The Power of Outliers in Research: What actually Works, and Does it Matter?

Dila Ram Bhandari
Lecturer, Nepal Commerce Campus, Tribhuvan University
Email: drbhandari2012@gmail.com
Kapil Shah
Lecturer, Nepal Commerce Campus, Tribhuvan University
Aayan Bhandari
Data Science, SENECA College, Toronto, Canada

https://doi.org/10.3126/pravaha.v30i1.76894

Abstract

Outliers have an important and diverse role in the social sciences, particularly when seen via a statistical lens. While outliers are frequently perceived as abnormalities or departures from the norm, they can contribute critical insights and improve our knowledge of social processes. Outliers, sometimes referred to as anomalies in datasets, play an important role in the development of research. While typically regarded as a threat to statistical integrity, their existence can produce surprising insights and breakthrough findings when managed correctly. This article investigates the varied nature of outliers, their influence on research methodology, and their contribution to significant scientific advances. We examine how to successfully discover, analyze, and use outliers, balancing their potential for innovation against the risk of drawing incorrect conclusions.

Keywords: *Outliers, Statistics, Detect, Power, Matter*

Cite this paper
Bhandari, D. R., Shah, K., & Bhandari, A. (2024). The Power of Outliers in Research: What Actually Works, and Does it Matter? *Pravaha, 30*(1), 84-91.

Introduction

Data consistency and correctness are critical for generating accurate study outcomes. However, not all data points follow predicted trends. These aberrations, known as outliers, are frequently ignored or corrected to ensure the robustness of statistical models. This strategy, although practical, ignores outliers' enormous ability to challenge old paradigms and generate new ideas. Outliers might be noise, mistakes, or unique happenings. Understanding their causes and consequences is crucial for making educated decisions regarding therapy. This essay investigates the dual nature of outliers, showing instances in which their inclusion has either enhanced or complicated research conclusions.

Outliers have an important and diverse role in the social sciences, particularly when seen via a statistical lens. While outliers are frequently perceived as abnormalities or departures from the norm, they can contribute critical insights and improve our knowledge of social processes. Outliers can cause exaggerated error rates and significant distortions in parameter and statistic estimations when applying parametric or nonparametric tests (e.g., Zimmerman, 1994, 1995, 1998). A cursory examination of the literature reveals that researchers seldom mention looking for outliers of any type. This conclusion is backed experimentally by Osborne, Christiansen, and Gunter (2001), who discovered that authors reported evaluating assumptions of the statistical procedure employed in their research, including screening for outliers, just 8% of the time.

Outliers are commonly characterized as data points that fall beyond a dataset's normal range, frequently many standard deviations from the mean. Outliers in social sciences can refer to distinctive individual behaviors, unusual events, or extraordinary situations. Traditional statistical approaches frequently strive to reduce the impact of

outliers to produce a more uniform dataset. However, disregarding outliers might result in the loss of valuable information and insights. This essay tries to emphasize the significance of outliers, demonstrating how they may improve our knowledge of social phenomena, detect structural cracks, and help construct more robust theories and policies. In income distribution research, outliers reflecting exceptionally high incomes can shed light on wealth concentration and economic inequality. These outliers emphasize the need for measures that promote wealth redistribution and social fairness. Outliers in health statistics, such as unexpected increases in illness incidence, can contribute to advances in understanding and avoiding health problems. Outlier cases identified in epidemiological studies, for example, might motivate more researc h into potential causes and risk factors. Outliers in educational performance, such as very high-achieving kids or those with major learning disabilities, can help shape educational policy and practices. These instances highlight the value of tailored learning methodologies and resource allocation.

