Practical Assessment, Research & Evaluation

A peer-reviewed electronic journal.

Copyright is retained by the first or sole author, who grants right of first publication to the *Practical Assessment, Research & Evaluation*. Permission is granted to distribute this article for nonprofit, educational purposes if it is copied in its entirety and the journal is credited.

Volume 15, Number 12, October, 2010 ISSN 1531-7714

Improving your data transformations: Applying the Box-Cox transformation

Jason W. Osborne, *North Carolina State University*

Many of us in the social sciences deal with data that do not conform to assumptions of normality and/or homoscedasticity/homogeneity of variance. Some research has shown that parametric tests (e.g., multiple regression, ANOVA) can be robust to modest violations of these assumptions. Yet the reality is that almost all analyses (even nonparametric tests) benefit from improved the normality of variables, particularly where substantial non-normality is present. While many are familiar with select traditional transformations (e.g., square root, log, inverse) for improving normality, the Box-Cox transformation (Box & Cox, 1964) represents a family of power transformations that incorporates and extends the traditional options to help researchers easily find the optimal normalizing transformation for each variable. As such, Box-Cox represents a potential best practice where normalizing data or equalizing variance is desired. This paper briefly presents an overview of traditional normalizing transformations and how Box-Cox incorporates, extends, and improves on these traditional approaches to normalizing data. Examples of applications are presented, and details of how to automate and use this technique in SPSS and SAS are included.

Data transformations are commonly-used tools that can serve many functions in quantitative analysis of data, including improving normality of a distribution and equalizing variance to meet assumptions and improve effect sizes, thus constituting important aspects of data cleaning and preparing for your statistical analyses. There are as many potential types of data transformations as there are mathematical functions. Some of the more commonly-discussed traditional transformations include: adding constants, square root, converting to logarithmic (e.g., base 10, natural log) scales, inverting and reflecting, and applying trigonometric transformations such as sine wave transformations.

While there are many reasons to utilize transformations, the focus of this paper is on transformations that improve normality of data, as both parametric and nonparametric tests tend to benefit from normally distributed data (e.g., Zimmerman, 1994, 1995, 1998). However, a cautionary note is in order. While transformations are important tools, they should be utilized thoughtfully as they fundamentally alter the nature of the variable, making the interpretation of the results somewhat more complex (e.g., instead of predicting student achievement test scores, you might be predicting the natural log of student achievement test scores). Thus, some authors suggest reversing the transformation once the analyses are done for reporting of means, standard deviations, graphing, etc. This decision ultimately depends on the nature of the hypotheses and analyses, and is best left to the discretion of the researcher.

Unfortunately for those with data that do not conform to the standard normal distribution, most statistical texts provide only cursory overview of best practices in transformation. Osborne (2002, 2008a) provides some detailed recommendations for utilizing traditional transformations (e.g., square root, log, inverse), such as anchoring the minimum value in a distribution at exactly 1.0, as the efficacy of some transformations are severely degraded as the minimum deviates above 1.0 (and having values in a distribution

less than 1.0 can cause mathematical problems as well). Examples provided in this paper will revisit previous recommendations.

The focus of this paper is streamlining and improving data normalization that should be part of a routine data cleaning process. For those researchers who routinely clean their data, Box-Cox (Box & Cox, 1964; Sakia, 1992) provides a family of transformations that will optimally normalize a particular variable, eliminating the need to randomly try different transformations to determine the best option. Box and Cox (1964) originally envisioned this transformation as a panacea for simultaneously correcting normality, linearity, and homoscedasticity. While these transformations often improve all of these aspects of a distribution or analysis, Sakia (1992) and others have noted it does not *always* accomplish these challenging goals.

Why do we need data transformations?

Many statistical procedures make two assumptions that are relevant to this topic: (a) an assumption that the variables (or their error terms, more technically) are normally distributed, and (b) an assumption of homoscedasticity or homogeneity of variance, meaning that the variance of the variable remains constant over the observed range of some other variable. In regression analyses this second assumption is that the variance around the regression line is constant across the entire observed range of data. In ANOVA analyses, this assumption is that the variance in one cell is not significantly different from that of other cells. Most statistical software packages provide ways to test both assumptions.

