Lecture 9: Sampling Distribution Theory

Prof. Magdy E. El-Adll & Assoc. Prof. Amany E. Aly

Department of Mathematics, Faculty of Science,
Helwan University, Ain Helwan, Cairo, Egypt.

16 / 4 / 2022

Chapter Objectives

Presentation Outline

Chapter Objectives

After completing this chapter, you should be able to:

Chapter Objectives

After completing this chapter, you should be able to:

1. Clarify the basic concepts of sampling theory.

Chapter Objectives

After completing this chapter, you should be able to:

1. Clarify the basic concepts of sampling theory.
2. Distinguish between statistic and parameter.

Chapter Objectives

After completing this chapter, you should be able to:

1. Clarify the basic concepts of sampling theory.
2. Distinguish between statistic and parameter.
3. Find the sampling distribution of $\overline{X}$.

Chapter Objectives

After completing this chapter, you should be able to:

1. Clarify the basic concepts of sampling theory.
2. Distinguish between statistic and parameter.
3. Find the sampling distribution of $\overline{X}$.
4. Identify student $T-$distribution.

The researcher performs an experiment and obtains some data. On the basis of this data, certain conclusions are drawn.

The researcher performs an experiment and obtains some data. On the basis of this data, certain conclusions are drawn.

In other words, researcher may generalize from a particular experiments to the class of all similar experiments.

The researcher performs an experiment and obtains some data. On the basis of this data, certain conclusions are drawn.

In other words, researcher may generalize from a particular experiments to the class of all similar experiments.

This sort of extension from the particular to the general is called *statistical inference.*

The researcher performs an experiment and obtains some data. On the basis of this data, certain conclusions are drawn.

In other words, researcher may generalize from a particular experiments to the class of all similar experiments.

This sort of extension from the particular to the general is called *statistical inference.*

Such inference is used to find new knowledge in the empirical sciences.

In this chapter, we are interested in sampling from populations or probability distributions and study some important quantities such as the sample mean , which will be of vital importance in later chapters.

Presentation Outline

Definition 1

A population is the totality of elements under consideration, from which information is desired.

Definition 1

A population is the totality of elements under consideration, from which information is desired.

- The number of observations in the population is defined to be the size of the population. This size may be finite or infinite.

Definition 1

A population is the totality of elements under consideration, from which information is desired.

- The number of observations in the population is defined to be the size of the population. This size may be finite or infinite.
- For any statistical investigation, it is often important to analyze the whole population.

Definition 1

A population is the totality of elements under consideration, from which information is desired.

- The number of observations in the population is defined to be the size of the population. This size may be finite or infinite.
- For any statistical investigation, it is often important to analyze the whole population.
- In order to overcome the difficulties involved restrictions of times and costs in the studding the whole population, we use the techniques of sampling.

Definition 2

The *sample* is a finite subset of the population. The number of individuals (objects) in the sample is called the *sample size* (the size of the sample). The process of obtaining suitable sample from a population is called *sampling*.

In drawing a sample, our main aim is to choose a representative sample to the population, so that from that sample we can obtain maximum information about the population with minimum effort and to measure and control the introduced error.

In drawing a sample, our main aim is to choose a representative sample to the population, so that from that sample we can obtain maximum information about the population with minimum effort and to measure and control the introduced error.

Certainly, the sample should be representative of the population. It should be a *random sample* in the sense that the observations are made independently and at random.

Definition of random sample

Definition 3

Let $X_1, X_2, \ldots, X_n$ *be* n *independent and identically distributed random variables each of which have the same cumulative distribution function* $F(x)$. *Then, we define* $X_1, X_2, \ldots, X_n$ *to be a random sample of size* n *from the population* $F(x)$.

Presentation Outline

What is the sampling theory?

What is the sampling theory?

- Sampling theory study the relationship between samples and their population.

What is the sampling theory?

- Sampling theory study the relationship between samples and their population.
- Statistical inference depends mainly on the sampling theory.

What is statistic?

What is statistic?

- A variable which computed from a sample is called a *statistic.*

What is statistic?

- A variable which computed from a sample is called a *statistic.*
- Since many random samples can be selected from the same population, we would expect that the statistic vary randomly from sample to sample.

What is statistic?

- A variable which computed from a sample is called a *statistic*.
- Since many random samples can be selected from the same population, we would expect that the statistic vary randomly from sample to sample.

Definition 4

A statistic is a random variable that depends only on the observed random sample.

