Mook, D.G. (1983). In defense of external invalidity. *American Psychologist, 38*, 379-387.

Oja, H. (1987). On permutation tests in multiple regression and analysis of covariance analysis problems. *Australian Journal of Statistics, 29*, 91-100.

Ostrom, T.M. (1984). The role of external invalidity in editorial decisions. *American Psychologist, 39*, 324.

Pitman, E.J.G. (1937a). Significance tests which may be applied to samples from any populations. *Journal of the Royal Statistical Society (Series B), 4*, 119-130.

Pitman, E.J.G. (1937b). Significance tests which may be applied to samples from any populations II. *Journal of the Royal Statistical Society (Series B), 4*, 225-232.

Pitman, E.J.G. (1938). Significance tests which may be applied to samples from any populations: III. The analysis of variance test. *Biometrika, 29*, 322-335.

Sawilowsky, S.S. (1990). Nonparametric tests of interaction in experimental design. *Review of Educational Research, 60*, 91-126.

ter Braak, C.J.F. (1992). Permutation versus bootstrap significance tests in multiple regression and ANOVA. In K.-H. Jöckel, G. Rothe, & W. Sendler (Eds.), *Bootstapping and related resampling techniques* (pp. 79-86). Berlin: Springer-Verlag.

van den Brink, W.P. & van den Brink, S.G.L. (1989). A comparison of the power of the *t* test, Wilcoxon's test, and the approximate permutation test for the two-sample location problem. *British Journal of Mathematical and Statistical Psychology, 42*, 183-189.

The Continuity Principle in Psychological Research: An Introduction to Robust Statistics

JOHN C. LIND
Alberta Hospital Edmonton

BRUNO D. ZUMBO
University of Ottawa

Abstract
Research in statistics has demonstrated that the classical estimates of means, variances and correlations are sensitive to small departures from the normal curve. Statisticians have urged caution in the use of classical statistics and have proposed a variety of alternatives which are robust with respect to departures from normality. Robust statistics continue, however, to be little used in psychological research. In this paper we describe common sources of nonnormality in psychological data and examine the distinction between data cleaning and robust estimation. Robust estimation using M-estimators is discussed and recommendations for using these techniques in practice are presented.

It is common practice among social and life scientists to adopt an implied continuity principle when interpreting the results of a statistical analysis. It is often assumed, for example, that data which are observed to deviate only slightly in form from that of the familiar normal curve, will only slightly distort the usual estimates of means, standard deviations, correlations and associated hypothesis tests. With increasing departure from an underlying normal model, the greater it is assumed, will be the inaccuracy of the computed statistics.

Over the past several decades, research in statistics has demonstrated that a continuity principle of the form described above for normal theory based statistics is invalid. The classical estimates of means, variances and correlations have been shown to be highly sensitive to even small departures from an underlying normal model. A single outlying observation, for example, can strongly bias these statistics and thereby provide misleading or invalid results (see for example, Huber, 1981; Hampel, Ronchetti, Rousseeuw, & Stahel, 1986; Zimmerman, & Zumbo, 1993). For an example where the presence of a single outlier in a sample of 29 observations results in a change of the correlation coefficient from .99 to 0 see Devlin, Gnanadesikan, and Kettenring (1981).

The sensitivity of classical statistics to small deviations from normality has important implications for the analysis of research data in psychology. The sensitivity of standard estimates of means and variances to nonnormality can adversely affect analysis of variance (ANOVA) results, while in the case of product moment correlations the lack of robustness will often bias results obtained from principal component analysis, common factor analysis and the analysis of covariance structures (i.e., structural equation modelling). Factor analysis results, for example, which initially appear to provide meaningful factors are often, on a closer examination of the data, simply the result of one or two outliers (Huber, 1981, p. 199).

