Ch1 - Estimation

(*Point and Interval Estimation)*

ECO 204
Statistics For Business and Economics - II

Shaikh Tanvir Hossain

East **W**est **U**niversity, Dhaka
Last updated: July 7, 2025

Outline

Descriptive vs Inferential Statistics

Motivating Picture

...a picture may be worth a thousand words.. - [Djikstra]

Consider following picture,

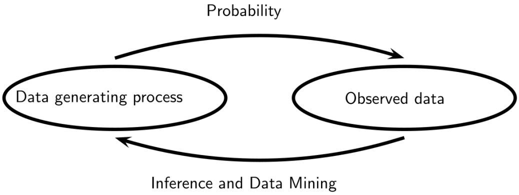


Figure 1: The figure is directly taken from Wasserman (2013), clearly explains what you did in ECO 104 (indicated by the arrow at the top going to right direction) and what you will do in ECO 204 (indicated by the arrow at the bottom going to left).

The work of *Probability Theory* is to describe - how the data / sample has been generated from a population, and the work of *Statistics* (or in particular Statistical Inference) is to make conclusions about the population using a sample.

Introduction to Statistical Inference

- Welcome to ECO 204!
- ECO 204 is about *Inferential Statistics* (as opposed to *Descriptive Statistics* which you did in ECO 104, as a side note - ECO 104 was about two things - i) Descriptive Statistics and ii) Probability Theory)
- Any idea about *Inferential Statistics*...?
- Roughly, It's is a branch of Statistics that helps us to *make conclusions about the population using a sample*
- Now couple of questions from ECO 104
  - What is *the population* for any particular study,
  - What is *a sample*, and
  - How do you make *Inference* using example?
- Let's see one example and answer these questions systematically, suppose we have the following data from 5 students studying currently at EWU (this is a hypothetical data, perhaps *randomly* collected!).

	Gender	Monthly Family Income (tk)
1.	Male	70,150
2.	Female	20,755
3.	Male	44,758
4.	Female	38,790
5.	Male	20,579

- You should already know that the columns are called *variables* and the rows are called *observations*. Let me stat with some questions.

Important Questions

Suppose our goal is to understand the Income and Gender of the current students at EWU, then consider following questions,

- Q1: What is the population of the study?
- Q2: Currently what proportion of students are female at EWU? Is it possible to calculate this?
- Q3: Currently what is the average income of all students at EWU? Is it possible to calculate this?
- Q4: Can we **estimate** proportion of students who are female at EWU? If yes then how?
- Q5: Can we **estimate** average monthly family income all students at EWU? If yes then how?
- Q6: What's the difference between sample and population quantities?
- Q7: If I have a population data, then *do I need a sample*? Why would I go for a sample?
- Q8: What are the possible issues in a sample?
- Q9: Are the sample quantities fixed or random?
- Q10: Do our predictions improve if we collect more samples?

Here are some answers,

- Ans 1: The population is the set of all units in the study, so in this case the population is *all currently enrolled EWU students*.
- Ans 2: We don't know this proportion exactly unless we have the data from all the students who are currently studying at EWU. In other words, we need data from the full population. Usually getting full population is very hard or time consuming. Although this is the quantity we are after, in other words this is one of our targets, but still we cannot calculate this exactly!
- Ans 3: Again the reasoning is same as Ans 2. Since we don't have the full population it's almost impossible to get the average income of all the current students.
- Ans 4: Yes, by using sample we can calculate **sample proportion** and then this sample proportion can work as an **estimate** for the **population proportion** of female students. In this sample proportion of female is 40%. So we can say, roughly, given that our sample represents the population, it is possible that the population proportion is close to 40%. Again, always remember here *population proportion* is for the entire population, and usually this is impossible to calculate. This is the first example of Statistical Inference, in particular we call this Statistical Estimation (more on this later)
- Ans 5: The answer is similar to the last answer. Yes we can estimate the **population mean** using *sample mean* of the students in the sample, in this case the sample mean is $39,006.4$ taka and we can take this **sample mean** as an **estimate** of the *population mean*. Again to make it clear here *population mean* is the average of the income of all the current students. This is the second example of Statistical Inference and in particular Statistical Estimation.

- Ans 6: We already answered this, but let's make it concrete, **sample quantities are estimates of the population quantities**. Population quantities are always our targets. Definitely chances are very low that they will be exactly same, however with a "good" sample we might be close to our target.

- Ans 7: NO, I have all information I need, so no need for sampling. I go for sampling because usually I don't have access to the population.

- Ans 8: Obvious issue - *biased sample* (sample doesn't represent population properly). Another issue - *small sample size*. From now on, we will avoid issues related to "biased sample", and assume our sample is a fairly good representative of the population....but ...?

- Ans 9: Definitely random since if we have a different sample then sample proportion or sample mean both will change.

- Ans 10: Of course ....

Introduction to Statistical Inference

- In the last example, we used a sample to calculate some sample quantities, in fact we calculated, *sample proportion* of female students in the sample, and *sample mean of family income*, and then we argued that we can use these objects to *"predict"*, or or *"conclude"* or *"infer"* about the unknown population quantities.
- This was an example of *Inferential Statistics* or *Statistical Inference*, in particular this is what we call *Statistical Estimation!*. The formal definition of *Statistical Inference* would require us to carefully define many things, but perhaps informally we can say -

Definition 1.1: Statistical Inference

Statistical Inference is a procedure where we have a **target parameter** θ defined for the population, and then we use a sample to make conclusions regarding the target parameter. There are two key branches of Statistical Inference,

- *Statistical Estimation (in short Estimation)*
- *Statistical Hypothesis Testing (in short Testing)*

We can also categorize estimation as *point estimation* and *interval estimation*. Typical examples of Statistical Inference include making conclusion about the population mean, population proportion, population variance, etc., using sample mean, sample proportion, sample variance.

