Which Wald statistic? Choosing a parameterization of the Wald statistic to maximize power in k-sample generalized estimating equations

David I. Warton *

Phone: (61)(2) 9385 7031

Fax: (61)(2) 9385 7123

email: David.Warton@unsw.edu.au

Running title – Wald statistics and power

* E-mail: *David.Warton@unsw.edu.au*

Abstract

The Wald statistic is known to vary under reparameterization. This raises the question: which parameterization should be chosen, in order to optimize power of the Wald statistic? We specifically consider k-sample tests of generalized linear models and generalized estimating equations in which the alternative hypothesis contains only two parameters. Amongst a general class of parameterizations, we find the parameterization that maximizes power via analysis of the non-centrality parameter, and show how the effect on power of reparameterization depends on sampling design and the differences in variance across samples. There is no single parameterization with optimal power across all alternatives. The Wald statistic commonly used, that under the canonical parameterization, is optimal in some instances but it performs very poorly in others. We demonstrate results by example and by simulation, and describe their implications for likelihood ratio statistics and score statistics. We conclude that due to poor power properties, the routine use of score statistics and Wald statistics under the canonical parameterization for generalized estimating equations is a questionable practice.

Keywords

Canonical parameterization, log-likelihood ratio statistic, power simulation, score statistic, skewness-reducing parameterization, variance-stabilizing parameterization.

1 Introduction

Wald statistics are commonly used in hypothesis testing of generalized linear models (GLMs) and generalized estimating equations (GEEs). Wald statistics are particularly useful in making inferences about parameters estimated using GEEs, for which the likelihood ratio statistic is undefined in general (Rotnitzky and Jewell, 1990). Another advantage of Wald statistics is computational efficiency. A Wald statistic is only a function of parameters estimated under the alternative model, so it can be calculated without fitting the null model to the data. Hence a Wald statistic can be computed after fitting only one model, even when conducting several hypothesis tests, as long as they involve the same alternative model. An example of such a situation is significance testing of multiple regression coefficients, and Wald statistics are standard output for generalized linear models in most statistics packages for this reason.

Whereas the likelihood ratio statistic and the score statistic of Rao (1948) are invariant under reparameterization, this is not the case for Wald statistics. The parameterization-variance of Wald statistics is usually considered to be a disadvantage (for example Barndorff-Nielsen and Cox, 1994, page 120), however it could be considered as an opportunity – which parameterization should we choose, in order to obtain a Wald statistic with good properties? In this paper, the property we are specifically interested in is power.

The study of properties of Wald and score statistics is an important issue, considering the widespread use of these statistics. Various authors have explored the Type I error

properties of these different statistics (Barnwal and Paul, 1988, for example) and improvements therein when using different variance estimates (Mancl and DeRouen, 2001; Guo, Pan, Connett, Hannan, and French, 2005, for example) or higher-order theory (Pierce and Peters, 1992, for example). Yet somewhat surprisingly, we found relatively little study of the power properties of these statistics. Building on previous literature (Peers, 1971; Hayakawa, 1975; Harris and Peers, 1980), Cordeiro, Botter, and Ferrari (1994) and Ferrari, Botter, and Cribari-Neto (1997) derived analytical expressions that could be used to compare the power of score, Wald and likelihood ratio statistics, to second order, for generalized linear models. We found no previous literature specifically studying the effect of reparameterization on power of Wald statistics.

We derive in this paper a relationship between parameterization and the non-centrality parameter of the Wald statistic, for a class of generalized estimating equations where the non-centrality parameter is a function of two model parameters. Hence we show that power properties depend on the extent of imbalance in the sampling design, and that some parameterizations can be described as "extreme" or "intermediate", depending on their power behavior under different sampling designs, for a fixed alternative hypothesis. Extreme statistics have very high power for some sampling designs, but very low power for others. The commonly used canonical parameterization is extreme in this sense, and we recommend that an intermediate statistic such as one based on the variance-stabilizing parameterization would be a better choice in most practical situations. We demonstrate these ideas by simulation, and demonstrate the implications for other commonly used test statistics –

score and likelihood ratio statistics.

We make recommendations regarding the usage of different statistics on the basis of our power results. It should be emphasized, however, that comparison of statistics on the basis of power is only relevant provided that the size of the tests has been suitably controlled – hence our conclusions are only relevant for situations where one either has a large enough sample size for Type I error properties to be sufficiently good, or where one uses resampling to ensure valid test sizes, as in section 2 and as in our simulations.

2 A motivating example

In this section we briefly describe an example application (Table 1) requiring a two-sample test of overdispersed count data. The data (obtained from the Key Centre for Biodiversity and Bioresources, Macquarie University) are counts of the abundance of two orders of invertebrates, sampled from the leaf litter at ten different sites near Sydney, Australia. Two of the ten sites are controls, and vegetation regeneration projects have been undertaken in the remaining eight. It is of interest to test whether the "regenerated" invertebrate communities are different to the controls in any way, and if so, how.

For the purposes of this study, it is important to note that the data are strongly overdispersed counts (Table 1), and that due to limited availability of control sites, the data have been sampled in an unbalanced design.

From a cursory glance at the data, it appears that for both orders of invertebrate

there is a clear effect of the "regeneration" treatment – to increase amphipod abundance from zero and to decrease cockroach abundance, usually to zero. However, whether or not we see this in a formal analysis depends which test statistic we use.

We fitted a negative binomial model to data, with mean-variance function $V(\mu) = \mu + \phi\mu^2$, for ϕ fixed across all observations. We tested the hypothesis of no difference between regeneration and control sites in terms of mean abundance ($H_0 : \mu_C = \mu_R$) separately for each invertebrate order, and obtain the results in Table 2. Because of the small sample sizes involved, we evaluated statistical significance using permutation tests, evaluating all 45 possible permutations of treatment labels.

Despite a substantial apparent difference in amphipod abundance between regenerated and control sites, the only statistic suggestive of this difference was the likelihood ratio statistic ($-2 \log L$). While $-2 \log L$ recorded a very large value for its test statistic, which was significant at the 0.05 level, the score statistic (S) recorded a small value which was not significant ($P = 0.222$), and the Wald statistic under the canonical parameterization (W_1) was undefined, because $\hat{\mu}_C = 0$. If we resolve this problem by replacing W_1 with its limiting value as $\hat{\mu}_C \to 0$, we obtain $W_1 = 0$ and $P = 1$! This effect has previously been explored by Væth (1985), who described how reparameterization can be used to solve this particular problem.

For the *Blattodea* data, again there were substantial discrepancies in the values taken by the test statistics. However, these discrepancies were not as large as for the *Amphipoda* data, and they did not lead to differences in interpretation of results – all statistics were significant at the 0.05 level for *Blattodea* abundances.

