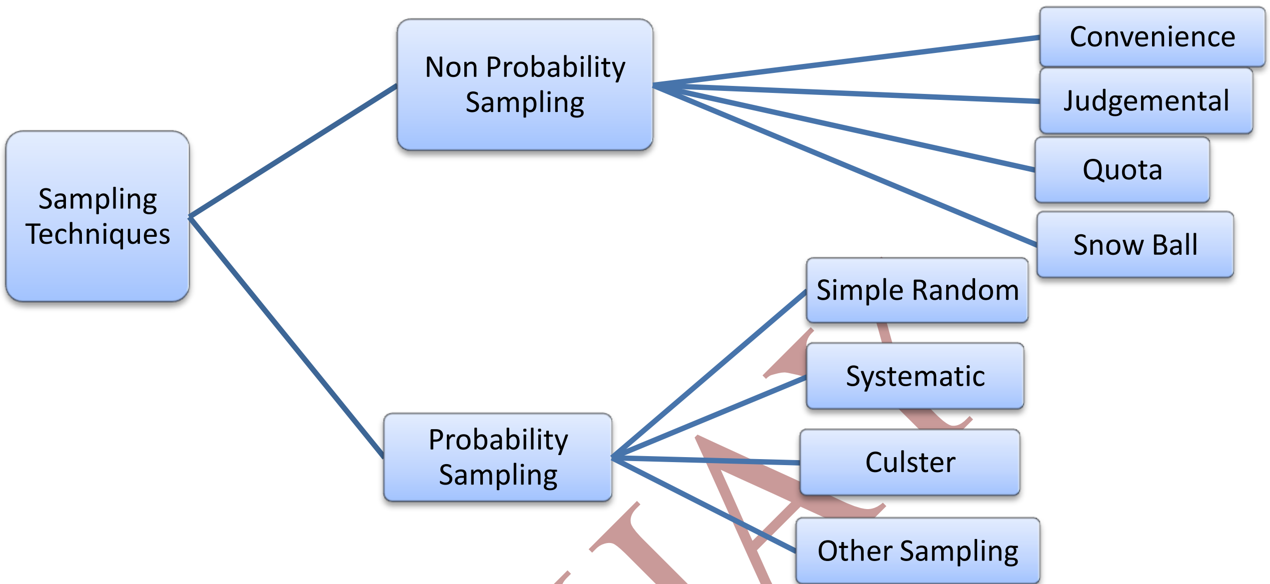
UNIT-II

SAMPLING DISTRIBUTION AND ESTIMATION

Sampling:

Design and Procedures



Convenience Sampling

Convenience sampling attempts to obtain a sample of convenient elements. Often, respondents are selected because they happen to be in the right place at the right time.

- use of students, and members of social organizations
- mall intercept interviews without qualifying the respondents
- department stores using charge account lists
- "people on the street" interviews

Advantages:

Least expensive, least time consuming and most convenient

Disadvantages:

Selection bias, sample not representative, not recommended for descriptive of casual research

Judgmental Sampling

Judgmental sampling is a form of convenience sampling in which the population elements are selected based on the judgment of the researcher.

- test markets

- purchase engineers selected in industrial marketing research
- bellwether precincts selected in voting behavior research
- expert witnesses used in court

Advantages:

Low cost, not time consuming and convenient

Disadvantages:

Does not allow generalization, subjective

Quota Sampling

Quota sampling may be viewed as two-stage restricted judgmental sampling.

- The first stage consists of developing control categories, or quotas, of population elements.
- In the second stage, sample elements are selected based on convenience or judgment.

Control Characteristic	Population Composition	Sample composition	
Sex	Percentage	Percentage	number
Male	48	48	480
Female	52	52	520
	100	100	1000

Advantages:

Sample can be controlled for certain characteristics

Disadvantages:

Selection bias, no assurance of representativeness

Snowball Sampling

In snowball sampling, an initial group of respondents is selected, usually at random.

- After being interviewed, these respondents are asked to identify others who belong to the target population of interest.
- Subsequent respondents are selected based on the referrals.

Advantages:

Can estimate rare characteristics

Disadvantages:

Time-consuming

Simple Random Sampling

- Each element in the population has a known and equal probability of selection.

- Each possible sample of a given size (n) has a known and equal probability of being the sample actually selected.
- This implies that every element is selected independently of every other element.

Advantages:

Easily understood and results projectable

Disadvantages:

Difficult to construct sampling frame, expensive lower precision no assurance of representativeness.

Systematic Sampling

- The sample is chosen by selecting a random starting point and then picking every ith element in succession from the sampling frame.
- The sampling interval, i, is determined by dividing the population size N by the sample size n and rounding to the nearest integer.
- When the ordering of the elements is related to the characteristic of interest, systematic sampling increases the representativeness of the sample.
- If the ordering of the elements produces a cyclical pattern, systematic sampling may decrease the representativeness of the sample.

For example, there are 100,000 elements in the population and a sample of 1,000 is desired. In this case the sampling interval, i, is 100. A random number between 1 and 100 is selected. If, for example, this number is 23, the sample consists of elements 23, 123, 223, 323, 423, 523, and so on.

Advantages:

Can increase representativeness, easier to implement than systematic random sampling, Sampling frame not necessary

Disadvantages:

Can decrease representativeness.