Significant violation of either assumption can increase your chances of committing either a Type I or II error (depending on the nature of the analysis and violation of the assumption). Yet few researchers test these assumptions, and fewer still report correcting for violation of these assumptions (Osborne, 2008b). This is unfortunate, given that in most cases it is relatively simple to correct this problem through the application of data transformations. Even when one is using analyses considered "robust" to violations of these assumptions or non-parametric tests (that do not explicitly assume normally distributed error terms), attending to these issues can improve the results of the analyses (e.g., Zimmerman, 1995).

How does one tell when a variable is violating the assumption of normality?

There are several ways to tell whether a variable deviates significantly from normal. While researchers tend to report favoring "eyeballing the data," or visual inspection of either the variable or the error terms (Orr, Sackett, & DuBois, 1991), more sophisticated tools are available, including tools that statistically test whether a distribution deviates significantly from a specified distribution (e.g., the standard normal distribution). These tools range from simple examination of skew (ideally between -0.80 and 0.80; closer to 0.00 is better) and kurtosis (closer to 3.0 in most software packages, closer to 0.00 in SPSS) to examination of P-P plots (plotted percentages should remain close to the diagonal line to indicate normality) and inferential tests of normality, such as the Kolmorogov-Smirnov or Shapiro-Wilk's W test (a $p > .05$ indicates the distribution does *not* differ significantly from the standard normal distribution; researchers wanting more information on the K-S test and other similar tests should consult the manual for their software (as well as Goodman, 1954; Lilliefors, 1968; Rosenthal, 1968; Wilcox, 1997)).

Traditional data transformations for improving normality

Square root transformation. Most readers will be familiar with this procedure-- when one applies a square root transformation, the square root of every value is taken (technically a special case of a power transformation where all values are raised to the one-half power). However, as one cannot take the square root of a negative number, a constant must be added to move the minimum value of the distribution above 0, preferably to 1.00. This recommendation from Osborne (2002) reflects the fact that numbers above 0.00 and below 1.0 behave differently than numbers 0.00, 1.00 and those larger than 1.00. The square root of 1.00 and 0.00 remain 1.00 and 0.00, respectively, while numbers above 1.00 always become smaller, and numbers between 0.00 and 1.00 become *larger* (the square root of 4 is 2, but the square root of 0.40 is 0.63). Thus, if you apply a square root transformation to a continuous variable that contains values between 0 and 1 as well as above 1, you are treating some numbers differently than others, which may not be desirable. Square root transformations are traditionally thought of as good for normalizing Poisson distributions (most common with

data that are counts of occurrences, such as number of times a student was suspended in a given year or the famous example of the number of soldiers in the Prussian Cavalry killed by horse kicks each year (Bortkiewicz, 1898) presented below) and equalizing variance.

Log transformation(s). Logarithmic transformations are actually a class of transformations, rather than a single transformation, and in many fields of science log-normal variables (i.e., normally distributed after log transformation) are relatively common. Log-normal variables seem to be more common when outcomes are influenced by many independent factors (e.g., biological outcomes), also common in the social sciences.

In brief, a logarithm is the power (exponent) a base number must be raised to in order to get the original number. Any given number can be expressed as y^x in an infinite number of ways. For example, if we were talking about base 10, 1 is 10^0, 100 is 10^2, 16 is $10^{1.2}$, and so on. Thus, $\log_{10}(100)=2$ and $\log_{10}(16)=1.2$. Another common option is the Natural Logarithm, where the constant e (2.7182818…) is the base. In this case the natural log of 100 is 4.605. As this example illustrates, a base in a logarithm can be almost any number, thus presenting infinite options for transformation. Traditionally, authors such as Cleveland (1984) have argued that a range of bases should be examined when attempting log transformations (see Osborne (2002) for a brief overview on how different bases can produce different transformation results). The argument that a variety of transformations should be considered is compatible with the assertion that Box-Cox can constitute a best practice in data transformation.

