Chapter- 5

**Introduction to Estimation**

**Objectives:**

* To learn how to estimate certain characteristics of a population from samples;
* To learn the strengths and shortcomings of point estimates and interval estimates;
* To calculate how accurate our estimates really are;
* To learn how to use the t distribution to make interval estimates in some cases when the normal distribution cannot be used; and
* To calculate the sample size required for any desired level of precision in estimation.

**Chapter Contents:**

* Some Important Terminology Used in this Chapter;
* Point Estimates;
* Interval Estimates: Basic Concepts;
* Interval Estimates and Confidence Intervals;
* Calculating Interval Estimates of the Mean from Large Samples;
* Calculating Interval Estimates of the Proportion from Large Samples;
* Interval Estimates Using the t Distribution; and
* Determining the Sample Size in Estimation.

**Basic Terminology**

***Estimate:***

A specific observed value of an estimator is called estimates. The objective of estimation is to determine the approximate value of a population parameter on the basis of a sample statistic. E.g., the sample mean (x) is employed to estimate the population mean (μ).

***Estimator:***

A sample statistic used to estimate a population parameter is called estimator. A good estimator should be unbiased, efficient, consistent and sufficient.

***Point Estimate:***

A single number used to estimate an unknown population parameter.

***Interval Estimate:***

A range of values to estimate an unknown population parameter is called interval estimate.

***Confidence Interval:***

A range of values that has some designated probability of including the true population parameter value.

***Confidence Level:***

The probability that statisticians associate with an interval estimate of a population parameter, indicating how confident they are that the interval estimate will include the population parameter.

***Confidence Limits:***

The upper and lower boundaries of a confidence interval are known as confidence limits.

***Degrees of Freedom:***

The number of values in a sample we can specify freely once we know something about that sample.

**Types of Estimates**

We can make two types of estimates about population: a *point estimate* and an *interval estimate*. A point estimate is a single number that is used to estimate an unknown population parameter. A point estimate is often insufficient, because it is either right or wrong. An interval estimate is a range of values used to estimate a population parameter. It indicates the error in two ways: by the extent of its range and by the probability of the true population parameter lying within that range.

**Point Estimates**

A ***point estimator*** draws inferences about a population by estimating the value of an unknown parameter using a single value or point.

The sample mean *x bar* is best estimator of the population mean μ. It is unbiased, consistent, the most efficient estimator, and, as long as the sample is sufficiently large, its sampling distribution can be approximated by the normal distribution.

*Sample mean,* x̅ =

*Sample Variance, s2* =

**Interval Estimates: Basic Concepts**

An interval estimate describes a range of values within which a population parameter is likely to lie.

An ***interval estimator*** draws inferences about a population by estimating the value of an unknown parameter using an interval.

That is we say (with some \_\_\_% certainty) that the population parameter of interest is between some lower and upper bounds.

*For example* suppose we want to estimate the mean summer income of a class of business students. For n=25 students, sample mean (x̅ = point estimate- certainty) is calculated to be 400 $/week.

*An alternative statement is:* The mean income is between 380 and 420 $/week (interval estimate- an uncertainty).

To provide such an uncertain statement, we need to find out the *standard error of the mean (S.E.)*. The formula for S.E. is given as-

σx̅ =

Where

σx̅ = Standard Error of Mean

Standard deviation of the population

n = no of sample observations.

**Confidence Interval Estimator for μ:**

The probability 1- α is called the ***confidence level (1- α).*** Usually represented as-

x̅±zα/2 = x̅±zα/2 σx̅ = x̅±zα/2SE

**Lower Confidence Level (LCL)** = x̅- zα/2 = x̅- zα/2 σx̅ = x̅- zα/2SE

**Upper Confidence Level (UCL)** = x̅+ zα/2 = x̅ + zα/2 σx̅ = x̅ + zα/2SE

**Note:** The probability is 0.955 that the mean of a sample will be within ± 2 standard errors of the population mean. In other words, 95.5 percent of all the sample means are within ± 2 standard errors (±2σx̅) from μ, and hence μ is within ±2 standard errors of 95.5 percent of all the sample means. So,