Stratified Sampling

- A two-step process in which the population is partitioned into subpopulations, or strata.
- The strata should be mutually exclusive and collectively exhaustive in that every population element should be assigned to one and only one stratum and no population elements should be omitted.
- Next, elements are selected from each stratum by a random procedure, usually SRS.
- A major objective of stratified sampling is to increase precision without increasing cost.
- The elements within a stratum should be as homogeneous as possible, but the elements in different strata should be as heterogeneous as possible.

- The stratification variables should also be closely related to the characteristic of interest.
- Finally, the variables should decrease the cost of the stratification process by being easy to measure and apply.
- In proportionate stratified sampling, the size of the sample drawn from each stratum is proportionate to the relative size of that stratum in the total population.
- In disproportionate stratified sampling, the size of the sample from each stratum is proportionate to the relative size of that stratum and to the standard deviation of the distribution of the characteristic of interest among all the elements in that stratum.

Advantages:

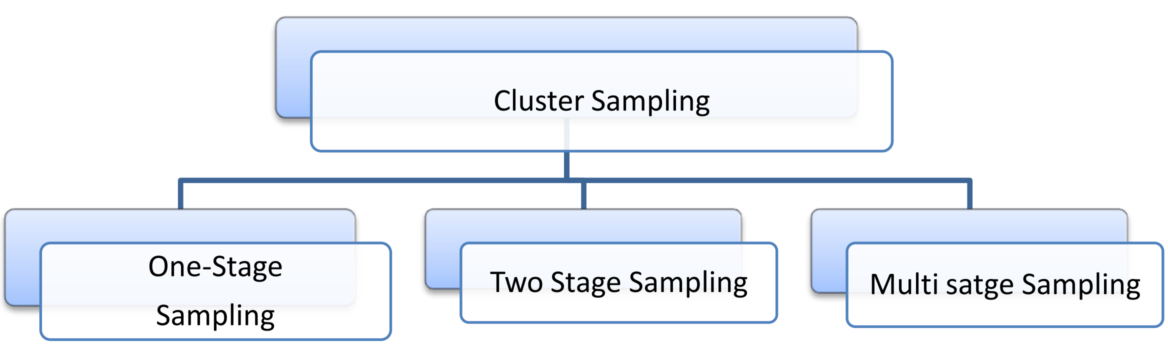
Include all important subpopulations, precision

Disadvantages:

Difficult to select relevant stratification variables, not feasible to stratify on many variables, expensive

Cluster Sampling

- The target population is first divided into mutually exclusive and collectively exhaustive subpopulations, or clusters.
- Then a random sample of clusters is selected, based on a probability sampling technique such as SRS.
- For each selected cluster, either all the elements are included in the sample (one-stage) or a sample of elements is drawn probabilistically (two-stage).
- Elements within a cluster should be as heterogeneous as possible, but clusters themselves should be as homogeneous as possible. Ideally, each cluster should be a small-scale representation of the population.
- In **probability proportionate to size sampling**, the clusters are sampled with probability proportional to size. In the second stage, the probability of selecting a sampling unit in a selected cluster varies inversely with the size of the cluster.



Advantages:

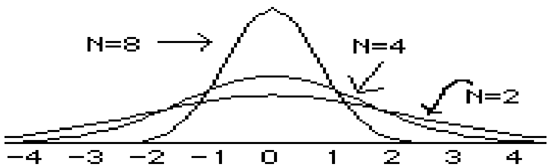
Easy to implement, cost effective

Disadvantages:

Imprecise, difficult to compute and interpret results

Sampling Distribution of the Mean

The sampling distribution of $\bar{x}$ is the probability distribution of all possible values of the sample mean $\bar{x}$. If a population is normal, the sampling distribution of the mean $\bar{x}$ is also normal for samples of all sizes, which can be seen from the following diagram



Properties

The following are important properties of the sampling distribution of mean.

1. It has a mean equal to the population mean

$$i.e., \quad \mu_{\bar{x}} = \mu$$

2. It has a standard deviation equal to the population standard deviation divided by the square root of the sample size.

$$i.e., \quad \sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

where $\sigma_{\bar{x}}$ is the standard error of the mean.

σ standard deviation of the population

n size of the sample

3. If a population is normal, the distribution of sample mean is normal, even if the sample size is small.

Example: 1

A random sample of size 9 is obtained from a normal population with $\mu = 25$. If the sample variance is equal to 100, find the probability that the sample mean exceeds 31.2

Solution:

Given

$$\mu_{\bar{x}} = \mu = 25$$

$$S^2 = 100; \quad S = 10$$

$$\sigma_{\bar{x}} = \frac{S}{\sqrt{n}} = \frac{10}{\sqrt{9}} = 3.33$$

Therefore the probability that the sample mean will exceed 31.2 is

$$P(\bar{x} > 31.2) = p\left(\frac{\bar{x} - \mu}{S/\sqrt{n}} > \frac{31.2 - 25}{3.33}\right)$$

$$= p(Z > 1.862) = 0.5 - 0.4686$$

$$P(\bar{x} > 31.2) = 0.0314$$

Example: 2

The mean strength of a certain cutting tool is 41.5 hrs with the standard deviation of 2.5 hrs. what is the probability that a random sample of size 50 drawn from this population will have a mean between 40.5 hrs and 42 hrs.