The poor performance of classical statistics in the presence of small departures from normality has led some statisticians (Tukey, 1977, pp. 103-106; Hogg, 1977, pp. 1-17) to warn that the routine use of classical statistics is unsafe. They recommend that classical estimates of means, variances and correlations only be used in conjunction with alternative methods that are robust with respect to departures from normality. Although there is an increasing amount of statistical software which incorporates robust methods, the use of these methods continues to be, despite some urging by statisticians (Stahel, 1989), little used in applied research. In the behavioural sciences, this is in part likely a result of undergraduate methodology courses that often describe the ANOVA as being robust with respect to type I error and nonnormality (see, for example, Glass & Stanley, 1970, p. 372; Glass, Peckham, & Sanders, 1972). Although ANOVA has some moderate robustness properties with respect to type I error and nonnormality, it is, in relation to type II error, very nonrobust (Hampel, et al., 1986, p. 344; Zimmerman & Zumbo, 1993). This places a researcher in an unusual situation when interpreting ANOVA results. In the presence of nonnormality, rejection of the null hypothesis may tend to be believed while less confidence should be placed in believing the failure of the rejection of the null hypothesis.

In this paper we examine the importance of robust statistics in psychological research (and also all of social science research) and provide an overview of some of these methods. We also provide some recommendations regarding the use of robust alternatives in conjunction with the usual methods of analysis.

Departures from Normality

Perhaps the most common form of nonnormality encountered in practice is that where the majority of the data follows the normal curve, however, the presence of errors in the data produces slight deviations from a truly normal model. This may result in sampling distributions which have heavy tails relative to the normal, or in some cases where most of the errors are situated on one side of the center, the tails may appear asymmetric.

In the following we consider the typical situation where a researcher is interested in estimating parameters such as the mean and variance of a normally distributed variable, however, the presence of errors in the data is suspected or known. The problem now becomes one of how to accurately estimate the mean and variance of the underlying normal distribution in the presence of these errors. Note that this situation differs – although it is often confused with – that where the underlying distribution is truly nonnormal. In the latter case, data analysis using nonparametric statistics is often the method of choice. However, using nonparametric statistics in the former means abandoning a normal model for the data and the parsimony such a model provides.

In practice, researchers usually assume that most data sets contain errors. In scientific data generally, the amount of errors commonly observed tends to be in the range of 4% to 7% (Hampel, et al., 1986, p. 27). In behavioural science data, Rosenthal (1978) has surveyed 15 data sets with numbers of observations that range from 96 to 89,980. In these studies observers recorded the data and provided whatever checking of their records they considered necessary prior to analysis. Observers were unaware the data would be subsequently examined for errors. Since observers were able to check their data prior to analysis, these data sets could be considered to be of average or better than average quality. Of the 15 data sets examined in the study, only one was found on subsequent inspection to be free from errors. The errors found in the remaining data sets ranged from 1% to 4%.

For the majority of data sets encountered in practice, the most common source of departures from underlying normality is the result of recording or data entry errors. Other forms of data contamination may arise from misclassified individuals or atypical subject responses. In psychological testing, sources of error are well known, among these being fatigue, motivation and failure to understand test items. In some cases, errors may be systematic. A proportion of individuals may, for example, try to perform poorly on a test if the test is being used to select individuals for an unpopular task (Cronbach, 1960, pp. 37-64).

Small departures from normality are often difficult to detect by visual inspection of the data. Heavy-tailed distributions often appear normal on examination of sample histograms or stem-leaf diagrams. Errors which result in distributions with heavy tails could be described as 'inliers' since they occur within what is judged to be the normal range of the data.

Data Cleaning Methods

Perhaps the most common approach to the problem of data contamination is, prior to analysis, to inspect the data for errors and when these errors are identified, remove them from the sample. Data analysis then proceeds by applying standard statistical procedures to the cleaned data. This strategy is in the class of what is generally referred to as outlier identification methods (Barnett, & Lewis, 1978; Hawkins, 1980; Tabachnick, & Fidell, 1983).

Data cleaning methods have several disadvantages which make them difficult to use in practice. One is the difficulty of specifying criteria for the removal of invalid data points. The familiar 'hard rejection rule', for example, is often invoked to reject data points which fall beyond three standard deviations from the mean. The mean and standard deviation in this case, however, are subject to the bias introduced by the same observations which they are intended to remove, making the accuracy of the procedure questionable. Further, setting the cutoff too high may result in the retention of invalid data points while setting the cutoff too low will result in the removal of data points which are valid.

