Positively Skewed Data: Revisiting the Box-Cox Power Transformation

Author:
Olivier, Jake; Norberg, Melissa

Publication details:
International Journal of Psychological Research
v. 3
Chapter No. 1
pp. 69-78
2011-2079 (ISSN)

Publication Date:
2010

International Journal of Psychological Research, 2010. Vo. 3. No. 1.
ISSN impresa (printed) 2011-2084
ISSN electrónica (electronic) 2011-2079

Olivier J., Norberg, M. M., (2010). Positively Skewed Data: Revisiting the Box-Cox Power Transformation. *International Journal of Psychological Research, 3*(1), 69-78.

Positively Skewed Data: Revisiting the Box-Cox Power Transformation.

Datos positivamente asimétricos: revisando la transformación Box-Cox.

Jake Olivier
University of New South Wales
Melissa M. Norberg
University of New South Wales

ABSTRACT

Although the normal probability distribution is the cornerstone of applying statistical methodology; data do not always meet the necessary normal distribution assumptions. In these cases, researchers often transform non-normal data to a distribution that is approximately normal. Power transformations constitute a family of transformations, which include logarithmic and fractional exponent transforms. The Box-Cox method offers a simple method for choosing the most appropriate power transformation. Another option for data that is positively skewed, often used when measuring reaction times, is the Ex-Gaussian distribution which is a combination of the exponential and normal distributions. In this paper, the Box-Cox power transformation and Ex-Gaussian distribution will be discussed and compared in the context of positively skewed data. This discussion will demonstrate that the Box-Cox power transformation is simpler to apply and easier to interpret than the Ex-Gaussian distribution.

Key words: Logarithmic transformations, geometric mean analysis, ex-Gaussian distribution, log-normal distribution.

RESUMEN

Aunque la distribución normal es la piedra angular de las aplicaciones estadísticas, los datos no siempre se ajustan a los criterios de la distribución normal. En tales casos, los investigadores a menudo transforman los datos no normales en datos que siguen una distribución aproximadamente normal. Las transformaciones de potencia constituyen una familia de transformaciones que incluye las transformaciones logarítmicas y fraccional exponente. El método de Box-Cox ofrece un método simple para elegir la transformación de potencia más apropiada. Otra opción que usa cuando los datos son positivamente asimétricos, e.g., los tiempos de reacción, es la distribución Ex-Gaussiana que es una combinación de las distribuciones exponenciales y normal. En este artículo, se discuten la transformación de potencia Box-Cox y la distribución Ex-Gaussiana en relación con datos positivamente asimétricos. La discusión demuestra que la transformación Box-Cox es más sencilla de aplicar e interpretar que la distribución Ex-Gaussiana.

Palabras clave: transformaciones logarítmicas, análisis de la media geométrica, distribución exponencial Gaussiana, distribución logarítmica normal.

Artículo recibido/Article received: December 15, 2009/Diciembre 15, 2009, Artículo aceptado/Article accepted: March 15, 2009/Marzo 15/2009
Dirección correspondencia/Mail Address: j.olivier@unsw.edu.au
Jake Olivier, School of Mathematics and Statistics, NSW Injury Risk Management Research Centre, University of New South Wales, Sydney NSW 2052, Australia,
Melissa M. Norberg, National Cannabis Prevention and Information Centre, University of New South Wales, Randwick NSW 2031, Australia

INTERNATIONAL JOURNAL OF PSYCHOLOGICAL RESEARCH esta incluida en PSERINFO, CENTRO DE INFORMACION PSICOLOGICA DE COLOMBIA, OPEN JOURNAL SYSTEM, BIBLIOTECA VIRTUAL DE PSICOLOGIA (ULAPSY-BIREME), DIALNET y GOOGLE SCHOLARS. Algunos de sus articulos aparecen en SOCIAL SCIENCE RESEARCH NETWORK y está en proceso de inclusion en diversas fuentes y bases de datos internacionales.
INTERNATIONAL JOURNAL OF PSYCHOLOGICAL RESEARCH is included in PSERINFO, CENTRO DE INFORMACIÓN PSICOLÓGICA DE COLOMBIA, OPEN JOURNAL SYSTEM, BIBLIOTECA VIRTUAL DE PSICOLOGIA (ULAPSY-BIREME), DIALNET and GOOGLE SCHOLARS. Some of its articles are in SOCIAL SCIENCE RESEARCH NETWORK, and it is in the process of inclusion in a variety of sources and international databases.

International Journal of Psychological Research, 2010. Vo. 3. No. 1.
ISSN impresa (printed) 2011-2084
ISSN electrónica (electronic) 2011-2079

Olivier J., Norberg, M. M., (2010). Positively Skewed Data: Revisiting the Box-Cox Power Transformation. *International Journal of Psychological Research, 3*(1), 69-78.

