

Chapter II

Estimation Theory

D. 2. 1. (*Estimation*)

The assignment of value(s) to a population parameter based on a value of the corresponding sample statistic is called *estimation*.

D. 2. 2. (*Estimate, Estimator*)

The value assigned to a population parameter based on the value of a sample statistic is called an *estimate*.

The sample statistic used to estimate a population parameter is called an *estimator*.

D. 2. 3. (*Unbiased Estimate, Unbiased Estimator*)

A statistic is called an *unbiased estimator* of a population parameter if the mean of the statistic is equal to the parameter. The corresponding value of the statistic is called an *unbiased estimate* of the parameter.

D. 2. 4. (*Efficient Estimate, Efficient Estimator*)

If the sampling distributions of two statistics have the same mean, the statistic with the smaller variance is called a *more efficient estimator* of the mean. The corresponding value of the efficient statistic is called an *efficient estimate*.

R. 2. 1. (*Estimation Steps*)

The estimation procedure involves the following steps:

1. Select a sample.
2. Collect the required information from the members of the sample.
3. Calculate the value of the sample statistic.
4. Assign value(s) to the corresponding population parameter.

R. 2. 2.

The procedures to be described in this chapter assume that the sample taken is a *simple* random sample.

If the sample is not a simple random sample, then the procedures to be used to estimate a parameter become more complex.

D. 2. 5. (*Point Estimates, Interval Estimates*)

An estimate of a population parameter given by a single number is called a *point estimate* of the parameter.

An estimate of a population parameter given by two numbers between which the parameter may be considered to lie is called an *interval estimate* of the parameter.

Ex. 2. 1.

If we say that the mean housing expenditure per month is 1370.00 €, we are giving a point estimate. If, on the other hand, we say that the mean housing expenditure lies in the interval $[1370.00 - 240.00, 1370.00 + 240.00] = [1130.00, 1610.00]$, we are giving an interval estimate.

D. 2. 6. (Margin of Error)

The number we add to and subtract from the point estimate is called the *margin of error*.

D. 2. 7. (Reliability of an Estimate)

A statement of the error or precision of an estimate is called its *reliability*.

R. 2. 3.

The question arises: What number should we subtract from and add to a point estimate to obtain an interval estimate?

The answer to this question depends on two considerations:

1. The standard deviation of the sample mean.
2. The “level of confidence” to be attached to the interval...

D. 2. 8. (Confidence Interval, Confidence Limits, Confidence Level, Critical Values)

The *confidence interval* is given as:

$$[\text{Point estimate}-\text{Margin of Error}, \text{Point estimate}+\text{Margin of Error}] .$$

The end numbers of the confidence interval are called *confidence limits*.

The *confidence level* associated with a confidence interval states how much confidence we have that this interval contains the true population parameter. The confidence level is denoted by $(1-\alpha)100\%$. When expressed as probability, it is called the *confidence coefficient* and is denoted by $1-\alpha$.

α is called the *significance level*.

R. 2. 4.

Although any value of the confidence level can be chosen to construct a confidence interval, the more common values are 90%, 95%, and 99%. The corresponding confidence coefficients are 0.90, 0.95, and 0.99.

R. 2. 5. (Estimation of Population Mean: Standard Deviation Known)

Here there are three possible cases:

Case I. If the following conditions are fulfilled:

1. The sample size is small (i.e., $n < 30$)
2. The population from which the sample is selected is normally distributed,

then we use the normal distribution to make the confidence interval for μ because from

Chapter 1 the sampling distribution of \bar{X} is normal with its mean equal to μ and the standard deviation equal to $\sigma_x = \frac{\sigma}{\sqrt{n}}$ assuming that $\frac{n}{N} \leq 0.05$.

Case II. If the following two conditions are fulfilled:

The sample size is large (i.e. $n \geq 30$),

then, again, we use the normal distribution to make the confidence interval for μ because from Chapter 1, due to the central limit theorem, the distribution of \bar{x} is (approximately normal with its mean equal to μ and the standard deviation equal to $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$ assuming that $\frac{n}{N} \leq 0.05$.

Case III. If the following three conditions are fulfilled:

1. The sample size is small (i.e., $n < 30$)
2. The population from which the sample is selected is not normally distributed or its distribution is unknown),

then we use a nonparametric method to make the confidence interval for μ . Such a procedure is not covered in this chapter.

