POLS 5377 Scope & Method of Political Science

Week 10 Inferential Statistics

Sampling and the Sampling Distribution

Healey. (2016) *Statistics: A Tool for Social Research,* Chapter 6

2

Key Questions:

- What is the inferential statistics?
- what are the sampling methods we use to conduct inferential statistics?
- What are the differences between the sampling distribution, the sample, and the population?
- How do we use sampling distribution to link the sample and the population?

3

Outline

- Inferential Statistics
- Probability Sampling
- The Sampling Distribution
- Central Limit Theorem

4

Inferential Statistics

- Descriptive Statistics: use of statistical techniques to describe and summarize data
- Inferential Statistics: use of statistical techniques to generalize from a sample to a population
- Researchers in the social sciences often try to deal with the problems that associate with a large population, such as:
  - the public's attitudes toward certain policy issues
  - levels of well-being of citizens
- It's almost impossible or too expensive to conduct studies to collect information from each individual.
- The common solution: use the samples to estimate the population

5

Inferential Statistics

- The goal of inferential statistics: to generalize the characteristics (parameters) of a population based on what is learned from the samples.
- There are two common applications for inferential statistics:
  - Estimation procedures (Lecture Week 10)
  - Hypothesis testing (Lecture Week 11 & 12)
- Example:
  - The City Council of Houston would like to know how its citizens feel about its public transportation system.
  - However, it is impossible for the city government to ask every single citizen of Houston how they feel about its public transportation. It will take forever and is very pricy.

6

Inferential Statistics

- What do we do?
- We randomly select a subset of people (say 100, 200, or 500) who live in Houston, and ask them how they feel about Houston's public transportation.
- From this sample of 100, 200, or 500 people we infer what the people of Houston think.
- But, how to choose these 100, 200 or 500 people from the 2 million citizens?
  - Do you simply go to the downtown of Houston and grab 100 people on the street? NO!!
  - If you remember our lecture in Week 3, there are different types of sampling method: probability sampling and non-probability sampling.

7

Probability Sampling

- Basic concepts and terms
  - A *population* is the total set of items that we are concerned with. In our example, the population is all the people who live in Houston.
  - A *sample* is a subset of a population. In our example the sample would be the 100, 200, or 500 people from Houston we interview.
  - Sampling is a process of selecting observations from a population.
- To ensure that a sample will accurately reflect the population from which it was drawn, we need to make sure that all samples were selected randomly.
- Be more precise, only *probability samples* can be used for inferential statistics.

8

Probability Sampling

- Probability Sampling
  - Each member of the population has a known probability of being selected in the sample. The distribution of samples is more likely to represent he distribution of the population.
- If a sample is not random, we have committed **sampling bias**. This means our sample does not accurately reflect the whole population.
- In the example of studying the citizens' attitude toward its public transit system:
  - if I stand on the corner of Houston downtown, and ask the first 100, 200 or 500 people that I meet
  - Will my sample be randomly (equal probability) selected? NO
  - Will my sample be biased? YES

9

Probability Sampling

- Why is my sample biased? There are several reasons:
  - not everyone I stop will be a Houston resident. Considering its downtown, many of them might be tourists.
  - not every resident of Houston is going to have an equal opportunity to be included.
- With samples, our goal is to select cases so that our sample is representative of the population as a whole.
  - if we know that 60% of Houston's population is female, our sample should contain approximately the same proportion of females.
- Fundamentally, to maximize the chance of obtaining a representative sample, we should follow the principle of EPSEM (the Equal Probability of Selection Method).
- The principle is that every member of a population must have an equal probability of being selected for the sample.

10

Probability Sampling

- EPSEM sampling techniques:
  - Simple random sampling
  - Systematic random sampling
  - Stratified random sampling
  - Cluster sampling

  (Refer to Lecture Week 3-3 or textbook Healey Chapter 6 for the details of each sampling method)
- In sum, we want to avoid *biased sample*. Only unbiased sample can be used in inferential statistics.
- With biased sample, we can describe the sample distribution, but we cannot conduct inferential statistical analysis to estimate the population.

