Preventing Bad/Biased Samples

Many so-called "standard" statistical analyses that are presented and discussed in introductory statistics courses make the assumption that the data of interest are **independent and identically distributed (or "i.i.d.") observations**. As discussed in the lectures earlier this week, **simple random sampling (SRS)** is the closest probability sampling analog to i.i.d., in that the sampling mechanism used to generate the observations will produce independent and identically distributed observations. While this type of sampling will produce samples with this nice "i.i.d." statistical property, facilitating "standard" statistical analyses, SRS is seldom used when sampling from real populations. One of the reasons for this is that SRS, while producing estimates that are unbiased in nature (which recall means that the estimates based on hypothetical repeated samples using SRS will have a mean equal to the true population mean), has the potential to generate "bad" samples with substantial sampling error (where an estimate based on the sample is quite different from the population parameter of interest).

Consider, for example, a national sample of 1,000 cell phone numbers selected using SRS. While in expectation any one given sample will include a representative random sampling of numbers from area codes across the nation, **all possible random samples using SRS are equally likely.** What this means is that a simple random sample of cell phone numbers that only includes area codes from Florida is just as likely as a simple random sample of numbers that includes a representative selection across the states. Ideally, we would like to use design strategies to reduce the chances of such a "bad sample" occurring, especially if our variable of interest tends to take on very different values in the state of Florida! The major statistical problem with the simple random "Florida" sample is that any estimate that we compute after collecting data from the sample will likely be very different from the true population parameter that we are trying to estimate (especially if the variable of interest tends to take on very different values in Florida relative to the rest of the nation). Because the probability of selecting these extreme samples is equal to the probability of selecting more representative samples, the sampling distribution for simple random samples can tend to be quite variable.

A very common sampling technique used to minimize the sampling variance that can arise from these so-called "bad samples" in SRS is **stratification**. You've already been introduced to stratification in an earlier lecture. When we conduct stratified sampling, we first allocate portions of our sample to all possible divisions (or "strata") of the population of interest (e.g., states). This ensures that some sample will be selected from all of these possible divisions, and that the overall sample will therefore be representative of the target population. For example, using a technique known as *proportionate allocation*, suppose that we knew that 55% of students enrolled in a particular college were females, and 45% were males. If we wanted to draw a sample of 1,000 students from this college, we would randomly selected 550 females from a list of all females enrolled, and 450 males from a list of all males enrolled. This ensures that our entire sample of size 1,000 won't include only females!

Consider our earlier gym example as well, and the web app that allowed us to visualize sampling distributions. If we only draw our sample from one stratum of the overall population (e.g., gym goers), and the units in that stratum tend to have values on a variable of interest that *differ* from the values for the variable in other strata, then the estimates that we compute based on that sample will be biased, and will not represent the overall target population. This is an example of **selection bias;** on average, estimates computed from repeated samples of gym goers will *not* be equal to the true population parameter of interest. Stratified sampling ensures that we would select a sample of gym goers *and* a sample of non-gym goers, increasing the representativeness of our sample and potentially reducing bias.

Another nice property of stratified sampling is that it shrinks the variance of sampling distributions. In SRS, all of the variance within strata and between strata in terms of the variable of interest contributes to the overall sampling variance. In stratified sampling, when we allocate a certain number of sampled units to be selected from each stratum, we r*emove the between-stratum variance from the overall sampling variance!* This is because every hypothetical repeated sample would use the same stratified design, and the same allocation; assuming reasonable response rates, we will have representation from each of the strata where we allocated a portion of the sample. There is no uncertainty in whether we will have sampled units from a particular stratum, and there is nothing random about the allocations; these are fixed by design! The only uncertainty arises from the random sampling that occurs *within* strata from one hypothetical sample to another. Each sample will always feature random selections from the same strata; what happens within the strata will change from sample to sample.

We will revisit the idea of stratified sampling in an upcoming lecture, but you will often hear sampling statisticians say **"when in doubt, stratify."** We can use this technique to prevent bad samples, and decrease the variance of our sampling distributions.

Bad Samples Arising from Nonresponse

When analyzing data, we always have to think carefully about the process used to ultimately produce the data that we are analyzing. We may dedicate substantial resources to a carefully designed stratified sample of some population that will produce unbiased estimates by design; *however, there is no guarantee that every unit sampled will agree to provide data.* If a sampled unit refuses to provide data after being sampled, this situation is known as **unit nonresponse**. Unit nonresponse can have a particularly negative impact on the quality of a given sample when the units that ultimately agree to provide data differ significantly from the units that do not agree to provide data on the variables of interest.

For example, suppose that people with lower income tend to respond to a survey of a nationally representative sample of individuals at higher rates than people with higher income. Because the resulting sample of *respondents* to the survey request tends to feature people with lower income, any estimates related to income (which will always be computed using data from the respondents, or the units that agree to provide data!) will be subject to another form of **selection bias**, namely **nonresponse bias**. In short, nonresponse bias occurs when there is a tendency for the units in a sample that agree to provide data to be systematically different from the units in the sample that do not provide data (in terms of the variable of interest). This type of bias can also occur for estimates based on specific variables, when sampled units may agree to provide data in general, but not on specific variables. For instance, a survey respondent may agree to participate in the survey, but refuse to share their income. This type of nonresponse is known as **item nonresponse**.

Whereas stratified sampling is a design tool that can be used to reduce selection bias from a sampling perspective, the selection bias introduced by unit or item nonresponse can either be addressed during the data collection process or via post-survey adjustments to the estimates based on a respondent sample. For example, sampled units reluctant to provide data may be offered additional incentives for their participation, or offered different methods for providing their data (e.g., over the web, rather than speaking to an interviewer). Such units may also receive additional effort from a data collection organization (e.g., more follow-up contact attempts). After the survey is over, if there is still evidence that the respondent sample somehow differs systematically from the full sample, respondents who had a lower probability of responding may receive larger weight in the overall analysis. Item nonresponse may be addressed via statistical models used to predict the missing values as a function of other observed data. There are all tools designed to reduce the type of selection bias that can arise from nonresponse.

It is essential that anyone computing estimates based on samples of populations carefully evaluate the steps that were taken to minimize the potential biases arising from these "bad" samples.

Additional Deep-Dive Readings on Stratified Sampling and Nonresponse Bias

Groves, R. M., Fowler Jr, F. J., Couper, M. P., Lepkowski, J. M., Singer, E., & Tourangeau, R. (2011). *Survey Methodology, Second Edition.* John Wiley & Sons.

Heeringa, S.G., West, B.T., and Berglund, P.A. (2017). *Applied Survey Data Analysis, Second Edition.* Chapman Hall / CRC Press.

Kish, Leslie. (1965). *Survey Sampling.* Wiley.

Lohr, Sharon. (1999). *Sampling: Design and Analysis, Second Edition.* Cengage Learning.