Data cleaning methods also have consequences for the randomness of the obtained data. The selective removal of observations from the sample compromises the randomness of the data and may compromise subsequent hypothesis tests (for possible consequences, see Zumbo & Zimmerman, 1991; Thomas, 1993). As well, any causal inferences from your data may be compromised due to the restricted sample (see, Wainer, 1989, and the accompanying responses in that issue). In any event, data sets are often large and frequently have a multivariate structure which makes visual inspection or the manual application of data screening methods impractical.

Robust Statistics

There now exists a variety of robust methods that avoid the disadvantages associated with data cleaning methods. These fall into three general classes; those based on rank statistics (R-estimates), those obtained from linear combinations of order statistics (L-estimates) and those based on maximum likelihood (M-estimates). A discussion of the theoretical properties of these statistics can be found in Huber (1964, 1977, 1981) or a less mathematically ladened discussion in Huynh (1982) or Wainer (1982). Due to space limitations we have chosen not to cover adaptive estimators and Bayesian estimators (see Ramsay & Novick, 1980).

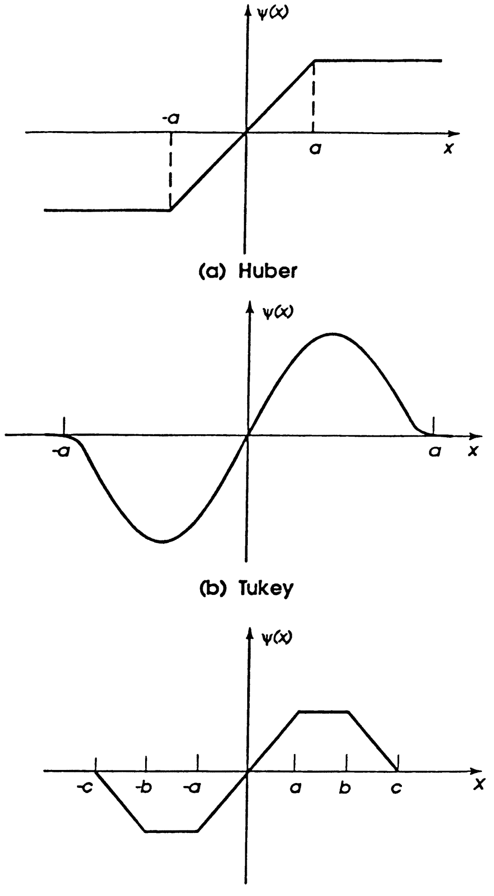
M-ESTIMATORS

In practice, M-estimates are computationally most tractable, particularly for multivariate procedures. They also often have a high breakdown point and in practice are often able to withstand contamination levels as high as 20 to 30 percent. Using M-estimation to estimate a location parameter, θ, i.e., the mean of an underlying normal distribution, involves solving an equation of the form

$$\sum_{i=1}^{n} \Psi\left(\frac{x_i - \Theta}{s}\right) = 0 \qquad (1)$$

where s is a robust estimate of the standard deviation (referred to as the scale parameter). The most commonly used scale estimate in (1) is the median absolute deviation (MAD), (Huber, 1981, p. 107) and is calculated as $s = \text{med}_i \mid x_i - \text{med}_i\, x_i \mid /.6745$.

Perhaps the best known M-estimator, discovered by Huber (1964), is of the form $\Psi(x) = \min(a, \max(x, -a))$. The form of Huber's Ψ-function is shown in Figure 1 (a). Note that observations within the interval [-a, a] remain unchanged while observations with a value greater than a are assigned the value a while all observations which are less than $-a$ are assigned a value of $-a$. By reducing the magnitude of extreme values to either $-a$ or a, the resulting estimate of θ is less influenced than the usual estimate of the mean in the presence of gross errors. A typical value for a in applied settings is 1.5. Figure 1 (b) displays the Ψ-function corresponding to the well known bi-weight (Beaton & Tukey, 1974) and Figure 1(c), a 3-part redescending Ψ-function. Note that redescending Ψ-functions remove the influence of extreme data points entirely.



(a) Huber

(b) Tukey

(c) Hampel

Fig. 1 Ψ-functions for some commonly used M-estimators.