We will **generally use** the Greek letter θ to represent the target parameter. In the examples above,

- for the first one θ is the population proportion. If we write population proportion with p, then in this case $\theta = p$
- for the second one θ is the population mean. If we write population mean with μ, then in this case $\theta = \mu$

From the next section, first we will focus on **Estimation** and then in the next chapter we will move to **Testing**. In both cases, we will always start with the some *numerical techniques* that you have already learned before, for example, sample mean $\overline{x}$, sample proportion $\bar{p}$, etc. But our goal is one step more - that is making *inference* about the population.

Introduction to Statistical Inference

- The practice of Statistics falls broadly into two categories *Descriptive and Inferential*
- Descriptive Statistics is about describing the data using both numerical and graphical techniques / methods,
  - **Numerical Methods:** Calculating Sample Mean, Median, Mode, Variance, Standard Deviation, etc.
  - **Graphical Methods:** Looking at Bar Charts, Pie Charts, Histograms, etc
- The goal in this case is to describe the data, not to make any inference (or predict) about the population. This is what you did in ECO 104. You will do some recap exercises in the first problem set (which I will have to prepare, sorry not done yet!)

Introduction to Statistical Inference

- However Inferential Statistics goes one step more, we try to make *"good" conclusions about the population using a sample*.
- There are two major themes of Inferential Statistics,
  - Statistical Estimation, in short we say *Estimation*
  - Hypothesis Testing, in short we say *Testing*
- You have already seen two examples of Estimation, we will see more. First we will focus on Estimation and then we will move to Hypothesis Testing.
- In both cases we will almost always start with the same *numerical techniques* that you have already learned in ECO104, that is we will use
  - sample mean,
  - sample median,
  - sample mode,
  - sample variance or sample standard deviation
  - sample quantiles or percentiles,
  - sample covariance, or sample correlation etc,

  but our goal is in this course will be one step more - that is making *inference* about the population.

  In the next section we will talk about *Estimation*.

Inferential Statistics - Part I, Estimation

Estimation, Estimate and Estimator

Estimator

We have already seen two examples of *Statistical Estimation*, in particular we saw

- we can use *sample proportion* $\bar{p}$ to estimate the unknown population proportion p.
- also, we can use *sample mean* $\bar{x}$ to estimate the unknown population mean μ.

Below we clearly define Statistical Estimation, in particular we will define what is *the target parameter, an estimate and an estimator, point estimation and interval estimation*,

Estimator

Definition 1.2: Statistical Estimation

Statistical Estimation is a Statistical Inferential procedure which assigns numerical values to the unknown population parameters θ (e.g. mean μ, proportion p, variance σ^2) using data from a **random sample** $X_1, X_2, \ldots, X_n$. There are two types of Statistical Estimation,

- **Point estimation:** Provide a single best-guess number $\hat{\theta}$ for θ.
- **Interval estimation:** Provide an interval of possible values that, in repeated sampling, contains θ with a specified coverage probability (e.g. a 95 % confidence interval).

Definition 1.3: Some Important Objects to Remember

- Any function of the random sample $(X_1, \ldots, X_n)$ is called a *Statistic*.
- When a statistic is used to infer a parameter it is called an *Estimator*.
- Both Statistic and Estimator are random variables, and their values change from sample to sample.
- For a fixed sample the value of the estimator is called an *Estimate*.
- The probability distribution of a Statistic (or an estimator) is called *Sampling Distribution*. Sampling distribution is a repeated sampling idea
- The Standard Deviation of any Statistic is called is called *Standard Error*.

There are several things to understand in the definition,

Estimator

Key Terms in the Definition

- Q1. What is a **Target Parameter** θ?
- Q2. What is a **Random Sample** $(X_1, \ldots, X_n)$?
- Q3. What is a **Statistic**?
- Q4. What is an **Estimator**?
- Q5. What is an **Estimate**?
- Q6. What is a **Sampling distribution of a Statistic / Estimator**?
- Q7. What is **Point Estimation** and **Interval Estimation**.

Again we will answer these questions and also understand the definition using an example. Consider the following sample, this is the same sample as above but just with one variable that is - Monthly family Income

sl.	Monthly Family Income (tk)	R.V.
1.	70,150	$X_1 = ?$
2.	20,755	$X_2 = ?$
3.	44,758	$X_3 = ?$
4.	38,790	$X_4 = ?$
5.	20,579	$X_5 = ?$

Remarks on Notation: Usually for fixed numbers we will use lower case letters $x_1, x_2, x_3, \ldots, x_n$, rather than numbers, just to make it more general...and for random variables we use upper case letters $X_1, X_2, X_3, \ldots, X_n$. Also generally when we think about n random variables, we write $X_1, X_2, X_3, \ldots, X_n$, and similarly for n fixed numbers we write $x_1, x_2, x_3, \ldots, x_n$.

Estimator

- **Ans 1 :** In this case the target parameter θ can be population mean of all current students at EWU. Since it's a mean in this case we often use μ for the target parameter.

Estimator

- **Ans 2 :** The idea of the random sample is when we think each row as a *random data point*. How this is random? The idea is before sampling it is possible to have different values in each row. So we can think we have a random variable $X_1 =?$ at first row, X_2 at the second row and so on... so in principle each row can take different values from the population, and each value before sampling is a random variable. Also note after the sampling, when we have **observed the sample**, a random data becomes a fixed number and in this case the random variable X_1 has already taken a value, so we get $X_1 = 70,150$, and similarly for $X_2 = 20,755$. Here we don't have any randomness, we call it *observed data or realized data*. Important, *there is no randomness after we have observed the sample!*. So in the sample we have 5 random variables, they are X_1, X_2, X_3, X_4, X_5, together we call it a *random sample*. And also we have 5 fixed numbers, and they are $70,150$; $20,755$; $44,758$; $38,790$ and $20,579$, together we call it an *observed sample* or a *realized sample*.

