BITS Pilani

Pilani|Dubai|Goa|Hyderabad

Introduction to Statistical Methods

Team ISM

Session No 8

Central Limit Theorem

(25th /26th June ,2022)

Central Limit Theorem

- If samples of size *n* are drawn randomly from a population that has a mean of μ and a standard deviation of σ, the sample means, $\bar{x}$, are approximately normally distributed for sufficiently large sample sizes (*n* >= 30) regardless of the shape of the population distribution.

- If the population is normally distributed, the sample means are normally distributed for any size sample.

- From mathematical expectation

$$\mu_{\bar{x}} = \mu \qquad \sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

Z score for sample means

- The central limit theorem states that sample means are normally distributed regardless of the shape of the population for large samples and for any sample size with normally distributed populations.
- Thus, **sample means** can be **analyzed** by using ***z* scores**
- The formula to determine z scores for individual values from a normal distribution:

$$z = \frac{x - \mu}{\sigma}$$

- If sample means are normally distributed, the z score formula applied to sample means would be

$$z = \frac{\bar{x} - \mu_{\bar{x}}}{\sigma_{\bar{x}}}$$

- The standard deviation of the statistic of interest is $\sigma_{\bar{x}}$, sometimes referred to as the **standard error of the mean.**

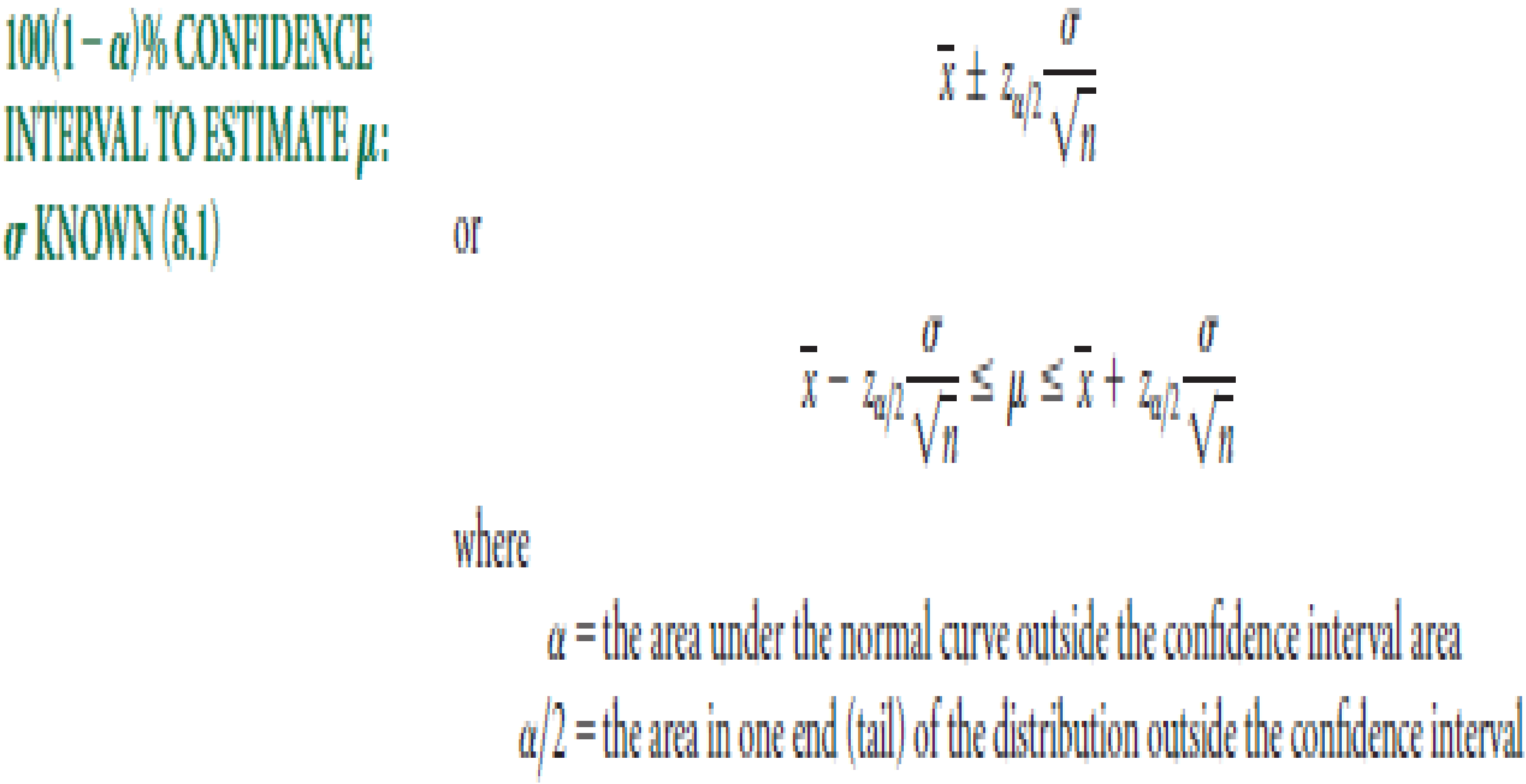
Point Estimate

- A **point estimate** is a statistic taken from a sample that is used to estimate a population parameter.
- A point estimate is only as good as the representativeness of its sample.
- If other random samples are taken from the population, the point estimates derived from those samples are likely to vary.

Interval Estimate

- Because of variation in sample statistics, estimating a population parameter with an interval estimate is often preferable to using a point estimate.
- An interval estimate (confidence interval) is a range of values within which the analyst can declare, with some confidence, the population parameter lies.

Confidence Interval to Estimate μ



100(1 − α)% CONFIDENCE INTERVAL TO ESTIMATE μ: σ KNOWN (8.1)

$$\bar{x} \pm z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$$

or

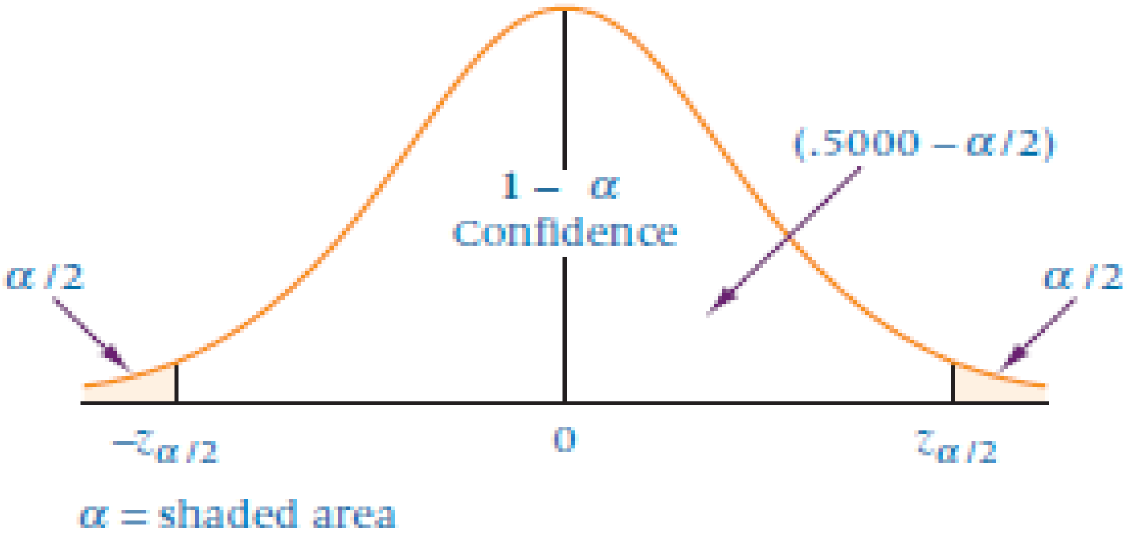
$$\bar{x} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \le \mu \le \bar{x} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$$

where

α = the area under the normal curve outside the confidence interval area

$\alpha/2$ = the area in one end (tail) of the distribution outside the confidence interval