Solution:

Given $\mu_{\bar{x}} = \mu = 41.5$

$$\sigma_{\bar{x}} = \frac{S}{\sqrt{n}} = \frac{2.5}{7.07} = 0.353$$

Therefore the probability that the sample mean between 40.5 hrs and 42 hrs is

$$P(40.5 < \bar{x} < 42) = p\left(\frac{40.5 - 41.5}{0.353} < \frac{\bar{x} - \mu}{S/\sqrt{n}} < \frac{40.5 - 42}{0.353}\right)$$

$$= p(-2.83Z < Z < 1.42)$$

$$= p(-2.83 < Z < 0) + P(0 < Z < 1.42)$$

$$= 0.4222 + 0.4977$$

$$= 0.9199$$

Example: 3

For a particular brand TV picture tube, it is known that the mean operating life of the tubes is 1000 hrs with a standard deviation of 250 hrs, what is the probability that the mean for a random sample of size 25 will be between 950 and 1050 hrs?

Solution:

$$Given\ \ \mu_{\bar{x}} = \mu = 1000\ hrs$$

$$\sigma_{\bar{x}} = \frac{S}{\sqrt{n}} = \frac{250}{\sqrt{25}}$$

Therefore the probability that the sample mean between 40.5 hrs and 42 hrs is

$$P(950 < \bar{x} < 1050) = p\left(\frac{950-1000}{\frac{250}{\sqrt{25}}} < \frac{\bar{x}-\mu}{S/\sqrt{n}} < \frac{1050-1000}{\frac{250}{\sqrt{25}}}\right)$$

$$= p(-1 < Z < 1) = p(-1 < z < 0) + p(0 < z < 1)$$

$$= 2 * p(0 < Z < 1)$$

$$= 2 * 0.3413$$

$$P(950 < \bar{x} < 1050) = 0.6826$$

Sampling Distribution of Difference between two Means

Suppose we have two populations, the first of size N_1 with mean μ_1 and S.D σ_1 and the second of size N_2 with mean μ_2 and S.D σ_2. The comparison is made on the basis of two independent random samples, with one of size n_1 drawn from the first population and the other of size n_2 drawn from the second population. If $\overline{x_1}$ and $\overline{x_2}$ are the two sample means, we can evaluate the possible difference between μ_1 and μ_2, by the difference of two sample means $(\overline{x_1} - \overline{x_2})$.

Properties

The important properties of the sampling distribution of difference between mean

1. Difference between of two sample mean $(\overline{x_1} - \overline{x_2})$ is equal to the difference between the population means.

$$i.e., \quad \mu_{\overline{x_1}-\overline{x_2}} = \mu_1 - \mu_2$$

2. The standard deviation of the sampling distribution of $\overline{x_1} - \overline{x_2}$ is given by

$$\sigma_{\overline{x_1}-\overline{x_2}} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

3. If $\overline{x_1}$ and $\overline{x_2}$ are the means of two independent samples drawn from the large populations, the sampling distribution of $\overline{x_1} - \overline{x_2}$ will be normal if the samples are sufficiently large size.

Example: 4

Strength of wire were produced by a company A has mean 4500 kg and S.D of 250 kg, company B has mean of 4000 kg and S.D of 300 kg. if 50 wires of a company A and 100 wires of company B are selected at random and tested for strength. What is the probability that the sample mean strength of A will be at least 600 kg more than that of B.

Solution: Given

$n_1 = 50 \qquad n_2 = 100$

$\mu_1 = 4550 \qquad \mu_2 = 4000$

$\sigma_1 = 200 \qquad \sigma_2 = 300$

For the sampling distribution of difference between means, we have

$$\mu_{\overline{x_1}-\overline{x_2}} = \mu_1 - \mu_2 = 4500 - 4000 = 500$$

$$\sigma_{\overline{x_1}-\overline{x_2}} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} = \sqrt{\frac{40000}{50} + \frac{90000}{100}}$$

$$\sigma_{\overline{x_1}-\overline{x_2}} = 41.231$$

$$p[(\overline{x_1} - \overline{x_2}) > 600] = p\left[\frac{(\overline{x_1} - \overline{x_2}) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \geq \frac{600 - 500}{41.23}\right]$$

$$= p(Z \geq 2.425) = 0.5 - p(0 < Z < 2.45)$$

$$= 0.5 - 0.4925$$

$$= 0.0075$$

Example: 5

Car stereo manufacturer of A have mean life time of 1400 hrs with a S.D of 200 hrs while those of manufacturer B have mean life time of 1200 hrs with S.D of 100 hrs. If a random sample of 120 stereos of each manufacturer are tested. (i). What is the probability that the manufacturer of $A's$ stereo's will have a mean life time of at least 160 hrs more than the manufacturer B's Stereo's (ii) and 250 hrs more than the manufacturer B stereos.

