Descriptive Statistics

The purpose of a descriptive statistic is to summarize data. Descriptive stats only make statements about the set of data from which they were calculated; they never go beyond the data you have. In other words, even though a sample is always taken from a population, and even though our long-term goal is usually to make a statement about the population (and not just the sample), descriptive stats only make statements about the sample.

⊕ Warning: the authors of SPSS seem to have missed this point entirely.

In general, to summarize a set of psychological data, we want to be able answer three questions: Where is the distribution located? How spread out are the values? And, what does the distribution look like? These are referred to as *center*, *spread*, and *shape*. Note that some people think of spread as being part of shape, which is fine, but not optimal. The reasons for using three attributes are that it better matches what people actually do in practice and that it better matches how we approach the issue of making assumptions, which will soon become crucial. For most variables, we usually provide a number for center, another number for spread, and then a verbal label for shape.

Central Tendency

The two most popular measures of central tendency are the arithmetic mean (symbol: $\overline{X}$; aka the first moment of the distribution) and the median (symbol X_{50}). If the distribution is symmetrical, then these two values are the same.

☞ Note: some people seem to believe that it is better to use the median than to use the mean when the sample is small or asymmetrical (i.e., skewed). In particular, I've seen it said that the median provides a better measure of center when the sample is small and/or skewed, because the median is less sensitive to outliers. This is, to be rude and blunt, utter nonsense. The median is never a better (or worse) measure of center; it is always a *different* measure of center, because it uses a different definition of *center*. Put another way: you should use whatever measure of center that you are interested in. If you are interested in the "center of gravity" of your data, then use the mean; if you are interested in the "middle value" within your data, then use the median.

☞ Another argument in favor of the median should probably be saved until we get to inferential statistics, but I'll include it here. The claim is that the sample median provides a better estimate of the population mean when the sample is small and/or skewed, again because the median is less sensitive to outliers. This sounds great (because many people worry about the effects of outliers and would love a "trick" to avoid their effects), but this also turns out to be false. As has been shown by Monte Carlo simulations (by Jeff Miller & Albano Lopes @ U.C. San Diego), the sample median actually becomes a worse and worse estimate of the population mean as the sample becomes smaller and/or more skewed. In general, if someone says that a certain "trick" does a good job of avoiding the unwanted effects of some bad thing (e.g., outliers), ask them for some evidence. Many such methods are elderly spouse stories with no basis in fact.

Spread

Paralleling the mean and the median, there are two main measures of spread. The one that goes hand-in-hand with the sample mean is the sample variance (symbol: S^2; aka the second moment of the distribution); alternatively, you can use the square-root of the sample variance, which is the sample standard deviation (symbol: S). I prefer S over S^2 because S has the same units as the mean, instead of the units squared. I know what a square-foot is, but I have great difficulty imagining what a square-millisecond is, so I avoid it as much as I can.

The measure of spread that goes with the median is the inter-quartile range (acronym: IQR); this is the difference between X_{75} and X_{25}.

Relative Spread

The coefficient of variation (symbol/acronym: CoV) is the ratio of the sample standard deviation to the sample mean (i.e., $S/\overline{X}$). This is considered to be useful (especially when making comparisons between samples) because it has no units. The units of the mean are exactly canceled by the units of the standard deviation. Cool fact: the CoV for response time is almost always right around 0.25, no matter what task the subject is doing.

Skewness

This is a measure of asymmetry. A distribution with positive skew has a long right (upper) tail; a distribution with negative skew has a long left (lower) tail. The normal distribution is symmetric, so it has a skewness value of zero. In general, if the absolute value of skewness is greater than 2.00, then the distribution will differ significantly from normal.

Kurtosis

This is a measure of "peakedness" or the extent to which observations cluster around a central point. Positive kurtosis indicates that the distribution is more clustered (and also has longer tails) than the normal distribution; negative kurtosis indicates less cluster (and shorter tails); more like a rectangle than a bell curve. As was true for skew, for a normal distribution, the value of the kurtosis statistic is zero. Warning: this is because we subtract 3.00 from the basic formula, so don't be surprised if you see the normal described as having a kurtosis of 3.00, instead, in some cases.

Getting Descriptive Stats from SPSS

There are several different ways to try to get the descriptive statistics (for a set of numerical data) out of SPSS. The most direct is to use **Analyze… Descriptive Statistics… Descriptives…** then "push over" the variable(s) that you are interested in, then click **Options** and choose the descriptives you want, then click **OK**. (You can get the same information using **Analyze… Descriptive Statistics… Frequencies…** if you then click on **Statistics**.) A more powerful set of tools is available under **Analyze… Descriptive Statistics… Explore…** which allows you to divide your data into groups (when you have one or more between-subject factors); this procedure automatically reports all of the standard descriptives and also provides some plots. My suggestion is that you always use **…Explore**.

Plotting Distributions

When plotting the distribution function for a univariate set of data, it is traditional to use a histogram when the data are discrete, and a polygon when the data are continuous. Note that some areas of research seem to employ cumulative distribution functions, instead of (plain) relative-frequency distribution functions.

When plotting the distribution of a bivariate set of data, use a scatter plot.

When plotting the distribution of a quantitative variable as a function of a qualitative variable, either make a series of small distribution functions and place them vertically (i.e., rotated 90° counter-clockwise; one for each value of the qualitative variable) or else include a series of vertical scatters (one for each value of the qualitative variable).

Larger data sets than the above examples cannot be plotted in a manner than most people can "read." You have to break them down into subsets and then use one of the methods given above for each of the subsets.

We will not be doing much plotting of samples, mostly because I think the plots that are produced by SPSS look awful.