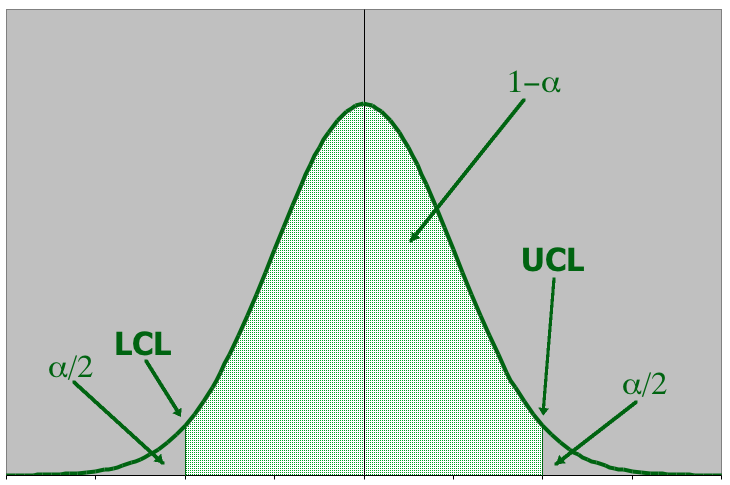
x̅ ± 1σx̅ covers 68.3 percent confidence level;

x̅ ± 2σx̅ covers 95.5 percent confidence level; and

x̅ ± 3σx̅ covers 99.7 percent confidence level.

As far as any particular interval is concerned, it is either contains the population mean or it does not, because the population mean is a fixed parameter.

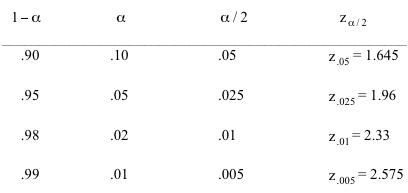
**The *actual* location of the population mean:**



|  |  |  |
| --- | --- | --- |
| … may be here | ….. or here | .or possibly here |

The population mean is a fixed but ***unknown*** quantity. It is incorrect to interpret the confidence interval estimate as a probability statement about. The interval acts as the lower and upper limits of the interval ***estimate*** of the population mean.

**Four commonly used confidence levels:**



***Example 1:***

For a population with a known variance (σ2) of 185, a sample of 64 individuals leads to 217 as an estimate of the mean.

1. Find the standard error of the mean.
2. Establish an interval estimate that should include the population mean 68.3 percent of the time.

**Solution: The Standard Error of Mean, σx̅** =  = = = 1.70

Interval Estimate at 68.3 percent = x̅ ± 1σx̅ = 217±1.70

Lower Confidence Level (LCL) = 217- 1.70 = 215.3

Upper Confidence Level (UCL) = 217+ 1.70 = 218.7

***Example 2:***

Muhammad is a frugal undergraduate management student at King Saud University who is interested in purchasing a used car. He randomly selected 125 want ads and found that the average price of a car in this sample was SR 10,000. He knows that the standard deviation of used- car prices in Riyadh is SR 200.

(a) Establish an interval estimate for the average price of a car so that Muhammad can be 68.3 percent certain that the population mean lies within this interval.

(b) Establish an interval estimate for the average price of a car so that Muhammad can be 95.5 percent certain that the population mean lies within this interval.

**Solution:** (a) **Interval Estimate at 68.3 percent** = x̅ ± 1σx̅

= 10,000± = 10,000± =10,000± = 10,000±17.88

Lower Confidence Level (LCL) = 10,000- 17.88 = 9982.12

Upper Confidence Level (UCL) = 10,000+ 17.88 = 10,017.88

(b) **Interval Estimate at 95.5 percent** = x̅ ± 2σx̅

= 10,000± = 10,000± =10,000± = 10,000±35.77

Lower Confidence Level (LCL) = 10,000- 35.77 = 9964.22

Upper Confidence Level (UCL) = 10,000+ 35.77 = 10,035.77

***Example 3:***

From a population known to have a standard deviation of 1.4, a sample of 60 individuals is taken. The mean for this sample is found to be 6.2.

