# Ch1 - Estimation

## (*Point and Interval Estimation)*

*ECO 204*
*Statistics For Business and Economics - II*

Shaikh Tanvir Hossain

**E**ast **W**est **U**niversity, Dhaka
Last updated: February 17, 2025

# Outline

## Outline

# Descriptive vs Inferential Statistics

# Motivating Picture

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

Consider following picture,

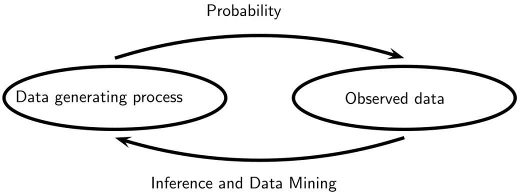


**Figure 1:** The figure is directly taken from Wasserman (2013), clearly explains what you did in ECO 104 (indicated by the arrow at the top going to right direction) and what you will do in ECO 204 (indicated by the arrow at the bottom going to left).

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: 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....

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

- 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.

- 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 ....

- 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*
- 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, Estimate and then Estimator*.

# Inferential Statistics - Part I, Estimation

## Estimation, Estimator and Sampling Distribution

# 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 $\mu$, so $\mu$ 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 $\mu$

# Estimator

- A *target parameter* can be different objects *calculated directly from the population*, for example,
  - the population mean (or average), usually denoted with $\mu$,
  - the population proportion, usually denoted with $p$,
  - the population variance, usually denoted with $\sigma^2$,
  - the population standard deviation, usually denoted with $\sigma$, etc.
- Can you think about more examples?

# Estimator

- Now let's understand the *Estimator*, first we can think about a each of the rows, or each data point in two ways, consider the first row
  - **Random Data:** First, we can think this 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, which in principle can take different values from the population, but *before sampling* there is no value yet.
  - **Observed Data:** Second, when we have observed the sample, this becomes a fixed number and in this case the random variable $X_1$ has already taken a value, so we get $X_1 = 70,150$. 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 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 $\bar{x}$, 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 *not random*, and it is the 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.

# Estimator

- Since $\bar{X}$ is a random variable, question is what is the probability distribution of $\bar{X}$? or Expectation of $\mathbb{E}(\bar{X})$. Do you know what is a Probability Distribution or Expectation or Variance?
- Let's recap some of the ideas from ECO104 .....

# 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.