WESTMINSTER
International University in Tashkent
An Accredited Institution of the University of Westminster (UK)

FUNDAMENTALS OF STATISTICS (4ECON013C)

TUTORIAL 5

Sampling methods and sampling distribution

Learning outcomes:

- Differentiate between probability and non-probability sampling methods.
- Understand the types of probability sampling methods.
- Define and describe the concept of a sampling distribution.
- Understand Central Limit Theorem (CLT) and apply it to calculate probabilities.

Non-probability vs probability sampling

In **non-probability sampling**, not every individual in the population has a known or equal chance of being selected. Selection is often based on convenience, judgment, or availability**,** rather than randomization. Because of this, results cannot be confidently generalized to the population — there's a higher risk of **sampling bias**.

In **probability sampling**, every member of the population has a known and non-zero chance of being selected in the sample. This means selection is based on randomization**,** which helps minimize bias and allows researchers to make valid **statistical inferences** about the population.

Key Characteristics of probability sampling:

- Random selection;
- Each element has a known probability of inclusion;
- Results can be generalized to the population;
- Sampling error can be measured.

Types of probability sampling methods:

1. **Simple Random Sampling** – every individual has an equal chance of selection (e.g., drawing names from a hat).
2. **Systematic Sampling** – selecting every *kth* element from a list after a random start.
3. **Stratified Sampling** – the population is divided into subgroups (strata) such as gender or region, and samples are taken proportionally from each.
4. **Cluster Sampling** – dividing the population into clusters (e.g., schools, districts), randomly selecting some clusters, and including all members within them.

Sampling distribution

A **sampling distribution** is the **probability distribution of a sample statistic** (like the sample mean, proportion, or variance) obtained from **many random samples** taken from the same population.

Example 5.1

Imagine a population of 100,000 students, and you want to know their **average height**. The true population mean (μ) might be 170 cm, but since you can't measure everyone, you take a sample.

- If you take one random sample of 50 students, you'll get a sample mean ($\bar{x}_1$).
- If you take **another sample** of 50 students, you'll get a slightly different mean ($\bar{x}_2$).
- If you repeat this process many times, you'll get many sample means ($\bar{x}_1, \bar{x}_2, \bar{x}_3, \ldots, \bar{x}_n$).

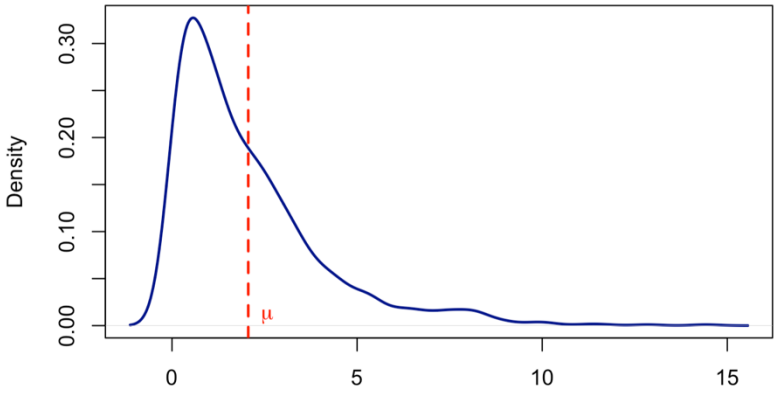
Now, if you plot all these sample means on a graph, you'll get the sampling distribution of the sample mean**.**

Central Limit Theorem (CLT)

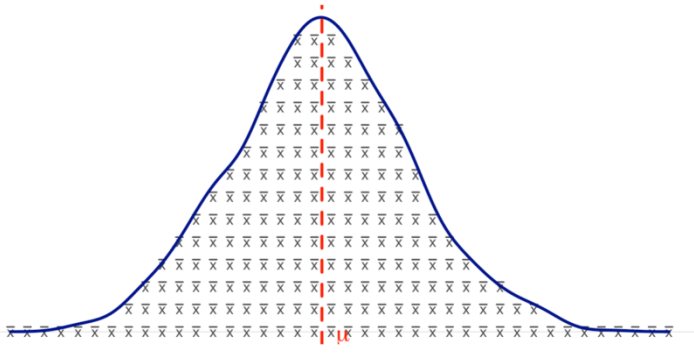
When we take many random samples of a sufficiently large size (usually $n \geq 30$) from any population (no matter its original shape), the **sampling distribution of the sample mean** will be **approximately normal** (bell-shaped).
The mean of this sampling distribution equals the population mean (μ), and its spread (standard error) equals the population standard deviation (σ) divided by $\sqrt{n}$.