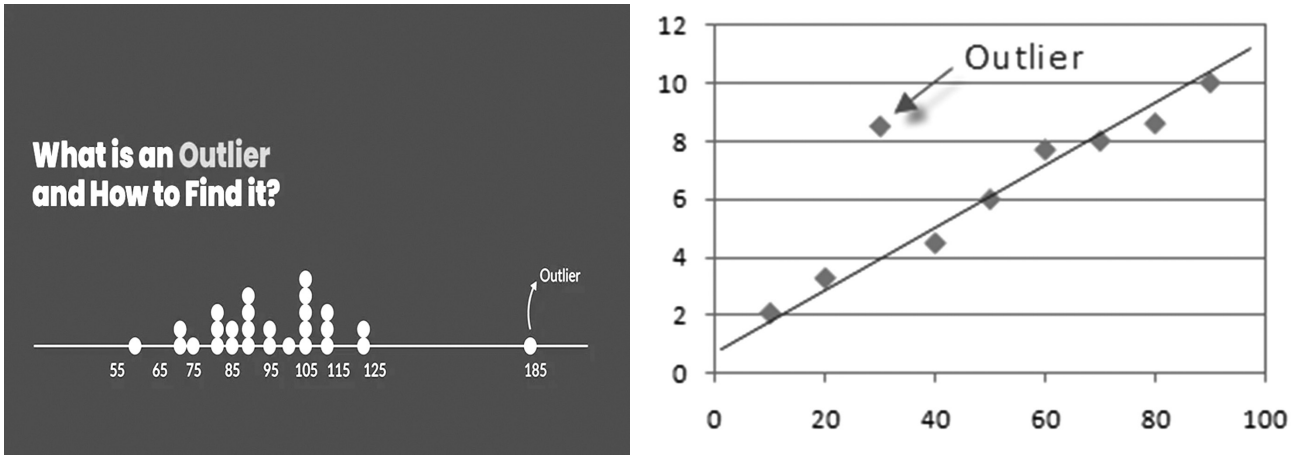


Figure 1: outlier

Defining Outliers

Outliers are data points that dramatically differ from the remainder of the dataset.

- Measurement errors: such as instrument imperfections or recording faults, might cause these issues.
- Errors in data entry: might be typographical or procedural.
- Natural Variability: refers to rare or exceptional events.
- While outliers are sometimes referred to as extreme numbers or anomalies, their classification is dependent on the context and research subject at hand.

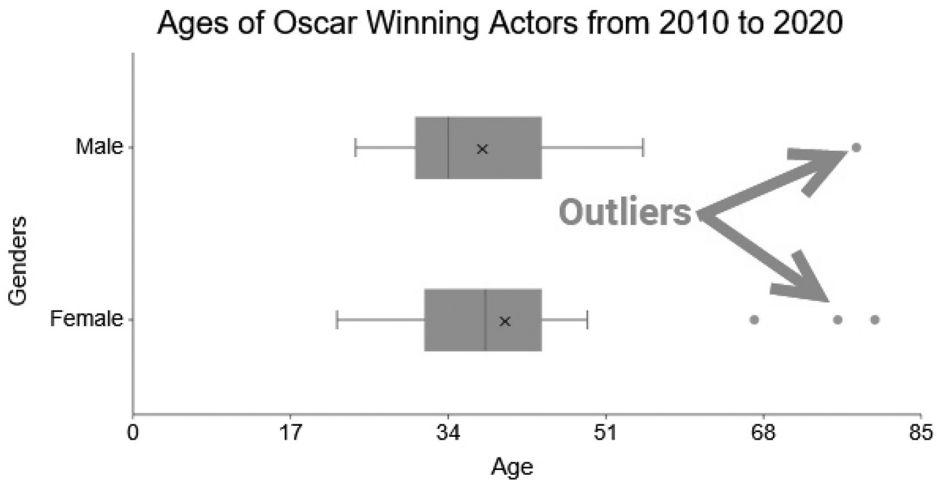


Figure 2: Outlier example

Care of Outliers

Although definitions differ, an outlier is commonly defined as a data point that deviates significantly from the norm for a variable or population (Jarrell, 1994; Rasmussen, 1988; Stevens, 1984). Hawkins defined an outlier as "an observation that deviates so significantly from other observations as to raise suspicions that it was generated by a different mechanism" (Hawkins, 1980, p.1). Outliers can also be characterized as values that are "dubious in the eyes of the researcher" (Dixon, 1950, p. 488) or pollutants (Wainer, 1976).