Solution:

Given $n_1 = 120 \quad n_2 = 120$

$\mu_1 = 1400 \quad \mu_2 = 1200$

$\sigma_1 = 200 \quad \sigma_2 = 100$

For the sampling distribution of difference between means, we have

$$\mu_{\overline{x_1}-\overline{x_2}} = \mu_1 - \mu_2 = 1400 - 1200 = 200$$

$$\sigma_{\overline{x_1}-\overline{x_2}} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} = \sqrt{\frac{40000}{120} + \frac{10000}{120}}$$

$$\sigma_{\overline{x_1}-\overline{x_2}} = 20.41$$

$$p[(\overline{x_1}-\overline{x_2})>160]=p\left[\frac{(\overline{x_1}-\overline{x_2})-(\mu_1-\mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}} \geq \frac{160-200}{20.41}\right]$$

$$=p(Z\geq 1.95)=0.5-p(0<z<1.95)$$

$$=0.5-0.475$$

$$p[(\overline{x_1}-\overline{x_2})>160]=0.9750$$

(ii) $\boldsymbol{p}[(\overline{\boldsymbol{x_1}}-\overline{\boldsymbol{x_2}})>250]$

$$=p\left[\frac{(\overline{x_1}-\overline{x_2})-(\mu_1-\mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}} \geq \frac{250-200}{20.41}\right]$$

$$=p(Z\geq 2.45)=0.5-p(0<Z<2.45)$$

$$=0.5-0.4929$$

$$p[(\overline{x_1}-\overline{x_2})>250]=0.0071$$

Sampling Distributions of Proportions

A population proportion is defined as $P=\frac{X}{N}$, where X is the number of elements which possess a certain attribute and N is the total number of item in the populations.

A sample proportion is defined as $p=\frac{x}{n}$, where is the number of items in the sample which possess a certain attribute and n is the sample size.

The sampling distribution of proportions whose mean and sampling distribution are given by

$$\mu_p=P \quad and \quad \sigma_p=\sqrt{\frac{PQ}{n}}$$

where σ_p is the standard error of proportion.

Sampling Distributions of the Difference between two Proportions

The mean and standard deviation of the difference between two proportion is given by

$$\mu_{p_1-p_2}=P_1-P_2 \quad and \quad \sigma_{p_1-p_2}=\sqrt{\frac{P_1Q_1}{n_1}+\frac{P_2Q_2}{n_2}}$$

Example: 6

In a quality department of manufacturing paints at the time of dispatch of decorators 30% of the containers are found to be defective. If a random sample of 500 is drawn with replacement from the population. What is the probability that the sample proportion will be less than 25% defective?

Solution:

Given $\mu_p = p = 0.3, \quad n = 500$

$$Q = 1 - P = 1 - 0.3 = 0.7$$

$$\sigma_p = \sqrt{\frac{PQ}{n}} = \sqrt{\frac{0.3 * 0.7}{500}} = 0.0205$$

The required probability is $prob(P \leq 0.25)$

$$= p\left[\frac{P-p}{\sqrt{\frac{PQ}{n}}} \leq \quad \frac{0.25 - 0.3}{0.0205}\right]$$

$$= p(Z \leq -2.44) = 0.5 - p(0 \leq Z \leq 2.44)$$

$$= 0.5 - 0.4927$$

$$prob(P \leq 0.25) = 0.0073$$

Example: 7

A manufacturer of watches has determined from experience that 3% of the watches he produces are defective. If a random sample of 300 watches is examined, what is the probability that, the proportion defective between 0.02 and 0.035.

Solution:

$Given \quad P = 0.03 \quad Q = 0.97 \quad n = 300$

$$\sigma_p = \sqrt{\frac{PQ}{n}} = \sqrt{\frac{0.03 * 0.97}{300}} = 0.0098$$

The required probability is

$$p(0.02 \leq Z \leq 0.035) = p\left[\frac{0.25 - 0.03}{0.0098} \leq \frac{P-p}{\sqrt{\frac{PQ}{n}}} \quad \leq \quad \frac{0.035 - 0.03}{0.0098}\right]$$

$$= p[-1.02 \leq \ Z \ \leq \ 0.51]$$

$$= 0.3451 + 0.1950$$

$$= 0.5411$$

Example: 8

If two proportions 10% of machine produced by a company A are defective and 5% of machine produced by a company B are defective. A random sample of 250 machines are taken from company A and has the random sample of 300 machines from company B. what is the probability that the difference in sample proportion is ≤ 0.02.

Solution:

Given $P_1 = 0.10 \quad P_2 = 0.05$

$n_1 = 250 \quad n_2 = 300$

$Q_1 = 0.90 \quad Q_2 = 0.95$

$$\sigma_p = \sqrt{\frac{P_1Q_1}{n_1} + \frac{P_2Q_2}{n_2}} = \sqrt{\frac{0.10 * 0.90}{250} + \frac{0.05 * 0.95}{300}}$$

$\sigma_p = 0.0228$

$\mu_p = P_1 - P_2 = 0.10 - 0.05 = 0.05$

The required probability is $p[(p_1 - p_2) \leq 0.02]$

$$p[(p_1 - p_2) \leq 0.02] = p\left[\frac{(p_1 - p_2) - (P_1 - P_2)}{\sqrt{\frac{P_1Q_1}{n_1} + \frac{P_2Q_2}{n_2}}} \leq \frac{0.02 - 0.05}{0.0228}\right]$$

$$= p[Z \leq -1.32]$$

$$= 0.5 - p(0 \leq Z \leq 1.32)$$

$$= 0.5 - 0.4066$$

$$p[(p_1 - p_2) \leq 0.02] = 0.0934.$$

Central Limit Theorem

Definition:

When sampling is done from a population with mean μ and finite standard deviation σ, the sampling distribution of the sample mean $\bar{x}$ will tend to a normal distribution with mean μ and standard deviation $\sigma/\sqrt{n}$ as the sample size n becomes large.

