Date : / /
Page No. :

* Statistical Inference.

The process of drawing Inference about population on the basis of sample data is called Statistical Inference.

→ It is also called Inferential statistics.

* Types of Statistical Inference.

1) Estimation.

2) Testing of Hypothesis.

(*) Estimation.

A process in which we obtain the values of unknown population parameters with the help of sample data

→ Population means totality of any data analysis.

→ sample means small part of population.

Estimate:

An estimate is the numeric value of the estimator.

Estimator:

It is a rule, formula or function that tells how to calculate an estimate.

Types of Estimation

1) Point Estimation 2) Interval Estimation.

⊛ Some Important Notations

→ For sample data.

1) n = Sample size.
2) $\bar{x}$ = sample mean
3) S^2 = Biased sample Variance. (Capital)
4) s^2 = Unbiased sample variance. (Small)
5) S = Sample standard deviation.
6) p = Sample proportion.
7) $S_{\bar{x}} = \frac{s}{\sqrt{n}}$ (sample standard Error)

→ For population Data

1) N = Population Size.
2) μ = Population mean.
3) σ^2 = Population Variance.
4) σ = Population standard deviation.
5) π = Population proportion.
6) α = Level of significance.
7) $1-\alpha$ = Level of confidence OR Confidence Interval.

* Point Estimation of Parameters and Sampling Distributions.

☆ Point Estimation:

When an estimate for the unknown population parameter is expressed by a single value it is called point Estimate.

☆ Interval Estimate:

When an estimate for the unknown population parameter is expressed by a range of values within which the population parameter is expected to occur is called an Interval Estimate.

e.g. when we find estimate for population mean (μ) we use sample mean ($\bar{x}$)

$$\bar{x} = \frac{\Sigma x}{n}$$ → Estimator. (Rule, formula or function)

Suppose $\bar{x} = 10$

i.e 10 is a numeric value of Estimator is called an Estimate & our answer is in single numeric form it is also called point estimator.

&

The whole process is called an Estimation.

Ex-1 A random sample of $n=6$ has the elements 6, 10, 13, 14, 18 and 20. compute a point estimate of
1) Population mean
2) The population standard deviation
3) the standard error of the mean

Sol

X
6
10
13
14
18
20
$\Sigma X = 81$

1) The sample mean is

$$\bar{x} = \frac{\Sigma X}{n} = \frac{81}{6} = 13.5$$

$$\boxed{\bar{x} = 13.5}$$

So the point estimate of population mean (μ) is 13.5 and $\bar{x}$ is estimator

X^2
36
100
169
196
324
400
$\Sigma X^2 = 1225$

2) The sample S.D is

$$s = \sqrt{\frac{\Sigma x^2}{n} - \left(\frac{\Sigma x}{n}\right)^2}$$

$$= \sqrt{\frac{1225}{6} - \left(\frac{81}{6}\right)^2}$$

$$= \sqrt{204.1667 - 182.25}$$

$$= \sqrt{21.9167}$$

$$= 4.68$$

So the point Estimate of the population standard deviation σ is 4.68 & s is

the sample standard deviation

3) The standard error of the mean.

the sample standard error of the mean is

$$S_{\bar{x}} = \frac{S}{\sqrt{n}} = \frac{4.68}{\sqrt{6}} = 1.91$$

∴ The point estimate of population standard error is 1.91.

Ex-2 A random sample of 35 airfare prices (in dollar) for a one way ticket from Atlanta to chicago Find a point estimate for the population mean.
99, 102, 105, 105, 104, 95, 100, 114, 108, 103, 94, 105, 101, 109, 103, 98, 96, 98, 104, 87, 101, 106, 103, 90, 107, 98, 101, 107, 105, 94, 111, 104, 87, 117, 101.

Ex-3 Manufacture of certain component requires three different maching operations. The total time for manufacturing one such component is known to have a normal distribution. However, the mean μ and variance σ^2 for the normal distribution are unknown. If we did an experiment in which we manufactured 10 components and record the operation in which we manufactured 10 components and record the operation time and the sample time is given as following
63.8, 60.5, 65.3, 65.7, 61.9, 68.2, 68.1, 64.8, 65.8, 65.4.
Find point estimate of population mean, population variance & the standard error of the mean.