Estimator

- **Ans 3:** A statistic is simply a function of the random sample, for example following are examples of statistic,
  - $\sum_{i=1}^{n} X_i$
  - $\overline{X} = \frac{1}{n}(X_1 + X_2 + \ldots, X_n) = \frac{1}{n}\sum_{i=1}^{n} X_i$
  - $\frac{\overline{X}-\mu}{\sigma/\sqrt{n}}$, where μ and σ are simply constants, μ is population mean and σ is population standard deviation
  - $\frac{\overline{X}-\mu}{S/\sqrt{n}}$ where S is the sample standard deviation and $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2$ and $S = \sqrt{S^2}$

  In a nutshell a statistic is any quantity made from the random sample $X_1, X_2, \ldots, X_n$
- **Ans 4:** When a Statistic is used for Estimation, we call it an **Estimator**. If Population mean μ is the target quantity, we can use $\overline{X}$ to calculate random sample mean. In this case $\overline{X}$ is a statistic and also it's an estimator.
- **Ans 5:** If we have a fixed sample then the value of the estimator becomes fixed, and on that case the numerical value of the estimator is called an **estimate**. For example if μ (or the population mean) is the target parameter then $\overline{X}$ is the estimator (this is the random sample mean) and for fixed sample we will get a value of $\overline{X}$ which will be an **estimate**. For example for this particular sample the estimate is $\bar{X} = 39,006$ taka. Again an estimate is a fixed number for a specific sample and an estimator is a random variable.

Estimator

- **Ans 6:** Now since **a statistic** or **an estimator** are random variables and their values will be different from sample to sample, it will have many possible values. The probability distribution over these values is called **sampling distribution**. For a concrete example, think about $\overline{X}$, this is an estimator when the target parameter is μ, and this will have many possible values for different samples. So we can think about a probability distribution over the values of $\overline{X}$, and this is the sampling distribution of $\overline{X}$. We will talk about sampling distribution detail in the next section, but there is a theorem which says if population is normally distributed with mean μ and variane σ^2, then the distribution of $\overline{X}$ is also normal with mean μ and variance σ^2/n. Using notation we write this as $\overline{X} \sim \mathcal{N}(\mu, \sigma^2/n)$. Again sampling distribution of sample mean $\overline{X}$ means distribution of different sample means, you should keep the shown picture (on the left) in your mind.

- **Ans 7:** When the parameter is μ, point estimation means we propose one best value, for example in this case sample mean $\bar{x}$ for a fixed sample. On the other hand interval estimation means we propose an interval of possible values with certain probability, such that in repeated sampling the parameter μ will be inside the interval. We will talk about interval estimation in the coming sections. But before that we need to discuss about sampling distribution of sample means.

Inferential Statistics - Part I, Estimation

Sampling Distribution of Sample Means (Assuming Normality)

Sampling Distribution of Sample Means With Normality Assumption

- Now when it comes to random variable $\overline{X}$, one of the most important question is,

  What is the probability distribution of $\overline{X}$ in repeated sampling?

- The answer to this question is not straightforward and it depends on probability distribution of the population data. In particular we can think about two cases,
  - **Case 1: Population is Normally distributed:** This means we are assuming the population data is normally distributed some unknown mean μ and naturally variance σ^2
  - **Case 2: Population has some unknown probability distribution:** This means we don't know the distribution of the population data (may or may not be normal).
- **I.I.D. Assumption:** In both cases we will always assume we have an i.i.d. random sample. This means two things
  - The random variables in the random sample are all distributed with same distribution, which means $X_1, X_2, \ldots, X_n$ *are all identically distributed*, e.g., all are normally distributed with same mean μ and variance σ^2.
  - And $X_1, X_2, \ldots, X_n$ *are independent random variables*, this means in our sample each row is independently sampled from other.
- This i.i.d. assumption is often a standard assumption in cross section data. However in time series data the independence assumption breaks down!

Sampling Distribution of Sample Means With Normality Assumption

- Below we write the results of the two cases separately.
- The *first case* is actually quite restrictive since in practice we never know whether the probability distribution of the population is normal or not.
- But we will see that there are some strong features of the results from here, that is it can be applied to any sample size.
- The *second case* is more practical, since in practice we never the probability distribution of the population. However there is an issue, we can only use the results from this case when we have a *large sample size*.
- Here is the result for Case - 1, in the next section we will see the results for the case 2.

Theorem 1.4: Sampling Distribution of $\overline{X}$ with Normality and I.I.D. Assumption

If the population data is Normally distributed with mean μ and variance σ^2. Note that this further means the random variables $X_1, X_2, \ldots, X_n$ are all Normal with same mean μ and variance σ^2 (this means *for all i we have $X_i = \mathcal{N}(\mu, \sigma^2)$*), and moreover they are all independent, then we can prove that,

$$\overline{X} \sim \mathcal{N}(\mu, \frac{\sigma^2}{n}) \tag{1}$$

Sampling Distribution of Sample Means With Normality Assumption

Proof.

When it comes to theorems we need to prove it. But for this one I will skip the proof, since this requires materials which are beyond the scope of this course! So you can also skip it. But at some point I will give the proof in the Appendix. □

In Words: The theorem means *if the population data is Normally distributed and the random sample have i.i.d. random variables, then the sampling distribution of $\overline{X}$ is also normal with mean μ and variance σ^2/n, where n is the sample size.*

Standard Error: As we already mentioned the standard deviation of any Statistic and Estimator is called **Standard Error**. In this case the standard deviation of the sampling distribution of $\overline{X}$ is called **Standard Error of the Sampling Distribution** and we write it as $\sqrt{\mathbb{V}(\overline{X})}$, where $\mathbb{V}(\overline{X})$ is the variance of the sampling distribution of $\overline{X}$.

So **Standard Error** of the sampling distribution is, $\sigma/\sqrt{n}$, If we replace σ with S, then we call it *estimate (or estimator in repeated sampling) of the Standard Error*, which is, $S/\sqrt{n}$

Sampling Distribution of Sample Means With Normality Assumption

- This theorem has some immediate consequences, in *Math the consequence are called corollaries*, in particular we can write following corollary from the theorem.

