

# Ch1 - Point and Interval Estimation

## Statistics For Business and Economics - II

Shaikh Tanvir Hossain

East West University, Dhaka

Last updated - October 29, 2023

# Outline

## Outline

1. **Statistical Inference - from Sample to Population**
2. **Point Estimator and Sampling Distribution**
  - Point Estimation
  - Sampling Distribution
  - Properties of Point Estimator
  - How to get Point Estimators
3. **Interval Estimation of Population Mean  $\mu$** 
  - Basic idea of Interval Estimation
  - Interval Estimation - First Example  $\sigma$  known case
  - Deriving Interval Estimator -  $\sigma$  known case
  - Interval Estimator -  $\sigma$  unknown case
  - Interval Estimator -  $\sigma$  unknown case with large samples

## 1. Statistical Inference - from Sample to Population

### 2. Point Estimator and Sampling Distribution

- Point Estimation
- Sampling Distribution
- Properties of Point Estimator
- How to get Point Estimators

### 3. Interval Estimation of Population Mean $\mu$

- Basic idea of Interval Estimation
- Interval Estimation - First Example  $\sigma$  known case
- Deriving Interval Estimator -  $\sigma$  known case
- Interval Estimator -  $\sigma$  unknown case
- Interval Estimator -  $\sigma$  unknown case with large samples

## **Statistical Inference - from Sample to Population**

# Population Data Vs. Sample Data

- ▶ Let's recap ECO104, do you know the difference between the Population and a Sample?
- ▶ Suppose we collected a data from 5 students studying currently at EWU (hypothetical data). You know that the columns are called *variables* and the rows are called *observations* or *units*.

	Gender	Monthly Income (tk)	ECO-101 Grade	# Retakes
Student A	Male	3615	B-	3
Student B	Female	49755	A	2
Student C	Male	44758	A	1
Student D	Female	3879	B	0
Student E	Male	22579	A+	2

- ▶ What is the Population of this study?
- ▶ What is the Population Proportion of Male Students? (Or Female Students?)
- ▶ What is the Population Mean of Income?
- ▶ We need to talk about what do we mean by the word "Population" and "Sample" in Statistics?
- ▶ What is the "Sample Proportion of Male" / "Sample Mean of Income" ?
- ▶ What's the difference?
- ▶ Last question - what is the sample size?

# Population Data Vs. Sample Data

## Definition 1.1: (Population and Sample)

The collection / set of *all observations* in a particular study is called the population. A sample is a subset of the population.

- ▶ Why Population matters?
- ▶ One answer could be

*"Maybe we are interested to get some idea about all of the students studying currently at EWU."*

- ▶ In this case, we say the *population* is the set of all current students at EWU. And the set of *5 students* is a sample of that population.
- ▶ Usually collecting population data is very time consuming and often impossible, here is an example, think about when the population is the set of all EWU students from the beginning of EWU.
- ▶ So we collect a sample of the population.

# Population Data Vs. Sample Data

- ▶ Note that a sample is supposed to be a good representative of the whole population.
- ▶ If the sample is not a good representative, then we say we have a *biased sample*.
- ▶ Biased sample is bad, why?... because any conclusion from a biased sample might lead to incorrect conclusion regarding the population (can you think about an example?)
- ▶ One way to get a good sample is - *Simple Random Sampling!* (details on board)
- ▶ What if the population is infinite ??? Is it even possible in reality ???
- ▶ One of the major tasks of statistical analysis is, using a sample to make some conclusions regarding the population, this process is called *Statistical Inference*, or we call this - *Inferential Statistics*.
- ▶ This is different than descriptive statistics, recall in descriptive statistics we have NO goal of making inference regarding population, we just *describe* the data, that's it!
- ▶ There were two types of descriptive statistics,
  - ▶ Graphical Methods (Bar Chart, Pie Chart, Histogram, Scatter Plot)
  - ▶ Numerical Methods (Sample Mean, Median, Mode, Variance, Percentile (or quantile), sample covariance, correlation etc)
- ▶ When we graph or calculate these things there is no goal of making inference! we just describe the data
- ▶ Here we will use the same techniques, but our goal is one step more - that is making inference about the population.
- ▶ For example, we can calculate the sample mean and make inference for the population mean.
- ▶ Does our estimation improve if we have larger sample size? yes... why? any law ???

