David Youngberg
BSAD 210—Montgomery College

LECTURE 02: GOOD AND BAD SAMPLES

I. Types of samples
   a. *Simple random sample*—Each population element has the same chance of being selected (becoming an observation). This is considered the gold standard of sampling. But it is not always practical.
      i. Requires a population list which isn't always available.
      ii. Fails to use all the information about a population.
      iii. Expensive and time-consuming to implement.
   b. *Systematic sampling*—sampling every element at a given increment (e.g. screening every 10^{th} person at the airport).
      i. Easy to implement and very flexible.
      ii. Periodicity is a potential problem. If you sample a restaurant's cleanliness at an increment of seven days, you're going to be looking at the cleanliness the same day each week. But some days of the week are busier than others; your sample won't tell you how clean the restaurant is all year around.
   c. *Stratified sampling*—sampling from a subpopulation, similar to panel data (e.g. a simple random sample from each major). We try to ensure the subpopulation are the same within and different compared to others.
      i. Gives you data on subpopulations and close to simple random sample.
      ii. Sampling here is usually proportional: what you take from each subpopulation is proportional to that subpopulation's share of the general population.
      iii. A big problem can occur if you don't know the subpopulation and a portion of the total population. If there's a major difference, you can get sampling bias.
   d. *Cluster sampling*—When we divide the population into some subgroup, trying to ensure that the groups are similar to each other and different within. We then select some of these "clusters" for further study.
      i. This is very cheap to do and easy without knowing much about the population.

ii. It's hard to get a sufficient level of diversity in the subgroups (e.g. neighborhoods geographically divided into clusters for study).

II. A Good Sample Is…

a. A good sample is *precise*—it minimizes the amount of error from the population due to random fluctuations
    i. S*ampling error* is unavoidable; there is always the chance that one gathered a disproportionate number of unusual observations.
    ii. There is no way to "fix" sample error. One can only make it unlikely and the only way to do that is to add observations to your sample.
    iii. Example: Two different dentists have tried to charge my wife for work she didn't need. This excessive charging was confirmed her by a third dentist. She therefore believes that most dentists aren't trustworthy. There is no reason to believe my wife's sample of three is inaccurate—so on one level this inference makes sense—but it's also quite likely she got two dishonest dentists by chance. Her sample size of three is probably imprecise.

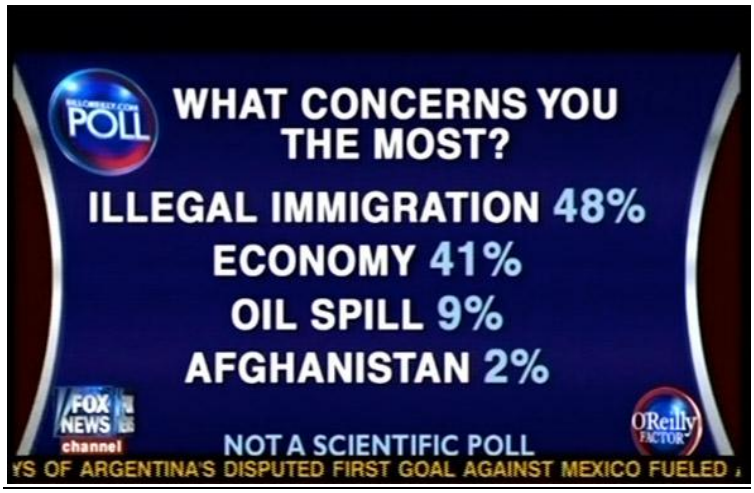
b. A good sample is *accurate*—it neither underestimates or overestimates the statistics of a population.
    i. By selecting randomly, you'll get some observations that are over the population average and some under. A good sample would make sure this natural variance cancels one-another out.
    ii. If you have *systematic variance*, then you have some issue of systematically overestimating or underestimating the population. Samples with systematic variance are *biased*.

c. A good sample is treated almost the same way (sometimes exactly the same way) as the whole population but because of the differences that can arise between them, we use different notation to describe the same characteristic, such as average ("$\bar{x}$" (x bar) for the sample and "μ" (mu) for the population).

III. Types of sampling bias

a. All biased samples contain a non-random component. This component creates systematic variance.

b. *Self-selection bias*—observations decide if they are gathered or not, resulting in a non-random element determines the sample, and thus possibly biasing the results
    i. Example: Every news outlet's online poll.



c. *Undercoverage bias*—when certain observations in the population cannot be included in the sample, excluding observations in a non-random way
    i. Example: *The Wisdom of Whores*
d. *Survivorship bias*—concentrating on observations that endured ("survived") some process, the reason for which is non-random
    i. Example: We're interested in determining how patients feel about their therapist. Because we want to make sure the patient's had a chance to get better, we want to include only the patients who been with their therapist for at least five years. But anyone who switched therapists isn't going to be included and such people are more likely to have a poor opinion of their therapist compared to those who stick around!
e. Survivorship bias and undercoverage bias are very similar. The difference is that :
    i. In undercoverage bias, segments of the population are being excluded and cannot be included by the design of the data-gathering process. You can tell from the beginning exactly which person(s)/object(s) will be excluded.
    ii. In survivorship bias, segments of the population could be included, but aren't because they didn't survive whatever criteria was set up. You can't tell from the beginning which person(s)/object(s) will be excluded.

IV. Scope of Inquiry
a. All biases come from a disconnect between what we want to figure out (information about the population) and the observations we have.
b. For example, a random sample of only Montgomery College students concerning how they think about the price of college.

i. This is a problem if we want to know what college students think about the price of college. (The population is all college students.)
ii. But this is fine if we are only interested in what Montgomery College students think about the price of college. (The population is all Montgomery College students.)

c. You need to a reason to think that the observations that are excluded will have a systemic impact on the results.
    i. Example: Online review sites. Because all reviews are voluntary; its samples could be biased because of self-selection. Perhaps people will only feel motivated to review if they had a bad time.
        1. For some things, this appears to be true. A lot of perfectly fine grocery stores have terrible reviews. But for other things, like restaurants, it doesn't seem to be the case. Perhaps because eating out is more novel than grocery shopping, people seem equally willing to review a restaurant regardless of the quality of their experience.
        2. It can go the other way, too. A lot of movies, TV shows, and books have a fair number of stars. That doesn't mean they are all good; the people who are interested in the type of show/movie/book are particularly enthusiastic about it and are thus more motivated to write a review. Those not interested will just stop reading or watching it.