Chapter 3

Estimation (Point and Interval Estimation)

Shaikh Tanvir Hossain

October 30, 2022

East West University

Outline

1. Point Estimation
2. Interval Estimation

1. Point Estimation

2. Interval Estimation

As we discussed one of the key tasks of Statistics is *making inference (guess)* about the true distribution of the data. It could be that we are interested in the distribution itself or we are interested in its parameters. There are two major themes of statistical inference

- 1. Estimation (Point Estimation and Interval Estimation)
- 2. Testing, also known as "Hypothesis Testing".

In this chapter we will focus on estimation. Here is an example of *Point Estimation*

- Suppose we are interested in the *Population mean μ* *
- But we cannot access it and we only have a sample $x_1, x_2, \dots, x_n$ (Recall these are fixed numbers for a sample of size n) †
- So we find the sample mean $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$
- This sample mean $\bar{x}$ is a *point estimate* of the *unknown Population mean μ*.
- This process is what we call *estimation*.
- And if we think about $\bar{X}_n$ (which is a function of n random variables) then this is what we call an *Estimator*.

---

*For example, in Chapter 2 we talked about population average income

†For example in chapter 2 we had $n = 10$ and we had 10 income levels and we can calculate sample average!

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 $\bar{x}$. Here $\bar{x}$ is a constant and it's not random. Some questions -

- Why did we take the average to estimate μ?
- Recall μ is actually the exptected value of X, $\mathbb{E}(X) = \mu$.
- Often it turns out we can get a "good" estimator, just by *replcing expectation with averages*. So

$$\mathbb{E}(X) = \mu, \text{ target object}$$

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

- 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

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

- This should be sample variance S^2, where

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

- Again note that, we replaced Expectation with averages.

- There are some properties of an estimator, these are often desirable properties. Two important properties are
  - ✓ Unbiasedness, and
  - ✓ Consistency
- If $\bar{X}_n$ is an *unbiased estimator* then we have $\mathbb{E}(\bar{X}_n) = \mu$
- If $\bar{X}_n$ is a *consistent estimator* then, if we have $n \to \infty$ then there is a very high probability that $\bar{X}_n$ will approach to μ.
- What about S^2 and σ^2?

1. Point Estimation

2. Interval Estimation

- Now, we discuss another type of estimation - *Interval Estimation*.
- Although point estimators are useful, interval estimators actually *convey more information* about the data that are used to obtain the point estimate.
- The purpose of using an interval estimator is to have some degree of confidence of securing the true parameter.
- For an interval estimator of a single parameter μ, we will use the random sample to find two quantities L and U such that $L \leq \mu \leq U$ *with some probability*.
- First note that L and U will depend on the sample values, so they will be random. So in this sense the better notation is $L(X_1, X_2, \ldots, X_n)$ 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.
- Ideally the interval $[L, U]$ should have two properties:
  - ✓ $P(L \leq \mu \leq U)$ *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.

- In summary, interval estimation actually does more than point estimation because along with an interval $[L, U]$ it also provides a measure of one's confidence in the accuracy of the estimate.
- The limits of the interval estimators are called, *the upper and lower confidence limits*. We denoted this with U and L respectively. The associated levels of confidence are determined by specified probabilities, this is called the *confidence coefficient*.
- When we calculate the interval estimators for a fixed sample, then the fixed range is called an *interval estimate*.
- The width of the confidence interval reflects the amount of variability inherent in the point estimate.
- As already mentioned our objective is to find a narrow interval with high probability of enclosing the true parameter, μ.
- The confidence coefficient tells us that what fraction of the time that the constructed interval will contain the true parameter, under *repeated sampling*.

- Let L and U be the lower and upper confidence limits for a parameter μ based on a random sample $X_1, \ldots, X_n$.
- Recall both L and U are functions of the sample, so we can write,

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

and we read it as we are $(1 - \alpha)100\%$ confident that the true parameter μ is located in the interval like $[L, U]$.

- Or we can also say

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

- 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)

- For example we already know that our iid random sample $X_1, X_2, \ldots, X_n \stackrel{iid}{\sim} \mathcal{N}(\mu, \sigma^2)$, then we know

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

- When we do standardization, we get

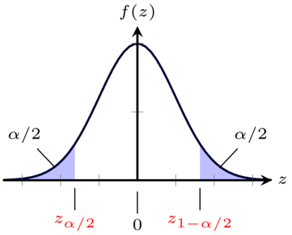
$$Z = \frac{\bar{X}_n - \mu}{\frac{\sigma}{\sqrt{n}}} \sim \mathcal{N}(0, 1)$$

- Recall $\bar{X}_n$ is a statistic. In this case Z is also a statistic, the benefit of transforming to Z is we can use standard normal.
- 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.

We will construct two-sided intervals, fix α, let's say $\alpha = 5\%$ then $1-\alpha$ is your confidence coefficient. Now since $Z \sim \mathcal{N}(0,1)$, we can write

$$P\left(z_{\alpha/2} \leq Z \leq z_{1-\alpha/2}\right) = 0.95$$

Visually this means,



Here $z_{\alpha/2}$ is a value such that $P(Z < z_{\alpha/2}) = \alpha/2$ and $z_{1-\alpha/2}$ is such a value such that $P(Z < z_{1-\alpha/2}) = 1-\alpha/2$. It is important to mention that because of the symmetry of the Normal distribution always we will have $z_{\alpha/2} = -z_{1-\alpha/2}$.

