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: March 20, 2025

Outline

Descriptive vs Inferential Statistics

Motivating Picture

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

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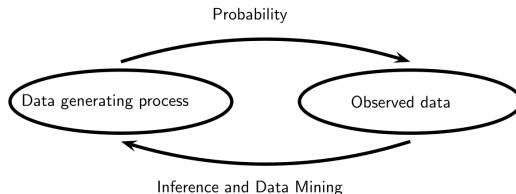
Figure 1: The figure is directly taken from Wasserman (2013), clearly explains what you did in ECO 104 (indicated by the arrow at the top going to right direction) and what you will do in ECO 204 (indicated by the arrow at the bottom going to left).

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 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: How would you calculate

    - ▶ % of female *proportion* of students *in this sample* and *in the population*

    - ▶ *mean family income* of the students *in this sample* and *in the population*

- ▶ Q3: What's the difference between sample and population quantities?
- ▶ Q4: If I have a population data, then *do I need a sample*?
- ▶ Q5: Why would I go for a sample?
- ▶ Q6: What are the possible issues in a sample?
- ▶ Q7: Do our predictions improve if we collect more samples....

Introduction to Statistical Inference

Here are some answers,

▶ Ans 1: All currently enrolled EWU students.

▶ Ans 2:

Sample Proportion and Population Proportion

▶ *Sample Proportion* of female students for this sample - 40%.
▶ *Population Proportion* for the entire population - we don't know, perhaps possible to calculate but quite hard!

Sample Mean and Population Mean

▶ *Sample Mean* of family income - $39,006.4$ taka.
▶ *Population Mean* of family income for the entire population - we don't know, perhaps possible to calculate, but quite hard

Introduction to Statistical Inference

- Ans 3:
    - Population quantities are something we would like to know, (*we call this target object*)
    - Sample quantities are something which is possible to calculate using a sample, we call this *estimate of the target*
    - Definitely chances are very low that they will be exactly same, however with a "good" sample we might be close to our target.
    - Interestingly even if we don't know about the population quantities, using a sample we can roughly make some conclusions about the population quantities, for example we can say our *population proportion* of female students is perhaps (possibly close to) 40%, or we can say our *population average / mean* of family income is perhaps (possibly close to) 39,006 tk.

    What we just saw, is an example of *Statistical Estimation* or in short *Estimation*, we will discuss this in detail in the coming section. Note *the target* is always something related to the population and *an estimate* is always something related to a sample.

Introduction to Statistical Inference

- Ans 4: NO, I have all information I need, so no need for sampling.
- Ans 5: Collecting population data is often hard, time consuming or sometimes impossible,
- Ans 6: 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 7: 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: Statistical Inference

Statistical inference is a process of making conclusions about the population using a sample, e.g., making conclusion about the population mean, proportion, variance, etc., using a sample mean, sample proportion, sample variance, etc.

Introduction to Statistical Inference

▶ The practice of Statistics falls broadly into two categories *Descriptive and Inferential*

▶ <u>Descriptive Statistics</u> 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 <u>Inferential Statistics</u> 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, Estimate and then Estimator*.

Inferential Statistics - Part I, Estimation

Estimation, Estimate and Estimator

Estimator

▶ We already have talked about *Estimation*, but let's discuss this in detail now, here is the same sample or data,

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	$X_5 =?$

▶ The *estimation* process is easy

  ▶ We fix a *target parameter*, in this case it is the population mean (or average) of the monthly income of all EWU students, let's write this unknown quantity with μ, so μ is our target.

  ▶ We think about a sample quantity, for example sample mean.

  ▶ And then we find / calculate the sample quantity using the sample we have, we call this *estimate*, in this case the estimate is

  $$\frac{70,150 + 20,755 + 44,758 + 38,790 + 20,579}{5} = \frac{195032}{5} = 39,006.4 \approx 39,006$$

  this is an estimate of the unknown quantity μ

Estimator

▶ A *target parameter* can be different objects *calculated directly from the population*, for example,

   ▶ the population mean (or average), usually denoted with μ,
   ▶ the population proportion, usually denoted with p,
   ▶ the population variance, usually denoted with σ^2,
   ▶ the population standard deviation, usually denoted with σ, etc.

▶ Can you think about more examples?

Estimator

▶ Now let's understand the *Estimator*, here is where *randomness* comes. There are two ways to think about the data set,

    ▶ **Random Data or Random Sample:** We can think each row as a *random data point* since before sampling, it is possible to have different values from the population. In this case we can think we have a random variable $X_1 =?$ at first row, X_2 as a second row and so on... which in principle can take different values from the population, but *before sampling* there is no value yet.