## 1. Statistical Inference - from Sample to Population

## 2. Point Estimator and Sampling Distribution

- Point Estimation
- Sampling Distribution
- Properties of Point Estimator
- How to get Point Estimators

## 3. Interval Estimation of Population Mean $\mu$

- Basic idea of Interval Estimation
- Interval Estimation - First Example  $\sigma$  known case
- Deriving Interval Estimator -  $\sigma$  known case
- Interval Estimator -  $\sigma$  unknown case
- Interval Estimator -  $\sigma$  unknown case with large samples



## Point Estimator and Sampling Distribution

# **Point Estimator and Sampling Distribution**

## **Point Estimation**

# Point Estimation

- ▶ We already saw one example of a point estimation. For example, it could be that we are interested in Population Mean, so this is our target, and we are guessing this target with sample mean!
- ▶ This process in general is known as *Point Estimation*, it's a concept in Inferential Statistics, where you just give one number as a guess!
- ▶ But there are other methods too!
- ▶ In fact there are two major themes of statistical inference
  - ▶ 1. Estimation - Point Estimation and Interval Estimation (a.k.a confidence interval)
  - ▶ 2. Testing (a.k.a Hypothesis Testing).
- ▶ Let's see a rough example on board about estimation and testing (more details will come in the coming chapters!)
- ▶ In this chapter we will focus on estimation, both point and interval, but we will discuss point estimation.

# Point Estimation

- ▶ *We will write the point estimation idea with notation* (please don't be scared!)
- ▶ Suppose we are interested in the *Population mean  $\mu$*
- ▶ But we cannot access it and we only have a sample  $x_1, x_2, \dots, x_n$  (These are fixed numbers for a sample of size  $n$ )
- ▶ So we find the sample mean  $\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$
- ▶ This sample mean  $\bar{x}$  is a *point estimate* of the *unknown Population mean  $\mu$* .
- ▶ This process is what we call *point estimation*.
- ▶ Question - Does the estimate changes with different sample? How do we write this? We need to think about random variables....

# **Point Estimator and Sampling Distribution**

## **Sampling Distribution**

# Point Estimator and Sampling Distribution

- Suppose we have an income data of 10 units / observations,

	Income	Random variable
1.	20	$X_1 = ?$
2.	60	$X_2 = ?$
3.	20	$X_3 = ?$
4.	-20	$X_4 = ?$
5.	-30	$X_5 = ?$
6.	-10	$X_6 = ?$
7.	80	$X_7 = ?$
8.	10	$X_8 = ?$
9.	30	$X_9 = ?$
10.	40	$X_{10} = ?$

Table: Income data

- We can think about a data of a single unit in two ways
  - A *realized data*, where the randomness is gone and we have observed the value, for example  $x_1$
  - Or a *random variable* for example,  $X_1$
- In this way for the whole data set, we can think as 10 fixed number  $x_1, x_2, \dots, x_{10}$  (this is when the sample is fixed and there is no randomness) or  $X_1, X_2, X_3, \dots, X_{10}$ , which are 10 random variables.

# Point Estimator and Sampling Distribution

- ▶ Generally when we think about  $n$  random variables,  $X_1, X_2, X_3, \dots, X_n$ , we will call it a *random sample* (the other one is the fixed sample!)
- ▶ The idea of *Estimator* comes when we think about a random sample.
- ▶ An Estimator is a function of a random sample, so this is random and hence this is a *random variable*.
- ▶ Since an *Estimator* is a random variable, it changes from sample to sample, but when we calculate it for a fixed sample, then we get  $\bar{x}$ . Here  $\bar{x}$  is a constant and it's not random.
- ▶ Here is the Estimator, which is a random variable!

$$\bar{X}_n = \frac{1}{n} \sum_{i=1}^n X_i \quad , \text{ an estimator}$$

- ▶ Since  $\bar{X}_n$  is a random variable, question is what is the probability distribution of  $\bar{X}_n$ ? or Expectation of  $\mathbb{E}(\bar{X}_n)$ . Do you know what is Expectation? or Variance.
- ▶ Let's recap some of the ideas from ECO104 .....
- ▶ Recall Expectation of a random variable  $X$ , written with  $\mathbb{E}(X)$  (which is like average) but its a weighted average, what is the formula?
- ▶ What is the variance of a random variable, written with  $\mathbb{V}(X)$ ?
- ▶ We are interested in 3 important questions,
  1. What is the Expectation of the random variable  $\bar{X}_n$ , written as  $\mathbb{E}(\bar{X}_n)$ ?
  2. What is the variance of the random variable  $\bar{X}_n$ , written  $\mathbb{V}(\bar{X}_n)$ ?
  3. What is the probability distribution of  $\bar{X}_n$  (this is what we call *sampling distribution*!)

# Point Estimator and Sampling Distribution

- ▶ The answer to the third question is what we call *Sampling Distribution of Means*.
- ▶ Note that, this is the distribution of sample means  $\bar{x}$ , that we get from repeated sampling!
- ▶ This is possibly the most important object for now, ...
- ▶ Definitely if we know the answer of 3, we know the answers of 1 and 2.
- ▶ We will now three important results ....
- ▶ *Important remarks regarding some notations:*
  - ▶ If we write  $\mathbb{E}(X)$ , this means  $X$  is a random variable and  $\mathbb{E}(X)$  is the *Expected Value* of  $X$ , or we say *Expectation* of  $X$ . Recall, the idea of Expectation is very similar to average, but it is a population average.
  - ▶ Similarly if we write  $\mathbb{V}(X)$ , this means the variance of  $X$ , again this is the population variance
  - ▶ Both Expectation and Variance depends on the population probability distribution.