Sampling distribution of $\bar{X} \sim N(\mu, \frac{\sigma}{\sqrt{n}})$

The density curve of the population values



Sampling Distribution of Sample Means



Watch the following link to understand CLT:
https://www.youtube.com/watch?v=UCmPmkHqHXk

Example 5.2

A factory produces light bulbs whose lifetimes are not normally distributed — they are *right-skewed* with a mean lifetime of 800 hours and a standard deviation of 120 hours.

a. Explain what will happen to the *shape* of the sampling distribution of the sample mean if we take random samples of size:

   i. n=5

   ii. n=50

b. Suppose we take a random sample with a size of 50. What is the probability that sample mean lifetime will be lower than 830 hours?

c. A merchant purchases boxes that contain 50 light bulbs each. Let the lifetime of individual bulbs have unknown standard deviation σ. The merchant wants the probability that a box's **sample mean** lifetime exceeds 780 hours to be at least 0.90. Assuming sampling is random and independent, what is the **largest** value of σ that meets this requirement?

Solution.

a. i. When n = 5, the sampling distribution of $\bar{X}$ will still show noticeable right skew (it will resemble the population shape). With such a small sample size the CLT has limited smoothing effect.

   ii. When n = 50, the sampling distribution will be very close to normal and noticeably narrower (less spread) than for smaller n.

b. $P(\bar{X} < 830) = P(z < \frac{830-800}{\frac{120}{\sqrt{50}}}) = P(z < 1.77) = 0.9616$

c. $P(\bar{X} > 780) = P(z > \frac{780-800}{\frac{\sigma}{\sqrt{50}}}) = P(z > -\frac{20\sqrt{50}}{\sigma}) = P(z < \frac{20\sqrt{50}}{\sigma}) \approx P(z < \frac{141.42}{\sigma}) = 0.90$

$\frac{141.42}{\sigma} = 1.28 \Rightarrow \sigma = \mathbf{110.5}$

TASKS

1. Multiple choice questions. Explain your choice.

1.1. Which of the following sampling methods would be appropriate when the population can conveniently be subdivided into relatively small and geographically compact units?
A) simple random sampling
B) cluster sampling
C) stratified sampling
D) systematic sampling

1.2. In which of the following sampling methods are the individuals of a population subdivided into mutually exclusive and collectively exhaustive separate subpopulations with a common characteristic?
A) cluster sampling
B) simple random sampling
C) stratified sampling
D) systematic sampling

1.3. Which of the following is an example of a nonprobability sampling technique?
A) simple random sampling
B) stratified random sampling
C) cluster sampling
D) quota sampling

1.4. An apartment complex manager randomly selects 10 buildings from the complex's 30 buildings, and then interviews one household member from each apartment in the 10 buildings. This is an example of _____.
A) cluster sampling
B) simple random sampling
C) stratified sampling
D) systematic sampling

1.5. In a study of stroke outcomes, we may group the population by gender, to ensure equal representation of men and women. The study sample is then obtained by taking equal sample sizes from each group. This is an example of _____.
A) cluster sampling
B) simple random sampling
C) stratified sampling
D) systematic sampling

1.6 The selection of 200 people to serve as potential jurors in a medical malpractice trial is conducted by assigning a number to each of 140,000 registered voters in the county. A computer software program is used to randomly select 200 numbers from the numbers 1 to 140,000. The people having these 200 numbers are sent a postcard notifying them of their selection for jury duty. This is an example for

A) cluster sampling
B) simple random sampling

C) stratified sampling
D) systematic sampling

1.7 Suppose you are selecting microchips from a production line for inspection for bent probes. As the chips proceed past the inspection point, every 100th chip is selected for inspection. This is an example for
A) cluster sampling
B) simple random sampling
C) stratified sampling
D) systematic sampling

1.8 The Internal Revenue Service wants to estimate the amount of personal deductions taxpayers made based on the type of deduction: home office, state income tax, property taxes, property losses, and charitable contributions. The amount claimed in each of these categories varies greatly depending on the adjusted gross income of the taxpayer. Therefore, a simple random sample would not be an efficient design. The IRS decides to divide taxpayers into five groups based on their adjusted gross incomes and then takes a simple random sample of taxpayers from each of the five groups. This is an example for
A) cluster sampling
B) simple random sampling
C) stratified sampling
D) systematic sampling

1.9 The USDA inspects produce for E. coli contamination. As trucks carrying produce cross the border, the truck is stopped for inspection. A random sample of five containers is selected for inspection from the hundreds of containers on the truck. Every apple in each of the five containers is then inspected for E. coli. This is an example for
A) cluster sampling
B) simple random sampling
C) stratified sampling
D) systematic sampling

2. **What is wrong?** Explain what is wrong in each of the following scenarios.

a. If the variance of a population is 10, then the variance of the mean for an SRS (Simple Random Sample) of 30 observations from this population will be $10/\sqrt{30}$.

b. The mean of a sampling distribution of $\bar{x}$ changes when the sample size changes.

c. The central limit theorem states that for large n, μ is approximately Normal.