Mathematically, the logarithm of number less than 0 is undefined, and similar to square root transformations, numbers between 0 and 1 are treated differently than those above 1.0. Thus a distribution to be transformed via this method should be anchored at 1.00 (the recommendation in Osborne, 2002) or higher.

Inverse transformation. To take the inverse of a number (x) is to compute 1/x. What this does is essentially make very small numbers (e.g., 0.00001) very large, and very large numbers very small, thus reversing the order of your scores (this is also technically a class of transformations, as inverse square root and inverse of other powers are all discussed in the literature). Therefore one must be careful to reflect, or reverse the distribution prior to (or after) applying an inverse transformation. To reflect, one multiplies a variable by -1, and then adds a constant to the distribution to bring the minimum value back above 1.00 (again, as numbers between 0.00 and 1.00 have different effects from this transformation than those at 1.00 and above, the recommendation is to anchor at 1.00).

Arcsine transformation. This transformation has traditionally been used for proportions, (which range from 0.00 to 1.00), and involves of taking the arcsine of the square root of a number, with the resulting transformed data reported in radians. Because of the mathematical properties of this transformation, the variable must be transformed to the range −1.00 to 1.00. While a perfectly valid transformation, other modern techniques may limit the need for this transformation. For example, rather than aggregating original binary outcome data to a proportion, analysts can use logistic regression on the original data.

Box- Cox transformation. If you are mathematically inclined, you may notice that many potential transformations, including several discussed above, are all members of a class of transformations called power transformations. Power transformations are merely transformations that raise numbers to an exponent (power). For example, a square root transformation can be characterized as $x^{1/2}$, inverse transformations can be characterized as x^{-1} and so forth. Various authors talk about third and fourth roots being useful in various circumstances (e.g., $x^{1/3}$, $x^{1/4}$). And as mentioned above, log transformations embody a class of power transformations. Thus we are talking about a potential continuum of transformations that provide a range of opportunities for closely calibrating a transformation to the needs of the data. Tukey (1957) is often credited with presenting the initial idea that transformations can be thought of as a class or family of similar mathematical functions. This idea was modified by Box and Cox (1964) to take the form of the Box-Cox transformation:

$$y_i^{\lambda} = (y_i^{\lambda} - 1) / \lambda \quad \text{where } \lambda \neq 0;$$

$$y_i^{\lambda} = \log_e(y_i) \text{ where } \lambda = 0.$$[1]

[1] Since Box and Cox (1964) other authors have introduced modifications of this transformations for special applications and circumstances (e.g., John & Draper, 1980), but for most researchers, the original Box-Cox suffices and is preferable due to computational simplicity.

While not implemented in all statistical packages[2], there are ways to estimate lambda, the Box-Cox transformation coefficient using any statistical package or by hand to estimate the effects of a selected range of λ automatically. This is discussed in detail in the appendix. Given that λ can take on an almost infinite number of values, we can theoretically calibrate a transformation to be maximally effective in moving a variable toward normality, regardless of whether it is negatively or positively skewed.[3] Additionally, as mentioned above, this family of transformations incorporates many traditional transformations:

λ = 1.00: no transformation needed; produces results identical to original data

λ = 0.50: square root transformation

λ = 0.33: cube root transformation

λ = 0.25: fourth root transformation

λ = 0.00: natural log transformation

λ = -0.50: reciprocal square root transformation

λ = -1.00: reciprocal (inverse) transformation

and so forth.