For 'large enough' n, $\bar{X} = N\left(\mu, \frac{\sigma^2}{n}\right)$

Example: 9

The life time of a certain brand of an electric bulb may be considered as a random variable with mean 1200 hours and standard deviation 250 hours. Find the probability using central limit theorem has the average life time of 60 bulbs exceed 1250 hours.

Solution:

Given that $n = 60$

$$Mean = \mu = 1200 \quad and\ S.D = \sigma = 250$$

Let X_i be the life time of an electric bulb.

$$\overline{X} = \frac{X_1 + X_2 + \cdots + X_{60}}{60}$$

By central limit theorem,

$$\therefore \quad \overline{X} \approx N\left(\mu, \frac{\sigma^2}{n}\right)$$

$$\therefore \quad p(\bar{X} > 1250) = p\left[\frac{\bar{X} - \mu}{\sigma/n} > \frac{1250 - 1200}{250/\sqrt{60}}\right]$$

$$= p(Z > 1.55)$$

$$= 0.5 - 0.4394$$

$$p(\bar{X} > 1250) = 0.0606$$

Example: 10

A random sample of size 100 is taken from a population whose mean is 60 and variance 400. Using central limit theorems find what probability that we can assert that the mean of the sample will not differ from μ more than 4.

Solution:

Given that $n = 60$

$$Mean = \mu = 60 \quad and\ S.D = \sigma = \sqrt{400} \qquad \overline{X} = Sample\ mean$$

By central limit theorem,

$$\therefore \quad \overline{X} \approx N\left(\mu, \frac{\sigma^2}{n}\right)$$

The required probability is

$$p(-4 < \overline{X} - \mu < 4) = p(-4 < \overline{X} - 60 < 4)$$

$$= p(56 < \overline{X} < 64)$$

$$= p\left[\frac{56 - 60}{2} < \frac{\bar{X} - \mu}{\sigma/n} < \frac{64 - 60}{2}\right]$$

$$= p[-2 < Z < +2]$$

$$= 2 * p(0 < Z < 2)$$

$$= 0.9544$$

Example: 11

An economist wishes to estimate the average family income in a certain population. The population standard deviation is known to be $4,500 and the economist uses the random sample of size n=225. What is the probability that the sample mean will fail within $800 of the population mean?

Solution:

Given that $n = 225$

$$Mean = \mu = 800 \quad and\ S.D = \sigma = 4{,}500 \qquad \overline{X} = Sample\ mean$$

By central limit theorem,

$$\therefore \quad \overline{X} \approx N\left(\mu, \frac{\sigma^2}{n}\right)$$

The required probability is

$$p(-800 < \overline{X} - \mu < 800) = p(-4 < \overline{X} - 60 < 4)$$

$$= p(56 < \overline{X} < 64)$$

$$= p\left[\frac{-800}{4500/\sqrt{225}} < \frac{\bar{X} - \mu}{\sigma/n} < \frac{800}{4500/\sqrt{225}}\right]$$

$$= p\left[\frac{-800}{300} < Z < \frac{800}{300}\right]$$

$$= p[-2.67 < Z < 2.67]$$

$$= 2 * p(0 < Z < 2.67)$$

$$= 0.9924.$$

Estimation

In statistics, estimation refers to the process by which one makes inferences about a population, based on information obtained from a sample.

Statisticians use sample statistics to estimate population parameters.

For example, sample means are used to estimate population means; sample proportions, to estimate population proportions.

An estimate of a population parameter may be expressed in two ways:

i) Point estimate ii) Interval estimates.

Point estimate:

A point estimate of a population parameter is a single value of a statistic. For example, the sample mean $\overline{x}$ is a point estimate of the population mean μ. Similarly, the sample proportion p is a point estimate of the population proportion P.

It is desirable for a point estimate to be:

(1) **Consistent:** The larger the sample size, the more accurate the estimate.

(2) **Unbiased:** The expectation of the observed values of many samples ("average observation value") equals the corresponding population parameter.

For example, the sample mean is an unbiased estimator for the population mean.

(3) **Most efficient or best unbiased**: The one possessing the smallest variance (a measure of the amount of dispersion away from the estimate).

In other words, the estimator that varies least from sample to sample. This generally depends on the particular distribution of the population.

For example, the mean is more efficient than the median (middle value) for the normal distribution but not for more "skewed" (asymmetrical) distributions.

Methods to calculate the estimator:

The most commonly used methods to calculate the estimator are

i) **The maximum likelihood method:** Maximum likelihood method uses differential calculus to determine the maximum of the probability function of a number of sample parameters.

ii) **The method of moments:** The moment's method equates values of sample moments (functions describing the parameter) to population moments.

Interval estimate:

An interval estimate is defined by two numbers, between which a population parameter is said to lie. For example, $a < \bar{x} < b$ is an interval estimate of the population mean μ. It indicates that the population mean is greater than a but less than b.

Confidence Intervals:

A confidence interval consists of three parts.

i) Confidence level, ii) Statistic iii) Margin of error.

