The New School Psychology Bulletin
2007, Vol. 5, No. 2

Print ISSN: 1931-793X; Online ISSN: 1931-7948

Research Report

Application of Data Screening Procedures in Stress Research

Daniel Cruz, M.A.
Seton Hall University

In the analysis of salivary cortisol data, researchers often perform statistical analysis for hypothesis testing in the absence of data mining procedures. In this article, I demonstrate the utility of screening data from a study investigating the effects of acute stress on salivary cortisol reactivity through the application of procedures recommended by Tabachnick and Fidell (2001). Specifically, an examination for the presence of both univariate and multivariate outliers (Study 1) and methods for correcting skewed distributions (Study 2) were used in order to demonstrate the efficacy of screening data prior to hypothesis testing. The results suggest that there were no outliers present in the data set. Application of algorithms from a family of transformations showed that they were effective in reducing skewness, kurtosis and variability.

Keywords: data screening, stress

When conducting statistical analysis for hypothesis testing, appropriate conclusions can only be drawn when the assumptions guiding the specific analysis are sound. Generally, the assumptions include independence of error, homogeneity of variance and normality. Normality, in particular, can often be addressed prior to hypothesis testing through data screening procedures. Assessing for normality and factors affecting the shape of the distribution prior to hypothesis testing can help researchers draw more accurate conclusions and can help diagnose potential problems early on that can affect the results of the statistical analysis and assumptions underlying the hypothesis tests. Furthermore, screening variables for normality is essential in the data analysis process for providing an optimal mathematical solution and improving generalizability of results.

Outliers

A skewed distribution refers to the degree to which the distribution is pulled in a direction away from the center, typically the result of extreme observations (Witte & Witte, 2008). Previous research has suggested that appropriate skewness values are those that range from -1 to +1 (Muthen & Kaplan, 1985). Kurtosis values close to zero suggest appropriate "peakedness," (Tabachnick & Fidell, 2001). One factor that can influence skewness and kurtosis is the presence of outliers. Undetected, outliers can distort phenomena inherent in the underlying population. Outliers are complex in that they can result from a variety of situations. Small samples may be particularly susceptible to extreme observations. With larger sample sizes, the influence of outliers is often reduced in accordance with the central limit theorem.

Transformations

Departures from normality is especially problematic in stress research when sample sizes are unduly small. As a result, the assumptions of hypothesis tests are often violated because of scale effects and extreme scores. Power transformation is useful in that it can reduce skewness and the effect of outliers on the shape of the distribution. Transformation of variables has been used across multiple studies and reliance on it is growing in the field of research and statistics (e.g., Peltier, Wilcox & Sharp, 1998; Guerrero & Johnson, 1982; Becerril, Wilcox, Wiggans & Sigmon, 1994). The goal of transformation is to rescale the original data so that the tails of the distribution are compressed and variance is minimized. Box and Cox (1964), for instance, developed complex computational formulas aimed at modifying skewed distributions so that they better conform to normality. These methods are further described in detail in a number of other studies (see Baker, 1934; Bartlett, 1947).

The present article sought to explore the application of Tabachnick and Fidell's (2001) recommended data screening procedures. The procedures have been developed for screening data for normality prior to hypothesis testing and provide strategies for correcting it. In particular, Tabachnick and Fidell's techniques for identifying outliers and transforming variables for skewed distributions were applied to a positively skewed data set. Specifically, Study 1 sought to explore the presence of univariate and multivariate outliers in a small positively skewed sample. The goal

Correspondence concerning this article should be addressed to Daniel Cruz, M.A., Professional Psychology and Family Therapy (PPFT), Department of Counseling Psychology, Seton Hall University, 400 South Orange Ave, South Orange, NJ 07079. Email: cruzdani@shu.edu

of Study 2 was to demonstrate the application of a family of transformation procedures to correct for the skewness present in the sample.

Study 1

Method

Data were taken from a pilot study investigating the effects of stress on salivary cortisol reactivity. Ten healthy young volunteers (8 women and 2 men; mean age = 19.1) were exposed to the Cold Pressor Task (CPT), a physiological stressor that has been shown to induce stress (see Patil, Apelbaum & Zacny, 1995; Alleva & Santucci, 2001).