Wainer (1976) also coined the term "fringelier," which refers to "unusual events that occur more often than seldom" (p. 286). These points are close to three standard deviations from the mean and hence may have a disproportionately high effect on parameter estimations, although they are less evident or easily identifiable than conventional outliers due to their closeness to the distribution center. Because fringeliers are a subset of outliers, throughout the rest of the paper, we shall use the word "outlier" to refer to any single data point of questionable provenance or excessive effect.

Outliers can skew statistical results. First, they tend to increase error variance and diminish the power of statistical tests. Second, if they are not randomly distributed, they can reduce normality (and, in multivariate analysis, violate the principles of sphericity and multivariate normality), increasing the likelihood of making Type I and Type II mistakes. Third, they have the potential to significantly distort or impact estimates that are of substantial importance (see Rasmussen, 1988; Schwager & Margolin, 1982; Zimmerman, 1994). Screening data for univariate, bivariate, and multivariate outliers is straightforward in today's ubiquitous computing. Failure to do so may have serious implications.

Causes Outliers

Outliers can result from a variety of reasons. Outliers, according to Anscombe (1960), are classified into two types: those caused by data mistakes and those caused by the data's intrinsic variability. Not all outliers are illicit pollutants, and not all illegitimate scores appear as outliers (Barnett and Lewis, 1994). It is so critical to analyze the variety of causes that may be responsible for outliers in any data collection.

Outliers from data errors

Outliers are frequently caused by human error, such as mistakes in data collecting, recording, or input. Data from an interview may be captured inaccurately during data entry. The first author participated in one study (published in Brewer, Nauenberg, & Osborne, 1998) that collected data on nurses' hourly pay, which at the time averaged around $14.00 per hour with a standard variation of roughly $2.00. In dataset, one nurse claimed an hourly pay of $44,000.00. This graphic depicted a data collecting mistake (particularly, the respondent's failure to read the question thoroughly). Errors of this kind are frequently addressed by returning to the source documents or, if required, the subjects and inputting the proper value.

In circumstances such as the nurse who earned $44,000 per hour, another alternative is to recalculate or re-estimate the accurate answer. We utilized anonymous questionnaires, but because the nature of the inaccuracy was clear, we were able to convert this nurse's income to an hourly compensation because we knew how many hours per week and how many weeks per year she worked. Thus, assuming enough information is available, recalculation provides a means of preserving vital data while removing a visible outlier. If outliers of this type cannot be fixed, they should be removed since they do not reflect legitimate population data points.

Outliers from misreporting

Sometimes individuals purposely submit erroneous data to experimenters or surveys. A participant may make a conscious effort to sabotage the research (Huck, 2000) or may behave for other reasons. Social desirability and self-presentation motivations can be significant. This can also occur for apparent reasons when data is sensitive (for example, teens under-reporting drug or alcohol usage, misreporting of sexual activity). If most youth under-report a behavior (e.g., the frequency of sexual fantasies experienced by teenage boys), the few honest replies may look like outliers, while being authentic and valid scores.

Outliers from sampling error

Outliers can also be caused by sampling. It is possible that a few individuals of a sample were accidentally picked from a different population than the remainder of the sample. For example, in the previously stated nurse salary survey, we selected from a database that included RNs who had moved into hospital administration, despite our preference for floor nurses. In education, it is possible to mistakenly sample academically exceptional or mentally retarded kids, which may result in unwanted outliers (depending on the study's purpose). These examples should be eliminated since they do not represent the intended demographic.