    ▶ **Observed / Fixed Data or Observed sample:** Also 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 continuing like this, for the whole data set, we can think

    ▶ 5 fixed numbers, and they are 70, 150; 20, 755; 44, 758; 38, 790 and 20, 579. These are observed data point, together we call it *observed sample* or *realized sample*.

    ▶ 5 random variables and they are X_1, X_2, X_3, X_4, X_5. These are random variables, together we call it *random sample*

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 $X_1, X_2, X_3, \ldots, X_n$.

▶ 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

▶ An Estimator is a *function of a random sample*, in this case the estimator is

$$\overline{X} = \frac{X_1 + X_2 + X_3 + X_4 + X_5}{5} = \frac{1}{n}\sum_{i=1}^{5} X_i$$

▶ So an *Estimator* is a random variable, it changes from sample to sample, but when we calculate it for a fixed sample, then we get $\overline{x}$, where

$$\overline{x} = \frac{x_1 + x_2 + x_3 + x_4 + x_5}{5} = \frac{1}{n}\sum_{i=1}^{5} x_i$$

this is a constant and it's *deterministic and not random*. Recall we call this an estimate.

▶ In our case $\overline{x} = 39,006$, in this case we also write $\overline{X} = 39,006$, since the random variable $\overline{X}$ has taken a value $39,006$ for this sample. Since $\overline{X}$ is a random variable, question is what is the probability distribution of $\overline{X}$? or Expectation of $\mathbb{E}(\overline{X})$. Do you know what is a Probability Distribution or Expectation or Variance?

▶ Let's recap some of the ideas from ECO104 ..... Pleas look at slides from Chapter 1.5

Inferential Statistics - Part I, Estimation

Estimator and Sampling Distribution

Estimator and Sampling Distribution

- ▶ Before continuing from here you are supposed to have a look at the recap slides of ECO 104, if not, then please do so before you proceed,
- ▶ Now when it comes to random variable $\overline{X}$, we are interested in 3 important questions,
  1. What is the Expectation of the random variable $\overline{X}$, written as $\mathbb{E}(\overline{X})$?
  2. What is the variance of the random variable $\overline{X}$, written $\mathbb{V}(\overline{X})$?
  3. What is the probability distribution of $\overline{X}$ (this is what we call *sampling distribution!*), for example is it the case that

  $$\overline{X} \sim \mathcal{N}(?, ?)$$

- ▶ The answer to the third question is what we call *Sampling Distribution of Sample Means*. Note that, this is the distribution of sample means $\overline{x}$, 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 (why?)
- ▶ Related to this question, will now see three important results ....,

Estimator and Sampling Distribution

Theorem 1.1: (Mean and Variance of $\overline{X}$ with only i.i.d assumption)

If we have i.i.d random variables $X_1, X_2, \ldots, X_n$ with the same mean μ and same variance σ^2, then

$$i) \quad \mathbb{E}\left(\overline{X}\right) = \mu \tag{1}$$

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

In Words: This says if all the random variables $X_1, X_2, \ldots, X_n$ are identically and independently distributed (in short we say i.i.d. random variables) with same mean and variance, this means

$$\mathbb{E}(X_1) = \mathbb{E}(X_2) = \ldots = \mathbb{E}(X_n) = \mu \text{ and } \mathbb{V}(X_1) = \mathbb{V}(X_2) = \ldots = \mathbb{V}(X_n) = \sigma^2$$

then no matter what is the distribution of the random variables, the expectation of the sample mean $\overline{X}$ will be equal to the population mean μ, i.e., $\mathbb{E}(\overline{X}) = \mu$, and the variance of the sample mean is $\mathbb{V}(\overline{X}) = \sigma^2/n$, (where n is the sample size. The key thing here is i.i.d. which means identically and independently distributed!).

Estimator and Sampling Distribution

Note the mean and variance could come from different distributions, not just normal. For example if the population follows Normal distribution $\mathcal{N}(\mu, \sigma^2)$ then the mean is μ and variance is σ^2, in this case theorem says

$$\mathbb{E}\left(\overline{X}\right) = \mu$$

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

If the data follows $\text{Bern}(p)$, then the mean is p and variance is $p(1-p)$, so the theorem says

$$\mathbb{E}\left(\overline{X}\right) = p$$

$$\mathbb{V}\left(\overline{X}\right) = \frac{p(1-p)}{n}$$

So in this case the μ becomes p and σ^2 is $p(1-p)$. Important is the theorem makes no assumption whether the data is Normal or Bernoulli.