Confidence Intervals

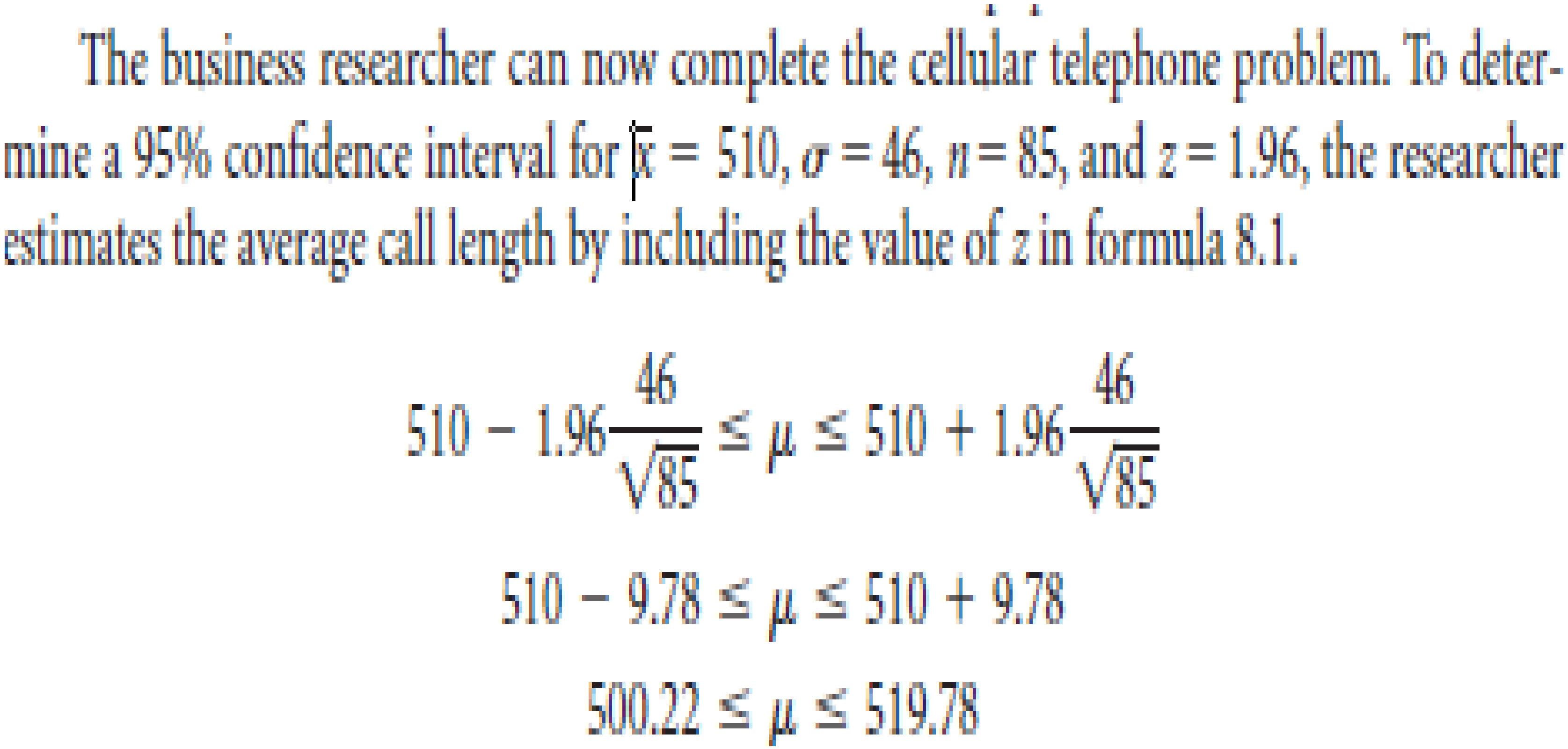


Example

- In the cellular telephone company problem of estimating the population mean number of minutes called per residential user per month, from the sample of 85 bills it was determined that the sample mean is 510 minutes.

- Suppose past history and similar studies indicate that the population standard deviation is 46 minutes.
- Determine a 95% confidence interval.

