01897	4.10	3.06	.246
89	.02685	.00135	.26015	.02090	3.81	2.92	.241
90	.02075	.00130	.27177	.02255	3.54	2.77	.233

A	P	SE(P)	Q	SE(Q)	E(N)	E(BC)	SE(E(BC))
81	.20516	.01065	.20516	.01065	6.61	3.82	.110
82	.17495	.00843	.22011	.01019	6.20	3.69	.127
83	.13851	.00377	.22344	.00772	5.80	3.59	.151
84	.10963	.00318	.22773	.00776	5.43	3.49	.180
85	.08664	.00334	.23305	.00973	5.07	3.38	.202
86	.06827	.00307	.23945	.01239	4.73	3.26	.217
87	.05356	.00251	.24699	.01510	4.41	3.13	.227
88	.04176	.00193	.25573	.01764	4.10	3.00	.232
89	.03230	.00150	.26574	.01993	3.81	2.86	.231
90	.02472	.00132	.27708	.02192	3.54	2.73	.226

A	P	SE(P)	Q	SE(Q)	E(N)	E(BC)	SE(E(BC))
82	.21353	.01112	.21353	.01112	6.20	3.63	.106
83	.18208	.00889	.23152	.01081	5.80	3.49	.123
84	.14223	.00389	.23533	.00816	5.43	3.40	.150
85	.11100	.00341	.24018	.00842	5.07	3.29	.177
86	.08643	.00353	.24612	.01077	4.73	3.18	.198
87	.06704	.00314	.25322	.01376	4.41	3.07	.214
88	.05171	.00250	.26155	.01675	4.10	2.94	.224
89	.03959	.00188	.27116	.01952	3.81	2.82	.228
90	.03002	.00150	.28211	.02197	3.54	2.69	.227

A	P	SE(P)	Q	SE(Q)	E(N)	E(BC)	SE(E(BC))
83	.22174	.01243	.22174	.01243	5.80	3.45	.109
84	.18996	.01007	.24408	.01233	5.43	3.30	.127
85	.14603	.00429	.24822	.00926	5.07	3.21	.156
86	.11211	.00390	.25349	.00978	4.73	3.11	.183
87	.08583	.00396	.25996	.01268	4.41	3.00	.207
88	.06540	.00341	.26769	.01622	4.10	2.89	.225
89	.04951	.00263	.27673	.01969	3.81	2.77	.235
90	.03716	.00197	.28715	.02284	3.54	2.65	.238

A	P	SE(P)	Q	SE(Q)	E(N)	E(BC)	SE(E(BC))
84	.22986	.01475	.22986	.01475	5.43	3.27	.119
85	.19858	.01219	.25785	.01504	5.07	3.11	.140
86	.14984	.00504	.26217	.01125	4.73	3.03	.169
87	.11291	.00468	.26774	.01199	4.41	2.93	.198
88	.08481	.00464	.27463	.01559	4.10	2.83	.226
89	.06337	.00389	.28289	.01989	3.81	2.72	.245
90	.04700	.00294	.29258	.02402	3.54	2.61	.255

A	P	SE(P)	Q	SE(Q)	E(N)	E(BC)	SE(E(BC))
85	.23797	.01804	.23797	.01804	5.07	3.10	.134
86	.20795	.01531	.27288	.01903	4.73	2.93	.155
87	.15360	.00609	.27722	.01417	4.41	2.85	.183
88	.11332	.00571	.28295	.01505	4.10	2.76	.219
89	.08332	.00555	.29015	.01946	3.81	2.67	.249
90	.06092	.00454	.29886	.02466	3.54	2.56	.269

APPENDIX B

TABLES BY STAGE, HISTOLOGY AND STAGE X HISTOLOGY

Unless otherwise stated, the first value of A in each of the following tables is the age of first observation of breast cancer.

STAGE 1

A	P	SE(P)	Q	SE(Q)	E(N)	E(BC)	SE(E(BC))
60	.05085	.00491	.05085	.00491	19.42	11.37	.301
61	.07248	.00384	.07637	.00403	18.65	10.95	.313
62	.06627	.00264	.07559	.00331	17.89	10.81	.340
63	.06078	.00185	.07500	.00279	17.15	10.66	.366
64	.05594	.00145	.07462	.00253	16.42	10.48	.389
65	.05165	.00139	.07446	.00253	15.71	10.29	.408
66	.04785	.00149	.07452	.00277	15.01	10.07	.422
67	.04446	.00160	.07483	.00314	14.33	9.84	.430
68	.04145	.00169	.07540	.00357	13.66	9.60	.435
69	.03876	.00172	.07626	.00403	13.01	9.34	.434
70	.03634	.00170	.07741	.00448	12.38	9.07	.430
71	.03417	.00164	.07888	.00492	11.77	8.79	.421
72	.03220	.00155	.08070	.00533	11.17	8.50	.410
73	.03040	.00142	.08289	.00572	10.59	8.21	.396
74	.02875	.00127	.08548	.00608	10.02	7.90	.379
75	.02723	.00112	.08850	.00640	9.48	7.60	.361
76	.02580	.00096	.09199	.00669	8.96	7.29	.342
77	.02444	.00081	.09598	.00695	8.45	6.97	.321
78	.02314	.00069	.10052	.00718	7.96	6.66	.300
79	.02187	.00063	.10564	.00738	7.49	6.35	.279
80	.02063	.00064	.11139	.00755	7.04	6.04	.258
81	.01939	.00070	.11782	.00769	6.61	5.74	.237
82	.01814	.00079	.12498	.00779	6.20	5.44	.217
83	.01688	.00090	.13293	.00787	5.80	5.14	.197
84	.01561	.00099	.14172	.00793	5.43	4.86	.178
85	.01431	.00107	.15141	.00795	5.07	4.58	.161
86	.01300	.00112	.16207	.00794	4.73	4.31	.144
87	.01168	.00115	.17376	.00791	4.41	4.04	.129
88	.01036	.00115	.18655	.00786	4.10	3.79	.115
89	.00906	.00112	.20049	.00777	3.81	3.55	.101
90	.00779	.00106	.21566	.00766	3.54	3.31	.089

A	P	SE(P)	Q	SE(Q)	E(N)	E(BC)	SE(E(BC))
60	.09467	.00686	.09467	.00686	19.42	8.52	.262
61	.13245	.00585	.14630	.00637	18.65	8.36	.283
62	.10473	.00310	.13551	.00485	17.89	8.71	.324
63	.08410	.00190	.12588	.00398	17.15	9.00	.365
64	.06853	.00166	.11733	.00376	16.42	9.23	.398
65	.05662	.00172	.10983	.00399	15.71	9.39	.423
66	.04741	.00175	.10331	.00441	15.01	9.49	.438
67	.04021	.00171	.09773	.00486	14.33	9.53	.443
68	.03455	.00162	.09304	.00526	13.66	9.51	.439
69	.03004	.00148	.08922	.00558	13.01	9.43	.427
70	.02645	.00133	.08623	.00581	12.38	9.31	.410
71	.02355	.00117	.08405	.00596	11.77	9.14	.387
72	.02122	.00101	.08265	.00603	11.17	8.94	.362
73	.01931	.00086	.08202	.00603	10.59	8.70	.334
74	.01776	.00073	.08217	.00598	10.02	8.43	.305
75	.01648	.00062	.08307	.00587	9.48	8.14	.277
76	.01542	.00055	.08474	.00573	8.96	7.83	.248
77	.01452	.00052	.08718	.00556	8.45	7.51	.221
78	.01374	.00053	.09041	.00536	7.96	7.18	.195
79	.01306	.00057	.09444	.00514	7.49	6.85	.171
80	.01243	.00062	.09931	.00491	7.04	6.51	.149
81	.01184	.00068	.10503	.00467	6.61	6.17	.129
82	.01127	.00074	.11165	.00442	6.20	5.84	.111
83	.01069	.00079	.11920	.00417	5.80	5.51	.095
84	.01009	.00082	.12772	.00392	5.43	5.19	.081
85	.00945	.00084	.13727	.00367	5.07	4.87	.068
86	.00879	.00084	.14789	.00343	4.73	4.57	.058
87	.00808	.00083	.15964	.00319	4.41	4.28	.048
88	.00734	.00080	.17256	.00295	4.10	4.00	.040
89	.00657	.00075	.18672	.00273	3.81	3.73	.033
90	.00579	.00069	.20218	.00251	3.54	3.47	.028

A	P	SE(P)	Q	SE(Q)	E(N)	E(BC)	SE(E(BC))
60	.14501	.01217	.14501	.01217	19.42	5.03	.254
61	.19287	.01039	.22558	.01172	18.65	4.81	.289
62	.14161	.00458	.21387	.00921	17.89	5.07	.362
63	.10565	.00285	.20297	.00809	17.15	5.32	.443
64	.08001	.00283	.19287	.00838	16.42	5.55	.525
65	.06147	.00285	.18356	.00960	15.71	5.76	.601
66	.04785	.00266	.17504	.01117	15.01	5.95	.668
67	.03773	.00236	.16729	.01278	14.33	6.11	.725
68	.03011	.00202	.16031	.01430	13.66	6.24	.770
69	.02430	.00169	.15410	.01565	13.01	6.34	.802
70	.01983	.00140	.14864	.01681	12.38	6.41	.821
71	.01635	.00115	.14394	.01779	11.77	6.45	.827
72	.01361	.00096	.14000	.01857	11.17	6.45	.822
73	.01144	.00081	.13683	.01919	10.59	6.42	.807
74	.00970	.00072	.13443	.01963	10.02	6.36	.782
75	.00830	.00067	.13281	.01993	9.48	6.27	.750
76	.00715	.00066	.13199	.02009	8.96	6.16	.711
77	.00621	.00066	.13199	.02012	8.45	6.02	.668
78	.00542	.00068	.13282	.02004	7.96	5.86	.622
79	.00476	.00070	.13452	.01986	7.49	5.68	.574
80	.00420	.00071	.13711	.01959	7.04	5.49	.525
81	.00372	.00073	.14063	.01924	6.61	5.29	.477
82	.00330	.00073	.14511	.01882	6.20	5.07	.430
83	.00292	.00073	.15060	.01834	5.80	4.85	.386
84	.00259	.00072	.15713	.01780	5.43	4.62	.343
85	.00229	.00070	.16476	.01722	5.07	4.39	.303
86	.00202	.00067	.17353	.01660	4.73	4.16	.266
87	.00176	.00064	.18351	.01595	4.41	3.93	.232
88	.00153	.00059	.19474	.01527	4.10	3.71	.201
89	.00131	.00054	.20728	.01457	3.81	3.49	.174
90	.00111	.00049	.22119	.01386	3.54	3.27	.149

A	P	SE(P)	Q	SE(Q)	E(N)	E(BC)	SE(E(BC))
60	.31195	.01735	.31195	.01735	19.42	2.62	.118
61	.22423	.01331	.32589	.01752	18.65	2.60	.155
62	.14957	.00517	.32247	.01384	17.89	2.63	.212
63	.10030	.00460	.31919	.01478	17.15	2.66	.280
64	.06762	.00429	.31607	.01943	16.42	2.69	.357
65	.04582	.00346	.31312	.02566	15.71	2.71	.437
66	.03119	.00257	.31035	.03243	15.01	2.74	.517
67	.02133	.00187	.30779	.03933	14.33	2.76	.596
68	.01466	.00144	.30543	.04621	13.66	2.78	.673
69	.01011	.00122	.30331	.05296	13.01	2.80	.748
70	.00700	.00111	.30144	.05957	12.38	2.81	.819
71	.00486	.00102	.29983	.06598	11.77	2.82	.885
72	.00339	.00093	.29851	.07220	11.17	2.82	.947
73	.00237	.00083	.29751	.07820	10.59	2.82	1.003
74	.00166	.00072	.29684	.08397	10.02	2.82	1.052
75	.00117	.00062	.29654	.08950	9.48	2.81	1.095
76	.00082	.00052	.29664	.09478	8.96	2.80	1.130
77	.00058	.00044	.29716	.09980	8.45	2.78	1.158
78	.00041	.00036	.29815	.10455	7.96	2.76	1.178
79	.00029	.00029	.29965	.10903	7.49	2.73	1.190
80	.00020	.00023	.30169	.11320	7.04	2.69	1.194
81	.00014	.00019	.30431	.11708	6.61	2.65	1.189
82	.00010	.00015	.30757	.12064	6.20	2.61	1.177
83	.00007	.00011	.31151	.12386	5.80	2.56	1.158
84	.00005	.00009	.31618	.12675	5.43	2.50	1.131
85	.00003	.00007	.32165	.12927	5.07	2.44	1.098
86	.00002	.00005	.32796	.13142	4.73	2.38	1.064
87	.00002	.00004	.33517	.13318	4.41	2.31	1.021
88	.00001	.00003	.34335	.13453	4.10	2.23	.974
89	.00001	.00002	.35255	.13545	3.81	2.16	.923
90	.00000	.00002	.36284	.13594	3.54	2.08	.869

A	P	SE(P)	Q	SE(Q)	E(N)	E(BC)	SE(E(BC))
60	.06593	.01182	.06593	.01182	19.42	10.78	.581
61	.08905	.00934	.09533	.00992	18.65	10.53	.608
62	.07706	.00589	.09119	.00785	17.89	10.59	.663
63	.06721	.00387	.08752	.00644	17.15	10.60	.716
64	.05908	.00306	.08431	.00580	16.42	10.57	.761
65	.05235	.00300	.08158	.00584	15.71	10.50	.794
66	.04673	.00317	.07930	.00633	15.01	10.39	.814
67	.04204	.00330	.07748	.00702	14.33	10.24	.820
68	.03811	.00334	.07613	.00776	13.66	10.06	.815
69	.03480	.00327	.07524	.00845	13.01	9.85	.799
70	.03200	.00312	.07483	.00906	12.38	9.61	.773
71	.02964	.00291	.07491	.00958	11.77	9.35	.741
72	.02763	.00265	.07548	.01001	11.17	9.07	.702
73	.02591	.00237	.07657	.01033	10.59	8.76	.660
74	.02443	.00207	.07819	.01057	10.02	8.45	.615
75	.02315	.00179	.08036	.01072	9.48	8.13	.568
76	.02202	.00153	.08312	.01080	8.96	7.79	.521
77	.02101	.00133	.08648	.01081	8.45	7.45	.475
78	.02008	.00122	.09049	.01075	7.96	7.11	.429
79	.01921	.00121	.09517	.01064	7.49	6.77	.386
80	.01836	.00129	.10057	.01049	7.04	6.43	.345
81	.01753	.00142	.10673	.01029	6.61	6.09	.306
82	.01668	.00158	.11369	.01005	6.20	5.76	.270
83	.01580	.00173	.12151	.00978	5.80	5.44	.238
84	.01488	.00186	.13024	.00948	5.43	5.12	.208
85	.01390	.00196	.13993	.00916	5.07	4.81	.181
86	.01287	.00202	.15064	.00882	4.73	4.52	.157
87	.01179	.00203	.16243	.00846	4.41	4.23	.135
88	.01066	.00199	.17537	.00809	4.10	3.96	.116
89	.00950	.00191	.18950	.00771	3.81	3.69	.099
90	.00832	.00179	.20491	.00733	3.54	3.44	.084

A	P	SE(P)	Q	SE(Q)	E(N)	E(BC)	SE(E(BC))
60	.09319	.00952	.09319	.00952	19.42	8.95	.369
61	.11352	.00736	.12519	.00801	18.65	8.82	.396
62	.09409	.00430	.11861	.00632	17.89	9.01	.442
63	.07879	.00269	.11269	.00524	17.15	9.16	.488
64	.06663	.00218	.10739	.00482	16.42	9.26	.529
65	.05688	.00218	.10271	.00496	15.71	9.32	.561
66	.04902	.00227	.09864	.00542	15.01	9.33	.583
67	.04262	.00230	.09517	.00601	14.33	9.30	.595
68	.03740	.00224	.09228	.00661	13.66	9.23	.597
69	.03310	.00212	.08998	.00715	13.01	9.11	.590
70	.02955	.00196	.08827	.00762	12.38	8.97	.575
71	.02660	.00178	.08714	.00801	11.77	8.79	.554
72	.02413	.00158	.08661	.00831	11.17	8.58	.528
73	.02206	.00138	.08669	.00852	10.59	8.35	.498
74	.02031	.00119	.08737	.00867	10.02	8.09	.465
75	.01881	.00102	.08869	.00874	9.48	7.82	.430
76	.01753	.00089	.09066	.00875	8.96	7.53	.395
77	.01640	.00081	.09331	.00870	8.45	7.23	.360
78	.01540	.00078	.09666	.00861	7.96	6.93	.326
79	.01450	.00080	.10074	.00847	7.49	6.62	.293
80	.01367	.00087	.10559	.00830	7.04	6.30	.261
81	.01288	.00094	.11124	.00809	6.61	5.99	.232
82	.01212	.00103	.11774	.00786	6.20	5.67	.204
83	.01136	.00110	.12514	.00761	5.80	5.37	.179
84	.01060	.00116	.13349	.00733	5.43	5.06	.156
85	.00983	.00120	.14283	.00705	5.07	4.77	.136
86	.00904	.00121	.15322	.00675	4.73	4.48	.117
87	.00823	.00120	.16473	.00644	4.41	4.20	.100
88	.00740	.00116	.17740	.00612	4.10	3.93	.086
89	.00657	.00110	.19131	.00580	3.81	3.67	.073
90	.00573	.00102	.20650	.00548	3.54	3.42	.062

A	P	SE(P)	Q	SE(Q)	E(N)	E(BC)	SE(E(BC))
60	.11054	.00596	.11054	.00596	19.42	7.32	.230
61	.12561	.00464	.14122	.00513	18.65	7.17	.256
62	.10476	.00260	.13715	.00406	17.89	7.27	.301
63	.08792	.00170	.13340	.00351	17.15	7.34	.345
64	.07425	.00159	.12999	.00356	16.42	7.40	.386
65	.06307	.00170	.12693	.00407	15.71	7.43	.421
66	.05389	.00177	.12422	.00481	15.01	7.44	.451
67	.04631	.00175	.12187	.00562	14.33	7.43	.475
68	.04000	.00165	.11990	.00643	13.66	7.39	.492
69	.03474	.00151	.11832	.00721	13.01	7.33	.503
70	.03033	.00134	.11715	.00792	12.38	7.25	.508
71	.02660	.00116	.11640	.00858	11.77	7.15	.507
72	.02345	.00098	.11609	.00917	11.17	7.03	.501
73	.02075	.00082	.11625	.00969	10.59	6.89	.491
74	.01844	.00067	.11690	.01015	10.02	6.73	.476
75	.01645	.00056	.11807	.01055	9.48	6.55	.458
76	.01472	.00050	.11979	.01089	8.96	6.36	.437
77	.01321	.00049	.12210	.01117	8.45	6.16	.414
78	.01187	.00052	.12502	.01139	7.96	5.95	.389
79	.01068	.00057	.12860	.01156	7.49	5.73	.363
80	.00962	.00063	.13289	.01168	7.04	5.50	.337
81	.00866	.00069	.13792	.01175	6.61	5.27	.310
82	.00778	.00074	.14375	.01177	6.20	5.04	.284
83	.00697	.00077	.15043	.01175	5.80	4.80	.259
84	.00622	.00079	.15801	.01169	5.43	4.57	.235
85	.00552	.00080	.16655	.01159	5.07	4.33	.211
86	.00487	.00079	.17611	.01145	4.73	4.10	.189
87	.00425	.00077	.18676	.01127	4.41	3.87	.169
88	.00368	.00073	.19855	.01106	4.10	3.64	.150
89	.00314	.00068	.21156	.01082	3.81	3.42	.132
90	.00264	.00062	.22583	.01055	3.54	3.21	.116

A	P	SE(P)	Q	SE(Q)	E(N)	E(RC)	SE(E(RC))
60	.03146	.00513	.03146	.00513	19.42	14.57	.680
61	.04306	.00739	.04446	.00763	18.65	14.03	.692
62	.04158	.00620	.04493	.00705	17.89	13.66	.623
63	.04027	.00515	.04556	.00652	17.15	13.28	.561
64	.03910	.00423	.04635	.00603	16.42	12.89	.508
65	.03807	.00343	.04732	.00559	15.71	12.49	.462
66	.03716	.00275	.04848	.00519	15.01	12.09	.421
67	.03636	.00217	.04985	.00484	14.33	11.68	.386
68	.03565	.00171	.05145	.00453	13.66	11.27	.356
69	.03503	.00137	.05329	.00427	13.01	10.85	.329
70	.03447	.00118	.05540	.00406	12.38	10.43	.305
71	.03398	.00114	.05780	.00390	11.77	10.02	.283
72	.03352	.00120	.06052	.00378	11.17	9.60	.264
73	.03308	.00131	.06358	.00370	10.59	9.19	.246
74	.03265	.00145	.06701	.00365	10.02	8.78	.229
75	.03221	.00159	.07085	.00364	9.48	8.37	.213
76	.03173	.00171	.07513	.00365	8.96	7.97	.198
77	.03121	.00182	.07989	.00368	8.45	7.58	.184
78	.03061	.00190	.08517	.00373	7.96	7.19	.170
79	.02992	.00197	.09101	.00378	7.49	6.82	.157
80	.02913	.00202	.09745	.00384	7.04	6.45	.145
81	.02821	.00204	.10456	.00390	6.61	6.09	.132
82	.02715	.00205	.11238	.00396	6.20	5.74	.121
83	.02594	.00203	.12096	.00402	5.80	5.41	.110
84	.02457	.00200	.13037	.00407	5.43	5.09	.100
85	.02306	.00195	.14067	.00411	5.07	4.77	.090
86	.02140	.00187	.15191	.00414	4.73	4.47	.081
87	.01961	.00178	.16416	.00416	4.41	4.19	.073
88	.01772	.00166	.17750	.00417	4.10	3.91	.065
89	.01577	.00154	.19198	.00417	3.81	3.65	.058
90	.01378	.00139	.20767	.00416	3.54	3.40	.051

HISTOLOGY 1 X STAGE 2

A	P	SE(P)	Q	SE(Q)	E(N)	E(BC)	SE(E(BC))
60	.05387	.01007	.05387	.01007	19.42	10.91	.719
61	.09419	.01119	.09955	.01178	18.65	10.51	.752
62	.08039	.00690	.09436	.00923	17.89	10.62	.824
63	.06926	.00458	.08977	.00765	17.15	10.67	.890
64	.06022	.00380	.08575	.00709	16.42	10.68	.944
65	.05284	.00382	.08230	.00736	15.71	10.63	.979
66	.04679	.00401	.07941	.00807	15.01	10.54	.997
67	.04180	.00411	.07705	.00893	14.33	10.41	.998
68	.03767	.00408	.07524	.00977	13.66	10.23	.983
69	.03424	.00394	.07396	.01052	13.01	10.03	.955
70	.03139	.00370	.07322	.01114	12.38	9.79	.916
71	.02901	.00339	.07302	.01163	11.77	9.52	.870
72	.02702	.00304	.07337	.01199	11.17	9.24	.817
73	.02535	.00267	.07428	.01222	10.59	8.93	.760
74	.02394	.00231	.07576	.01235	10.02	8.60	.701
75	.02273	.00196	.07783	.01238	9.48	8.27	.641
76	.02168	.00167	.08051	.01233	8.96	7.92	.582
77	.02076	.00147	.08383	.01219	8.45	7.57	.525
78	.01992	.00139	.08781	.01199	7.96	7.22	.470
79	.01914	.00144	.09250	.01174	7.49	6.87	.418
80	.01839	.00159	.09791	.01144	7.04	6.52	.370
81	.01763	.00178	.10410	.01109	6.61	6.17	.325
82	.01686	.00199	.11112	.01072	6.20	5.83	.284
83	.01605	.00218	.11899	.01032	5.80	5.50	.247
84	.01519	.00233	.12779	.00989	5.43	5.18	.214
85	.01426	.00244	.13756	.00946	5.07	4.86	.184
86	.01326	.00250	.14835	.00901	4.73	4.56	.158
87	.01220	.00250	.16023	.00855	4.41	4.27	.135
88	.01108	.00244	.17326	.00809	4.10	3.99	.114
89	.00991	.00233	.18750	.00763	3.81	3.72	.097
90	.00872	.00217	.20301	.00717	3.54	3.46	.081

HISTOLOGY 1 X STAGE 3

A	P	SE(P)	Q	SE(Q)	E(N)	E(RC)	SE(E(RC))
60	.08028	.01643	.08028	.01643	19.42	6.64	.515
61	.15601	.01716	.16963	.01841	18.65	6.26	.543
62	.12442	.00905	.16292	.01500	17.89	6.44	.612
63	.10018	.00494	.15671	.01264	17.15	6.60	.697
64	.08140	.00382	.15099	.01151	16.42	6.73	.784
65	.06672	.00398	.14576	.01157	15.71	6.85	.867
66	.05514	.00419	.14103	.01250	15.01	6.93	.941
67	.04595	.00417	.13681	.01392	14.33	6.99	1.002
68	.03858	.00396	.13308	.01551	13.66	7.02	1.048
69	.03264	.00364	.12987	.01711	13.01	7.02	1.080
70	.02781	.00326	.12719	.01862	12.38	7.00	1.098
71	.02387	.00288	.12504	.02000	11.77	6.95	1.101
72	.02061	.00250	.12344	.02122	11.17	6.87	1.091
73	.01792	.00217	.12241	.02228	10.59	6.77	1.069
74	.01567	.00188	.12197	.02317	10.02	6.65	1.037
75	.01378	.00166	.12213	.02390	9.48	6.50	.997
76	.01217	.00149	.12294	.02448	8.96	6.34	.949
77	.01080	.00140	.12440	.02491	8.45	6.16	.896
78	.00962	.00136	.12656	.02521	7.96	5.96	.839
79	.00860	.00136	.12946	.02537	7.49	5.76	.780
80	.00770	.00139	.13312	.02541	7.04	5.54	.720
81	.00690	.00143	.13760	.02534	6.61	5.31	.659
82	.00618	.00146	.14293	.02517	6.20	5.08	.600
83	.00553	.00149	.14917	.02490	5.80	4.85	.542
84	.00493	.00149	.15636	.02453	5.43	4.61	.489
85	.00438	.00148	.16457	.02409	5.07	4.38	.436
86	.00386	.00145	.17385	.02357	4.73	4.14	.388
87	.00338	.00139	.18425	.02299	4.41	3.91	.342
88	.00293	.00131	.19585	.02234	4.10	3.68	.301
89	.00251	.00121	.20869	.02164	3.81	3.46	.263
90	.00212	.00110	.22285	.02088	3.54	3.25	.229