If we wanted to test simultaneously for a difference in abundance of both inverte-

brate orders, between control and regeneration sites, then one approach would be to use generalized estimating equations (GEEs). A problem that then arises, however, is that this is a semi-parametric modeling approach for which the likelihood is undefined. Hence $-2\log L$, the only statistic which behaved well in the above univariate tests, is not available to us for GEEs, and we must choose between S and the Wald statistic under some parameterization. We explore in this paper the effect of reparameterization on the power of the Wald statistic. This gives us guidance as to which Wald statistic to use in hypothesis tests for GEEs, and it also sheds light on why S had poor properties in the above example.

3 Parameterizations of the Wald statistic

Consider N observations $y = (y_1, \ldots, y_N)^T$ that satisfy a k-sample generalized linear model, *i.e.* $var(y_i) = V(\mu_i)$ and $E(y_i) = \mu_i$ satisfies:

$$h_\alpha(\mu_i) = \sum_{j=1}^{k} \mathcal{I}(i \in \mathcal{S}_j)\beta_j$$

for some link function $h_\alpha(\cdot)$, where $\mathcal{I}$ denotes the indicator function, and $\mathcal{S}_j$ denotes the set of all observations belonging to the jth sample. We also define $\beta = (\beta_1, \ldots, \beta_k)^T$, the vector storing the k mean parameters.

Note that for k-sample models, changing the link function does not change the form of the model being fitted, and so has the effect of reparameterizing β. This is not the case for generalized linear models as they are usually specified. Consider, for example, a model in which the linear predictor contains an explanatory variable X that is not an indicator variable, and takes more than two distinct values. In such a

case, changing the link function would change the form of the relationship between X and μ.

We specifically consider a class of parameterizations formed via changing α in the link function, of the form

$$h_\alpha(\mu) = \int_0^\mu V(t)^{-\alpha} dt$$

for $\alpha \in [0, 1]$. For exponential families, these parameterizations always exist, although their solution does not always have a closed form (Hougaard, 1982, for example).

Four parameterizations are of particular interest, due to their special properties:

Canonical (h_1) The canonical parameterization (McCullagh and Nelder, 1989) of μ, for which $\alpha = 1$, gives the special property that

$$\frac{\partial l(\beta; y)}{\partial \beta_j} = \sum_{i=1}^{N} \mathcal{I}(i \in \mathcal{S}_j)(y_i - \mu_i)$$

Skewness reducing ($h_{2/3}$) The skewness reducing (DiCiccio, 1984) or vanishing third derivative (Slate, 1994) parameterization, for which $\alpha = 2/3$, satisfies

$$\left. \frac{\partial^3 l(\beta; y)}{\partial \beta_j \beta_{j'} \beta_{j''}} \right|_{\beta=\hat{\beta}} = 0. \tag{1}$$

Variance stabilizing ($h_{1/2}$) The variance stabilizing parameterization (DiCiccio, 1984; Slate, 1994), for which $\alpha = 1/2$, ensures that the expected information is not a function of μ:

$$- E\left(\frac{\partial^2 l(\beta; y)}{\partial \beta_j \beta_{j'}} \right) \Bigg|_{\beta=\hat{\beta}} \propto \begin{cases} \sum_{i=1}^{N} \mathcal{I}(i \in \mathcal{S}_j) & \text{if } j = j' \\ 0 & \text{if } j \neq j' \end{cases}$$

Mean-value (h_0) The mean-value parameterization (Væth, 1985), for which $\alpha = 0$, satisfies $h_0(\mu) = \mu$.

We denote as W_α the Wald statistic under the parameterization using link function $h_\alpha(\mu)$. For the test $H_0 : L\beta = 0$ with some matrix of constraints L, the Wald statistic can be written as

$$W_\alpha = (L\hat{\beta})^T \left(L\hat{var}(\hat{\beta})L^T\right)^{-1} (L\hat{\beta})$$

Some interesting relations are known between Wald statistics under various parameterizations and other likelihood-based statistics. For example, consider the score statistic due to Rao (1948), in which the expected information under the null hypothesis is used to estimate the variance term. This is known to have the same power as W_1, to second order, under local alternatives (Cordeiro et al., 1994). A similar relationship holds for the (log-)likelihood ratio statistic: under local alternatives, it is known to be equivalent to $W_{2/3}$ to order $O_p(N^{-1})$ for k-sample tests (Warton and Hudson, 2006), and these statistics have the same power to second order in more general settings (Cordeiro et al., 1994).

4 Power of Wald statistics

To first order under local alternatives, it is well known that the power of a Wald statistic can be calculated using a non-central chi-squared distribution (Barndorff-Nielsen and Cox, 1994, for example). To second order, the power of a Wald statistic can be calculated using a linear combination of non-central chi-squared distributions with differing degrees of freedom, yet with a constant non-centrality parameter

(Hayakawa, 1975). Hence we study the power of Wald statistics under reparameterization via their non-centrality parameter Φ.

The central result of this paper is Theorem 1 below.

Theorem 1. *Consider a scenario in which the true model is a function of parameters* $\beta_1 = b_1 1_{a_1 \times 1}$ *and* $\beta_2 = b_2 1_{a_2 \times 1}$*, where* $1_{a \times b}$ *is an* $a \times b$ *matrix of ones, and the non-centrality parameter of* W_α *is a function of*

$$\Phi = (b_1 - b_2)^2 \left(\frac{V(m_1)^{1-2\alpha}}{N_1} + \frac{V(m_2)^{1-2\alpha}}{N_2} \right)^{-1} \tag{2}$$

where $h_\alpha(m_i) = b_i$ *and* N_i *is the number of observations that satisfy* $\mu_j = m_i$*. Assume (without loss of generality) that* $m_2 > m_1$ *such that* $m_2 = m + \delta$ *for some* $\delta > 0$*, and* $m_1 = m$*.*

The non-centrality parameter can be written as

$$\Phi = \frac{\delta^2}{V(m)\left(\frac{1}{N_1} + \frac{1}{N_2}\right)} \left\{ 1 + \delta\psi + O(\delta^2) \right\} \tag{3}$$

where

$$\psi = \frac{V'(m)}{V(m)} \frac{1}{\frac{1}{N_1} + \frac{1}{N_2}} \left\{ \alpha \left(\frac{1}{N_2} - \frac{1}{N_1} \right) - \frac{1}{N_2} \right\} \tag{4}$$

Based on the above expansion of the non-centrality parameter, for $\alpha \in [0, 1]$*:*

- *If* $N_1 < N_2$*, power is maximized at* $\alpha = 0$*.*
- *If* $N_1 = N_2$*, all parameterizations have equal power.*
- *If* $N_1 > N_2$*, power is maximized at* $\alpha = 1$*.*

Further, because the non-centrality parameter is a linear function of α*, to second order, for* $N_1 \neq N_2$ *power increases smoothly as* α *approaches its optimal value.*

Proof. A Taylor expansion of b_2 about b_1 gives

$$b_2 = b_1 + \delta h'(m) + \frac{\delta^2}{2} h''(m) + O(\delta^2)$$

and so

$$\begin{aligned}(b_2 - b_1)^2 &= \delta^2 h'(m)^2 \left(1 + \delta \frac{h''(m)}{h'(m)} + O(\delta^2)\right) \\ &= \delta^2 V(m)^{-2\alpha} \left(1 - \delta\alpha \frac{V'(m)}{V(m)} + O(\delta^2)\right)\end{aligned}$$

since $h'(m) = V(m)^{-\alpha}$.