Some Location Statistics: Sample mean

Definition 5

If $X_1, X_2, \ldots, X_n$ *represent a random sample of size* n, *then the* *sample mean* *is defined by the statistic*

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i.$$

Note that the statistic $\overline{X}$ *is a random variable assumes the value* $\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ *when* X_i *assumes the value* x_i, $i = 1, 2, \ldots, n$.

The sample variance

Definition 6

If $X_1, X_2, \ldots, X_n$ *is a random sample of size* n, *then the* *sample variance* *is defined by the statistic*

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2.$$

The sample variance

Definition 6

If $X_1, X_2, \ldots, X_n$ is a random sample of size n, then the sample variance is defined by the statistic

$$S^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \overline{X})^2.$$

The sample standard deviation, denoted by S is defined to be the positive square root of the sample variance.

Sampling distribution and the standard error

Definition 7

The distribution of a statistic is called a sampling distribution.

Sampling distribution and the standard error

Definition 7

The distribution of a statistic is called a sampling distribution.

Definition 8

The standard deviation of the sampling distribution of a statistic is called a standard error of the statistic.

The Sampling Distribution of $\overline{X}$

The Distribution of $\overline{X}$

Suppose that a random sample of size n is drawn from a normal population with mean μ and variance σ^2. Hence,

$$\boxed{\overline{X} \quad \text{follows the normal distribution} \quad N\left(\mu, \sigma^2/n\right).}$$

If we are sampling from a population with unknown distribution, the sampling distribution of $\overline{X}$ will be approximately normal with mean μ and variance σ^2/n provided that the sample size is large, that is $n \geq 30$. This result is an immediate consequence of the *central limit theorem*

Central Limit Theorem

Theorem 9 (Central Limit Theorem)

If $\overline{X}$ is the mean of a random sample $X_1, X_2, \ldots, X_n$ of size n from a distribution with finite mean μ and finite positive variance σ^2, then the distribution of

$$Z = \frac{\overline{X} - \mu}{\sigma/\sqrt{n}},$$

is $N(0,1)$ in limit as $n \to \infty$.

Presentation Outline

An Important Theorem

Theorem 10

Suppose that $X_1, X_2, \cdots, X_n$ is a random sample of size n from a normal distribution $N(\mu, \sigma^2)$. Then

An Important Theorem

Theorem 10

Suppose that $X_1, X_2, \cdots, X_n$ is a random sample of size n from a normal distribution $N(\mu, \sigma^2)$. Then

1. *The sample mean $\bar{X}$ and the sample variance S^2 are independent.*

An Important Theorem

Theorem 10

Suppose that $X_1, X_2, \cdots, X_n$ is a random sample of size n from a normal distribution $N(\mu, \sigma^2)$. Then

1. *The sample mean $\bar{X}$ and the sample variance S^2 are independent.*
2. *The quantity $\dfrac{(n-1)S^2}{\sigma^2}$ follows χ^2 (read chi-square) distribution with $n-1$ degrees of freedom.*

What is the distribution of $\bar{X}$ if σ is unknown and $n < 30$?

If σ is unknown, $n < 30$, and we sampling from the normal distribution, $N(\mu, \sigma^2)$, a natural statistic can be considered to deal with inferences on μ is

$$T = \frac{\bar{X} - \mu}{S/\sqrt{n}},$$

where S denotes the sample standard deviation.

- If the sample size is large enough, say $n \geqslant 30$, the distribution of T does not differ considerably from the standard normal.

- If the sample size is large enough, say $n \geqslant 30$, the distribution of T does not differ considerably from the standard normal.
- for $n < 30$, it is useful to deal with the exact distribution of T.

In developing the sampling distribution of T, we shall assume that our random sample was selected from a normal population.

Theorem 11

The probability density function of the random variable $T = \frac{\bar{X}-\mu}{S/\sqrt{n}}$, *is given by*

$$f(t) = \frac{\Gamma[(\nu+1)/2]}{\sqrt{\pi\nu}\Gamma(\nu/2)}\left(1+\frac{t^2}{\nu}\right)^{-(\nu+1)/2}, \quad -\infty < t < \infty.$$

The distribution is known as the Student $t-$*distribution with* $\nu = n-1$ *degrees of freedom.*

What Does the $t-$Distribution Look Like?

- The distribution of T is similar to the distribution of Z in that they both are symmetric about a mean of zero.

- The distribution of T is similar to the distribution of Z in that they both are symmetric about a mean of zero.
- Both distributions are bell shaped, but the $t-$distribution is more variable, owing to the fact that the $T-$values depend on of two quantities, $\bar{X}$ and S^2,

- the Z-values depend only on the changes in $\bar{X}$ from sample to sample.