## Theorem 1.2: (Mean and Variance of $\bar{X}_n$ with only i.i.d assumption)

If we have i.i.d random variables  $X_1, X_2, \dots, X_n$  with the same mean  $\mu$  and same variance  $\sigma^2$ , then

$$i) \quad \mathbb{E}(\bar{X}_n) = \mu \quad (1)$$

$$ii) \quad \mathbb{V}(\bar{X}_n) = \frac{\sigma^2}{n} \quad (2)$$



# Point Estimator and Sampling Distribution

## Theorem 1.3: (Distribution of $\bar{X}_n$ with normality and i.i.d assumption)

If we have i.i.d random variables  $X_1, X_2, \dots, X_n$  where they all are distributed with  $\mathcal{N}(\mu, \sigma^2)$ , then

$$i) \quad \mathbb{E}(\bar{X}_n) = \mu$$

$$ii) \quad \mathbb{V}(\bar{X}_n) = \frac{\sigma^2}{n}$$

$$iii) \quad \bar{X}_n \sim \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right)$$

$$iv) \quad Z_n \sim \mathcal{N}(0, 1) \text{ where } Z_n = \frac{\bar{X}_n - \mu}{\frac{\sigma}{\sqrt{n}}} \quad (3)$$

$$v) \quad T_n \sim t_{n-1} \text{ where } T_n = \frac{\bar{X}_n - \mu}{\frac{S}{\sqrt{n}}} \text{ and } S^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X}_n)^2 \quad (4)$$

# Point Estimator and Sampling Distribution

## Theorem 1.4: (Central Limit Theorem (CLT) and related results)

Let  $X_1, X_2, \dots, X_n$  be i.i.d random variables with population mean  $\mu$  and variance  $\sigma^2$ . Then for *large  $n$  (technically we need  $n \rightarrow \infty$ )*, we get following results:

$$i) \quad Z_n \overset{\text{approx}}{\sim} \mathcal{N}(0, 1) \text{ where } Z_n = \frac{\bar{X}_n - \mu}{\frac{\sigma}{\sqrt{n}}} \quad [\text{CLT}] \quad (5)$$

$$ii) \quad \bar{X}_n \overset{\text{approx}}{\sim} \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right)$$

$$iii) \quad T_n \overset{\text{approx}}{\sim} \mathcal{N}(0, 1) \text{ where } T_n = \frac{\bar{X}_n - \mu}{\frac{S}{\sqrt{n}}}$$

# Statistic, Point Estimator and Sampling Distribution - Important Remarks

- ▶ So we understood that *the idea of the sampling distribution is a repeated sampling idea*. In real life you can only have one sample, so you can never calculate this using a sample data.
- ▶ And the last three results tell us that, we can only know the sampling distribution of means under certain assumptions (in particular we need either normality or large sample size)
- ▶ If we assume normality (this means our data is normally distributed), then the distribution of the sample means is also normal and this result is for any sample size! This is called the *exact distribution*!
- ▶ If we don't assume normality for the population, then usually we have no hope, except for large  $n$ .
- ▶ The standard deviation of sampling distribution is called *standard error*! This is standard deviation, but this name is special for sampling distribution.
- ▶ In general *any function* of the random sample is called a "*Statistic*", so an estimator is also a *Statistic*. The difference is Estimator is a type of Statistic where we are estimating some target! A statistic might not have any goal, it's just a function of random variables  $X_1, X_2, X_3, \dots, X_n$ ! The distribution of statistic is called *sampling distribution*.
- ▶ For example,  $\bar{X}_n, Z_n$  are both examples statistics but  $\bar{X}_n$  is an estimator for  $\mu$ ,  $Z_n$  is just a statistic.
- ▶ Another example is  $S^2$ , where  $S^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X}_n)^2$ . This is a statistic since it's a function of the random sample. And this is also an estimator for  $\sigma^2$ , because it is targeting population variance  $\sigma^2$ . Note that  $S^2$  is just a sample variance.

# **Point Estimator and Sampling Distribution**

## **Properties of Point Estimator**

# Properties of Point Estimators

- ▶ Why did we take the average to estimate  $\mu$ ? Why not median, or maximum? These all are examples of estimators for  $\mu$ , so why sample mean  $\bar{X}_n$ ?
- ▶ The answer is, the sample average is a “good” estimator for the population mean  $\mu$
- ▶ What do we mean by “good”?
- ▶ One answer is - it is an “unbiased” and a “consistent estimator”?
- ▶ What does “unbiasedness” mean? In notation this means

$$\mathbb{E}(\bar{X}_n) = \mu$$

- ▶ The interpretation is, if we calculate, sample means many times, *on average* we are not doing a bad job, even if our sample size  $n$  is not that big.
- ▶ Draw the dart picture.... on board
- ▶ Notice this result does not depend on the sample sizes, so we say unbiasedness is a finite sample property of an estimator.

