Cornell University
Cornell Statistical Consulting Unit

StatNews #67

Censored Data
February 2005

Censoring occurs when observations have some information available for a variable but the information is not complete. This lack of information arises when a variable can be measured precisely only within a certain range. Outside of this range, the only information available is that it is greater or smaller than a specific value or that it lies between two values. Analyzing a censored variable requires procedures designed to account for the censoring. This newsletter describes censoring, some common examples in which it occurs, and some of the disadvantages of dealing with it inappropriately.

Censoring is probably most well known because of survival analysis, which studies time until an event. There are usually some individuals who do not experience the event during the study, so the time to event is incomplete for these cases. The researcher knows it is greater than the length of time these individuals were studied, though not how much greater.

But censoring occurs in other situations as well. For example, when the demand for a concert is recorded as the number of tickets sold, the data are censored when the tickets are sold out. Likewise, censoring happens with the use of equipment that has a lower or higher threshold or a lack of refinement in the measurement it provides.

There are three main types of censoring: right, left, and interval. Censoring could occur, for example, when administering a survey to mothers every other month asking if they are still breast feeding. Right censoring occurs when mothers are still breast feeding after the last survey, since we do not know exactly how long they will continue. Left censoring occurs when mothers enter the study after they have stopped breast feeding. We do not know exactly when they stopped breast feeding, although we know that it happened before their entry to the study. Interval censoring occurs if the breast feeding ended between two successive surveys since one can only say that breast feeding ended somewhere within the past two months.

Simple approaches researchers might choose to deal with censored data are to set the censored observations to missing or replace the unobserved value of the variable by zero, the minimum, maximum, mean value, or a randomly assigned value from the range of possible values. When the censoring is minimal, using one of these approaches can be reasonable. When it is not, these simple solutions can, however, cause serious bias in estimates and standard errors obtained in subsequent statistical analysis, can discard potentially important information, and can create a sample that is not representative of the population studied.

Several superior solutions exist. These will be addressed in greater detail in a future newsletter.

Author: Francoise Vermeylen fmv1@cornell.edu