Examples of application and efficacy of the Box-Cox transformation

Bortkiewicz's data on Prussian cavalrymen killed by horse-kicks. This classic data set has long been used as an example of non-normal (poisson, or count) data. In this data set, Bortkiewicz (1898) gathered the number of cavalrymen in each Prussian army unit that had been killed each year from horse-kicks between 1875 and 1894. Each unit had relatively few (ranging from 0-4 per year), resulting in a skewed distribution (presented in Figure 1; skew = 1.24), as is often the case in count data. Using square root, $\log_e$, and $\log_{10}$, will improve normality in this variable (resulting in skew of 0.84, 0.55, and 0.55, respectively). By utilizing Box-Cox with a variety of λ ranging from -2.00 to 1.00, we can determine that the optimal transformation after being anchored at 1.0 would be a Box-Cox transformation with λ = - 2.00 (see Figure 2) yielding a variable that is almost symmetrical (skew = 0.11; note that although transformations between λ = - 2.00 and λ = - 3.00 yield slightly better skew, it is not substantially better).

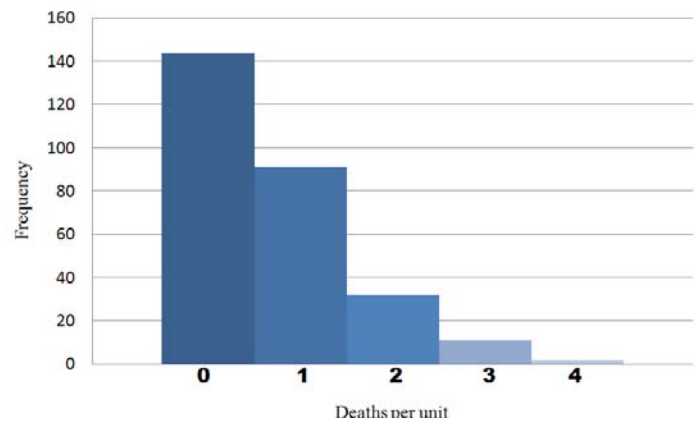

Figure 1. *Deaths from horse kicks, Prussian Army 1875-1894*

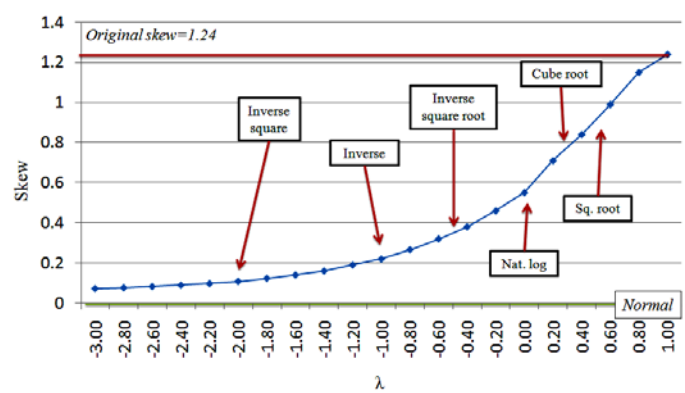

Figure 2.*Box-Cox transforms of horse-kicks with various λ*

University size and faculty salary in the USA. Data from 1161 institutions in the USA were collected on the size of the institution (number of faculty) and average faculty salary by the AAUP (American Association of University Professors) in 2005. As Figure 3 shows, the variable *number of faculty* is highly skewed (skew = 2.58), and Figure 4 shows the results of Box-Cox transformation after being anchored at 1.0 over the range of λ from -2.00 to 1.00. Because of the nature of these data (values ranging from 7 to over 2000 with a strong skew), this transformation attempt produced a wide range of outcomes across the thirty-two examples of Box-Cox transformation, from extremely bad outcomes (skew < -30.0 where λ < -1.20) to very positive

[2] For example, SAS has a convenient and very well done implementation of Box-Cox within *proc transreg* that iteratively tests a variety of λ and identifies the best options for you. Many resources on the web, such as http://support.sas.com/rnd/app/da/new/802ce/stat/chap15/sect8.htm provide guidance on how to use Box-Cox within SAS.