# Properties of Point Estimators

- Now let's focus on *consistency*, we say an estimator is a *consistent estimator* then, if we have  $n \rightarrow \infty$  then there is a very high probability that  $\bar{X}_n$  will approach to  $\mu$ . So we can say

if  $n \rightarrow \infty$  then  $\bar{X}_n \rightarrow \mu$  happens with very high probability

- So this says, even if for small sample our sample mean is doing a bad job, as we increase the sample size we will eventually go very close to  $\mu$ .
- If you contrast consistency to unbiasedness, you will notice that this is a limiting property or we say *asymptotic property* (because we are saying  $n \rightarrow \infty$ , recall limit...), as opposed to finite sample property!
- The estimator sample mean  $\bar{X}_n$  is both an unbiased and a consistent estimator for the population mean  $\mu$ .

# Properties of Point Estimators

- ▶ So we talked about unbiasedness and consistency of an estimator, there is another thing called *variance* of an estimator, for sample mean this is  $\mathbb{V}(\bar{X}_n) = \frac{\sigma^2}{n}$  (In Anderson you will see the notation  $\sigma_{\bar{X}}^2$  to represent the same object, but I will not use this notation).
- ▶ You will not see too much discussion of the variance of an estimator here, but in higher courses this theme will come a lot!
- ▶ Looking at the variance of estimators is useful if we want to compare *two or more estimators*.
- ▶ It is possible that two estimators are unbiased, one has very high variance.
- ▶ This means on average both are doing fine, but one has very high uncertainty!
- ▶ Again the dart picture!

# **Point Estimator and Sampling Distribution**

## **How to get Point Estimators**



# How to get Estimators

- ▶ There are many techniques to get estimators. For example, here are some common techniques,
  - ▶ Method of Least Squares
  - ▶ Method of Maximum Likelihood
  - ▶ Method of Moments
- ▶ Sadly in this course we will not cover systematically any of these technique, but in higher courses you will see all of these methods.
- ▶ But we will talk about *method of least squares* when we talk about regression, and already we have seen some examples of *method of moment* techniques, ... roughly you can think the word “moment” is same as “expectation”...
- ▶ In the method of moment idea, we can get estimators by *replacing population expectation with averages*.

$\mathbb{E}(X) = \mu$ , if this is our target object

$$\bar{X}_n = \frac{1}{n} \sum_{i=1}^n X_i \quad , \text{ replace the expectation, then we have an estimator}$$

- ▶ Can you think about an estimator of  $\sigma^2$ , recall  $\sigma^2$  is actually the population variance of  $X$ , so  $\text{Var}(X) = \sigma^2$ , and for variance we have the following formula

$$\sigma^2 = \text{Var}(X) = \mathbb{E} \left[ (X - \mathbb{E}(X))^2 \right]$$

# How to get Estimators

- This should be sample variance  $S^2$ , where

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

- Again note that, we replaced Expectation with averages.
- $S^2$  is an unbiased estimator of  $\sigma^2$ .
- What if we divide by  $n$  rather than  $n-1$ ? this is also an estimator of  $\sigma^2$ , but unfortunately this is a biased estimator!

## 1. Statistical Inference - from Sample to Population

## 2. Point Estimator and Sampling Distribution

- Point Estimation
- Sampling Distribution
- Properties of Point Estimator
- How to get Point Estimators

## 3. Interval Estimation of Population Mean $\mu$

- Basic idea of Interval Estimation
- Interval Estimation - First Example  $\sigma$  known case
- Deriving Interval Estimator -  $\sigma$  known case
- Interval Estimator -  $\sigma$  unknown case
- Interval Estimator -  $\sigma$  unknown case with large samples