(*) Central Limit th^m.

If samples of size n are drawn randomly from a population having mean μ and S.D. σ (Not necessarily be normal) then the sampling distribution of the mean $\bar{x}$ is approximately distributed with

mean $\mu_{\bar{x}} = \mu$ and

S.D $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$ (Also known as standard error of mean)

Provided that the sample size n is large enough (usually $n \geq 30$) (as $n \to \infty$)

Hence, the variable

$$\boxed{Z = \frac{\bar{x} - \mu_{\bar{x}}}{\sigma_{\bar{x}}}} = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} \sim N(0,1)$$ (Z formula for sample mean)

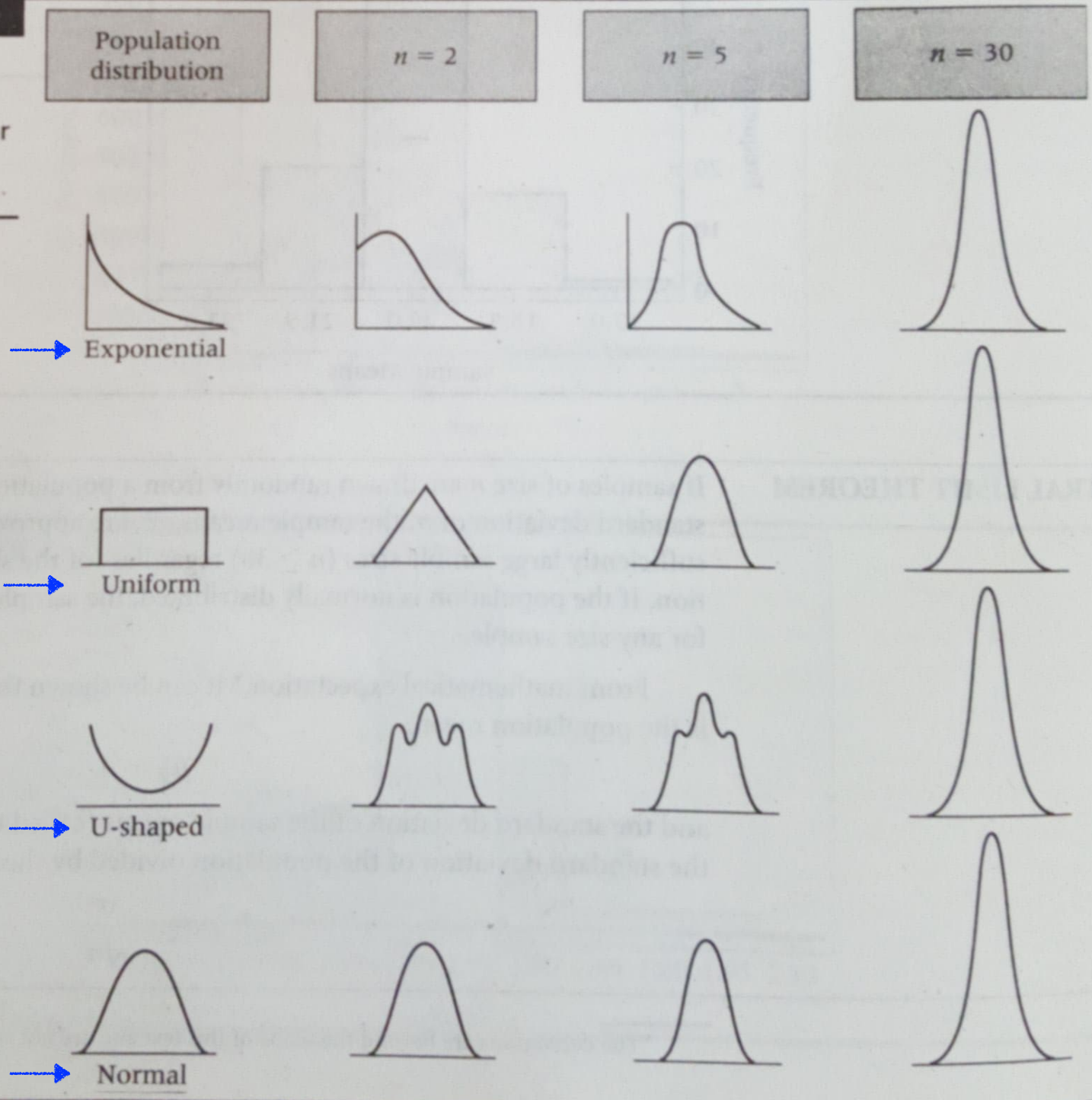
has a standard Normal distribution.