1. Find the standard error of mean.
2. Establish an interval estimate around the sample mean, using one standard error of the mean.

**Solution: (a) The Standard Error of Mean, σx̅** =  = = = 0.18

(b) **Interval Estimate at 1 S.E.** = x̅ ± 1σx̅ = 6.2±0.18

Lower Confidence Level (LCL) = 6.2- 0.18 = 6.02

Upper Confidence Level (UCL) = 6.2+ 0.18 = 6.38

***Example 4:***

A computer company samples demand during lead time over 25 time periods:



It is known that the standard deviation of demand over lead time is 75 computers. We want to estimate the ***mean*** demand over lead time with 95% confidence in order to set inventory levels…

“We want to estimate the ***mean*** demand over lead time with 95% confidence in order to set inventory levels…”

**Solution:** Thus, the parameter to be estimated is the population mean: μ

And so our confidence interval estimator will be: x̅±zα/2

In order to use our confidence interval estimator, we need the following pieces of data:

|  |  |  |
| --- | --- | --- |
| x̅ | 370.16 | Calculated from the data |
| zα/2 | 1.96 | 1- α = 95% = 0.95, hence α/2 = 0.025, so zα/2 = z0.025 = 1.96 |
| σ | 75 | Given |
| n | 25 | Given |

Therefore:

x̅±zα/2 = : 370.16±z0.025 = 370.16 ± 1.96 = 370.16 ± 29.40

The **lower** and **upper** confidence limits are 340.76 and 399.56.

**When the Population Standard Deviation is Unknown**

***Example 5:***

When we are interested in estimating the mean annual income of 700 families living in four-square- block section of a community. We take a simple random sample and find these results:

n = 50 (sample size)

x̅ = SR 11,800 (Sample mean)

s = SR 950 (Sample standard deviation)

You have to calculate an interval estimate of the mean annual income of all 700 families so that it can be 90 percent confident that the population mean falls within that interval.

**Solution:** The sample size is over 30, so the central limit theorem enables us to use the normal distribution as the sampling distribution. Here, we do not know the population standard deviation, and so we will use the sample standard deviation to estimate the population standard deviation:

σ̂ = s =;

Where σ̂ = Estimate of the population standard deviation

Now we can estimate the standard error of the mean. Because we have a finite population size of 700, and because our sample is more than 5 percent of the population, we will use the formula for deriving the standard error of the mean of finite populations:

σx**̅** =  ×;

But because we are calculating the standard error of the mean using estimate of the standard deviation of the population, we must rewrite this equation so that it is correct symbolically:

σ̂x**̅**= ×

= × = 129.57 (Estimate of the standard error of the mean of a finite population)

Now, we consider the 90 percent confidence level, which would include 45 percent of the area on either side of the mean of the sampling distribution. Looking in the Z- table (area under the standard normal probability distribution between the mean and positive values of z) for the 0.45 value, we find that about 0.45 of the area under the normal curve is located between the mean and a point 1.64 standard errors away from the mean. Therefore, 90 percent of the area is located between plus and minus 1.64 standard errors away from the mean, and our confidence limits are:

x̅ ± 1.64 σ̂x**̅**= 11,800±1.64(129.57)

= 11,800±212.50

Lower Confidence Level (LCL) = 11,800- 212.50 = 11,587.50

Upper Confidence Level (UCL) = 11,800+ 212.50 = 12,012.50

So, we can report with 90 percent confidence that the average annual income of all 700 families living in the four-square- block section falls between SR 11,587.50 and SR 12,012.50.

**Calculating Interval Estimates of the Proportion from Large Samples**

Mean of the Sampling Distribution of the Proportion, μp̅ = p

Standard Error of the Proportion, σp̅ =

Estimated Standard Error of the Proportion, σ̂p̅ =

***Example 6:***

When a sample of 70 retail executives was surveyed regarding the poor November performance of the retail industry, 66 percent believed that decreased sales were due to unseasonably warm temperatures, resulting in consumers’ delaying purchase of cold- weather items.