- the Z-values depend only on the changes in $\bar{X}$ from sample to sample.
- The distribution of T differs from that of Z in that the variance of T depends on the sample size n and is always greater than 1.

- the Z-values depend only on the changes in $\bar{X}$ from sample to sample.
- The distribution of T differs from that of Z in that the variance of T depends on the sample size n and is always greater than 1.
- Only when the sample size $n \to \infty$ will the two distributions become the same.

- the Z-values depend only on the changes in $\bar{X}$ from sample to sample.
- The distribution of T differs from that of Z in that the variance of T depends on the sample size n and is always greater than 1.
- Only when the sample size $n \to \infty$ will the two distributions become the same.

Figure 4.7, page 59 show the relationship between a standard normal distribution $(\nu = \infty)$ and $t-$distributions with 2 and 5 degrees of freedom.

What Is the $t-$Distribution Used For?

The $t-$distribution is used extensively in problems that deal with inference about the population mean.

Presentation Outline

Example

Example 12

Traveling between two campuses of a university in a city via shuttle bus takes, on average, 28 minutes with a standard deviation of 5 minutes. In a given week, a bus transported passengers 40 times. What is the probability that the average transport time was more than 30 minutes? Assume the mean time is measured to the nearest minute.

Solution.

In this case, $\mu = 28$ and $\sigma = 5$. We need to calculate the probability $P(\overline{X} > 30)$ with $n = 40$. Hence,

Solution.

In this case, $\mu = 28$ and $\sigma = 5$. We need to calculate the probability $P(\overline{X} > 30)$ with $n = 40$. Hence,

$$P(\overline{X} > 30)$$

Solution.

In this case, $\mu = 28$ and $\sigma = 5$. We need to calculate the probability $P(\overline{X} > 30)$ with $n = 40$. Hence,

$$P(\overline{X} > 30) = P\left(\frac{\overline{X} - 28}{5/\sqrt{40}} \geqslant \frac{30 - 28}{5/\sqrt{40}}\right)$$

Solution.
In this case, $\mu = 28$ and $\sigma = 5$. We need to calculate the probability $P(\overline{X} > 30)$ with $n = 40$. Hence,

$$P(\overline{X} > 30) = P\left(\frac{\overline{X} - 28}{5/\sqrt{40}} \geqslant \frac{30 - 28}{5/\sqrt{40}}\right)$$

$$= P(Z \geqslant 3.16) = 0.0008.$$

Example 13

The heights of students are normally distributed with a mean of 174.5 and a standard deviation of 6.9. Suppose 1000 random samples of size 25 are drawn from this population. Determine

Example

Example 13

The heights of students are normally distributed with a mean of 174.5 and a standard deviation of 6.9. Suppose 1000 random samples of size 25 are drawn from this population. Determine

1. the mean and standard deviation of the sampling distribution of $\overline{X}$;

Example

Example 13

The heights of students are normally distributed with a mean of 174.5 and a standard deviation of 6.9. Suppose 1000 random samples of size 25 are drawn from this population. Determine

1. the mean and standard deviation of the sampling distribution of $\overline{X}$;
2. the probability of sample mean that fall between 172.5 and 175.8 cm;

Example

Example 13

The heights of students are normally distributed with a mean of 174.5 and a standard deviation of 6.9. Suppose 1000 random samples of size 25 are drawn from this population. Determine

1. the mean and standard deviation of the sampling distribution of $\overline{X}$;
2. the probability of sample mean that fall between 172.5 and 175.8 cm;
3. the number of sample means falling below 172.0 cm.

Solution.

1. $\mu_{\overline{X}} = \mu = 174.5$ and $\sigma_{\overline{X}} = \sigma/\sqrt{n} = 6.9/\sqrt{25} = 1.38$;
2. $$P(172.5 \leqslant \overline{X} \leqslant 175.8)$$

Solution.

1. $\mu_{\overline{X}} = \mu = 174.5$ and $\sigma_{\overline{X}} = \sigma/\sqrt{n} = 6.9/\sqrt{25} = 1.38$;
2. $$P(172.5 \leqslant \overline{X} \leqslant 175.8) \;=\; P\left(\frac{172.5 - 174.5}{6.9/\sqrt{25}} \leqslant \frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \leqslant \frac{175.8 - 174.5}{6.9/\sqrt{25}}\right)$$

Solution.