The confidence level describes the uncertainty of a sampling method. The statistic and the margin of error define an interval estimate that describes the precision of the method. The interval estimate of a confidence interval is defined by the sample statistic $\pm$ margin of error.

For example, we might say that we are 95% confident that the true population mean falls within a specified range. This statement is a confidence interval. It means that if we used the same sampling method to select different samples and compute different interval estimates, the true population mean would fall within a range defined by the *sample statistic* $\pm$ *margin of error* 95% of the time. Confidence intervals are preferred to point estimates, because confidence intervals indicate (a) the precision of the estimate and (b) the uncertainty of the estimate.

Confidence Level:

The probability part of a confidence interval is called a confidence level. The confidence level describes how strongly we believe that a particular sampling method will produce a confidence interval that includes the true population parameter.

Here is how to interpret a confidence level. Suppose we collected many different samples, and computed confidence intervals for each sample. Some confidence intervals would include the true population parameter; others would not. A 95% confidence level means that 95% of the intervals contain the true population parameter; a 90% confidence level means that 90% of the intervals contain the population parameter; and so on.

Margin of Error:

In a confidence interval, the range of values above and below the sample statistic is called the margin of error.

For example, suppose the local newspaper conducts an election survey and reports that the independent candidate will receive 30% of the vote. The newspaper states that the survey had a 5% margin of error and a confidence level of 95%. These findings result in the following confidence interval: We are 95% confident that the independent candidate will receive between 25% and 35% of the vote.

Note: Many public opinion surveys report interval estimates, but not confidence intervals. They provide the margin of error, but not the confidence level. To clearly interpret survey results you need to know both! We are much more likely to accept survey findings if the confidence level is high (say, 95%) than if it is low (say, 50%).

Standard Error of the Mean:

The standard error of the mean is designated as: σ_M . It is the standard deviation of the sampling distribution of the mean. The formula for the standard error of the mean is:

$$\sigma_M = \frac{\sigma}{\sqrt{N}}$$

where σ is the standard deviation of the original distribution and n is the sample size (the number of scores each mean is based upon). This formula does not assume a normal distribution. However, many of the uses of the formula do assume a normal distribution. The formula shows that the larger the sample size, the smaller the standard error of the mean. More specifically, the size of the standard error of the mean is inversely proportional to the square root of the sample size.

Confidence interval:

Confidence interval for the population mean for large samples ($\boldsymbol{\sigma}$ $\boldsymbol{is\ known}$)

To estimate the mean of some characteristic or event in a population:

$$Estimate \pm Z_{\frac{\alpha}{2}} \times SE$$

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

The probability that the true population mean will fall in the interval

$$\bar{x} - Z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} \leq \bar{x} \leq \bar{x} + Z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} \quad is \quad 1 - \alpha$$

where $\bar{x}$ is the sample mean,

n is the sample size,

σ is the standard deviation of the population,

$Z_{\frac{\alpha}{2}}$ is the point on the standard normal curve area beyond which is $\frac{\alpha}{2}$%

$1 - \alpha$ is the confidence level.

Example: 12

A machine produces components, which have a standard deviation of 1.6 cm in length. A random sample of 64 parts is selected from the output and this sample has a mean length of 90 cm. the customer will reject the part if it is either less than 88 cm or more than 92 cm. does the 95% confidence interval for the true mean length of all the components produced ensure acceptance by the customer?

Solution:

Here μ is the mean length of the components in the population. The formula for the confidence interval is

$$\overline{x} - Z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} \leq \overline{x} \leq \overline{x} + Z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} \quad is \quad 1 - \alpha$$

$$Given \quad \sigma = 1.6, \quad Z_{\frac{\alpha}{2}} = 1.96, \quad \overline{x} = 90 \quad and \quad n = 64$$

$$\therefore \quad 90 - 1.96\left(\frac{1.6}{\sqrt{64}}\right) \leq \mu \leq 90 + 1.96\left(\frac{1.6}{\sqrt{64}}\right)$$

$$90 - 1.96(0.2) \leq \mu \leq 90 + 1.96(0.2)$$

$$89.61 \leq \mu \leq 90.39$$

This implies that the probability that the true value of the population mean length of the components will fall in this interval is 95%. Hence we infer that the 95% confidence interval ensures acceptance by the customer.

Example: 13

A sample of 100 measurements at breaking strength of cotton threads gave a mean of 7.4 and a standard deviation of 1.2 gms. Find 95% confidence limits for the mean breaking strength.

Solution:

$$Given \quad Sample\ S.D = S = 1.2, \quad Z_{\frac{\alpha}{2}} = 1.96, \quad \overline{x} = 7.4 \quad and \quad n = 100$$

$$\frac{\sigma}{\sqrt{n}} = \frac{S}{\sqrt{n}} = \frac{1.2}{\sqrt{100}} = 0.12$$

Hence the 95% confidence limits for the population are

$$\overline{x} \pm Z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}$$

$$= 7.4 \pm 1.96\ (0.12) = 7.4 \pm 0.2352$$

$$\mu = (7.1648, \ \ 7.6352)$$

Example: 14

A mining company needs to estimate the average amount of copper are per ton mined. A random sample of 50 tons gives a sample mean of 146.75 pounds. The population standard deviation is assumed to be 35.2 pounds. Give a 95% confidence interval for the average amount of copper in the “population” of tons mined.