Solution



The business researcher can now complete the cellular telephone problem. To determine a 95% confidence interval for $\bar{x} = 510$, $\sigma = 46$, $n = 85$, and $z = 1.96$, the researcher estimates the average call length by including the value of z in formula 8.1.

$$510 - 1.96\frac{46}{\sqrt{85}} \le \mu \le 510 + 1.96\frac{46}{\sqrt{85}}$$

$$510 - 9.78 \le \mu \le 510 + 9.78$$

$$500.22 \le \mu \le 519.78$$

Solution

- The confidence interval is constructed from the point estimate, which in this problem is 510 minutes, and the error of this estimate, which is 9.78 minutes.

- The resulting confidence interval is 500.22 <= µ <= 519.78.

- The cellular telephone company researcher is 95%, confident that the average length of a call for the population is between 500.22 and 519.78 minutes.

Exercise

- A survey was taken of U.S. companies that do business with firms in India. One of the questions on the survey was: Approximately how many years has your company been trading with firms in India?

- A random sample of 44 responses to this question yielded a mean of 10.455 years. Suppose the population standard deviation for this question is 7.7 years.

- Using this information, construct a 90% confidence interval for the mean number of years that a company has been trading in India for the population of U.S. companies trading with firms in India.

Solution

Here, n= 44, $\bar{x}$= 10.455 and σ= 7.7. To determine the value of $z_{\alpha/2}$, divide the 90% confidence in half, or take .5000 - α/2= .5000 - .0500= 0.45 where α= 10%.

Z table yields a z value of 1.645 for the area of .45

The confidence interval is

$$\bar{x} - z\frac{\sigma}{\sqrt{n}} \le \mu \le \bar{x} + z\frac{\sigma}{\sqrt{n}}$$

$$10.455 - 1.645\frac{7.7}{\sqrt{44}} \le \mu \le 10.455 + 1.645\frac{7.7}{\sqrt{44}}$$

$$10.455 - 1.910 \le \mu \le 10.455 + 1.910$$

$$8.545 \le \mu \le 12.365$$

Exercise

A study is conducted in a company that employs 800 engineers. A random sample of 50 engineers reveals that the average sample age is 34.3 years. Historically, the population standard deviation of the age of the company's engineers is approximately 8 years.

Construct a 98% confidence interval to estimate the average age of all the engineers in this company.

Solution

- This problem has a finite population. The sample size, 50, is greater than 5% of the population, so the finite correction factor may be helpful.

- In this case N = 800, n = 50, $\bar{x}$= 34.3 and σ= 8

- The z value for a 98% confidence interval is 2.33

$$34.30 - 2.33\frac{8}{\sqrt{50}}\sqrt{\frac{750}{799}} \le \mu \le 34.30 + 2.33\frac{8}{\sqrt{50}}\sqrt{\frac{750}{799}}$$

$$34.30 - 2.55 \le \mu \le 34.30 + 2.55$$

$$31.75 \le \mu \le 36.85$$

Estimating The Population Proportion

- Methods similar to those used earlier can be used to estimate the population proportion.
- The central limit theorem for sample proportions led to the following formula

$$z = \frac{\hat{p} - p}{\sqrt{\frac{p \cdot q}{n}}}$$

- where q = 1 - p. Recall that this formula can be applied only when n*p and n*q are greater than 5.

- for confidence interval purposes only and for large sample sizes— is substituted for p in the denominator, yielding

$$z = \frac{\hat{p} - p}{\sqrt{\frac{\hat{p} \cdot \hat{q}}{n}}}$$

Confidence Interval To Estimate P

$$\hat{p} - z_{\alpha/2}\sqrt{\frac{\hat{p}\cdot\hat{q}}{n}} \le p \le \hat{p} + z_{\alpha/2}\sqrt{\frac{\hat{p}\cdot\hat{q}}{n}}$$