Corollary 1.5: More Results Assuming Normality

Following the last theorem, we can also get following results if the assumptions of the theorem holds,

$$i) \quad \mathbb{E}\left(\overline{X}\right) = \mu \tag{2}$$

$$ii) \quad \mathbb{V}\left(\overline{X}\right) = \frac{\sigma^2}{n} \tag{3}$$

$$iii) \quad Z \sim \mathcal{N}(0,1) \text{ where } Z = \frac{\overline{X} - \mu}{\frac{\sigma}{\sqrt{n}}} \tag{4}$$

$$iv) \quad T \sim t_{n-1} \tag{5}$$

$$\text{where } T = \frac{\overline{X} - \mu}{\frac{S}{\sqrt{n}}} \text{ and } S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2 \tag{6}$$

Sampling Distribution of Sample Means With Normality Assumption

Proof.

Again we will skip the proof for now □

- *In Words:* Number i) and ii) says that *the expectation of the sample mean $\overline{X}$ (i.e., the average of the averages) will be equal to the population mean μ, i.e., $\mathbb{E}(\overline{X}) = \mu$, and the variance of the sample mean is $\mathbb{V}(\overline{X}) = \sigma^2/n$.*

- Number (iii) is a direct consequence of the relationship between standard normal and normal distributions in general. Note that here

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

is actually *a Statistic*, so in this sense this distribution of Z is also sampling distribution.

Sampling Distribution of Sample Means With Normality Assumption

- The last one is a result for T statistic and *t distribution*, here $n-1$ is the parameter of the distribution, we call it *degrees of freedom*. Let's talk about the T statistic. First of all, note that T is a Statistic and a random variable, since it's a function of $\overline{X}$, in particular we have,

$$T = \frac{\overline{X} - \mu}{\frac{S}{\sqrt{n}}}$$

  number iv) says the distribution of this Statistic is a new distribution which is called t distribution. We write t_{n-1}, to write the parameter of the distribution which is in this case is $n-1$ and the name of this parameter is degrees of freedom.

- Interestingly there is a common pattern for these two statistics, that is

$$Z = \frac{\overline{X} - \mathbb{E}(\overline{X})}{\sqrt{\mathbb{V}(\overline{X})}} \text{ and } T = \frac{\overline{X} - \mathbb{E}(\overline{X})}{\sqrt{\widehat{\mathbb{V}}(\overline{X})}}$$

Inferential Statistics - Part I, Estimation

Interval Estimation Assuming Normality

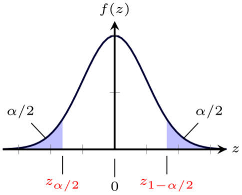
Interval Estimation With Normality Assumption (σ-known)

- Now we have the sampling distribution of sample means $\overline{X}$, and we can use this to construct *confidence intervals* for the population mean μ where the normality assumption is there. Let's see how do we get the formula for the confidence interval,
- Again recall,

$$\overline{X} \sim \mathcal{N}(\mu, \frac{\sigma^2}{n}) \quad \text{equivalently} \quad Z \sim \mathcal{N}(0,1) \text{ where } Z = \frac{\overline{X} - \mu}{\frac{\sigma}{\sqrt{n}}}$$

- Now we start from the Z statistic, and write

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



Interval Estimation With Normality Assumption (σ-known)

- Where α is any probability $0 < \alpha < 1$, and $z_{\alpha/2}$ and $z_{1-\alpha/2}$ are values such that $\alpha/2$ and $1-\alpha/2$ are the probabilities to the left of $z_{\alpha/2}$ and $z_{1-\alpha/2}$ respectively (these are actually quantiles, if you know what is a quantile...)
- Let's say, $\alpha = 0.05$, or 5%, then means if we randomly pick 100 values of Z from the distribution of Z (which is standard normal), then 95 of the values will be in between $z_{\alpha/2}$ and $z_{1-\alpha/2}$.
- We can also re-arrange this and write (try to get this)

$$\mathbb{P}\left(\mu - z_{\alpha/2} \leq \overline{X} \leq \mu + z_{1-\alpha/2}\right) = 1 - \alpha \tag{7}$$

- This means if we *randomly sample* 100 *times* and get 100 sample means then *95 the sample means will be in between $\mu - z_{\alpha/2}$ and $\mu + z_{1-\alpha/2}$*.
- When we construct confidence interval we actually make the interval random, in particular we will derive the following,

$$\mathbb{P}\left(\overline{X} - \frac{\sigma}{\sqrt{n}}\, z_{1-\alpha/2} \leq \mu \leq \overline{X} + \frac{\sigma}{\sqrt{n}}\, z_{1-\alpha/2}\right) = 1 - \alpha \tag{8}$$

- There is a big difference between (7) and (8), the intervals are fixed, what falls inside are random. But in the second one the intervals are random and what falls inside is fixed, which is μ.