Now we will do some algebra with $z_{\alpha/2} \leq Z \leq z_{1-\alpha/2}$,

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

So this means, writing

$$P\left(z_{\alpha/2} \leq Z \leq z_{1-\alpha/2}\right) = 0.95$$

is same as

$$P\left(\bar{X}_n - \frac{\sigma}{\sqrt{n}} z_{1-\alpha/2} \leq \mu \leq \bar{X}_n + \frac{\sigma}{\sqrt{n}}\, z_{1-\alpha/2}\right) = 0.95$$

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

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

So the interval estimator is

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

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

- It is important to note that the interpretation of the the *Interval estmator* is going to be a *probabilistic interpretation*. This is what we mentioned in slide - 9. But once we calculate the *interval estimate* using a particular sample, the interpretation should not have any probabilistic statement.
- In Anderson et al. (2020) the *interval estimates* are also called *confidence interval*.
- Let's do an example from Anderson et al. (2020)

Example 1.1 (Interval Estimator and Interval Estimate/Confidence Interval, Example from Anderson et al. (2020))

Suppose we have $\bar{x} = 82$, population standard deviation $\sigma = 20$, sample size $n = 100$, and we are asked to compute the 95% *confidence interval of the population mean μ*, then since $z_{1-\alpha/2} = 1.96$, the interval estimator is

$$[\bar{X}_n - 1.96\frac{20}{\sqrt{100}}, \quad \bar{X}_n + 1.96\frac{20}{\sqrt{100}}]$$
$$[82 - 3.92, \quad 82 + 3.92]$$

The interval estimate or confidence interval is

$$[82 - 1.96\frac{20}{\sqrt{100}}, \quad 82 + 1.96\frac{20}{\sqrt{100}}]$$
$$[82 - 3.92, \quad 82 + 3.92]$$

Now note, the first one is a random interval since $\bar{X}_n$ is random but the second one is a determinnistic interval (there is no randomness here!). So in the second one either our population mean μ is there or it is not there. So we can say "for this particular sample, *the interval estimate* is $[78, 85.92]$".

- Can we construct intervals when we do not know the population standard deviation σ. The answer is YES (there are some details in Chapter 2. We need to use T Statistic and t-distribution (why?)
- 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 (degrees of freedom (DF) is a parameter for t-distribution).

$$T = \frac{\bar{X}_n - \mu}{\frac{S}{\sqrt{n}}} \sim t_{(n-1)}$$

here S is the sample standard deviation, Recall S^2 is

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

- And $S = \sqrt{S}$
- If we follow the exact same approach to construct an interval estimator, we will find

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

- And then we can also calculate the interval estimate for a fixed sample but using T distribution.
- We will use the following two notations throughout the lecture, so make sure you understand it clearly.
- $z_{1-\alpha/2}$ is a value of the random variable Z such that

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

- Similarly $t_{1-\alpha/2}$ is a value of the random variable T such that

$$P(T < t_{1-\alpha/2}) = 1 - \alpha/2$$

- In the last two cases, We have constructed confidence interval using the *exact distribution*, we call it *exact confidence Intervals*.

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*. Cengage, Boston, MA, 14th edition.

Blitzstein, J. K. and Hwang, J. (2015). *Introduction to Probability*.

Casella, G. and Berger, R. L. (2002). *Statistical Inference*. Thomson Learning, Australia ; Pacific Grove, CA, 2nd edition.

Dalgaard, P. (2008). *Introductory Statistics with R*. Statistics and Computing. Springer New York, New York, NY.

DeGroot, M. H. and Schervish, M. J. (2012). *Probability and Statistics*. Addison-Wesley, Boston, 4th edition.

Griffiths, W. E., Hill, R. C., and Judge, G. G. (1993). *Learning and Practicing Econometrics*. Wiley, New York.

Gujarati, D. N. (2005). *Basic Econometrics*. Tata McGraw-Hill, New Delhi, 4th edition.

Hansen, B. (2022). *Econometrics*. Princeton University Press, Princeton.

References II

Hill, R. C., Griffiths, W. E., and Lim, G. C. (2017). *Principles of Econometrics.* Wiley, Hoboken, 5th edition.

Newbold, P., Carlson, W. L., and Thorne, B. M. (2020). *Statistics for Business and Economics.* Pearson, Harlow, England, 9th, global edition.

Ramachandran, K. M. and Tsokos, C. P. (2020). *Mathematical Statistics with Applications in R.* Elsevier, Philadelphia, 3rd edition.

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

Verbeek, M. (2004). *A Guide to Modern Econometrics.* John Wiley & Sons, Southern Gate, Chichester, West Sussex, England Hoboken, NJ, 2nd edition.

Verzani, J. (2014). *Using R for Introductory Statistics.* Chapman & Hall/CRC the R series. CRC Press, Taylor & Francis Group, Boca Raton, 2nd edition.

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