

The Unit of Analysis: Facts About What?

O.K., if you didn't already know how, then now you know how to compute a median and a mean. Those are the centers. Corresponding to each of these concepts of the center there is an associated concept of the variation: Corresponding to the median, the hinges provide a way to specify, verbally, where the central half (not just the center) of the data lies. In words, using the median and the hinges: "The middle 50% lie between ___ (lower hinge) and ___ (the upper hinge). Corresponding to the mean, the standard deviation estimates the mean deviation: "The mean is ___ with a standard deviation of ___." The verbal summary provides the mean and the standard deviation from which the reader of the summary is supposed to construct a mental image a bell-shaped distribution with the central peak at the mean and the center of the distribution lying between values which are one standard deviation below the mean and one standard deviation above the mean.

Corresponding to each of these concepts of variation there is also a concept of *too much* variation: Corresponding to the median and the quartiles, too much variation is marked by the fences: The inner and outer fences mark the limits of routine variation. Value beyond the fences are sufficiently unusual to be suspicious — not just different from the central values of the data, but different in kind — or, at the least, that is a possibility to be investigated. Corresponding to the mean and the standard deviation, too much variation is marked by the value of the mean plus or minus two standard deviations or three standard deviations (or, sometimes, more precisely plus or minus 1.96 and 2.81 standard deviations). In either system "too much variation" designates variation so large and unexpected that the analyst may either leave it out entirely, applying the name "outlier", or do just the opposite by focusing in on these special cases as extreme examples of the general principles at work in the data (Gerber's High Protein).

In addition to learning the concepts for the center, for the variation, and for too much variation, you have taken steps toward sorting all of this out computationally, and you've begun to get Excel and the Mac to do your bidding.

With these "mechanics" under control, it is time to delve in to the art of using these things, time to think. The first question is "Facts about what?" A datum is an attribute of something: A number of grams of protein is an attribute of a cereal. A cause of death is an attribute of a person. A literacy rate is an attribute of a country. I want to draw your attention to that thing: What is this thing to which the numbers are attached or, in the language of the trade "What is the correct unit of analysis?"

The routine answer to that question is usually easy enough to answer. The deeper answer is more interesting: I asked "What *is the correct* unit of analysis?" There are choices. And there is no compelling reason why the data analyst should automatically accept the unit of measure that was convenient to the person who organized the data. That is up to us — analyst's choice. And the choice may make a difference. For example, beginning with the breakfast cereal, the original data show grams of protein as an attribute of a commonly used portion of breakfast cereal. But grams of protein could be recomputed as an attribute 100 grams of breakfast cereal — changing the unit of analysis by standardizing the data to a common weight of cereal. Analyst's choice.

To emphasize the importance of the unit of analysis and to encourage active choice of the unit of analysis on the part of the data analyst, I am going to take you through five sets of mental gymnastics. The job of these gymnastics is to create doubt, doubt with respect to passive acceptance of the data as given, and then to improve the focus of the analysis by asking "What is the 'correct' unit of analysis?"