## Interval Estimation of Population Mean $\mu$

## Interval Estimation of Population Mean $\mu$

### Basic idea of Interval Estimation

# Interval Estimators

- ▶ Before we proceed, recall  $\bar{X}_n$  is the *point estimator* of the population mean  $\mu$ , and for a fixed sample  $\bar{x}$  is what we call *an estimate* of the unknown parameter  $\mu$ .
- ▶ Point estimator is nice, but it is rather crude! we are just giving one number as a guess.
- ▶ Now we will discuss another type of estimation, known as *Interval estimation*. Here also we will have *Interval estimators* (which is a random interval) and an *Interval estimate* for a fixed sample.
- ▶ Interval estimators is a little bit flexible, because it gives a range of possible values of the parameter (not just one value)!
- ▶ In particular, given  $\alpha$ , where  $0 < \alpha < 1$ , we say a  *$100(1 - \alpha)\%$  interval estimator for  $\mu$  is a random interval  $[L, U]$  such that*

$$\mathbb{P}(L \leq \mu \leq U) = 1 - \alpha \quad (6)$$

- ▶ For example, if we want to construct a 95% interval estimator, then  $1 - \alpha = .95$ , and we want to find  $L$ , and  $U$  such that, there is a 95% possibility that the true parameter will fall in this interval. So this means,

$$\mathbb{P}(L \leq \mu \leq U) = .95 \quad (7)$$

- ▶ Recall the *the frequency interpretation of probability*.
- ▶ Using the frequency interpretation means, if we construct the interval  $[L, U]$  around 100 times, roughly 95 times the true  $\mu$  fall in this interval.

# Interval Estimators

- ▶ We can only think about probability for a random object, so where is this probability coming from?
- ▶ First of all note that,  $\mu$  here is not random (in classical statistics the parameter is never a random object, it is always fixed!), so what is random inside the probability?
- ▶ Actually, we will see that the  $L$  and  $U$  are random in repeated sampling.
- ▶ In fact, the random  $L$  and  $U$  depends on the random sample  $X_1, X_2, X_3, \dots, X_n$ , so we should write,  $L(X_1, X_2, \dots, X_n)$  and  $U(X_1, X_2, \dots, X_n)$ . But just to make our life easier, we will use  $L$  and  $U$ . You should understand that these are functions of the random sample.
- ▶ We will see that we will construct interval estimator of the type,

$$\mathbb{P}(L \leq \mu \leq U) = 1 - \alpha$$

- ▶ The interpretation is,

*"In a repeated sampling, 95 out 100 times the interval constructed using  $[L, U]$  will contain the true parameter  $\mu$ "*

# Interval Estimators

- ▶ “Ideally” the interval  $[L, U]$  should have two properties:
  - ▶  $\mathbb{P}(L \leq \mu \leq U)$  *should be high*, that is, the true parameter  $\mu \in [L, U]$  will happen with high probability.
  - ▶ The length of the interval  $[L, U]$  should be relatively narrow on average.
- ▶ How do we find a interval estimator? There are different methods, but definitely we need to use a statistic and the distribution of the statistic (i.e., the sampling distribution)



## Interval Estimation of Population Mean $\mu$

Interval Estimation - First Example  $\sigma$  known case

# Interval Estimate / Confidence Interval

$\sigma$  known

Let's do an example first where we will calculate interval estimate for a fixed sample.

**Example 1.5:** (Interval Estimator and Interval Estimate/Confidence Interval)

Suppose we have  $\bar{x} = 82$ , population standard deviation  $\sigma = 20$ , sample size  $n = 100$ , and we are asked to compute the 95% *confidence interval or interval estimate of the population mean  $\mu$* , then since  $z_{1-\alpha/2} = z_{0.975} = 1.96$  (this is  $1 - \alpha/2$  quantile of the standard normal distribution and you can calculate this using R function `qnorm(.975)`), the *interval estimator* is

$$[\bar{X}_n - 1.96 \frac{20}{\sqrt{100}}, \quad \bar{X}_n + 1.96 \frac{20}{\sqrt{100}}] \quad (8)$$

The *interval estimate or confidence interval* is

$$\begin{aligned} & [82 - 1.96 \frac{20}{\sqrt{100}}, \quad 82 + 1.96 \frac{20}{\sqrt{100}}] \\ &= [82 - 3.92, \quad 82 + 3.92] \\ &= [78, \quad 85.92] \end{aligned} \quad (9)$$

Now note, the first one at (8) is a *random interval* since  $\bar{X}_n$  is random but the second one (9) is a deterministic interval (there is no randomness here!), this is the interval estimate, [Anderson et al. \(2020\)](#) called this *confidence interval*.

So in the second one either our population mean  $\mu$  is there or it is not there. If you say that there is a 95% probability that true parameter  $\mu$  will fall inside  $[78, \quad 85.92]$ , this is a *wrong interpretation*. We can say “for this particular sample, *the interval estimate* is  $[78, 85.92]$ ”.


# Interval Estimate / Confidence Interval

$\sigma$  known

- ▶ So what is the correct interpretation? - You should say - *if we construct these kinds of intervals 100 times, then roughly 95 times our true parameter will fall inside.*
- ▶ So now we have a probabilistic interpretation.
- ▶ *A Side Note:* Note that when we constructed the interval estimate, we added and subtracted the following same number with  $\bar{x}$

$$\frac{\sigma}{\sqrt{n}} \times z_{1-\alpha/2}$$

Here  $\sigma/\sqrt{n}$  is the standard error and the whole term is called the *margin of error* of the point estimate.