Interval Estimation With Normality Assumption (σ-known)

- This gives us the formula for the interval estimation, in particular we write

$$\text{lower limit} = \overline{X} - \frac{\sigma}{\sqrt{n}} z_{1-\alpha/2} \text{ and upper limit} = \overline{X} + \frac{\sigma}{\sqrt{n}}\, z_{1-\alpha/2}$$

Interval Estimation With Normality Assumption (σ-known)

- Let's see how do we get the formula,
- We start with Z statistic and do some algebra such that μ comes in between

$$
\begin{aligned}
z_{\alpha/2} \le Z \le z_{1-\alpha/2} &= -z_{1-\alpha/2} \le Z \le z_{1-\alpha/2} \text{ [using symmetry of the normal]} \\
&= -z_{1-\alpha/2} \le \frac{\overline{X} - \mu}{\frac{\sigma}{\sqrt{n}}} \le z_{1-\alpha/2} \\
&= -\frac{\sigma}{\sqrt{n}}\, z_{1-\alpha/2} \le \overline{X} - \mu \le \frac{\sigma}{\sqrt{n}}\, z_{1-\alpha/2} \text{ [multiplying all sides by } \sigma/\sqrt{n}\text{]} \\
&= \frac{\sigma}{\sqrt{n}}\, z_{1-\alpha/2} \ge -\overline{X} + \mu \ge -\frac{\sigma}{\sqrt{n}}\, z_{1-\alpha/2} \text{ [multiplying all sides by } -1\text{]} \\
&= -\frac{\sigma}{\sqrt{n}}\, z_{1-\alpha/2} \le -\overline{X} + \mu \le \frac{\sigma}{\sqrt{n}}\, z_{1-\alpha/2} \text{ [rewriting the inequalities]} \\
&= \overline{X} - \frac{\sigma}{\sqrt{n}}\, z_{1-\alpha/2} \le \mu \le \overline{X} + \frac{\sigma}{\sqrt{n}}\, z_{1-\alpha/2} \text{ [adding } \overline{X} \text{ to all sides]}
\end{aligned}
$$

- This gives us following probability

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

Interval Estimation With Normality Assumption (σ-known)

- And this is the formula for the interval estimation, in particular,

$$\text{lower limit} = \overline{X} - \frac{\sigma}{\sqrt{n}} z_{1-\alpha/2} \text{ and upper limit} = \overline{X} + \frac{\sigma}{\sqrt{n}}\ z_{1-\alpha/2}$$

  So the *interval estimator* is

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

  where $\overline{X}$ is a random quantity,

  and the *interval estimate* is

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

  where $\overline{x}$ is a fixed number from a specific sample.

Interval Estimation With Normality Assumption (σ-known)

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

Example 1.6: (Interval Estimation With Normality Assumption)
Suppose we have calculated a sample mean of a sample of size $n = 100$, and the sample mean $\overline{x} = 82$. Population data is normally distributed with population standard deviation $\sigma = 20$. Calculate the 95% *confidence interval estimate of the population mean μ*,

Solution:
We will simply apply the formula,

$$\overline{x} \pm \frac{\sigma}{\sqrt{n}} z_{1-\alpha/2}$$

Here we have

$$\begin{aligned} \bar{x} &= 82, \\ \sigma &= 20, \\ n &= 100, \\ z_{1-\alpha/2} = z_{0.975} &= 1.96 \quad \text{(this is } 1-\alpha/2 \text{ quantile of the standard normal distribution)} \end{aligned}$$

We can calculate $z_{1-\alpha/2} = z_{0.975} = 1.96$ using Excel with the formula `NORM.INV(.975, 0, 1)` and R function `qnorm(.975)`).

Interval Estimation With Normality Assumption (σ-known)

the interval estimate is

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

and *the interval estimator is*

$$[\overline{X} - 3.92, \quad \overline{X} + 3.92] \tag{10}$$

Interval Estimation With Normality Assumption (σ-known)

Interpretation:

Question: Why we call this $(1-\alpha)\times 100\%$ confidence Interval estimate?In particular in our last example, we have $\alpha = 0.05$ then why we call this a 95% confidence interval estimate, what is the meaning of 95%?

First note, if we calculate the interval a fixed sample we will call it an *interval estimate* which will be

$$[\bar{x} - \frac{\sigma}{\sqrt{n}} z_{1-\alpha/2} \quad , \quad \bar{x} + \frac{\sigma}{\sqrt{n}}\ z_{1-\alpha/2}] = [78, \quad 85.92]$$

For an interval estimate, there is no probabilistic interpretation. This is a fixed interval and our target parameter μ is either inside the interval or not. So probability is either 0 or 1.

But for the *interval estimator*, which is

$$[\overline{X} - 3.92, \quad \overline{X} + 3.92] \tag{11}$$

we can think about the probability, since $\overline{X}$ is random. Recall probability is coming from sampling distribution, and sampling distribution is a repeated sampling idea. So we can interpret the random interval as following

if we do repeated sampling 100 times, and construct intervals with $[\overline{X} - 3.92, \quad \overline{X} + 3.92]$ for 100 times, then 95 out 100 times the intervals will contain the true parameter μ and in the remaining 5 times it won't.

Interval Estimation With Normality Assumption (σ-known)

- **Important:** If you say that there is a 95% probability that true parameter μ will fall inside $[78, \quad 85.92]$, this is a *wrong interpretation*. We can say "for this particular sample, *the interval estimate* is $[78, 85.92]$". Again this is a fixed interval and our target parameter μ is either inside the interval or not. So probability is either 0 or 1.
- *A Side Note:* Note that when we constructed the interval estimate, we added and subtracted the following same number with $\overline{x}$

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

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

Interval Estimation With Normality Assumption (σ-unknown)

- So in the last example we assumed that we know the population variance σ, and the formula for the confidence interval estimate is

$$\overline{x} \pm z_{1-\alpha/2} \frac{\sigma}{\sqrt{n}}$$

- There is an issue with this formula, that is it is strange to assume that we know the population variance σ^2 (or standard deviation σ).
- There is a solution to this problem. If we have a sample we can always calculate S^2 which is the **sample variance** and S is the **sample standard deviation**, then we can find the interval estimate with the following formula

$$\overline{x} \pm t_{1-\alpha/2,n-1} \frac{s}{\sqrt{n}}$$

where $t_{1-\alpha/2,n-1}$ is the t value coming from the t_{n-1} distribution. Like standard normal $Z \sim \mathcal{N}(0,1)$, t_{n-1} is also a probability distribution, and it looks very similar to normal distribution,

Interval Estimation With Normality Assumption (σ-unknown)

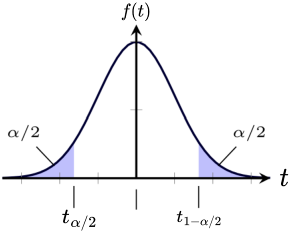


Figure 2: t distribution with $n-1$ degrees of freedom

Normal distribution has two parameters, mean μ and variance σ^2. The t distribution has a single parameter which is $n-1$, to specify the parameter we write t_{n-1}.