11

Sampling Distribution

- Let's say we use one of the Equal Probability Selection Methods, and have an unbiased sample. Will our sample accurately reflect the population? No!
- *Important*: just because a researcher follows EPSEM religiously, does not always mean a sample will be representative.
- Despite our best efforts, we always assume that there is some error in our sample that occurs by chance.
- The sampling error doesn't mean the researcher made a mistake. Every time you draw a random sample, you always have the possibility of sampling error.

12

Sampling Distribution

- Example: toss a coin in the air ten times
- It will generate random outcomes. In theory, we will get a head five times, and a tail five times.

- Try it yourself. Did you get five heads and five tails?
  - I got a head six times and a tail four times.
  - Is it possible that I could toss the coin the times and get heads for ten times? Sure it is possible.
- Just because the sample is properly selected does not eliminate the possibility that it does not accurately reflect the population.

13

Sampling Distribution

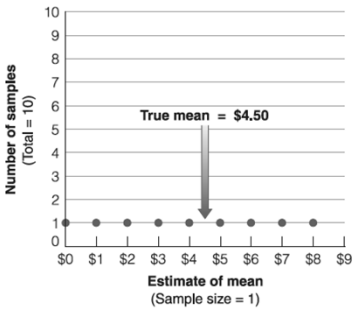
- Example:

  There are 10 students in the room. Each student has different amount of money in their pocket, from 0 to 9 dollars. Assume, a guest walks in, and randomly sample one student from this room to estimate the average amount of money the 10 students have in their pocket, based on the sample. What are the possible outcome the guest may get?

14

Sampling Distribution

- We know the true average amount of money is $4.5.
- Since each student has the equal probability to be selected, the guest can get any one of students with amount of money from 0 to 9.
  - Based on the student the guest selects, the estimated average amount of money could be from 0 to 9.
  - Even though the process is random, the guest still have the probability to make a wrong estimation about the population.



15

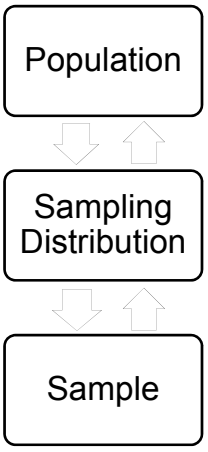
Sampling Distribution

- The diagram in the last slide, you see there is a distribution of all possible sampling outcomes.
- Sampling distribution is most important concept in inferential statistics.
- **Definition**: The theoretical, probabilistic distribution of a statistic for all possible samples of a given size (N)
- Sampling distribution
  - is a theoretical concept
  - is a distribution of the mean of all possible sampling results.
  - is related to the size of the sample.

16

Sampling Distribution

- Every application of inferential statistics involves 3 different distributions
  - Population: empirical; unknown
  - Sampling Distribution: theoretical; known
  - Sample: empirical; known
- Information from the sample is linked to the population via the sampling distribution

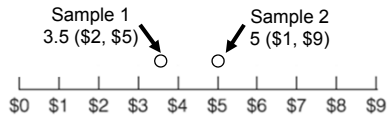


17

Sampling Distribution

Example:

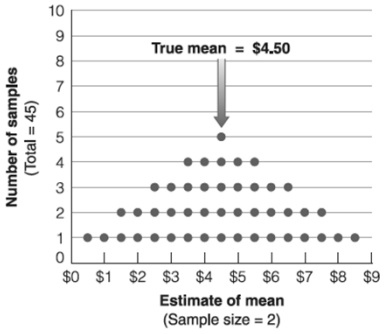
- 10 students with 0 to 9 dollars in pocket. We sample 2 students from the 10 students.
- We select a first sample of 2 students and plot its mean, 3.5 (one has $2, another one has $5), then replace the selected students back into the population
- Then we select a second sample of 2 students and plot that sample's mean, 5 (one has $1, another one has $9), again replacing the students in the second sample back into the population



18

Sampling Distribution

- We repeat this procedure (sampling and replacing) until we have exhausted every possible different combination of 2 students from the population of 10. There should be 45 possible samples. The sampling distribution looks like this:



- The sampling distribution will be:
  - Normal in shape
  - Has a mean equal to the population mean (μ)
  - Has a standard deviation (standard error, σ) equal to the population standard deviation divided by the square root of N