[3] Most common transformations reduce positive skew but may exacerbate negative skew unless the variable is reflected prior to transformation. Box –Cox eliminates the need for this.

outcomes of λ = 0.00 (equivalent to a natural log transformation) achieved the best result. (skew = 0.14 at λ = 0.00) . Figure 5 shows results of the same analysis when the distribution is anchored at the original mean (132.0) rather than 1.0. In this case, there are no extremely poor outcomes for any of the transformations, and one (λ = - 1.20) achieves a skew of 0.00. However, it is not advisable to stray too far from 1.0 as an anchor point. As Osborne (2002) noted, as minimum values of distributions deviate from 1.00, power transformations tend to become less effective. To illustrate this, Figure 5 shows the same data anchored at a minimum of 500. Even this relatively small change from anchoring at 132 to 500 eliminates the possibility of reducing the skew to near zero.

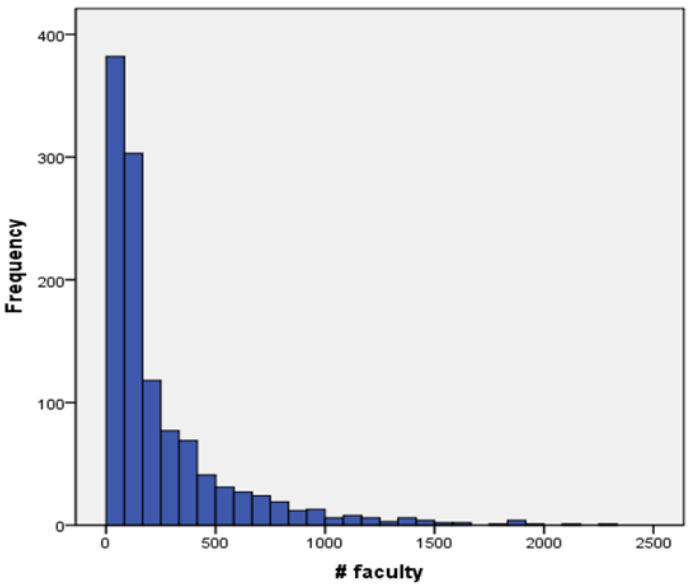

Figure 3. *Number of faculty at institutions in the USA*

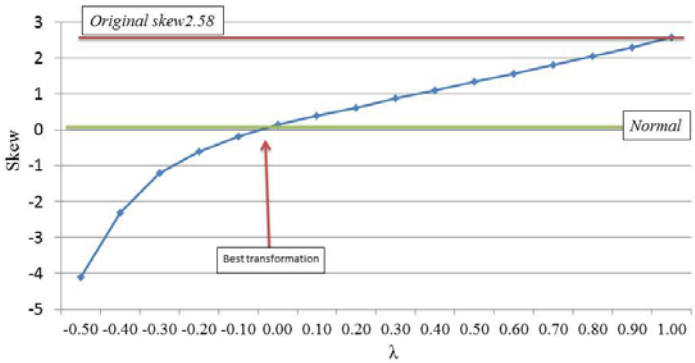

Figure 4. *Box-Cox transform of university size with various λ, anchored at 1.00*

Faculty salary (associate professors) was more normally distributed to begin with, with a skew of 0.36. A Box-Cox transformation with λ = 0.70 produced a skew of -0.03.

To demonstrate the benefits of normalizing data via Box-Cox a simple correlation between number of faculty and associate professor salary (computed prior to any transformation) produced a correlation of $r_{(1161)} = 0.49$, $p < .0001$. This represents a coefficient of determination (% variance accounted for) of 0.24, which is substantial yet probably under-estimates the true population effect due to the substantial non-normality present. Once both variables were optimally transformed, the simple correlation was calculated to be $r_{(1161)} = 0.66$, $p < .0001$. This represents a coefficient of determination (% variance accounted for) of 0.44, or an 81.5% increase in the coefficient of determination over the original.