Materials

A container of ice water chilled to a temperature of 4 degrees Celsius was used. Temperature was assessed intermittently by a laboratory-grade thermometer. Subjects were instructed to immerse their left hand, with the assistance of an experimenter, up to the wrist, for a period of 2 minutes. Salivary cortisol samples were collected at baseline and 20 minutes after exposure to the CPT procedure and provided a measure of physiological stress. Samples were sent to Salimetrics Laboratory (State College, PA) for duplicate biochemical analysis. Data were then examined for shape and for the presence of extreme observations using SPSS (Version 15 for Windows).

Preliminary analysis of the data suggested that the distributions of both the baseline and stress condition salivary cortisol were positively skewed. The baseline condition resulted in a skewness value of 2.366 and kurtosis at 6.322. The stress condition resulted in a skewness value of 2.258 and kurtosis value of 5.450.

Analysis of Univariate Outliers. Univariate outliers are cases with extreme scores on a single variable (Tabachnick & Fidell, 2001). In order to determine whether univariate outliers were present, a standardized residual was calculated for each case where the difference between each individual's raw score and sample mean is divided by the sample standard deviation. This resulted in a standardized residual value for each case (see Table 1). Values that exceed 3 standard deviations from the mean were considered univariate outliers.

Multivariate Outliers. Multivariate outliers suggest that the individual is responding differently compared to other participants across multiple dimensions. Specifically, they are cases with extreme scores on two or more variables (Tabachnick & Fidell, 2001). Multivariate outliers were evaluated based on Mahalanobis distance values. According to Tabachnick and Fidell, Mahalanobis distance refers to the degree to which a case differs from the centroid created as a function of means for the combination of variables across multidimensional space. Mahalanobis distance values were calculated for each combination of scores

Table 1
Summary of Outlier Analysis for Salivary Cortisol Data by Condition

	Salivary cortisol (Ug/Dl)[a]		**Standardized residual[b]**		**Mahalanobis distance[c]**
Participant	**Baseline**	**Stress**	**Baseline**	**Stress**	
1	0.324	0.249	0.10	-0.09	0.73081
2	0.341	0.435	0.17	0.71	6.17814
3	0.380	0.274	0.32	0.02	1.89018
4	0.240	0.198	-0.23	-0.31	0.19727
5	0.966	0.867	2.63	2.57	6.92853
6	0.157	0.131	-0.56	-0.60	0.37233
7	0.178	0.157	-0.47	-0.48	0.23344
8	0.088	0.088	-0.83	-0.78	0.69793
9	0.165	0.168	-0.52	-0.44	0.38315
10	0.140	0.127	-0.62	-0.61	0.38821

[a] Ug/dL = microgram per deciliter. [b]Standard residual scores above or below 3 are defined as univariate outliers. [c]Multivariate outliers are defined as Mahalanobis distance > 13.816 based on $\chi^2(2, N = 10)$ at $p < .001$.

by evoking the SPSS regression function. The Mahalanobis distance is defined by: Mahalanobis distance = $(x_i - y_i)' \Sigma^{-1} (x_i - y_i)$ where the sample mean vector y_i and covariance matrix Σ assigns a weight to x_i, and provides a measure of how far x_i is from the mean vector (Yuan, Fung & Reise, 2004)

Results

Analysis of the standardized residuals to assess univariate outliers suggested that there were no outliers present for either the baseline or stress condition (see Table 1). None of the standardized residual values exceeded 3 standard deviations from the mean. In order to determine whether any multivariate outliers were present, Mahalanobis distance values were assessed using x^2 (2, N = 10) = 13.816, $p < .001$. The results show that the Mahalanobis distance values were all below the obtained x^2 value of 13.816 respectively (see Table 1).

Study 2

Methods

Skewness. Skewness and kurtosis were assessed through SPSS (Version 15 for Windows) descriptive analysis for both the baseline and CPT salivary cortisol conditions. The results of the output suggest that both the baseline and stress condition salivary cortisol values were positively skewed. As mentioned above, the baseline condition resulted in a skewness value of 2.366 and a kurtosis value of 6.322. The stress condition resulted in a skewness value of 2.258 and a kurtosis value of 5.450. To address the positively skewed distribution, three mathematical transformations were performed.