Solution:

$Given \quad S.D = \sigma = 35.2, \; Z_{\frac{\alpha}{2}} = 1.96, \quad \bar{x} = 146.75 \quad and \quad n = 50$

$$\frac{\sigma}{\sqrt{n}} = \frac{35.2}{\sqrt{50}} = 4.98$$

Hence the 95% confidence limits for the population are

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

$$= 46.75 \pm 1.96\ (4.98) = 146.75 \pm 9.76$$

$$\mu = (136.99, \;\; 156.51)$$

Example: 15

A server channel monitored for an hour was found to have an estimated mean of 20 transactions transmitted per minute. The variance is known to be 4. Find the standard error. Establish an interval estimate that includes a population mean 95% of the time and 99% of the time.

Solution:

Given $\sigma^2 = 4 \;\; \Longrightarrow \;\; \sigma = 2$

$$n = 1\ hour = 60\ minutes \qquad \therefore\ \bar{x} = 20\ per\ mint$$

$$Standard\ error \;\; \sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{2}{\sqrt{60}} = 0.2582$$

Also $Z_{\frac{\alpha}{2}} = 1.96,$

Hence the 95% confidence interval is

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

$$= 20 \pm 1.96\ (0.2582) = 20 \pm 0.506072$$

$$\mu = (19.4939, \;\; 20.5061)$$

Level of confidence = 99%

Here $Z_{\frac{\alpha}{2}} = 2.58$

Hence the 99% confidence interval is

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

$$= 20 \pm 2.58(0.2582)$$

$$\mu = (19.33, \;\; 20.67)$$

Confidence Interval for the Population Mean for Small Samples ($\boldsymbol{\sigma}$ ***Unknown***)

When the population standard deviation is not known, we may use the sample standard deviation s in its place. If the population is normally distributed, the standardized statistic

$$t = \frac{\overline{x} - \mu}{s/\sqrt{n}}$$

has a t- distribution with $n - 1$ degrees of freedom. The t distribution is also called student's t distribution. The confidence interval for μ when σ is not known (assuming a normally distributed population) is

$$\overline{x} \pm t_{\alpha/2} \frac{s}{\sqrt{n}}$$

where $t_{\alpha/2}$ is the value of the t –distribution with $n - 1$ degrees of freedom.

Example: 16

A stock market analyst wants to estimate the average return on a certain stock. A random sample of 15 days yields an average return of $\bar{x} = 10.37\%$ and standard deviation of $s = 3.5\%$. Assuming a normal population of returns, give a 95% of confidence interval for the average return on this stock.

Solution:

Since the sample size $n = 15$, we want to use the t- distribution with $n - 1$ degrees of freedom.

Here $t_{\alpha/2} = 2.145, \quad \overline{x} = 10.37 \quad and \quad s = 3.5$

The 95% of confidence interval is

$$\overline{x} \pm t_{\alpha/2} \frac{s}{\sqrt{n}}$$

$$= 10.37 \pm 2.145 \frac{3.5}{\sqrt{15}}$$

$$\mu = (8.43, \ \ 12.31)$$

Thus, the analyst may be 95% sure that the average annualized return on the stock is anywhere from 8.43% to 12.31%.

Example: 17

The average travel time based on the random sample of 100 people working in a company to reach the office is 40 minutes with S.D of 10 minutes. Establish the 95% confidence interval for the mean travel time of everyone in the company to design the working hours.

Solution:

Given $n = 10, \ \bar{x} = 40, \ \ s = 10 \ \ and \ \ t_{\alpha/2} = 2.262$

The 95% confidence interval for the mean travel time of everyone in the company is

$$\overline{x} \pm t_{\alpha/2}\,\frac{s}{\sqrt{n}}$$

$$= 40 \pm 2.262\,\frac{10}{\sqrt{10}}$$

$$= 40 \pm 7.153$$

$$\mu = (32.85,\ \ 47.15)$$

Example: 18

A transportation company wants to estimate the average length of time goods are in transit across the country. A random sample of 20 shipments gives $\overline{x} = 2.6$ days and $s = 0.4$ day. Give a 99% of confidence interval for the average transit time.

Solution:

Given $n = 20,\ \ \bar{x} = 2.6,\ \ \ s = 0.4$

The value of $t_{\alpha/2}$ with $20 - 1 = 19$ degrees of freedom is 2.861.

Hence 95% confidence interval is

$$\overline{x} \pm t_{\alpha/2}\,\frac{s}{\sqrt{n}}$$

$$= 2.6 \pm 2.861\,\frac{0.4}{\sqrt{20}}$$

$$= 2.6 \pm 0.256$$

$$\mu = (2.344,\ \ 2.856)$$

Example: 19

A management consulting agency needs to estimate the average number of years of experience of executives in a given branch of management. A random sample of 28 executives gives $\bar{x} = 6.7$ years and $s = 2.4$ years. Give a 99% of confidence interval for the average number of years of experience for all executives in this branch.