R. 2. 6. (Confidence Interval for μ using the Normal Distribution)

The $(1 - \alpha)100\%$ confidence interval for μ under Cases I and II above is

$$\bar{x} \pm z \sigma_{\bar{x}}$$

where

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} .$$

The value of z here is obtained from the standard normal distribution for the given confidence interval

In the following table we give values of z corresponding to various confidence levels used in practice. For confidence levels not presented in the table, the values of z can be found from the normal curve area table:

Confidence Level	z
99.73%	3.0000
99.00%	2.5800
98.00%	2.3300
96.00%	2.0500
95.45%	2.0000
95.00%	1.9600
90.00%	1.6450
80.00%	1.2800
68.27%	1.0000
50.00%	0.6745

D. 2. 9. (Margin of Error)

The *margin of error* for the estimate for μ , denoted by E , is the quantity that is subtracted from and added to the value of \bar{x} to obtain a confidence interval for μ . Thus,

$$E = z \cdot \frac{\sigma}{\sqrt{n}}$$

R. 2. 7. (The Minimum Required Sample Size)

$$n = \left(\frac{z \cdot \sigma}{E} \right)^2.$$

E : maximum error of estimate.

(When solving for example size, any fractional result is always rounded up. Further, unless σ is known *and* the population is normally distributed, any computed sample size below 30 should be increased to 30, because the above formula is based on use of normal distribution.)

R. 2. 8. (Estimation of Population Mean: Standard Deviation Unknown)

Here there are three possible cases:

Case I. If the following conditions are fulfilled:

1. The sample size is small (i.e., $n < 30$)
2. The population from which the sample is selected is normally distributed,

then we use the t distribution to make the confidence interval for μ .

Case II. If the following two conditions are fulfilled:

The sample size is large (i.e., $n \geq 30$),

then we use the t distribution to make the confidence interval for μ .

(Alternatively z can be used as an approximation of t if $n \gg 30$)

Case III. If the following three conditions are fulfilled:

1. The sample size is small (i.e., $n < 30$)
2. The population from which the sample is selected is not normally distributed or its distribution is unknown),

then we use a nonparametric method to make the confidence interval for μ . Such a procedure is not covered in this chapter.

R. 2. 9. (The t Distribution)

The t distribution was developed by W.S. Gosset in 1908 and published under the pseudonym *Student*. The t distribution, also called *Student's t distribution* is a specific type of bell-shaped distribution with a lower height and a wider spread than the standard normal

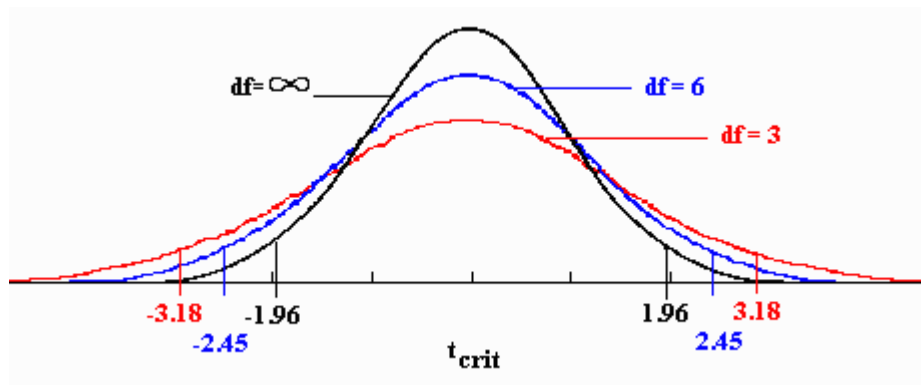
distribution. As the sample size becomes larger, the t distribution approaches the standard normal distribution.

The t distribution has only one parameter, called the *degree of freedom*

$$df := n - 1.$$

The mean of the t distribution is equal to 0 and its standard deviation is $\sqrt{df / (df - 2)} > 1$.

The effect of df on the t distribution is illustrated in the four t distributions below.



R. 2. 10. (*Confidence Interval for μ using the t Distribution*)

The $(1 - \alpha)100\%$ confidence interval for μ is

$$\bar{x} \pm t s_{\bar{x}}$$

where

$$s_{\bar{x}} = \frac{s}{\sqrt{n}}.$$

The value of t is obtained from the t distribution table for $n - 1$ degrees of freedom and the given confidence level.