Estimator and Sampling Distribution

Theorem 1.2: (Distribution of $\overline{X}$ with normality and i.i.d assumption)

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

$$i) \quad \mathbb{E}\left(\overline{X}\right) = \mu \quad , \quad ii) \quad \mathbb{V}\left(\overline{X}\right) = \frac{\sigma^2}{n}$$

$$iii) \quad \overline{X} \sim \mathcal{N}(\mu, \frac{\sigma^2}{n}) \quad , \quad iv) \quad Z \sim \mathcal{N}(0,1) \text{ where } Z = \frac{\overline{X} - \mu}{\frac{\sigma}{\sqrt{n}}} \tag{3}$$

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

In Words: First note, we added one more condition here that is now we are assuming i.i.d. and also assuming Normality in the data, this gives the distribution of $\overline{X}$ (which is more than what we got before) and the distribution is normal with mean μ and variance σ^2/n. This is a very important result, and we will do some problems related to this. The number (iv) is a direct consequence of the relationship between standard normal and normal distributions in general. 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*

Estimator and Sampling Distribution

Let's talk about the T statistic,

- first of all, note that T is a random variable, since $\overline{X}$ is a random variable, (and also S is a random variable!)

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

number v) says the distribution of this random variable is t_{n-1}, where $n-1$ is the degrees of freedom (which is the parameter of the distribution)

Estimator and Sampling Distribution

Theorem 1.3: (Central Limit Theorem (CLT) and related results)

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

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

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

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

Estimator and Sampling Distribution

▶ This theorem is similar to the first one, it says if we have i.i.d. random variables, with mean μ and variance σ^2, then when *n is large*,

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

here again this is distribution free result, for example when we have $\mathcal{N}(\mu, \sigma^2)$, Z is same as above, but when the data is $\text{Bern}(p)$, then

$$Z \overset{approx}{\sim} \mathcal{N}(0,1) \text{ where } Z = \frac{\overline{X} - p}{\sqrt{\frac{p(1-p)}{n}}}$$

In general, note that we can always write

$$Z = \frac{\overline{X} - \mathbb{E}(\overline{X})}{\sqrt{\mathbb{V}(\overline{X})}} \text{ where } Z \overset{approx}{\sim} \mathcal{N}(0,1)$$

Estimator and Sampling Distribution

▶ Similarly for T Statistic we can write,

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

And in general we can write,

$$\widehat{\mathbb{V}}(\overline{X}) = \frac{S^2}{n}, \text{ where } S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2$$

But for Bernoulli, we can also write,

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

Note $\bar{p}$ is same as $\overline{X}$ which is sample proportion.

▶ Note that in repeated sampling, T is a random variable (this is because both $\overline{X}$ and S is random), but you will see sometimes people use lower case t to denote the t-statistic. It doesn't matter as long as you understand when we are talking about random variable and when we are talking about a fixed number (i.e., realized value of the random variable). We will use uppercase T to denote this random variable and lowercase t to denote the realized value of the random variable.

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*.

Statistic, Point Estimator and Sampling Distribution - Important Remarks

▶ 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 $X_1, X_2, X_3, \ldots, X_n$! The distribution of statistic is called *sampling distribution*.

▶ For example, $\overline{X}$, Z are both examples statistics but $\overline{X}$ is an estimator for μ, Z is just a statistic.

▶ Another example is S^2, where $S^2 = \frac{1}{n-1} \sum_{i=1}^{n} \left(X_i - \overline{X} \right)^2$. This is a statistic since it's a function of the random sample. And this is also an estimator for σ^2, because it is targeting population variance σ^2. Note that S^2 is just a sample variance.

Inferential Statistics - Part I, Estimation

Properties of Point Estimator

Properties of Point Estimators

- Why did we take the average to estimate μ? Why not median, or maximum? These all are examples of estimators for μ, so why sample mean $\overline{X}$?
- The answer is, the sample average is a "good" estimator for the population mean μ
- 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

$$\mathbb{E}(\overline{X}) = \mu$$

- 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.

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

$$\text{if } n \to \infty \text{ then } \overline{X} \to \mu \text{ happens with very high probability}$$

▶ 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 μ.