1. Estimate the standard error of the proportion of retail executives who blame warm weather for low sales.
2. Find the upper and lower confidence limits for this proportion, given a 95 percent confidence level.

**Solution:** n = 70; p̅ = 0.66; so q̅ = 0.34

1. Standard Error of the Proportion, σ̂p̅ = = = 0.0566
2. p̅±1.96 σ̂p̅ = 0.66 ± 1.96 (0.0566) = 0.66 ± 0.111 = (0.549, 0.771)

**Interval Width…**

***A wide interval provides little information***.

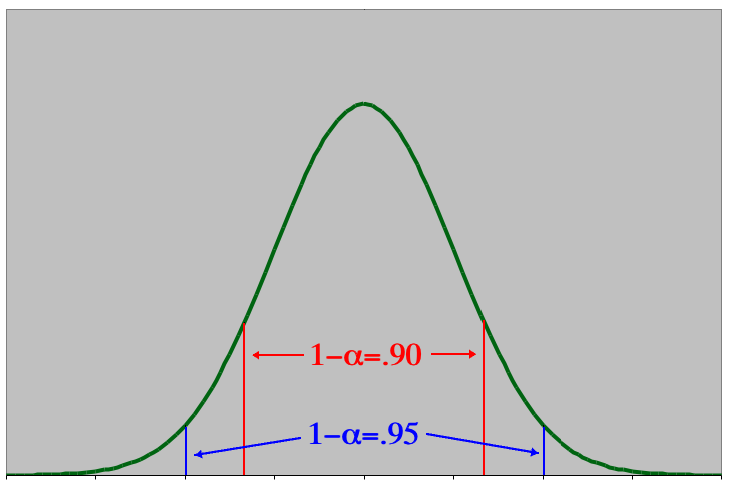
*For example*, suppose we estimate with 95% confidence that an accountant’s average starting salary is between $15,000 and $100,000.

***Contrast*** this with: a 95% confidence interval estimate of starting salaries between $42,000 and $45,000.

The second estimate is much narrower, providing accounting students more precise information about starting salaries.

The width of the confidence interval estimate is a function of the confidence level, the population standard deviation, and the sample size; *x̅±zα/2*where *zα/2*indicates confidence level, ***σ*** shows population standard deviation and ***n*** denotes sample size.

A larger confidence level produces a **wider** confidence interval.



Increasing the sample size decreases the width of the confidence interval while the confidence level can remain unchanged.

**Note:** this also increases the ***cost*** of obtaining additional data

**Selecting the Sample Size:**

We can control the width of the interval by determining the sample size necessary to produce narrow intervals.

Suppose we want to estimate the mean demand “to within 5 units”; i.e. we want to the interval estimate to be: *x̅±5*

Since: *x̅±zα/2*

It follows that *zα/2* = 5

Solving this equation for n, we get

n = (*zα/2*)2 = ()2 = 865

That is, to produce a 95% confidence intervals estimate of the mean (±5 units), we need to sample 865 lead time periods (vs. the 25 data points we have currently).

**Sample Size to Estimate a Mean:**

The general formula for the sample size needed to estimate a population mean with an interval estimate of: *x̅±W*

Requires a sample size of at least this large: ***n = (zα/2)2***

***Question 1:*** A lumber company must estimate the mean diameter of trees to determine whether or not there is sufficient lumber to harvest an area of forest. They need to estimate this to within 1 inch at a confidence level of 99%. The tree diameters are normally distributed with a standard deviation of 6 inches. How many trees need to be sampled?

***Solution:***

**Interval Estimates Using the t Distribution**

How can we handle estimates where the normal distribution is not the appropriate sampling distribution, in other words when we are estimating the population standard deviation and the sample size is 30 or less? In that case we will use t distribution to solve these types of questions.

W. S. Gosset (his pen name was *student*) gave the concept of *t distribution*. It is also known as *student’s t distribution* or simply *student’s distribution*.