HISTOLOGY 2 X STAGE 1

A	P	SE(P)	Q	SE(Q)	E(N)	E(RC)	SE(E(RC))
60	.04155	.00547	.04155	.00547	19.42	12.49	.577
61	.06073	.00766	.06336	.00798	18.65	12.01	.592
62	.05676	.00607	.06322	.00729	17.89	11.79	.548
63	.05320	.00476	.06326	.00667	17.15	11.55	.512
64	.05000	.00370	.06347	.00613	16.42	11.29	.482
65	.04713	.00286	.06388	.00567	15.71	11.03	.458
66	.04454	.00222	.06449	.00531	15.01	10.74	.437
67	.04221	.00179	.06533	.00502	14.33	10.45	.419
68	.04010	.00154	.06640	.00482	13.66	10.15	.403
69	.03819	.00145	.06774	.00471	13.01	9.83	.387
70	.03645	.00145	.06936	.00466	12.38	9.51	.372
71	.03486	.00151	.07128	.00467	11.77	9.18	.357
72	.03340	.00157	.07352	.00474	11.17	8.85	.341
73	.03204	.00163	.07613	.00484	10.59	8.51	.326
74	.03076	.00168	.07911	.00497	10.02	8.17	.309
75	.02955	.00171	.08251	.00511	9.48	7.83	.293
76	.02837	.00173	.08637	.00526	8.96	7.49	.276
77	.02723	.00173	.09071	.00542	8.45	7.15	.259
78	.02609	.00173	.09558	.00557	7.96	6.82	.241
79	.02493	.00172	.10101	.00571	7.49	6.48	.224
80	.02376	.00171	.10707	.00584	7.04	6.16	.207
81	.02255	.00170	.11379	.00597	6.61	5.84	.191
82	.02129	.00168	.12123	.00607	6.20	5.52	.175
83	.01998	.00166	.12945	.00616	5.80	5.21	.160
84	.01861	.00164	.13849	.00624	5.43	4.92	.145
85	.01718	.00160	.14843	.00629	5.07	4.63	.131
86	.01570	.00155	.15933	.00632	4.73	4.35	.118
87	.01419	.00149	.17124	.00634	4.41	4.08	.106
88	.01265	.00141	.18423	.00634	4.10	3.82	.095
89	.01111	.00132	.19838	.00631	3.81	3.57	.084
90	.00960	.00121	.21374	.00626	3.54	3.33	.075

A	P	SE(P)	Q	SE(Q)	E(N)	E(BC)	SE(E(BC))
60	.07526	.00959	.07526	.00959	19.42	9.11	.503
61	.11990	.00989	.12966	.01061	18.65	8.86	.535
62	.09793	.00559	.12168	.00825	17.89	9.11	.603
63	.08096	.00350	.11452	.00685	17.15	9.31	.669
64	.06770	.00296	.10816	.00643	16.42	9.45	.725
65	.05725	.00304	.10256	.00676	15.71	9.53	.766
66	.04894	.00315	.09769	.00746	15.01	9.57	.791
67	.04229	.00314	.09355	.00826	14.33	9.55	.801
68	.03692	.00303	.09010	.00901	13.66	9.49	.797
69	.03256	.00283	.08733	.00966	13.01	9.38	.781
70	.02900	.00259	.08523	.01018	12.38	9.23	.754
71	.02608	.00231	.08380	.01057	11.77	9.04	.719
72	.02368	.00203	.08303	.01083	11.17	8.82	.678
73	.02169	.00175	.08292	.01098	10.59	8.58	.632
74	.02002	.00149	.08348	.01103	10.02	8.31	.584
75	.01862	.00127	.08472	.01099	9.48	8.02	.535
76	.01743	.00110	.08665	.01087	8.96	7.72	.486
77	.01641	.00102	.08929	.01069	8.45	7.40	.437
78	.01551	.00101	.09267	.01045	7.96	7.08	.391
79	.01470	.00107	.09680	.01017	7.49	6.75	.347
80	.01395	.00118	.10172	.00984	7.04	6.42	.306
81	.01324	.00130	.10747	.00949	6.61	6.10	.269
82	.01254	.00142	.11408	.00911	6.20	5.77	.234
83	.01184	.00152	.12160	.00872	5.80	5.45	.203
84	.01113	.00160	.13008	.00831	5.43	5.14	.175
85	.01039	.00165	.13956	.00789	5.07	4.83	.150
86	.00961	.00166	.15010	.00747	4.73	4.53	.128
87	.00880	.00165	.16175	.00704	4.41	4.25	.109
88	.00797	.00159	.17458	.00662	4.10	3.97	.092
89	.00710	.00151	.18864	.00621	3.81	3.71	.077
90	.00623	.00139	.20398	.00580	3.54	3.45	.065

A	P	SE(P)	Q	SE(Q)	E(N)	E(BC)	SE(E(BC))
60	.11445	.01572	.11445	.01572	19.42	5.50	.337
61	.18281	.01456	.20643	.01603	18.65	5.15	.364
62	.13866	.00682	.19732	.01288	17.89	5.36	.428
63	.10651	.00364	.18881	.01087	17.15	5.57	.507
64	.08279	.00327	.18092	.01019	16.42	5.75	.590
65	.06508	.00347	.17364	.01070	15.71	5.91	.670
66	.05171	.00346	.16697	.01195	15.01	6.05	.742
67	.04151	.00325	.16090	.01353	14.33	6.17	.805
68	.03365	.00292	.15545	.01518	13.66	6.26	.856
69	.02754	.00256	.15061	.01677	13.01	6.33	.894
70	.02273	.00219	.14639	.01822	12.38	6.36	.920
71	.01893	.00186	.14280	.01951	11.77	6.37	.932
72	.01589	.00158	.13984	.02063	11.17	6.35	.932
73	.01344	.00134	.13754	.02157	10.59	6.30	.921
74	.01146	.00117	.13591	.02234	10.02	6.23	.899
75	.00983	.00104	.13496	.02295	9.48	6.14	.869
76	.00849	.00096	.13472	.02340	8.96	6.02	.832
77	.00737	.00092	.13521	.02371	8.45	5.88	.789
78	.00643	.00091	.13646	.02388	7.96	5.72	.741
79	.00564	.00091	.13851	.02393	7.49	5.55	.690
80	.00496	.00092	.14138	.02387	7.04	5.36	.638
81	.00437	.00093	.14512	.02370	6.61	5.16	.585
82	.00386	.00093	.14977	.02343	6.20	4.95	.533
83	.00340	.00093	.15537	.02308	5.80	4.74	.484
84	.00300	.00092	.16198	.02264	5.43	4.52	.435
85	.00263	.00089	.16964	.02214	5.07	4.30	.388
86	.00230	.00086	.17842	.02157	4.73	4.08	.344
87	.00199	.00081	.18836	.02094	4.41	3.86	.304
88	.00171	.00075	.19952	.02026	4.10	3.64	.266
89	.00146	.00069	.21197	.01954	3.81	3.43	.232
90	.00122	.00062	.22576	.01877	3.54	3.22	.202

HISTOLOGY 3 X STAGE 1

A	P	SE(P)	Q	SE(Q)	E(N)	E(BC)	SE(E(BC))
60	.04806	.00549	.04806	.00549	19.42	11.31	.443
61	.06820	.00664	.07164	.00696	18.65	10.86	.456
62	.06357	.00524	.07194	.00646	17.89	10.65	.416
63	.05936	.00407	.07237	.00599	17.15	10.44	.382
64	.05551	.00310	.07297	.00556	16.42	10.22	.353
65	.05200	.00230	.07373	.00517	15.71	9.98	.328
66	.04878	.00167	.07468	.00482	15.01	9.74	.307
67	.04583	.00122	.07582	.00451	14.33	9.49	.289
68	.04312	.00096	.07718	.00426	13.66	9.22	.273
69	.04062	.00090	.07878	.00406	13.01	8.95	.260
70	.03830	.00097	.08064	.00391	12.38	8.68	.249
71	.03614	.00110	.08278	.00382	11.77	8.39	.239
72	.03413	.00122	.08522	.00377	11.17	8.11	.229
73	.03224	.00133	.08800	.00378	10.59	7.81	.220
74	.03045	.00141	.09114	.00382	10.02	7.52	.210
75	.02875	.00146	.09468	.00390	9.48	7.23	.201
76	.02712	.00149	.09864	.00400	8.96	6.93	.192
77	.02554	.00150	.10307	.00412	8.45	6.63	.183
78	.02401	.00149	.10801	.00426	7.96	6.34	.173
79	.02250	.00147	.11350	.00441	7.49	6.05	.163
80	.02102	.00144	.11958	.00456	7.04	5.76	.153
81	.01954	.00139	.12631	.00471	6.61	5.47	.144
82	.01808	.00134	.13373	.00486	6.20	5.19	.134
83	.01662	.00128	.14190	.00500	5.80	4.92	.124
84	.01516	.00121	.15088	.00513	5.43	4.65	.115
85	.01372	.00114	.16074	.00526	5.07	4.39	.106
86	.01228	.00107	.17152	.00537	4.73	4.13	.097
87	.01088	.00099	.18330	.00546	4.41	3.89	.088
88	.00950	.00090	.19613	.00554	4.10	3.65	.080
89	.00818	.00082	.21010	.00560	3.81	3.42	.073
90	.00693	.00073	.22525	.00564	3.54	3.20	.065

A	P	SE(P)	Q	SE(Q)	E(N)	E(BC)	SE(E(BC))
60	.08892	.00851	.08892	.00851	19.42	7.35	.373
61	.13373	.00916	.14678	.00997	18.65	7.08	.401
62	.11017	.00540	.14173	.00843	17.89	7.21	.438
63	.09146	.00330	.13708	.00742	17.15	7.32	.481
64	.07648	.00246	.13283	.00697	16.42	7.41	.523
65	.06440	.00232	.12900	.00703	15.71	7.47	.563
66	.05460	.00235	.12557	.00747	15.01	7.50	.596
67	.04661	.00235	.12257	.00813	14.33	7.51	.622
68	.04004	.00227	.12000	.00889	13.66	7.49	.641
69	.03461	.00214	.11787	.00966	13.01	7.45	.652
70	.03009	.00196	.11620	.01042	12.38	7.38	.655
71	.02632	.00177	.11500	.01111	11.77	7.28	.651
72	.02315	.00158	.11428	.01174	11.17	7.17	.641
73	.02047	.00140	.11408	.01230	10.59	7.03	.624
74	.01818	.00124	.11440	.01278	10.02	6.87	.602
75	.01623	.00112	.11529	.01319	9.48	6.69	.576
76	.01454	.00103	.11676	.01352	8.96	6.50	.547
77	.01307	.00098	.11884	.01378	8.45	6.30	.515
78	.01178	.00096	.12158	.01397	7.96	6.08	.482
79	.01064	.00097	.12500	.01409	7.49	5.85	.447
80	.00962	.00100	.12915	.01415	7.04	5.62	.413
81	.00870	.00103	.13408	.01415	6.61	5.38	.378
82	.00786	.00106	.13983	.01409	6.20	5.14	.344
83	.00708	.00108	.14645	.01399	5.80	4.89	.311
84	.00635	.00109	.15400	.01383	5.43	4.65	.281
85	.00567	.00109	.16253	.01363	5.07	4.40	.251
86	.00503	.00107	.17210	.01339	4.73	4.16	.224
87	.00442	.00103	.18277	.01311	4.41	3.93	.198
88	.00385	.00097	.19461	.01279	4.10	3.70	.174
89	.00331	.00090	.20767	.01243	3.81	3.47	.153
90	.00280	.00082	.22202	.01205	3.54	3.25	.133

A	P	SE(P)	Q	SE(Q)	E(N)	E(BC)	SE(E(BC))
60	.13599	.01418	.13599	.01418	19.42	4.61	.237
61	.19486	.01419	.22553	.01601	18.65	4.26	.261
62	.14722	.00674	.22002	.01387	17.89	4.36	.265
63	.11211	.00310	.21481	.01222	17.15	4.45	.280
64	.08602	.00233	.20990	.01113	16.42	4.54	.308
65	.06648	.00261	.20532	.01066	15.71	4.62	.341
66	.05174	.00276	.20107	.01078	15.01	4.68	.377
67	.04053	.00269	.19716	.01138	14.33	4.74	.413
68	.03195	.00248	.19361	.01229	13.66	4.79	.447
69	.02534	.00221	.19043	.01340	13.01	4.82	.479
70	.02022	.00193	.18764	.01461	12.38	4.85	.506
71	.01622	.00167	.18527	.01584	11.77	4.85	.529
72	.01307	.00143	.18333	.01705	11.17	4.85	.547
73	.01059	.00123	.18185	.01821	10.59	4.83	.560
74	.00862	.00107	.18084	.01931	10.02	4.80	.567
75	.00704	.00093	.18036	.02033	9.48	4.75	.568
76	.00577	.00082	.18041	.02127	8.96	4.69	.565
77	.00475	.00073	.18104	.02212	8.45	4.61	.557
78	.00391	.00066	.18229	.02288	7.96	4.52	.546
79	.00323	.00060	.18418	.02354	7.49	4.42	.529
80	.00268	.00055	.18678	.02412	7.04	4.31	.509
81	.00221	.00050	.19012	.02460	6.61	4.19	.486
82	.00183	.00046	.19425	.02498	6.20	4.06	.461
83	.00151	.00042	.19923	.02527	5.80	3.92	.434
84	.00125	.00039	.20511	.02548	5.43	3.78	.406
85	.00103	.00035	.21194	.02559	5.07	3.63	.377
86	.00084	.00031	.21979	.02561	4.73	3.48	.348
87	.00068	.00028	.22873	.02554	4.41	3.32	.319
88	.00055	.00024	.23880	.02538	4.10	3.16	.290
89	.00044	.00021	.25008	.02514	3.81	3.01	.263
90	.00034	.00018	.26264	.02481	3.54	2.85	.237

AGE AT FIRST OBSERVATION	H1 E(BC)	H2 E(BC)	H3 E(BC)
35	20.04	15.82	10.97
36	19.65	15.58	10.81
37	19.32	15.34	10.70
38	18.98	15.13	10.60
39	18.67	14.92	10.57
40	18.37	14.74	10.50
41	18.10	14.61	10.44
42	17.85	14.45	10.39
43	17.61	14.25	10.35
44	17.75	14.38	10.58
45	17.33	14.10	10.37
46	16.83	13.76	10.13
47	16.33	13.38	9.87
48	15.81	12.98	9.60
49	15.28	12.57	9.33
50	14.76	12.19	9.10
51	14.26	11.77	8.84
52	13.77	11.38	8.60
53	13.30	10.99	8.37
54	12.86	10.63	8.17
55	12.45	10.30	7.98
56	12.07	9.98	7.81
57	11.71	9.69	7.67
58	11.38	9.43	7.53
59	11.07	9.18	7.42
60	10.78	8.95	7.32
61	10.50	8.74	7.23
62	10.25	8.55	7.14
63	10.00	8.36	7.06
64	9.75	8.18	6.98
65	9.51	8.01	6.89
66	9.27	7.84	6.80
67	9.05	7.66	6.70
68	8.79	7.48	6.59
69	8.53	7.29	6.46
70	8.25	7.09	6.32
71	7.96	6.87	6.15
72	7.65	6.64	5.96
73	7.31	6.39	5.75
74	6.97	6.12	5.51
75	6.48	5.68	5.15
76	6.17	5.39	4.92
77	5.84	5.11	4.69
78	5.54	4.85	4.47
79	5.24	4.60	4.26
80	4.95	4.36	4.05
81	4.68	4.14	3.85
82	4.42	3.92	3.66
83	4.16	3.71	3.47
84	3.92	3.51	3.29
85	3.69	3.32	3.12

AGE AT FIRST OBSERVATION	S1 E(BC)	S2 F(BC)	S3 E(BC)	S4 F(BC)
35	19.70	15.54	6.99	2.91
36	19.29	15.32	6.97	2.95
37	18.93	15.05	6.96	3.00
38	18.59	14.84	6.94	3.04
39	18.29	14.65	6.93	3.07
40	18.03	14.46	6.92	3.09
41	17.79	14.28	6.91	3.11
42	17.58	14.12	6.90	3.11
43	17.39	13.95	6.89	3.11
44	17.79	13.97	6.97	3.13
45	17.29	13.75	6.92	3.09
46	16.78	13.51	6.84	3.05
47	16.25	13.17	6.72	3.00
48	15.72	12.79	6.58	2.95
49	15.21	12.39	6.43	2.90
50	14.72	11.98	6.28	2.84
51	14.26	11.57	6.12	2.80
52	13.82	11.16	5.96	2.76
53	13.42	10.76	5.81	2.72
54	13.04	10.38	5.67	2.69
55	12.70	10.02	5.53	2.67
56	12.39	9.67	5.41	2.65
57	12.10	9.35	5.30	2.63
58	11.84	9.06	5.20	2.62
59	11.60	8.78	5.11	2.62
60	11.37	8.52	5.04	2.62
61	11.15	8.29	4.97	2.63
62	10.93	8.07	4.91	2.64
63	10.71	7.86	4.85	2.65
64	10.49	7.67	4.80	2.67
65	10.26	7.49	4.76	2.69
66	10.02	7.31	4.71	2.70
67	9.76	7.14	4.67	2.71
68	9.49	6.97	4.62	2.72
69	9.20	6.79	4.56	2.72
70	8.89	6.61	4.49	2.71
71	8.56	6.42	4.42	2.69
72	8.21	6.22	4.33	2.65
73	7.84	6.01	4.22	2.60
74	7.46	5.78	4.10	2.53
75	6.98	5.34	3.83	2.40
76	6.64	5.05	3.66	2.31
77	6.31	4.79	3.51	2.23
78	5.99	4.54	3.36	2.16
79	5.67	4.31	3.23	2.09
80	5.36	4.09	3.10	2.02
81	5.05	3.89	2.98	1.96
82	4.75	3.70	2.86	1.90
83	4.47	3.51	2.74	1.84
84	4.17	3.34	2.64	1.79
85	3.93	3.17	2.53	1.74

AGE AT FIRST OBSERVATION	H1XS1 E(BC)	H1XS2 E(BC)	H1XS3 E(BC)
35	27.81	20.58	9.84
36	27.15	20.17	9.73
37	26.53	19.79	9.63
38	25.96	19.43	9.55
39	25.43	19.09	9.48
40	24.95	18.78	9.42
41	24.49	18.49	9.37
42	24.05	18.22	9.34
43	23.65	17.96	9.31
44	23.88	18.06	9.52
45	23.14	17.65	9.39
46	22.40	17.19	9.21
47	21.66	16.69	9.01
48	20.94	16.17	8.80
49	20.23	15.64	8.57
50	19.55	15.11	8.35
51	18.91	14.59	8.04
52	18.31	14.09	7.83
53	17.74	13.61	7.63
54	17.20	13.15	7.44
55	16.70	12.71	7.27
56	16.23	12.30	7.12
57	15.78	11.92	6.98
58	15.36	11.56	6.85
59	14.95	11.23	6.74
60	14.56	10.91	6.64
61	14.18	10.62	6.55
62	13.81	10.34	6.46
63	13.43	10.07	6.38
64	13.04	9.81	6.38
65	12.65	9.55	6.30
66	12.25	9.30	6.22
67	11.83	9.04	6.13
68	11.41	8.78	6.04
69	10.98	8.51	5.93
70	10.53	8.23	5.81
71	10.07	7.94	5.68
72	9.61	7.64	5.52
73	9.14	7.33	5.28
74	8.65	7.00	5.16
75	8.12	6.51	4.83
76	7.71	6.17	4.63
77	7.30	5.81	4.43
78	6.90	5.51	4.24
79	6.51	5.21	4.05
80	6.13	4.93	3.87
81	5.76	4.66	3.70
82	5.41	4.43	3.52
83	5.07	4.18	3.36
84	4.74	3.94	3.20
85	4.42	3.70	3.04

AGE AT FIRST OBSERVATION	H2XS1 E(BC)	H2XS2 E(BC)	H2XS3 E(BC)
35	22.28	16.53	7.59
36	21.79	16.24	7.55
37	21.34	16.04	7.52
38	20.93	15.79	7.48
39	20.56	15.57	7.46
40	20.23	15.36	7.44
41	19.93	15.17	7.42
42	19.66	14.98	7.41
43	19.41	14.81	7.40
44	19.80	14.90	7.53
45	19.22	14.57	7.46
46	18.62	14.24	7.35
47	18.02	13.86	7.22
48	17.43	13.45	7.07
49	16.86	13.02	6.91
50	16.31	12.59	6.74
51	15.79	12.16	6.57
52	15.30	11.74	6.41
53	14.84	11.34	6.26
54	14.42	10.96	6.12
55	14.04	10.59	5.99
56	13.68	10.25	5.87
57	13.35	9.93	5.76
58	13.05	9.64	5.66
59	12.76	9.36	5.58
60	12.49	9.11	5.50
61	12.22	8.87	5.43
62	11.96	8.65	5.37
63	11.70	8.44	5.32
64	11.42	8.25	5.27
65	11.14	8.05	5.22
66	10.85	7.87	5.17
67	10.54	7.68	5.11
68	10.22	7.53	5.05
69	9.87	7.33	4.99
70	9.52	7.12	4.91
71	9.14	6.91	4.81
72	8.75	6.63	4.71
73	8.34	6.39	4.58
74	7.92	6.17	4.44
75	7.42	5.73	4.16
76	7.06	5.43	3.99
77	6.70	5.15	3.82
78	6.35	4.88	3.66
79	6.00	4.63	3.51
80	5.66	4.40	3.37
81	5.33	4.17	3.23
82	5.01	3.96	3.10
83	4.70	3.75	2.96
84	4.41	3.55	2.84
85	4.12	3.36	2.71

AGE AT FIRST OBSERVATION	H3XS1 E(BC)	H3XS2 E(BC)	H3XS3 E(BC)
35	18.57	11.20	5.59
36	18.19	11.10	5.59
37	17.86	10.98	5.59
38	17.56	10.87	5.58
39	17.30	10.78	5.58
40	17.07	10.70	5.59
41	16.88	10.63	5.59
42	16.71	10.57	5.60
43	16.57	10.53	5.61
44	17.10	10.78	5.74
45	16.59	10.59	5.68
46	16.08	10.36	5.59
47	15.56	10.11	5.50
48	15.06	9.84	5.40
49	14.58	9.56	5.30
50	14.13	9.23	5.20
51	13.71	8.97	5.10
52	13.33	8.72	5.01
53	12.98	8.49	4.93
54	12.67	8.27	4.86
55	12.38	8.08	4.79
56	12.13	7.90	4.74
57	11.90	7.74	4.69
58	11.69	7.60	4.66
59	11.50	7.47	4.63
60	11.31	7.35	4.61
61	11.13	7.25	4.59
62	10.95	7.15	4.58
63	10.77	7.06	4.57
64	10.57	6.97	4.56
65	10.36	6.88	4.55
66	10.13	6.78	4.53
67	9.89	6.72	4.51
68	9.62	6.60	4.48
69	9.32	6.47	4.44
70	9.01	6.32	4.38
71	8.67	6.16	4.31
72	8.31	5.93	4.23
73	7.94	5.72	4.12
74	7.54	5.54	3.99
75	7.08	5.18	3.77
76	6.75	4.95	3.64
77	6.42	4.72	3.51
78	6.09	4.51	3.38
79	5.77	4.30	3.25
80	5.45	4.09	3.13
81	5.14	3.90	3.01
82	4.84	3.70	2.89
83	4.54	3.52	2.77
84	4.26	3.34	2.66
85	3.96	3.16	2.54

APPENDIX C

TABLES BY TREATMENT, HISTOLOGY X TREATMENT AND STAGE X TREATMENT

The first value of A in each of the following tables is the age of first observation of breast cancer.