Outliers from faulty distribution

Incorrect assumptions regarding data distribution can also result in the emergence of suspected outliers (Iglewicz & Hoaglin, 1993). Blood sugar levels, disciplinary referrals, well-prepared classroom test scores, and self-reports of low-frequency behaviors (e.g., the number of times a student has been suspended or held back a grade) may result in bimodal, skewed, asymptotic, or flat distributions, depending on the sampling design. Similarly, the data structure may differ from what the researcher anticipated, and long or short-term patterns may have an unexpected impact on the data. For example, in the following dataset, the times recorded for Employee ID 6 (180 minutes) and Employee ID 9 (250 minutes) are significantly higher than the rest. These values are outliers and could be the result of a faulty distribution, such as errors in data recording or measurement.

Employee ID	12	3	4	5	6	7	8	9
Task Completion (minutes)	40	50	53	55	180 outlier	45	65	250 Outlier

Outliers as legitimate cases

An outlier might result from a valid random sampling of the population. It is vital to note that sample size influences the likelihood of outlier values. In a normally distributed population, a particular data point is more likely to come from the most densely concentrated portion of the distribution rather than one of the tails (Evans, 1999; Sachs, 1982). As a researcher casts a broader net and the data collection grows, the sample becomes more representative of the population from which it was obtained, increasing the risk of outlying results.

In other words, there is only around a 1% probability of getting an outlier data point from a normally distributed population; this indicates that approximately 1% of subjects should be 3 standard deviations off the mean. Outliers occur because of the data's intrinsic unpredictability, and perspectives on what to do vary greatly. Because outliers and fringeliers can have negative effects on power, accuracy, and error rates, it may be desirable to use a transformation or recoding strategy to keep the individual in the data set while minimizing the harm to statistical inference (Osborne, 2002).

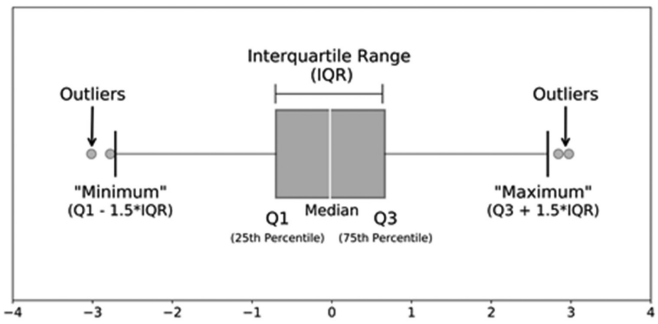
Outliers as a potential focus of inquiry

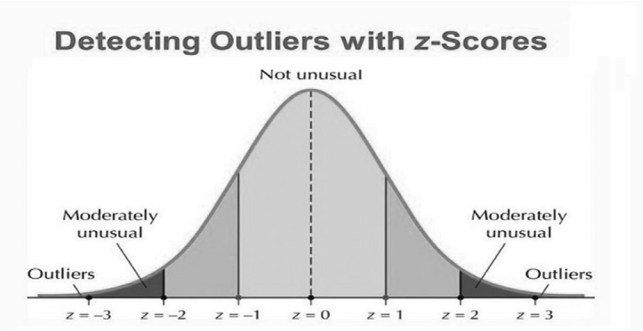
Outliers might be considered annoyance, mistake, or real data. They can also spark curiosity. When researchers in Africa discovered that some women had been living with HIV for years and years without treatment, they were anomalies in comparison to most untreated women, who died quickly. They could have been dismissed as noise or mistake, but instead, they serve as a source of inspiration for further research: what distinguishes these women, and what can we learn from them? In research in which the first author participated, a teenager claimed to have 100 close friends. Is this possible? Yes. Is this likely? Not typically, based on any plausible definition of "close friends." So, this data point may indicate intentional misreporting, a data recording or input error (it wasn't), a protocol fault caused by a misunderstanding of the inquiry, or something more fascinating. This exceptional score may give insight into a key concept or issue. Before rejecting outliers, researchers must assess if the data contains valuable information that may not be relevant to the targeted study but is important in a broader sense.