- ▶ The idea is  $\bar{x}$  is our point estimate, but of course there might be some error, so we say that with  $1 - \alpha$  confidence roughly the error is  $\frac{\sigma}{\sqrt{n}} \times z_{1-\alpha/2}$
- ▶ Let's see how we can do the whole calculation of Example 1.5 in 
- ▶ First note, we have following information
  - ▶  $n = 100$
  - ▶  $\bar{x} = 82$
  - ▶  $\alpha = 0.05$  (this is because we are asked to construct 90% confidence interval)
  - ▶  $\sigma = 20$

# Interval Estimate / Confidence Interval

$\sigma$  known

## code - sigma known (confidence interval)

```
# First create some objects with the information given
n <- 100
xbar <- 82
alpha <- 0.05
sigma <- 20

# calculate sderror and moe and save them as objects
sderror <- sigma/sqrt(n)
moe <- qnorm(1 - alpha/2) * sderror

# upper limit
xbar + moe
# [1] 85.91993

# lower limit
xbar - moe
# [1] 78.08007
```

- So the interval estimate or the confidence interval in this case is (78.08 , 85.91)

## Interval Estimation of Population Mean $\mu$

Deriving Interval Estimator -  $\sigma$  known case

# Deriving Interval Estimators

$\sigma$  known

- ▶ Now how did we get that interval?
- ▶ Suppose we already know that our population data is normally distributed,
- ▶ This means all the random variables  $X_1, X_2, \dots, X_n$  are also normally distributed with the distribution  $\mathcal{N}(\mu, \sigma^2)$ . Additionally assume they are independent.
- ▶ This means we have [applying Theorem 1.3 (iii)]

$$\bar{X}_n \sim \mathcal{N}(\mu, \sigma^2 / n)$$

- ▶ But this also means [applying Theorem 1.3 (iv)] (this is just doing standardization)

$$Z_n = \frac{\bar{X}_n - \mu}{\frac{\sigma}{\sqrt{n}}} \sim \mathcal{N}(0, 1)$$

- ▶ Recall  $\bar{X}_n$  is a statistic and an estimator of  $\mu$ . In this case  $Z_n$  is also a statistic, the benefit of transforming  $\bar{X}_n$  to  $Z_n$  is we can now use standard normal. Why did we do this? We will see that here  $Z_n$  also plays an important role to find the interval estimator for  $\mu$ .
- ▶ Now, let's derive the interval estimator for  $\mu$ . You can skip the derivation for exam but I recommend you to do it at least once in your lifetime, actually this is not difficult at all.

# Deriving Interval Estimators

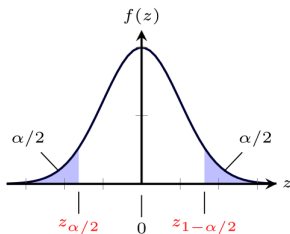
$\sigma$  known

We will construct two-sided interval (it is also possible to construct one sided interval, I will write something in the Appendix, but you can skip it for exam!).

To get the two sided interval, first fix  $\alpha$ , let's say  $\alpha = 5\%$  then  $1 - \alpha$  is what we call *confidence coefficient / confidence level / nominal coverage*. Now since  $Z_n \sim \mathcal{N}(0, 1)$ , we can write

$$\mathbb{P}(z_{\alpha/2} \leq Z_n \leq z_{1-\alpha/2}) = 1 - \alpha \quad (10)$$

Visually this means,



Here  $z_{\alpha/2}$  is a value such that  $\mathbb{P}(Z_n < z_{\alpha/2}) = \alpha/2$  and  $z_{1-\alpha/2}$  is a value such that  $\mathbb{P}(Z_n < z_{1-\alpha/2}) = 1 - \alpha/2$ . It is important to mention that because of the *symmetry* of the Normal distribution always we will have  $z_{\alpha/2} = -z_{1-\alpha/2}$  (note the two tail probabilities are equal, and it is  $\alpha/2$ ).

# Deriving Interval Estimators

$\sigma$  known

Now we will do some algebra with the term inside the probability in (10), recall we had

$$z_{\alpha/2} \leq Z_n \leq z_{1-\alpha/2},$$

$$z_{\alpha/2} \leq Z_n \leq z_{1-\alpha/2} = -z_{1-\alpha/2} \leq Z_n \leq z_{1-\alpha/2} \text{ [using symmetry of the normal]}$$

$$= -z_{1-\alpha/2} \leq \frac{\bar{X}_n - \mu}{\frac{\sigma}{\sqrt{n}}} \leq z_{1-\alpha/2}$$

$$= -\frac{\sigma}{\sqrt{n}} z_{1-\alpha/2} \leq \bar{X}_n - \mu \leq \frac{\sigma}{\sqrt{n}} z_{1-\alpha/2} \text{ [multiplying all sides by } \sigma/\sqrt{n} \text{]}$$