TREATMENT 1

A	P	SE(P)	Q	SE(Q)	E(N)	E(BC)	SE(E(BC))
60	.08935	.00586	.08935	.00586	19.42	9.40	.245
61	.10483	.00446	.11511	.00484	18.65	9.27	.263
62	.08837	.00267	.10966	.00382	17.89	9.41	.292
63	.07516	.00171	.10476	.00317	17.15	9.51	.321
64	.06449	.00139	.10041	.00292	16.42	9.56	.347
65	.05580	.00140	.09658	.00301	15.71	9.58	.366
66	.04869	.00146	.09328	.00331	15.01	9.55	.380
67	.04284	.00150	.09051	.00369	14.33	9.48	.386
68	.03799	.00148	.08825	.00408	13.66	9.37	.387
69	.03396	.00141	.08652	.00445	13.01	9.23	.383
70	.03059	.00132	.08532	.00476	12.38	9.06	.373
71	.02776	.00120	.08465	.00503	11.77	8.86	.360
72	.02537	.00107	.08453	.00524	11.17	8.64	.343
73	.02335	.00094	.08497	.00541	10.59	8.39	.324
74	.02162	.00081	.08598	.00552	10.02	8.12	.304
75	.02013	.00070	.08759	.00560	9.48	7.84	.282
76	.01883	.00060	.08982	.00563	8.96	7.54	.260
77	.01769	.00053	.09268	.00563	8.45	7.24	.237
78	.01666	.00051	.09623	.00560	7.96	6.93	.215
79	.01573	.00052	.10048	.00554	7.49	6.61	.194
80	.01485	.00056	.10547	.00545	7.04	6.29	.174
81	.01401	.00062	.11125	.00534	6.61	5.98	.155
82	.01319	.00068	.11786	.00521	6.20	5.67	.137
83	.01238	.00074	.12535	.00507	5.80	5.36	.121
84	.01155	.00078	.13377	.00491	5.43	5.05	.106
85	.01071	.00081	.14317	.00474	5.07	4.76	.092
86	.00985	.00083	.15362	.00456	4.73	4.47	.080
87	.00896	.00082	.16516	.00438	4.41	4.19	.069
88	.00806	.00080	.17786	.00418	4.10	3.92	.059
89	.00714	.00076	.19178	.00398	3.81	3.67	.050
90	.00623	.00071	.20699	.00378	3.54	3.42	.043

A	P	SE(P)	Q	SE(Q)	E(N)	E(BC)	SE(E(BC)
60	.19224	.01278	.19224	.01278	19.42	4.41	.211
61	.19589	.00994	.24250	.01169	18.65	4.35	.251
62	.14103	.00424	.23049	.00915	17.89	4.59	.323
63	.10324	.00278	.21926	.00815	17.15	4.82	.403
64	.07676	.00279	.20880	.00869	16.42	5.05	.484
65	.05791	.00272	.19911	.01017	15.71	5.25	.563
66	.04430	.00247	.19017	.01199	15.01	5.44	.634
67	.03433	.00212	.18200	.01383	14.33	5.60	.696
68	.02694	.00177	.17458	.01554	13.66	5.74	.748
69	.02139	.00144	.16790	.01708	13.01	5.86	.788
70	.01717	.00117	.16198	.01842	12.38	5.94	.815
71	.01393	.00096	.15682	.01956	11.77	6.00	.830
72	.01141	.00080	.15241	.02049	11.17	6.02	.833
73	.00944	.00069	.14876	.02124	10.59	6.02	.826
74	.00788	.00063	.14588	.02181	10.02	5.99	.808
75	.00664	.00060	.14379	.02222	9.48	5.93	.782
76	.00563	.00059	.14251	.02247	8.96	5.84	.748
77	.00481	.00059	.14204	.02258	8.45	5.73	.709
78	.00414	.00060	.14242	.02257	7.96	5.60	.665
79	.00358	.00060	.14367	.02244	7.49	5.45	.619
80	.00311	.00061	.14582	.02221	7.04	5.28	.570
81	.00272	.00061	.14890	.02189	6.61	5.10	.522
82	.00237	.00060	.15296	.02148	6.20	4.91	.474
83	.00208	.00059	.15803	.02100	5.80	4.70	.428
84	.00182	.00057	.16416	.02046	5.43	4.50	.383
85	.00159	.00055	.17139	.01986	5.07	4.28	.341
86	.00138	.00052	.17978	.01921	4.73	4.07	.301
87	.00119	.00049	.18939	.01851	4.41	3.85	.264
88	.00102	.00045	.20026	.01779	4.10	3.64	.231
89	.00087	.00041	.21245	.01703	3.81	3.43	.200
90	.00073	.00036	.22601	.01625	3.54	3.22	.173

TREATMENT 3

A	P	SE(P)	Q	SE(Q)	E(N)	E(BC)	SE(E(BC)
60	.12112	.00525	.12112	.00525	19.42	7.83	.163
61	.13103	.00391	.14909	.00436	18.65	7.84	.179
62	.10462	.00211	.13989	.00339	17.89	8.13	.205
63	.08463	.00127	.13158	.00279	17.15	8.37	.232
64	.06933	.00107	.12411	.00259	16.42	8.57	.256
65	.05747	.00110	.11746	.00270	15.71	8.71	.276
66	.04819	.00113	.11160	.00299	15.01	8.81	.289
67	.04086	.00111	.10651	.00332	14.33	8.85	.297
68	.03502	.00105	.10217	.00364	13.66	8.85	.300
69	.03033	.00097	.09856	.00392	13.01	8.80	.297
70	.02654	.00088	.09568	.00415	12.38	8.71	.290
71	.02346	.00077	.09350	.00433	11.77	8.58	.280
72	.02093	.00067	.09204	.00446	11.17	8.41	.266
73	.01885	.00058	.09128	.00454	10.59	8.21	.250
74	.01712	.00050	.09122	.00457	10.02	7.99	.233
75	.01567	.00043	.09189	.00457	9.48	7.74	.215
76	.01444	.00038	.09328	.00454	8.96	7.48	.197
77	.01340	.00035	.09542	.00448	8.45	7.19	.178
78	.01249	.00035	.09832	.00439	7.96	6.90	.160
79	.01168	.00036	.10202	.00428	7.49	6.60	.143
80	.01095	.00039	.10653	.00416	7.04	6.29	.127
81	.01028	.00042	.11189	.00402	6.61	5.98	.112
82	.00964	.00046	.11815	.00387	6.20	5.68	.098
83	.00902	.00048	.12534	.00371	5.80	5.37	.085
84	.00840	.00050	.13350	.00355	5.43	5.07	.074
85	.00778	.00051	.14269	.00338	5.07	4.77	.064
86	.00715	.00051	.15297	.00320	4.73	4.49	.055
87	.00651	.00051	.16437	.00303	4.41	4.21	.046
88	.00586	.00049	.17697	.00285	4.10	3.94	.039
89	.00520	.00046	.19081	.00268	3.81	3.68	.033
90	.00454	.00042	.20596	.00251	3.54	3.43	.028

A	P	SE(P)	Q	SE(Q)	E(N)	E(BC)	SE(E(BC)
60	.03474	.01404	.03474	.01404	19.42	11.70	1.013
61	.07168	.01320	.07426	.01364	18.65	11.10	1.037
62	.06578	.00923	.07361	.01134	17.89	10.95	1.113
63	.06055	.00651	.07315	.00962	17.15	10.78	1.189
64	.05592	.00503	.07289	.00865	16.42	10.59	1.258
65	.05181	.00463	.07284	.00850	15.71	10.38	1.315
66	.04815	.00484	.07301	.00906	15.01	10.16	1.357
67	.04489	.00521	.07342	.01010	14.33	9.92	1.384
68	.04197	.00549	.07409	.01139	13.66	9.67	1.397
69	.03936	.00564	.07504	.01279	13.01	9.40	1.395
70	.03701	.00562	.07628	.01421	12.38	9.13	1.380
71	.03488	.00546	.07784	.01560	11.77	8.84	1.354
72	.03295	.00518	.07974	.01692	11.17	8.54	1.318
73	.03119	.00480	.08200	.01816	10.59	8.24	1.273
74	.02956	.00434	.08467	.01932	10.02	7.93	1.220
75	.02804	.00385	.08776	.02038	9.48	7.62	1.162
76	.02662	.00334	.09132	.02134	8.96	7.31	1.100
77	.02526	.00287	.09537	.02220	8.45	6.99	1.035
78	.02395	.00248	.09996	.02297	7.96	6.68	.968
79	.02268	.00226	.10514	.02364	7.49	6.36	.900
80	.02141	.00223	.11094	.02421	7.04	6.05	.832
81	.02015	.00239	.11742	.02468	6.61	5.75	.765
82	.01887	.00267	.12463	.02507	6.20	5.44	.700
83	.01758	.00298	.13262	.02536	5.80	5.15	.638
84	.01627	.00329	.14145	.02556	5.43	4.86	.578
85	.01493	.00355	.15119	.02567	5.07	4.58	.522
86	.01357	.00373	.16188	.02569	4.73	4.31	.469
87	.01219	.00382	.17360	.02562	4.41	4.04	.419
88	.01082	.00381	.18642	.02547	4.10	3.79	.373
89	.00946	.00371	.20039	.02523	3.81	3.55	.330
90	.00814	.00352	.21559	.02491	3.54	3.31	.291

A	P	SE(P)	Q	SE(Q)	E(N)	E(BC)	SE(E(BC
60	.15533	.05406	.15533	.05406	19.42	5.67	.965
61	.13329	.03510	.15780	.04031	18.65	5.62	1.083
62	.11244	.02004	.15806	.03364	17.89	5.58	1.230
63	.09490	.01198	.15844	.02911	17.15	5.54	1.412
64	.08012	.00986	.15895	.02764	16.42	5.49	1.607
65	.06766	.01040	.15961	.02955	15.71	5.43	1.800
66	.05715	.01095	.16042	.03415	15.01	5.37	1.986
67	.04828	.01087	.16141	.04042	14.33	5.31	2.158
68	.04078	.01023	.16259	.04763	13.66	5.24	2.314
69	.03444	.00923	.16397	.05531	13.01	5.16	2.451
70	.02908	.00805	.16559	.06322	12.38	5.07	2.568
71	.02454	.00686	.16745	.07122	11.77	4.98	2.666
72	.02069	.00577	.16958	.07921	11.17	4.89	2.742
73	.01743	.00488	.17200	.08712	10.59	4.79	2.797
74	.01466	.00424	.17475	.09493	10.02	4.68	2.832
75	.01231	.00387	.17786	.10259	9.48	4.57	2.847
76	.01032	.00373	.18134	.11007	8.96	4.45	2.843
77	.00863	.00371	.18525	.11735	8.45	4.33	2.820
78	.00720	.00375	.18961	.12441	7.96	4.21	2.791
79	.00598	.00379	.19447	.13122	7.49	4.08	2.736
80	.00495	.00378	.19986	.13777	7.04	3.95	2.668
81	.00408	.00371	.20584	.14402	6.61	3.81	2.586
82	.00335	.00359	.21244	.14995	6.20	3.67	2.494
83	.00272	.00341	.21973	.15553	5.80	3.53	2.392
84	.00220	.00318	.22775	.16075	5.43	3.39	2.282
85	.00177	.00292	.23655	.16556	5.07	3.25	2.166
86	.00140	.00263	.24621	.16994	4.73	3.11	2.045
87	.00110	.00232	.25677	.17386	4.41	2.97	1.922
88	.00086	.00201	.26829	.17728	4.10	2.82	1.797
89	.00066	.00171	.28084	.18017	3.81	2.68	1.672
90	.00050	.00142	.29447	.18249	3.54	2.55	1.548

HISTOLOGY 1 X TREATMENT 3

A	P	SE(P)	Q	SE(Q)	E(N)	E(BC)	SE(E(BC))
60	.07841	.02110	.07841	.02110	19.42	9.73	.832
61	.09368	.01515	.10165	.01627	18.65	9.54	.876
62	.08171	.00969	.09869	.01323	17.89	9.57	.953
63	.07168	.00628	.09606	.01101	17.15	9.56	1.037
64	.06323	.00469	.09374	.00977	16.42	9.52	1.115
65	.05609	.00441	.09177	.00955	15.71	9.46	1.181
66	.05004	.00465	.09013	.01015	15.01	9.36	1.232
67	.04488	.00491	.08884	.01124	14.33	9.24	1.266
68	.04047	.00502	.08792	.01256	13.66	9.10	1.283
69	.03668	.00498	.08737	.01392	13.01	8.93	1.285
70	.03341	.00480	.08721	.01524	12.38	8.73	1.271
71	.03059	.00451	.08746	.01647	11.77	8.52	1.245
72	.02813	.00415	.08814	.01758	11.17	8.29	1.208
73	.02598	.00374	.08928	.01856	10.59	8.04	1.161
74	.02409	.00331	.09088	.01940	10.02	7.78	1.106
75	.02241	.00289	.09300	.02012	9.48	7.51	1.046
76	.02090	.00251	.09564	.02071	8.96	7.23	.982
77	.01954	.00220	.09886	.02118	8.45	6.95	.915
78	.01829	.00200	.10267	.02154	7.96	6.65	.848
79	.01712	.00192	.10714	.02178	7.49	6.36	.780
80	.01602	.00197	.11228	.02192	7.04	6.06	.713
81	.01497	.00211	.11817	.02196	6.61	5.77	.648
82	.01394	.00229	.12483	.02192	6.20	5.47	.585
83	.01294	.00248	.13233	.02178	5.80	5.18	.526
84	.01194	.00265	.14071	.02157	5.43	4.90	.470
85	.01094	.00277	.15005	.02128	5.07	4.62	.419
86	.00994	.00284	.16039	.02092	4.73	4.35	.371
87	.00894	.00284	.17181	.02050	4.41	4.08	.326
88	.00794	.00279	.18435	.02002	4.10	3.83	.286
89	.00696	.00267	.19810	.01948	3.81	3.58	.250
90	.00600	.00250	.21310	.01890	3.54	3.35	.217

A	P	SE(P)	Q	SE(Q)	E(N)	E(BC)	SE(E(BC)
60	.05932	.01248	.05932	.01248	19.42	9.46	.578
61	.09371	.01017	.09962	.01073	18.65	9.03	.602
62	.08343	.00676	.09850	.00902	17.89	8.97	.652
63	.07450	.00452	.09758	.00776	17.15	8.90	.707
64	.06673	.00338	.09685	.00705	16.42	8.81	.761
65	.05995	.00311	.09633	.00696	15.71	8.70	.809
66	.05401	.00325	.09604	.00742	15.01	8.58	.850
67	.04880	.00345	.09599	.00825	14.33	8.44	.883
68	.04421	.00355	.09620	.00930	13.66	8.28	.905
69	.04016	.00355	.09668	.01044	13.01	8.11	.919
70	.03657	.00344	.09746	.01162	12.38	7.93	.924
71	.03337	.00324	.09856	.01278	11.77	7.73	.920
72	.03052	.00299	.10000	.01390	11.17	7.52	.909
73	.02797	.00270	.10180	.01497	10.59	7.30	.890
74	.02566	.00239	.10401	.01597	10.02	7.07	.865
75	.02358	.00209	.10664	.01691	9.48	6.84	.835
76	.02167	.00182	.10974	.01778	8.96	6.59	.801
77	.01993	.00161	.11333	.01857	8.45	6.35	.763
78	.01831	.00148	.11746	.01929	7.96	6.09	.722
79	.01681	.00144	.12217	.01993	7.49	5.84	.680
80	.01540	.00148	.12751	.02049	7.04	5.58	.636
81	.01407	.00158	.13352	.02098	6.61	5.33	.591
82	.01281	.00170	.14026	.02140	6.20	5.07	.547
83	.01160	.00183	.14778	.02174	5.80	4.82	.503
84	.01045	.00193	.15614	.02200	5.43	4.57	.462
85	.00934	.00199	.16540	.02219	5.07	4.32	.421
86	.00827	.00202	.17561	.02230	4.73	4.08	.381
87	.00726	.00200	.18685	.02234	4.41	3.85	.344
88	.00629	.00195	.19918	.02230	4.10	3.62	.309
89	.00538	.00184	.21267	.02219	3.81	3.40	.276
90	.00453	.00171	.22737	.02200	3.54	3.19	.246

HISTOLOGY 2 X TREATMENT 2

A	P	SE(P)	Q	SE(Q)	E(N)	E(BC)	SE(E(BC)
60	.19797	.04416	.19797	.04416	19.42	4.44	.790
61	.24090	.03933	.30037	.04616	18.65	4.42	.947
62	.14954	.01346	.26650	.03403	17.89	5.13	1.295
63	.09729	.00984	.23637	.03000	17.15	5.82	1.669
64	.06594	.00958	.20980	.03208	16.42	6.43	2.009
65	.04634	.00859	.18658	.03630	15.71	7.07	2.281
66	.03363	.00728	.16646	.04028	15.01	7.58	2.465
67	.02512	.00599	.14920	.04320	14.33	8.00	2.558
68	.01928	.00487	.13455	.04490	13.66	8.32	2.565
69	.01516	.00393	.12229	.04549	13.01	8.54	2.498
70	.01221	.00317	.11220	.04513	12.38	8.67	2.373
71	.01006	.00258	.10407	.04403	11.77	8.70	2.206
72	.00846	.00212	.09774	.04235	11.17	8.66	2.013
73	.00727	.00180	.09305	.04026	10.59	8.54	1.807
74	.00637	.00161	.08986	.03790	10.02	8.37	1.599
75	.00568	.00152	.08806	.03537	9.48	8.15	1.396
76	.00515	.00153	.08756	.03277	8.96	7.88	1.205
77	.00474	.00159	.08828	.03016	8.45	7.59	1.030
78	.00441	.00169	.09017	.02760	7.96	7.28	.872
79	.00415	.00180	.09317	.02512	7.49	6.95	.732
80	.00392	.00190	.09726	.02276	7.04	6.62	.609
81	.00373	.00199	.10242	.02053	6.61	6.28	.504
82	.00355	.00206	.10865	.01844	6.20	5.94	.414
83	.00338	.00211	.11595	.01650	5.80	5.60	.339
84	.00320	.00212	.12434	.01471	5.43	5.27	.275
85	.00302	.00210	.13384	.01307	5.07	4.95	.222
86	.00282	.00205	.14448	.01158	4.73	4.63	.178
87	.00261	.00196	.15631	.01022	4.41	4.33	.142
88	.00239	.00185	.16935	.00899	4.10	4.04	.113
89	.00215	.00170	.18367	.00788	3.81	3.77	.089
90	.00191	.00154	.19929	.00689	3.54	3.51	.070

A	P	SE(P)	Q	SE(Q)	E(N)	F(BC)	SE(F(BC
60	.09321	.01497	.09321	.01497	19.42	9.22	.542
61	.10732	.01091	.11835	.01187	18.65	9.12	.578
62	.08998	.00653	.11256	.00944	17.89	9.28	.635
63	.07615	.00409	.10734	.00777	17.15	9.39	.696
64	.06503	.00315	.10269	.00697	16.42	9.46	.751
65	.05603	.00308	.09859	.00695	15.71	9.49	.794
66	.04869	.00322	.09505	.00745	15.01	9.47	.825
67	.04267	.00330	.09206	.00821	14.33	9.42	.841
68	.03771	.00327	.08960	.00902	13.66	9.32	.845
69	.03360	.00314	.08769	.00980	13.01	9.19	.836
70	.03018	.00294	.08633	.01049	12.38	9.03	.816
71	.02731	.00270	.08551	.01107	11.77	8.83	.787
72	.02490	.00243	.08526	.01154	11.17	8.61	.752
73	.02287	.00215	.08558	.01190	10.59	8.37	.710
74	.02113	.00188	.08649	.01215	10.02	8.11	.665
75	.01964	.00163	.08800	.01231	9.48	7.85	.617
76	.01835	.00142	.09014	.01238	8.96	7.54	.568
77	.01721	.00127	.09294	.01237	8.45	7.23	.519
78	.01620	.00120	.09642	.01229	7.96	6.93	.471
79	.01527	.00121	.10061	.01215	7.49	6.61	.425
80	.01441	.00127	.10555	.01195	7.04	6.30	.380
81	.01359	.00138	.11129	.01170	6.61	5.98	.338
82	.01279	.00149	.11787	.01141	6.20	5.67	.299
83	.01200	.00159	.12532	.01108	5.80	5.36	.263
84	.01120	.00168	.13372	.01073	5.43	5.06	.230
85	.01038	.00174	.14310	.01035	5.07	4.76	.200
86	.00954	.00176	.15352	.00995	4.73	4.47	.174
87	.00868	.00175	.16505	.00953	4.41	4.19	.150
88	.00781	.00170	.17774	.00909	4.10	3.92	.128
89	.00692	.00162	.19165	.00865	3.81	3.67	.109
90	.00604	.00150	.20685	.00820	3.54	3.42	.093

A	P	SE(P)	Q	SE(Q)	E(N)	E(BC)	SE(E(BC))
60	.10756	.01144	.10756	.01144	19.42	8.24	.488
61	.10767	.00842	.12064	.00930	18.65	8.18	.539
62	.09221	.00504	.11750	.00743	17.89	8.23	.618
63	.07941	.00332	.11466	.00635	17.15	8.27	.694
64	.06875	.00292	.11212	.00621	16.42	8.27	.762
65	.05983	.00309	.10989	.00686	15.71	8.26	.821
66	.05233	.00328	.10799	.00796	15.01	8.22	.869
67	.04600	.00334	.10641	.00924	14.33	8.15	.904
68	.04063	.00327	.10518	.01055	13.66	8.06	.928
69	.03605	.00309	.10432	.01182	13.01	7.95	.940
70	.03214	.00284	.10383	.01301	12.38	7.82	.941
71	.02878	.00254	.10373	.01411	11.77	7.67	.932
72	.02587	.00222	.10406	.01510	11.17	7.50	.915
73	.02335	.00189	.10483	.01599	10.59	7.32	.890
74	.02115	.00159	.10606	.01678	10.02	7.12	.858
75	.01922	.00132	.10779	.01746	9.48	6.90	.820
76	.01750	.00112	.11005	.01804	8.96	6.68	.779
77	.01598	.00101	.11287	.01853	8.45	6.44	.734
78	.01460	.00101	.11629	.01893	7.96	6.20	.688
79	.01336	.00109	.12034	.01923	7.49	5.95	.639
80	.01221	.00121	.12509	.01946	7.04	5.70	.591
81	.01115	.00135	.13055	.01960	6.61	5.44	.543
82	.01016	.00148	.13680	.01966	6.20	5.19	.496
83	.00922	.00160	.14388	.01965	5.80	4.93	.450
84	.00833	.00168	.15185	.01957	5.43	4.67	.408
85	.00748	.00172	.16077	.01943	5.07	4.42	.366
86	.00667	.00174	.17069	.01922	4.73	4.18	.327
87	.00589	.00171	.18168	.01895	4.41	3.94	.291
88	.00514	.00165	.19380	.01862	4.10	3.70	.258
89	.00443	.00155	.20712	.01823	3.81	3.47	.227
90	.00376	.00143	.22170	.01780	3.54	3.25	.199

A	P	SE(P)	Q	SE(Q)	E(N)	E(BC)	SE(E(BC)
60	.20006	.03005	.20006	.03005	19.42	3.92	.459
61	.20173	.02419	.25219	.02872	18.65	3.79	.554
62	.14657	.01012	.24501	.02221	17.89	3.90	.748
63	.10758	.00801	.23820	.02161	17.15	4.01	.954
64	.07974	.00835	.23176	.02635	16.42	4.12	1.167
65	.05966	.00784	.22570	.03358	15.71	4.22	1.376
66	.04503	.00672	.22003	.04152	15.01	4.31	1.578
67	.03428	.00542	.21477	.04944	14.33	4.40	1.766
68	.02631	.00422	.20992	.05705	13.66	4.47	1.939
69	.02035	.00325	.20550	.06423	13.01	4.53	2.092
70	.01586	.00256	.20152	.07090	12.38	4.57	2.222
71	.01244	.00215	.19802	.07706	11.77	4.61	2.329
72	.00983	.00196	.19500	.08268	11.17	4.63	2.411
73	.00781	.00190	.19249	.08778	10.59	4.63	2.467
74	.00624	.00190	.19052	.09236	10.02	4.62	2.497
75	.00501	.00190	.18910	.09643	9.48	4.59	2.514
76	.00405	.00189	.18828	.10000	8.96	4.55	2.496
77	.00328	.00185	.18808	.10308	8.45	4.50	2.455
78	.00267	.00179	.18855	.10569	7.96	4.43	2.394
79	.00218	.00172	.18971	.10785	7.49	4.34	2.316
80	.00179	.00163	.19161	.10955	7.04	4.25	2.221
81	.00146	.00153	.19430	.11082	6.61	4.14	2.115
82	.00120	.00142	.19782	.11167	6.20	4.02	1.998
83	.00098	.00131	.20223	.11212	5.80	3.89	1.874
84	.00081	.00120	.20757	.11216	5.43	3.76	1.745
85	.00066	.00108	.21392	.11183	5.07	3.62	1.614
86	.00054	.00096	.22131	.11112	4.73	3.47	1.483
87	.00043	.00085	.22983	.11005	4.41	3.32	1.354
88	.00035	.00074	.23952	.10862	4.10	3.16	1.228
89	.00028	.00064	.25045	.10687	3.81	3.01	1.108
90	.00022	.00054	.26269	.10478	3.54	2.85	.993

A	P	SE(P)	Q	SE(Q)	E(N)	E(BC)	SE(E(BC)
60	.10490	.00924	.10490	.00924	19.42	7.22	.313
61	.12851	.00723	.14357	.00794	18.65	7.01	.343
62	.10705	.00408	.13965	.00637	17.89	7.10	.399
63	.08972	.00255	.13604	.00545	17.15	7.17	.456
64	.07565	.00221	.13276	.00528	16.42	7.23	.511
65	.06414	.00234	.12980	.00579	15.71	7.26	.560
66	.05469	.00245	.12719	.00668	15.01	7.27	.602
67	.04689	.00244	.12492	.00775	14.33	7.26	.636
68	.04040	.00232	.12302	.00886	13.66	7.22	.662
69	.03500	.00214	.12149	.00994	13.01	7.17	.680
70	.03046	.00192	.12036	.01096	12.38	7.09	.689
71	.02663	.00167	.11965	.01191	11.77	6.99	.691
72	.02339	.00143	.11937	.01277	11.17	6.88	.685
73	.02063	.00121	.11954	.01355	10.59	6.74	.673
74	.01826	.00102	.12020	.01424	10.02	6.59	.655
75	.01622	.00087	.12136	.01485	9.48	6.42	.633
76	.01446	.00077	.12307	.01537	8.96	6.24	.606
77	.01291	.00074	.12536	.01581	8.45	6.05	.576
78	.01155	.00076	.12826	.01617	7.96	5.85	.544
79	.01035	.00081	.13181	.01646	7.49	5.64	.510
80	.00928	.00087	.13605	.01668	7.04	5.42	.474
81	.00831	.00094	.14104	.01683	6.61	5.19	.439
82	.00743	.00099	.14681	.01691	6.20	4.97	.403
83	.00662	.00103	.15343	.01693	5.80	4.74	.368
84	.00588	.00106	.16095	.01688	5.43	4.51	.336
85	.00520	.00106	.16942	.01678	5.07	4.28	.303
86	.00456	.00104	.17892	.01662	4.73	4.05	.272
87	.00396	.00101	.18949	.01641	4.41	3.83	.243
88	.00341	.00096	.20120	.01615	4.10	3.61	.217
89	.00290	.00089	.21412	.01583	3.81	3.39	.192
90	.00243	.00081	.22830	.01547	3.54	3.18	.169

