Recitation, Week 7

Ye Wang

New York University

POL-850

Spring 2018

Outline

1. Review: Midterm exam (common mistakes)
2. Review: Sampling distribution and CLT

Mid-term exam score descriptive statistics (Stata Exercise)

Mid-term exam score descriptive statistics (Stata Exercise)

- There are 6 free points
- More difficult than previous years (so don't be discouraged!)
- Probably stricter grading rules than other courses

Mid-term exam score descriptive statistics (Stata Exercise)

- There are 6 free points
- More difficult than previous years (so don't be discouraged!)
- Probably stricter grading rules than other courses

- Mean: 33.08/50
- Median: 34/50
- Min: 21/50
- Max: 42/50

To whom claim?

- Feel free to ask any questions to us
- 1 (True/False) and 2-5 (multiple choice): Jason
- 6 (short answer): Dongil
- 7 (short answer) and 8 (STATA output): Ye

- Remark: for 2-5 (multiple choice), you lose one point for each incorrect answer (i.e. wrong answer chosen or correct answer not chosen) from the total given for the question

Some common mistakes

In Question 6 (short answer)

- In 6. (b), researchers distributed a small bag of rice to villagers on a first-come-first-serve basis in exchange for using an early voting ballot in order to prevent vote-buying before the election. What is the problem with this design in terms of causal inference?

In Question 7 (short answer)

- Many people forgot to specify the unit of the answer

In Question 8 (STATA output)

- There is one really tricky statement about the box graph
- We cannot know for sure that there are *exactly two* outliers.

Review: Sampling distribution

You should be able to answer these questions:

1. What's the sampling distribution of the sample mean?

Review: Sampling distribution

You should be able to answer these questions:

1. What's the sampling distribution of the sample mean?
   A **probability distribution of sample mean** $\overline{x}$ obtained through a large number of samples drawn from a specific population

Review: Sampling distribution

You should be able to answer these questions:

1. What's the sampling distribution of the sample mean?
   A **probability distribution of sample mean** $\overline{x}$ obtained through a large number of samples drawn from a specific population
2. What's the standard error of the sample mean?

Review: Sampling distribution

You should be able to answer these questions:

1. What's the sampling distribution of the sample mean?
   A **probability distribution of sample mean** $\overline{x}$ obtained through a large number of samples drawn from a specific population
2. What's the standard error of the sample mean?
   The **standard deviation** of its sampling distribution
3. How can we calculate the standard error?

Review: Sampling distribution

You should be able to answer these questions:

1. What's the sampling distribution of the sample mean?
   A **probability distribution of sample mean** $\overline{x}$ obtained through a large number of samples drawn from a specific population
2. What's the standard error of the sample mean?
   The **standard deviation** of its sampling distribution
3. How can we calculate the standard error?

$$SE = \frac{\sigma}{\sqrt{n}} = \frac{\text{Population standard deviation}}{\text{Square root of the sample size}}$$

Review: Central Limit Theorem

What's the Central Limit Theorem (CLT)?

Review: Central Limit Theorem

What's the Central Limit Theorem (CLT)?

Hint: three key words are

1. the sampling distribution of the sample mean
2. large sample size
3. normal distribution

Review: Central Limit Theorem

Answer: If we generate a reasonably large number of samples (large n, typically greater than 30) from any population, the **sampling distribution of the sample mean** is normally distributed

i.e. CLT implies, with large n, $\overline{X} \sim N(\mu, SE)$: the mean of $\overline{X}$ is μ (population mean) and the standard deviation of $\overline{X}$ is SE.

Review: Central Limit Theorem

Why is **CLT** one of the important results in statistical inference?

Review: Central Limit Theorem

Why is **CLT** one of the important results in statistical inference?

1. Because we know which distribution the sampling distribution takes, we can make statements about how close our sample mean ($\overline{X}$) is to our population mean (μ: true but unknown).

Review: Central Limit Theorem

Why is **CLT** one of the important results in statistical inference?

1. Because we know which distribution the sampling distribution takes, we can make statements about how close our sample mean ($\overline{X}$) is to our population mean (μ: true but unknown).
2. Whichever distribution the population has, the **sampling distribution** is always a normal distribution given large n.

Individual Exercise 1

The engines made by Ford for speedboats have an average power of 220 horsepower (HP) and standard deviation of 15 HP.

A potential buyer intends to take a sample of thirty six engines and will not place an order if the sample mean is less than 215 HP. What is the probability that the buyer will not place an order?

Hint: We want to find $P(\overline{X} < 215)$.

Source: https://onlinecourses.science.psu.edu/stat500/node/27

Individual Exercise 1

Note that the sample size is large enough: $n = 36$ **and we can use the CLT!** (typically n must be greater than 30 to apply the CLT).

By CLT, we can conclude that $\overline{X}$ has a normal distribution with mean 220 HP and its standard error is $SE = \frac{\sigma}{\sqrt{n}} = \frac{15}{\sqrt{36}} = 2.5HP$.

i.e. we know tha $\overline{X} \sim N(220, 2.5^2) \rightarrow$ Standardization!

$$\begin{aligned} P[\overline{X} < 215] &= P\Big[\frac{\overline{X} - \mu}{SE} < \frac{215 - 220}{2.5}\Big] \\ &= P[Z < -2] \\ &= 0.0228 = 2.28\% \end{aligned}$$

Individual Exercise 2

Average household income in America in 2018 is known to be \$61,500. Its standard deviation is \$15,000. A POL-850 student

takes a sample of 100 households and wants to calculate the probability that its sample mean is greater than \$63,500. What's the probability?

Individual Exercise 2

Again the sample size is large enough to use the CLT. CLT implies

that $\overline{X}$ has a normal distribution with its mean \$61,500 and its standard error $SE = \frac{\sigma}{\sqrt{n}} = \frac{15000}{\sqrt{100}} = \frac{15000}{10} = 1500$ (i.e. \$1500) i.e.

we know tha $\overline{X} \sim N(61500, 1500^2) \rightarrow$ Standardization!

$$\begin{aligned} P[\overline{X} > 63500] &= P\Big[\frac{\overline{X} - \mu}{SE} > \frac{63500 - 61500}{1500}\Big] \\ &= P[Z > 1.33] \\ &= 0.0918 = 9.18\% \end{aligned}$$