INTRODUCTION

Numerous significance tests assume data are normally distributed such as t-tests, chi-square tests and F-tests. This is often reasonable, as many real-world measurements/observations follow a normal distribution; however, there are several situations in which the normal distribution assumption is not appropriate (e.g., immunologic and reaction time data). In these cases, data transformation is a common technique used to modify non-normal data to a distribution that makes the normal assumption more reasonable and, in turn, makes significance tests based on a normal assumption more appropriate (Olivier, Johnson & Marshall, 2008).

The normal distribution, sometimes called a *Gaussian distribution*, is characterised by a symmetric, bell-like shape. Significance tests based on a normal assumption are not appropriate for asymmetric data. This is because skewness is most reflected in the tails of distributions, which are where p-values are calculated. This usually results in p-values that are less than expected (and thus more likely to be incorrectly significant). When a data set does not follow the shape of a normal distribution, an appropriate function is sometimes chosen that transforms the data to a distribution that is reasonably normal-shaped.

A common misconception in statistics is that data must be sampled from a normal distribution for significance tests based on a normal assumption to be appropriate. In truth, the normal assumption applies to the distribution of the sample mean $\bar{X}$, called a *sampling distribution*, and not the distribution from which the data are sampled. There is partial truth in the misconception because data that are sampled from a normal distribution implies that $\bar{X}$ is also normally distributed. When data is not sampled from a normal distribution, the central limit theorem (CLT) ensures that $\bar{X}$ is approximately normally distributed for a large enough sample size. Although many texts mention sample sizes above 30 constitutes a large enough sample size for the CLT to apply, there exists no "magic" sample size for every situation. There are instances where data sampled from a highly skewed data set would require a sample of size of 1000 for $\bar{X}$ to be approximately normally distributed. In practice, a data analyst would be wise to check for normality with an understanding that data need only look reasonably normal to be suitable. A visual method for testing normality will be discussed later in the paper.

There are a few broad, well-established guidelines for choosing the most appropriate transformation. For instance, a logarithmic transformation is recommended for positively skewed data, while negatively skewed data is often transformed using the square root function. However, these guidelines are not appropriate for every situation. Box and Cox (1964) introduced a method for choosing the best transformation from a family of power transformations. In this instance, the data is raised to a power chosen to best approximate a normal distribution. If the data has been transformed to a distribution that is reasonably normal, the data analyst would then perform significance test(s) on the transformed data using methods based on a normal assumption.

The ex-Gaussian distribution is another method, often used for non-negative, positively skewed data. This distribution is defined by three parameters; and, an ex-Gaussian analysis involves the estimation of these parameters usually by either method of moments or maximum likelihood estimation (Heathcote, 1996). It should be noted that this is only a method for estimating a probability distribution and does not lend itself easily to statistical inference.

Another option is to utilise nonparametric statistical techniques such as Mann-Whitney or Kruskal-Wallis tests. Nonparametric methods do not make explicit distributional assumptions and are less powerful than parametric tests when a distributional assumption is reasonable (nonparametric tests are more powerful when distributional assumptions are inappropriate for parametric analysis). A full discussion of nonparametric statistics is beyond the scope of this paper. Conover (1998) provides a good overview of nonparametric statistics.

This paper will first introduce basic statistical methodology that is essential in understanding the transformation of data and the ex-Gaussian distribution, then the Box-Cox family of transformations and methods for estimating the ex-Gaussian distribution will be introduced, and will conclude with a discussion regarding the usefulness of both methods.

STATISTICAL BACKGROUND

Symmetric data are often best described by a measure of centre (e.g., mean) and by a measure of spread (e.g., standard deviation). Skewed data, on the other hand, benefit from describing the direction (positive or negative) and magnitude of skew. Describing the skewness of a data set is a concept that is often overlooked. In theoretical terms, a distribution's skew is defined as

$$\gamma_1 = \frac{\mu_3}{\sigma^3}$$

Olivier J., Norberg, M. M., (2010). Positively Skewed Data: Revisiting the Box-Cox Power Transformation. *International Journal of Psychological Research, 3*(1), 69-78.

where σ is the standard deviation and μ_3 is the third moment about the mean (variance is the second moment about the mean, i.e., $\mu_2 = \sigma^2$; and $\gamma_2 = \mu_4/\sigma^4$ is known as *kurtosis*). Kurtosis is a measure of the thickness of a distribution's tails and thus affects how p-values are computed. However, kurtosis is not as important as skewness when performing normal distribution-based significance tests such as the t-test. T-tests are protective against excess kurtosis (i.e., heavier tails) because probabilities computed in the tails of the t-distribution are larger than those from the normal distribution making the analysis more conservative.