Detection and Analysis of Outliers

Outliers are data points that differ dramatically from the overall trend or pattern of the dataset. They have a substantial influence on statistical analyses and machine learning models. Each technique has advantages and is appropriate for various data kinds and settings. It's critical to understand the nature of your data and select the best strategy for finding outliers.

- *Statistical Tests:* Methods like Grubbs' test, Dixon's Q test, and the Z-score can help identify outliers.
- *The Z-Score* (Standard Score) is a way of determining how much a data point deviates from the mean. Data points having a Z-score larger than 3 or less than -3 are frequently considered outliers.
- *Graphical Methods:* Scatter plots, box plots, and histograms are useful for visual identification of outliers.
- *The interquartile range (IQR)* is the difference between the data's first and third quartiles (Q1 and Q3). Outliers are commonly characterized as data points that fall below Q1 - 1.5*IQR or above Q3 + 1.5*IQR. A box plot depicts the distribution of data and emphasizes outliers as isolated points outside the whiskers.
- *Advanced machine learning* approaches, such as Isolation Forests, Local Outlier Factor (LOF), and One-Class Support Vector Machines (SVM), may efficiently model regular data while detecting abnormalities. Outliers are data points that differ dramatically from the remainder of the dataset. Their existence can cause distortions in analysis, therefore finding and deleting them can enhance data integrity.
- *Robust Statistics:* Focuses on statistical methods that are not unduly affected by outliers, such as the use of median and interquartile range instead of mean and standard deviation.





There is as much controversy over what constitutes an outlier as whether to remove them or not. Simple rules of thumb (e.g., data points three or more standard deviations from the mean) are good starting points. Some researchers prefer visual inspection of the data. Others (Lornez, 1987) argue that outlier detection is merely a special case of the examination of data for influential data points.

Simple rules like z = 3 are simple and relatively effective, but Miller (1991) and Van Selst and Jolicoeur (1994) demonstrated that this procedure (no recursive elimination of extreme scores) can cause problems with certain distributions (for example, highly skewed distributions characteristic of response latency variables), especially when the sample is small. To assist researchers in dealing with this issue, Van Selst and Jolicoeur (1994) provide a table of proposed cutoff scores for researchers to employ with varied sample sizes to reduce concerns with very non-normal distributions. We often utilize a z = 3 guideline as a first screening tool, then study the data more thoroughly and adjust the outlier identification technique accordingly.

Most current statistical software can provide a variety of statistics for ANOVA paradigms, including standardized residuals. The most significant difficulty in ANOVA after screening for univariate outliers is the issue of within-cell outliers, or an individual's distance from the subgroup. Standardized residuals show the distance from the subgroup and are thus useful in aiding analysts in identifying multivariate outliers. Tabachnick and Fidell (2000) examine data cleansing about other analyses.

Outliers as Sources of Error

On the flip side, outliers can distort statistical analyses, leading to:

- *Skewed Means:* Extreme values disproportionately influence averages.

- *Model Misfit:* Outliers can disrupt regression models, leading to incorrect predictions.
- *False Positives/Negatives:* In hypothesis testing, outliers may generate misleading results, compromising validity.

Importance and Role of Outliers

- *Identifying hidden patterns:* Outliers may indicate the presence of underlying patterns or tendencies that are not immediately evident. Outliers in income statistics, for example, may disclose the presence of extreme wealth or poverty, necessitating more research into socioeconomic discrepancies.
- *Testing hypotheses:* Outliers can question established beliefs and theories. In the social sciences, they might reflect situations that differ from predicted behavior, forcing researchers to reassess their models or assumptions. This careful study can provide more solid and complete theories.
- *Improving Model Accuracy:* o Identifying and resolving outliers improves statistical model accuracy. Understanding why specific data points differ dramatically allows researchers to adjust their models to account for these anomalies, resulting in more accurate predictions and conclusions.
- *Policy implications:* Outliers in policymaking can identify exceptional occurrences that necessitate focused solutions. For example, recognizing places with abnormally high unemployment rates might aid in the development of targeted economic strategies to meet local difficulties.

Do Outliers Matter?