A	P	SE(P)	Q	SE(Q)	E(N)	E(BC)	SE(E(BC))
60	.03756	.01664	.03756	.01664	19.42	11.99	1.274
61	.06657	.00799	.06917	.00822	18.65	11.43	1.313
62	.06173	.00550	.06891	.00659	17.89	11.24	1.435
63	.05740	.00455	.06881	.00614	17.15	11.04	1.537
64	.05351	.00479	.06890	.00689	16.42	10.82	1.617
65	.05003	.00544	.06918	.00838	15.71	10.58	1.676
66	.04690	.00605	.06966	.01018	15.01	10.33	1.715
67	.04408	.00646	.07038	.01207	14.33	10.07	1.734
68	.04153	.00666	.07133	.01394	13.66	9.79	1.736
69	.03923	.00664	.07255	.01575	13.01	9.51	1.722
70	.03713	.00643	.07405	.01747	12.38	9.21	1.693
71	.03522	.00606	.07585	.01909	11.77	8.91	1.652
72	.03346	.00556	.07798	.02061	11.17	8.60	1.600
73	.03184	.00495	.08047	.02202	10.59	8.28	1.539
74	.03032	.00426	.08334	.02333	10.02	7.97	1.471
75	.02889	.00352	.08663	.02452	9.48	7.64	1.398
76	.02753	.00277	.09037	.02561	8.96	7.32	1.320
77	.02621	.00209	.09460	.02659	8.45	7.00	1.240
78	.02492	.00162	.09936	.02747	7.96	6.68	1.158
79	.02365	.00157	.10469	.02824	7.49	6.36	1.076
80	.02238	.00195	.11063	.02891	7.04	6.05	.995
81	.02109	.00253	.11725	.02948	6.61	5.74	.915
82	.01978	.00315	.12458	.02994	6.20	5.44	.838
83	.01845	.00372	.13268	.03031	5.80	5.14	.763
84	.01708	.00422	.14162	.03057	5.43	4.85	.692
85	.01568	.00461	.15145	.03074	5.07	4.57	.626
86	.01425	.00487	.16224	.03081	4.73	4.30	.563
87	.01281	.00500	.17404	.03077	4.41	4.04	.504
88	.01136	.00500	.18693	.03065	4.10	3.78	.449
89	.00993	.00486	.20096	.03042	3.81	3.54	.399
90	.00854	.00460	.21621	.03009	3.54	3.31	.352

A	P	SE(P)	Q	SE(Q)	E(N)	F(BC)	SE(E(BC)
60	.07185	.02553	.07185	.02553	19.42	6.53	.922
61	.14422	.02667	.15538	.02841	18.65	6.00	.983
62	.12059	.01452	.15383	.02299	17.89	6.01	1.160
63	.10112	.00877	.15244	.01986	17.15	6.02	1.351
64	.08503	.00803	.15124	.01973	16.42	6.01	1.541
65	.07169	.00882	.15023	.02234	15.71	6.00	1.720
66	.06059	.00926	.14942	.02665	15.01	5.97	1.886
67	.05133	.00910	.14883	.03179	14.33	5.93	2.032
68	.04359	.00849	.14849	.03726	13.66	5.89	2.159
69	.03710	.00760	.14840	.04279	13.01	5.83	2.265
70	.03163	.00658	.14858	.04825	12.38	5.76	2.348
71	.02702	.00554	.14907	.05357	11.77	5.68	2.410
72	.02312	.00459	.14987	.05871	11.17	5.59	2.450
73	.01980	.00380	.15103	.06362	10.59	5.49	2.468
74	.01698	.00322	.15256	.06831	10.02	5.37	2.466
75	.01457	.00291	.15449	.07275	9.48	5.25	2.446
76	.01251	.00283	.15687	.07693	8.96	5.12	2.407
77	.01074	.00291	.15973	.08085	8.45	4.99	2.353
78	.00922	.00307	.16309	.08450	7.96	4.84	2.284
79	.00790	.00323	.16702	.08787	7.49	4.69	2.203
80	.00676	.00337	.17154	.09097	7.04	4.53	2.111
81	.00577	.00345	.17672	.09377	6.61	4.37	2.018
82	.00491	.00347	.18259	.09628	6.20	4.21	1.912
83	.00416	.00343	.18922	.09849	5.80	4.04	1.801
84	.00350	.00333	.19665	.10039	5.43	3.87	1.687
85	.00293	.00318	.20495	.10197	5.07	3.69	1.571
86	.00244	.00298	.21418	.10324	4.73	3.52	1.455
87	.00201	.00275	.22440	.10417	4.41	3.35	1.341
88	.00163	.00249	.23568	.10477	4.10	3.18	1.230
89	.00131	.00220	.24808	.10502	3.81	3.01	1.122
90	.00104	.00191	.26165	.10492	3.54	2.84	1.018

A	P	SE(P)	Q	SE(Q)	E(N)	F(BC)	SE(E(BC)
60	.05778	.00857	.05778	.00857	19.42	11.17	.485
61	.06634	.00581	.07040	.00613	18.65	10.83	.506
62	.06215	.00416	.07096	.00516	17.89	10.61	.548
63	.05830	.00298	.07165	.00442	17.15	10.39	.590
64	.05476	.00233	.07249	.00400	16.42	10.15	.628
65	.05149	.00217	.07349	.00397	15.71	9.90	.661
66	.04847	.00231	.07467	.00431	15.01	9.65	.688
67	.04568	.00253	.07604	.00491	14.33	9.39	.708
68	.04308	.00271	.07762	.00567	13.66	9.12	.722
69	.04066	.00283	.07942	.00652	13.01	8.85	.730
70	.03840	.00286	.08148	.00741	12.38	8.57	.731
71	.03628	.00280	.08382	.00832	11.77	8.28	.727
72	.03429	.00268	.08645	.00922	11.17	8.00	.717
73	.03239	.00250	.08941	.01011	10.59	7.71	.703
74	.03059	.00227	.09272	.01097	10.02	7.41	.685
75	.02886	.00201	.09642	.01181	9.48	7.12	.663
76	.02719	.00173	.10054	.01262	8.96	6.83	.638
77	.02557	.00147	.10512	.01340	8.45	6.54	.611
78	.02399	.00124	.11020	.01415	7.96	6.25	.581
79	.02244	.00110	.11582	.01485	7.49	5.96	.550
80	.02090	.00107	.12203	.01552	7.04	5.67	.518
81	.01938	.00115	.12887	.01614	6.61	5.39	.486
82	.01787	.00131	.13640	.01672	6.20	5.12	.453
83	.01637	.00149	.14467	.01726	5.80	4.85	.421
84	.01488	.00166	.15374	.01774	5.43	4.58	.389
85	.01340	.00180	.16368	.01818	5.07	4.33	.358
86	.01195	.00189	.17453	.01856	4.73	4.08	.328
87	.01054	.00194	.18636	.01888	4.41	3.84	.300
88	.00916	.00194	.19925	.01914	4.10	3.61	.272
89	.00785	.00188	.21325	.01934	3.81	3.38	.246
90	.00662	.00178	.22843	.01948	3.54	3.17	.222

A	P	SE(P)	Q	SE(Q)	E(N)	E(BC)	SE(E(BC
60	.17114	.02733	.17114	.02733	19.42	5.01	.615
61	.20820	.02469	.25119	.02861	18.65	4.94	.725
62	.14248	.00958	.22957	.02127	17.89	5.44	.954
63	.10040	.00659	.20997	.01861	17.15	5.93	1.195
64	.07266	.00674	.19233	.01997	16.42	6.38	1.420
65	.05387	.00649	.17654	.02314	15.71	6.78	1.611
66	.04083	.00582	.16252	.02652	15.01	7.13	1.758
67	.03160	.00500	.15017	.02947	14.33	7.42	1.857
68	.02492	.00419	.13939	.03177	13.66	7.65	1.907
69	.02002	.00346	.13010	.03340	13.01	7.81	1.912
70	.01636	.00282	.12221	.03441	12.38	7.91	1.876
71	.01359	.00230	.11566	.03487	11.77	7.94	1.808
72	.01147	.00188	.11038	.03485	11.17	7.92	1.713
73	.00983	.00157	.10630	.03444	10.59	7.84	1.598
74	.00854	.00136	.10338	.03370	10.02	7.71	1.472
75	.00752	.00126	.10157	.03270	9.48	7.55	1.339
76	.00671	.00124	.10085	.03150	8.96	7.34	1.205
77	.00605	.00129	.10118	.03015	8.45	7.11	1.073
78	.00552	.00137	.10255	.02869	7.96	6.86	.946
79	.00507	.00146	.10495	.02716	7.49	6.58	.828
80	.00468	.00156	.10838	.02559	7.04	6.30	.718
81	.00435	.00164	.11285	.02400	6.61	6.00	.618
82	.00405	.00171	.11837	.02242	6.20	5.71	.528
83	.00377	.00176	.12496	.02087	5.80	5.40	.448
84	.00350	.00178	.13265	.01935	5.43	5.11	.379
85	.00324	.00177	.14147	.01787	5.07	4.81	.318
86	.00297	.00173	.15146	.01646	4.73	4.52	.266
87	.00271	.00167	.16266	.01510	4.41	4.24	.220
88	.00244	.00158	.17511	.01381	4.10	3.97	.182
89	.00217	.00147	.18887	.01259	3.81	3.71	.149
90	.00190	.00133	.20397	.01143	3.54	3.46	.122

STAGE 2 X TREATMENT 1

A	P	SE(P)	Q	SE(Q)	E(N)	E(BC)	SE(E(BC)
60	.07951	.01072	.07951	.01072	19.42	8.41	.438
61	.11968	.00916	.13002	.00984	18.65	8.14	.466
62	.09990	.00535	.12475	.00789	17.89	8.28	.525
63	.08408	.00333	.11996	.00663	17.15	8.39	.586
64	.07132	.00268	.11563	.00616	16.42	8.47	.642
65	.06097	.00272	.11176	.00640	15.71	8.51	.690
66	.05251	.00286	.10837	.00707	15.01	8.52	.727
67	.04556	.00290	.10545	.00795	14.33	8.50	.752
68	.03981	.00283	.10300	.00887	13.66	8.44	.767
69	.03503	.00267	.10104	.00975	13.01	8.36	.770
70	.03103	.00246	.09957	.01055	12.38	8.24	.764
71	.02767	.00222	.09860	.01126	11.77	8.10	.749
72	.02483	.00196	.09816	.01186	11.17	7.93	.726
73	.02241	.00171	.09825	.01237	10.59	7.74	.698
74	.02034	.00147	.09889	.01278	10.02	7.53	.664
75	.01856	.00126	.10012	.01309	9.48	7.30	.626
76	.01700	.00110	.10195	.01332	8.96	7.06	.586
77	.01564	.00100	.10441	.01347	8.45	6.80	.544
78	.01443	.00097	.10754	.01354	7.96	6.54	.501
79	.01333	.00100	.11137	.01354	7.49	6.26	.459
80	.01234	.00107	.11594	.01348	7.04	5.99	.417
81	.01141	.00116	.12130	.01336	6.61	5.71	.377
82	.01054	.00126	.12748	.01319	6.20	5.43	.338
83	.00970	.00134	.13454	.01297	5.80	5.15	.302
84	.00890	.00140	.14254	.01271	5.43	4.87	.268
85	.00811	.00144	.15152	.01241	5.07	4.60	.237
86	.00734	.00145	.16155	.01207	4.73	4.34	.208
87	.00657	.00143	.17268	.01171	4.41	4.08	.182
88	.00583	.00138	.18497	.01131	4.10	3.83	.158
89	.00510	.00130	.19849	.01090	3.81	3.58	.137
90	.00439	.00120	.21330	.01046	3.54	3.35	.118

A	P	SE(P)	Q	SE(Q)	E(N)	F(BC)	SE(E(BC
60	.07620	.00920	.07620	.00920	19.42	9.15	.390
61	.12727	.00840	.13777	.00899	18.65	8.87	.413
62	.10144	.00453	.12736	.00681	17.89	9.21	.468
63	.08209	.00277	.11810	.00553	17.15	9.48	.522
64	.06740	.00238	.10995	.00515	16.42	9.69	.565
65	.05611	.00246	.10284	.00541	15.71	9.82	.596
66	.04734	.00253	.09670	.00595	15.01	9.89	.612
67	.04046	.00250	.09150	.00654	14.33	9.90	.615
68	.03503	.00238	.08719	.00705	13.66	9.85	.605
69	.03070	.00220	.08374	.00745	13.01	9.74	.586
70	.02725	.00199	.08110	.00774	12.38	9.59	.558
71	.02447	.00176	.07925	.00791	11.77	9.39	.525
72	.02222	.00153	.07818	.00798	11.17	9.15	.487
73	.02040	.00131	.07786	.00796	10.59	8.89	.447
74	.01892	.00111	.07830	.00786	10.02	8.60	.407
75	.01770	.00095	.07948	.00770	9.48	8.29	.366
76	.01669	.00083	.08141	.00749	8.96	7.96	.327
77	.01584	.00078	.08410	.00724	8.45	7.62	.290
78	.01510	.00078	.08757	.00696	7.96	7.28	.255
79	.01445	.00084	.09183	.00666	7.49	6.93	.222
80	.01385	.00092	.09690	.00634	7.04	6.58	.193
81	.01327	.00101	.10282	.00600	6.61	6.23	.166
82	.01269	.00110	.10962	.00567	6.20	5.89	.142
83	.01210	.00118	.11734	.00533	5.80	5.55	.121
84	.01147	.00123	.12603	.00499	5.43	5.22	.103
85	.01080	.00126	.13572	.00466	5.07	4.90	.087
86	.01007	.00127	.14648	.00433	4.73	4.59	.073
87	.00929	.00125	.15835	.00402	4.41	4.30	.061
88	.00846	.00120	.17140	.00371	4.10	4.01	.051
89	.00760	.00113	.18567	.00342	3.81	3.74	.042
90	.00671	.00104	.20123	.00314	3.54	3.48	.034

STAGE 3 X TREATMENT 1

A	P	SE(P)	Q	SE(Q)	E(N)	E(BC)	SE(E(BC
60	.11091	.02467	.11091	.02467	19.42	7.73	.780
61	.13885	.02008	.15617	.02216	18.65	7.63	.849
62	.10945	.01060	.14588	.01722	17.89	7.96	.970
63	.08753	.00621	.13659	.01414	17.15	8.23	1.097
64	.07096	.00517	.12825	.01300	16.42	8.46	1.213
65	.05828	.00530	.12083	.01343	15.71	8.63	1.306
66	.04846	.00542	.11429	.01470	15.01	8.76	1.372
67	.04079	.00532	.10861	.01622	14.33	8.82	1.409
68	.03473	.00503	.10375	.01769	13.66	8.84	1.420
69	.02991	.00462	.09969	.01896	13.01	8.80	1.406
70	.02604	.00415	.09641	.01997	12.38	8.72	1.370
71	.02292	.00367	.09390	.02073	11.77	8.60	1.317
72	.02038	.00319	.09215	.02124	11.17	8.44	1.250
73	.01830	.00275	.09115	.02152	10.59	8.25	1.172
74	.01659	.00236	.09089	.02160	10.02	8.03	1.088
75	.01516	.00204	.09139	.02149	9.48	7.78	1.000
76	.01397	.00182	.09265	.02124	8.96	7.52	.911
77	.01295	.00170	.09468	.02085	8.45	7.23	.823
78	.01207	.00169	.09750	.02035	7.96	6.94	.738
79	.01130	.00176	.10113	.01976	7.49	6.64	.656
80	.01061	.00188	.10560	.01910	7.04	6.33	.579
81	.00997	.00202	.11093	.01837	6.61	6.02	.508
82	.00936	.00216	.11717	.01760	6.20	5.70	.443
83	.00877	.00227	.12436	.01680	5.80	5.39	.384
84	.00818	.00235	.13253	.01598	5.43	5.09	.331
85	.00759	.00240	.14174	.01514	5.07	4.79	.284
86	.00699	.00240	.15203	.01430	4.73	4.50	.242
87	.00637	.00235	.16346	.01346	4.41	4.22	.205
88	.00574	.00226	.17609	.01262	4.10	3.95	.173
89	.00510	.00212	.18997	.01180	3.81	3.69	.146
90	.00447	.00195	.20516	.01099	3.54	3.44	.122

STAGE 3 X TREATMENT 2

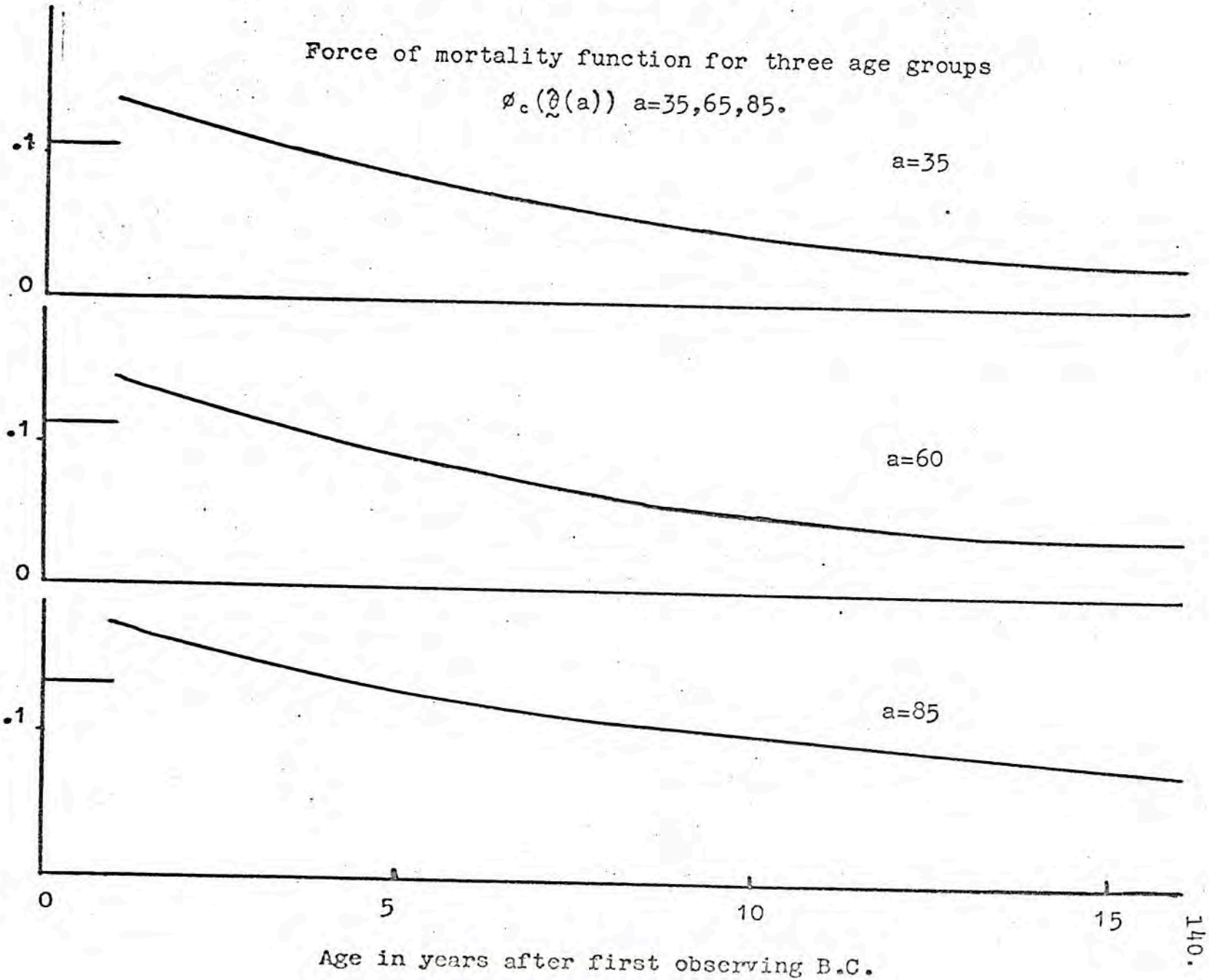
A	P	SE(P)	Q	SE(Q)	E(N)	E(BC)	SE(E(BC))
60	.14859	.02026	.14859	.02026	19.42	4.05	.269
61	.20546	.01783	.24132	.02014	18.65	3.67	.300
62	.15493	.00767	.23984	.01619	17.89	3.69	.373
63	.11711	.00522	.23850	.01496	17.15	3.70	.463
64	.08873	.00556	.23730	.01685	16.42	3.71	.558
65	.06738	.00550	.23626	.02088	15.71	3.72	.655
66	.05127	.00490	.23538	.02595	15.01	3.72	.749
67	.03909	.00408	.23469	.03147	14.33	3.72	.843
68	.02985	.00325	.23420	.03714	13.66	3.71	.930
69	.02283	.00257	.23392	.04284	13.01	3.70	1.011
70	.01749	.00209	.23388	.04849	12.38	3.68	1.085
71	.01341	.00180	.23409	.05404	11.77	3.66	1.152
72	.01029	.00166	.23458	.05946	11.17	3.64	1.210
73	.00790	.00158	.23537	.06475	10.59	3.60	1.259
74	.00607	.00153	.23648	.06987	10.02	3.57	1.300
75	.00467	.00147	.23795	.07482	9.48	3.52	1.331
76	.00358	.00139	.23981	.07959	8.96	3.47	1.353
77	.00275	.00129	.24209	.08416	8.45	3.42	1.365
78	.00211	.00119	.24483	.08852	7.96	3.36	1.368
79	.00161	.00107	.24805	.09267	7.49	3.29	1.362
80	.00123	.00095	.25182	.09659	7.04	3.22	1.347
81	.00094	.00083	.25617	.10027	6.61	3.14	1.325
82	.00071	.00072	.26114	.10369	6.20	3.06	1.294
83	.00054	.00062	.26679	.10684	5.80	2.98	1.257
84	.00040	.00052	.27317	.10972	5.43	2.89	1.214
85	.00030	.00043	.28033	.11229	5.07	2.79	1.165
86	.00022	.00036	.28833	.11455	4.73	2.70	1.112
87	.00016	.00029	.29724	.11648	4.41	2.60	1.056
88	.00012	.00023	.30710	.11807	4.10	2.50	.996
89	.00008	.00018	.31798	.11928	3.81	2.39	.939
90	.00006	.00014	.32994	.12012	3.54	2.29	.877

STAGE 3 x TREATMENT 3

A	P	SE(P)	Q	SE(Q)	E(N)	E(BC)	SE(E(BC))
60	.12933	.02044	.12933	.02044	19.42	5.52	.502
61	.19798	.01876	.22739	.02087	18.65	5.27	.561
62	.14172	.00799	.21068	.01607	17.89	5.68	.704
63	.10376	.00493	.19542	.01386	17.15	6.07	.856
64	.07757	.00488	.18157	.01412	16.42	6.43	1.001
65	.05911	.00487	.16906	.01582	15.71	6.75	1.127
66	.04586	.00455	.15784	.01798	15.01	7.02	1.229
67	.03618	.00405	.14787	.02007	14.33	7.25	1.302
68	.02900	.00350	.13909	.02187	13.66	7.42	1.346
69	.02359	.00296	.13145	.02330	13.01	7.55	1.362
70	.01947	.00248	.12491	.02437	12.38	7.61	1.352
71	.01629	.00206	.11944	.02509	11.77	7.63	1.319
72	.01382	.00172	.11500	.02549	11.17	7.60	1.268
73	.01186	.00145	.11157	.02561	10.59	7.52	1.202
74	.01031	.00126	.10911	.02549	10.02	7.41	1.125
75	.00906	.00114	.10763	.02516	9.48	7.26	1.041
76	.00804	.00110	.10709	.02465	8.96	7.07	.953
77	.00721	.00111	.10751	.02401	8.45	6.86	.863
78	.00652	.00115	.10888	.02324	7.96	6.63	.775
79	.00593	.00121	.11121	.02239	7.49	6.38	.690
80	.00543	.00128	.11451	.02146	7.04	6.11	.609
81	.00499	.00134	.11880	.02048	6.61	5.84	.534
82	.00459	.00139	.12411	.01947	6.20	5.56	.465
83	.00423	.00142	.13046	.01844	5.80	5.28	.402
84	.00388	.00143	.13788	.01739	5.43	5.00	.346
85	.00356	.00142	.14643	.01635	5.07	4.72	.296
86	.00324	.00140	.15613	.01532	4.73	4.44	.251
87	.00292	.00134	.16704	.01430	4.41	4.18	.212
88	.00261	.00127	.17920	.01331	4.10	3.91	.178
89	.00230	.00118	.19267	.01234	3.81	3.66	.149
90	.00200	.00107	.20749	.01140	3.54	3.42	.124

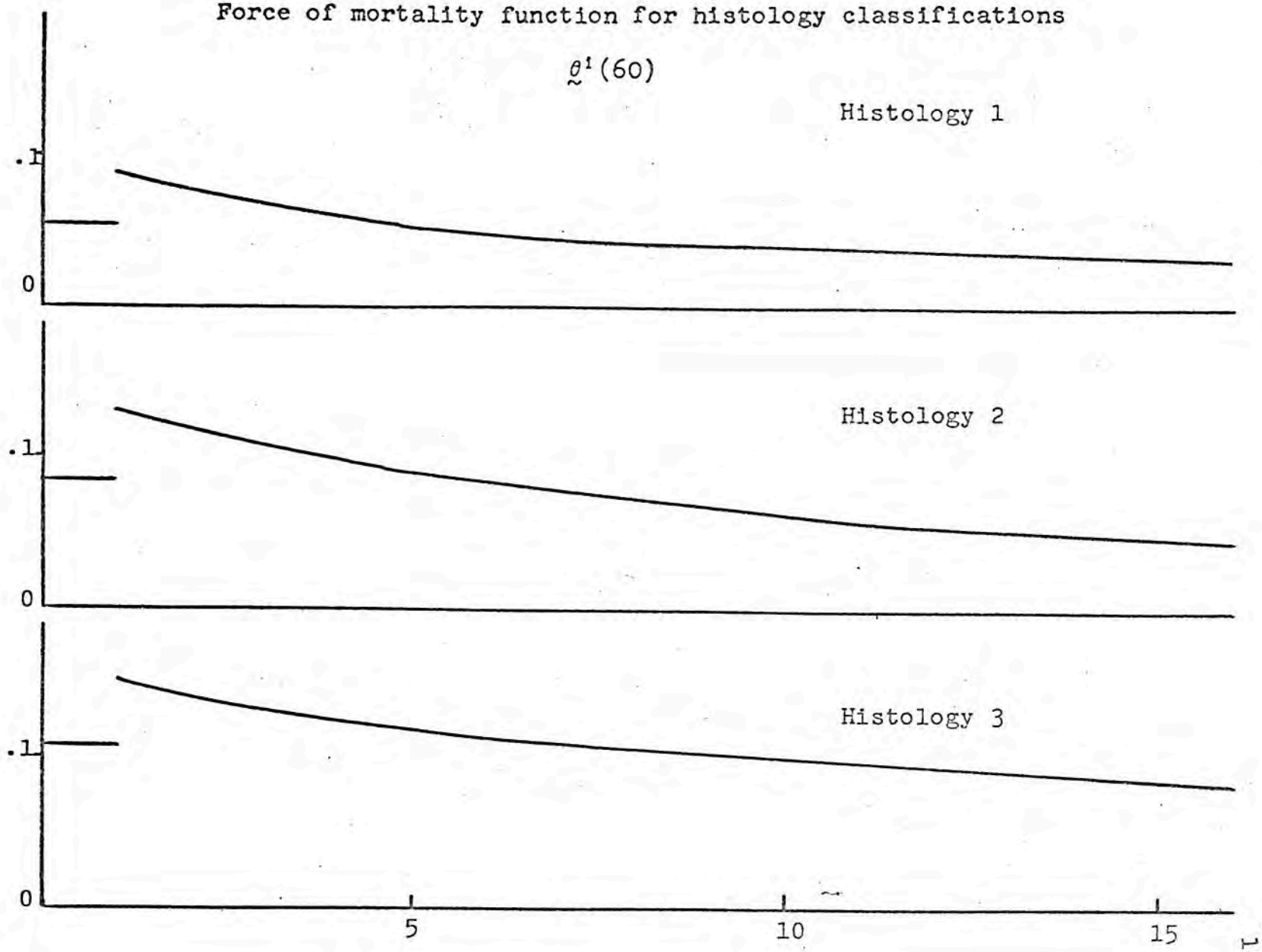
APPENDIX D

GRAPHS OF THE FORCE OF MORTALITY DUE TO BREAST CANCER


Force of mortality function for three age groups
$\phi_c(\underset{\sim}{\theta}(a))$ a=35,65,85.
.1
0
a=35
.1
0
a=60
.1
a=85
0
5
10
15
Age in years after first observing B.C.