Note

1) If population distribution is normal distribution then for any sample size distribution is normal
2) If population distribution is not normal then take sample size $n \geq 30$ to normalized the data.
3) When sample size increases to infinity, the distribution of sample means literally becomes normal in shape and S.D will decreases.

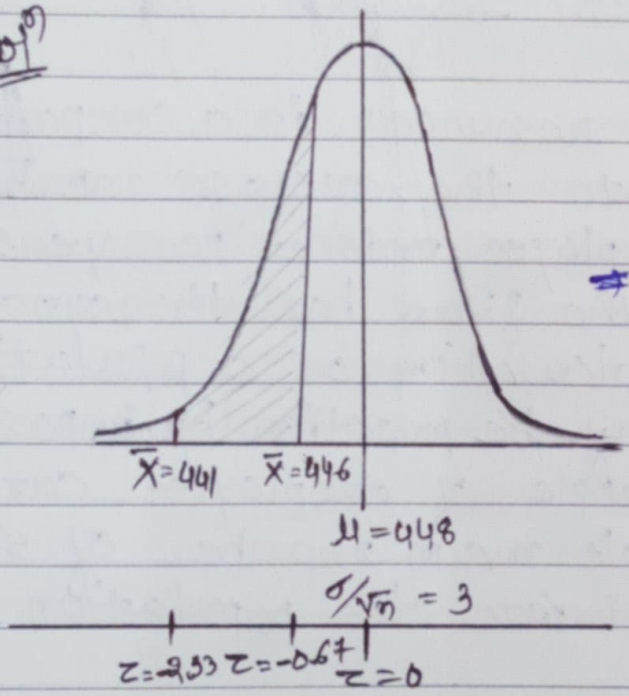
FIGURE 7.6

Shapes of the Distributions of Sample Means for Three Sample Sizes Drawn from Four Different Population Distributions



Ex-1 Suppose that during any hour in a large department store, the average number of shoppers is 448, with a standard deviation of 21 shoppers. What is the probability that a random sample of 49 different shopping hours will yield a sample mean bet^{n} 441 and 446 of shoppers?

Sol^{n}



Here $\mu = 448$, $\sigma = 21$, $n = 49$

$$P(441 \le \bar{x} \le 446)$$

$z = (\bar{x} - \mu\bar{x}) / (\sigma/\sqrt{n})$

$$= P\left(\frac{441-448}{21/\sqrt{49}} \le z_{\bar{x}} \le \frac{446-448}{21/\sqrt{49}}\right)$$

$$= P(-2.33 \le z_{\bar{x}} \le -0.67)$$

$$= P(0.67 \le z_{\bar{x}} \le 2.33)$$

$$= P(0 \le z_{\bar{x}} \le 2.33) - P(0 \le z_{\bar{x}} \le 0.67)$$

$$= 0.4901 - 0.2486$$

$$= 0.2415$$

x 100

∴ There is a 24.15% chance of randomly selecting 49 hourly periods for which the sample mean is bet^{n} 441 and 446 shoppers.

(*) Estimation for single population

→ Estimating the population mean using the z-statistic.

on many occasions estimating the population mean is useful in business research.

e.g. The manager of human resources in a company might want to estimate the average number of days of work an employee misses per year because of illness. If the firm has thousands of employees, direct calculation of a population mean such as this may be practically impossible. Instead, a random sample of employee can be taken, and the sample mean number of sick days can be used to estimate the population mean.

→ A point estimate is a statistic taken from a sample that used to estimate a population parameter. A point estimate is only as good as the representativeness of its sample. If other sample (random) are taken from population, the point estimates derived from those samples are likely to vary. 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 intervals can be two-sided or one-sided.
How are confidence intervals constructed ?.

As a result of the central limit theorem, the following z formula for sample means can be used when sample size are large, regardless of the shape of population distribution, or for smaller size if the population is normally distributed.

$$Z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}}$$

Rearranging this formula algebraically to solve for μ gives.