Now

$$\begin{aligned}\left(\frac{V(m_1)^{1-2\alpha}}{N_1} + \frac{V(m_2)^{1-2\alpha}}{N_2}\right)^{-1} &= \left(\frac{V(m)^{1-2\alpha}}{N_1} + \frac{V(m+\delta)^{1-2\alpha}}{N_2}\right)^{-1} \\ &= V(m)^{2\alpha-1} \left\{\left(\frac{1}{N_1} + \frac{1}{N_2}\right) + \delta \frac{1-2\alpha}{N_2} \frac{V'(m)}{V(m)} + O(\delta^2)\right\}^{-1} \\ &= V(m)^{2\alpha-1} \left(\frac{1}{N_1} + \frac{1}{N_2}\right)^{-1} \left\{1 - \delta \frac{1-2\alpha}{N_2} \left(\frac{1}{N_1} + \frac{1}{N_2}\right)^{-1} \frac{V'(m)}{V(m)} + O(\delta^2)\right\}\end{aligned}$$

And multiplying together the above two expressions we get

$$\Phi = \frac{\delta^2}{V(m)\left(\frac{1}{N_1} + \frac{1}{N_2}\right)} \left(1 - \delta\alpha \frac{V'(m)}{V(m)} + O(\delta^2)\right) \left\{1 - \delta \frac{1-2\alpha}{N_2} \left(\frac{1}{N_1} + \frac{1}{N_2}\right)^{-1} \frac{V'(m)}{V(m)} + O(\delta^2)\right\}$$

This simplifies readily to the form of equation 3, where the coefficient of δ is

$$\begin{aligned}\psi &= -\frac{1-2\alpha}{N_2} \left(\frac{1}{N_1} + \frac{1}{N_2}\right)^{-1} \frac{V'(m)}{V(m)} - \alpha \frac{V'(m)}{V(m)} \\ &= \frac{V'(m)}{V(m)\left(\frac{1}{N_1} + \frac{1}{N_2}\right)} \left\{\frac{2\alpha - 1}{N_2} - \alpha\left(\frac{1}{N_1} + \frac{1}{N_2}\right)\right\} \\ &= \frac{V'(m)}{V(m)\left(\frac{1}{N_1} + \frac{1}{N_2}\right)} \left\{\alpha\left(\frac{1}{N_2} - \frac{1}{N_1}\right) - \frac{1}{N_2}\right\}\end{aligned}$$

□

Remark 1. *Note that the coefficient of α in the second order term ψ of equation 4 is proportional to*

$$\frac{\frac{1}{N_2} - \frac{1}{N_1}}{\frac{1}{N_1} + \frac{1}{N_2}} = \frac{N_1 - N_2}{N_1 + N_2}$$

This term is a measure of imbalance in sampling. More specifically, it is the difference in sample sizes as a proportion of total sample size. So the effect on power of reparameterization increases with the extent of imbalance in sampling.

Remark 2. *Note, from equation 4, that the second order term $\delta\psi$ is proportional to*

$$\delta\frac{V'(m)}{V(m)} \approx \frac{V(m_2) - V(m_1)}{V(m_1)}$$

i.e. *a linear approximation to the proportional difference in variance. So the effect on power of reparameterization increases with the proportional difference in variance.*

One might attempt to use Theorem 1 to choose a parameterization of the Wald statistic that is expected to maximize power, for $N_1 \neq N_2$. However, a practical problem doing so is that one would need to know *a priori* which of $V(m_1)$ and $V(m_2)$ is larger. This is possible in some situations, such as when interested in a one-sided alternative, but otherwise this will not be the case. Hence it makes sense as a general rule to choose an intermediate value such as $\alpha = 1/2$, to "bet-hedge": this value will always have intermediate power between W_0 and W_1 and one would expect this statistic to always perform reasonably and never perform poorly. In contrast, for some alternatives W_0 and W_1 will have maximum power amongst the W_α, $\alpha \in [0, 1]$, but for other alternatives they will have minimum power.

Although we have only considered Wald statistics in the above, our results also have relevance for other members of the "Holy Trinity", the log-likelihood ratio statistic

$(-2\log L)$ and the score statistic using expected information evaluated under the null hypothesis (S). The log-likelihood ratio statistic and its power are known to approximate $W_{2/3}$ to second order (Cordeiro et al., 1994; Warton and Hudson, 2006), and the power of the score statistic is known to approximate W_1 to second order (Cordeiro et al., 1994). Hence the properties we describe for "intermediate" statistics such as $W_{2/3}$ have relevance for $-2\log L$: we expect it to have relatively good power across all alternatives. We expect S to inherit the properties of the "extreme" statistic W_1: it is expected to similarly have either very high or very low power for unbalanced designs, depending on the nature of the imbalance.

5 Generality of Theorem 1

In order to derive a closed form result in theorem 1, it was necessary to restrict attention to models in which there were only two unique true parameters. The form of the non-centrality parameter then had the familiar form of a two-sample test for a generalized linear model. It should be noted however that the theorem is applicable in a broader range of models, such as those for which we use generalized estimating equations, in the specific case of a two-parameter alternative:

Theorem 2. *Consider the non-centrality parameter of W_α in the following cases:*

(1) A two sample test of a generalized linear model i.e. *where for all i, $h_\alpha(\mu_i) \in \{b_1, b_2\}$ and we test $H_0 : b_1 = b_2$ against $H_1 : b_1 \neq b_2$, or against a one-sided alternative, such as $H_1' : b_1 > b_2$.*

(2) A two-parameter alternative in k-sample tests of a generalized linear model. We test $H_0 : \beta_i = \beta$ for all i against $H_1 : \beta_i \neq \beta$ for some i. We consider power of

the specific alternative for which all k parameters satisfy $\beta_i \in \{b_1, b_2\}$.