OPTIMALITY OF M-ESTIMATES

An important aspect of M-estimation is a result of certain optimality properties which make them a better choice in applied research than other methods. Since these methods are an extension of classical maximum likelihood methods they incorporate much of the rigor that maximum likelihood theory provides.

In situations where outliers are distributed symmetrically about the mean of a normally distributed variable, for example, Huber's Ψ-function provides an estimate of the mean which is minimum variance, i.e., this estimate will show less variability between samples drawn for the same population than any other robust estimate defined by equation (1). In the case of asymmetrically distributed errors and an underlying normal distribution, the optimal Ψ-function is described in Collins (1976).

The optimality properties of M-estimators has important implications for the practitioner. It means in the case of symmetrically distributed errors for example, that other commonly used estimates of the mean such as the median or various estimates based on winsorizing or trimming procedures will fail to perform as well as the M-estimator.

MULTIVARIATE M-estimators

Regression. The sensitivity of least-squares estimation of regression parameters is well known. Further, invalid data points are often difficult to detect using residual plots (see for example, Agee & Turner, 1977, p. 113). Residual plots from least-squares estimation mask the effect of outliers over all of the data points, again making there detection difficult. That is, the residuals, themselves based on the least-squares estimators, are biased.

M-estimates of regression parameters are obtained by the straight-forward extension of equation (1). For a p-variate regression model of the form

$$y_i = b_1x_{i1} + b_2x_{i2} + \ldots + b_px_{ip} + e_i,\ i = 1, \ldots n, \quad (2)$$

M-estimates of the regression parameters b_j, j = 1,   p, are obtained by solving an equation of the form

$$\sum_{i=1}^{n} \psi\left[\frac{y_i - \beta_1x_{i1} - \beta_2x_{i2} - \ldots - \beta_px_{ip}}{s}\right] x_{ij} = 0 \quad (3)$$

where s is an initial robust estimate of the scale of the residuals such as the MAD. In this case note that robust estimates of the regression parameters are obtained by applying a Ψ-function to the residuals, thereby reducing the influence of a large residual or in the case of a redescending Ψ-function, possibly removing it entirely.

Correlation matrices. The simultaneous estimation of correlation coefficients presents several problems for robust estimation. If the elements of a correlation matrix are estimated individually, the resulting matrix may fail to be positive definite. In other cases, the resulting estimates may lack the invariance properties associated with the familiar product-moment correlations. For a comparison of robust estimates of correlations including element-wise procedures, ellipsoidal trimming methods and M-estimates see Devlin, Ganadesikan and Kettenring (1981).

Devlin et al. (1981) observe that M-estimators of correlation matrices have a low breakdown point, i.e., they are only able to withstand small amounts of contamination in high dimensional cases. More recent developments (Li & Chen, 1985; Huber, 1985), have overcome this problem using M-estimators with projection pursuit methods. Project pursuit methods for correlation matrices involve robustly estimating the eigenvalues and corresponding eigenvectors of the correlation matrix. This is accomplished by projecting the multidimensional data onto a single dimension and then applying M-estimators to the form in (1) to the unidimensional data. The unidimensional estimates are then combined using straightforward matrix operations to obtain a final estimate of the correlation matrix. This procedure has the advantage of reducing an estimation problem involving a large number of parameters to a sequence of steps each involving a univariate estimation problem.

ROBUST INFERENCE

By surveying the variety of robust statistics that are now available to researchers, it is clear that most parameter estimation and statistical modelling problems encountered in the behavioural and social sciences can be treated with robust methods.

The importance of robust parameter estimates for summarizing data, describing the performance of experimental groups and developing predictive models is self-evident. Because of theoretical complexity, however, the development of robust hypothesis testing procedures has lagged behind the development of methods for robust parameter estimation. Clearly, robust estimates of means and variances cannot be used in the standard formulas for testing hypotheses in ANOVA or regression since this will alter the distribution of the familiar *F* and Student's *t* statistics. Some approaches to robust hypothesis testing may be found in Schrader & Hettmansperger (1980), Schrader & McKean (1977), Tiku, Tan & Balakrishnan (1986) and Hampel et al. (1986).