Here the *margin of error of the estimate* is

$$E := t s_{\bar{x}}.$$

R. 2. 11.

The following table summarises the interval estimation of the population mean:

Interval Estimation of the population Mean

Population	Sample size	σ known	σ unknown
Normally distributed	Large ($n \geq 30$)	$\bar{x} \pm z\sigma_{\bar{x}}$	$\bar{x} \pm ts_i$ or $\bar{x} \pm zs_{\bar{x}}^{**}$
	Small ($n < 30$)	$\bar{x} \pm z\sigma_{\bar{x}}$	$\bar{x} \pm ts_{\bar{x}}$
Not normally distributed	Large ($n \geq 30$)	$\bar{x} \pm z\sigma_{\bar{x}}^*$	$\bar{x} \pm ts_{\bar{x}}^*$ or $\bar{x} \pm zs_{\bar{x}}^{***}$
	Small ($n < 30$)	Nonparametric procedures directed toward the median generally would be used.	

* Central limit theorem is invoked.

** z is used as an approximation of t .

*** Central limit theorem is invoked, and z is used as an approximation of t .

Ex. 2. 2.

We wish to find the average annual salary of all university lecturers in the United States. We take a random sample and find the mean \$53145.

The average salary of all lecturers is assumed to be normally distributed with a standard deviation of \$16451.

1.

Find 95% confidence intervals for the following sample sizes:

a) $n = 36$, b) $n = 64$, c) $n = 144$.

2.

Does the margin of error increase or decrease as the sample size gets larger?

3.

Does the confidence interval get larger or shorter as the sample size increases?

Solution:

We have:

$$\sigma = 16451, \quad \bar{x} = 53145, \quad z = 1.96.$$

a)

$$E = 1.96 \cdot \frac{16451}{\sqrt{36}} = 5374$$

$$\mu \in [53145 - 5374, 53145 + 5374] = [47771, 58519]$$

b)

$$E = 1.96 \cdot \frac{16451}{\sqrt{64}} = 4030$$

$$\mu \in [53145 - 4030, 53145 + 4030] = [49115, 57175].$$

c)

$$E = 1.96 \cdot \frac{16451}{\sqrt{144}} = 2687$$

$$\mu \in [53145 - 2687, 53145 + 2687] = [50458, 55832]$$

2.

The margin of error decreases as the sample size increases.

3.

The confidence interval gets shorter as the sample size increases.

Ex. 2. 3.

Suppose the time it takes for a supermarket cashier to serve customers follows a normal distribution with an unknown mean and a standard deviation of 0.5 minutes. For a random sample of 25 customers, the average time is 5.2 minutes, Calculate the confidence interval at 95% for the average time it takes to serve all customers.

Solution:

We have:

$$\sigma = 0.5, \quad \bar{x} = 5.2, \quad n = 25, \quad z = 1.96.$$

$$\mu \in \left[5.2 - 1.96 \cdot \frac{0.5}{\sqrt{25}}, 5.2 + 1.96 \cdot \frac{0.5}{\sqrt{25}} \right] = [5.004, 5.396].$$

Ex. 2. 4.

Suppose that the standard deviation of the tube life for a particular brand of TV picture tube is known to be 500, but that the mean operating life is not known. The operating life of the tubes cannot be assumed to be normally distributed. For a sample of 35 the mean operating life is 8900 hr.

Determine the 95 percent confidence interval for estimating the population mean.

Solution:

$$\sigma = 500, \quad \bar{x} = 8900, \quad n = 35, \quad z = 1.96.$$

$$\mu \in \left[8900 - 1.96 \cdot \frac{500}{\sqrt{35}}, 8900 + 1.96 \cdot \frac{500}{\sqrt{35}} \right] \approx [8734.35, 9065.65].$$

Ex. 2. 5.

A scientist is interested in monitoring the chemical containments in food. He chooses 50 adults and based on this sample finds that the average intake of food is 756 grams with a standard deviation of 35 grams. Suppose that the population of adults consuming the food can be assumed to be normally distributed.

Construct a 95 percent interval for the mean intake of food for an adult.