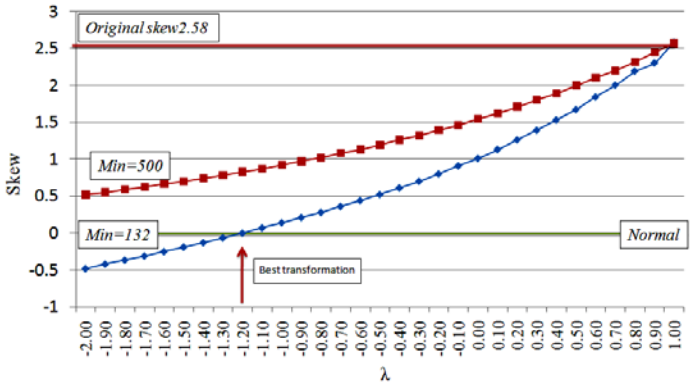

Figure 5. *Box-Cox transform of university size with various λ anchored at 132, 500*

Student test grades. Positively skewed variables are easily dealt with via the above procedures. Traditionally, a negatively skewed variable had to be *reflected* (reversed), anchored at 1.0, transformed via one of the traditional (square root, log, inverse) transformations, and reflected again. While this reflect-and-transform procedure also works fine with Box-Cox, researchers can merely use a different range of λ to create a transformation that deals with negatively skewed data. In this case I use data from a test in an undergraduate Educational Psychology class several years ago. These 174 scores range from 48% to 100%, with a mean of 87.3% and a skew of -1.75. Anchoring the distribution at 1.0 by subtracting 47 from all scores, and applying Box-Cox transformation from 1.0 to 4.0, we get the results presented in Figure 6,

indicating a Box-Cox transformation with a λ = 2.70 produces a skew of 0.02.

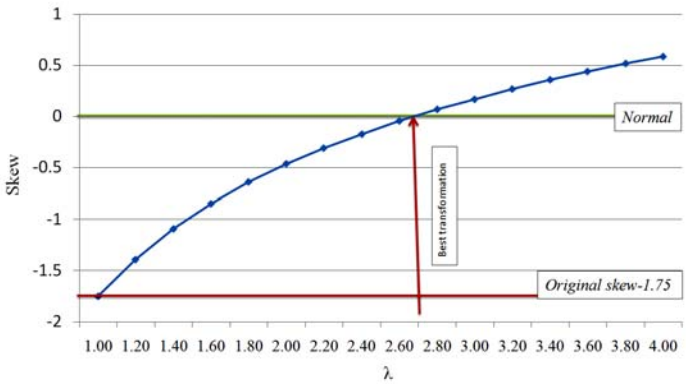

Figure 6. *Box-Cox transform of student grades, negatively skewed*

SUMMARY AND CONCLUSION

The goal of this paper was to introduce Box-Cox transformation procedures to researchers as a potential best practice in data cleaning. Although many of us have been briefly exposed to data transformations, few researchers appear to use them or report data cleaning of any kind (Osborne, 2008b). Box-Cox takes the idea of having a range of power transformations (rather than the classic square root, log, and inverse) available to improve the efficacy of normalizing and variance equalizing for both positively- and negatively-skewed variables.

As the three examples presented above show, not only does Box-Cox easily normalize skewed data, but normalizing the data also can have a dramatic impact on effect sizes in analyses (in this case, improving the effect size of a simple correlation over 80%).

Further, many modern statistical programs (e.g., SAS) incorporate powerful Box-Cox routines, and in others (e.g., SPSS) it is relatively simple to use a script (see appendix) to automatically examine a wide range of λ to quickly determine the optimal transformation.

Data transformations can introduce complexity into substantive interpretation of the results (as they change the nature of the variable, and any λ less than 0.00 has the effect of reversing the order of the data, and thus care should be taken when interpreting results.). Sakia (1992) briefly reviews the arguments revolving around this issue, as well as techniques for utilizing variables that have been power transformed in prediction or converting results back to the original metric of the variable. For example, Taylor (1986) describes a method of approximating the results of an analysis following transformation, and others (see Sakia, 1992) have shown that this seems to be a relatively good solution in most cases. Given the potential benefits of utilizing transformations (e.g., meeting assumptions of analyses, improving generalizability of the results, improving effect sizes) the drawbacks do not seem compelling in the age of modern computing.