You might be wondering - how is the t distribution coming... answer is ... this is also coming via repeated sampling, with fixed sample of size n.

Interval Estimation With Normality Assumption (σ-unknown)

We can derive the confidence interval formula using the similar technique, but in this case we need to use T statistic as opposed to Z statistic and the result from the corollary.

Remember T statistic is following,

$$T = \frac{\overline{X} - \mu}{\frac{S}{\sqrt{n}}}$$

and we have the result

$$T \sim t_{n-1}$$

So now we can derive in a similar fashion (do this)

$$\mathbb{P}\Big(t_{n-1,\alpha/2} \leq T \leq t_{n-1,1-\alpha/2}\Big) = 1 - \alpha$$

$$\mathbb{P}\Big(t_{n-1,\alpha/2} \leq \frac{\overline{X} - \mu}{\frac{S}{\sqrt{n}}} \leq t_{n-1,1-\alpha/2}\Big) = 1 - \alpha$$

$$\vdots$$

$$\mathbb{P}\Big(\overline{X} - \frac{S}{\sqrt{n}}\, t_{n-1,1-\alpha/2} \leq \mu \leq \overline{X} + \frac{S}{\sqrt{n}}\, t_{n-1,1-\alpha/2}\Big) = 1 - \alpha$$

Interval Estimation With Normality Assumption (σ-unknown)

The last probability is calculated using the t distribution. Now this gives us the desired formula for the confidence interval estimate,

$$\text{lower limit} = \overline{x} - \frac{s}{\sqrt{n}}\, t_{1-\alpha/2,n-1} \text{ and upper limit} = \overline{x} + \frac{s}{\sqrt{n}}\, t_{1-\alpha/2,n-1}$$

and the interval estimator is

$$[\overline{X} - \frac{S}{\sqrt{n}}\, t_{1-\alpha/2,n-1} \quad , \quad \overline{X} + \frac{S}{\sqrt{n}}\, t_{1-\alpha/2,n-1}]$$

Interval Estimation With Normality Assumption (σ-unknown)

We can now modify the last example,
Example 1.7: (Interval Estimation With Normality Assumption - unknown σ)
Suppose we have calculated a sample mean of a sample of size $n = 100$, and the sample mean $\overline{x} = 82$ and sample standard deviation is $s = 18$. Population data is normally distributed. Calculate the 95% *confidence interval estimate of the population mean μ*,

Solution:
We will simply apply the formula,

$$\overline{x} \pm \frac{s}{\sqrt{n}} t_{1-\alpha/2,n-1}$$

Here we have

$$\begin{aligned} \bar{x} &= 82, \\ s &= 18, \\ n &= 100, \\ t_{n-1,1-\alpha/2} &= t_{99,\,0.975} = 1.98 \quad \text{(this is } 1-\alpha/2 \text{ quantile of the } t_{n-1} \text{ distribution)} \end{aligned}$$

We can calculate $t_{n-1,1-\alpha/2} = t_{99,\,0.975} = 1.98$ using Excel with the formula `T.INV(.975, 99)` and R function `qt(.975, 99)`).

Interval Estimation With Normality Assumption (σ-unknown)

the interval estimate is

$$\begin{aligned} &[82 - 1.98 \frac{18}{\sqrt{100}}, \quad 82 + 1.98 \frac{18}{\sqrt{100}}] \\ &= [82 - 3.56, \quad 82 + 3.56] \\ &= [78.44, \quad 85.56] \end{aligned} \tag{12}$$

and *the interval estimator is*

$$[\overline{X} - 0.198\ S, \quad \overline{X} + 0.198\ S] \tag{13}$$

Inferential Statistics - Part I, Estimation

Sampling Distribution of Sample Means (Without Assuming Normality)

Sampling Distribution of Sample Means Without Normality Assumption

- Without the Normality assumption, there are two ways to construct confidence intervals,
  - **Using Large Sample Results or Asymptotic Results:** Assuming we have a large sample, and then we can apply *Central Limit Theorem*.
  - **Using Resampling Techniques, e.g., bootstrap or jackknife:** This is not part of the syllabus ... if you are curious we can talk about this later.

  We will focus on the first one and learn how to apply Central Limit Theorem. But before that let's recall some rules of expectation and variance, two rules are important for us (there are many rules, but for now these two are fine...)

Sampling Distribution of Sample Means Without Normality Assumption

Rules 1.8: Some Rules of Expectation and Variance

- **Rule 1: Linearity of Expectation:** For any random variable X, Y and constant a and b we have

$$\mathbb{E}(aX + bY) = a\mathbb{E}(X) + b\mathbb{E}(Y)$$

- **Rule 2:** For any random variable X, Y and constant a and b we have

$$\mathbb{V}(aX + bY) = a^2\mathbb{V}(X) + b^2\mathbb{V}(Y) + 2ab\,\mathbb{C}\text{ov}(X, Y)$$

  Now the $\mathbb{C}\text{ov}(X, Y) = 0$ if the random variables X and Y are independent, and in that case we can write,

$$\mathbb{V}(aX + bY) = a^2\mathbb{V}(X) + b^2\mathbb{V}(Y)$$

Sampling Distribution of Sample Means Without Normality Assumption

Now if we have i.i.d. random sample, we can apply these rules,

$$\begin{aligned}\mathbb{E}(\overline{X}) &= \mathbb{E}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) \\ &= \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}(X_i) \\ &= \frac{1}{n} n\mathbb{E}(X_i) = \mathbb{E}(X_i)\end{aligned}$$

