User's Guide: Statistical Issues in Working with Secondary Data

A number of publicly-available datasets identified respondents using a multistage probabilistic sampling strategy. Understanding what this means is best illustrated by an example from the National Ambulatory Medical Care Survey (NAMCS). NAMCS seeks to create nationally representative estimates of visits to community-based physician offices. The most straightforward way to do this – identifying a simple random sample of office visits throughout the country – would be very difficult and expensive. Instead, NAMCS samples physicians in a series of defined geographic areas, with over-sampling of relatively underrepresented specialties, and then asks each sampled physician to collect data on a systematic random sample of patient visits over the course of a week. In and of itself, this process creates a non-random sample of patient visits. However, by adjusting for the sampling strategy – in particular, the likelihood that a given visit was sampled – the results can be extrapolated to create national estimates of patient visits.

This process of adjustment generally involves two factors. First, one must adjust for the likelihood that a given visit was sampled. This is adjusted for by the sampling weight (sometimes called patient weight). Second, one must adjust for clustering. Clustering represents the observation that there is a relationship between the care patterns and outcomes of patients who share a common physician, hospital, health care plan, or geographic area. For example, consider 3 patients sampled by a survey like NAMCS – Mr. A, Mr. B, and Mr. C. Mr. A and Mr. B both live in Nashville and are cared for by the same family physician, while Mr. C. lives in Los Angeles and sees a different doctor. They each go to their doctor for a common cold. Mr. A. is prescribed an antihistamine and acetaminophen. Because Mr. B lives in the same city and sees the same doctor as Mr. A., he is more likely to receive the same treatment for his cold than is Mr. C. This is the effect of clustering – in this case, the care that Mr. B receives is to a certain degree correlated (non-independent) with the care that Mr. A receives because they share the same physician and geographic area, which was a consequence of the sampling strategy. Because NAMCS selects its patient visits from within certain geographic areas and certain physicians, it needs to account for the lack of independence of care patterns between patients who share a common physician or area of residence. Statistical adjustment for clustering is complex, but the main practical effect is that it reduces the effective sample size of the population when evaluating tests of statistical significance between two groups. Finally, NAMCS contains a 3rd level of sampling complexity, stratification, which refers to another stage of the purposeful sampling process.

Fortunately, many datasets using complex sampling designs provide instructions to account for weight, clustering, and stratification effects to generate representative estimates. Often, dataset documentation provides sample code for common statistical programs such as SAS or Stata to correctly set parameters for adjusting for the survey design and sampling weights. That said,

there are many exceptions to these rules, and it is important to consult and collaborate with an experienced statistician to ensure that the analyses are conducted properly. Of note, statistical packages such as SAS, Stata, and SPSS are becoming increasingly sophisticated at handling complex survey designs, but in some cases more specialized programs such as SAS-callable SUDAAN are necessary for properly analyzing the data. Such complex issues should not be a cause for despair, since many of these skills can be learned (particularly in collaboration with the support of a statistician or experienced researcher). Nonetheless, it is essential to closely read the dataset documentation and seek experienced help to ensure the analytic plan is approached and executed properly.