Values of γ_1 can be either negative or positive which correspond respectively to negative or positive skewness (sometimes referred to as left and right skewed respectively). Skewness is different for each probability distribution and can be computed from a distribution's probability density function. To get an understanding about how to analyse a skewed data set, we will next discuss the properties of four probability distribution functions, namely the normal, log-normal, exponential and ex-Gaussian distributions. This is an abbreviated description and the authors recommend Wackerly, Mendenhall and Scheaffer (2007) for a more complete discussion of mathematical statistics.

The Normal Distribution

The first probability distribution that will be discussed is the normal distribution, the backbone of statistical analysis. The normal distribution (i.e., Gaussian distribution) is a symmetric distribution characterised by a bell-like shape with a probability density function (pdf) that can be written as

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{\sigma^2}}, \quad -\infty < x < \infty$$

where the mean μ and variance σ^2 uniquely define the distribution of the random variable X. A probability distribution is often written as a pdf because this function is used to graph the probability distribution as a smooth curve for continuous data and useful information such as the mean, variance and skewness are computed from this function.

The normal distribution has no skew (i.e., $\gamma_1 = 0$), so it is a very poor choice to model positively skewed data. On the other hand, non-normal data can often be transformed to a distribution that is reasonably normal. As mentioned above, transforming a non-normal set of data to a distribution that is reasonably normal can be very useful as many significance tests rely on the assumption $\bar{X}$ is normally distributed (which is certainly true when data has been sampled from a normal distribution). These tests include various forms of t-tests and analyses of variance (ANOVA). There are even several non-parametric tests that rely on a relaxed normality assumption (Mann-Whitney, Kruskal-Wallis and chi-square tests).

The Log-normal Distribution

A variant of the normal distribution is the log-normal distribution. In statistical jargon, a random variable W with parameters μ and σ has a log-normal distribution if the random variable $U = \ln(W)$ has a normal distribution, where ln is the logarithm with base e, i.e. $\log_e$ and sometimes called the *natural logarithm*. In broad terms, a log-normal distribution is a distribution that is normal when log-transformed. The pdf of the log-normal distribution can be written as

$$f(w) = \frac{1}{w\sigma\sqrt{2\pi}} e^{-\frac{(\ln(w)-\mu)^2}{\sigma^2}}, \quad 0 < w < \infty$$

where μ and σ are the mean and standard deviation of the normal random variable U respectively. The mean and variance of the log-normal random variable W are

$$\mu_W = e^{\mu+\sigma^2/2} \text{ and } \sigma_W^2 = \mu_W^2\left(e^{\sigma^2} - 1\right).$$

Unlike the normal distribution, the log-normal distribution is positively skewed with $\gamma_1 = \left(e^{\sigma^2} + 2\right)\sqrt{e^{\sigma^2} - 1}$, a value that is always positive when $\sigma > 0$. In practice, positively skewed data are often transformed using logarithms in an effort to make the data more reasonably normally distributed due to the relationship between the normal and log-normal distributions.

The Ex-Gaussian Distribution

The ex-Gaussian distribution is aptly named as it is the sum of two independent probability distributions, the exponential and the normal. The exponential distribution is a positively skewed distribution whose pdf for a random variable Y is

$$f(y) = \frac{1}{\lambda} e^{-y/\lambda}, \quad 0 < y < \infty$$

where the mean λ uniquely defines this distribution. The variance and skewness of the exponential distribution are λ^2 and 2 respectively.

If X and Y represent independent normal and exponential random variables (as defined above), then

$Z = X + Y$ is an ex-Gaussian random variable with mean and variance

$$\mu_Z = \mu + \lambda \quad \text{and} \quad \sigma_Z^2 = \sigma^2 + \lambda^2 .$$

The skewness parameter for Z is $\gamma_1 = 2\lambda^3/(\sigma^2 + \lambda^2)^{3/2}$ which can be shown to never exceed 2, the skewness of the exponential distribution. The pdf of the ex-Gaussian random variable Z is

$$f(z) = \frac{1}{\lambda}\Phi\left(\frac{z-\mu}{\sigma} - \frac{\sigma}{\lambda}\right) e^{\left\{\frac{\sigma^2}{2\lambda^2} - \frac{z-\mu}{\lambda}\right\}}, \qquad -\infty < x < \infty$$

where $\Phi(\cdot)$ is the cumulative distribution function of the standard normal distribution. Notably, the pdf of the ex-Gaussian distribution can be viewed as a weighted average of the shifted exponential distribution, where