For variance we get,

$$\begin{aligned}\mathbb{V}(\overline{X}) &= \mathbb{V}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) \\ &= \frac{1}{n^2}\sum_{i=1}^{n}\mathbb{V}(X_i) \\ &= \frac{1}{n^2} n\mathbb{V}(X_i) = \frac{\mathbb{V}(X_i)}{n}\end{aligned}$$

Now we can see the central limit theorem,

Sampling Distribution of Sample Means Without Normality Assumption

Theorem 1.9: Central Limit Theorem

Whatever the population distribution is, if $X_1, X_2, \ldots, X_n$ are i.i.d. random variables with the same population mean $\mathbb{E}(X)$ and variance $\mathbb{V}(X)$, and we form a **Statistic** Z, where

$$Z = \frac{\overline{X} - \mathbb{E}(\overline{X})}{\sqrt{\mathbb{V}(\overline{X})}} = \frac{\overline{X} - \mathbb{E}(X_i)}{\sqrt{\mathbb{V}(X_i)/n}}$$

then for *large n (technically we need $n \to \infty$)*, the distribution of Z becomes approximately standard normal, this means

$$Z \overset{approx}{\sim} \mathcal{N}(0, 1) \text{ as } n \to \infty \tag{14}$$

And following the relation between Normal and Standard Normal, we can also write

$$\overline{X} \overset{approx}{\sim} \mathcal{N}\left(\mathbb{E}(\overline{X}), \mathbb{V}(\overline{X})\right) \text{ as } n \to \infty$$

In Words: This is really large sample result, which says whatever the population distribution is, if we take the sample size to ∞, then the Statistic Z, which is the standardized sample mean will follow Standard Normal distribution approximately. Note that it's an approximate result, since $n \to \infty$ is really a hypothetical matter, we can never take sample size n to ∞. This result also a special name, in statistics this is called **Central**

Sampling Distribution of Sample Means Without Normality Assumption

Limit Theorem, and this possibly one of the most influential results in Statistics. Note that, the result means - *No matter what is your population distribution, the sampling distribution of the standardized mean will become Standard Normal, as we increase the sample size*.

Sampling Distribution of Sample Means Without Normality Assumption

Example 1: Applying CLT to Bernoulli Random Sample
Let's see an example, suppose we have Bernoulli random sample, this means,

$$X_i \sim \text{Bernoulli}(p) \text{ for } i = 1, 2, \ldots, n$$

where p is a population proportion of 1, and this is also the probability of success and moreover we have,

$$\mathbb{E}(X_i) = p \qquad \mathbb{V}(X_i) = p(1-p)$$

Now we can calculate the mean and variance of $\overline{X}$. Note here $\overline{X}$ is the sample proportion of 1's, so we will write this with $\overline{p}$, but the idea is same. Here we have $\overline{X} = \overline{p} = \frac{1}{n}\sum_{i=1}^{n} X_i$. Now,

$$\mathbb{E}(\overline{X}) = \mathbb{E}(\overline{p}) = \mathbb{E}(X_i) = p, \qquad \mathbb{V}(\overline{X}) = \mathbb{V}(\overline{p}) = \frac{\mathbb{V}(X_i)}{n} = \frac{p(1-p)}{n}$$

Now the central limit theorem says,

$$Z = \frac{\overline{p} - p}{\sqrt{p(1-p)/n}} \overset{approx}{\sim} \mathcal{N}(0, 1) \text{ as } n \to \infty$$

or we can also write, $\overline{p} \overset{approx}{\sim} \mathcal{N}\left(p, \frac{p(1-p)}{n}\right)$ as $n \to \infty$

Sampling Distribution of Sample Means Without Normality Assumption

Example 2: Applying CLT to Normal Random Sample

This is already you know

$$X_i \sim \mathcal{N}(\mu, \sigma^2) \text{ for } i = 1, 2, \ldots, n$$

Since this is a Normal random variable, we have

$$\mathbb{E}(X_i) = \mu$$
$$\mathbb{V}(X_i) = \sigma^2$$

Now we can calculate the sample mean and variance of $\overline{X}$ (do it!)

$$\mathbb{E}(\overline{X}) = \mu$$
$$\mathbb{V}(\overline{X}) = \frac{\sigma^2}{n}$$

Now the central limit theorem says,

$$Z = \frac{\overline{X} - \mu}{\sqrt{\sigma^2/n}} \overset{approx}{\sim} \mathcal{N}(0, 1) \text{ as } n \to \infty$$

This is actually unnecessary, since we have the exact distribution result for normality assumption.

Sampling Distribution of Sample Means Without Normality Assumption

One important limitation for CLT is **it's not clear what does large n mean?** Actually we don't have any satisfactory answer, here more is better, but usually the rule of thumb is *the sample size more than 30 is considered to be a large sample...*

Sampling Distribution of Sample Means Without Normality Assumption

If you notice carefully in the CLT, we used the population variance, for Bernoulli we used $p(1-p)$ and for Normal we used σ^2, but in practice we don't know the population variance, so we need to use sample variance... there is also a large sample result for this, which is following

Sampling Distribution of Sample Means Without Normality Assumption

Corollary 1.10: Large Sample Result using T Statistic

Following the Theorem 1.5. we also get following results:

$$T \overset{approx}{\sim} \mathcal{N}(0,1) \text{ as } n \to \infty \tag{15}$$

$$\text{where } T = \frac{\overline{X} - \mathbb{E}(\overline{X})}{\sqrt{\widehat{V}(\overline{X})}} \tag{16}$$

This is the old T statistic type result, usually it's hard to calculate $\mathbb{V}(\overline{X})$ (e.g., for Bernoulli this is $\frac{p(1-p)}{n}$ and for Normal this σ^2/n). So in this case we will use this result. Note that in this case we use the *estimator of the variance of $\overline{X}$*, when it comes to Bernoulli we can use

$$\widehat{V}(\overline{X}) = \frac{\bar{p}(1-\bar{p})}{n}$$

then we can use the T statistic which is

$$T = \frac{\overline{p} - p}{\sqrt{\widehat{V}(\overline{X})}} = \frac{\overline{p} - p}{\sqrt{\bar{p}(1-\bar{p})/n}}$$