USING ROBUST METHODS IN PRACTICE

A general procedure for using robust statistics in practice has been suggested by Hogg (1977, pp. 15-16). This involves the following steps.

(a) Perform the usual analysis of the data.
(b) Perform an analysis of the data using robust statistics such as M-estimators.
(c) If the results obtained from (a) and (b) agree, then report the results associated with (a) in the usual manner.
(d) If the results obtained from (a) and (b) fail to agree then the data should be reexamined for the presence of errors. If obvious errors are found, they can be removed and the data reanalyzed. Again observe the warning mentioned above for this strategy.

In part (d) it should be noted, however, that observations which are invalid are often difficult to identify and in these situations results of the robust analysis should be reported rather than those obtained using classical methods.

Discussion

In applied research, the increasing availability of computers and software has resulted in an acceleration of activities associated with data gathering and analysis. The use of microcomputers for data gathering and analysis is now common practice for the majority of researchers in the behavioural sciences and the application of standard computing packages to acquired data is routine. An important concern for researchers, however, is that standard methods of statistical analysis are unsafe, and in situations where data contains even small proportion of errors, the usual methods of analysis will often produce biased and misleading results.

Robust alternatives to classical statistics has research implications for several areas of psychology. In the area of psychological measurement, little research has been directed toward investigating the impact of outlying observations on item response theory or reliability measures such as alpha coefficients. Generalizability theory has much in common with mixed-model ANOVA, and therefore share the lack of robustness associated with classical variance estimates and ANOVA models (see, Crocker & Algina, 1986, for an introduction to these methods). Finally, since causal modeling, structural equation modeling and factor analysis methods are based on estimates of covariance or correlation matrices the reliability of these methods would undoubtedly be improved using robust covariance or correlation matrix estimates. In fact, anyone using structural equation modeling (path analysis or causal modeling) would benefit greatly by a series of papers in Volume 12, Number 2 of the *Journal of Educational Statistics* wherein the noted statistician D.A. Freedman and others make some poignant comments on the *casual* use of these *causal* modeling techniques.

In order to facilitate the use of robust methods in applied research, there is a current need for more training and software. A wider variety of robust methods, particularly

M-estimators, should be included in statistical software packages to make them more accessible to researchers. In regard to research training, a substantial portion or undergraduate and graduate methodology courses should be devoted to robust estimation methods. This would provide researchers with the background necessary to effectively apply robust methods in research problems. Training in the use of robust statistics in combination with improving the availability of the necessary computer software will contribute to the increased use of these methods and thereby aid in improving the overall reliability and quality of research in psychology and the life sciences.

We would like to thank Dr. Jerome N. Sheahan of the Department of Statistics, University of Alberta for comments on an earlier draft of this paper. This research was made possible by a Social Sciences and Humanities Research Council of Canada Fellowship and a Carleton University Research Grant to BDZ.

Send Correspondence to: Bruno D. Zumbo, Measurement and Evaluation, Faculty of Education, University of Ottawa, 145 Jean-Jacques Lussier, Ottawa, Ontario, Canada, K1N 6N5 or John C. Lind, Alberta Hospital Edmonton, Box 307, Edmonton, Alberta, Canada, T5J 2J7.

References

Agee, W.S. & Turner, R.S. (1977). Applications of robust regression to trajectory data reduction. In R. L. Launer & G. N. Wilkinson (Eds.), *Robustness in statistics.* New York: Academic Press.

Barnett, V., & Lewis, T. (1978). *Outliers in statistical data.* New York: Wiley.

Beaton, A.E., & Tukey, J.W. (1974). The fitting of power series, meaning polynomials, illustrated on band-spectroscopic data. *Technometrics, 16*, 147-185.

Collins, J.R. (1976). Robust estimation of a location parameter in the presence of asymmetry. *Annals of Statistics, 4*, 66-85.

Crocker, L., & Algina, J. (1986). *Introduction to classical & modern test theory.* New York: Holt, Rinehart, & Winston.

Cronbach, L.J. (1960). *Essentials of psychological testing* (2nd ed.). New York: Harper & Row.