(3) A two sample test of generalized estimating equations, using a naïve estimator of $var(\hat{\beta}_i)$*, which is correctly specified. We have two p-variate samples, where all p variables have marginal distributions belonging to an exponential family with the same mean-variance function* $var(y_{ij}) = V(\mu_{ij})$*. Mean parameters are stored in the p-vectors* β_1 *and* β_2*. We estimate the* β_i *using generalized estimating equations, and test* $H_0 : \beta_1 = \beta_2$ *against* $H_1 : \beta_1 \neq \beta_2$*. We calculate power for a two-parameter alternative in which* $\beta_1 = b_1 1_{p\times 1}$ *and* $\beta_2 = b_2 1_{p\times 1}$.

(4) A two sample test of generalized estimating equations, using a sandwich estimator of the variance matrix. We require the same conditions here as for 3, except we do not require that $var(\hat{\beta}_i)$ *is correctly specified, but instead the two groups must share a common correlation matrix,* i.e. $E(\hat{R}_1) = E(\hat{R}_2)$.

(5) A two-parameter alternative in k-sample tests of generalized estimating equations. Either a naïve estimator of the variance matrix could be used, or a sandwich estimator, although for the former we require $var(\hat{\beta}_i)$ *to be correctly specified and for the latter we require* $E(\hat{R}_i) = R$ *for all i. The setup is as for 3, except that there are now k samples, k p-vectors of parameters* β_i*, and we test* $H_0 : \beta_i = \beta$ *for all i against* $H_1 : \beta_i \neq \beta$ *for some i. We consider the two parameter alternative* $\beta_i \in \{b_1 1_{p\times 1}, b_2 1_{p\times 1}\}$ *for all i.*

W_α *has a non-centrality parameter that simplifies to a multiple of equation 2 in each of these cases. Hence Theorem 1 is applicable in each case.*

Proof. 1. The proof is straightforward in this case. $var(\hat{b}_i) = V(m_i)\left\{h'(m_i)\right\}^2 / N_i$ and the result follows, given that the non-centrality parameter of the Wald statistic

is $(b_1 - b_2)^2 var(\hat{b}_1 - \hat{b}_2)^{-1}$.

2. Let the sample sizes in each of the k groups be stored in N. Consider a partition of the k-vector of parameters $\beta = (\beta_1, \beta_2)$ where $\beta_1 = 1_{(k_1+1)\times 1} b_1$ and $\beta_2 = 1_{k_2 \times 1} b_2$. Similarly, let $N = (n_1, n_2)$, but let us further partition $n_1 = (n_{11}, n_{12})$ where n_{11} is the sample size in the first group, and n_{12} is the sample size of the remaining k_1 groups in the first partition. Let $N_1 = 1_{1\times(k_1+1)} n_1$ and $N_2 = 1_{1\times k_2} n_2$. Hence there are $k_1 + 1$ groups with $\beta_i = b_1$, and the total sample size across all these groups is N_1, while there are k_2 groups with $\beta_i = b_2$, and the total sample size across these groups is N_2.

The Wald statistic can be written as

$$(L\hat{\beta})^T \left(L\hat{var}(\hat{\beta}) L^T \right)^{-1} L\hat{\beta}$$

where $L = [1_{k-1\times 1},\ -I_{k-1}]$, and I_a is the $a \times a$ identity matrix. In this case the non-centrality parameter can be written as

$$\begin{pmatrix} 0_{k_1\times 1} \\ (b_1 - b_2) 1_{k_2 \times 1} \end{pmatrix}^T \begin{pmatrix} v_1 \text{diag}(n_{12}^{-1}) + \frac{v_1}{n_{11}} 1_{k_1\times k_1} & \frac{v_1}{n_{11}} 1_{k_1 \times k_2} \\ \frac{v_1}{n_{11}} 1_{k_2 \times k_1} & v_2 \text{diag}(n_2^{-1}) + \frac{v_1}{n_{11}} 1_{k_2 \times k_2} \end{pmatrix}^{-1} \begin{pmatrix} 0_{k_1\times 1} \\ (b_1 - b_2) 1_{k_2 \times 1} \end{pmatrix}$$

$$= (b_2 - b_1)^2 1_{1\times k_2} V^{22} 1_{k_2 \times 1}$$

where V^{22} is the bottom-right term of the inverse of $Lvar(\hat{\beta})L^T$, and $v_i = V(m_i)^{1-2\alpha}$.

Now

$$
\begin{aligned}
(V^{22})^{-1} &= V_{22} - V_{21}\left(V_{11}\right)^{-1} V_{12} \\
&= v_2 \text{diag}(n_2^{-1}) + \frac{v_1}{n_{11}} 1_{k_2 \times k_2} - \frac{v_1^2}{n_{11}^2} 1_{k_2 \times k_1} \left(v_1 \text{diag}(n_2^{-1}) + \frac{v_1}{n_{11}} 1_{k_1 \times k_1} \right)^{-1} 1_{k_1 \times k_2}
\end{aligned} \tag{5}
$$

Now it can be shown using (Petersen and Pedersen, 2006, page 16, for example) that

$$
(A + b1_{k \times k})^{-1} = A^{-1} - \frac{bA^{-1}1_{k \times k}A^{-1}}{1 + b1_{1 \times k}A^{-1}1_{k \times 1}}
$$

Also, $1_{a \times k_1} \text{diag}(n_{12}) 1_{k_1 \times b} = (N_1 - n_{11}) 1_{a \times b}$ and so

$$
\begin{aligned}
&\frac{v_1}{n_{11}} 1_{k_2 \times k_1} \left(v_1 \text{diag}(n_{12}^{-1}) + \frac{v_1}{n_{11}} 1_{k_1 \times k_1} \right)^{-1} 1_{k_1 \times k_2} \\
&\qquad = \frac{1}{n_{11}} 1_{k_2 \times k_1} \left(\text{diag}(n_{12}) - \frac{\text{diag}(n_{12}) 1_{k_1 \times k_1} \text{diag}(n_{12})}{n_{11} + 1_{1 \times k_1} \text{diag}(n_{12}) 1_{k_1 \times 1}} \right) 1_{k_1 \times k_2} \\
&\qquad = \frac{1}{n_{11}} \left(N_1 - n_{11} - \frac{(N_1 - n_{11})^2}{n_{11} + (N_1 - n_{11})} \right) 1_{k_2 \times k_2} \\
&\qquad = \frac{N_1 - n_{11}}{N_1} 1_{k_2 \times k_2}
\end{aligned}
$$

Equation 5 then simplifies to

$$
\begin{aligned}
(V^{22})^{-1} &= v_2 \text{diag}(n_2^{-1}) + \frac{v_1}{n_{11}} 1_{k_2 \times k_2} - \frac{v_1}{n_{11}} \frac{N_1 - n_{11}}{N_1} 1_{k_2 \times k_2} \\
&= v_2 \text{diag}(N_2^{-1}) + \frac{v_1}{N_1} 1_{k_2 \times k_2}
\end{aligned}
$$

hence

$$
\phi = (b_2 - b_1)^2 1_{1 \times k_2} \left(v_2 \text{diag}(N_2^{-1}) + \frac{v_1}{N_1} 1_{k_2 \times k_2} \right)^{-1} 1_{k_2 \times 1}
$$

Using the same identities above, this simplifies to

$$\begin{aligned}\phi &= (b_2 - b_1)^2 \left\{ \frac{N_2}{v_2} - \frac{v_1 N_2^2}{N_1 v_2^2} \left(1 + \frac{v_1 N_2}{N_1 v_2}\right)^{-1} \right\} \\ &= (b_2 - b_1)^2 \frac{N_2}{v_2} \left(1 + \frac{v_1 N_2}{N_1 v_2}\right)^{-1} \\ &= (b_2 - b_1)^2 \left(\frac{v_1}{N_1} + \frac{v_2}{N_2}\right)^{-1}\end{aligned}$$

which has the same form as equation 2.