where

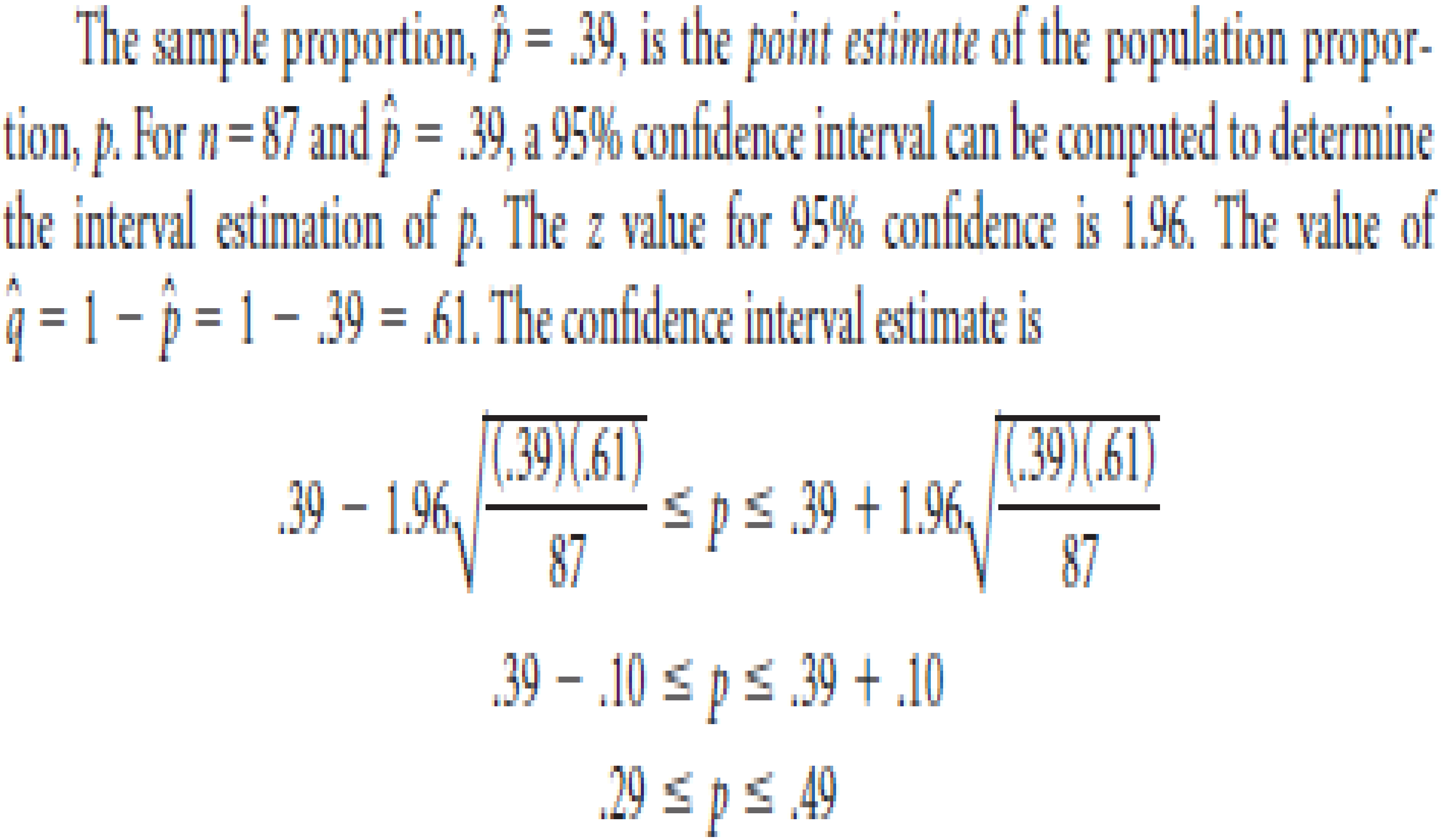
$\hat{p} =$ sample proportion
$\hat{q} = 1 - \hat{p}$
$p =$ population proportion
$n =$ sample size

In this formula, $\hat{p}$ is the point estimate and $\pm z_{\alpha/2}\sqrt{\frac{\hat{p}\cdot\hat{q}}{n}}$ is the error of the estimation.

Example

- A study of 87 randomly selected companies with a telemarketing operation revealed that 39% of the sampled companies used telemarketing to assist them in order processing. Using this information, how could a researcher estimate the *population* proportion of telemarketing companies that use their telemarketing operation to assist them in order processing?