$$\frac{1}{\lambda} e^{-(z-\mu)/\lambda}$$

is an exponential distribution shifted by the normal mean parameter μ and weight

$$\Phi\left(\frac{z-\mu}{\sigma} - \frac{\sigma}{\lambda}\right) e^{\sigma^2/2\lambda^2}$$

which, in a sense, is similar to a probability from the normal distribution. As the name implies, the shifted exponential distribution is an exponential distribution that has been horizontally displaced by the shift parameter. In statistical shorthand, which will be used throughout the remainder of the text, the ex-Gaussian distribution can be written as

$$W \sim exGaussian(\mu, \sigma, \lambda).$$

For illustrative purposes, data was simulated first from a normal distribution with $\mu = 10$ and $\sigma = 2$ as well as an exponential distribution with $\lambda = 5$. The sum of values from these data sets is an ex-Gaussian distribution which has skewness

$$\gamma_1 = 2(5)^3/(2^2 + 5^2)^{3/2} \approx 1.60.$$

Histograms of these distributions are given in Figure 1.

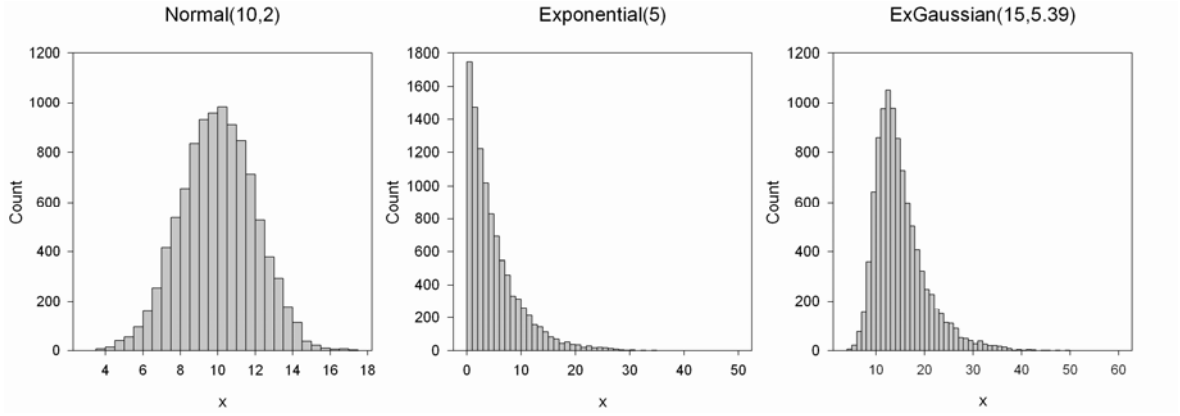


Figure 1: *Histograms of normal and exponential distributions, and the distribution of their sum (ex-Gaussian distribution)*

Box-Cox Transformation and Estimating the Ex-Gaussian Distribution

The previous section discussed many important statistical concepts for dealing with positively skewed data. In this section, we will discuss methods for dealing with a realised sample of observations, say $y_{1,}y_{2,}, \ldots, y_n$, that are possibly positively skewed.

The skewness of a data set can be checked by inspection, using either a histogram or a quantile-quantile (Q-Q) plot, or by estimating the skewness parameter. There is no shortage of software packages that can create histograms, so it will not be discussed in this paper. A Q-Q plot is a method of comparing sample data with a known distribution such as the normal distribution. The first step in creating a Q-Q plot is to compute the mean and standard deviation of a sample, namely

$$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i \quad \text{and} \quad s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2} .$$

The next step is to sort the data in ascending order and create a scatterplot of the values

$$\left(y_{1,}\Phi^{-1}\left(\frac{1}{n+1}\right)\right), \left(y_{2,}\Phi^{-1}\left(\frac{2}{n+1}\right)\right), \ldots, \left(y_{n,}\Phi^{-1}\left(\frac{n}{n+1}\right)\right)$$

where $\Phi^{-1}(\cdot)$ is the inverse cumulative distribution function of the standard normal distribution. If the sample data follows a normal distribution, these values will lie in a straight line. A reference line passing through the points