3. A naïve estimator for $\hat{var}(\hat{\beta}_i)$ has the form

$$\hat{var}_{\text{naïve}}(\hat{\beta}_i) = \frac{1}{N_i} \text{diag}(V(\hat{M}_i)^{1-2\alpha})^{1/2} \, \hat{R}_w \, \text{diag}(V(\hat{M}_i)^{1-2\alpha})^{1/2} \tag{6}$$

where $\hat{R}_w$ is the working correlation structure, and M_i is the vector of means for the ith group.

Now if $\beta_i = 1_{p\times 1} b_i$, then $M_i = 1_{p\times 1} m_i$ and $\text{diag}(V(M_i)) = V(m_i) I_p$, hence

$$var_{\text{naïve}}(\hat{\beta}_i) = \frac{1}{N_i} V(m_i)^{1-2\alpha} E(R_w)$$

and so, provided that the variance has been correctly specified, Φ simplifies to

$$\begin{aligned}&(\beta_1 - \beta_2)^T \left(var(\hat{\beta}_1) + var(\hat{\beta}_2)\right)^{-1} (\beta_1 - \beta_2) \\ &\qquad = (b_1 - b_2)^2 \left(\frac{V(m_1)^{1-2\alpha}}{N_1} + \frac{V(m_2)^{1-2\alpha}}{N_2}\right)^{-1} 1_{1\times p} E(R_w)^{-1} 1_{p\times 1}\end{aligned}$$

which is proportional to equation 2.

4. A sandwich estimator for $var(\hat{\beta}_i)$ has the form

$$\hat{var}_{\text{sandwich}}(\hat{\beta}_i) = \hat{var}_{\text{naïve}}(\hat{\beta}_i) \, \hat{\kappa} \, \hat{var}_{\text{naïve}}(\hat{\beta}_i)$$

where

$$\hat{\kappa} = N_i \, \mathrm{diag}(V(\hat{M}_i)^{1-2\alpha})^{-1/2} \, \hat{R}_w^{-1} \, \hat{R}_i \hat{R}_w^{-1} \, \mathrm{diag}(V(\hat{M}_i)^{1-2\alpha})^{-1/2}$$

and $\hat{R}_i$ is the sample correlation matrix of Pearson residuals calculated within the ith group, $\mathrm{diag}(M_i)^{-1/2}(Y_i - 1_{N_1 \times 1} M_i)$.

$var_{\mathrm{sandwich}}(\hat{\beta}_i)$ simplifies to

$$\hat{var}_{\mathrm{sandwich}}(\hat{\beta}_i) = \frac{1}{N_i} \mathrm{diag}(V(\hat{M}_i)^{1-2\alpha})^{1/2} \, \hat{R}_i \, \mathrm{diag}(V(\hat{M}_i)^{1-2\alpha})^{1/2}$$

This has the same form as equation 6, except that the working correlation matrix $\hat{R}_w$ has been replaced by a within-group unstructured correlation matrix, $\hat{R}_i$. So we can use the same approach as for 3 to show that provided that $E(\hat{R}_1) = E(\hat{R}_2)$, the non-centrality parameter is proportional to equation 2.

5. This follows by generalizing result 2 using the approach of results 3 and 4. □

While the above only represent a restricted class of GEE's and GLM's, they give some indication of how Theorem 1 generalizes to more complex contexts. Hence, for example, we suggest that in k-sample testing of GEE's in general, if the more variable groups tend to be more intensively sampled, then a Wald statistic will have higher power when α is smaller. This corresponds to the $N_1 < N_2$ situation. Conversely, when the more variable groups tend to be less intensively sampled, larger α gives higher power. Finally, for a GEE based on a balanced sampling design, we expect all parameterizations to have similar power.

6 Simulations

In this section, the above results are demonstrated via simulation, focussing specifically on the case of two-sample or three-sample designs for overdispersed count data.

To ensure that any differences among statistics were due to power and not due to accuracy of the test, we measured statistical significance using permutation tests rather than through reference to the chi-squared distribution. This ensured that all tests were approximately exact, the approximation being due to a very small amount of Monte Carlo error.

In all cases, power was estimated at the 0.05 level from 1000 simulated datasets. Statistical significance at the 0.05 level was evaluated using permutation tests, where the group labels of each observation were permuted. We used 999 randomly chosen permutations of the data to estimate significance levels. There were two levels of sampling involved in simulations – sampling datasets, then resampling them to estimate P-values. This was computationally burdensome, and required approximately 20 hours of total time when using an AMD Opteron 246 processor (2.6 GHz).

We calculated six test statistics: W_0, $W_{1/2}$, $W_{2/3}$, W_1, S and $-2\log L$. Theorem 1 applies directly to Wald statistics only, but given that $S \approx W_1$ and $-2\log L \approx W_{2/3}$, it has implications for these statistics also. We did not calculate $-2\log L$ for GEEs, as it is undefined.

In the case of GEEs, we used a sandwich estimator of the variance matrix with slight modifications. The use of a sandwich estimator has the potential to introduce

considerable inefficiencies, and is often not a good idea for small samples (Drum and McCullagh, 1993). This is pertinent in our case, where a sandwich estimator would require the use of $\hat{R}_i$ in calculations, where N_i is typically 15 or less. Hence we replaced $\hat{R}_i$ with a pooled estimate $\hat{R}$ obtained by moment estimation using all Pearson residuals, not just those from the ith group.

We considered 9 power simulations in a 3×3 design, in which we varied the data structure and the parameter that was varied in simulations.

Each of the following three data structures were considered:

- Two-sample tests for data from a negative binomial distribution, with a total of 30 observations.
- Two samples tests for data from a bivariate negative binomial distribution, fitted using generalized estimating equations, with a total of 30 observations.
- Three-sample tests for negative binomial counts, for a total of 45 observations, with the sample size and mean of group 2 fixed at $N_2 = 15$ and $m_2 = 2$.

