6-5-2018

Winsorizing

Bruce E. Blaine
St. John Fisher College, bblaine@sjfc.edu

Publication Information

Blaine, Bruce E. (2018). "Winsorizing." *The SAGE Encyclopedia of Educational Research, Measurement, and Evaluation* , 1817-1818.

Winsorizing

Abstract

In lieu of an abstract, here is the entry's first paragraph:

Winsorizing is a procedure that moderates the influence of outliers on the mean and variance and thereby creates more robust estimators of location and variability. The procedure is named for biostatistician Charles P. Winsor. Parametric inferential procedures that rely on the mean and variance (e.g., t test) become more robust when they incorporate Winsorized estimators. Winsorizing is an important tool for educational and social science researchers for two reasons. First, significance tests based on the mean and variance are very common procedures for significance testing in the social sciences. Second, surveys of the educational and psychological literature show that nonnormally distributed data are the rule rather than the exception, and even modest departures from normality disproportionately affect the mean and variance compared with other more robust estimators of location (e.g., median) and variability (e.g., median absolute deviation)

Disciplines

Statistics and Probability

Comments

This is an entry in: Frey, B. (2018). *The SAGE encyclopedia of educational research, measurement, and evaluation* (Vols. 1-4). Thousand Oaks,, CA: SAGE Publications, Inc.: https://dx.doi.org/10.4135/9781506326139

WINSORIZING

Winsorizing is a procedure that moderates the influence of outliers on the mean and variance and thereby creates more robust estimators of location and variability. The procedure is named for biostatistician Charles P. Winsor. Parametric inferential procedures that rely on the mean and variance (e.g., *t* test) become more robust when they incorporate Winsorized estimators. Winsorizing is an important tool for educational and social science researchers for two reasons. First, significance tests based on the mean and variance are very common procedures for significance testing in the social sciences. Second, surveys of the educational and psychological literature show that nonnormally distributed data are the rule rather than the exception, and even modest departures from normality disproportionately affect the mean and variance compared with other more robust estimators of location (e.g., median) and variability (e.g., median absolute deviation).

Winsorizing reassigns values to a percentage of cases in both tails of a distribution using the next highest (in the lower tail) and lowest (in the upper tail) value; the resultant variable is said to have been Winsorized. The Winsorized mean is the mean of the Winsorized values, and the Winsorized variance is the average squared deviation of the Winsorized values from the Winsorized mean. Consider, as an example, a variable with the following values:

2, 2, 3, 3, 3, 4, 5, 8, 15, 25.

The mean and variance of this variable are 7.0 and 55.6, respectively. Owing to the presence of at least one unusually large score, these estimators perform poorly in characterizing location and variability. A 20% Winsorizing procedure identifies the lowest (e.g., 2, 2) and highest (e.g., 15, 25) 20% of the cases. These cases reassigned the value of the adjacent upper (e.g., 3) or lower (e.g., 8) case, producing this Winsorized variable:

3, 3, 3, 3, 3, 4, 5, 8, 8, 8.

The 20% Winsorized mean and variance are 4.8 and 5.3, respectively. From intuitive and statistical perspectives, these statistics are better estimators of the location and variability of the original variable.

For moderating the influence of outliers on the mean and variance and creating better estimators from skewed or heavy tailed data, Winsorizing is an alternative to trimming. The most common

levels of Winsorizing are 10% and 20%, although this decision is up to the researcher; greater levels of Winsorizing create more robust estimators. Relative to trimmed means and variances, Winsorizing does not cast aside data and create associated issues (e.g., reduced degrees of freedom). Functions for computing Winsorized means and variances, other Winsorized estimators (e.g., Winsorized product–moment correlation coefficient), and inferential procedures incorporating Winsorized estimators are available in some statistical software packages. Notably, there are numerous packages in R for doing robust descriptive and inferential data analysis, including Winsorized procedures.

Bruce E. Blaine

See also Missing Data Analysis; Robust Statistics

Further Readings

Erceg-Hurn, D., & Mirosevich, V. (2008). Modern robust statistical methods: An easy way to maximize the accuracy and power of your research. *American Psychologist, 63*(7), 591–601. doi:10.1037/0003-066X.63.7.591.

Wilcox, R., & Keselman, H. (2003). Modern robust data analysis methods: Measures of central tendency. *Psychological Methods, 8*, 254–274.