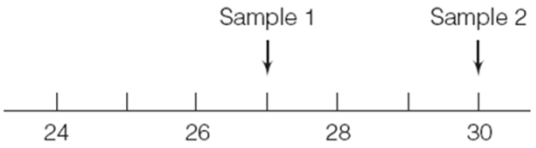
$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{N}}$$

19

Sampling Distribution

The example in the textbook:

- Suppose we want to gather information on the age of a community of 10,000 individuals
- We select a first sample of 100 people and plot its mean, 27, then replace the people in the sample back into the population
- Then we select a second sample of 100 people and plot that sample's mean, 30, again replacing the people in the second sample back into the population



20

Sampling Distribution

- We repeat this procedure (sampling and replacing) until we have exhausted every possible different combination of 100 people from the population of 10,000
- Eventually, the sampling distribution will be
  - Normal in shape
  - Has a mean equal to the population mean (μ)
  - Has a standard deviation (standard error, σ) equal to the population standard deviation divided by the square root of N

$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{N}}$$

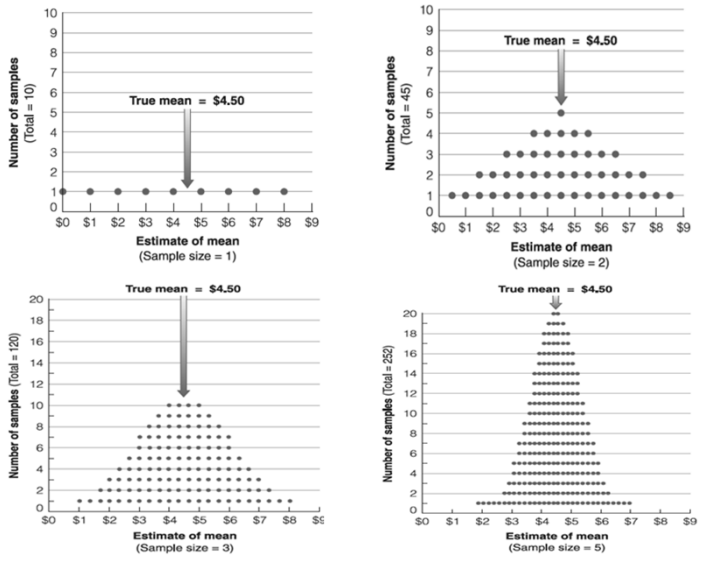
21

Central Limit Theorem

- If repeated random samples of size N are drawn from any population with mean μ and standard deviation σ, then, as N becomes large, the *sampling distribution* of sample means will approach normality, with a mean μ and standard deviation of $\frac{\sigma}{\sqrt{N}}$
- For any trait or variable, even those that are not normally distributed in the population, as sample size grows larger, the sampling distribution of sample means will become normal in shape

22

Central Limit Theorem



23

Central Limit Theorem

- The importance of the Central Limit Theorem is that it indicates that even the population is not normally distributed, the sampling distribution is still have a **normal distribution**.
- It applies to large samples (N≥100), but if the sample is small (N<100) we must have information on the normality of the population before we can assume the sampling distribution is normal

24

Implication of Sampling Distribution

- The sampling distribution is normal so we can use the Normal Curve Table to find areas
- We do not know the value of the population mean (μ) but the mean of the sampling distribution is the same value as μ
- We do not know the value of the population standard deviation (σ) but the standard deviation of the sampling distribution is equal to σ divided by the square root of N ($\frac{\sigma}{\sqrt{N}}$)

25

Three Distributions

	Shape	Central Tendency	Dispersion
Sample	Varies	$\bar{X}$	S
Sampling Distribution	Normal	$\mu_{\bar{x}} = \mu$	$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{N}}$
Population	Varies	μ	σ

26

After this lecture:

You should learn the following key concepts:

- The purpose of inferential statistics
- Need to EPSEM to sample to have unbiased sample for inferential statistics
- Even an unbiased sample may not represent the population by chance
- Understand the meaning and characteristics of sampling distribution
- Use the sampling distribution to link the sample and the population
- Understand the Central Limit Theorem in inferential statistics