▶ If you contrast consistency to unbiasedness, you will notice that this is a limiting property or we say *asymptotic property* (because we are saying $n \to \infty$, recall limit...), as opposed to finite sample property!

▶ The estimator sample mean X_n is both an unbiased and a consistent estimator for the population mean μ.

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 $\mathbb{V}(\overline{X}) = \frac{\sigma^2}{n}$ (In Anderson you will see the notation $\sigma_{\overline{X}}^2$ 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!

Inferential Statistics - Part I, Estimation

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*.

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

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

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

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

How to get Estimators

- This should be sample variance S^2, where
$$S^2 = \frac{1}{n-1} \sum_{i=1}^{n} \left(X_i - \overline{X} \right)^2$$
- Again note that, we replaced Expectation with averages.
- S^2 is an unbiased estimator of σ^2.
- What if we divide by n rather than $n-1$? this is also an estimator of σ^2, but unfortunately this is a biased estimator!

Interval Estimation of Population Mean μ

Interval Estimation of Population Mean μ

Basic idea of Interval Estimation

Interval Estimators

▶ Before we proceed, here is the recap of the terminologies,
  ▶ $\overline{X}$ is the *point estimator* of the population mean μ (note this is a random object!)
  ▶ When we calculate this for one fixed sample, then this is what we call *an estimate* of the unknown parameter μ.
  ▶ The whole process is called estimation.
  ▶ The probability distribution of the sample means (in repeated sampling) is what we call the sampling distribution of the sample means.

Interval Estimators

▶ 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 α, where $0 < \alpha < 1$, we say a $100(1 - \alpha)\%$ *interval estimator for μ is a random interval* $[L, U]$ *such that*

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

▶ 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,

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

▶ What is the interpretation

Interval Estimators

▶ Recall the *the frequency interpretation of probability*.

▶ Using the frequency interpretation means, if we construct the interval $[L.U]$, then around 100 times, roughly 95 times the true μ fall in this interval.

▶ We can only think about probability when we have a random object, what is random here?

▶ First of all note that, μ 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 $X_1, X_2, X_3, \ldots, X_n$, so we should write, $L(X_1, X_2, \ldots, X_n)$ and $U(X_1, X_2, \ldots, X_n)$. 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 μ"

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 μ

Interval Estimation - First Example σ **known case**

Interval Estimate / Confidence Interval
σ known

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

Example 1.4: (Interval Estimator and Interval Estimate/Confidence Interval)
Suppose we have $\overline{x} = 82$, 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* μ, 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

$$[\overline{X} - 1.96\frac{20}{\sqrt{100}}, \quad \overline{X} + 1.96\frac{20}{\sqrt{100}}] \tag{8}$$

The *interval estimate or confidence interval* is

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

Now note, the first one at (8) is a *random interval* since $\overline{X}$ 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 μ is there or it is not there. 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 $[78, 85.92]$".

Interval Estimate / Confidence Interval
σ 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 $\overline{x}$

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

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

▶ The idea is $\overline{x}$ 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 **R**

▶ First note, we have following information
  ▶ $n = 100$
  ▶ $\overline{x} = 82$
  ▶ $\alpha = 0.05$ (this is because we are asked to construct 90% confidence interval)
  ▶ $\sigma = 20$

Interval Estimate / Confidence Interval
σ **known**

ℝ **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 μ

Deriving Interval Estimator - σ known case

Deriving Interval Estimators
σ 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 $X_1, X_2, \ldots, X_n$ 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)]

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

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

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

- Recall $\overline{X}$ is a statistic and an estimator of μ. In this case Z is also a statistic, the benefit of transforming $\overline{X}$ to Z is we can now use standard normal. Why did we do this? We will see that here Z also plays an important role to find the interval estimator for μ.
- Now, let's derive the interval estimator for μ. 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
σ 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 gte the two sided interval, first fix α, let's say $\alpha = 5\%$ then $1 - \alpha$ is what we call *confidence coefficient / confidence level / nominal coverage*. Now since $Z \sim \mathcal{N}(0,1)$, we can write

$$\mathbb{P}\left(z_{\alpha/2} \leq Z \leq z_{1-\alpha/2}\right) = 1 - \alpha \tag{10}$$

