# Estimation

Lecture # Milind Tirmare

- ***Estimation*** is the assignment of a numerical value to a population parameter or the construction of an interval of numerical values likely to contain a population parameter.

## POINT ESTIMATE

- The value of a sample statistic assigned to an unknown population parameter is called a ***point estimate*** of the parameter.

EXAMPLE 1

The mean starting salary for 10 randomly selected new graduates with a Masters of Business Administration (MBA) at Fortune *500* companies is found to be ***$56,000.*** Fifty-six thousand dollars is a point estimate of the mean starting salary for all new MBA degree graduates at Fortune 500 companies.

The median cost for 350 homes selected from across the United States is found to equal **$1** 15,OOO. The value of the sample median, $1 **15,000,** is a point estimate of the median cost of a home in the United States.

**A** survey of 950 households finds that 35% have a home computer. Thirty-five percent is a point estimate of the percentage of homes that have a home computer

# INTERVAL ESTIMATE

- In addition to a point estimate, it is desirable to have some idea of the size of the sampling error, that is the difference between the population parameter and the point estimate. **By** utilizing the standard error of the sample statistic and its sampling distribution, *ainterval estimate* for the population parameter may be developed.

- A*confidence interval* is an interval estimate that consists of an interval of numbers obtained from the point estimate of the parameter along with a percentage that specifies how confident we are that the value of the parameter lies in the interval. The confidence percentage **is** called the *confidence level*

# CONFIDENCE INTERVAL FOR THE POPULATION MEAN: LARGE SAMPLES

According to the central limit theorem, the sample mean, $\bar{x}$, has a normal distribution provided the sample size is 30 or more. Furthermore, the mean of the sample mean equals the mean of the population and the standard error of the mean equals the population standard deviation divided by the square root of the sample size.

The variable z has a standard normal distribution provided n ⩾ **30**

$$z = \frac{\bar{x} - \mu}{\sigma_{\bar{x}}} \qquad (1)$$

Since 95% of the area under the standard normal curve is between z = −1.96 and **z** = 1.96, and since the variable z in formula (1) has a standard normal distribution,

$$P(-1.96 < \frac{\bar{x} - \mu}{\sigma_{\bar{x}}} < 1.96) = .95$$

The inequality, $-1.96 < \frac{\bar{x} - \mu}{\sigma_{\bar{x}}} < 1.96$, is solved for $\mu$ and the result is

$$\bar{x} - 1.96\sigma_{\bar{x}} < \mu < \bar{x} + 1.96\sigma_{\bar{x}} \qquad (2)$$

The interval given in formula (2) is called a 95% confidence interval for the population mean, $\mu$

The general form for the interval is shown in formula *(3),* where z represents the proper value from the standard normal distribution table as determined by the desired confidence level

$$\bar{x} - z\sigma_{\bar{x}} < \mu < \bar{x} + z\sigma_{\bar{x}} \qquad (3)$$

EXAMPLE 2

The mean age of policyholders at Mutual Insurance Company is estimated by sampling the records of 75 policyholders. The standard deviation of ages is known to equal 5.5 years and has not changed over the years. However, it is unknown if the mean age has remained constant. The mean age for the sample of 75 policyholders is 30.5 years

The standard error of the ages is $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{5.5}{\sqrt{75}} = .635$ years.

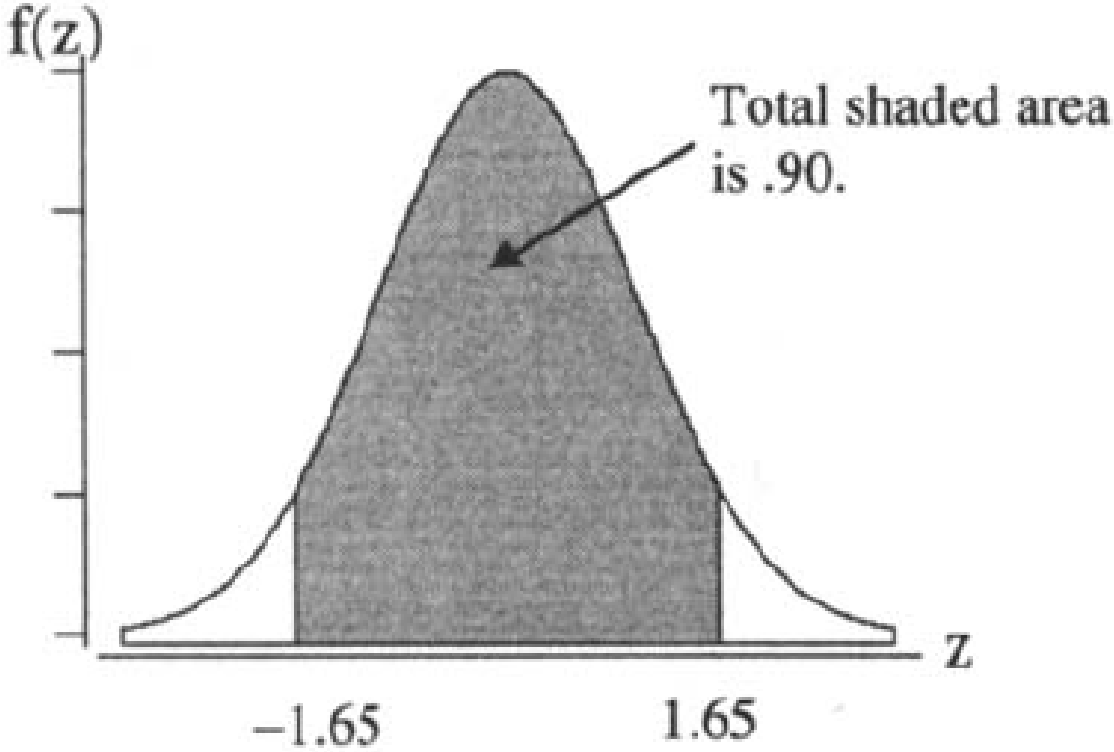
In order to find a 90% confidence interval for $\mu$, it is necessary to find the value of z in formula (3) for confidence level equal to 90%. If we let c be the correct value for z, then we are looking for that value of c which satisfies the equation

