Heaped Data in Count Models

Tammy Harris

Abstract

Heaped data result when subjects who recall the frequency of events prefer for reporting from a limited set of rounded responses or preferred digits over reporting exact counts. These rounded responses and digit preferences (also referred to as data coarsening) could be characterized by reported frequencies (or counts) favoring multiples of 20, reporting counts ending with 0 or 5, or a preference for reporting an even number over an odd number. This mixture of values is a type of measurement error (pattern of misreporting) that can lead to biased estimation and imprecision in discrete quantitative data. Sometimes this pattern in data can be explained or understood, but its effect on the statistical inference may be harder to anticipate. A visual representation of heaped data can be seen in a frequency distribution (histogram) where the heaps are represented as periodic peaks or spikes within the overall data layout.

We develop statistical models to model heaped count data using a mixture of likelihood functions for heaped and nonheaped count data. For the heaped count data, we consider that the reported outcome is actually censored over the half width of the heaping multiple. Simultaneously, we consider that nonheaped data follow the count distribution’s likelihood for exact counts; that they are not censored. The investigator specifies the heaping multiples over which heaped values are censored via an interval regression approach in our approach. We also create new Stata commands to model these heaped data.