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

Because a sample mean can be greater than or less than the population mean, z can be positive or negative.
Thus the preceding expression takes the form

$$\bar{x} \pm z\frac{\sigma}{\sqrt{n}}$$

Rewriting this expression yields the confidence Interval formula for estimating μ with large sample sizes.

$100(1-\alpha)\%$ confidence interval to estimate μ.

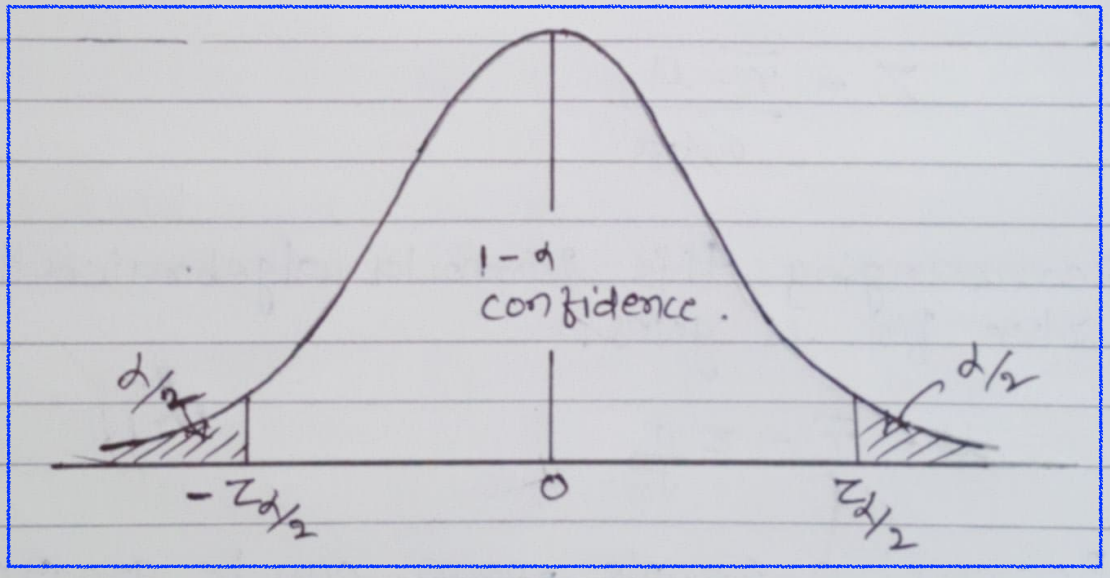
$$\bar{x} \pm z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$$

OR

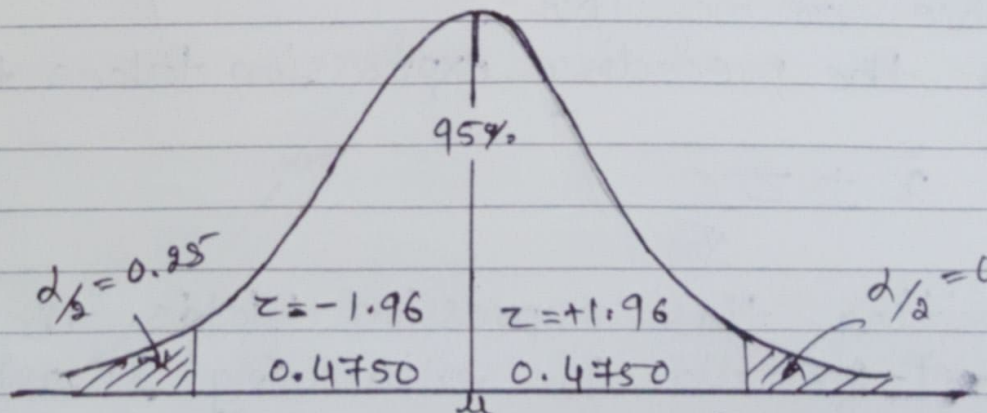
$$\bar{x} - z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{x} + z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$$

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.



α = Shaded area.



distribution of sample means for 95% confidence.

