David Youngberg

Acct/Busa/Econ 222—Bethany College

**Lecture 04: Sampling**

1. Why sample?
	1. Look: whenever we want to figure something out, we want to know what’s going on for all instances, not just a few.
		1. Pepsi doesn’t care how popular a new drink for a few people. They want to know how popular it will be for everyone.
		2. Scientists don’t care that much about how a drug affects a few people. They want to know how it affects everyone.
		3. 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.
	2. But checking a whole population is really hard. So we take a sample, or a subset of the population that represents the population.
		1. Note 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.
	3. Sampling has a lot of advantages:
		1. It’s cheaper;
		2. It allows greater depth in questioning;
		3. It’s faster;
		4. It’s often 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)
2. Good and Bad Samples
	1. A good sample is *accurate*—it neither underestimates or overestimates the population
		1. 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.
		2. If you have *systematic variance*, then you have some issue of systematically overestimating or underestimating the population. Samples with systematic variance are *biased*.
		3. Example 1: If you ask your American TV audience to submit votes on what they think, you won’t get an accurate view of Americans or even your audience!



**CORRECT!**

* + 1. Example 2: *The Wisdom of Whores*
	1. A good sample is *precise*—it minimizes the amount of error from the population due to random fluctuations
		1. Once we’ve accounted for all systematic error, only *sampling error* remains.
		2. We naturally want a small error, though this is not always possible.
		3. The sampling error is the standard error of the estimate. Only use it if you’re confident the error is accurate.
1. Types of samples
	1. *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.
		1. Requires a population list which isn’t always available.
		2. Fails to use all the information about a population.
		3. Expensive and time-consuming to implement.
	2. *Systematic sampling*—sampling every element at a given increment (e.g. screening every 10th person at the airport).
		1. Easy to implement and very flexible.
		2. 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.
	3. *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.
		1. Gives you data on subpopulations and close to simple random sample.
		2. Sampling here is usually proportional: what you take from each subpopulation is proportional to that subpopulation’s share of the general population.
		3. 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.
	4. *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.
		1. This is very cheap to do and easy without knowing much about the population.
		2. It’s hard to get a sufficient level of diversity in the subgroups (e.g. neighborhoods geographically divided into clusters for study).
2. Discrete probability distributions
	1. It’s useful to think about sampling as a discrete probability distribution.
	2. A *discrete probability distribution* is when there are only possible results: “success” and “no success.” We start with binomial probability distributions.
	3. Binomial distributions assume the probability of success is constant. This either means:
		1. You are replacing each selection.
		2. Your sample size is so small compared to the population, you don’t affect the probability when you perform a trial. The threshold for “big enough” is if your sample is less than 5% of your population.
	4. Hypergeometric Distributions
		1. What happens if your sample is more than 5%? If each trial affects the likelihood of success, we need to use a different discrete probability function: hypergeometric.
	5. Poisson Distribution
		1. This type of distribution describes the number of times some event occurs during a specific interval (such as time, distance, area, volume, etc).
		2. If we know how often something occurs on average, we can use Poisson to figure out how often something other than the average occurs.
3. Continuous Probability Distribution
	1. Where a discrete distribution has “chunky” values (1,2,3,4,etc.), a continuous distribution has infinitely divisible values. For a limited range (e.g. 1 through 6), there are limited values (e.g. 6) for a discrete distribution. For continuous distribution, there are an infinite number of values in a limited range.
	2. *Uniform distribution*—we’ve discussed this one before: each outcome has the same likelihood of appearing. Its distribution has a rectangular shape.
		1. Its mean is $\overbar{x}=\frac{Max+Min}{2}$
		2. Its standard deviation is $σ=\sqrt{\frac{\left(Max-Min\right)^{2}}{12}}$
	3. *Standard Normal Distribution*—there are many types of normal distribution, but “standard” distribution has a mean of zero and a standard deviation of one. Like all normal distributions, it is a bell curve.
		1. Lots of things follow a normal distribution so it would be nice to have a systematic way to evaluate them.
		2. We can convert any normal distribution into a standard normal using a z-score.

$$z=\frac{x\_{i}-\overbar{x}}{σ}$$

* + - 1. Where *xi* is the value in question.
		1. The resulting z-score tells us the standard deviation from the mean of the value in question.
		2. It’s also used to tell us the probability that we will get that value. (The latter exercise requires a table of references.)
		3. For example, if people make an average of $40,000 a year and the standard deviation is $4,000, how many standard deviations is $50,000 from the mean?

$$z=\frac{\$50,000-\$40,000}{\$4,000}=2.5$$

* + 1. $50,000 is 2.5 standard deviations above the mean. (If it was a negative number, it would be below the mean.)
1. Proposal
	1. Go to my website and download the Excel data
		1. This data is taken from [www.gapminder.org](http://www.gapminder.org), a website which also provides scatterplots
	2. Of these variables, select one variable you are curious why it is what it is. This should be a “big think” variable, one where there are many possible reasons for its value. This is your dependent variable.
		1. GDP per capita, HDI, and suicide rates are good examples, but not the only ones.
		2. Then select three possible variables which could *cause* that one.
			1. These variables should not be obvious; it is not worth arguing why electrical consumption causes electrical generation or vice versa.
			2. These variables should not be inter-related. Choosing BMI (female) is functionally the same thing as choosing BMI (male). Different is better.
	3. For Friday, write a paragraph describing how these three variables cause the dependent variable. This does not have to be an in-depth analysis, just a logical thought process for why, if you change one of these three variables, your dependent variable also changes.

**Lab Section**

1. Chapter 7
2. Homework
	1. Chapter 7: 1-3,4,6; page 88-89