

Descriptive Statistics

A Evans

May 2003

Means vs. Medians

Median is better than Mean as descriptor of the center of a distribution if there are extreme values at one end.

Variance

The variance is the average squared difference from the mean. Calculate by summing the squared differences and dividing by the number of observations minus one ($n-1$). For the best estimate of the *population* variance, you must divide by $n-1$, instead of n . This adjustment makes it a little bigger, and thus accounts for the extra uncertainty in estimating the population variance when the sample mean, rather than the true population mean, is used.

Standard Deviation

Square root of the variance; it describes the spread of the data in terms of the original units. For a normal population, 68% of observations are within 1 standard deviation of the mean and 95% are within 2 standard deviations. However, for non-normal distributions, this isn't true; instead, you only know that the proportion of the population within k standard deviations from the mean is at least $\geq 1 - (1/k)^2$ (Chebychev's Inequality). Therefore, for any distribution (regardless of the shape), at least 75% of all observations will fall within 2 standard deviations of the mean.

Range, Interquartile Range, SD

Range often not helpful because of extreme values.

Standard Deviation not helpful if distribution skewed.

Reporting the interquartile range (25th and 75th percentiles) or reporting 10th and 90th percentiles is preferable for skewed distributions.

Compare the use of SD vs. use of percentiles to describe distributions:

SPSS: Analyze: Descriptive: Frequencies: Statistics: Percentiles:

Type in 2.5%, click ADD; do the same for 97.5%. Thus, 95% of all observations will fall between these values. Compare, however, these values with the values you calculate from $\text{mean} \pm 1.96 \times \text{standard deviation}$. (Unselect *Display Frequency Tables* if this is interval data; otherwise there will be a long table with each possible value.)

Standard Error of the Mean

This is equivalent to the standard deviation of sample *means*, rather than standard deviation of individual observations. ($SE \text{ or } SEM = SD/\sqrt{n}$). It describes the spread of sample means around the true population mean. Therefore, 95% of sample means will be within 2 standard errors of the true population mean. Since you usually have only one sample and one sample mean, it makes more sense to consider SE to be a measure of the precision in estimating the population mean (using the sample mean as your point estimate). Therefore, you can be 95% confident that the true population mean falls within 2 standard errors of the sample mean.

Standardizing Values

To “standardize” a variable means that you change the units of measurement from the original units (years, packs-per-year, mm Hg, etc) to a common metric: standard deviations. Change each value into the distance (in standard deviations) from the Mean. Do this by first subtracting the Mean from each value and then dividing by the standard deviation. A standardized value of 0.8 would mean that the observation is 0.8 standard deviations above the mean. A value of -1.3 would mean that the observation is 1.3 standard deviations below the mean. Standardized values are often called *z-scores*.