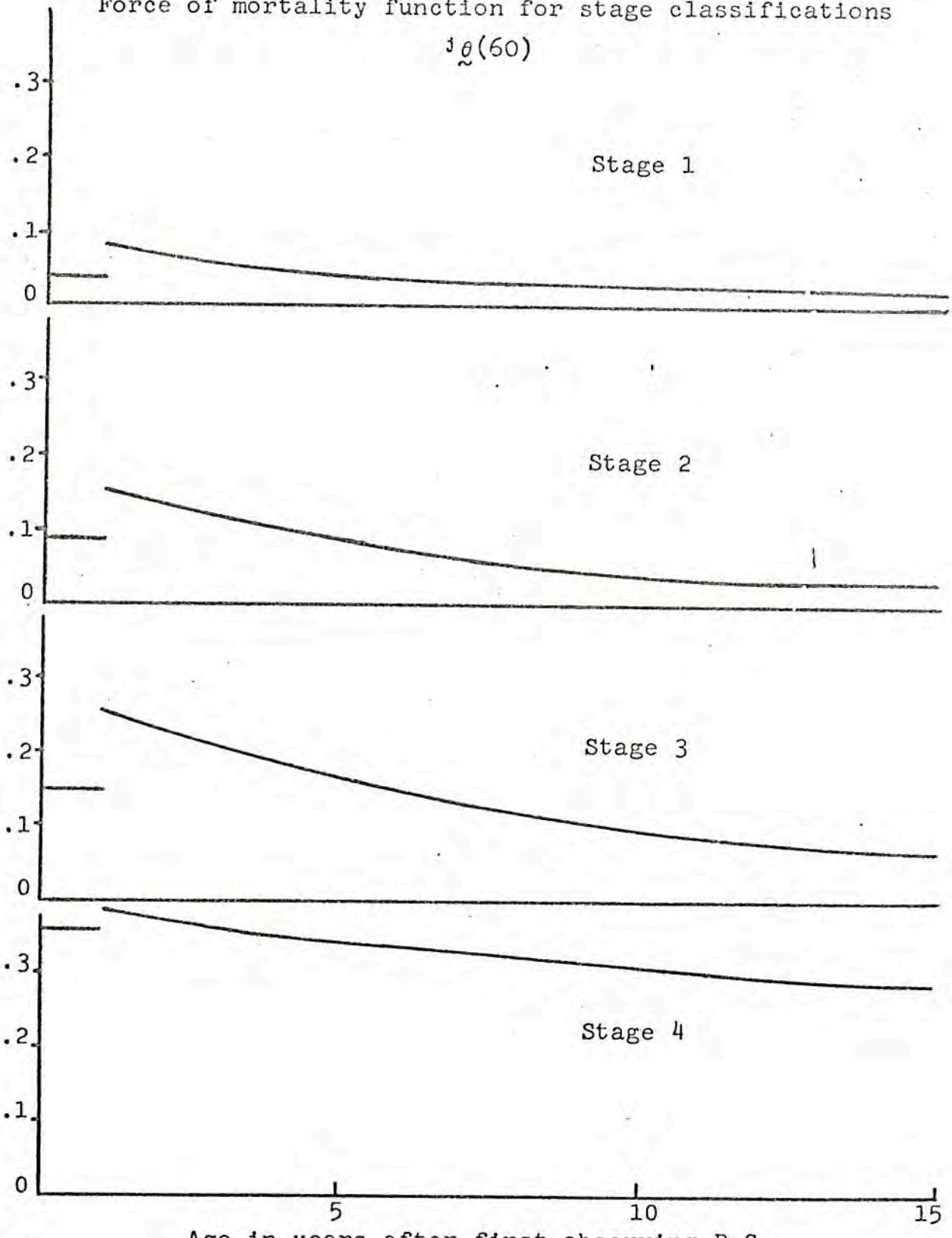
Force of mortality function for histology classifications

$\underset{\sim}{\varrho}^{1}(60)$



Force of mortality function for stage classifications

$^{j}\underset{\sim}{\theta}(60)$



B [10]

THE RELATIONSHIP BETWEEN A CLASS OF ASYMPTOTICALLY NORMAL ESTIMATORS AND GOODNESS OF FIT TESTS

G.M. TALLIS AND P.L. CHESSON
DEPARTMENT OF STATISTICS,
UNIVERSITY OF ADELAIDE.

To appear in
"The Australian Journal of Statistics"

SUMMARY

This paper considers a class of asymptotically normal estimators defined implicitly in terms of asymptotically normal statistics. The estimators are used as a basis for a general goodness of fit test where asymptotic null and non-null distributions are obtained. Certain standard tests are obtained as special cases of these results.

I Introduction

Let $\underset{\sim}{X}_n$ be a sequence of random vectors with components $X_i^{(n)}$, $i=1,2,\ldots,k$. It is assumed that $\sqrt{n}\,(\underset{\sim}{X}_n-\underset{\sim}{\mu}(\underset{\sim}{\theta})) \xrightarrow{L} N(\underset{\sim}{0},V(\underset{\sim}{\theta}))$ where $\underset{\sim}{\theta} \in \Omega$ is a q dimensional vector of unknown parameters and $q < \text{rank } V \equiv \rho(V) = \ell \leqslant k$.

Define the general class of estimators of $\underset{\sim}{\theta}$, $\underset{\sim}{\theta}_n$, implicitly by

$$\underset{\sim}{\Phi}(\underset{\sim}{X}_n,\underset{\sim}{\theta}_n) = \underset{\sim}{0}, \quad \underset{\sim}{\Phi}' = (\Phi_1,\ldots,\Phi_q)$$

where $\underset{\sim}{\Phi}$ satisfies the Implicit Function Theorem at $(\underset{\sim}{\mu}(\underset{\sim}{\theta}),\underset{\sim}{\theta})$ i.e. $\underset{\sim}{\Phi}(\underset{\sim}{\mu}(\underset{\sim}{\theta}),\underset{\sim}{\theta}) = \underset{\sim}{0}$ and $|\partial\Phi_i(\underset{\sim}{\mu}(\underset{\sim}{\theta}),\underset{\sim}{\theta})/\partial\theta_j| \neq 0$. Then $\underset{\sim}{\theta}_n$ is consistent for $\underset{\sim}{\theta}$ and $\sqrt{n}(\underset{\sim}{\theta}_n-\underset{\sim}{\theta})$ is asymptotically normally distributed.

In this paper the use of $\underset{\sim}{\theta}_n$ in certain goodness of fit tests is examined. The aim of the work is to establish general procedures which include some standard tests as special cases. The structure is also used to discuss minimum quadratic form estimation first introduced, in a limited context, by Gurland and Dahiya (1972).

II Main Result

If $\rho(V) = \ell < k$ then there is a $(k-\ell) \times k$ matrix H such that $VH' = 0$ and $\rho\begin{pmatrix} V \\ H \end{pmatrix} = k$. This fact will be used below.

Theorem 1

Let $\underset{\sim}{\mu}(\underset{\sim}{\theta})$ have continuous first order partial derivatives with respect to $\underset{\sim}{\theta} \in \Omega$ and $\Phi_i(\underset{\sim}{X}_n,\underset{\sim}{\theta})$ have continuous first order partial derivatives with respect to $\underset{\sim}{X}_n$ and $\underset{\sim}{\theta} \in \Omega$. Suppose that

(1) $M = [m_{ij}]$, $m_{ij} = \partial\Phi_i/\partial X_j$, $N = [n_{ir}]$,

$n_{ir} = -\partial\Phi_i/\partial\theta_r$; $Q = [q_{jr}]$, $q_{jr} = \partial\mu_j(\underset{\sim}{\theta})/\partial\theta_r$

$i,r = 1,2,\ldots,q$, $j = 1,\ldots,k$; $\Sigma = (I-A')V(I-A)$,

$A' = QN^{-1}M$; range of $Q(\underset{\sim}{\theta}) \subset$ range of $V(\underset{\sim}{\theta})$.

(2) N^{-1} exists for $\underset{\sim}{\theta} \in \Omega$;

Then $W_n = (\underset{\sim}{X}_n - \underset{\sim}{\mu}(\underset{\sim}{\theta}_n))' (\Sigma + M'M + H'H)^{-1} (\underset{\sim}{X}_n - \underset{\sim}{\mu}(\underset{\sim}{\theta}_n))$

is asymptotically distributed as $\chi^2(\ell-q)$.

Note: (i) M and N are evaluated at $\underset{\sim}{\mu}(\underset{\sim}{\theta})$, $\underset{\sim}{\theta}$ and Q at $\underset{\sim}{\theta}$, and for the purpose of differentiation all X_j are treated as independent variables.

(ii) Implicit differentiation of $\Phi(\underset{\sim}{\mu}(\underset{\sim}{\theta}),\underset{\sim}{\theta})$ with respect to $\underset{\sim}{\theta}$ shows that $MQ=N$ and $\rho(Q)=q$.

(iii) Since $\begin{bmatrix} M \\ H \end{bmatrix} [Q,H'] = \begin{bmatrix} N & MH' \\ O & HH' \end{bmatrix}$, $\rho(M',H') = q+k-1$.

The assumption that N^{-1} exists is a restatement of the assumed Implicit Function Theorem property of $\underset{\sim}{\Phi}$.

Proof We make use of the asymptotic methods described by Rao 1965, Chpt. 6.

For $i = 1,2,\ldots,q$,

$$\sqrt{n}\,\Phi_i(\underset{\sim}{X}_n,\underset{\sim}{\theta}_n) - \Big[\sqrt{n}\,\Phi_i(\underset{\sim}{\mu}(\underset{\sim}{\theta}),\underset{\sim}{\theta}) + \sqrt{n}\sum_{j=1}^{k}\frac{\partial}{\partial X_j}\Phi_i(\underset{\sim}{\mu}(\underset{\sim}{\theta}),\underset{\sim}{\theta})(X_j^{(n)}-\mu_j(\underset{\sim}{\theta}))$$

$$+ \sqrt{n}\sum_{r=1}^{q}\frac{\partial}{\partial\theta_r}\Phi_i(\underset{\sim}{\mu}(\underset{\sim}{\theta}),\underset{\sim}{\theta})(\theta_r^{(n)}-\theta_r)\Big] \xrightarrow{P} 0 .$$

or in matrix notation

$$\sqrt{n}\,[M(\underset{\sim}{X}_n - \underset{\sim}{\mu}(\underset{\sim}{\theta})) - N(\underset{\sim}{\theta}_n - \underset{\sim}{\theta})] \xrightarrow{P} \underset{\sim}{0}.$$

Since N^{-1} exists for $\underset{\sim}{\theta} \in \Omega$,

$$\sqrt{n}\,(\underset{\sim}{\theta}_n - \underset{\sim}{\theta}) - \sqrt{n}\,N^{-1}M(\underset{\sim}{X}_n - \underset{\sim}{\mu}(\underset{\sim}{\theta})) \xrightarrow{P} \underset{\sim}{0}.$$

Now

$$\sqrt{n}\,(\underset{\sim}{X}_n - \underset{\sim}{\mu}(\underset{\sim}{\theta}_n)) - [\sqrt{n}(\underset{\sim}{X}_n - \underset{\sim}{\mu}(\underset{\sim}{\theta})) - \sqrt{n}\,Q(\underset{\sim}{\theta}_n - \underset{\sim}{\theta})] \xrightarrow{P} \underset{\sim}{0}$$

which becomes upon substituting for $\sqrt{n}(\underset{\sim}{\theta}_n - \underset{\sim}{\theta})$

$$\sqrt{n}\,(\underset{\sim}{X}_n - \underset{\sim}{\mu}(\underset{\sim}{\theta}_n)) - \sqrt{n}\,(I - QN^{-1}M)(\underset{\sim}{X}_n - \underset{\sim}{\mu}(\underset{\sim}{\theta})) \xrightarrow{P} \underset{\sim}{0}.$$

Hence $\sqrt{n}\,(\underset{\sim}{X}_n - \underset{\sim}{\mu}(\underset{\sim}{\theta}_n)) \xrightarrow{L} N(\underset{\sim}{0},\Sigma)$ .....(*)

where $\Sigma = (I-A')V(I-A)$, $A' = QN^{-1}M$ and

$$n(\underset{\sim}{X}_n-\underset{\sim}{\mu}(\underset{\sim}{\theta}_n))'\Sigma^g(\underset{\sim}{X}_n-\underset{\sim}{\mu}(\underset{\sim}{\theta}_n)) \xrightarrow{L} \chi^2(d)$$

with $\Sigma\Sigma^g\Sigma = \Sigma$ and $d = \rho(\Sigma)$. This follows since, if $\underset{\sim}{Z}$ is distributed as $N(0,\Sigma)$, $\underset{\sim}{Z}'B\underset{\sim}{Z}$ is distributed as $\chi^2(d)$ if $\Sigma B\Sigma B\Sigma = \Sigma B\Sigma$ and $tr.B\Sigma = d$, Searle (1971, page 69). Choosing $B = \Sigma^g$, then $\Sigma\,\Sigma^g\Sigma\,\Sigma^g\Sigma = \Sigma\,\Sigma^g\Sigma$ and $\rho(\Sigma) = tr.\,\Sigma^g\Sigma$.

Now $V = P'I^{(\ell)}I^{(\ell)}P$ where P is non-singular and $I^{(\ell)} = diag(1,\ldots,1,0,\ldots 0)$, there being ℓ 1's, and $\Sigma = (I-A')P'I^{(\ell)}I^{(\ell)}P(I-A) = E'E$, say.

Clearly $\rho(\Sigma) = \rho(E)$ and $EM' = 0$, $EH' = 0$ and hence the null space of E has dimension greater than or equal to $\rho(M' \vdots H') = q+k-\ell$. Thus $\rho(E) \leqslant k-q-k+\ell = \ell-q$.

But for $k\times k$ matrices E_1 and E_2, $\rho(E_1) = r$ and $\rho(E_2) = s$, $\rho(E_1E_2) \geqslant r+s-k$. Put $E_1 = I^{(\ell)}P$, $E_2 = (I-A)$, then $\rho(I^{(\ell)}P) = \ell$ and $\rho(I-A) = tr(I-A) = k-q$ and $\rho(E) \geqslant \ell + k-q-k = \ell-q$. Finally, $\rho(\Sigma) = \ell-q$ and

$$\rho\begin{bmatrix} E \\ M \\ H \end{bmatrix} = k$$

and a generalised inverse of Σ, Σ^g, is $(\Sigma+M'M+H'H)^{-1}$, Rao (1965, page 30).

Note that Σ and M are usually continuous functions of $\underset{\sim}{\theta}$, but in $(\Sigma+M'M+H'H)^{-1}$ $\underset{\sim}{\theta}$ can be replaced by any consistent estimator, for example $\underset{\sim}{\theta}_n$, without altering the asymptotic result.

III The Non-Null Case

Results in this section are applied in IV in connection with the asymptotic power of goodness of fit tests.

We define the non-null case as

$$\sqrt{n}\ (\underset{\sim}{X}_n - \underset{\sim}{\mu}(\underset{\sim}{\theta})) \xrightarrow{L} N(\underset{\sim}{\gamma}, G).$$

Note that the property $\underset{\sim}{X}_n \xrightarrow{P} \underset{\sim}{\mu}(\underset{\sim}{\theta})$ is unaltered and so the estimator $\underset{\sim}{\theta}_n$ can be defined as in I.

If $\underset{\sim}{Y}_n = \underset{\sim}{X}_n - \frac{1}{\sqrt{n}}\underset{\sim}{\gamma}$, it follows that

$$\sqrt{n}\ (\underset{\sim}{Y}_n - \underset{\sim}{\mu}(\underset{\sim}{\theta})) \xrightarrow{L} N(\underset{\sim}{0}, G).$$

We need the following lemma.

Lemma

If $\underset{\sim}{f}$ is a function with continuous first order partial derivatives then

$$\sqrt{n}\ (\underset{\sim}{f}(\underset{\sim}{X}_n) - \underset{\sim}{f}(\underset{\sim}{Y}_n)) - J\underset{\sim}{\gamma} \xrightarrow{P} \underset{\sim}{0}$$

where J is the matrix $\left(\frac{\partial f_i}{\partial x_j}\right)$ evaluated at $\underset{\sim}{\mu}$.

Proof

Let g be a component of $\underset{\sim}{f}$. From the Mean Value Theorem

$$\sqrt{n}\ (g(\underset{\sim}{X}_n) - g(\underset{\sim}{Y}_n)) = \underset{\sim}{\gamma}' \nabla g(\underset{\sim}{Z}_n)$$ where $\underset{\sim}{Z}_n$ is on the line segment joining $\underset{\sim}{X}_n$ and $\underset{\sim}{Y}_n$; and ∇g is the gradient vector.

This can be rearranged to give

$$\sqrt{n}\ (g(\underset{\sim}{X}_n) - g(\underset{\sim}{Y}_n)) - \underset{\sim}{\gamma}' \nabla g(\underset{\sim}{\mu}) = \underset{\sim}{\gamma}' (\nabla g(\underset{\sim}{Z}_n) - \nabla g(\underset{\sim}{\mu})).$$

Now $\underset{\sim}{X}_n$ and $\underset{\sim}{Y}_n \xrightarrow{P} \underset{\sim}{\mu}$ implying $\underset{\sim}{Z}_n \xrightarrow{P} \underset{\sim}{\mu}$. The continuity of ∇g implies

$$\underset{\sim}{\gamma}' (\nabla g(\underset{\sim}{Z}_n) - \nabla g(\underset{\sim}{\mu})) \xrightarrow{P} 0.$$

Hence $\sqrt{n}\ (g(\underset{\sim}{X}_n) - g(\underset{\sim}{Y}_n)) - \underset{\sim}{\gamma}' \nabla g(\underset{\sim}{\mu}) \xrightarrow{P} 0.$

This holds for every component of $\underset{\sim}{f}$ giving the result.

Theorem 2.

If the conditions on $\underset{\sim}{\mu}$ and $\underset{\sim}{\Phi}$ stated in Theorem 1 hold and

(1) $G = V(\underset{\sim}{\theta})$

(2) $\underset{\sim}{\gamma}$ is in the range of $V(\underset{\sim}{\theta})$

then W_n is asymptotically distributed as $\chi^2(\ell-q,\lambda)$ where

$$\lambda = \tfrac{1}{2}\,\underset{\sim}{\gamma}'(I-A)\,(\Sigma+M'M+H'H)^{-1}\,(I-A')\underset{\sim}{\gamma}$$

is the non-centrality parameter.

Proof

Define $\underset{\sim}{\varphi}_n$ by

$$\Phi(\underset{\sim}{Y}_n,\underset{\sim}{\varphi}_n) = 0.$$

Theorem 1 applies to $\underset{\sim}{Y}_n$ and $\underset{\sim}{\varphi}_n$ and so we obtain from statement (*)

$$\sqrt{n}\,(\underset{\sim}{Y}_n-\underset{\sim}{\mu}(\underset{\sim}{\varphi}_n)) \xrightarrow{L} N(\underset{\sim}{0},\Sigma).$$

Moreover we note that

$$\sqrt{n}(\underset{\sim}{Y}_n-\mu(\underset{\sim}{\varphi}_n)) = \sqrt{n}\,(\underset{\sim}{X}_n-\underset{\sim}{\mu}(\underset{\sim}{\theta}_n)) - \underset{\sim}{\gamma} + \sqrt{n}\,(\underset{\sim}{\mu}(\underset{\sim}{\theta}_n)-\underset{\sim}{\mu}(\underset{\sim}{\varphi}_n)).$$

Now $\underset{\sim}{\mu}(\underset{\sim}{\theta}_n) = \underset{\sim}{\mu}(\underset{\sim}{h}(\underset{\sim}{X}_n))$; $\underset{\sim}{\mu}(\underset{\sim}{\varphi}_n) = \underset{\sim}{\mu}(\underset{\sim}{h}(\underset{\sim}{Y}_n))$

where $\underset{\sim}{h}$ is the function determined implicitly by $\underset{\sim}{\Phi}$ at the point $(\underset{\sim}{\mu}(\underset{\sim}{\theta}),\underset{\sim}{\theta})$. The lemma applies giving

$$\sqrt{n}\,(\underset{\sim}{\mu}(\underset{\sim}{\theta}_n)-\underset{\sim}{\mu}(\underset{\sim}{\varphi}_n)) \xrightarrow{P} \Omega N^{-1}M\underset{\sim}{\gamma} = A'\underset{\sim}{\gamma}.$$

Hence $\sqrt{n}\,(\underset{\sim}{X}_n-\underset{\sim}{\mu}(\underset{\sim}{\theta}_n)) \xrightarrow{L} N((I-A')\underset{\sim}{\gamma},\Sigma)$.

If $\underset{\sim}{Z}$ is distributed $N(\underset{\sim}{\delta},\Sigma)$ then $\underset{\sim}{Z}'B\underset{\sim}{Z}$ is distributed $\chi^2(\mathrm{tr}B\Sigma,\ \tfrac{1}{2}\underset{\sim}{\delta}'B\underset{\sim}{\delta})$ if and only if

(i) $\Sigma B\Sigma B\Sigma = \Sigma B\Sigma$

(ii) $\underset{\sim}{\delta}'B\Sigma = \underset{\sim}{\delta}'B\Sigma B\Sigma$

(iii) $\underset{\sim}{\delta}'B\underset{\sim}{\delta} = \underset{\sim}{\delta}'B\Sigma B\underset{\sim}{\delta}$, Searle (loc. cit.).

With $B = \Sigma^g$ (i) holds as in Theorem 1. $\Sigma^g\Sigma\Sigma^g\Sigma = \Sigma^g\Sigma$ and so (ii) holds. Finally $(I-A')\underset{\sim}{\gamma}$ belongs to the range of $\Sigma = (I-A')V(I-A)$ since $\underset{\sim}{\gamma}$ is in the range of V. Hence

$$\Sigma\Sigma^g(I-A')\underset{\sim}{\gamma} = (I-A')\underset{\sim}{\gamma}$$

and so

$$\underset{\sim}{\gamma}'(I-A)\Sigma^g(I-A')\underset{\sim}{\gamma} = \underset{\sim}{\gamma}'(I-A)\Sigma^g\Sigma\Sigma^g(I-A')\underset{\sim}{\gamma}.$$

Thus (iii) holds with $\underset{\sim}{\delta} = (I-A')\underset{\sim}{\gamma}$.

It follows from these observations that the asymptotic distribution of W_n is $\chi^2(\ell-q,\lambda)$.

If $G \neq V(\underset{\sim}{\theta})$ or $\underset{\sim}{\gamma}$ is not in the range of $V(\underset{\sim}{\theta})$ the asymptotic distribution of W_n is not, in general, a non-central χ^2. It can be written as the distribution of $\underset{\sim}{Z}'\Sigma^g\underset{\sim}{Z}$ where $\underset{\sim}{Z}$ has the distribution $N((I-A')\underset{\sim}{\gamma}, (I-A')G(I-A))$. Such distributions are discussed by Johnson and Kotz (1970, Chapter 29).

Note

The conditions (i), (ii), and (iii) above could have been written into Theorems 1 and 2 directly to obtain necessary and sufficient conditions for asymptotic χ^2. However this more general approach has less practical advantage.