Solution:

Using the t -distribution, we have

$$s = 35, \quad \bar{x} = 756, \quad n = 50, \quad df = 49, \quad t_{49,0.05} = 2.011.$$

Therefore

$$\mu \in \left[756 - 2.011 \cdot \frac{35}{\sqrt{50}}, 756 + 2.011 \cdot \frac{35}{\sqrt{50}} \right] \approx [746.05, 765.95].$$

Approximating by the normal distribution, we obtain

$$\mu \in \left[756 - 1.96 \cdot \frac{35}{\sqrt{50}}, 756 + 1.96 \cdot \frac{35}{\sqrt{50}} \right] \approx [746.30, 765.70].$$

Ex. 2. 6.

A sample of 10 accounts drawn at random from a credit department with a population of accounts assumed to be normally distributed in a certain company yield the following results (in euro):

16.39, 16.31, 17.58, 17.21, 18.48,

15.51, 17.68, 18.46, 14.25, 16.88

Obtain a 95% confidence interval for the average credit.

Solution:

We have

$$n = 10, \quad df = 9,$$

$$\bar{x} = \frac{16.39 + 16.31 + 17.58 + 17.21 + 18.48 + 15.51 + 17.68 + 18.46 + 14.25 + 16.88}{10} = 16.875,$$

$$s^2 = \frac{(16.39 - 16.875)^2 + (16.31 - 16.875)^2 + \dots + (16.88 - 16.875)^2}{9} = 1.73931667.$$

$$s = 1.31883155 \approx 1.32$$

Thus, we obtain

$$\mu \in \left[16.875 - 2.26 \cdot \frac{1.32}{\sqrt{10}}, 16.875 + 2.26 \cdot \frac{1.32}{\sqrt{10}} \right] \approx [15.93, 17.82].$$

Ex. 2. 7.

The manager of a restaurant wants to estimate the average amount a customer spends on lunch Monday through Friday.

A random sample of 115 customers' lunch tab gave a mean of 9.74 € with a standard deviation of 2.93 €.

1. Find a 95% confidence interval for the average amount spent on lunch by all customers.
2. For a day when the restaurant has 115 customers, estimate the range of Euro values for the total lunch income that day.

Solution:

$$s = 2.93, \quad \bar{x} = 9.74, \quad n = 115, \quad df = 114, \quad t_{114,0.05} = 1.984$$

1.

$$\mu \in \left[9.74 - 1.984 \cdot \frac{2.93}{\sqrt{115}}, 9.74 + 1.984 \cdot \frac{2.93}{\sqrt{115}} \right] \approx [9.20, 10.28].$$

Approximating by the normal distribution, we get

$$\mu \in \left[9.74 - 1.96 \cdot \frac{2.93}{\sqrt{115}}, 9.74 + 1.96 \cdot \frac{2.93}{\sqrt{115}} \right] \approx [9.204, 10.276].$$

2.

$$[9.20 \cdot 115, 10.28 \cdot 115] = [1058, 1182].$$

Ex. 2. 8.

An airways company advertises that, on average, its New York to London flights is 5 hours and 30 minutes.

A consumer advocacy group randomly selects ten such flights and finds that the sample mean is 5.67 hours with a sample standard deviation of 0.57 hours.

The population is supposed to be normally distributed.

1. What is the point estimate of the population mean?
2. Construct a 95% confidence interval for the population mean.
3. Verify the company's claim.

Solution:

$$s = 0.57, \quad \bar{x} = 5.67, \quad n = 10, \quad df = 9, \quad t_{9,0.05} = 2.262.$$

1.

$$\mu = 5.67.$$

2.

$$\mu \in \left[5.67 - 2.262 \cdot \frac{0.57}{\sqrt{10}}, 5.67 + 2.262 \cdot \frac{0.57}{\sqrt{10}} \right] \approx [5.262, 6.078].$$

3.

$$5.5 \in [5.262, 6.078].$$

\therefore The claim is reasonable.

Ex. 2. 9.

The monthly credit card debts for individual accounts are normally distributed with a standard deviation of 43 dollars. A researcher wishes to estimate the mean monthly credit debt for all individual accounts.

Find the sample size needed to assure with 99% confidence that the sample mean will not differ from the population mean by more than 3 units.