In each case we sampled data from geometric distributions, such that data satisfied the mean-variance function $V(\mu) = \mu + \phi\mu^2$ where $\phi = 1$. For bivariate data, we sampled X_1, X_2 and Z independently from a negative binomial distribution with $V(\mu) = \mu + \mu^2/2$ and $\mu = m_i/2$. We then calculated the two response variables as $Y_1 = X_1 + Z$ and $Y_2 = X_2 + Z$, whose correlation is 0.5, and whose marginal distributions are geometric with mean m_i (Johnson et al., 1997).

For each data structure, we explored the effect on power of varying key parameters in three settings:

Sample size ratio varied The relative sample size in two groups was varied, while keeping means and variances fixed. For k-sample tests, we varied N_k/N_1 while keeping $N_1+N_k = 30$. In the two sample case, $m_1 = 1$, $m_2 = 3$, hence $V(m_2)/V(m_1) = 6$. In the three sample case, $N_2 = 15$ and $m_1 = 1$, $m_2 = 2$, and $m_3 = (73^{1/2}-1)/2$, hence $V(m_3)/V(m_2) = V(m_2)/V(m_1) = 3$.

Variance ratio varied, balanced sampling The ratio of variances across groups $V(m_k)/V(m_1)$ was varied, by varying the means, while keeping sample sizes fixed in a balanced sampling design with $N_i = 15$ for all i. For the two-sample case, we set $m_2 = 3$ and varied m_1. For the three-sample case, we set $m_2 = 2$, and varied m_1 and m_3 such that $V(m_2)/V(m_1) = V(m_3)/V(m_2)$.

Variance ratio varied, unbalanced sampling The ratio of variances across groups $V(m_k)/V(m_1)$ was varied, by varying the means, while keeping sample sizes fixed in an unbalanced design. We repeated the conditions of the previous variance ratio simulation, except that now $N_1 = 10$ and $N_k = 20$.

In modeling the data we assumed that the dispersion parameter ϕ was unknown. For bivariate data, we estimated ϕ separately for each response variable.

For bivariate data, we estimated parameters using an independence estimating equations approach, *i.e.* using the identity matrix as the working correlation structure. This is much more computationally efficient than joint estimation of β and R, in fact it leads to important computational savings. This was an important consideration due to the extensive computation times in conducting these simulations.

Simulations varying the sample size ratio were designed to demonstrate remark 1, and simulations changing the variance ratio were designed to demonstrate remark 2.

While the simulations using a two-sample design represent a demonstration of Theorem 1, the three-sample design considered here represents a situation in which Theorem 1 does not apply directly. In this simulation, the true model is a three-parameter alternative. It is included to demonstrate how lessons learnt from Theorem 1 for two-parameter alternatives provide a guide for performance in other situations. Specifically, we expected that for the three-sample case, just as in the two-sample case, $\alpha = 1$ would have highest power when more variable groups were less intensively sampled, $\alpha = 0$ would have highest power when more variable groups were more intensively sampled, and that reparameterization would have negligible effect on power when sampling was balanced.

In all simulations, we observed the patterns predicted by Theorem 1 (Figure 1). These patterns were just as evident for three-sample tests (Figure 1c) as for the two sample situations that are directly covered by Theorem 1.

In the first and third columns of Figure 1, W_1 had highest power when the more variable group was less intensively sampled, and lowest power when the more variable group was more intensively sampled. W_0 had the opposite behavior of W_1. The extent of differences in power increased as N_k/N_1 or $V(m_k)/V(m_1)$ moved away from 1, as expected from remarks 1 and 2, respectively. Differences in power with parameterization were most extreme in the first column of Figure 1, where W_1 often had over twice the power or less than half the power of W_0 and $W_{1/2}$, depending whether $N_1 > N_k$ or $N_1 < N_k$, respectively.

$W_{1/2}$ and $W_{2/3}$ were intermediate between W_0 and W_1, with $W_{1/2}$ closer in power to the former and $W_{2/3}$ closer to the latter. $W_{1/2}$ was the only statistic whose power

was roughly symmetrical when plotted against $\log(N_k/N_1)$ and $\log\{V(m_k)/V(m_1)\}$ (Figure 1). This approximate symmetry can be understood from equation 4, where the second order term $\delta\psi$ is symmetric with respect to N_1 and N_2 when $\alpha = 1/2$.

When sampling was balanced, all statistics had almost identical power. This was evident in the almost coincident power curves when sampling was balanced (second column of Figure 1), and in the almost perfect intersection of the power curves at $N_k/N_1 = 1$ (first column of Figure 1).

As expected, we found that S and $-2\log L$ behaved like their respective Wald statistics in power simulations (Figure 2). Power curves for S and W_1 were almost perfectly co-incident, and similarly for $-2\log L$ and $W_{2/3}$, with the exception of extremely unbalanced designs, where $N_k/N_1 > 3$ or $N_k/N_1 < 1/3$. Because we sampled such that $N_1 + N_k = 30$, such unbalanced designs correspond to situations where $N_i \leq 7$ for some i. Hence we suspect that the divergence of power curves is due to small sample size rather than being related to imbalanced design *per se*.

7 Discussion

We find our results quite concerning, from the perspective that the two test statistics routinely used with generalized estimating equations – W_1 and S – both have some quite undesirable power properties. For k-sample tests with unbalanced sampling designs, these statistics have extreme power behavior, either performing very well or very badly. Which of these two possibilities will eventuate can only be known *a priori* if a restricted form of alternative model is expected, *e.g.* $m_1 < m_2$. In

other situations, it would be more prudent to use a test statistic with intermediate properties – one that has relatively high power across a range of alternatives, such as $W_{2/3}$, $-2\log L$ where it is applicable, or perhaps $W_{1/2}$.

We expect that the effect of reparameterization on power of W_α will be larger for some types of data than for others, and quite substantial for overdispersed count data in particular. This is because the extent of effects of α on power is a function of the amount of variation in $V(m_i)$ across samples (remark 2, and third column of Figure 1), which is related to the derivative $V'(m_i)$. When modeling overdispersed count data in particular, $V'(m_i) > 1$, hence $V(m_i)$ can vary considerably across samples and the properties of W_α might change considerably as α changes. In contrast, when modeling binomial data, $|V'(m_i)| \leq 1$ and $V(m_i)$ will only vary considerably across samples when we encounter cells with rare responses, *i.e.* some m_i near 0 or 1. Hence we might expect that reparameterization will have comparatively little effect on power of W_α for most binary problems. When fitting a normal model with constant variance, of course, we expect negligible effect of reparameterization on power.