IV Specialisations

(a) Standard Goodness of Fit Tests

Let Y be a random variable with distribution function $F(y,\underset{\sim}{\theta})$, $\underset{\sim}{\theta} \in \Omega$ a q-vector of unknown parameters. For $y_0 = -\infty < y_1 < y_2 < \ldots < y_k = \infty$ put

$$p_i(\underset{\sim}{\theta}) = F(y_i,\underset{\sim}{\theta}) - F(y_{i-1},\underset{\sim}{\theta}), \quad i=1,\ldots,k.$$

A random sample of size n is taken from $F(y,\underset{\sim}{\theta})$ and N_i is the number in this sample for which $y_{i-1} < Y \leqslant y_i$. The theory of II will be used to establish certain classical results with regard to the asymptotic distribution of the test statistic $W_n = \sum_{i=1}^{k}(N_i-np_i(\underset{\sim}{\theta}_n))^2/np_i(\underset{\sim}{\theta}_n)$, where $\underset{\sim}{\theta}_n$

represents certain estimators for $\underset{\sim}{\theta}$.

It is known for instance that when $\underset{\sim}{\theta}_n = \hat{\underset{\sim}{\theta}}_n$, the maximum likelihood estimator, then $W_n \xrightarrow{L} \chi^2(k-q-1)$. This result holds whenever $\hat{\underset{\sim}{\theta}}_n$ is replaced by an asymptotically efficient estimator $\underset{\sim}{\theta}_n^*$ since then $\sqrt{n}\ (\underset{\sim}{\theta}_n^* - \hat{\underset{\sim}{\theta}}_n) \xrightarrow{P} 0$.

In the notation of II put $\underset{\sim}{\mu}(\underset{\sim}{\theta}) = \underset{\sim}{p}(\underset{\sim}{\theta})$, $X_i^{(n)} = \frac{N_i}{n}$, $V = \Lambda^{-1}(I - \underset{\sim}{\phi}\underset{\sim}{\phi}')\Lambda^{-1}$, $\Lambda^{-1} = \mathrm{diag}(p_1^{\frac{1}{2}}(\underset{\sim}{\theta}), \ldots, p_k^{\frac{1}{2}}(\underset{\sim}{\theta}))$ $\underset{\sim}{\phi} = \Lambda^{-1}\underset{\sim}{1}$, $H = \underset{\sim}{1}' = (1, \ldots, 1)$. Theorem 1 then states that

$$n(\underset{\sim}{X}_n - \underset{\sim}{p}(\underset{\sim}{\theta}_n))'(\Sigma + M'M + \underset{\sim}{1}\underset{\sim}{1}')^{-1}(\underset{\sim}{X}_n - \underset{\sim}{p}(\underset{\sim}{\theta}_n)) \xrightarrow{L} \chi^2(k-q-1).$$

For the non-null case consider a sequence of alternatives obtained by replacing $F(y, \underset{\sim}{\theta})$ with $F_n(y) = F(y, \underset{\sim}{\theta}) + \frac{1}{\sqrt{n}} G(y)$. Under the alternative $n\underset{\sim}{X}_n$ is a multinomial with mean $np(\underset{\sim}{\theta}) + \sqrt{n}\ \underset{\sim}{\gamma}$ in which $\underset{\sim}{\gamma}'\underset{\sim}{1} = 0$. It can be shown that $\sqrt{n}\ (\underset{\sim}{X}_n - \underset{\sim}{p}(\underset{\sim}{\theta})) \xrightarrow{L} N(\underset{\sim}{\gamma}, V)$. Note that $\underset{\sim}{\gamma}$ belongs to the range of V since $\rho(V) = k-1$ and $V\underset{\sim}{1} = \underset{\sim}{0}$. Hence Theorem 2 applies giving the asymptotic distribution of $n(\underset{\sim}{X}_n - \underset{\sim}{p}(\underset{\sim}{\theta}_n))'(\Sigma + M'M + \underset{\sim}{1}'\underset{\sim}{1})^{-1}(\underset{\sim}{X}_n - \underset{\sim}{p}(\underset{\sim}{\theta}_n))$ as $\chi^2(k-q-1, \lambda)$.

(i) Maximum Likelihood, $\underset{\sim}{\theta}_n = \hat{\underset{\sim}{\theta}}_n$.

When $\underset{\sim}{\theta}$ is estimated by maximum likelihood Φ_i takes the form

$$\Phi_i(\underset{\sim}{X}_n, \hat{\underset{\sim}{\theta}}_n) = \sum_{j=1}^{k} X_j^{(n)} \frac{\partial}{\partial \theta_i} \ln p_j(\hat{\underset{\sim}{\theta}}_n) = 0.$$

The condition $\underset{\sim}{\Phi}(\underset{\sim}{p}(\underset{\sim}{\theta}), \underset{\sim}{\theta}) = \underset{\sim}{0}$ of Theorem 1 is satisfied and

a(1) $M = R'\Lambda$ for $R = \Lambda Q$ and
$N = R'R = J$, the information matrix;

a(2) J^{-1} is assumed to exist for $\underset{\sim}{\theta} \in \Omega$;

a(3) $\rho(M' \vdots \underset{\sim}{1}) = \rho(\Lambda^{-1}M' \vdots \underset{\sim}{\phi})$ and
$\Lambda^{-1}M' = R = \Lambda Q$; thus $R'\underset{\sim}{\phi} = Q'\Lambda\underset{\sim}{\phi} = Q'\underset{\sim}{1} = \underset{\sim}{0}$
and $\rho(M' \vdots \underset{\sim}{1}) = q+1$.

Consider $W_n = n(\underset{\sim}{X}_n - \underset{\sim}{p}(\hat{\underset{\sim}{\theta}}_n)' \Lambda^2 (\underset{\sim}{X}_n - \underset{\sim}{p}(\hat{\underset{\sim}{\theta}}_n))$, then $W_n \xrightarrow{L} \chi^2(k-q-1)$ if $\Sigma\Lambda^2\Sigma\Lambda^2\Sigma = \Sigma\Lambda^2\Sigma$ and $\operatorname{tr} \Lambda^2\Sigma = k-q-1$. Using the special forms of M, Q and N

$$\Sigma = \Lambda^{-1}[I - RJ^{-1}R'][I - \underset{\sim}{\phi}\underset{\sim}{\phi}'][I - RJ^{-1}R']\Lambda^{-1}$$

and the matrix condition is obviously satisfied while

$$\operatorname{tr} \Lambda^2\Sigma = \operatorname{tr}[I - RJ^{-1}R' - \underset{\sim}{\phi}\underset{\sim}{\phi}'] = k-q-1.$$

This establishes the classical result $W_n \xrightarrow{L} \chi^2(k-q-1)$. In the non-null case $W_n \xrightarrow{L} \chi^2(k-q-1, \lambda)$ where $\lambda = \tfrac{1}{2}\underset{\sim}{\gamma}'\Lambda(I - RJ^{-1}R')\Lambda\underset{\sim}{\gamma}$.

(ii) A Moment Estimator, $\underset{\sim}{\theta}_n = \bar{\underset{\sim}{\theta}}_n$.

Let $T = [t_{ij}]$, $t_{ij} = \bar{y}_j^i$, $j=1,2,\ldots,k$, $i=1,2,\ldots,q$, $\bar{y}_j \in (y_{j-1}, y_j]$ then a moment estimator of $\underset{\sim}{\theta}$ is defined by

$$\underset{\sim}{\Phi}(\underset{\sim}{X}_n, \bar{\underset{\sim}{\theta}}_n) = T(\underset{\sim}{X}_n - \underset{\sim}{p}(\bar{\underset{\sim}{\theta}}_n)) = \underset{\sim}{0},$$

where $T\underset{\sim}{p}(\underset{\sim}{\theta}) = \underset{\sim}{\mu}(\underset{\sim}{\theta})$. In this case, again, $\underset{\sim}{\Phi}(\underset{\sim}{p}(\underset{\sim}{\theta}), \underset{\sim}{\theta}) = 0$, $M = T$, $N = TQ$ and N^{-1} is assumed to exist. It follows that $\rho(T' \vdots \underset{\sim}{1}) = q+1$ and the general results of Theorems 1 and 2 can then be applied using these special forms of M and N.

(iii) Minimum chi-square $\underset{\sim}{\theta}_n^c$.

This case is dealt with in the next section.

(b) Minimum Quadratic Form Estimators

Let

$$A(\underset{\sim}{\theta}, S(\underset{\sim}{\theta})) = n(\underset{\sim}{X}_n - \underset{\sim}{\mu}(\underset{\sim}{\theta}))' S(\underset{\sim}{\theta}) (\underset{\sim}{X}_n - \underset{\sim}{\mu}(\underset{\sim}{\theta}))$$

where $S(\underset{\sim}{\theta})$ is symmetric and of full rank, $\underset{\sim}{\theta} \in \Omega$. Then the minimum quadratic form estimator for $\underset{\sim}{\theta}$, $\underset{\sim}{\theta}_n^*$, is defined by

$$A(\underset{\sim}{\theta}_n^*, S(\underset{\sim}{\theta}_n^*)) = \min_{\underset{\sim}{\theta} \in \Omega} A(\underset{\sim}{\theta}, S(\underset{\sim}{\theta}))$$

Thus $A(\underset{\sim}{\theta}_n^*, S(\underset{\sim}{\theta}_n^*))$ is a measure of the agreement between $\underset{\sim}{X}_n$ and $\underset{\sim}{\mu}(\underset{\sim}{\theta})$. The asymptotic distribution of a certain quadratic form in $\underset{\sim}{X}_n - \underset{\sim}{\mu}(\underset{\sim}{\theta}_n^*)$ will be derived with the help of Theorems 1 and 2.

Set

$$\Phi_i(\underset{\sim}{X}_n, \underset{\sim}{\theta}) = (2n)^{-1} \frac{\partial}{\partial \theta_i} A(\underset{\sim}{\theta}, S(\underset{\sim}{\theta}))$$

$$= -(\underset{\sim}{X}_n - \underset{\sim}{\mu}(\theta))' S(\underset{\sim}{\theta}) \mu_i(\underset{\sim}{\theta}) + \tfrac{1}{2} (\underset{\sim}{X}_n - \underset{\sim}{\mu}(\underset{\sim}{\theta}))' S_i(\underset{\sim}{\theta}) (\underset{\sim}{X}_n - \underset{\sim}{\mu}(\underset{\sim}{\theta}))$$

$\underset{\sim}{\mu}_i(\underset{\sim}{\theta}) = \frac{\partial}{\partial \theta_i} \underset{\sim}{\mu}(\underset{\sim}{\theta})$, $S_i(\underset{\sim}{\theta}) = \frac{\partial}{\partial \theta_i} S(\underset{\sim}{\theta})$. Put $\underset{\sim}{\theta} = \underset{\sim}{\theta}_n^*$, then

$$\Phi_i(\underset{\sim}{X}_n, \underset{\sim}{\theta}_n^*) = 0, \quad i=1,2,\ldots,q$$

or

$$\underset{\sim}{\Phi}(\underset{\sim}{X}_n, \underset{\sim}{\theta}_n^*) = \underset{\sim}{0}.$$

Here $M = -Q'S$, $N = Q'SQ$ and

$$\Sigma = (I - Q(Q'SQ)^{-1}Q'S)V(I - SQ(Q'SQ)^{-1}Q').$$

Theorems 1 and 2 can now be applied.

If V is of full rank, choose $S = V^{-1}$, then $\Sigma V^{-1} \Sigma V^{-1} \Sigma = \Sigma V^{-1} \Sigma$, $\operatorname{tr}(V^{-1}\Sigma) = k-q$ and, in the null case,

$$n(\underset{\sim}{X}_n - \underset{\sim}{\mu}(\underset{\sim}{\theta}_n^*))' (V(\underset{\sim}{\theta}_n^*))^{-1} (\underset{\sim}{X}_n - \underset{\sim}{\mu}(\underset{\sim}{\theta}_n^*)) \xrightarrow{L} \chi^2(k-q)$$

In the non-null case $\underset{\sim}{\gamma}$ will belong to the range of V, since V is of full rank, and so a non-central χ^2 will result. Here the non-centrality parameter is $\tfrac{1}{2}\underset{\sim}{\gamma}' (V - Q(Q'V^{-1}Q)^{-1}Q'$

<u>Note</u>

If $\tilde{\underset{\sim}{\theta}}_n$ is defined by

$$\min_{\underset{\sim}{\theta} \in \Omega} A(\underset{\sim}{\theta}, V^{-1}) = A(\tilde{\underset{\sim}{\theta}}_n, V^{-1})$$

where V is the true asymptotic variance matrix, then $A(\tilde{\underset{\sim}{\theta}}_n, V^{-1}(\tilde{\underset{\sim}{\theta}}_n))$ has the same asymptotic distribution as $A(\underset{\sim}{\theta}_n^*, V^{-1}(\underset{\sim}{\theta}_n^*))$. This result is needed in the example.

On the other hand, if $\underset{\sim}{\mu}(\underset{\sim}{\theta}) = \underset{\sim}{p}(\underset{\sim}{\theta})$ and V takes the form $\Lambda^{-1}(I-\underset{\sim}{\phi}\underset{\sim}{\phi}')\Lambda^{-1}$, put $S = \Lambda^2$. In this case Σ reduces to the special form of a(i). This shows that if $\hat{\underset{\sim}{\theta}}_n$ is replaced by the minimum chi-square estimator, $\underset{\sim}{\theta}^c_n$, defined by

$$\min_{\theta\in\Omega} A(\underset{\sim}{\theta},\Lambda^2(\underset{\sim}{\theta})) = A(\underset{\sim}{\theta}^c_n,\Lambda^2(\underset{\sim}{\theta}^c_n))$$

the results of a(i) are unaltered i.e. $W_n \xrightarrow{L} \chi^2(k-q-1)$, in the null case, and $W_n \xrightarrow{L} \chi^2(k-q-1,\lambda)$ in the non-null case with the same non-null parameter.

For earlier results on generalised quadratic form estimation see Gurland and Dahiya (1972).

Example

Consider a continuous and strictly monotone distribution function which is specified up to a q dimensional vector of unknown parameters, $F(y,\underset{\sim}{\theta})$, $\underset{\sim}{\theta} \in \Omega$. A random sample $Y_1, Y_2, \ldots, Y_n$ is drawn and let X_{pj}, $j = 1,2,\ldots,k$, be the p_jth quantile statistic defined in the usual way, $p_i < p_j$, $i < j$.

It is well known that, under mild conditions the vector $\underset{\sim}{X}'_n = (X^{(n)}_{p_1},\ldots,X^{(n)}_{pk})$ is asymptotically normally distributed with mean vector $\underset{\sim}{y}' = (y_{p_1},\ldots,y_{pk})$ and covariance matrix $n^{-1}V$, where for $i < j$ $v_{ij} = p_iq_j/f(y_{pi})f(y_{pj})$, $F(y_{pi}) = p_i, p_i+q_i = 1$ and $F'(y) = f(y)$, Cramér (1946, page 367). Since F depends on $\underset{\sim}{\theta}$, it follows that $y_{pi} = F^{-1}(p_i)$ and V also depend on $\underset{\sim}{\theta}$, $y_{pi}(\underset{\sim}{\theta})$ and $V(\underset{\sim}{\theta})$ say. Then, it has been shown that

$$W_n = \min_{\underset{\sim}{\theta}\in\Omega} n\,(\underset{\sim}{X}_n-\underset{\sim}{y}(\underset{\sim}{\theta}))'V^{-1}(\underset{\sim}{\theta})\,(\underset{\sim}{X}_n-\underset{\sim}{y}(\underset{\sim}{\theta}))$$

is asymptotically distributed as $\chi^2(k-q)$.

For the non-null case consider a sequence of alternatives of the form

$$F_n(y) = F(y,\underset{\sim}{\theta}) + \frac{1}{\sqrt{n}} G(y)$$

With suitable conditions on G it can be shown that

$$\sqrt{n}\ (\underset{\sim}{X}_n - \underset{\sim}{y}(\underset{\sim}{\theta})) \xrightarrow{L} N(\underset{\sim}{\gamma}, V)$$

where

$$\gamma_i = -G(y_{pi})/f(y_{pi})$$

Since V is of full rank $\underset{\sim}{\gamma}$ belongs to the range of V. Theorem 2 applies and so W_n is asymptotically distributed $\chi^2(k-q,\lambda)$ where $\lambda = \frac{1}{2}\underset{\sim}{\gamma}'(V-Q(Q'V^{-1}Q)^{-1}Q')\underset{\sim}{\gamma}$.

In the case where F is normal, $\underset{\sim}{\theta}' = (\mu,\sigma)$, $y_{pi}(\underset{\sim}{\theta}) = \mu+\sigma a_i$ and $V(\underset{\sim}{\theta}) = \sigma^2 V_1$ where $\frac{1}{\sqrt{2\pi}} \int_{-\infty}^{a_i} e^{\frac{-t^2}{2}} dt = p_i$ and V_1 is a constant matrix.

Define $\underset{\sim}{\tilde{\theta}}_n$ as the value which minimises the quadratic form

$$(\underset{\sim}{X}_n - \underset{\sim}{y}(\underset{\sim}{\theta}))'\ V_1^{-1}(\underset{\sim}{X}_n - \underset{\sim}{y}(\underset{\sim}{\theta})).$$

Since V is proportional to V_1 this procedure is equivalent to minimising the quadratic form with the true V replacing V_1. By the note in IV(b)

$$(\underset{\sim}{X}_n - \underset{\sim}{y}(\underset{\sim}{\tilde{\theta}}_n))'(\tilde{\sigma}^2 V_1)^{-1}(\underset{\sim}{X}_n - \underset{\sim}{y}(\underset{\sim}{\tilde{\theta}}_n))$$

has the same asymptotic distribution as W_n.

Explicit expressions for $\tilde{\mu}$ and $\tilde{\sigma}$ are given by

$$\tilde{\mu} = \frac{\underset{\sim}{1}'V_1^{-1}\underset{\sim}{X}_n\underset{\sim}{a}'V_1^{-1}\underset{\sim}{a} - \underset{\sim}{a}'V_1^{-1}\underset{\sim}{X}_n\underset{\sim}{1}'V_1^{-1}\underset{\sim}{a}}{\underset{\sim}{1}'V_1^{-1}\underset{\sim}{1}\underset{\sim}{a}'V_1^{-1}\underset{\sim}{a} - (\underset{\sim}{1}'V_1^{-1}\underset{\sim}{a})^2}$$

$$\tilde{\sigma} = \frac{-\underset{\sim}{1}'V_1^{-1}\underset{\sim}{a}\underset{\sim}{1}'V_1^{-1}\underset{\sim}{X}_n + \underset{\sim}{1}'V_1^{-1}\underset{\sim}{1}\underset{\sim}{a}'V_1^{-1}\underset{\sim}{X}_n}{\underset{\sim}{1}'V_1^{-1}\underset{\sim}{1}\underset{\sim}{a}'V_1^{-1}\underset{\sim}{a} - (\underset{\sim}{1}'V_1^{-1}\underset{\sim}{a})^2}$$

where $\underset{\sim}{1}' = (1,1,\ldots,1)$. These results hold whenever θ_1 and θ_2 are location and scale parameters.

The idea of using order statistics to measure goodness of fit has been presented from a different point of view by Bofinger (1973).

ACKNOWLEDGEMENTS

We wish to thank Mr. T.J. O'Neill, for his helpful comments on an early draft of this paper, and the C.S.I.R.O. for support during the project.

REFERENCES

Bofinger, Eve. (1973) Goodness-of-fit test using sample quantiles.
J. Roy. Statist. Soc., B., 35, p277-284.

Cramér, H. (1946) Mathematical Methods of Statistics.
Princeton University Press.

Gurland, J. and Dahiya, R.C. (1972) Tests of fit for continuous distributions based on generalised minimum chi-square.
Statistical Papers in Honor of George W. Snedecor.
p115-128 Iowa State University Press.

Johnson, N.L. and Kotz, S. (1970) Distributions in Statistics, Continuous Univariate Distributions -2.
Houghton Mifflin, Boston.

Rao, C.R. (1965) Linear Statistical Inference and its Applications. Wiley, New York.

Searle, S.R. (1971) Linear Models. Wiley, New York.

B [11]

MATHEMATICAL BIOSCIENCES 19, 371–376 (1974)

On the Distribution of the Time to Reporting Cancers with Application to Breast Cancer in Women

G. M. TALLIS
Department of Statistics, University of Adelaide, Adelaide, S. A.

AND

G. SARFATY
Endocrine Research Unit, Peter McCallum Clinic, Melbourne, Victoria, Australia

Communicated by K. E. F. Watt

ABSTRACT

This paper explores a model for the delay in reporting some cancers. The basic assumption of the model is that the infinitesimal, conditional probability of reporting the disease is proportional to the rate of tumor growth. Under the hypothesis of exponential tumor growth, the distribution function of the time to reporting is tied to the clinical staging of the disease. From suitable data on breast cancer in women collected at the Central Cancer Registry of Victoria, estiamtes of tumor growth rate for several tumor types are obtained with the model. Estimates of the average delay from the onset of the disease to when it is first reported range from 5 to 14 years depending on tumor type.

I. INTRODUCTION

For people contracting various forms of cancer there is an inevitable delay between the time of onset of the disease and the time it is first presented for treatment. This period of delay which will be referred to as the time to reporting, T, varies according to the rate of tumor growth, the patient's awareness of the significance of symptoms and their reactions to them.

This note proposes a mathematical model for the distribution function, d.f., of the random variable T. The model is based on the assumption that the instantaneous rate of reporting the disease is proportional to its rate of progress. The analysis parallels actuarial methods with force of reporting playing the roll of the usual force of mortality.

II. THE MODEL

Let the d.f. of T be $F(t)=\Pr\{T\leqslant t\}$ with derivative $F'(t)=f(t)$, then the force of reporting function is defined by

$$f(t)[1-F(t)]^{-1}=\phi(t)=\begin{cases}0, & 0\leqslant t<t_0,\\ \alpha V'(t), & t_0\leqslant t,\end{cases} \tag{1}$$

where $V(t)$ is the tumor volume t time units after onset, α is a constant of proportionality and t_0 is defined by the equation $V(t_0)=v_0$. The volume v_0 is some predetermined, specified tumor volume around which the analysis is pivoted. For instance, in some case v_0 may be the minimum clinically detectable tumor size for the particular type of cancer. In other cases it may be required to take v_0 as the initial cell volume of the particular tissue concern $t_0=0$ and the necessary specializations and interpretations are given below.

From (1) it follows that

$$F(t)=\begin{cases}0, & 0\leqslant t<t_0,\\ 1-\exp\{-\alpha V(t)\}, & t_0\leqslant t.\end{cases} \tag{2}$$

However, at this stage it is convenient to introduce the conditional d.f.

$$G(t)=\frac{F(t)-F(t_0)}{1-F(t_0)},$$

which is the d.f. of $T|T\geqslant t_0$. Now make the change of variable $R=T-t_0|T\geqslant t_0$, then $0\leqslant R<\infty$ and the new d.f. is

$$\begin{aligned}H(r)&=G(r+t_0)\\ &=[1-F(t_0)]^{-1}[F(r+t_0)-F(t_0)],\\ &=1-\exp\{-\alpha V(t_0+r)+\alpha v_0\}, \qquad 0\leqslant r.\end{aligned} \tag{3}$$

Four stages can be recognized in the clinical development of cancers. These are referred to as S_1, S_2, S_3 and S_4. It is assumed that the various stages are associated with certain volumes of tumor growth. In fact, suppose all cases where $v_0\leqslant V<v_1$ are classified as S_1, $v_1\leqslant V<v_2$ as S_2, $v_2\leqslant V<v_3$ as S_3 and $v_3\leqslant V$ as S_4 where the v_i, $i=1,2,3$ are volume boundaries for the stages. From the definition of S_1 [4], v_1 is known to be approximately 50 cc.

Let the numbers r_i be such that $V(t_0+r_i)=v_i$, $i=0,1,2,3$ ($r_0=0$); then taking V as a monotone increasing function, if P_i is the probability of

reporting the disease in S_i,

$$P_1 = \Pr\{R < r_1\} = 1 - x\{-\alpha v_1 + \alpha v_0\},$$

$$P_2 = \Pr\{r_1 \leqslant R < r_2\} = [\exp\{-\alpha v_1\} - \exp\{-\alpha v_2\}]e^{\alpha v_0},$$

$$P_3 = \Pr\{r_2 \leqslant R < r_3\} = [\exp\{-\alpha v_2\} - \exp\{-\alpha v_3\}]e^{\alpha v_0},$$

$$P_4 = \Pr\{r_3 \leqslant R\} = \exp\{-\alpha v_3 + \alpha v_0\}, \tag{4}$$

by (3).

We use a standard model for tumor growth

$$V(t_0 + r) = v_0 e^{\beta r}, \tag{5}$$

where β is the rate of growth. With this parameterization the final working form of H is

$$H(r) = 1 - \exp\{-\alpha v_0 e^{\beta r} + \alpha v_0\}, \qquad 0 \leqslant r, \tag{6}$$

and the expectation of R, $E[R]$, is

$$E[R] = e^{\alpha v_0}\int_0^\infty \exp\{-\alpha v_0 e^{\beta r}\}\, dr,$$

$$= e^{\alpha v_0}\beta^{-1}E_1(\alpha v_0), \tag{7}$$

where $E_1(x) = \int_x^\infty y^{-1}e^{-y}dy$ is well tabulated [1]. However, αv_0 is extremely small and for small x

$$E_1(x) = -\ln x - 0.5772,$$

which provides the necessary approximation formula.

If the expected tumor volume at reporting is required, this is provided by the formula

$$E[V(R)] = \int_0^\infty v_0 e^{\beta r}\alpha\beta v_0 e^{\beta r} e^{-\alpha v_0 e^{\beta r} + \alpha v_0} dr,$$

$$= \alpha^{-1}(1 + \alpha v_0) \cong \alpha^{-1}. \tag{8}$$

Two points emerge from the above argument. Firstly, if tumor growth is near exponential, then the assumption of (1) is equivalent to requiring the force of reporting, ϕ, to be proportional to tumor volume for $t > t_0$. Secondly, R is the time from when the disease can be first detected clinically. In fact, R is measured from t_0, i.e., from when $V(t) = v_0$ and this is explicitly stated by the formula $R = T - t_0 | T \geqslant t_0$. As is shown by the example below,

t_0 need not be specified.