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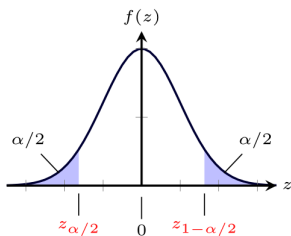
Here $z_{\alpha/2}$ is a value such that $\mathbb{P}(Z < z_{\alpha/2}) = \alpha/2$ and $z_{1-\alpha/2}$ is a value such that $\mathbb{P}(Z < z_{1-\alpha/2}) = 1 - \alpha/2$. It is important to mention that because of the *symmetry* of the Normal distribution always we will have $z_{\alpha/2} = -z_{1-\alpha/2}$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

Deriving Interval Estimators
σ known

Now we will do some algebra with the term inside the probability in (10), recall we had $z_{\alpha/2} \leq Z \leq z_{1-\alpha/2}$,

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

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

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

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

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

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

So this means, writing

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

is same as

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

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

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

So the interval estimator is

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

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

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

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* μ

Interval Estimation of Population Mean μ

Interval Estimator - σ unknown case

Interval Estimators

▶ Can we construct intervals when we do not know the population standard deviation σ. The answer is YES!

▶ We need to use the statistic T from the Theorem 1.3 (v). This is called *t-statistic*, on the other hand when we used Z, that is called *z-statistic*.

▶ Note that in this case the statistic T 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, 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).

- T or the t-statistic is written as,

$$T = \frac{\overline{X} - \mu}{\frac{S}{\sqrt{n}}} \tag{11}$$

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

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

- This means T 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

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

Interval Estimators
σ **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,

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

▶ The *interval estimate* in this case is,

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

▶ Here $t_{(n-1),1-\alpha/2}$ is the $1-\alpha/2$ quantile of the t distribution with parameter $(n-1)$, and $\overline{x}$ is the sample mean, and s is the sample standard deviation.

- ▶ Let's see a concrete example.
- ▶ Suppose in Example 1.5, we don't know σ, rather we know sample standard deviation $s = 18.5$.
- ▶ This means everything is same except the information of σ is not known to us,
    - ▶ $n = 100$
    - ▶ $\overline{x} = 82$
    - ▶ $\alpha = 0.05$ (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 σ known case.
    - ▶ As mentioend we need to use t_{n-1} distribution
    - ▶ We need to use s, which is the sample standard deviation
    - ▶ Because we don't know σ, 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
σ unknown

ℝ 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 μ

Interval Estimator - σ 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 σ, we get a new statistic, that is what we called *t statistic*,

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

- 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

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

- 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,

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

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

Proportion Estimation

Proportion Estimation

▶ The idea of the proportion estimation is same as mean estimation.

▶ If we do point estimation, in this case our *target object* is population proportion p and the point estimate is the sample proportion $\overline{x}$, which we sometimes also write as $\bar{p}$

▶ For interval estimation, the general formula is,

$$\overline{p} \pm z_{1-\alpha/2} \times \sqrt{\frac{\overline{p}(1-\overline{p})}{n}}$$

▶ Where $z_{1-\alpha/2}$ is the $1-\alpha/2$ quantile of the standard normal distribution, and $\overline{p}$ is the sample proportion and $\sqrt{\frac{\overline{p}(1-\overline{p})}{n}}$ is the estimate of the standard error, since in this case the standard error is

$$\sqrt{\mathbb{V}(\overline{p})} = \sqrt{\frac{p(1-p)}{n}}$$

▶ If we use $\overline{p}$ then we have an estimate of the standard error, which is

$$\sqrt{\hat{\mathbb{V}}(\overline{p})} = \sqrt{\frac{\bar{p}(1-\bar{p})}{n}}$$

Proportion Estimation

▶ The code is very simple, here is an example

ℝ code - Confidence Interval for Population Proportion

```
# First create some objects with the information given
n <- 100
pbar <- 82/100 # you can also write this as xbar
alpha <- 0.05


# calculate the estimate of the sderror and moe and save them as objects
sderror_est <- sqrt((pbar*(1-pbar))/n) # this is for Bernoulli
moe <- qnorm(1 - alpha/2) * sderror_est

# upper limit
pbar + moe


# lower limit
pbar - moe
```

Proportion Estimation

▶ Although a theoretical matter, but what happens here is, the statistic (which we wrote below) becomes approximately Normal when the sample size becomes very large, i.e.,

$$\frac{\bar{p} - p}{\sqrt{\frac{\bar{p}(1-\bar{p})}{n}}} \overset{\text{approx.}}{\sim} \mathcal{N}(0, 1)$$

▶ This is the reason we use the standard normal distribution in the above code.

▶ This is an application of Theorem 1.3 (which is known as Central Limit Theorem) and also another theorem known as Continuous Mapping Theorem.

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.