$$P(-c < z < c) = .90$$

Or

because of the symmetry of the z curve, we are looking for that value of c which satisfies the equation $P(0 < z < c) = .45$. From the standard normal distribution table, we find $P(0 < \mathbf{z} < 1.64) = .4495$ and $P(0 < z < 1.65) = .4505$

The interpolated value for c is 1.645, which we round to 1.65




The 90% confidence interval is ,

The lower limit of the interval is , $\bar{x} - 1.65\sigma_{\bar{x}} = 30.5 - 1.65 \times .635 = 29.5$ years.

The upper limit is $\bar{x} + 1.65\sigma_{\bar{x}} = 31.5.$

We are 90% confident that the mean age of all 250,000 policyholders is between 29.5 and 3 1.5 years.

To say we are 90% confident that the mean age of all policyholders is between 29.5 and 31.5 years means that if this study were conducted a large number of times and a confidence interval were computed each time, then 90% of all the possible confidence intervals would contain the true value of $\mu$

Since it is time consuming to determine the correct value of **z** in formula *(3)*, the values for the most often used confidence levels are given in Table 1

Table 1

| Confidence level | Z value |
|---|---|
| 80 | 1.28 |
| 90 | 1.65 |
| 95 | 1.96 |
| 99 | 2.58 |

EXAMPLE 3

The 99% response time for terrorists bomb threats was investigated. No historical data existed concerning the standard deviation or mean for the response times. A sample of 35 response times was obtained and it was found that the sample mean was 8.5 minutes and the standard deviation was 4.5 minutes.

When $\sigma$ is unknown and the sample size is 30 or more, the standard deviation of the sample itself is used in place of $\sigma$ when constructing a confidence interval

The estimated standard error of the mean is represented by $s_{\overline{x}}$

$$s_{\overline{x}} = \frac{s}{\sqrt{n}} = \frac{4.5}{\sqrt{35}} = .76$$

From Table 1, the value for z is 2.58

The lower limit for the 99% confidence interval is

$$\bar{x} - 2.58 \times s_{\bar{x}} = 8.5 - 2.58 \times .76 = 6.54$$

The upper limit is $8.5 + 2.58 \times .76 = 10.46.$

We are 99% confident that the mean response time is between 6.54 minutes and 10.46 minutes

- **For large samples, no assumption is made concerning the shape of the population distribution. If the population standard deviation is known, it is used in formula ( 3). If the population standard deviation is unknown, it is estimated by using the sample standard deviation**

## MAXIMUM ERROR OF ESTIMATE FOR THE POPULATION MEAN

The inequality in formula (3) may be expressed as shown:

$$|\bar{x} - \mu| < z\sigma_{\bar{x}} \qquad (4)$$

The left-hand side of formula ( 4) is the *sampling error* when $\bar{x}$ is used as a point estimate of $\mu$. The right-hand side of formula ( 4) is the *maximum error of estimate or margin of error* when $\bar{x}$ is used as a point estimate of $\mu$.

That is, when $\bar{x}$ is used as a point estimate of $\mu$, the maximum error of estimate or margin of error, E, is

$$E = z\sigma_{\bar{x}} \qquad (5)$$

When the confidence level is 95%, $z = 1.96$ and $E = 1.96\sigma_{\bar{x}}$. This value of E, $1.96\sigma_{\bar{x}}$, is called the 95% *margin of error* or simply *margin of error* when $\bar{x}$ is used as a point estimate of $\mu$.

EXAMPLE **4**

The annual college tuition costs for 40 community colleges selected from across the United States are given in Table 2. The mean for these 40 sample values is $1396, the standard deviation of the 40 values is $655

Table 2

| | | | |
|---|---|---|---|
| 1200 | 850 | 1750 | 930 |
| 850 | 3000 | 1650 | 1640 |
| 1700 | 2100 | 900 | 1320 |
| 1500 | 500 | 2050 | 1750 |
| 700 | 500 | 1780 | 2500 |
| 1200 | 1950 | 675 | 2310 |
| 1500 | 1000 | 1080 | 2900 |
| 2000 | 950 | 680 | 1875 |
| 1950 | 560 | 900 | 1450 |
| 750 | 500 | 1500 | 950 |

the estimated standard error is $s_{\bar{x}} = \frac{s}{\sqrt{n}} = \frac{655}{\sqrt{40}} = \$104.$

The mean tuition cost for all community colleges in the United States is represented by $\mu$

A point estimate of $\mu$ is given by \$1,396.

The margin of error associated with this estimate is 1.96 x 104 = \$204

The 95% confidence interval for $\mu$, based upon these data goes from

1396 - 204 = \$1,192 to 1396 + 204 = \$1,600.

The margin of error is actually ±\$204, since the error may occur in either direction