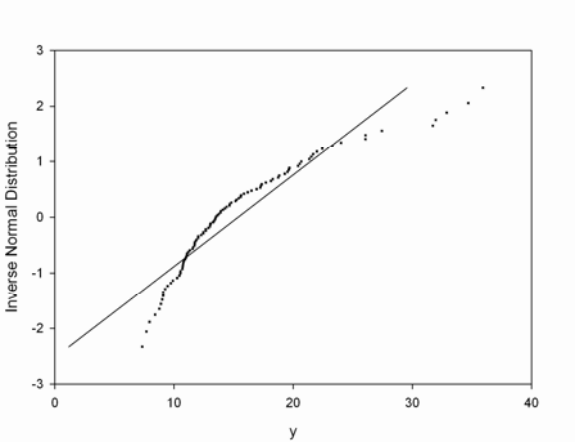
$$\left(\bar{y} + s\Phi^{-1}\left(\frac{1}{n+1}\right), \Phi^{-1}\left(\frac{1}{n+1}\right)\right)$$

and

$$\left(\bar{y} + s\Phi^{-1}\left(\frac{n}{n+1}\right), \Phi^{-1}\left(\frac{n}{n+1}\right)\right)$$

is commonly inserted into a Q-Q plot. This line represents an exact fit to the normal distribution. For demonstrative purposes, a Q-Q plot was created using 100 observations simulated from an ex-Gaussian distribution with parameters 10, 2 and 5 (as above). The simulated data is given in appendix 1.

Figure 2: *Quantile-quantile plot for data simulated from* $exGaussian(10,2,5)$



The scatterplot shown in Figure 2 indicates the sample does not follow a normal distribution very well. In fact, the specific curvature of points indicates the data is positively skewed.

The skewness parameter also can be estimated. The method of moments estimator for skew is

$$\widehat{\gamma_1} = \frac{1}{s^3(n-1)}\sum_{i=1}^{n}(y_i - \bar{y})^3 .$$

For the simulated data set, the estimate of skewness is $\widehat{\gamma_1} \approx 1.46$. Since the normal distribution has no skewness, i.e. $\gamma_1 = 0$, values of $\widehat{\gamma_1}$ near 0 are indicative of normality and values far from 0 indicate skewness. Brown (1997) gives some discussion regarding skewness; however, there is much debate regarding values of $\widehat{\gamma_1}$ that indicate a distribution is skewed (i.e., how far should $\widehat{\gamma_1}$ be from 0 to indicate skewness).

Box-Cox Transformation

Box and Cox (1964) introduced a family of transformations as powers of the variable y, and is often referred to as *power transformations*. The goal is to transform non-normal data to a data set that is reasonably normal. Its simplest form is given by

$$y(\nu) = \begin{cases} \dfrac{y^{\nu} - 1}{\nu} & \nu \neq 0 \\ \ln(y) & \nu = 0 \end{cases}$$

where ν is chosen to best represent a transformation to the normal distribution. Note that $\nu = 1/2$ and $\nu = 0$ correspond to the square root and logarithmic transformation respectively. A commonly used method is to choose values of ν from -3 to 3 in increments of 0.25 (SAS Institute Inc., 2008). Then for each choice of ν, a regression line is fit through the points in a Q-Q plot using the method discussed above. The resulting regression lines are then compared using either the coefficient of determination R^2 or the log-likelihood statistic (larger is better in both cases). If the best choice is $\nu = 1$, there is no indication the data should be transformed.

Transforming the data and computing the log-likelihood in each case can be very time consuming. SAS is one of the few statistical packages that will do this for you. SAS uses the PROC TRANSREG procedure to perform a Box-Cox transformation. Sample SAS code using this procedure with the simulated data is given in Appendix 2. Abbreviated results are given in Table 1.

Table 1: *Results from Box-Cox transformations using SAS PROC TRANSREG*

λ	R^2	$log-like$
-2.0	0.90	-58.086
-1.5	0.94	-23.048
-1.0	0.97	8.271
-0.5	0.97	9.586
0.0	0.94	-22.607
0.5	0.89	-61.077
1.0	0.81	-97.276
1.5	0.73	-131.329
2.0	0.64	-164.289

The results from Table 1 suggest the optimal transformation is $y(-0.5) = -2(y^{-0.5} - 1)$. A linear

transformation of a normal random variable is also a normal random variable, so, in practice, the constant term is omitted and the data transformed as $x = y^{-0.5}$.