REFERENCES

Bortkiewicz, L., von. (1898). *Das Gesetz der kleinen Zahlen.* Leipzig: G. Teubner.

Box, G. E. P., & Cox, D. R. (1964). An analysis of transformations. *Journal of the Royal Statistical Society, B,, 26*(211-234).

Cleveland, W. S. (1984). Graphical methods for data presentation: Full scale breaks, dot charts, and multibased logging. *The American Statistician, 38*(4), 270-280.

Goodman, L. A. (1954). Kolmogorov-Smirnov tests for psychological research. . *Psychological Bulletin, 51*, 160-168.

John, J. A., & Draper, N. R. (1980). An alternative family of transformations. *applied statistics, 29*, 190-197.

Lilliefors, H. W. (1968). On the kolmogorov-smirnov test for normality with mean and variance unknown. *Journal of the American Statistical Association, 62*, 399-402.

Orr, J. M., Sackett, P. R., & DuBois, C. L. Z. (1991). Outlier detection and treatment in I/O Psychology: A survey of researcher beliefs and an empirical illustration. *Personnel Psychology, 44*, 473-486.

Osborne, J. W. (2002). Notes on the use of data transformations. *Practical Assessment, Research, and Evaluation., 8*, Available online at http://pareonline.net/getvn.asp?v=8&n=6 .

Osborne, J. W. (2008a). Best Practices in Data Transformation: The overlooked effect of minimum values. In J. W. Osborne (Ed.), *Best Practices in Quantitative Methods.* Thousand Oaks, CA: Sage Publishing.

Osborne, J. W. (2008b). Sweating the small stuff in educational psychology: how effect size and power reporting failed to change from 1969 to 1999, and what that means for the future of changing practices. *Educational Psychology, 28*(2), 1 - 10.

Rosenthal, R. (1968). An application of the kolmogorov-smirnov test for normality with estimated mean and variance. *Psychological-Reports, 22*(570).

Sakia, R. M. (1992). The Box-Cox transformation technique: A review. *The statistician, 41*, 169-178.

Taylor, M. J. G. (1986). the retransformed mean after a fitted power transformation. *Journal of the American Statistical Association, 81*, 114-118.

Tukey, J. W. (1957). The comparative anatomy of transformations. *Annals of Mathematical Statistics, 28*, 602-632.

Wilcox, R. R. (1997). Some practical reasons for reconsidering the Kolmogorov-Smirnov test. *British Journal of Mathematical and Statistical Psychology, 50*(1), 71-78.

Zimmerman, D. W. (1994). A note on the influence of outliers on parametric and nonparametric tests. *Journal of General Psychology, 121*(4), 391-401.

Zimmerman, D. W. (1995). Increasing the power of nonparametric tests by detecting and downweighting outliers. *Journal of Experimental Education, 64*(1), 71-78.

Zimmerman, D. W. (1998). Invalidation of parametric and nonparamteric statistical tests by concurrent violation of two assumptions. *Journal of Experimental Education, 67*(1), 55-68.

Citation:

Osborne, Jason (2010). Improving your data transformations: Applying the Box-Cox transformation. *Practical Assessment, Research & Evaluation*, 15(12). Available online: http://pareonline.net/getvn.asp?v=15&n=12.

Note:

The author wishes to thank to Raynald Levesque for his web page: http://www.spsstools.net/Syntax/Compute/Box-CoxTransformation.txt, from which the SPSS syntax for estimating lambda was derived.