The significance of outliers depends on the research context:

- In exploratory research, outliers often illuminate hidden patterns and new directions.
- In confirmatory studies, they may undermine the reliability of results.

Ultimately, outliers are important when they lead to useful discoveries or a better knowledge of events. Their importance is heightened in subjects such as medicine, physics, biology and social sciences, where unusual events can have disproportionate consequences.

Public Health and Disease Outbreaks

Consider the COVID-19 epidemic, which is a notable anomaly in global health data. The number of cases, hospitalizations, and fatalities differed significantly from the regular yearly patterns seen in respiratory infections. The study of this anomaly underscores the need for pandemic planning and response tactics. It teaches important lessons about healthcare infrastructure, legislation, and the socioeconomic implications of global health crises.

Consumer Behavior

In a study of online purchasing behavior, an outlier might be a buyer who makes an exceptionally significant number of purchases within a given period, such as Black Friday, festival, Christmas or Cyber Monday.

Investigating this anomaly can reveal information on customer behavior during peak purchasing seasons. It may help firms improve their marketing strategy, inventory management, and customer interaction procedures. There are a lot of disputes over what to do with detected outliers.

It is just basic sense to eliminate outliers from data that have been inserted illegitimately (Barnett and Lewis, 1994). When the outlier is either a valid component of the data or the reason is unknown, the situation becomes more complicated. Judd and McClelland (1989) argue that even in these circumstances, elimination is necessary to provide the best accurate estimate of population characteristics (see also Barnett & Lewis, 1994). However, not all scholars agree (see Orr, Sackett, & DuBois, 1991). In this scenario, researchers must utilize their training, intuition, reasoned reasoning, and careful deliberation to make choices.

Outliers: to remove, or not to remove?

Although some writers believe that removing extreme scores results in bad outcomes, they are in the minority,

particularly when the outliers are illegitimate. When data points are thought to be real, some writers (Orr, Sackett, & DuBois, 1991) claim that if outliers are not deleted, the data is more likely to be typical of the population. There are compelling conceptual justifications for removing or altering outliers. Both correlations and t-tests exhibited substantial statistical changes as outliers were removed, and in most analyses, estimate accuracy improved. In most cases, inference errors were greatly decreased, highlighting the importance of screening for and removing outliers. Although these were two basic statistical processes, it is easy to argue that the benefits of data cleansing apply to simple and multiple regression, as well as many forms of ANOVA techniques.

Ethical Considerations

Researchers must describe the presence of outliers, as well as their judgments on whether to include, exclude, or change such data. Manipulating outliers to attain desired results is unethical action. Removing outliers without adequate rationale might distort data and lead to incorrect conclusions. Ethical research necessitates thorough evaluation to identify whether outliers are legitimate findings or mistakes. Outlier decisions must be thoroughly recorded for the research to be repeated. The lack of openness in dealing with outliers affects the credibility of the conclusions. Outliers can indicate unusual but noteworthy instances, especially in sectors such as medicine or environmental research. Ignoring them can result in policies that fail to consider vulnerable people or severe events. It is immoral to intentionally eliminate outliers to improve statistical findings, such as making a treatment appear more beneficial than it is.

Handling Outliers

Outliers are an important part of data analysis, and knowing their theoretical foundation enables improved data interpretation and decision-making.

- Ignoring Outliers: Outliers may be ignored if they have minimal impact on the analysis.
- Data modifications, such as log transformation, can mitigate the impact of outliers.
- Outliers may need to be removed, particularly if they are caused by mistakes.
- Adjusting Analysis: Use robust regression or non-parametric approaches to reduce the impact of outliers on results.