Devlin, S.J., Gnanadesikan, R., & Kettenring, J.R. (1981). Robust estimation of dispersion matrices and principal components. *Journal of the American Statistical Association, 76*, 354-362.

Freedman, D.A. (1987). As others see us: A case study in path analysis. *Journal of Educational Statistics, 12*, 101-128.

Glass, G., Peckham, P., & Sanders, J. (1972). Consequences of failure to meet assumptions underlying the fixed effects analysis of variance and covariance. *Review of Educational Research, 42*, 237-288.

Glass, G.V., & Stanley, J.C. (1970). *Statistical methods in education and psychology.* Englewood Cliffs: Prentice-Hall.

Hampel, F.R., Ronchetti, E.M., Rousseeuw, P.J., & Stahel, W.A. (1986). *Robust Statistics: The approach based on influence functions.* New York: Wiley.

Hawkins, D.M. (1980). *Identification of outliers.* London: Chapman & Hall.

Hogg, R.V. (1977). An introduction to robust estimation. In R.L. Launer, & G.N. Wilkinson (Eds.), *Robustness in statistics.* New York: Academic Press.

Huber, P.J. (1985). Projection pursuit. *Annals of Statistics, 13*, 435-525.

Huber, P.J. (1981). *Robust Statistics.* New York: Wiley.

Huber, P.J. (1977). *Robust Statistical Procedures.* Philadelphia: Society for Industrial and Applied Mathematics.

Huber, P.J. (1964). Robust estimation of a location parameter. *Annals of Mathematical Statistics, 35*, 73-101.

Huynh, H. (1982). A comparison of four approaches to robust regression. *Psychological Bulletin, 92*, 505-512.

Li, G., & Chen, Z. (1985). Projection-pursuit approach to robust dispersion matrices and principal components: Primary theory and monte carlo. *Journal of the American Statistical Association, 80*, 759-766.

Ramsay, J.O, & Novick, M.R. (1980). PLU robust Bayesian decision theory: Point estimation. *Journal of the American Statistical Association, 75*, 901-907.

Rosenthal, R. (1978). How often are our numbers wrong? *American Psychologist, 33*, 1005-1008.

Schrader, R.M., & Hettmansperger, T.P. (1980). Robust analysis of variance based upon a likelihood ratio criterion. *Biometrika, 67*, 93-101.

Schrader, R.M., & McKean, J.W. (1977). Robust analysis of variance. *Communications in Statistics, (A)6*, 879-894.

Stahel, W.A. (1989). *Robust statistics: From an intellectual game to a consumer product.* Minneapolis: Institute of Mathematics and its Applications.

Tabachnick, B.G., & Fidell, L.S. (1983). *Using multivariate statistics*. New York: Harper & Row.

Tiku, M.L., Tan, W.Y., & Balakrishnan, N. (1986). *Robust inference*. New York: Dekker.

Thomas, D.R. (1993). Hypothesis testing using data from clustered samples. *Canadian Psychology, 34*(4), 415-431.

Tukey, J.W. (1977). Robust techniques for the user. In R.L. Launer, & G.N. Wilkinson (Eds.), *Robustness in statistics*. New York: Academic Press.

Wainer, H. (1982). Robust statistics: A survey and some prescriptions. In G. Keren (Ed.), *Statistical and Methodological Issues in Psychology and Social Sciences Research.* Hillsdale, NJ: Lawrence Erlbaum.

Wainer, H. (1989). Eelworms, bullet holes, and Geraldine Ferraro: Some problems with statistical adjustment and some solutions. *Journal of Educational Statistics, 14*, 121-140.

Zimmerman, D.W., & Zumbo, B.D. (1993). Relative power of parametric and nonparametric statistical methods. In Gideon Keren & Charlie Lewis (Eds.), *A handbook for data analysis in the behavioral sciences, Vol. 1: Methodological issues* (pp. 481-517). Hillsdale, NJ: Lawrence Erlbaum.

Zumbo, B.D. & Zimmerman, D.W. (1991). Further evidence for Coren & Hakstian's, "Methodological implications of interaural correlation: Count heads not ears" and an alternative analysis. *Perception & Psychophysics, 50*, 297-301.