Data Transformations. The transformations explored to modify the skewness of the distribution were as follows: A square root function transformation, a logarithmic transformation and an inverse transformation (Tabachnick & Fidell, 2001). Square root function transformations convert each individual raw score into its square root equivalent. In doing so, skewness is reduced by compressing the negative and positive tails of the distribution (Nolan & Heinzin, 2008). The logarithmic (base 10) transformation computes a log for each score (see Hamilton, 1992). Skewness is reduced by compressing the positive side of the distribution and extending the smaller values on the negative side (Nolan & Heinzen, 2008). Inverse transformations are calculated by dividing each individual raw score into a value of 1 and calculating the inverse of the original power (Hamilton, 1992).

Results

The results of the data transformations are summarized in Table 2. In terms of the baseline salivary cortisol

Table 2
Results of a Family of Transformations on Distribution Skewness and Kurtosis

Descriptives	Salivary cortisol (Ug/dL)	SQRT[a]	LOG[b]	INV[c]
		Baseline Condition		
Skewness	2.366	2.203	2.025	-1.637
Kurtisis	6.322	5.613	4.862	3.329
Mean	0.298	1.135	0.107	0.791
SD	0.25	0.1	0.07	0.12
		Stress Condition		
Skewness	2.258	2.129	1.993	-1.707
Kurtisis	5.45	4.845	4.231	3.025
Mean	0.269	1.123	0.098	0.807
SD	0.23	0.1	0.07	0.11

Note. A constant was added to transformations to ensure that all arguments are greater than zero.
[a]Square root transformation defined as $f(x) = \sqrt{(x+c)}$. [b]Logarithmic transformation defined as $f(x) = \log_{10}(x+c)$.
[c]Inverse transformation defined as $f(x) = 1/(x+c)$.

condition, skewness was originally calculated at 2.366 and kurtosis was calculated at 6.322 (*M* = .298, *SD* = .25). The square root transformation reduced the skewness of the distribution to 2.203 and the kurtosis to 5.613 (*M* = 1.135, *SD* = .10). The logarithmic transformation (Log10) reduced the skewness to 2.025 and kurtosis to 4.862 (*M* = .107, *SD* = .07) and the inverse transformation reduced the skewness to -1.637 and kurtosis to 3.329 (*M* = .791, *SD* = .12).

In regards to the cold stress condition, skewness was initially calculated at 2.258 and kurtosis at 5.450 (*M* = .269, *SD* = .23). Application of the square root transformation reduced the skewness to 2.129 and kurtosis to 4.845 (*M* = 1.123, *SD* = .10). The logarithmic transformation (log_{10}) reduced the skewness to 1.993 and kurtosis to 4.231 (*M* = .098, *SD* = .07) and the inverse transformation reduced the skewness to -1.707 and kurtosis to 3.025 (*M* = .807, *SD* = .11), respectively. In both conditions, square root transformation minimized the variability within groups as did the logarithmic transformation.

Discussion

Study 1 was conducted to determine whether outliers were present in the data. The results of the standardized residuals indicated the absence of univariate outliers suggesting that this was not a factor affecting the shape of the distribution. Further analysis of outliers by way of Mahalanobis distance values provided additional evidence of this finding. The methods for detecting univariate and multivariate outliers used in this article are further described by Tabachnick and Fidell (2001).

In Study 2, algorithms from a family of transformations were applied in order to examine the effects of a square root transformation, logarithmic transformation and an inverse transformation on the shape of the skewed distribution. There were differential effects of transformation functions on the shape of the distributions. The square root function transformation successfully reduced the skewness and kurtosis of the distribution and minimized the variability within the sample in both the baseline and stress salivary cortisol conditions. The logarithmic transformation also minimized the skewness and kurtosis as well as compressed the variability within groups. The inverse transformation moderately improved the shape of the distribution and reversed the direction of the skew.

The purpose of this article was to describe the application of Tabachnick and Fidell's (2001) data screening procedures with data from a pilot study investigating the effects of physiological stress on salivary cortisol reactivity. The use of this data set was especially useful because of the small sample size, a circumstance that is often the case in biological research dissociating the complex nature of stress. The procedures described in this article have received support from other studies were violations of normality are present (Bocchino, Hartman & Foley, 2003; Peltier et al., 1998) and provide a plausible mechanism by which researchers can address these complex issues.