Solution:

$$n \geq \left(\frac{2.58 \cdot 43}{3} \right)^2 = 1367.5204 = 1368.$$

R. 2. 12. (Estimation of Population Proportion: Large Samples)

From Chapter 1, we know that for large samples (In the case of a proportion, a sample is considered to be large if $n \cdot p > 5 \wedge n \cdot (1 - p) > 5$.):

1. The sampling distribution of the sample proportion, p , is (approximately) normal.
2. The mean, μ_p , of the sampling distribution of p is equal to the population proportion, P :

$$\mu_p = P$$

3. The standard deviation of the sample proportion, p , is

$$\sigma_p = \sqrt{\frac{P(1-P)}{n}}.$$

When estimating the value of a population proportion, we do not know the value of P . Consequently, we cannot compute σ_p . Therefore, in the estimation of a population proportion, we use the value of s_p as an estimate of σ_p .

D. 2. 10. (Estimator of the Standard Deviation of p)

s_p is called an *estimator* of σ_p :

$$s_p := \sqrt{\frac{p \cdot (1-p)}{n}}.$$

R. 2. 13. (Confidence Interval for the Proportion, P)

The $(1 - \alpha)100\%$ confidence interval for the population proportion, P , is

$$p \pm z s_p.$$

The value of z used here is obtained from the standard normal distribution table for the given

confidence level, and $s_p := \sqrt{\frac{p \cdot (1 - p)}{n}}$.

(The term $z s_p$ is called the *margin of error*, E .)

Ex. 2. 10

Of 80 adults selected from one town in the United States, 66 have health insurance. Find a 90% confidence interval for all adults in the town who have health insurance.

Solution:

$$p = \frac{66}{80} = 0.825.$$

$$n \cdot p = 66 > 0, \quad n(1 - p) = 80 \cdot (1 - 0.825) = 14 > 5.$$

$$s_p = \sqrt{\frac{0.825 \cdot 0.175}{80}} = 0.0425,$$

$$P \in [0.825 - 1.645 \cdot 0.0425, 0.825 + 1.645 \cdot 0.0425] = [0.755, 0.895].$$

R. 2. 14. (Determining the Sample Size for the Estimation of P)

Given the confidence level and the value of p , the sample size that will produce a predetermined margin of error E of the confidence interval *estimate of P* is

$$n = \frac{z^2 \cdot p \cdot q}{E^2}.$$

We can observe from this formula that to find n , we need to know the value of p . However, the value of p is not known to us. In such a situation, we can choose one of the following alternatives:

1. We make the *most conservative estimate* of the sample size n by using $p = 0.50$. For a given E , this value of p will give us the largest sample size in comparison to any other value of p because the product $p \cdot q = 0.50 \cdot 0.50$ is greater than the product of any other pair of values for p and q .
2. We take a *preliminary sample* (of arbitrarily determined size) and evaluate p and q for this sample. Then we use this value of p to find n .

Ex. 2. 11.

A company has just installed a new machine that makes a part that is used in clocks. The company wants to estimate the proportion of these parts produced by this machine that are defective. The company manager wants this estimate to be within 0.02 of the population proportion for a 95% confidence level.

What is the most conservative estimate of the sample size that will limit the margin of error to within 0.02 of the population proportion?

Solution:

We have $E = \pm 0.02$. Therefore, $E = 0.02$. The z value for a 95% confidence level is 1.96. Hence, the required sample size is

$$n \geq \frac{z^2 \cdot p \cdot q}{E^2} = \frac{(1.96)^2 \cdot 0.5 \cdot 0.5}{(0.02)^2} = 2401.$$

Ex. 2. 12.

Consider example Ex. 2.12. again. Suppose a preliminary sample of 200 parts produced by this machine showed that 7% of them are defective.

How large a sample should the company select so that the 95% confidence interval for P will be within 0.02 of the population proportion?

Solution:

Again, we have $p = \pm 0.02$ and hence, $E = 0.02$. The z value for a 95% confidence level is 1.96.

From the preliminary sample, $p = 0.07$, $q = 0.93$. Using these value, we obtain

$$n = \frac{z^2 \cdot p \cdot q}{E^2} = \frac{(1.96)^2 \cdot 0.07 \cdot 0.93}{(0.02)^2} \approx 625.$$