For k-sample tests of generalized estimating equations where a single test statistic is required, we recommend $W_{1/2}$, the Wald statistic under the variance stabilizing parameterization. This statistic has been shown to have relatively good power across a range of scenarios. A further important advantage of this statistic is that it is defined when $\mu = 0$, and does not have the undesirable properties as $\mu \to 0$ described by Væth (1985). Returning to the invertebrate data introduced in Section 2, $W_{1/2}$ took large, significant values for both variables, when analyzing each separately,

whereas $W_{2/3}$ had problems for the amphipod data because $\hat{\mu}_C = 0$. Similar results were obtained when fitting GEEs to the bivariate data.

We have demonstrated the relevance of our results for other members of the "Holy Trinity", the score statistic (S) and likelihood ratio statistic ($-2 \log L$). We conclude that the score statistic, like W_1, is an "extreme statistic" which should not be used for unbalanced designs, unless testing a one-sided alternative for which it is known to be favorable. The log-likelihood ratio statistic, like $W_{2/3}$, is an "intermediate statistic" which has reasonably high power across a broad range of alternatives. Hence $-2 \log L$ is a natural choice of test statistic for generalized linear models, as it will have more reliable power than the other commonly used statistics S and W_1.

Our conclusions contrast with those of Barnwal and Paul (1988), who recommended the use of a score statistic in place of a log-likelihood statistic in the analysis of overdispersed count data, because of more reliable Type I error properties. We reached a very different conclusion, by controlling Type I error and comparing statistics on the basis of power. We emphasize that our results are only relevant for situations where one either has a large enough sample size for Type I error properties to be good, or where one uses resampling to ensure valid test sizes.

Our results for S and W_1 are analogous to results found in the analysis of variance literature for the Behrens-Fisher problem (Miller, 1986). An ANOVA F statistic tends to take smaller values when more variable groups are more intensively sampled, and larger values when less intensively sampled, just as S and W_1 did in this paper. In fact, it can be shown that the score statistic is proportional to an ANOVA statistic in one-way classifications (Cordeiro et al., 1994, page 715, for example). Hence it is

unsurprising that S has poor properties when variances differ considerably across groups, for unbalanced samples, and that these properties co-incide with those of an ANOVA F-statistic in the same scenario.

It is unclear exactly how to generalize these results beyond k-sample tests. As mentioned previously, in the k-sample case, changing the link function $h_\alpha(\mu)$ does not change the model. This is not the case in general. While we can still reparameterize β without changing the model, this can not be done directly through the link function, which complicates the situation.

It is currently standard practice when using GEE's to base hypothesis tests either on the score statistic (S) or the Wald statistic under the canonical parameterization (W_1). Our results suggest that certainly in the case of k-sample testing of GEE's, these statistics should not be used as the standard test statistics, because they cannot be relied upon to have good power properties when sampling is unbalanced. This is especially the case when dealing with overdispersed counts, for which variances can differ considerably under the alternative, exacerbating the poor properties of S and W_1. For this situation, we propose using the Wald statistic based on the variance stabilizing transformation ($W_{1/2}$), or when it is defined, the log-likelihood ratio statistic ($-2\log L$). A priority in future research is to establish whether or not the undesirable properties of W_1 and S observed here can also be seen in more general marginal models – and if so, what statistics can be used in their place.

Acknowledgements

Thanks to Malcolm Hudson for advice and many helpful discussions. This work was partially completed while visiting the Department of Statistics at Macquarie University.

REFERENCES

Barndorff-Nielsen, O. E., Cox, D. R., 1994. Inference and asymptotics. Chapman & Hall, London.

Barnwal, R. K., Paul, S. R., 1988. Analysis of one-way layout of count data with negative binomial variation. Biometrika 75, 215–222.

Cordeiro, G. M., Botter, D. A., Ferrari, S. L. D. P., 1994. Nonnull asymptotic distributions of three classic criteria in generalised linear models. Biometrika 81 (4), 709–720.

DiCiccio, T. J., 1984. On parameter transformations and interval estimation. Biometrika 71, 477–485.

Drum, M., McCullagh, P., 1993. Regression models for discrete longitudinal responses: Comment. Statistical Science 8 (3), 300–301.

Ferrari, S. L. P., Botter, D. A., Cribari-Neto, F., 1997. Local power of three classic criteria in generalised linear models with unknown dispersion. Biometrika 84 (2), 482–485.

Guo, X., Pan, W., Connett, J. E., Hannan, P. J., French, S. A., 2005. Small-sample performance of the robust score test and its modifications in generalized estimating

equations. Statistics in Medicine 24, 3479–3495.

Harris, P., Peers, H. W., 1980. The local power of the efficient scores test statistic. Biometrika 67 (3), 525–529.

Hayakawa, T., 1975. The likelihood ratio criterion for a composite hypothesis under a local alternative. Biometrika 62 (2), 451–460.

Hougaard, P., 1982. Parameterizations of non-linear models. Journal of the Royal Statistical Society B 44, 244–252.

Johnson, N. L., Kotz, S., Balakrishnan, N., 1997. Discrete Multivariate Distributions. John Wiley & Sons, New York.

Mancl, L. A., DeRouen, T. A., 2001. A covariance estimator for GEE with improved small-sample properties. Biometrics 57 (1), 126–134.

McCullagh, P., Nelder, J. A., 1989. Generalized linear models, 2nd Edition. Chapman & Hall, London.

Miller, Jr, R. G., 1986. Beyond ANOVA, basics of applied statistics. John Wiley & Sons, New York.

Peers, H. W., 1971. Likelihood ratio and associated test criteria. Biometrika 58 (3), 577–587.

Petersen, K. B., Pedersen, M. S., 2006. The matrix cookbook, version February 16, 2006. Tech. rep., `http://matrixcookbook.com`.

Pierce, D. A., Peters, D., 1992. Practical use of higher order asymptotics for multiparameter exponential families. Journal of the Royal Statistical Society B 54 (3), 701–737.

Rao, C. R., 1948. Large sample tests of statistical hypotheses concerning several parameters with applications to problems of estimation. Proceedings of the Cam-

bridge Philosophical Society 44, 50–57.

Rotnitzky, A., Jewell, N. P., 1990. Hypothesis testing of regression parameters in semiparametric generalized linear models for cluster correlated data. Biometrika 77, 485–497.

Slate, E. H., 1994. Parameterizations for natural exponential families with quadratic variance functions. Journal of the American Statistical Association 89, 1471–1482.

Væth, M., 1985. On the use of Wald's test in exponential families. International Statistical Review 53, 199–214.

Warton, D. I., Hudson, H. M., 2006. A generalised linear model parameterisation matching Wald and likelihood ratio statistics to order N^{-1}. Tech. rep., Department of Statistics, Macquarie University.