Conclusion

The Dual Role of Outliers is both disruptive and illuminating. Their combined position as statistical outliers and windows into deeper truths renders them useful in the study. In criminology, investigating extreme criminal behavior (outliers) frequently results in novel profiling and prevention measures. Outliers are more than just statistical inconveniences; they are valuable resources for discovery in the social sciences. As the frontiers of social phenomena increase, outliers play an increasingly important role in changing our knowledge. Outliers play an important role in social sciences because they challenge popular knowledge, expose hidden patterns, and stimulate additional research. Outliers may lead to more accurate models, complete theories, and informed policy decisions when carefully identified, analyzed, and handled statistically. While outliers might be difficult to examine, they can provide useful insights into a variety of phenomena across several domains. By investigating these high numbers, researchers might discover hidden patterns, confirm or challenge current theories, and guide practical applications and policy choices. Implications for Policy and Practice Policymakers may utilize lessons from outliers to create customized solutions that address unique needs and issues. Identifying outlier locations with high unemployment rates, for example, can result in the creation of targeted job training and economic development strategies. Recognizing the importance of outliers fosters inclusive research techniques that consider varied viewpoints and experiences. This method encourages a more thorough knowledge of social issues while also promoting fairness in research and policymaking. By accepting outliers as useful sources of knowledge, researchers may build stronger hypotheses, increase data quality, and influence successful policy.

References

Anscombe, F. J. (1960). Rejection of outliers. *Technometrics, 2,* 123-147.

Barnett, V, & Lewis, T. (1994). *Outliers in statistical data* (3rd ed.). New York: Wiley.

Brewer, C. S., Nauenberg, E., & Osborne, J. W. (1998, June). *Differences among hospital and non-hospital RNs participation, satisfaction, and organizational commitment in western New York.* Paper presented at the National meeting of the Association for Health Service Research, Washington DC.

Dixon, W. J. (1950). Analysis of extreme values. *Annals of Mathematical Statistics, 21,* 488-506.

Evans, V.P. (1999). Strategies for detecting outliers in regression analysis: An introductory primer.

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

Huck, S.W. (2000). *Reading statistics and research* (3rd ed.). New York: Longman.

Iglewicz, B., & Hoaglin, D.C. (1993). *How to detect and handle outliers.* Milwaukee, WI.: ASQC Quality

Jarrell, M. G. (1994). A comparison of two procedures, the Mahalanobis Distance and the Andrews-Pregibon Statistic, for identifying multivariate outliers. *Research in the schools, 1,* 49-58.

Judd, C. M., & McClelland, G. H. (1989). *Data analysis: A model comparison approach.* San Diego, CA.: Harcourt Brace Jovanovich.

Lornez, F. O. (1987). Teaching about influence in simple regression. *Teaching Sociology, 15*(2), 173-177.

Miller, J. (1991). Reaction time analysis with outlier exclusion: Bias varies with sample size. *The quarterly journal of experimental psychology, 43*(4), 907-912.

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://ericae.net/pare/getvn.asp?v=8&n=6.

Press.

Rasmussen, J. L. (1988). Evaluating outlier identification tests: Mahalanobis D Squared and Comrey D. *Multivariate Behavioral Research, 23*(2), 189-202.

Sachs, L. (1982). *Applied statistics: A handbook of techniques* (2nd ed). New York: Springer-Verlag.

Schwager, S. J., & Margolin, B. H. (1982). Detection of multivariate outliers. *The annals of statistics, 10,* 943954.

Stevens, J. P. (1984). Outliers and influential data points in regression analysis. *Psychological Bulletin, 95,* 334344.

Tabachnick, B.G., & Fidell, L. S. (2000). Using multivariate statistics, 4th edition. Pearson Allyn & Bacon.

Thompson (Ed.), *Advances in social science methodology:* (Vol. 5, pp. 213-233). Stamford, CT.: JAI Press.

Van Selst, M., & Jolicoeur, P. (1994). A solution to the effect of sample size on outlier elimination. *The quarterly journal of experimental psychology, 47*(3), 631-650.

Wainer, H. (1976). Robust statistics: A survey and some prescriptions. *Journal of Educational Statistics, 1*(4), 285-312.

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 nonparametric statistical tests by concurrent violation of two assumptions. *Journal of Experimental Education, 67*(1), 55-68.