If it is required to find the distribution of T, v_0 must obviously be put equal to the cell volume of the tissue concerned, when $t_0=0$. For $t_0>0$ the model assigns $\Pr\{0 \leqslant T < t_0\}=0$ and $\Pr\{t_i \leqslant T \leqslant t_{i+1}\}=P_i$ for $i=0,1,2,3$, and $t_i=t_0+r_i$, $r_0=0$, $r_4=\infty$. This conforms with the postulated threshold quality of detection around v_0.

III. APPLICATION

From the records of women with breast cancer in the Central Cancer Registry, Melbourne, Australia [3], estimates of the P_i were as follows;

$$\bar{P}_1=2001/5372=0.3725, \qquad \bar{P}_2=1821/5372=0.3390,$$

$$\bar{P}_3=837/5372=0.1558, \qquad \bar{P}_4=713/5372=0.1327.$$

Using equations (4) estimators of α, v_2, and v_3 for given v_1 are

$$\bar{\alpha}=-\ln(1-\bar{P}_1)/(v_1-v_0),$$

$$\bar{v}_2=-\ln(\bar{P}_3+\bar{P}_4)/\bar{\alpha}+v_0,$$

$$\bar{v}_3=-\ln\bar{P}_4/\bar{\alpha}+v_0.$$

As stated eariler $v_1=50$ cc and, with this value the estimates were, neglecting v_0,

$$\bar{\alpha}=0.00932, \qquad \bar{v}_2=133.4 \text{ cc}, \qquad \bar{v}_3=216.7 \text{ cc}.$$

By (8), $E\overline{[V(R)]} \cong 1/0.00932=107.3$ cc.

A recent paper by Lee and Sprott [2] gives ranges for the doubling time, D.T., for various untreated metastatic lesions of breast cancer. Using the 95% limits from their Table 3 and the formula $\beta^{-1}=\text{D.T.}/\ln 2$, it is found that, for these data, $5 \leqslant \beta^{-1} \leqslant 100$ approximately. If v_0 is taken as 0.5 cc then $E_1(\alpha v_0)=5.37-0.58=4.79$ and, in days, $25 \leqslant E[R] \leqslant 480$.

Since the average delay from first observing the disease to registration is known in a large number of women in the Central Cancer Registry, the average delay for women in S_1 and S_2, $\bar{L}_{12}$, and for women in S_3 and S_4, $\bar{L}_{34}$, can be calculated.Define

$$P_{12}=P_1+P_2, \; P_{34}=P_3+P_4$$

and

$$H_{12}(r)=\begin{cases} P_{12}^{-1}H(r), & 0\leqslant r\leqslant r_2,\\ 1, & r_2\leqslant r,\end{cases}$$

and

$$H_{34}(r)=\begin{cases} 0, & 0\leqslant r\leqslant r_2,\\ P_{34}^{-1}[H(r)-H(r_2)], & r_2\leqslant r,\end{cases}$$

the conditional d.f.'s for R for the pooled groups S_1+S_2 and S_3+S_4. Let

$$E^{12}[R]=\int_0^\infty [1-H_{12}(r)]\,dr,$$

$$=P_{12}^{-1}e^{\alpha v_0}\int_0^{r_2}[e^{-\alpha v_0 e^{\beta r}}-e^{-\alpha v_2}]\,dr,$$

$$=(\beta P_{12})^{-1}e^{\alpha v_0}\int_{\alpha v_0}^{\alpha v_2}[e^{-y}-e^{-\alpha v_2}]\,y^{-1}\,dy,$$

$$=\theta\gamma_{12}$$

where $\theta=\beta^{-1}$ and $\gamma_{12}=P_{12}^{-1}e^{\alpha v_0}[E_1(\alpha v_0)-E_1(\alpha v_2)-e^{-\alpha v_2}\ln(v_2/v_0)]$, and

$$E^{34}[R]=\int_0^\infty [1-H_{34}(r)]\,dr,$$

$$=\int_0^{r_2}dr+\int_{r_2}^\infty e^{\alpha v_2-\alpha v_0 e^{\beta r}}\,dr$$

$$=\theta\gamma_{34}$$

where $\gamma_{34}=\ln(v_2/v_0)+e^{\alpha v_2}E_1(\alpha v_2)$. A moment estimator of θ is now provided by the equation $E^{34}[R]-E^{12}[R]=\bar{L}_{34}-\bar{L}_{12}$, i.e., $\bar{\theta}=(\bar{L}_{34}-\bar{L}_{12})/(\gamma_{34}-\gamma_{12})$. Using $v_0=0.5$, $v_1=50$ and the previous estimates of α and v_2, since $\bar{L}_{12}=0.90$ and $\bar{L}_{34}=1.58$ based on 3353 and 1213 registrants respectively,

$$\bar{\theta}=0.68/1.81=0.376.$$

Moreover, $E[R]=\theta E_1(\alpha v_0)e^{\alpha v_0}$ and an estimate of $E[R]$ is

$$\bar{E}[R]=0.376\times 4.79\times 1.005=1.81 \text{ years.}$$

The above analysis can be applied to the data when classified according to the degree of malignancy. Because of the lack of numbers for tumors

graded as well differentiated and moderate and because the two seemed to have similar delay patterns, they were pooled and then compared with the anaplastic tumor group. The pooled group had delays $\bar{L}_{12}=0.89$ and $\bar{L}_{34}=2.02$ based on 655 and 152 registrants and for anaplastic $\bar{L}_{12}=0.81$ and $\bar{L}_{34}=1.18$ based on 1100 and 313. Thus, for the two slower growing tumors $\theta = 1.13/1.81 \cong 0.624$ and $\bar{E}[R] \cong 3.00$ years, and for anaplastic tumors, $\theta \cong 0.205$ and $\bar{E}[R] \cong 0.98$.

It is interesting to notice that only the anaplastic grade has a growth rate ($\theta = \beta^{-1} = 75$ days) falling within the limits reported by Lee and Sprott [2]. It would appear that these authors were dealing with particularly rapidly growing tumors.

If one takes the tumor cell diameter as about 20 μ, cell volume would be of the order of 10^4 μ^3. Hence, using this value for v_0, $\alpha v_0 \cong 10^{-10}$ cc, and $E_1(\alpha v_0) \cong 22.45$. Thus, using the above estimates of θ for slow and fast growing tumors $E[T]$ is 14.0 years and 4.6 years respectively. The average, over all tumor grades is 8.4 years.

Thus, if one measures the delay to reporting from the time of onset of the disease (i.e., T), it appears that the expected reporting time may be of the order of 5–15 years. If true, this may have an important bearing on the attitude towards early detection of the disease. Of course, the estimates are based on the assumptions of exponential growth of the tumor and a nonvarying rate.

REFERENCES

1 Milton Abramowitz and Irene A. Stegun, *Handbook of Mathematical Functions*, Dover, New York (1964).
2 Yeu-Tsu Lee and John S. Sprott, Rate of growth of soft tissue metastases of breast cancer, *Cancer* 29, 344–48 (1972).
3 D. W. Rankin, The Central Cancer Registry, Melbourne 1940–1970, *Med. J. Aust.* 1, 750–754 (1971).
4 T. N. M. Classification of Malignant Tumors, V. I. C. C., Geneva, 1968.

B[12]

SURVIVAL RATES AS A FUNCTION OF TUMOR VOLUME FOR WOMEN WITH BREAST CANCER

G.M. Tallis*, G. Sarfaty# and P. Leppard*

To appear in

"Mathematical Biosciences"

* Department of Statistics, University of Adelaide, Adelaide, S.A.

Endocrine Research Unit, Cancer Institute, Melbourne, Vic.

ABSTRACT

This paper indicates how the proportion of women suffering from breast cancer, and who have a normal life expectancy, can be estimated from the volume of tumor present at first reporting. These results are then used to predict the effect of screening procedures for early detection. Model predictions are compared with recently published results of a New York screening trial.

I. INTRODUCTION

In a recent paper, (Tallis and Sarfaty, 1974 (TS)), the time delay to first reporting breast cancer was investigated. In (TS), estimates of the tumor volumes associated with particular stagings were obtained.

Breast cancer is classified into four clinical stages, which will be labelled S_1, S_2, S_3 and S_4; S_1 being the least severe. It was suggested that for S_1, the tumor volume V could range from .5ccs to 50ccs; for S_2, $50 \leqslant V \leqslant 133$; for S_3, $133 \leqslant V \leqslant 217$ and for S_4, $217 < V$. By using the density derived for V in that paper in conjunction with the above boundaries, the average tumor volumes for the four stages are estimated as $\bar{v}_1 = 25$ccs, $\bar{v}_2 = 85$ccs, $\bar{v}_3 = 170$cc and $\bar{v}_4 = 325$ccs respectively.

In a subsequent paper, (Tallis, Sarfaty and Leppard, 1975 (TSL)), it was shown how estimates of the proportions of women with normal survival could be obtained for the various stages. Using relevant data from the Victorian Cancer Registry (1973) and a clinical trial undertaken by the Endocrine Research Unit, Cancer Institute, estimates of the proportions of women with normal survival, p_i, were $p_1 = .378$, $p_2 = .268$, $p_3 = .107$ and $p_4 = .018$ for the four stages respectively.

It is the purpose of this paper to show how the above information can be used to design patient examination frequencies for the early detection of small tumor volumes.

II. A MODEL FOR NORMAL SURVIVAL RATES

The first requirement for designing screening frequencies is a model which assigns a normal survival rate to given volumes, V, of tumor. Since we only have four $(p_i, \bar{v}_i)$ pairs, which are

approximations, the best that can be obtained is a set of inspection frequency guidelines suitable for clinical application.

Let $q(v)$ be the probability of normal survival given that $V=v$. We assume the following model :

$$q(v) = \begin{cases} \exp\{-(v/25)^{\delta}\} & 0 \leq v \leq 25 \\ \exp\{-(\gamma_0+\gamma_1 (v/100)+\gamma_2 (v/100)^2)\} & 25>v. \end{cases} \quad (1)$$

The unknown parameters $\gamma_0, \gamma_1, \gamma_2,$ and δ have to be established. To the order of the accuracy required here, $-\ln q(\bar{v}_1) = 1.0$, $-\ln q(\bar{v}_2) = 1.3$, $-\ln q(\bar{v}_3) = 2.2$ and $-\ln q(\bar{v}_4) = 4.0$. We choose γ_0 so that $-\ln \hat{q}(\bar{v}_1) = 1.0$: thus

$$\gamma_0 = 1 - \gamma_1 (25/100) - \gamma_2 (25/100)^2.$$

A plot of $-\ln q(\bar{v}_i)$ against v is given in Figure 1 with the least squares fit

$$-\ln \hat{q}(v) = \hat{\gamma}_0 + \hat{\gamma}_1 (v/100) + \hat{\gamma}_2 (v/100)^2,$$

where $\hat{\gamma}_1 = .51$, $\hat{\gamma}_2 = .14$ and $\hat{\gamma}_0 = 1 - .51(25/100) - .14(25/100) = .86$.

The parameter δ cannot be estimated but must be assigned. We reason as follows. From the slope of $-\ln \hat{q}(v)$, it is unlikely that $\delta>>1$, otherwise $-\ln \hat{q}(v)$ would show unreasonable discontinuity at $v=25$ (see Fig. 1). Possibly $\delta<1$, although the intuitively most reasonable shape for $-\ln \hat{q}(v)$ over the range 0 to 25 is sigmoid, thus providing more of a threshold effect of tumor volume on survival. Thus $\delta=1$ may be a reasonable compromise, and we assume this value from now on.

III. SCREENING DESIGNS

The aim of large population screening of women for breast cancer is to detect tumors at an earlier stage with the consequent possibility of a higher rate of normal survival. At the first screening, women are examined by multiple procedures, and we foresee that a volume v_C can be assigned such that all women with tumors of size greater than v_C will be detected.

After the first screening then it is anticipated that all women with breast tumors greater than v_C will have been identified and treated. Now define t_C by the equation $V(t_C)=v_C$; if V is strictly monotone then t_C is uniquely defined. Moreover, if the time interval between successive screenings is t_I, the volume of tumors detected after the initial screening should lie between $V(t_C)=v_C$ and $V(t_C+t_I)=v_I$, say. The situation is illustrated in Figure 2.

In order to calculate the average volume of tumors detected by the screening we make two assumptions :

(a) $V(t)=v_0\exp(\beta t)$, where v_0 is the initial cell volume at $t=0$;

(b) the average tumor volume of those women detected by the screening procedure is essentially $V(t_C+t_I/2)$.

Let R be the time delay in reporting the disease, (see (TS)). Then the value of R for those women detected by the screening procedure must obviously be greater than t_C+t_I, and hence we use the notation

$$E_{CI}[\,V \mid R \geq t_C+t_I] = V(t_C+t_I/2)$$

Note that this is the expected value of V for those women detected at the time of screening.

In order to estimate the tumor volume associated with a particular set of screening parameters v_C and t_I, two situations must be considered :

(1) a proportion of cases report before screening;

(2) the remaining cases are detected at screening.

Thus, in usual notation,

$$E_{CI}[V] = E[V|R \leq t_C+t_I]Pr\{R \leq t_C+t_I\} + E[V|R \geq t_C+t_I]Pr\{R \geq t_C+t_I\}$$

$$= [1-\exp(-\alpha v_I)]/\alpha - v_I\exp(-\alpha v_I) + \exp(-\alpha v_I)V(t_C+t_I/2$$

$$= \bar{v}_{CI} \text{ , say,}$$

using the densities of V and R proposed in (TS), where estimates of about .01 for α and 2.66 for β were obtained.

The proportion of women with normal survival for a particular screening design is estimated by $q(\bar{v}_{CI})$. The values of this function are given for various v_C and t_I in Table 1. Here, the critical dependence of the average chance of survival on the frequency of inspection and the tumor volume is emphasise

Although these findings may be intuitively obvious, the modelling given here allows the influence of v_C and t_I on $q(\bar{v}_{CI})$ to be examined quantatively.

IV. EXAMPLE

We compare the above modelling with experimental results. Shapiro et al (1973) reported the five year results of a screening study in New York. They compared the 1-5 year survival proportions, $P(t)$, of a sample of screened women against those of a non-screened control group. Their results are summarised in the last two columns of Table 2.

In order to approximately model their situation we write;

$$P(t|\bar{v}_{CI}) = \Pr\{\text{death within } t \text{ years of reporting}\}$$
$$= q(\bar{v}_{CI})P_C(t) + [1-q(\bar{v}_{CI})]P_{NC}(t) \qquad (2)$$

where $P_C(t)$ and $P_{NC}(t)$ are the probabilities of death within t years of reporting given that a woman is "cured" and "not cured" respectively.

The average age of the women in the New York study was 55 and $P_C(t)$ was calculated directly from The Australian Life Tables (1962) for women aged 55. Using (2), $q(\bar{v}_{CI}) = .25$ as estimated in (TSL) for the average cure rate over all stages at reporting, and $P(t)-P(t-1)$ given in Tallis, Sarfaty and Leppard (1973, page 64, column 2), it was possible to solve for an estimate of $P_{NC}(t), t=1,2,..,5$. These figures are also recorded in Table 2.

In the Shapiro study $t_I=1.5$ but the value of v_C is not stated. Table 2 lists the values of $P(t|\bar{v}_{CI})$ calculated with values of $v_C=.5, 2, .25$. A comparison of these estimates and the New York findings is rewarding. Although the two populations appear to differ in that the standard mortality figures for Victoria are slightly higher, and in spite of the approximate methods used in the modelling and estimation, the quite reasonable value of $v_C= 1.0$ gives estimates of $P(t)$ which are in good agreement with the New York figures.

REFERENCES

Australian Life Tables (1962)
Commonwealth Bureau of Census and Statistics.

Shapiro, S., Strax, P., Venet, L. and Venet, W. (1973)
Changes in 5-year breast cancer mortality in a breast cancer screening program.
Proc. Nat. Can. Conf., 663-678.

Tallis, G.M., Sarfaty, G. and Leppard, P. (1973)
The use of a probability model for the construction of age specific life tables for women with breast cancer.
University of Adelaide and Cancer Institute, Melbourne.

Tallis, G.M. and Sarfaty, G. (1974)
On the distribution of the time to reporting cancers with application to breast cancer in women.
Mathematical Biosciences, 19, 371-376.

Tallis, G.M., Sarfaty, G. and Leppard, P. (1975)
The random introduction of patients into a clinical trial in relation to estimates of survival.
To be published.

TABLE 1

Estimated proportion of normal survivors for various screening designs

V_C (ccs)	t_I (years)			
	.50	1.00	1.50	2.00
.50	.96	.93	.80	.36
.75	.94	.89	.68	.32
1.00	.93	.85	.57	.29
1.25	.91	.81	.47	.27
1.50	.89	.78	.39	.25
1.75	.87	.73	.36	.23
2.00	.86	.70	.35	.23

TABLE 2

Comparison of P(t) for Victorian data and New York study

t	RESULTS OBTAINED FROM VICTORIAN DATA									SHAPIRO ET. AL.	
	P(t) Control	$P_C(t)$	$P_{NC}(t)$	$P(t\|\bar{v}_{CI})$ for t_I=1.5 and v_C tabulated Screened						P(t) Control	P(t) Screened
				.50	.75	1.00	1.25	1.50	1.75	2.00	
1	.11	.01	.14	.04	.05	.07	.08	.09	.09	.09	.08	.08
2	.24	.01	.32	.07	.11	.14	.17	.20	.21	.21	.22	.14
3	.33	.02	.43	.10	.14	.20	.24	.26	.28	.29	.28	.19
4	.41	.03	.54	.13	.19	.25	.30	.34	.36	.36	.37	.23
5	.48	.04	.63	.16	.23	.29	.35	.40	.42	.42	.42	.28

FIGURE 1

Least square estimate of q(v)

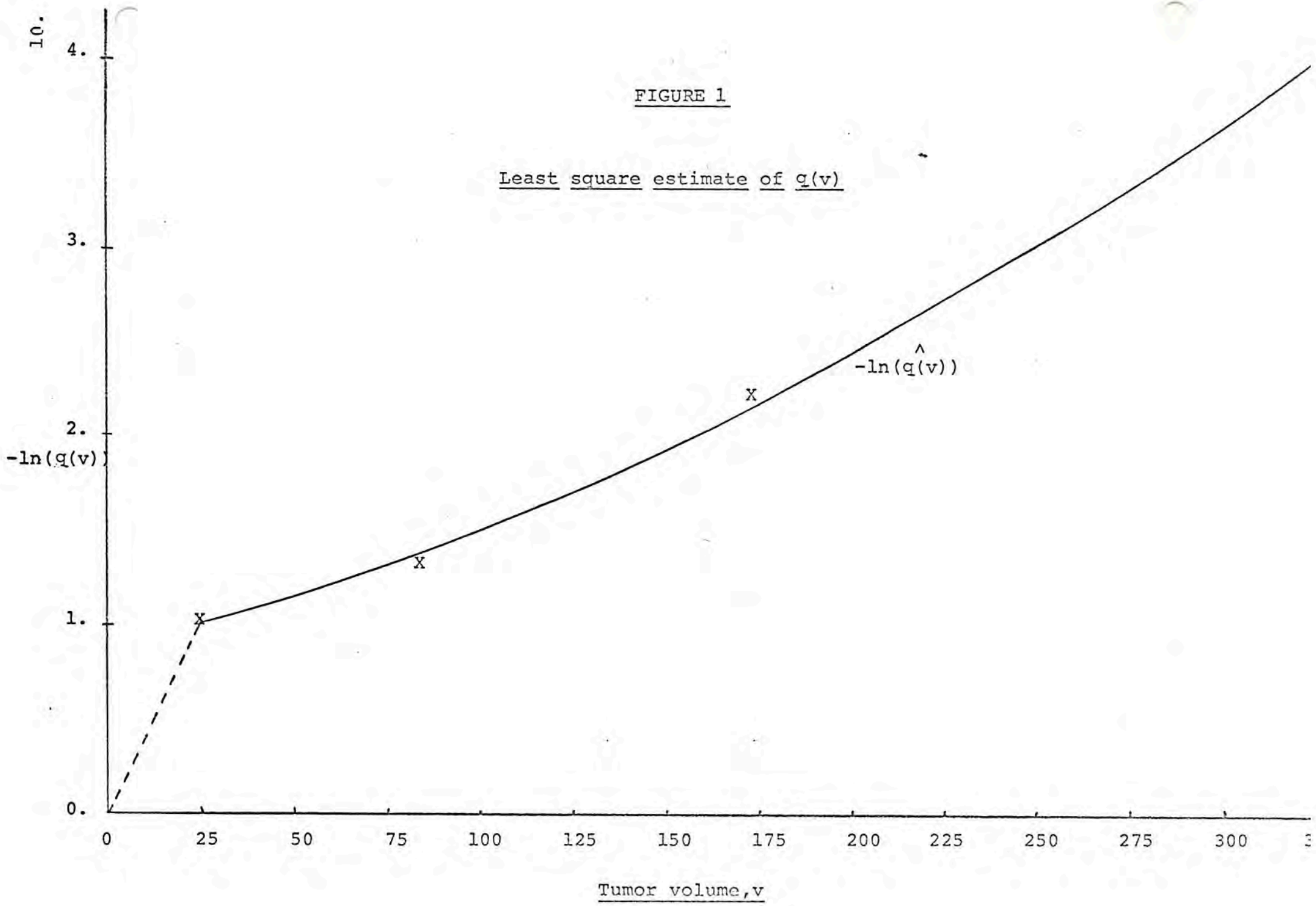
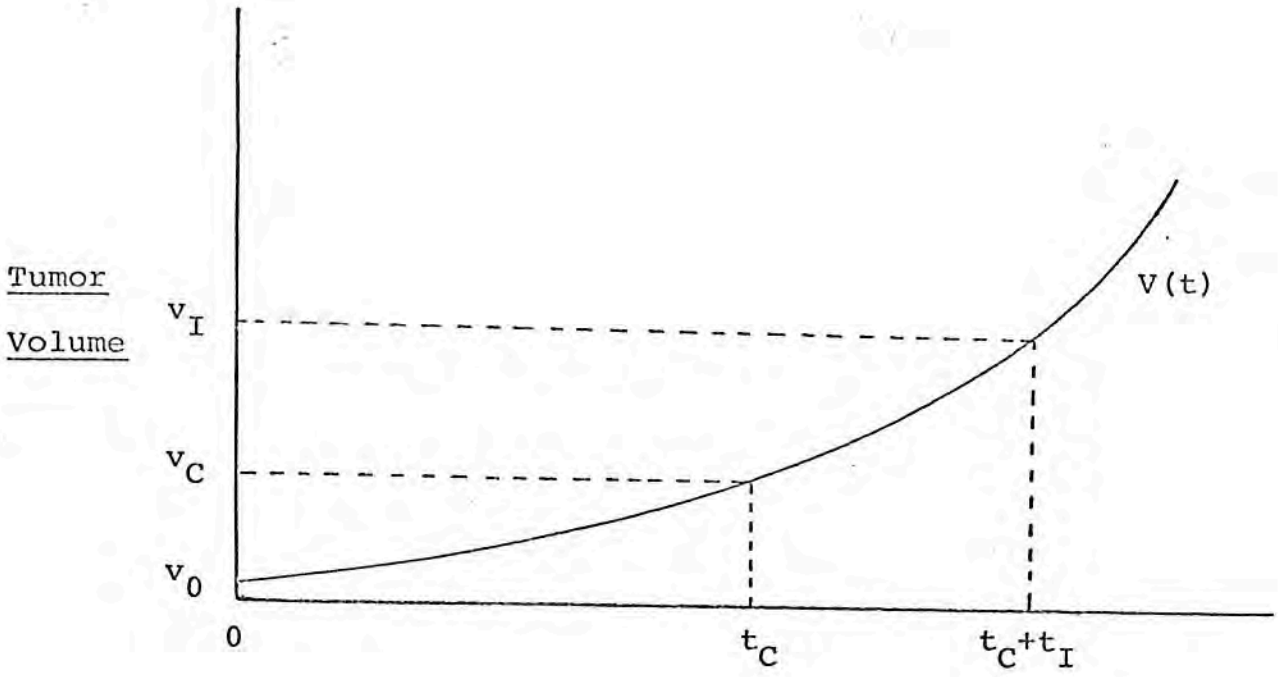


FIGURE 2

Relationship between tumor volumes and time intervals



B[13]

THE RANDOM INTRODUCTION OF PATIENTS INTO A CLINICAL TRIAL: VARIOUS ASPECTS OF ESTIMATES AND PROBABILITIES OF SURVIVAL

BY

G.M. TALLIS*, G. SARFATY# and P. LEPPARD*

To appear in

"Computers and Biomedical Research"

* Department of Statistics, University of Adelaide, Adelaide, S.A.

Endocrine Research Unit, Cancer Institute, Melbourne, Vic.

ABSTRACT

A feature of some medical studies is that patients enter the trial at different times. At the close of the trial they have been observed for varying periods and any analysis of survival must consider this feature of random entry.

We propose a simple model to cope with the above situation. One advantage of this approach is that it allows for the model assumptions to be tested statistically. The model is illustrated with the survival data from a breast cancer investigation conducted in Victoria, Australia.

Survival is studied further and is related to the response to the treatment of adrenalectomy for women with advanced disease. This work bears on previous results reported here (TSL).

The idea that a certain proportion of the women who report breast cancer have a normal life expectancy is developed. This proportion is examined in the four clinical stages of the disease and is found to depend dramatically on this staging.

Further results which are developed are an estimate of the tumour volume at death and personalised, conditional probabilities and expectations of survival following adrenalectomy. The latter functions have been programmed for use in a clinical context.

I. INTRODUCTION

A medical trial involving 131 women in advanced stages of breast cancer was initiated at the Endocrine Research Unit, Cancer Institute, Melbourne, in September 1967. Women entered the trial following adrenalectomy and they were subsequently classified as to whether or not they showed clinical remission to the disease. Response to the operation has been assessed in relation to certain hormones and other physiological factors. Details of the variables are given in Sarfaty, Pitt, Tallis (1973) and the model of response as a threshold phenomena (TSL) is discussed in Tallis, Sarfaty and Leppard (1975).