Log-transformed data is a special case of the Box-Cox transformation (i.e., $\nu = 0$). In this instance, data is transformed as $x_i = \ln(y_i)$ for $i = 1, \ldots, n$. The sample is then analysed using the x_i values. When reporting results from log-transformed data (or data from any other type of transformation), the summary statistics are not easily interpreted on the original scale. Instead, summary statistics computed on the transformed scale are back-transformed to the original scale. For instance, summary statistics from log-transformed data are exponentiated (i.e., $y = e^x$), and summary statistics from square root-transformed data are squared (i.e., $y = x^2$). A statistical analysis using log-transformed data is sometimes referred to as a *geometric mean analysis* because when the sample mean of the transformed data

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{1}{n}\sum_{i=1}^{n} \ln(y_i)$$

is exponentiated, the result is the geometric mean, i.e.,

$$\bar{y}_g = e^{\bar{x}} = e^{\frac{1}{n}\sum_{i=1}^{n} \ln(y_i)} = \sqrt[n]{x_1 \cdot x_2 \cdots x_n}\,.$$

The standard deviation from the transformed data, on the other hand, is not interpretable on the original scale because e^s is not an applicable measure of variability on the original scale. In general, data that is presented as $\bar{x} \pm s$ or $\bar{x} \pm s.e.$ is only useful for symmetric data. A much better option is to back-transform confidence limits. For log-transformed data, a $(1-\alpha)100\%$ confidence interval computed from the transformed data is

$$\bar{x} \pm t_{\frac{\alpha}{2},n-1}\frac{s}{\sqrt{n}}$$

which on the original scale is

$$\left(e^{\bar{x}-t_{\alpha/2,n-1}s/\sqrt{n}}, e^{\bar{x}+t_{\alpha/2,n-1}s/\sqrt{n}}\right).$$

If a log-transformation is appropriate for the ex-Gaussian data, the summary statistics would be reported as $\bar{y} = 14.4$ and 95% confidence interval $(13.41, 15.42)$.

The log-transformation, by design, does not allow for data that can take on negative or zero values since logarithms are undefined on the interval $(-\infty, 0]$. However, a data set can be shifted to the right so that all values are positive. This is especially useful for count data which may have 0 values. The instinctive log-transformation in this instance is $x_i = \ln(y_i + 1)$ for $i = 1, \ldots, n$; however, this shift must be replicated when back-transforming summary statistics where

$$\bar{y}_g = e^{\bar{x}} - 1 \quad \text{and}$$
$$\left(e^{\bar{x}-t_{\alpha/2,n-1}s/\sqrt{n}} - 1, e^{\bar{x}+t_{\alpha/2,n-1}s/\sqrt{n}} - 1\right)$$

since the inverse of this transformation is $y_i = e^{x_i} - 1$. The log-transformation for count data has a very nice interpretation in that values that are 0 on the original scale are also 0 on the log scale.

Estimating Ex-Gaussian Parameters

As mentioned previously, the ex-Gaussian distribution is defined by the parameters from the normal and exponential distributions, specifically μ, σ and λ. The goal in this type of analysis is to estimate these parameters using either method of moments or maximum likelihood estimation. The method of moments estimates for the mean, variance and skewness for any data set are $\bar{y}$, s^2 and $\widehat{\gamma_1}$ respectively. To find estimates for μ, σ and λ, the method of moments estimates are set equal to mean, variance and skewness of the ex-Gaussian distribution, i.e.,

$$\begin{aligned}\bar{y} &= \mu + \lambda \\ s^2 &= \sigma^2 + \lambda^2 \\ \widehat{\gamma_1} &= 2\lambda^3/(\sigma^2 + \lambda^2)^{3/2}\end{aligned}$$

then these three equations are solved simultaneously for μ, σ and λ in terms of $\bar{y}$, s^2 and $\widehat{\gamma_1}$. The resulting estimates for the ex-Gaussian parameters are

$$\begin{aligned}\hat{\mu} &= \bar{y} - s\left(\frac{\widehat{\gamma_1}}{2}\right)^{1/3} \\ \widehat{\sigma^2} &= s^2\left[1 - \left(\frac{\widehat{\gamma_1}}{2}\right)^{2/3}\right] \\ \hat{\lambda} &= s\left(\frac{\widehat{\gamma_1}}{2}\right)^{1/3}\end{aligned}$$

where $\hat{\mu}$, $\hat{\sigma}$ and $\hat{\lambda}$ are estimates of μ, σ and λ respectively. For the simulated data set, the summary statistics are $\bar{y} \approx 15.36$, $s \approx 6.09$ and $\widehat{\gamma_1} \approx 1.46$ which result in ex-Gaussian parameter estimates $\hat{\mu} \approx 9.87$, $\hat{\sigma} \approx 2.64$ and $\hat{\lambda} \approx 5.49$. SAS code for these computations is given in Appendix 3.

The maximum likelihood estimates for the ex-Gaussian parameters are more efficient (smaller variance) than method of moments estimates (Heathcote, 1996); however, this method is beyond the scope of this paper since it does not have a mathematical closed form, is

computationally intensive and optimisation methods are chosen ad hoc.

