David Youngberg
BSAD 210—Montgomery College

LECTURE 03: SAMPLING

I. Why sample?
   a. Look: whenever we want to figure something out, we want to know what's going on for all instances, not just a few.
      i. Pepsi doesn't care how popular a new drink for a few people. They want to know how popular it will be for everyone.
      ii. Scientists don't care that much about how a drug affects a few people. They want to know how it affects everyone.
      iii. Policy makers aren't interested if just a few criminals commit crimes after a rehabilitation program. They want to know how effective that program will be for all criminals.
   b. But checking a whole population is really hard. So we take a sample, or a subset of the population that, ideally, represents the population.
      i. By *population* we mean all possible subjects of interest. Note this can include subjects which don't exist yet (like the future recipients of a drug treatment).
      ii. A *parameter* is a characteristic about a population.
      iii. A *statistic* is a characteristic about a sample.
      iv. We care about this smaller size not because we're interested in how change affects just the sample but because the sample *represents* a larger population that we do care about.
   c. Sampling has a lot of advantages:
      i. It's cheaper;
      ii. It allows greater depth in questioning;
      iii. It's faster;
      iv. It's more practical (you have to use a sample for crash testing cars, or you'll smash all your cars and have none left to sell)

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 parameter.
   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*.

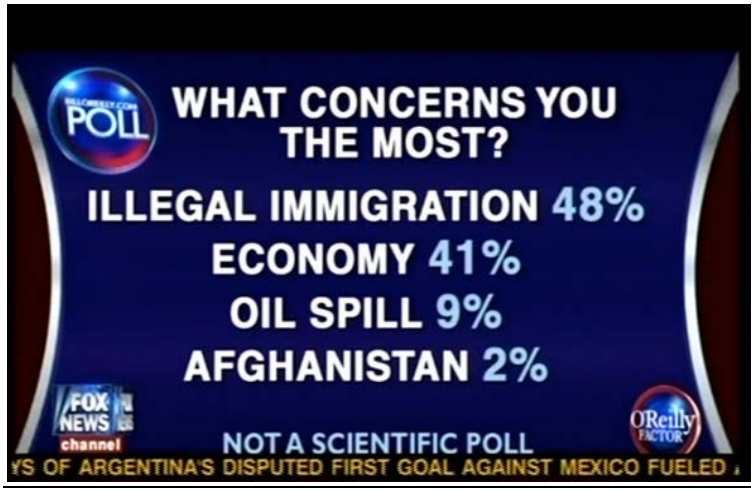
c. A good sample is treated almost the same way (sometimes exactly the same way) as the whole population but because of the differences that can arise between them, we use different notation to describe the same characteristic, such as average ("$\bar{x}$" (x bar) for the sample and "μ" (mu) for the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

d. The gold standard for a good sample is called a *simple random sample*. It means that every element in the population has an equal chance of being a part of the sample.
   i. True simple random samples are often practically impossible but thinking about what a sample ideally should be is helpful because it gives a useful benchmark to evaluate actual samples.
   ii. A good illustration of this is sampling bias, an area where samples often go bad.

III. Sampling Bias: How Good Samples Go Bad

a. All biased samples contain a non-random component. This component creates systematic variance. Here are three common kinds of sampling bias.

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. Again, this only an issue if the group that's opting out of the sample is systematically different than the group that's opting in.
   ii. (Possible) example: The polls for the 2016 election (Trump voters were possibly more likely to opt out of the polling)
   iii. Example: Every news outlet's online polls.



c. *Undercoverage bias*—when certain observations in the population cannot be included in the sample, excluding observations in a non-random way
    i. Example: Polls for 2016 election (too much emphasis on landlines)
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!
    ii. ***Not*** an example: Thanos randomly kills half of the people on a planet. He then asks all the survivors if they had a healthy breakfast or not that day and uses this data to determine what percent of the whole planet (before half of them died) had a healthy breakfast.

        Although Thanos only asked survivors, there's no survivorship bias here because the survivors were randomly determined. There is no bias; this is effectively a simple random sample.
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.