*Use of the t distribution for estimating is required whenever the sample size is 30 or less and the population standard deviation is not known. Furthermore, in using the t distribution, we assume that the population is normal or approximately normal.*

**Characteristics of the t Distribution:**

|  |  |
| --- | --- |
| * Like normal distribution, t distribution is also bell- shaped symmetric distribution. * In general, the t distribution is flatter than the normal distribution, and there is a different t distribution for every possible sample size. As the sample size gets larger, the shape of the t distribution loses its flatness and becomes approximately equal to the normal distribution. * A t distribution is lower at the mean and higher at the tails than normal distribution. | http://image.slidesharecdn.com/samplingexcel-13332144521611-phpapp02-120331125837-phpapp02/95/sampling-excel-11-728.jpg?cb=1333199544 |

**Degrees of Freedom (n – 1):**

It can be define as the number of values we can choose freely. For example, assume that we are dealing with two sample values, *a* and *b*, and we know that they have a mean of 20. Symbolically, the situation is = 20. How can we find what values a and b can take on this situation? The answer is that *a* and *b* can be any values whose sum is 40, because 40÷2= 20.

Suppose we learn that *a* has a value of 15. Now b is no longer free to take on any value but must have the value of 25, because

If a = 15, then = 20

= 20 = 15+b = 40 => b = 40 – 15 = 25.

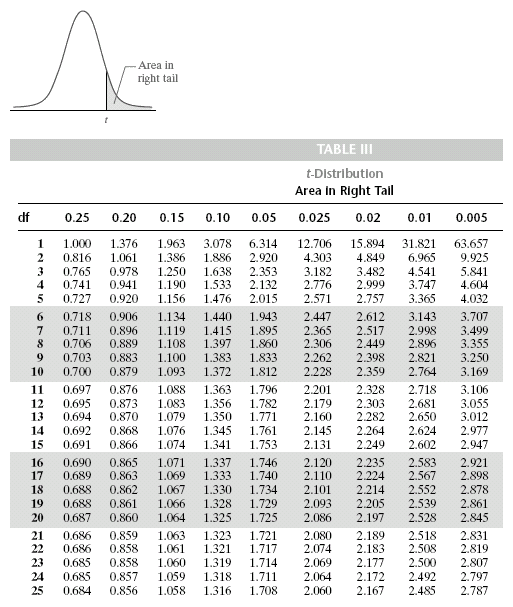
So, we can say that the degree of freedom, or the number of variables we can specify freely, is (n-1) = 2-1 = 1.

**Using the t Distribution Table:**

The t table is more compact and shows areas and t values for only a few percentages (10, 5, 2 and 1 percent). Because there is a different t distribution for each number of degrees of freedom, a more complete table would be quite lengthy.

A second difference in the t table is that it does not focus on the chance that the population parameter being estimated will fall within our confidence interval. Instead, it measures the chance that the population parameter we are estimating will not be within our confidence interval (i.e., it will lie outside it).

A third difference in using the t table is that we must specify the degrees of freedom with which we are dealing. Suppose we make an estimate at 90 percent confidence level with a sample of 15, which is 14 degrees of freedom (n-1). Look in the following table under the 0.10 column until you encounter the row labeled 14. Like a z value, the t value there of 1.345 shows that if we mark off plus and minus 1.345 σ̂x̄’s (estimated standard error of x̅) on either side of the mean, the area under the curve between these two limits will be 90 percent, and the area outside these limits (the chance of error) will be 10 percent.



**Summary of Confidence Limits under Various Conditions**

|  |  |  |
| --- | --- | --- |
|  | When the Population is Finite  (and n/N > 0.05) | When the Population is Infinite  (or n/N > 0.05) |
| Estimating μ (the population mean): when σ (the population standard deviation) is known | Confidence Limits = x̅ ± z × | Confidence Limits = x̅ ± z |
| When σ (the population standard deviation) is not known (σ̂ = s): when n > 30 | Confidence Limits = x̅ ± z × | Confidence Limits = x̅ ± z |
| When n is 30 or less and the population is normal or approximately normal | This case is beyond your course. | Confidence Limits = x̅ ± t |
| Estimating p (the population proportion): when n > 30  σ̂p̅ = | This case is beyond your course. | Confidence Limits = p̅ ± z |