UC San Diego

Econometrics 120A: Discussion Section

Week 7

Natalia Madrid & Lapo Bini

Department of Economics

Chapter 5: Sampling

Chapter 5: Sampling

Central Limit Theorem (CLT)

If $X_1, X_2, \ldots, X_n$ are independent random variables with mean μ and variance σ^2, then the sample mean $\bar{X}$ is approximately normally distributed for large n:

$$\bar{X} \overset{a}{\sim} N\left(\mu, \frac{\sigma^2}{n}\right)$$

This means that as n increases, the distribution of the sample mean approaches a normal distribution with mean μ and variance $\frac{\sigma^2}{n}$.

Law of Large Numbers

If $X_1, X_2, \ldots, X_n$ are independent and identically distributed random variables with expected value μ, then:

$$\lim_{n \to \infty} P\left(\mu - \epsilon < \bar{X} < \mu + \epsilon\right) = 1$$

This indicates that as the sample size n approaches infinity, the probability that the sample mean $\bar{X}$ falls within any arbitrary distance ϵ of the population mean μ approaches 1.

Kind of PS4 Question 3

A study suggests that college students spend an average of 4 hours per day on social media platforms, with a standard deviation of 1.5 hours, based on a sample of 50 students.

(a) What is the probability that, given the true average time spent on social media is 5 hours per day, a randomly chosen student spends less than 3 hours on social media in a day?

(b) A blogger writes an article questioning the accuracy of the study, stating that despite the reported average of 4 hours, they found a student who only spends 1 hour per day on social media. They use this example to argue that the average reported is misleading. What is the probability that, if the true average time spent was actually 3 hours per day, the sample average might still be reported as 4 hours just by chance?

(c) Does the example provided by the blogger (a student spending only 1 hour) give useful information about the overall accuracy of the study's reported average? Why or why not?

Kind of PS4 Question 3

A study suggests that college students spend an average of 4 hours per day on social media platforms, with a standard deviation of 1.5 hours, based on a sample of 50 students.

(a) What is the probability that, given the true average time spent on social media is 5 hours per day, a randomly chosen student spends less than 3 hours on social media in a day?

Kind of PS4 Question 3

A study suggests that college students spend an average of 4 hours per day on social media platforms, with a standard deviation of 1.5 hours, based on a sample of 50 students.

(b) A blogger writes an article questioning the accuracy of the study, stating that despite the reported average of 4 hours, they found a student who only spends 1 hour per day on social media. They use this example to argue that the average reported is misleading. What is the probability that, if the true average time spent was actually 3 hours per day, the sample average might still be reported as 4 hours just by chance?

Kind of PS4 Question 3

A study suggests that college students spend an average of 4 hours per day on social media platforms, with a standard deviation of 1.5 hours, based on a sample of 50 students.

(c) Does the example provided by the blogger (a student spending only 1 hour) give useful information about the overall accuracy of the study's reported average? Why or why not?

Kind of PS4 Question 4

You are planning to buy an electric bike for commuting to the university next year. You currently have $2000 saved and are considering investing this money to reach your goal of $2500. You can choose between different investment options with varying levels of risk. Assume you have two options:

Option 1: Invest the entire $2000 in a new startup company that promises high returns but is risky. The expected annual return is 12%, with a standard deviation of 15%.

Option 2: Diversify by investing equal amounts in a mix of 20 well-established companies. Each company has an average annual return of 8%, with a standard deviation of 5%. The returns across companies are independent.

(a) If you choose Option 1, what is the probability that you will have less than $2000 at the end of the year? Assume that the returns are normally distributed.

(b) If you choose Option 2 and invest equally in the 20 companies, what is the probability that your total investment will result in less than $2000 at the end of the year?

(c) For Option 2, do you need to assume that the returns are normally distributed? Why or why not?

Kind of PS4 Question 4

Option 1: Invest the entire $2000 in a new startup company that promises high returns but is risky. The expected annual return is 12%, with a standard deviation of 15%.

(a) If you choose Option 1, what is the probability that you will have less than $2000 at the end of the year? Assume that the returns are normally distributed.

Kind of PS4 Question 4

Option 2: Diversify by investing equal amounts in a mix of 20 well-established companies. Each company has an average annual return of 8%, with a standard deviation of 5%. The returns across companies are independent.

(b) If you choose Option 2 and invest equally in the 20 companies, what is the probability that your total investment will result in less than $2000 at the end of the year?

Kind of PS4 Question 4

Option 2: Diversify by investing equal amounts in a mix of 20 well-established companies. Each company has an average annual return of 8%, with a standard deviation of 5%. The returns across companies are independent.