In terms of reaction time data, the ex-Gaussian distribution is easily interpretable in that it is the sum of two independent constructs with one having a symmetric distribution (Gaussian) and the other being positively skewed (Exponential). The choice of the normal and exponential distributions is also beneficial in that it allows for more than two constructs to influence reaction time since the sum of independent normal random variables is another normal random variable and the sum of independent exponential random variables is another exponential random variable. In statistical terms, if X_1 and X_2 are independent normal random variables with means μ_1 and μ_2 and variances σ_1^2 and σ_2^2, and independent random variables Y_1 and Y_2 are exponential random variables with means λ_1 and λ_2, then if each random variable is pairwise independent the sum of these random variables is an ex-Gaussian distribution, i.e.,

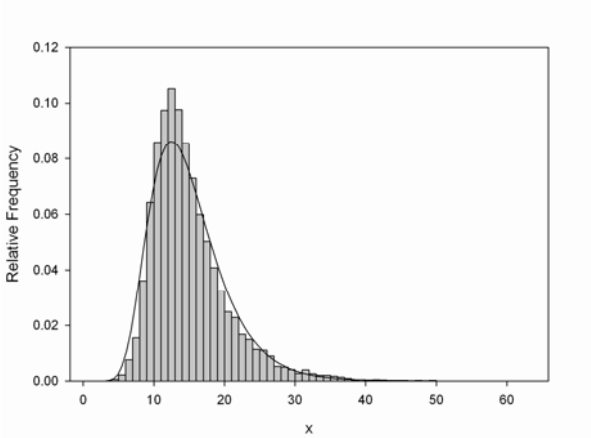
$$X_1 + X_2 + Y_1 + Y_2 \sim exGaussian\left(\mu_1 + \mu_2, \sqrt{\sigma_1^2 + \sigma_2^2}, \lambda_1 + \lambda_2\right).$$

If two constructs influence reaction time in a symmetric way and another in a positively skewed way, estimating an ex-Gaussian distribution would be appropriate. However, the major drawback to the ex-Gaussian is that it does not lend itself very easily to statistical inference unlike normally distributed data (or data that can be transformed to the normal distribution), because commonly used significance tests do not exist for the ex-Gaussian distribution. In practical terms, common statistical techniques such as the t-test or ANOVA are not possible with the ex-Gaussian distribution. Distributional assumptions can be made on the ex-Gaussian parameter estimates to create a significance test such as the likelihood ratio test; however, there are many issues estimating these parameters mentioned above which, in the authors' opinion, makes any procedure questionable.

DISCUSSION

The Box-Cox transformation method and estimation of the ex-Gaussian distribution parameters have been presented for positively skewed data. Although there is a strong theoretical justification for using the ex-Gaussian distribution to estimate reaction time data, it has very little benefit in terms of statistical inference. On the other hand, the Box-Cox transformation, if appropriate, is a method for determining the best power transformation to the normal distribution. Further, it is well known that many significance tests based on a normal assumption are robust to minor deviations from normality (see for example Heeren and D'Agostino, 1987). In other words, if a data set looks reasonably normal then these statistical tests are appropriate even if data is not sampled from a normal distribution. For instance, a histogram of the data simulated from the $exGaussian(10,2,5)$ is given in Figure 3. The curve drawn over the histogram is a log-normal distribution estimated from the data set. Although the data was simulated from an ex-Gaussian distribution, the log-normal would appear to be a reasonable approximation to this distribution. This certainly does not exhaustively prove the Box-Cox transformation is a better method for estimating the distribution of a positively skewed data set than estimating the ex-Gaussian distribution. However, the vast world of statistical inference is open to the data analyst if a power transformation results in reasonably normal data.

Figure 3: *Histogram of simulated data from* $exGaussian(10,2,5)$ *with estimated log-normal distribution overlain*



REFERENCES

Box, G.E.P., & Cox, D. R. (1964). An analysis of transformations. *Journal of the Royal Statistical Society, Series B, 26*, 211–252.

Brown, J. D. (1997). Statistics corner: Questions and answers about language testing statistics: Skewness and kurtosis. *Shiken, 1*, 20-23. Available online at www.jalt.org/test/bro_1.htm. [16 Aug. 1997].

Conover, W.J. (1998). *Practical Nonparametric Statistics*. Hoboken, NJ: Wiley.

Heathcote, A. (1996). RTSYS: A DOS application for the analysis of reaction time data. *Behavior Research Methods, Instruments, and Computers, 28*, 427-445.

Heeren, T. & D'Agostino R. (1987). Robustness of the two independent samples t-test when applied to ordinal scaled data. *Statistics in Medicine, 6*, 79-90.