1. $\mu_{\overline{X}} = \mu = 174.5$ and $\sigma_{\overline{X}} = \sigma/\sqrt{n} = 6.9/\sqrt{25} = 1.38$;
2. $$\begin{aligned} P(172.5 \leqslant \overline{X} \leqslant 175.8) &= P\left(\frac{172.5 - 174.5}{6.9/\sqrt{25}} \leqslant \frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \leqslant \frac{175.8 - 174.5}{6.9/\sqrt{25}}\right) \\ &= P(-1.45 \leqslant Z \leqslant 0.942) \end{aligned}$$

Solution.

1. $\mu_{\overline{X}} = \mu = 174.5$ and $\sigma_{\overline{X}} = \sigma/\sqrt{n} = 6.9/\sqrt{25} = 1.38$;
2. $$\begin{aligned} P(172.5 \leqslant \overline{X} \leqslant 175.8) &= P\left(\frac{172.5 - 174.5}{6.9/\sqrt{25}} \leqslant \frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \leqslant \frac{175.8 - 174.5}{6.9/\sqrt{25}}\right) \\ &= P(-1.45 \leqslant Z \leqslant 0.942) \\ &= \Phi(0.94) - \Phi(-1.45) \end{aligned}$$

Solution.

1. $\mu_{\overline{X}} = \mu = 174.5$ and $\sigma_{\overline{X}} = \sigma/\sqrt{n} = 6.9/\sqrt{25} = 1.38$;
2. $$\begin{aligned} P(172.5 \leqslant \overline{X} \leqslant 175.8) &= P\left(\frac{172.5 - 174.5}{6.9/\sqrt{25}} \leqslant \frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \leqslant \frac{175.8 - 174.5}{6.9/\sqrt{25}}\right) \\ &= P(-1.45 \leqslant Z \leqslant 0.942) \\ &= \Phi(0.94) - \Phi(-1.45) \\ &= 0.8264 - (1 - 0.9265) = 0.7529. \end{aligned}$$

3. First find $P(\overline{X} > 172.0)$ and then multiply the result by $n = 1000$.

$$
\begin{aligned}
P(\overline{X} > 172.0) &= P\left(\frac{\overline{X}-\mu}{\sigma/\sqrt{n}} > \frac{\overline{X}-\mu}{\sigma/\sqrt{n}}\right) \\
&= P\left(Z > \frac{172.0 - 174.5}{6.9/\sqrt{25}}\right) \\
&= P(Z > -1.81) = 0.9649.
\end{aligned}
$$

Hence, the number of sample means falling below 172.0 cm. is $(0.9649 \times 1000) \approx 965$ samples.

Example

Example 14

Suppose that the IQ scores for a given population follows $N(105, \sigma^2)$ where σ is unknown. If a random sample of size $n = 25$ is drown from this population have $s = 10$. Find $P(\overline{X} > 109.128)$.

Solution.

Since the population is normally distributed, $n < 30$, and σ is unknown,

$$\frac{\overline{X} - \mu}{S/\sqrt{n}} \sim t(n-1).$$

Solution.

Since the population is normally distributed, $n < 30$, and σ is unknown,

$$\frac{\overline{X} - \mu}{S/\sqrt{n}} \sim t(n-1).$$

$$P(\overline{X} > 109.128)$$

Solution.

Since the population is normally distributed, $n < 30$, and σ is unknown,

$$\frac{\overline{X} - \mu}{S/\sqrt{n}} \sim t(n-1).$$

$$P(\overline{X} > 109.128) = P\left(\frac{\overline{X} - \mu}{S/\sqrt{n}} > \frac{109.128 - 105}{10/5}\right)$$

Solution.

Since the population is normally distributed, $n < 30$, and σ is unknown,

$$\frac{\overline{X} - \mu}{S/\sqrt{n}} \sim t(n-1).$$

$$\begin{aligned} P(\overline{X} > 109.128) &= P\left(\frac{\overline{X} - \mu}{S/\sqrt{n}} > \frac{109.128 - 105}{10/5}\right) \\ &> P(T > 2.064) \end{aligned}$$

Solution.

Since the population is normally distributed, $n < 30$, and σ is unknown,

$$\frac{\overline{X} - \mu}{S/\sqrt{n}} \sim t(n-1).$$

$$\begin{aligned} P(\overline{X} > 109.128) &= P\left(\frac{\overline{X} - \mu}{S/\sqrt{n}} > \frac{109.128 - 105}{10/5}\right) \\ &> P(T > 2.064) \\ &= 0.025 \end{aligned}$$