Corresponding Author:

Jason W. Osborne
Curriculum, Instruction, and Counselor Education
North Carolina State University
Poe 602, Campus Box 7801
Raleigh, NC 27695-7801
919-244-3538
Jason_osborne@ncsu.edu

APPENDIX

Calculating Box-Cox λ by hand

If you desire to estimate λ by hand, the general procedure is to:

- divide the variable into at least 10 regions or parts,
- calculate the mean and s.d. for each region or part,
- Plot log(s.d.) vs. log(mean) for the set of regions,
- Estimate the slope of the plot, and use the slope (1-b) as the initial estimate of λ

As an example of this procedure, we revisit the second example, number of faculty at a university. After determining the ten cut points that divides this variable into even parts, selecting each part and calculating the mean and standard deviation, and then taking the $\log_{10}$ of each mean and standard deviation, Figure 7 shows the plot of these data. I estimated the slope for each segment of the line since there was a slight curve (segment slopes ranged from -1.61 for the first segment to 2.08 for the last) and averaged all, producing an average slope of 1.02. Interestingly, the estimated λ from this exercise would be -0.02, very close to the empirically derived 0.00 used in the example above.

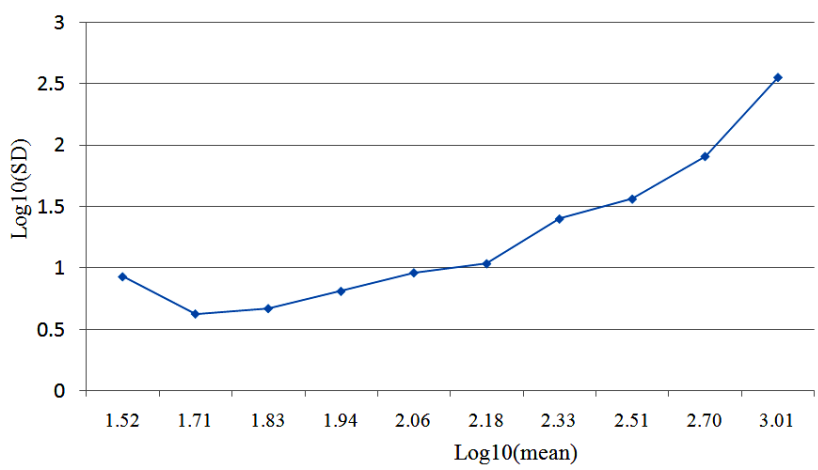

Figure 7. *Figuring λ by hand*

Estimating λ empirically in SPSS

Using the syntax below, you can estimate the effects of Box-Cox using 32 different lambdas simultaneously, choosing the one that seems to work the best. Note that the first COMPUTE anchors the variable (NUM_TOT) at 1.0, as the minimum value in this example was 7. You need to edit this to move your variable to 1.0.

```
****************************.
*** faculty #, anchored 1.0
****************************.

COMPUTE var1=num_tot-6.
execute.

VECTOR lam(31) /xl(31).
LOOP idx=1 TO 31.
- COMPUTE lam(idx)=-2.1 + idx * .1.
- DO IF lam(idx)=0.
-   COMPUTE xl(idx)=LN(var1).
- ELSE.
-   COMPUTE xl(idx)=(var1**lam(idx) - 1)/lam(idx).
- END IF.
END LOOP.
EXECUTE.
```

```
FREQUENCIES VARIABLES=var1 xl1 xl2 xl3 xl4 xl5 xl6 xl7 xl8 xl9 xl10 xl11 xl12 xl13 xl14 xl15
xl16 xl17 xl18 xl19 xl20 xl21 xl22 xl23 xl24 xl25 xl26 xl27 xl28 xl29 xl30 xl31
  /format=notable
  /STATISTICS=MINIMUM MAXIMUM SKEWNESS
  /HISTOGRAM
  /ORDER=ANALYSIS.
```

Note that this syntax tests λ from -2.0 to 1.0, a good initial range for positively skewed variables. There is no reason to limit analyses to this range, however, so that depending on the needs of your analysis, you may need to change the range of lamda tested, or the interval of lambda. To do this, you can either change the starting value on the above line:

```
- COMPUTE lam(idx)=-2.1 + idx * .1.
```

For example, changing -2.1 to 0.9 starts lambda at 1.0 for exploring variables with negative skew. Changing the number at the end (0.1) changes the interval SPSS examines—in this case it examines lambda in 0.1 intervals, but changing to 0.2 or 0.05 can help fine-tune an analysis.