Olivier, J., Johnson, W.D., Marshall, G.D. (2008). The logarithmic transformation and the geometric mean in reporting experimental IgE results: What are they and when and why to use them? *Annals of Allergy, Asthma and Immunology*, *100*, 333-337. Erratum in: *Annals of Allergy, Asthma and Immunology*, 100, 625-626.

SAS Institute Inc. (2008). *SAS/STAT® 9.2 User's Guide*. Cary, NC: SAS Institute Inc.

Wackerly, D., Mendenhall W., and Scheaffer R. (2007). *Mathematical Statistics with Applications, 7th Edition*. Belmont, CA: Duxbury

Appendix 1: Simulated data from $exGaussian(10,2,5)$

7.320513	10.77892	12.66247	14.80155	19.62323
7.682358	10.81963	12.66346	15.16745	19.67403
7.943175	10.928	12.89001	15.2489	20.41566
8.410734	11.00682	13.02891	15.42959	20.54978
8.763813	11.0686	13.03188	15.59962	20.68714
8.897307	11.09093	13.0755	15.61826	21.3443
9.012962	11.19313	13.2747	15.86461	21.5189
9.081573	11.32054	13.35536	16.18985	21.6779
9.089867	11.56637	13.44562	16.50195	21.98603
9.248693	11.60094	13.47057	16.92947	22.44573
9.461362	11.70802	13.53161	17.2424	23.29426
9.734749	11.73127	13.65484	17.28401	24.01103
9.910837	11.73771	13.76778	17.35752	26.05911
10.26082	11.83094	13.89841	17.70572	26.06802
10.44308	11.89868	13.93583	18.12848	27.44567
10.53079	12.02192	14.10027	18.28239	31.72515
10.54709	12.03848	14.27582	18.74268	31.9653
10.70932	12.27987	14.41111	18.82539	32.88067
10.72175	12.44038	14.65361	19.29076	34.69273
10.74654	12.4649	14.674	19.52527	35.92378

Appendix 2: SAS Code for Box-Cox Transformation

```
/* data entry step */

data exgauss;
input y @@;
datalines;
7.320512959 7.682357787 7.94317508 8.41073393 8.763812698
8.897307347 9.012962076 9.081572559 9.089866975 9.248692656
9.461362445 9.734748626 9.910837237 10.26081667 10.44308261
10.53078925 10.54708712 10.70931577 10.72175008 10.74654487
10.77891849 10.81963062 10.92799959 11.00681543 11.06859594
11.09092509 11.19312526 11.3205358 11.56636891 11.60093853
11.70801564 11.73126699 11.73770999 11.83094198 11.89868016
12.02191788 12.03847837 12.27987236 12.44037757 12.464899
12.66247168 12.66346481 12.89001118 13.02890989 13.03188066
13.0754981 13.27470156 13.35536191 13.44562317 13.47057383
13.53160551 13.65484362 13.76777744 13.89840783 13.9358281
14.10026604 14.27581726 14.41110619 14.65360921 14.67400029
14.80154825 15.1674475 15.24889965 15.42958947 15.59962124
15.61826302 15.86460957 16.18984529 16.50195202 16.92947431
17.24240075 17.28401156 17.35752387 17.70571722 18.12847532
18.28239245 18.74267966 18.82539369 19.29075742 19.52527166
19.62322993 19.67402743 20.41566003 20.54977655 20.68713776
21.34429578 21.51889836 21.67790301 21.9860314 22.44573172
23.29425731 24.01102625 26.059114 26.06802137 27.4456749
31.72515393 31.96530324 32.88066716 34.6927323 35.92378141
;

/* create values from standard normal distribution */

data percentages;
do i = 1 to 100;
      pct = i/(100+1);
      z = probit(pct);
      output;
end;
keep pct z;
run;

/* sort data in ascending order and merge data sets */

proc sort data=exgauss out=exgauss;
by y;
run;
data boxcox;
merge exgauss percentages;
run;

/* Box-Cox transformations for lambda=-2,-1.5,...,2 */

proc transreg data=boxcox;
model boxcox(y / lambda=-2 to 2 by 0.5) = identity(z);
run;
```

Appendix 3: SAS Code for Method of Moments Estimates of Ex-Gaussian Distribution

```
/* Summary Statistics */

proc means mean std skew data=simdata;
var y;
output out=sumstats mean=m std=s skew=sk;
run;

/* Compute Method of Moments Estimates */

data exgauss;
set sumstats;
mu = m - s*(sk/2)**(1/3);
sigma = s*sqrt(1-(sk/2)**(2/3));
sigma2 = sigma**2;
lambda = s*(sk/2)**(1/3);
run;
proc print data=exgauss;
var mu sigma sigma2 lambda;
run;
```