(c) For Option 2, do you need to assume that the returns are normally distributed? Why or why not?

Chapter 6: Data Generation

Questioning The Assumptions

⇨ Main assumption of chapter five: $\{X_1, \ldots, X_n\}$ VSRS with $X_i \sim \text{iid}(\mu, \sigma^2)$. What does IID mean? How can we compose a joint PDF?

$p_{X_1,\ldots,X_n}(x_1, \ldots, x_n) =$

⇨ All the asymptotic results from chapter 5 based on VSRS assumption:

CLT:

WLLN:

What If Our Assumptions Fail?

⇨ Internal validity: Does the study measure what we aim to measure?

⇨ External validity: in most cases, the target population cannot be sampled, but a subpopulation can. Is the subpopulation representative of the population? Does the study extend to the larger population?

When IID assumption is valid?

Consider the following scenario (**similar to Question 6 - PS4**): Researchers conducted a study using medical records from a sample of children admitted to UC San Diego Jacobs Medical Center with suspected flu. The study focuses on analyzing three key metrics: the mean respiration rate, the average glucose level, and the average heart rate for patients with the seasonal flu.

(a) The sample was collected between January and May 2009 and consists of **30 observations**. Do you believe the assumption of a simple random sample is reasonable in this context?

- ⇨ Independent?
- ⇨ Identically Distributed?
- ⇨ Can we extrapolate?

When IID assumption is valid?

Consider the following scenario (**similar to Question 6 - PS4**): Researchers conducted a study using medical records from a sample of children admitted to UC San Diego Jacobs Medical Center with suspected flu. The study focuses on analyzing three key metrics: the mean respiration rate, the average glucose level, and the average heart rate for patients with the seasonal flu.

(b) The sample was collected between January and May 2009 and consists of **30000 observations**. Do you believe the assumption of a simple random sample is reasonable in this context?

- ⇨ Independent?
- ⇨ Identically Distributed?
- ⇨ Can we extrapolate?

When IID assumption is valid?

Consider the following scenario (**similar to Question 6 - PS4**): Researchers conducted a study using medical records from a sample of children admitted to UC San Diego Jacobs Medical Center with suspected flu. The study focuses on analyzing three key metrics: the mean respiration rate, the average glucose level, and the average heart rate for patients with the seasonal flu.

(c) The sample was collected between January and May 2009 and consists of **30000 observations**. It is important to note that all children in the sample **tested positive for seasonal flu** prior to their inclusion in the sample.

- ⇨ Independent?
- ⇨ Identically Distributed?
- ⇨ Can we extrapolate?

When IID assumption is valid?

Consider the following scenario (**similar to Question 6 - PS4**): Researchers conducted a study using medical records from a sample of children admitted to UC San Diego Jacobs Medical Center with suspected flu. The study focuses on analyzing three key metrics: the mean respiration rate, the average glucose level, and the average heart rate for patients with the seasonal flu.

(d) The sample was collected between January and May 2009 and consists of **30000 observations**. It is important to note that all children in the sample **tested positive for seasonal flu** prior to their inclusion in the sample. Moreover, the sample is constructed using data from **all the medical centers** in San Diego.

- ⇨ Independent?
- ⇨ Identically Distributed?
- ⇨ Can we extrapolate?

Kind of PS4 Question 7

A member of the Economics department wants to assess student appreciation for ECON120A. Two potential sources of information are available:

(a) 30 reviews submitted by concerned students after final grades were released.

(b) Evaluations from a randomized focus group of 15 students.

What is the main issue likely to arise when using the first source of information?

⇨ Are we measuring a conditional or a marginal distribution?

Kind of PS4 Question 7

A member of the Economics department wants to assess student appreciation for ECON120A. Two potential sources of information are available:

(a) 30 reviews submitted by concerned students after final grades were released.

(b) Evaluations from a randomized focus group of 15 students.

What is the main issue likely to arise when using the second source of information?

Midterm Review

Question 4

A Pew poll found that 67% of adults have paid to download music. The sample had 1000 respondents and was random amongst all adults using the internet. Suppose that the real proportion is 65

(a) What is the chance of seeing a poll with between 63% and 67%?

Question 4

A Pew poll found that 67% of adults have paid to download music. The sample had 1000 respondents and was random amongst all adults using the internet. Suppose that the real proportion is 65

(b) Is it likely that the truth is 65% when we see an outcome of the poll as in the Pew poll?

Question 4

A Pew poll found that 67% of adults have paid to download music. The sample had 1000 respondents and was random amongst all adults using the internet. Suppose that the real proportion is 65

(c) If we obtained the same 67% with 2000 observations, how would you answer (b)?