Ex-1 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

Soln. Here, $n = 44$, $\bar{x} = 10.455$, $\sigma = 7.7$.

to determine the value of $z_{\alpha/2}$ divide the 90% confidence in half, or take

.9

$0.5000 - \alpha/2 = 0.5000 - 0.0500 = 0.4500$

alpha = .45

$\Rightarrow \mu$ contains 0.4500 of the area on each side of μ

see Z-score table and find corresponding value

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

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

.45 => 1.6 and between (.04 and .05) ~ 1.6 + .045 = 1.645

$$\Rightarrow 10.455 - 1.91 \le \mu \le 10.455 + 1.91$$

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

$$\Rightarrow P(8.545 \le \mu \le 12.365) = 0.90.$$

The analyst is 90% confident that if a census of all U.S. companies trading with firms in India were taken at the time of this survey, the actual population mean number of years a company would have been trading with firms in India would be bet^n 8.545 and 12.365 (interval estimation)
∴ The point estimation is 10.455 years.

Confidence interval to estimate μ using the finite correction factor.

$$\bar{x} - z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \sqrt{\frac{N-n}{N-1}} \leq \mu \leq \bar{x} + z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \sqrt{\frac{N-n}{N+1}}$$

→ If the sample is taken from a finite population, a finite correction factor may be used to increase the accuracy of the solution. In this case of interval estimation, the finite correction factor is used to reduce the width of the interval. If the sample size is less than 5% of the population, the finite correction factor does not significantly alter the solution.

Ex. 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.

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

Here, $N = 800$, $n = 50$, $\bar{x} = 34.3$, $\sigma = 8$

z value for a 98% confidence interval is 2.33 (0.98 divided into two equal parts yields 0.4900; the z value is obtained from table by using 0.4900).

$$\bar{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}} \le \mu \le \bar{x} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}}$$

$$\Rightarrow 34.3 - 2.33\left(\frac{8}{\sqrt{50}}\right)\sqrt{\frac{750}{799}} \le \mu \le 34.3 + 2.33\left(\frac{8}{\sqrt{50}}\right)\sqrt{\frac{750}{799}}$$

$$\Rightarrow 34.3 - 2.55 \le \mu \le 34.3 + 2.55$$

$$\Rightarrow \boxed{31.75 \le \mu \le 36.85}$$

Without the finite correction factor,

$$34.3 - 2.64 \le \mu \le 34.3 + 2.64 \Rightarrow \boxed{31.66 \le \mu \le 36.94}$$

The finite correction factor takes into account the fact that the population is only 800 instead of being infinitely large. The sample, $n = 50$, is a greater proportion of the 800 than it would be of a larger population, and thus the width of the confidence interval is reduced.

(*) Confidence Interval to Estimate μ when σ is Unknown and n is large.

Many business researchers and statisticians believe that the sample standard deviation is good enough estimate of the population standard deviation when the sample size is large ($n \geq 30$) to use the sample standard deviation in the z formula for estimating a mean.

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

OR

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

Ex. Suppose a U.S. car rental firm wants to estimate the average number of miles traveled per day by each of its cars rented in California. A random sample of 110 cars rented in California reveals that the sample mean travel distance per day is 85.5 miles, with a sample s.D. of 19.3 miles. For a 99% ~~level of~~ confidence, ~~or z value of 2.575~~ Interval to estimate μ.

Soln

Here $n = 110$, $\bar{x} = 85.5$, $s = 19.3$.

For a 99% level of confidence, a z value of 2.575 is obtained.

The confidence interval is

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

$$\Rightarrow 85.5 - 2.575\left(\frac{19.3}{\sqrt{110}}\right) \leq \mu \leq 85.5 + 2.575\left(\frac{19.3}{\sqrt{110}}\right)$$

$$\Rightarrow 85.5 - 4.7 \leq \mu \leq 85.5 + 4.7$$

$$\Rightarrow 80.8 \leq \mu \leq 90.2.$$

The point estimate indicates that the average number of miles traveled per day by a rental car in California is 85.5 with 99% confidence, we estimate that the population mean is somewhere betⁿ 80.8 and 90.2 miles per day.

Note: Using the computer to construct z confidence Intervals for the mean.

It is possible to construct a z confidence interval for the mean with either Excel or MINITAB. Excel yields the ± error portion of the confidence interval that must be placed with the sample mean to construct the complete confidence Interval MINITAB constructs.