Solution



The sample proportion, $\hat{p} = .39$, is the *point estimate* of the population proportion, p. For $n = 87$ and $\hat{p} = .39$, a 95% confidence interval can be computed to determine the interval estimation of p. The z value for 95% confidence is 1.96. The value of $\hat{q} = 1 - \hat{p} = 1 - .39 = .61$. The confidence interval estimate is

$$.39 - 1.96\sqrt{\frac{(.39)(.61)}{87}} \le p \le .39 + 1.96\sqrt{\frac{(.39)(.61)}{87}}$$

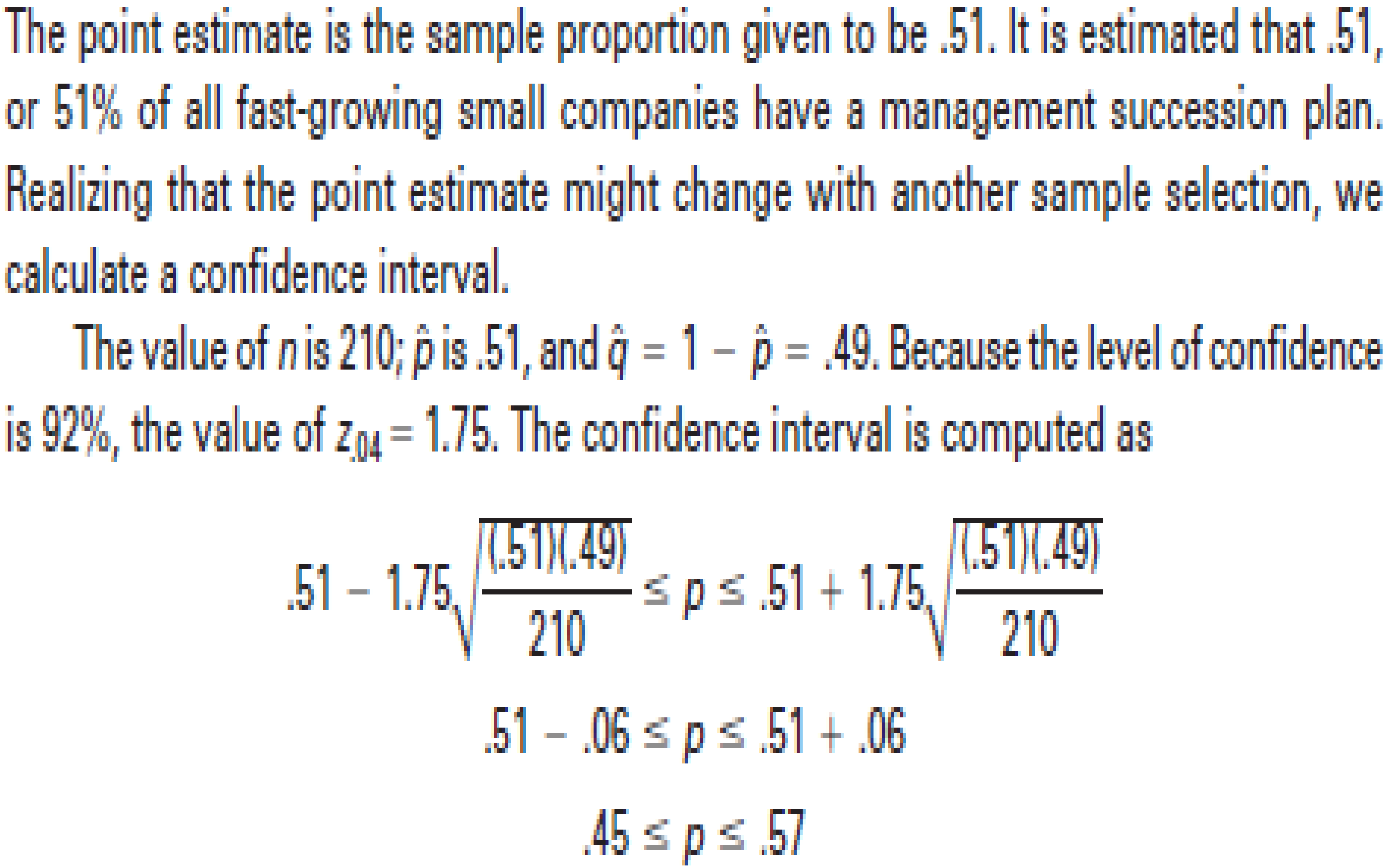
$$.39 - .10 \le p \le .39 + .10$$

$$.29 \le p \le .49$$

Exercise

- Coopers & Lybrand surveyed 210 chief executives of fast-growing small companies. Only 51% of these executives had a management succession plan in place. A spokesperson for Cooper & Lybrand said that many companies do not worry about management succession unless it is an immediate problem. However, the unexpected exit of a corporate leader can disrupt and unfocus a company for long enough to cause it to lose its momentum. Use the data given to compute a 92% confidence interval to estimate the propor

Solution



The point estimate is the sample proportion given to be .51. It is estimated that .51, or 51% of all fast-growing small companies have a management succession plan. Realizing that the point estimate might change with another sample selection, we calculate a confidence interval.

The value of n is 210; $\hat{p}$ is .51, and $\hat{q} = 1 - \hat{p} = .49$. Because the level of confidence is 92%, the value of $z_{.04} = 1.75$. The confidence interval is computed as

$$.51 - 1.75\sqrt{\frac{(.51)(.49)}{210}} \le p \le .51 + 1.75\sqrt{\frac{(.51)(.49)}{210}}$$