A feature of this type of continuing study is that patients enter the trial at different times and at the close of the trial have been observed for varying periods. The length of survival, Y, is an important assessment in the study and a way therefore must be developed to deal with the fact that the value of Y for some or all patients may be unknown at any fixed time, t, say at the close of the study. Some patients who entered the trial at an early stage may be dead; others may still be alive. The chance that they are still alive at t increases as their time of entry approaches t.

Sections II, III and IV of this paper are devoted to establishing a simple model for this type of truncation and to showing how estimates of the required parameters can be found. The techniques are applied to the survival data of the adrenalectomy trial and illustrate the effect of truncation caused by the random method of introducing patients.

In Section V the estimated survival time is used to investigate the disease further. Response to treatment has been postulated as a continuous variable X_0, with clinical remission a threshold phenomenon (TSL). The relationship of $W=\ln Y$, which is assumed to be normally distributed, to X_0 is pursued in part (a). If the vector $\underset{\sim}{X}_0'=(X_1,..,)$ represents the physiological variables of interest, then from the work in this section, and from previous results, the joint distribution of

W, X_0 and $\underset{\sim}{X}$ can be found. Relevant correlations are reported in Tab II. The expectations of survival for the clinical groups "remission" and "non-remission" are also obtained.

Part (b) of this section develops the idea that a certain proportion of women who report breast cancer have a normal survival or life expectancy. The proportion of women having normal survival is examined in the four clinical stages of the disease, and is found to depend dramatically on the staging.

From the above results it is also possible to estimate the volume of tumor at death, and this is dealt with in part (c).

In part (d) the general results relating survival following adrenalectomy as a function of $\underset{\sim}{X}=\underset{\sim}{x}$ have been used to obtain "personalised" conditional expectations. These in conjunction with individu probabilities of clinical remission to the disease can aid the decisi of whether or not to perform an adrenalectomy.

The results given in this paper illustrate the various statistical models which have been designed specifically to extract the required information.

II. MODEL AND NOTATION

Consider a trial in which subjects enter randomly over a period θ. The results are assessed in terms of survival at a time h after the trial concludes. i.e. at $\theta+h$, where $h\geq 0$. Let L be the time from entry to assessment and Y the length of survival from entry, with distribution functions $B(\ell)$ and $F(y)$ respectively. Note that $B(\ell)=0$ for $\ell<h$. Moreover, let $\underset{\sim}{X}$ be a random m-vector which is to be used in the predictor $E[Y|\underset{\sim}{X}=\underset{\sim}{x}]$. Denote the marginal distribution function of $\underset{\sim}{X}$ by $K(\underset{\sim}{x})$ and the joint distribution function of $\underset{\sim}{X}$

and Y by $R(\underset{\sim}{x},y)$. The variable L is assumed to be independent of both $\underset{\sim}{X}$ and Y.

Because θ and h are finite time lengths it is not always possible to observe Y directly. What can be observed, however, is $V=\min[Y,L]$, and we require the joint distribution of $\underset{\sim}{X}$ and V. Thus, if $\underset{\sim}{X}\leqslant\underset{\sim}{x}$ represent $X_1\leqslant x_1, X_2\leqslant x_2, \ldots, X_m\leqslant x_m$, then

$$\begin{aligned} G(\underset{\sim}{x},v) &= \Pr\{\underset{\sim}{X}\leqslant\underset{\sim}{x}, V\leqslant v\} \\ &= \Pr\{\underset{\sim}{X}\leqslant\underset{\sim}{x}, Y\leqslant v, L\leqslant v\} + \Pr\{\underset{\sim}{X}\leqslant\underset{\sim}{x}, Y\leqslant v, L>v\} + \Pr\{\underset{\sim}{X}\leqslant\underset{\sim}{x}, Y>v, L\leqslant v\} \\ &= 1 - \Pr\{\underset{\sim}{X}\leqslant\underset{\sim}{x}, Y>v, L>v\} \\ &= 1 - \Pr\{\underset{\sim}{X}\leqslant\underset{\sim}{x}, Y>v\}.\Pr\{L>v\} \\ &= 1 - [K(\underset{\sim}{x})-R(\underset{\sim}{x},v)][1-B(v)] \end{aligned} \tag{1}$$

The marginal distribution function of V is then

$$P(v) = G(\underset{\sim}{x}=\underset{\sim}{\infty},v) = 1 - [1 - F(v)][1 - B(v)] \tag{2}$$

III. APPLICATION OF THE MODEL

The records from September 1967 to November 1972 of the trial group of 131 were analysed in April 1973. Thus, $\theta=5.16$ years and $h=.42$ years.

In fact, entry into the trial was approximately uniform : hence it is assumed

$$B(\ell) = \begin{cases} 0 & 0\leqslant\ell\leqslant h \\ 1 - \dfrac{\theta+h-\ell}{\theta} & h<\ell\leqslant\theta+h \\ 1 & \ell>\theta+h \end{cases} \tag{3}$$

and Y, the length of survival following adrenalectomy, is lognormal with density

$$f(y;\gamma,\tau) = F'(y;\gamma,\tau) = \exp\{-(\ln y - \gamma)^2/2\tau^2\}/(\sqrt{2}\pi\tau y),\ y\geqslant 0. \tag{4}$$

This latter density was chosen because of the positive and skewed nature of Y. As in (TSL) the vector $\underset{\sim}{X}$ consists of 11 transformed

measurements where the transformation has been chosen so that $\underset{\sim}{X} \sim MVN(\underset{\sim}{\mu}, \Sigma$

Following the assumption (4), it is more convenient to consider the variate $W=\ln Y$ rather than Y. The joint distribution of W and $\underset{\sim}{X}$ is again assumed to be multinormal; that is

$$R(\underset{\sim}{x},w) = MVN\left(\begin{bmatrix} \underset{\sim}{\mu} \\ \gamma \end{bmatrix}, \begin{bmatrix} \Sigma & \underset{\sim}{c} \\ \underset{\sim}{c}' & \tau^2 \end{bmatrix} \right) \qquad (5$$

and the joint distribution of any X_i and W (needed at a later stage) is then bivariate normal and is denoted by

$$r_i(x,w;\ \mu_i,\gamma,\sigma_{ii},\tau^2,c_i). \qquad (6$$

IV. ESTIMATION

The parameters of the model in the preceding section were estimated in a variety of ways, with all numerical calculations carried out on the University of Adelaide CDC6400 in FORTRAN. The parameters $\underset{\sim}{\mu}$ and Σ were estimated in the usual manner from the 131 sample values of $\underset{\sim}{X}$. In order to estimate γ, τ and $\underset{\sim}{c}$ the density of V and the joint density of V and $X_i, i=1,\ldots,11$, are needed.

By differentiating (2)

$$p(v) = P'(v) = f(v)[1-B(v)] + [1-F(v)]b(v),\ 0 \leqslant v \leqslant \theta+h, \qquad (7$$

where B and f are defined in (3) and (4) respectively, and $b(\ell)=B'(\ell)$. Since p(v) is a function of γ and τ only, $p(v;\gamma,\tau)$ say, these parameters were estimated by maximising the likelihood

$$L(\gamma,\tau;\underset{\sim}{v}) = \prod_{i=1}^{131} p(v_i;\gamma,\tau)$$

for γ and τ using the Nelder and Mead (1965) simplex method. The estimates obtained are $\hat{\gamma}=-.1108$ and $\hat{\tau}=1.1937$. These values were then used to calculate the mean and variance of V by numerically integrating

$$\int_0^{\theta+h} v.p(v;\hat{\gamma},\hat{\tau})dv \text{ and } \int_0^{\theta+h} v^2 p(v;\hat{\gamma},\hat{\tau})dv.$$

The results,with corresponding sample estimates,are shown in Table I.

As an alternative fitting procedure,γ and τ were also estimated by using a minimum χ^2 procedure developed by Tallis (1973). This method measures,in terms of a χ^2 value,the difference between sample percentiles and percentiles obtained from the hypothesised distribution function. Using the specialised form of the distribution function (2),a minimum χ^2 value of 13.1 on 7 d.f. was obtained and this is not significant at the 5% level. The estimates of γ and τ,γ^* and τ^* say,were $\gamma^*=-.07$ and $\tau^*=1.38$. Although we continue to use the maximum likelihood estimates in subsequent work,the non-significant value and the reasonably close agreement between the two sets of estimates gives confidence in both the estimates and the model.

The density (4) was also used directly to obtain values for γ and τ by maximising the likelihood

$$\prod_i f(v_i;\gamma,\tau) . \prod_j [1-F(v_j;\gamma,\tau)]$$

for γ and τ. The products over i and j respectively consist of those women who died during the trial and those who were still alive at the end. This method has been used by others(Armitage 1966). The numerical estimates were again in good agreement with those stated above,but there is a major advantage in working directly with p(v) since the agreement of the model with the data can be tested.

The estimation of the vector $\underset{\sim}{c}$ (as a vector) was found to be computationally impossible. However,by applying the principle leading to (1),the joint density of each X_i and V is,in the notation of (6)

$$t_i(x,v) = r_i(x,v)[1-B(v)] + b(v)n(x;\gamma,\tau^2)[1-N(\ln v;a,s^2)] \qquad (8$$

where $a = \gamma+c_i\tau(x-\mu_i)/\sqrt{\sigma_{ii}}$, $s = \tau\sqrt{1-c_i^2}$,and B and $r_i$are given by (3)

and (6) respectively. The density and distribution functions of a variable Z distributed normally with mean α and variance β will be written as $n(z;\alpha,\beta)$ and $N(z;\alpha,\beta)$. For the special case $\alpha=0$ and $\beta=1$, the standard notation $\phi(z)$ and $\Phi(z)$ are used.

Each c_i was then estimated by maximising a conditional likelihood function of the form

$$L(c_i|\hat{\mu}_i,\hat{\gamma},\hat{\sigma}_{ii},\hat{\tau}^2) = \prod_{i=1}^{131} t_i(x_{ij},v_j;\hat{\mu}_i,\hat{\gamma},\hat{\sigma}_{ii},\hat{\tau}^2)$$

with respect to c_i, i=1,..,11. These estimates are given in Table II.

V. SOME APPLICATIONS OF THE ESTIMATES

(a) Remission and Survival

In (TSL) we assumed that remission after adrenalectomy is a continuous variable and it was labelled X_0. If X_0 is greater than some fixed quantity a, a clinical remission of the disease is recorded and if $X_0 \leqslant a$ then no significant reduction in the rate of progress of the tumor is noticed. The estimate of a was $\hat{a}=.466$. We have obtained above $\text{cor}(W,X_i)$, i=1,..11, and it is also possible to estimate $\rho=\text{cor}(W,X_0)$ under the assumption that these variables have a binormal distribution.

For each woman we have the information V=v and whether or not she has had a remission; that is $X_0 \leqslant a$ or $X_0 > a$. Therefore the appropriate density for this situation is obtained from (8) as

$$t_0^*(v;\rho) = \begin{cases} \int_{-\infty}^{a} t_0(x,v)\,dx & \text{non-remitters} \\ \int_{a}^{\infty} t_0(x,v)\,dx & \text{remitters} \end{cases} \qquad (9)$$

where $0 \leqslant v \leqslant \theta+h$, and ρ is the only unestimated parameter. A likelihood function using the observed sample values of V, the known clinical class

ifications and the density of (9) was maximised for ρ, with $\hat{\rho}$ shown in Table II.

To obtain the expectation of Y, the survival time following adrenalectomy, we use the formula

$$\mu_Y = E[Y] = E[\exp\{W\}] = \exp\{\gamma+\tau^2/2\}$$

where γ and τ are defined in (4). Using the values $\hat{\gamma}=-.1108$ and $\hat{\tau}=1.193$ the estimate of μ_Y is $\hat{\mu}_Y=1.83$. This can be compared with $\bar{v}=1.15$ (Table I) and emphasises the necessity of correcting the data for truncation effects.

It is also possible to estimate the expectation of Y given that a remission occurs. Specialising the methods for truncated multinormal distributions of Tallis (1961), formula (2) with $b_1=-\infty, b_2=a-\rho\tau$, we get

$$\begin{aligned}\mu_{Y|R} &= E[Y|\text{Remission}] \\ &= E[Y|X_0>a] \\ &= \exp\{\gamma+\tau^2/2\}[1-\Phi(a-\rho\tau)]/[1-\Phi(a)]. \qquad (10)\end{aligned}$$

This follows since we seek the expectation of exp{W} over the appropriate region.

The expected value of Y given a non-remission, $\mu_{Y|NR}=E[Y|X_0\leq a]$ is then obtained from the relationship

$$E[Y] = \Phi(a)E[Y|X_0\leq a] + [1-\Phi(a)]E[Y|X_0>a]. \qquad (11)$$

Using the relevant parameter estimates obtained earlier, numerical estimates of $\mu_{Y|R}$ and $\mu_{Y|NR}$ are $\hat{\mu}_{Y|R}=3.98$ and $\hat{\mu}_{Y|NR}=.81$.

(b) The Probability of Normal Survival

For women of age T, the population expectation of survival from the time of first clinical diagnosis of breast cancer, D_T, is given in the life tables by Tallis, Sarfaty and Leppard (1973) (hereafter called

the tables). We assume that this population is a mixture of those women who have a normal life expectancy, NLE, and those who do not. The following model then holds

$$E[D_T|BC] = pE[D_T|NLE] + (1-p)E[D_T|\text{not NLE}],\ 0 \leq p \leq 1. \quad (12$$

In (12), BC≡breast cancer, $E[D_T|BC]$ is the population expectation of survival for women with breast cancer, $E[D_T|NLE]$ is the normal life expectancy for women of age T, and $E[D_T|\text{not NLE}]$ is the expectatio of all women who become eligible for adrenalectomy. Both $E[D_T|BC]$ and $E[D_T|NLE]$ can be found from the tables. To obtain an estimate of $E[D_T|\text{not NLE}]$ we use the data on women having an adrenalectomy as described in Section III. Note that we have assumed that this grou is representative of the population of women with breast cancer who do not have a normal life expectancy.

Specifically, for these women D_T is the sum of two intervals; diagnosis to adrenalectomy, I_T, and adrenalectomy to death, Y_T. The joint distributions of I_T and T, and Y_T and T were assumed to be bivariate lognormal, thus allowing the calculation of the conditional expectation at age T,

$$E[D_T|\text{not NLE}] = E[I_T] + E[Y_T] \quad (13$$

The practical details relating to the fitting and testing of the above joint distributions will be omitted, since standard methods only were used.

The estimated correlation between Y_T and T is essentially 0 as it must be by intuition, since the disease and not age is the prime cause of death for women in this trial.

Equation (12) has been solved for T=50, the nearest tabular value to the average age of the group, and the estimate of p is .249. (se Table III). Moreover, since the women in the trial had first reported with their disease in one of four clinical stages, stage four being th

most advanced, formula (12) could be applied to these groups. The values of $E[D_T|BC]$ by stage are only available for T=60, requiring the calculation of $E[D_{60}|$ not NLE] by stage, see Table III.

Notice that the estimate of p on the combined group at T=60 compares favourably with that at T=50, giving confidence in the model and age correcting procedure.

(c) Final Tumor Volume at Death

In previous work (Tallis and Sarfaty (1974) (TS)), tumor volumes were estimated for each of the four clinical stages of the disease. As a further application of the above results we can estimate the tumor volume at death.

It has been assumed above that W and X_0 have a bivariate normal distribution $BVN(w,x;\gamma,0,\tau^2,1,\rho)$, and estimates $\hat{\gamma}=-.1108$, $\hat{\tau}=1.1937$ and $\hat{\rho}=.85$ obtained. Thus the expected survival for a given response x_0 is

$$E[Y|x_0] = t(x_0) = \exp\{\gamma+\rho x_0+\tau^2(1-\rho^2)/2\} \qquad (14)$$

Now $t(x_0)$ will be in some way related to the rate of tumor growth, $\beta(x_0)$, given an approximately constant tumor volume, v_d, at death for all patients. Then if v_4 is the tumor volume on entry to stage four, and assuming exponential growth,

$$v_4.\exp\{\beta(x_0).t(x_0)\} \overset{x_0}{\equiv} v_d. \qquad (15)$$

Using the previously estimated values for the parameters in (14), we find $t(x_0)=1.091\exp\{\rho x_0\}$, and the only way in which (15) can be satisfied is if $\beta(x_0)=\beta\exp\{-\rho x_0\}$ for some fixed constant $\beta>0$.

We choose β so that $E_{NR}[\beta(X_0)]=2.7$, the average rate of tumor growth estimated in (TS) and which agrees with values quoted elsewhere. Since

$$E_{NR}[\beta(X_0)] = \beta\int_{-\infty}^{a}\exp\{-\rho x_0\}\phi(x_0)dx_0/\Phi(a) = \beta\exp\{\rho^2/2\}\Phi(a+\rho)/\Phi(a) \qquad (16)$$

then substituting the various estimates in (16) gives $\hat{\beta}=1.41$. Thus, finally we get $\hat{v}_4 \exp\{1.091 \times 1.41\} = \hat{v}_d$, and since $\hat{v}_4=217$ from (TS), then $\hat{v}_d=1006$ccs.

(d) Individual Probabilities of Survival and Life Expectations.

Methods were developed in (TSL) to estimate conditional probabilities of showing a clinical remission to adrenalectomy given $\underset{\sim}{X}=\underset{\sim}{x}$. It is of further interest to know the conditional expectation of survival after the operation given $\underset{\sim}{X}=\underset{\sim}{x}$. i.e. $E[Y|\underset{\sim}{x}]$ where Y is defined in Section II.

Since we have estimated all the parameters for $\underset{\sim}{X}$ and W under the assumption of multinormality, $E[Y|\underset{\sim}{x}]$ is easily calculated by standard regression theory (Rao 1965) and also the conditional distribution of $W|\underset{\sim}{x}$, where $Y=\exp\{W\}$.

In the same way, for fixed $\underset{\sim}{X}=\underset{\sim}{x}$, we can calculate $P(i|j)$, the probability of dying within i years after adrenalectomy given the patient has already survived j years.

The formula for $E[Y|\underset{\sim}{x}]$, $P(i|j)$ and $P(\underset{\sim}{x})$ from (TSL) have been programmed so that individual prognoses to adrenalectomy can be obtained. A sample patient output is given in Table IV. This procedure is currently under test at the Endocrine Research Unit, Cancer Institute, Melbourne.

REFERENCES

ARMITAGE, P. and ZIPPIN, C. (1966).
Use of concomitant variables and incomplete survival information in the estimation of an exponential survival parameter.
Biometrics 22, 665-672.

NELDER, J.A. and MEAD, R. (1965).
A simplex method for function minimization.
Comp. J. 7, 308-313.

RAO, C.R. (1965).
Linear Statistical Inference and its Applications.
Wiley, New York.

SARFATY, G., PITT, P. and TALLIS, G.M. (1973).
Basic results of a study of bilateral adrenalectomy for advanced breast cancer.
Med. J. Aust. 2, 877-881.

TALLIS, G.M. (1961).
The moment generating function of the truncated multinormal distribution.
J.R. Statist. Soc. B., 23, 223-229.

TALLIS, G.M. (1973).
Goodness of fit tests for a class of asymptotic normal estimators.
Univ. of Adelaide, Dept. of Stats., Tech. Paper 3.

TALLIS, G.M., SARFATY, G. and LEPPARD, P. (1973).
The use of a probability model for the construction of age specific life tables for women with breast cancer.
The University of Adelaide and The Cancer Institute, Melb.

TALLIS, G.M. and SARFATY, G. (1974).
On the distribution of the time to reporting cancers with application to breast cancer in women.
Math. Bios. 19, 371-376.

TALLIS, G.M., SARFATY, G. and LEPPARD, P. (1975).
A general classification model with specific application to response to adrenalectomy in women with breast cancer.
Comp. and Biomed. Res. (to appear).

TABLE I

Mean and variance of V

	Mean	Variance
Sample	1.15	1.11
Model	1.17	1.22

TABLE II

Correlation table for the X_i and W

X_0	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}	X_{11}	W
1	-.09	.24	.39	.44	.36	.30	.39	.39	.30	.17	.27	.85
	1	-.15	-.24	-.25	-.12	-.10	-.21	-.15	-.14	-.08	-.47	-.06
		1	.20	.17	.16	.09	.23	.13	.14	.05	-.10	-.14
			1	.77	.46	.56	.82	.49	.56	.13	.13	-.16
				1	.53	.64	.79	.49	.42	-.11	.06	.01
					1	.40	.51	.69	.62	-.05	-.02	.14
						1	.67	.50	.36	-.10	-.03	.00
							1	.53	.59	-.05	.10	.12
								1	.71	-.12	.04	.20
									1	-.07	.10	.12
										1	-.21	.07
											1	.29
												1

TABLE III

Estimated values of p

GROUP	T	$E[I_T]$	$E[Y_T]$	E[BC]	E[BC|NLE]	p
Combined	50	4.12	1.82	11.37	27.72	.249
Combined	60	2.65	1.75	8.22	19.42	.254
Stage 1	60	4.77	1.71	11.37	19.42	.378
Stage 2	60	2.76	1.76	8.52	19.42	.268
Stage 3	60	1.58	1.74	5.03	19.42	.107
Stage 4	60	.56	1.73	2.62	19.42	.018

TABLE IV

An individual prognosis to adrenalectomy

SUBJECT NO.= 7777

VARIATES PRESENT

X 1 = .99
X 2 = -.05
X 3 = -.98
X 4 = -.71
X 5 = 2.16
X 6 = -.53
X 7 = -.51
X 8 = 2.33
X 9 = 2.70
X10 = .64
X11 = .95

MULTIPLE CORRELATION (X0,X1,...,X11) = .7121
MULTIPLE CORRELATION (X1,...,X11,W) = .6546

PROBABILITY OF REMISSION = .885

POST ADRENALECTOMY LIFE TABLE

E(D,J)=EXPECTED TIME TO DEATH,GIVEN J YEARS SURVIVAL AFTER ADRENALECTOMY

P(I,J)=PROBABILITY OF DEATH WITHIN I YEARS,GIVEN J YEARS SURVIVAL AFTER ADRENALECTOMY

J	E(D,J)	P(1,J)	P(2,J)	P(3,J)	P(4,J)	P(5,J)	P(10,J)
0	11.158	.0131	.0730	.1576	.2465	.3306	.6292
1	10.297	.0607	.1464	.2364	.3216	.3987	.6639
2	9.926	.0912	.1871	.2778	.3599	.4325	.6791
3	9.871	.1055	.2053	.2956	.3755	.4455	.6825
4	9.977	.1116	.2125	.3018	.3801	.4484	.6800
5	10.168	.1136	.2141	.3022	.3791	.4461	.6744

B[13]
Addendum

ACCESSING RECORD SYSTEMS ACCORDING TO DATE OF DEATH.

G.M. Tallis

Introduction

In on-going clinical trials, or in any continuously up-dated record bank, patients enter the system sequentially. If length of survival is an important variable for study, some problems in obtaining unbiased estimates of, say, the expected survival time may be experienced.

Suppose records have been kept over a period $[o,b]$, then some patients will have entered the system and may have died during the interval of observation, and others may have survived. Clearly, accessing the records via the time of entry into the system will lead to complications in any analysis of survival since the incomplete data on surviving patients must be dealt with somehow or another.

An obvious question, therefore, is; what sort of trouble does one experience if patients are accessed via their time of death? Intuitively, it seems that any snags in the analysis may be a direct consequence of the patient input scheme.

This problem is examined briefly in this note. It is found that an unbiased estimate of the density of survival time is obtainable if and only if patients enter the system according to a uniform distribution, and the system is sufficiently old.

Results

According to the introduction, patients enter the system during a period $[o,b]$. Moreover, all the records are inspected over an interval $[a, b]$, $0 < a < b$. We note that b may be large, as would be the case, for instance, in some registry systems. All patients who died in $[a,b]$ are identified and their survival time, Y, calculated. If Y has a (population) density $g(y)$, we seek conditions under which the above sampling is in fact from $g(y)$.

Let the time that patients enter the system, X, have density $f(x)$ and assume X and Y independent. The latter assumption seems reasonable in any situation which is essentially stationary.

Now the method of accessing implies that for each patient examined, $a \leqslant X + Y \leqslant b$; the original sample space is $[o,b] \times [o,\infty)$ and we look at the restricted space $R = \{x,y;\ a \leqslant x+y \leqslant b\}$, the joint density of X and Y in R being

$$\phi(x,y|R) = f(x)g(y)/K(a,b)$$

where $K(a,b) = \iint_R f(x)g(y)dxdy = \int_o^b [F(b-y)-F(a-y)]g(y)dy.$

The conditional density of Y (in R) is

$$g(y|R) = \int_{a-y}^{b-y} \phi(x,y|R)dx = g(y)[F(b-y) - F(a-y)]/K(a,b)$$

What is required is for $g(y|R) = g(y)$, i.e.

$$F(b-y) - F(a-y) = K(a,b) \qquad (*)$$

for all $a < b$ and y.

Suppose (*) holds, then put $y=o$ to get $K(a,b) = F(b)-F(a)$. Now put $y=a$ and use $F(o)=o$ to show that $F(b-a) = F(b)-F(a)$. This is a Cauchy functional equation with a unique solution $F(x) = cx$. Clearly $c = b^{-1}$ and

$$b^{-1}(b-a) = K(a,b) = \int_0^b b^{-1}[(b-y)-(a-y)]\, g(y)dy = b^{-1}(b-a)G(b)$$

and hence $G(b) = 1$.

Thus a necessary condition for $g(y|R) = g(y)$ is that

(1) $f(x) = 1/b$

(2) $G(b) = 1$

and this condition is also obviously sufficient.

From a practical viewpoint the upshot is clear. Provided the record system has been in existence considerably longer than reasonable survival experience ($G(b)=1$) and provided that entrance into the system is approximately uniform, ($f(x) = 1/b$), the method of accessing records suggested above should give accurate and direct information concerning $g(y)$.