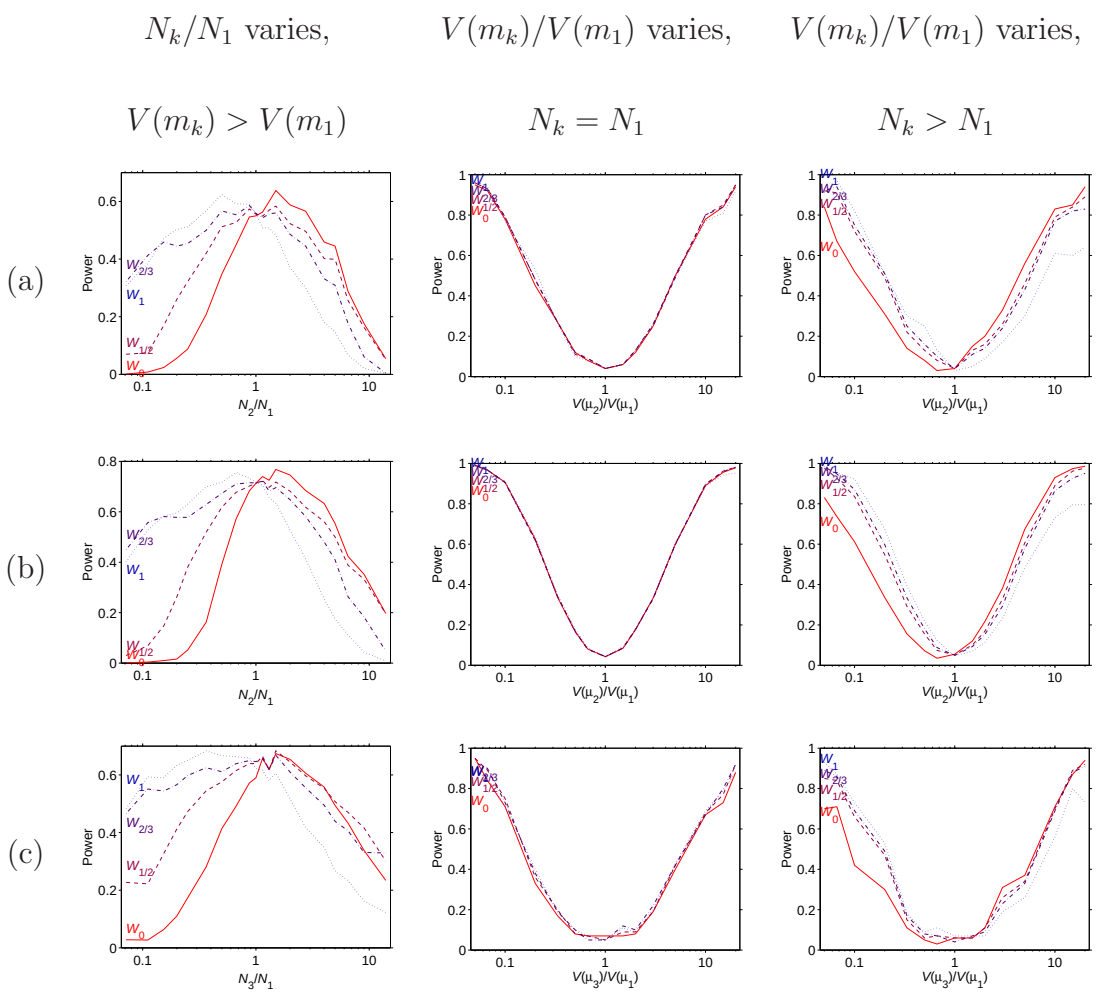

Fig. 1. Power of four parameterizations of the Wald statistic. Simulations vary the sample size while holding means and variances fixed (first column), or vary the variance ratio whilst holding sample sizes fixed in a balanced (second column) or unbalanced (third column) sampling design. The data generating mechanism was varied in different rows: (a) Two sample negative binomial simulations; (b) Two sample bivariate negative binomial simulations; (c) Three sample negative binomial simulations.

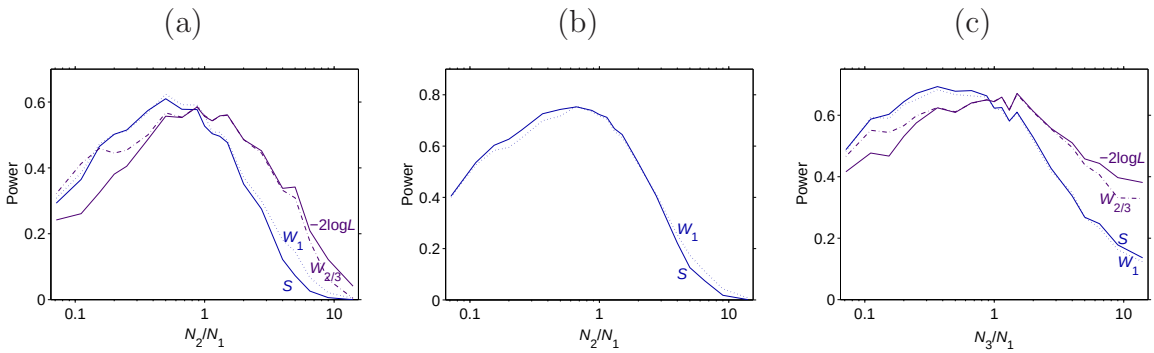

Fig. 2. Comparison of the power of S with that of W_1, and comparison of $-2\log L$ with $W_{2/3}$, in simulations varying the sample size ratio whilst holding means and variances fixed. In all cases, $m_1 < m_2$ $(< m_3)$, similarly for variances. (a) Two sample negative binomial simulations; (b) Two sample bivariate negative binomial simulations; (c) Three sample negative binomial simulations. Note that $-2\log L$ is undefined for GEE's, so in (b) only S and W_1 are shown.

Table 1

Counts of the abundance of amphipods, order Amphipoda*, and cockroaches, order* Blattodea*, obtained from 10 sites near Sydney, Australia. Two sites were controls (C), eight had undergone regeneration (R).*

Site	C1	C2	R1	R2	R3	R4	R5	R6	R7	R8
Amphipoda abundance	0	0	156	31	1	52	376	159	21	11
Blattodea abundance	3	4	0	0	0	1	0	0	0	0

Table 2

Two-sample test statistics and their significance, for the invertebrate data of Table 1. We fitted a negative binomial log-linear model, as described in the text, and compared the log-likelihood ratio statistic ($-2\log L$), the score statistic (S) and the Wald statistic under the canonical parameterization (W_1). P-values were obtained using exact permutation tests.

Statistic:	$-2\log L$	S	W_1
Amphipoda abundance			
Observed value	13.9	1.12	?
P	0.022	0.222	?
Blattodea abundance			
Observed value	17.0	8.37	9.72
P	0.022	0.022	0.022