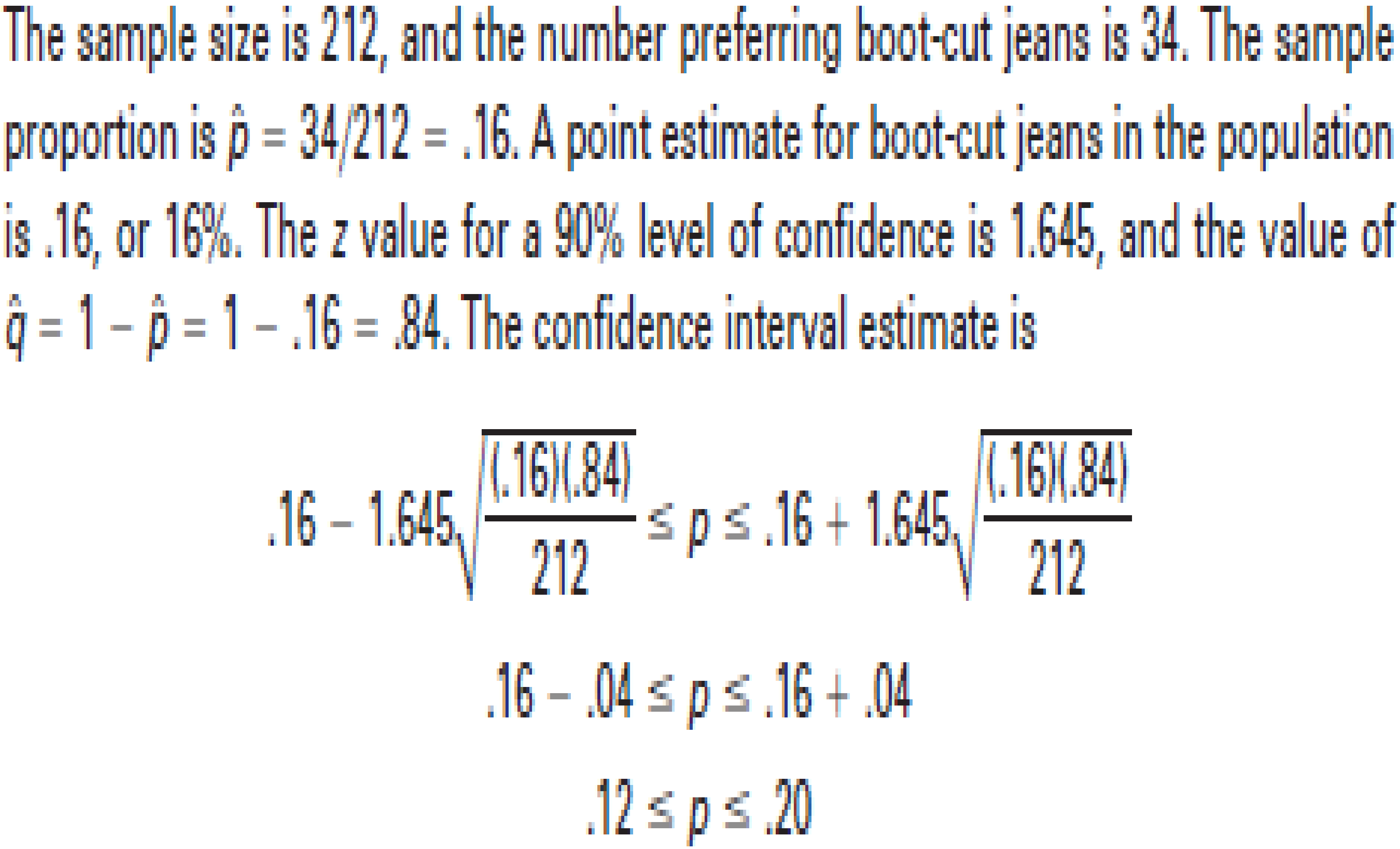
$$.51 - .06 \le p \le .51 + .06$$

$$.45 \le p \le .57$$

Exercise

- A clothing company produces men’s jeans. The jeans are made and sold with either a regular cut or a boot cut. In an effort to estimate the proportion of their men’s jeans market in Oklahoma City that prefers boot-cut jeans, the analyst takes a random sample of 212 jeans sales from the company’s two Oklahoma City retail outlets. Only 34 of the sales were for boot-cut jeans. Construct a 90% confidence interval to estimate the proportion of the population in Oklahoma City who prefer boot-cut jeans.

Solution



The sample size is 212, and the number preferring boot-cut jeans is 34. The sample proportion is $\hat{p} = 34/212 = .16$. A point estimate for boot-cut jeans in the population is .16, or 16%. The z value for a 90% level of confidence is 1.645, and the value of $\hat{q} = 1 - \hat{p} = 1 - .16 = .84$. The confidence interval estimate is

$$.16 - 1.645\sqrt{\frac{(.16)(.84)}{212}} \le p \le .16 + 1.645\sqrt{\frac{(.16)(.84)}{212}}$$

$$.16 - .04 \le p \le .16 + .04$$

$$.12 \le p \le .20$$