3. You want to find out what percentage of undergraduate WIUT students work full-time in addition to their studies. You plan to survey 150 students. Because higher-level students may be more likely to be working than lower-level students, you decide to collect a representative proportion of students from each level. State which probability sampling method is appropriate, and determine how many students should be selected from each level if there are 1,500 Level 3, 1,450 Level 4, 1,400 Level 5, and 1,250 Level 6 students.

4. Suppose that a market research analyst for a cell phone company conducts a study of their customers who exceed the time allowance included on their basic cell phone contract; the analyst finds that for those people who exceed the time included in their basic contract, the excess time used follows a normal distribution with a mean of 20 minutes and standard deviation of 10 minutes. Find the probability that the average excess time used by the 16 customers in the sample is longer than 25 minutes.

5. Human Resource Consulting (HRC) surveyed a random sample of 60 Twin Cities construction companies to find information on the costs of their health care plans. One of the items being tracked is the annual deductible that employees must pay. The Minnesota Department of Labor reports that historically the mean deductible amount per employee is $502 with a standard deviation of $100.

a. Compute the standard error of the sample mean for HRC.

b. What is the chance HRC finds a sample mean between $477 and $527?

c. Calculate the likelihood that the sample mean is between $492 and $512.

d. What is the probability the sample mean is greater than $550?

6. A study involving stress is done on a college campus among the students. The stress scores follow a uniform distribution with the lowest stress score equal to 1 and the highest equal to 5. Using a sample of 40 students, find:

a. The probability that the average stress score for the 40 students is less than 2.5.

b. The 90th percentile for the average stress score for the sample of 40 students.

7. Let $X_1, X_2, \ldots, X_{36}$ and $Y_1, Y_2, \ldots, Y_{49}$ be monthly salaries of full-time English teachers from Uzbekistan and Kazakhstan, respectively. They are independent random samples from distributions with means $\mu_X = \$600$ and $\mu_Y = 700$ and with standard deviations $\sigma_X = 150$ and $\sigma_Y = 180$. What is the approximate value of $P(\bar{X} > \bar{Y})$?

Note: If X and Y are independent**,** then Var(X−Y) = Var(X) + Var(Y)

HOMEWORK

8. Refer to The Sport Journal (Winter, 2007) analysis of critical part failures at NASCAR races, Recall that researchers found that the time x (in hours) until the first critical part failure is exponentially distributed with $\mu = 0.10$ and $\sigma = 0.10$. Now, consider a random sample of n = 50 NASCAR races and let $\bar{x}$ represent the sample mean time until the first critical part failure.

a. Find $E(\bar{x})$ and $Var(\bar{x})$.

b. Find the probability that the sample mean time until the first critical part failure exceeds 0.13 hour.

9. A random sample of n = 100 observations is selected from a population with $\mu = 30$ and $\sigma = 16$.

a. Find $\mu_{\bar{x}}$ and $\sigma_{\bar{x}}$.

b. Describe the shape of the sampling distribution of $\bar{x}$.

c. Find $P(\bar{X} \geq 28)$.

d. Find $P(22.1 \leq \bar{X} \leq 26.8)$

10. Information from the American Institute of Insurance indicates the mean amount of life insurance per household in the United States is $165,000. This distribution follows the normal distribution with a standard deviation of $40,000. If we select a random sample of 49 households,

a. what is the standard error of the mean?

b. what is the probability of selecting a sample with a mean of at least $167,000?

11. The length of time, in hours, it takes an "over 40" group of people to play one soccer match is normally distributed with a mean of 2 hours and a standard deviation of 0.5 hours. A sample of size n = 50 is drawn randomly from the population.

Find the probability that the sample mean is between 1.8 hours and 2.3 hours.

12. According to an IRS study, it takes a mean of 330 minutes for taxpayers to prepare, copy, and electronically file a 1040 tax form. This distribution of times follows the normal distribution and the standard deviation is 80 minutes. A consumer watchdog agency selects a random sample of 40 taxpayers.

a. What is the standard error of the mean in this example?

b. What is the probability the sample mean is greater than 320 minutes?

c. What is the probability the sample mean is between 320 and 350 minutes?

d. What is the probability the sample mean is greater than 340 minutes if it is known to be less than 350 minutes?