$$= \frac{\sigma}{\sqrt{n}} z_{1-\alpha/2} \geq -\bar{X}_n + \mu \geq -\frac{\sigma}{\sqrt{n}} z_{1-\alpha/2} \text{ [multiplying all sides by } -1 \text{]}$$

$$= -\frac{\sigma}{\sqrt{n}} z_{1-\alpha/2} \leq -\bar{X}_n + \mu \leq \frac{\sigma}{\sqrt{n}} z_{1-\alpha/2} \text{ [rewriting the inequalities]}$$

$$= \bar{X}_n - \frac{\sigma}{\sqrt{n}} z_{1-\alpha/2} \leq \mu \leq \bar{X}_n + \frac{\sigma}{\sqrt{n}} z_{1-\alpha/2} \text{ [adding } \bar{X}_n \text{ to all sides]}$$



# Deriving Interval Estimators

$\sigma$  known

So this means, writing

$$\mathbb{P}(z_{\alpha/2} \leq Z_n \leq z_{1-\alpha/2}) = 1 - \alpha$$

is same as

$$\mathbb{P}\left(\bar{X}_n - \frac{\sigma}{\sqrt{n}} z_{1-\alpha/2} \leq \mu \leq \bar{X}_n + \frac{\sigma}{\sqrt{n}} z_{1-\alpha/2}\right) = 1 - \alpha$$

So we have found our *upper and lower confidence limits*, these are

$$L = \bar{X}_n - \frac{\sigma}{\sqrt{n}} z_{1-\alpha/2} \text{ and } U = \bar{X}_n + \frac{\sigma}{\sqrt{n}} z_{1-\alpha/2}$$

So the interval estimator is

$$\left[\bar{X}_n - \frac{\sigma}{\sqrt{n}} z_{1-\alpha/2} \quad , \quad \bar{X}_n + \frac{\sigma}{\sqrt{n}} z_{1-\alpha/2}\right]$$

Now if we calculate this for a fixed sample we will call it an *interval estimate* which will be

$$\left[\bar{x} - \frac{\sigma}{\sqrt{n}} z_{1-\alpha/2} \quad , \quad \bar{x} + \frac{\sigma}{\sqrt{n}} z_{1-\alpha/2}\right]$$

For an interval estimate, there is no probabilistic interpretation.

But for the interval estimator, can think about the frequency interpretation of probability, that is, *if we do repeated sampling 100 times, then 95 out 100 times the intervals that we constructed will contain the true parameter  $\mu$*

## Interval Estimation of Population Mean $\mu$

Interval Estimator -  $\sigma$  unknown case

# Interval Estimators

$\sigma$  unknown

- ▶ Can we construct intervals when we do not know the population standard deviation  $\sigma$ . The answer is YES!
- ▶ We need to use the statistic  $T_n$  from the Theorem 1.3 (v). This is called *t-statistic*, on the other hand when we used  $Z_n$ , that is called *z-statistic*.
- ▶ Note that in this case the statistic  $T_n$  follows a new sampling distribution, it is called *t-distribution, with parameter  $n - 1$  (where  $n$  is the sample size!), there is a special name of this parameter, it is called degrees of freedom*.
- ▶ The idea is if we use the sample standard deviation  $S$ , which is possible to calculate using the sample. Then we get a new Statistic  $T_n$ , which is distributed with *t-distribution* with  $n - 1$  degrees of freedom (Again to emphasize, degrees of freedom (df) is a parameter for the *t-distribution*).

# Interval Estimators

$\sigma$  unknown

- $T_n$  or the  $t$ -statistic is written as,

$$T_n = \frac{\bar{X}_n - \mu}{\frac{S}{\sqrt{n}}} \quad (11)$$

- And according to the Theorem 1.3 (v), we have

$$T_n \sim t_{(n-1)}$$

- This means  $T_n$  is distributed with  $t$  distribution with parameter  $(n-1)$ , or degrees of freedom  $(n-1)$ .
- Note that in (11)  $S$  is the sample standard deviation, Recall  $S^2$  is

$$S^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X}_n)^2 \text{ and } S = \sqrt{S^2}$$

# Interval Estimators

$\sigma$  unknown

- Now how do we get an interval estimator in this case? The steps are actually same as page 28, except now you need to use quantile from  $t$  distribution with parameter  $n - 1$ .
- If you do, then you should get the following *interval estimator* using  $t_{n-1}$  distribution,


$$\left[ \bar{X}_n - \frac{s}{\sqrt{n}} t_{1-\alpha/2} \quad , \quad \bar{X}_n + \frac{s}{\sqrt{n}} t_{1-\alpha/2} \right]$$