Estimating The Population Variance

- Sample variance is computed by using the formula $$s^2 = \frac{\sum(x - \overline{x})^2}{n - 1}$$

- Mathematical adjustment is made in the denominator by using n - 1 to make the sample variance an unbiased estimator of the population variance.

- Suppose a researcher wants to estimate the population variance from the sample variance in a manner that is similar to the estimation of the population mean from a sample mean.

- The relationship of the sample variance to the population variance is captured by the **chi-square distribution**

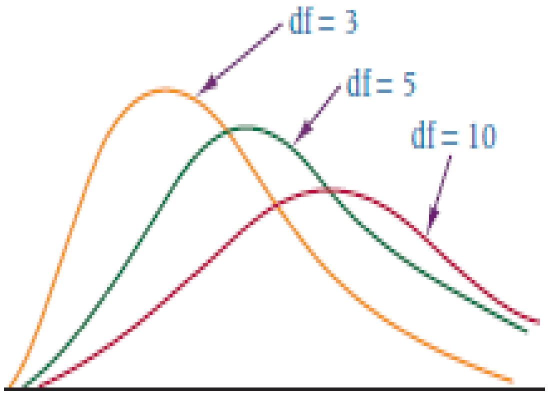
Chi-square Statistic

- Although the technique is still rather widely presented as a mechanism for constructing confidence intervals to estimate a population variance, you should proceed with extreme caution and **apply** the technique **only** in cases **where** the **population** is known to be **normally distributed**. We can say that this technique lacks robustness.

χ^2 FORMULA FOR SINGLE VARIANCE (8.5)

$$\chi^2 = \frac{(n-1)s^2}{\sigma^2}$$

$$\text{df} = n - 1$$



Three Chi-Square Distributions

Thanks