In large sample this statistic follows Normal distribution, hence we can write

Sampling Distribution of Sample Means Without Normality Assumption

$$T \overset{approx}{\sim} \mathcal{N}(0, 1) \text{ as } n \to \infty$$

Inferential Statistics - Part I, Estimation

Interval Estimation Without Assuming Normality

Interval Estimation Without Normality Assumption

- Since in large samples we can use the T statistic and it follows approximately normal distribution, we can start from

$$\mathbb{P}\Big(z_{\alpha/2} \leq T \leq z_{1-\alpha/2}\Big) \approx 1-\alpha \quad \text{as we take } n \to \infty$$

$$\mathbb{P}\left(z_{\alpha/2} \leq \frac{\overline{X}-\mathbb{E}(\overline{X})}{\sqrt{\widehat{V}(\overline{X})}} \leq z_{1-\alpha/2}\right) \approx 1-\alpha$$

$$\vdots$$

$$\mathbb{P}\left(\overline{X}-\sqrt{\widehat{V}(\overline{X})}\, z_{1-\alpha/2} \leq \mu \leq \overline{X}+\sqrt{\widehat{V}(\overline{X})}\, z_{1-\alpha/2}\right) \approx 1-\alpha$$

- For large sample applying these results give us the following formula for the confidence interval estimate

$$\text{lower limit} = \overline{x}-\sqrt{\widehat{V}(\overline{X})}\, z_{1-\alpha/2} \text{ and upper limit} = \overline{x}+\sqrt{\widehat{V}(\overline{X})}\, z_{1-\alpha/2}$$

- Let's do an example

Example 1.11: In the spring of 2017, the Consumer Reports National Research Center conducted a survey of 1007 adults to learn about their major health-care concerns. The survey results showed that 574 of the respondents lack confidence they will be able to afford health insurance in the future.

- a. What is the point estimate of the population proportion of adults who lack confidence they will be able to afford health insurance in the future.
- b. Develop a 90% confidence interval for the population proportion of adults who lack confidence they will be able to afford health insurance in the future.
- c. Develop a 95% confidence interval for this population proportion.

Solution:
Here we have

- Our target is population proportion p, this means this is the population proportion of 1, or population proportion of lack of confidence.
- If we think about the Bernoulli random variables $X_i \sim \text{Bern}(p)$ for all $i = 1, 2, \ldots, n$, then p is probability of success, which is the probability that a respondent lacks confidence they will be able to afford health insurance in the future.
- Population mean is also $\mathbb{E}(X_i) = p$, and Population variance is $\mathbb{V}(X_i) = p(1-p)$, both are unknown since p is unknown.

- Sample size $n = 1007$.
- Sample mean or sample proportion $\overline{p} = \frac{574}{1007} = 0.570$.

Interval Estimation Without Normality Assumption

- Sample Standard deviation, $\widehat{\mathbb{V}}(X_i) = \overline{p}(1-\overline{p}) = 0.570 * (1-0.570) = 0.2451$

- Mean of sample proportion $\mathbb{E}(\overline{X}) = \mathbb{E}(\overline{p}) = p$.Variance of sample proportion $\mathbb{V}(\overline{X}) = \mathbb{V}(\overline{p}) = \frac{p(1-p)}{n}$ and Standard Error is $\sqrt{\mathbb{V}(\overline{X})} = \sqrt{\frac{p(1-p)}{n}}$. All of these quantities cannot be calculated since we don't know p

- Estimate of the standard error is $\sqrt{\widehat{\mathbb{V}}(\overline{X})} = \sqrt{\frac{\overline{p}(1-\overline{p})}{n}} = 0.0156$. This is the most important object for us since we need this to construct confidence interval.

  We do the 90% confidence interval, in this case $\alpha = 0.10$, the formula is

$$\overline{p} \pm z_{1-\alpha/2}\sqrt{\widehat{\mathbb{V}}(\overline{X})}$$

- We can find $z_{1-\alpha/2} = z_{0.95} = 1.645$ using Excel with the formula `NORM.INV(.95, 0, 1)` and R function `qnorm(.95)`).

- Now the upper limit is

$$\overline{p} + z_{1-\alpha/2}\sqrt{\widehat{\mathbb{V}}(\overline{X})} = 0.570 + 1.645 \times 0.0156 = 0.570 + 0.0257 = 0.5957$$

- And the lower limit is

$$\overline{p} - z_{1-\alpha/2}\sqrt{\widehat{\mathbb{V}}(\overline{X})} = 0.570 - 1.645 \times 0.0156 = 0.570 - 0.0257 = 0.5443$$

- So the interval estimate is

$$[0.5443, 0.5957]$$

- Interpretation: This is a fixed interval, this means the population proportion p of adults who lack confidence they will be able to afford health insurance in the future, is either in this interval or not. However in repeated sampling, if we construct intervals like this 100 times then roughly 90 times of the time the intervals will contain the true population proportion p.

References

Abraham, B. and Ledolter, J. (2006), *Introduction to Regression Modeling*, Duxbury applied series, Thomson Brooks/Cole, Belmont, CA.

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.

James, G., Witten, D., Hastie, T. and Tibshirani, R. (2023), *An introduction to statistical learning*, Vol. 112, Springer.

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.

References

Rice, J. A. (2007), *Mathematical Statistics and Data Analysis*, Duxbury advanced series, 3rd edn, Thomson/Brooks/Cole, Belmont, CA.

Wasserman, L. (2013), *All of statistics: a concise course in statistical inference*, Springer Science & Business Media.

Wooldridge, J. M. (2009), *Introductory Econometrics: A Modern Approach*, 4th edn, South Western, Cengage Learning, Mason, OH.