While the techniques described in this article have received empirical support, it only represents an introduction to what is considered by many statisticians as a complex topic. Data mining procedures take considerable time and effort and statisticians differ in the methods they employ. In this article, for instance, standardized residuals were used to screen for univariate outliers. However, there are other methods for screening univariate outliers which researchers in the area of stress may consider including graphical methods such as stem-and-leaf plots. Multivariate outliers can also be detected by application of the jackknife procedure (Ott & Longnecker, 2001) which involves calculating successive regression coefficients, removing one data point at a time and analyzing the model change that results from the exclusion of the single case. Additionally, leverage, Cook's distance and influence are other statistics used to detect multivariate outliers in addition to Mahalanobis distance values (Tabachnick & Fidell, 2001).

The observations support the utility of data screening procedures in detecting univariate and multivariate outliers. Linear transformation procedures have been found to be successful in correcting for departure from normality. Transformation procedures are versatile, can be applied to grouped and ungrouped data and can greatly improve assumptions of hypothesis tests including heterogeneity of error variance (Peltier et al., 1998). Nonnormal distributions can be problematic and if not detected prior to hypothesis testing, researchers risk drawing erroneous and inaccurate conclusions. The procedures demonstrated in this article are described in more detail elsewhere (Tabachnick & Fidell, 2001) and provide promising techniques for researchers and statisticians for screening data prior to hypothesis testing.

References

Alleva, E., & Santucci, D. (2001). Psychosocial vs. "physical" stress situations in rodents and humans: Role of neurotrophins. *Physiology & Behavior, 73*, 313-320.

Baker, G. A. (1934). Transformations of non-normal frequency distribution into normal distributions. *The Annals of Mathematical Statistics, 5*, 113-123.

Bartlett, M. S. (1947). The use of transformation. *Biometrics, 3*, 39-52.

Becerril, C. M., Wilcox, C. J., Wiggans, G. R., & Sigmon, K. N. (1994). Transformation of measurements of white coat color for Holsteins and estimation of heritability.

Journal of Dairy Science, 77, 2651-2657.

Bocchino, C. C., Hartman, B. W., & Foley, P. F. (2003). The relationship between person-organizational congruence, perceived violations of the psychological contract, and occupational stress symptoms. *Consulting Psychology Journal: Practice and Research, 55*, 203-214.

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

Guerrero, V. M., & Johnson, R. A. (1982). Use of Box-Cox transformations with binary response models. *Biometrika, 69*, 309-314.

Hamilton, L. C. (1992). *Regression with graphics: A second course in applied statistics.* Belmont, CA: Duxbury

Muthen, B., & Kaplan, D. (1985). A comparison of some methodologies for the factor analysis of non-normal Likert variables. *British Journal of Mathematical and Statistical Psychology, 38*, 171-189.

Nolan, S. A., & Heinzen, T. E. (2008). Statistics for the behavioral sciences. New York, NY: Worth Publishers Ott, R. L., & Longnecker, M. (2001). *An introduction to statistical methods and data analysis* (5th ed.). Pacific Grove, CA: Duxbury.

Patil, P. G., Apelbaum, J. L., Zacny, J. P. (1995). Effects of a cold-water stressor on psychomotor and cognitive functioning in humans. *Physiology &. Behavior, 58*, 1281-1286.

Peltier, M. R., Wilcox, C. J., & Sharp, D. C. (1998). Technical note: Application of the Box-Cox data transformation to animal science experiments. *Journal of Animal Science, 76*, 847-849.

Tabachnick, B. G., & Fidell, L. S. (2001). *Using multivariate statistics* (4th ed.). Needham Heights, MA: Allyn & Bacon.

Witte, R. S., & Witte, J. S. (2008). *Statistics* (8th ed.). New York: Harcourt Brace College Publishers.

Yuan, K., Fung, W. K., & Reise, S. P. (2004). Three Ma halanobis distances and their role in assessing unidemensionality. *British Journal of Mathematical and Statistical Psychology, 57*, 151-165.