- The *interval estimate* in this case is,

$$\left[ \bar{x} - \frac{s}{\sqrt{n}} t_{1-\alpha/2} \quad , \quad \bar{x} + \frac{s}{\sqrt{n}} t_{1-\alpha/2} \right] \quad (12)$$

# Interval Estimators

$\sigma$  unknown

- ▶ Let's see a concrete example.
- ▶ Suppose in Example 1.5, we don't know  $\sigma$ , rather we know sample standard deviation  $s = 18.5$ .
- ▶ This means everything is same except the information of  $\sigma$  is not known to us,
  - ▶  $n = 100$
  - ▶  $\bar{x} = 82$
  - ▶  $\alpha = 0.05$  (this is because we are asked to construct 90% confidence interval)
  - ▶  $s = 18.5$
- ▶ Now we will do the calculation in 
- ▶ You will see following important differences compared to  $\sigma$  known case.
  - ▶ As mentioned we need to use  $t_{n-1}$  distribution
  - ▶ We need to use  $s$ , which is the sample standard deviation
  - ▶ Because we don't know  $\sigma$ , we cannot calculate the standard error  $\sigma/\sqrt{n}$ , however we can calculate the *estimate of the standard error*, which is  $\frac{s}{\sqrt{n}}$

# Interval Estimators

$\sigma$  unknown

## code - sigma unknown (confidence interval)

```
# First create some objects with the information given
n <- 100
xbar <- 82
alpha <- 0.05
s <- 18.5

# calculate the estimate of the sderror and moe and save them as objects
sderror_est <- s/sqrt(n)
moe <- qt(1 - alpha/2, n-1) * sderror_est

# upper limit
xbar + moe
# [1] 85.6708

# lower limit
xbar - moe
# [1] 78.3292
```

► So the 95% interval estimate or the confidence interval in this case is (78.33 , 85.67)

## Interval Estimation of Population Mean $\mu$

Interval Estimator -  $\sigma$  unknown case with large samples



## Large Sample results for $t$ statistic

- ▶ There is one last important result before we go to the next section.
- ▶ Recall, from the last section we learned that, when we use  $S$ , rather than  $\sigma$ , we get a new statistic, that is what we called  *$t$  statistic*,

$$T_n = \frac{\bar{X}_n - \mu}{\frac{S}{\sqrt{n}}}$$

- ▶ And we already learned that this statistic is distributed with  $t$  distribution.
- ▶ Now there is a very interesting result. Have a look at the result in Theorem 1.4 (iii), this says if we have a very large sample, then

$$T_n \overset{\text{approx}}{\sim} \mathcal{N}(0, 1)$$

- ▶ This means for very large  $n$ , the  $t$  statistic is approximately normally distributed!
- ▶ This means, if the sample size  $n$  is large then, we can forget about anything called  $t$  distribution, and just use the normal distribution with  $t$  statistic.

## Large Sample results for $t$ statistic

- ▶ What's the implication of this result for our confidence interval construction?
- ▶ The answer is, for large sample size we can construct the confidence interval in the following way,

$$\left[ \bar{x} - \frac{s}{\sqrt{n}} z_{1-\alpha/2} \quad , \quad \bar{x} + \frac{s}{\sqrt{n}} z_{1-\alpha/2} \right] \quad (13)$$

- ▶ Now if you compare (12) and (13), you will understand the difference.
- ▶ What's the difference?
- ▶ Again why does this happen?

# References

- Anderson, D. R., Sweeney, D. J., Williams, T. A., Camm, J. D., Cochran, J. J., Fry, M. J. and Ohlmann, J. W. (2020), *Statistics for Business & Economics*, 14th edn, Cengage, Boston, MA.
- Bertsekas, D. and Tsitsiklis, J. N. (2008), *Introduction to probability*, 2nd edn, Athena Scientific.
- Blitzstein, J. K. and Hwang, J. (2015), *Introduction to Probability*.
- Casella, G. and Berger, R. L. (2002), *Statistical Inference*, 2nd edn, Thomson Learning, Australia ; Pacific Grove, CA.
- DeGroot, M. H. and Schervish, M. J. (2012), *Probability and Statistics*, 4th edn, Addison-Wesley, Boston.
- Hansen, B. (2022), *Econometrics*, Princeton University Press, Princeton.
- Newbold, P., Carlson, W. L. and Thorne, B. M. (2020), *Statistics for Business and Economics*, 9th, global edn, Pearson, Harlow, England.
- Pishro-Nik, H. (2016), *Introduction to probability, statistics, and random processes*.
- Ramachandran, K. M. and Tsokos, C. P. (2020), *Mathematical Statistics with Applications in R*, 3rd edn, Elsevier, Philadelphia.
- Rice, J. A. (2007), *Mathematical Statistics and Data Analysis*, Duxbury advanced series, 3rd edn, Thomson/Brooks/Cole, Belmont, CA.