Handling Those Pesky Statistical Outliers

A recent experience of mine (C.R.K.) when exploring a secondary data set was the impetus for this editorial. I found that three meaningful predictor variables accounted for 42% of the variance in future health problems of nursing home residents. Buoyed by this discovery, I set out to conduct a more formal analysis that included checking the distributions of each variable. After removing one extreme outlier, the $R^2$ dropped a whopping 20 points to 22%. This experience forced some reflection regarding the ethical, conceptual, and practical considerations involved in deciding whether to remove extreme scores. So, let’s revisit this well-known statistical quirk and explore a few considerations when deciding how to manage these mavericks in data sets.

An outlier is an extreme value that differs greatly from other points in a data set. Outliers can distort statistical analyses in a number of ways. Outliers tend to increase error variance and reduce the power of statistical tests, and often skew the distribution of scores. Although simple statistics, such as the $t$ test, are robust to some violations of the distribution from normality, extreme scores increase the error rates and odds of both Type I and II errors. Analyses based on correlations are particularly vulnerable to extreme scores because even one extreme case can significantly change the relationships between two or more variables. In multivariate analyses, outliers can violate assumptions of sphericity and multivariate normality, which can warp analyses.

There are two categories of outliers: univariate and multivariate. Univariate outliers are extreme values on a single variable. Multivariate outliers are extreme combinations of scores on two or more variables. Modern statistical software packages often include tests for outliers. For example, when running SPSS software, univariate outliers can be identified using Analyze->Descriptive Statistics->Explore procedure. Multivariate outliers can be identified using Cook’s distance, which can be calculated in analysis of variance or regression analysis (Weinberg & Abramowitz, 2002). A univariate outlier may also be determined through visual inspection of a distribution or from cutoffs, such as being three or more standard deviations from the mean or sitting 1.5 interquartile ranges below the first or above the third quartile (Dawson, 2011; Osborne & Overbay, 2004).

Outliers are caused by errors in the data or inherent variability in the data. Errors in data collection, recording, or entry can often be easily corrected by returning to the original data collection document, checking the statistical file for incorrect entries, or recontacting the responding participant. If an outlier caused by errors in the data cannot be corrected, it should be deleted, because it does not represent a valid value in the sample or population.

Multiple factors can contribute to outliers caused by variability in the data. A participant can provide socially desirable responses or be motivated by another factor to misrepresent responses. Variations in research procedures or extraneous factors, such as fatigue or stress, that confound the study or contribute to measurement error can create extreme scores. If several participants are fundamentally different than the target population, this sampling error could result in outliers. An outlier can also come from the population being sampled by random chance. There is no systematic bias involved and these data may represent a legitimate part of the population.

There are both conceptual and practical reasons for dropping some outliers that seem to represent a legitimate part of the population. Results from one or a few participants who are extremely different from the sample as a whole may indicate that a fundamentally different mechanism of action is at play. Outliers also tend to reduce power,
distort results, and abrogate the proper use of common, easily understandable parametric statistical tests. If outliers are dropped from a study, consider whether those cases raise interesting questions that should be examined in a future study, or point to a possible serendipitous finding.

In what instances should we decide that, rather than being interlopers, extreme scores hold a rightful place in our data set? Outliers should be retained if the conceptual logic or ethical standards of the study are compromised by deleting these cases. If the motive for retaining the cases is to intentionally distort or misrepresent results in the direction of the principal investigator’s biases, this constitutes a breach of ethics. There is no standard rule or agreement about how to manage outliers that are a legitimate part of a data set. Transformations and truncation are two methods used when outliers remain in a data set. Transforming data will maintain the relative ranking of scores, while decreasing the skew and error variance. However, results from transformed data are more difficult to interpret and present clearly when disseminated. Truncation is the recoding of extreme scores to the highest or lowest reasonable score. A variety of more complex and robust estimation methods are available for managing outliers. The trimmed mean (Anscombe, 1960) and winsorized mean (Barnett & Lewis, 1994) can be used for managing univariate outliers, whereas the least trimmed squares regression can be used to manage multivariate outliers (Rousseeuw & Van Driesen, 2006).

Outliers are common and often unwelcome data points—annoying nuisances that represent error. On the other hand, extreme scores can represent legitimate findings that lead to a new research question or area of inquiry. After investigating the outlier in my data set, I decided it was a legitimate case and used the truncation method to avoid unfairly inflating and misrepresenting the results. The quality of our science and the public health are jeopardized if we fail to thoughtfully identify and manage outliers.

REFERENCES

Christine R. Kovach, PhD, RN, FAAN, FGSA
Editor
Jewish Home and Care Center Research Professor in Aging

Weiming Ke, PhD
Associate Researcher/Statistician
University of Wisconsin–Milwaukee
Milwaukee, Wisconsin

The authors have disclosed no potential conflicts of interest, financial or otherwise.

doi:10.3928/19404921-20160902-01