Solution:

Given $n = 28,\ \ \bar{x} = 6.7,\ \ \ s = 2.4\ \ \ and \ \ \ \ t_{\alpha/2} = 2.861$

Hence 95% confidence interval is

$$\overline{x} \pm t_{\alpha/2}\,\frac{s}{\sqrt{n}}$$

$$= 6.7 \pm 2.861\,\frac{2.4}{\sqrt{28}}$$

$$= 6.7 \pm 1.298$$

$$\mu = (5.402,\ \ 7.998)$$

Confidence Interval for the Difference Between Two Population Means for Large Samples (σ Known)

Let $\overline{x_1}\ and\ \ \overline{x_2}$ be the means of two samples of sizes $n_1\ \ and\ \ n_2$ respectively. Then the confidence limits for the difference between two population means $(\mu_1 - \mu_2)$ is given by

$$(\overline{x_1}\ - \overline{x_2}) \pm\ Z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

Note:

Suppose that $\sigma_1^2\ \ and\ \ \sigma_2^2$ are not known and the sample sizes $n_1\ \ and\ \ \ n_2$ are larger (> 30), one can use $s_1^2\ and\ \ s_2^2$ as the sample variances and the confidence limits are

$$(\overline{x_1}\ - \overline{x_2}) \pm\ Z_{\alpha/2}\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

Example: 20

In a certain factory there are two independent processes manufacturing that same item. The average weight in a sample of 250 items produced from one process is found to be 120 O_{zs} with a standard deviation of 12 O_{zs}. While the corresponding figures in a sample of 400 items from the other process are 124 O_{zs} and 14 O_{zs}. Find the 99% confidence limits for the difference in the average weight of items produced by the two process respectively.

Solution:

Given that $n_1 = 250,\quad \overline{x_1}\ = 120,\ \ S_1 = 12$

$n_2 = 400,\quad \overline{x_2}\ = 124,\ \ S_2 = 14$

$Z_{\alpha/2} = 2.58$

$\therefore$ 99% confidence for $(\mu_1 - \mu_2)$ are

$$(\overline{x_1}\ - \overline{x_2}) \pm\ Z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} = (\overline{x_1}\ - \overline{x_2}) \pm\ Z_{\alpha/2}\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

$$(120 - 124) \pm\ 2.58\sqrt{\frac{12^2}{250} + \frac{14^2}{400}} = 4 \pm (2.58 * 1.0324)$$

$$(\mu_1 - \mu_2) = (1.34,\ \ 6.66)$$

Example: 21

In order to compare the I.Q of students, two schools were selected. A random sample of 90 students from each school. At school A, the mean I.Q is 109 and the S.D is 11. At school B, the mean I.Q is 98 and S.D is 9. Construct 95% confidence interval for the difference between mean I.Q of two schools.

Solution:

Given that $n_1 = n_2 = 90, \quad \overline{x_1} = 109, \quad \overline{x_2} = 98, \; S_1 = 11, \quad S_2 = 9$

$Z_{\alpha/2} = 1.96$

$\therefore$ 95% confidence for the difference between mean I.Q of two schools are $(\mu_1 - \mu_2)$

$$(\overline{x_1} - \overline{x_2}) \pm Z_{\alpha/2} \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} = (\overline{x_1} - \overline{x_2}) \pm Z_{\alpha/2} \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

$$= (109 - 98) \pm 1.96 \sqrt{\frac{12^2}{90} + \frac{9^2}{90}} = 11 \pm (1.96 * 1.4979)$$

$$(\mu_1 - \mu_2) = (8.06, \; 13.94)$$

Confidence Interval for the Difference between two Population Means for Small Samples (σ unknown)

Let $\overline{x_1}$ *and* $\overline{x_2}$ be the means of two small samples of sizes(< 30) n_1 *and* n_2 respectively. Then the confidence limits for the difference between two population means $(\mu_1 - \mu_2)$ is given by

$$(\overline{x_1} - \overline{x_2}) \pm t_{\alpha/2} \; S \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

where $t_{\alpha/2}$ is the tabulated value of t with $n_1 + n_2 - 2$ degrees of freedom at α level of significance and

$$S^2 = \frac{1}{n_1 + n_2 - 2}\left[\sum_i (x_i - \bar{x})^2 + \sum_j (y_i - \bar{y})^2\right]$$

Is an unbiased estimator of σ^2.

Example: 22

In a test given to two groups of students the marks obtained were as follows.

First Group	18	20	36	50	49	36	34	49	61
Second Group	29	28	26	35	30	44	46		

Construct a 95% confidence interval on the mean marks secured by the students of the above two groups.

Solution:

Given that $n_1 = 9, \quad n_2 = 7, \quad \overline{x_1} = \frac{333}{9}, \quad \overline{x_2} = \frac{238}{7} = 34,$

$$S^2 = \frac{1}{n_1 + n_2 - 2}\left[\sum_i (x_i - \bar{x})^2 + \sum_j (y_i - \bar{y})^2\right]$$

$$S^2 = \frac{1}{14}[1134 + 386]$$

$$S^2 = 108.57$$

$$S = 10.42 \quad and \quad t_{\frac{\alpha}{2}} \text{ with } n_1 + n_2 - 2 = 14\ d.o.f\ \ is\ 1.76$$

$\therefore$ 95% confidence interval is

$$(\overline{x_1} - \overline{x_2}) \pm t_{\alpha/2}\ S\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

$$= (37 - 34) \pm (1.76 * 10.42)\sqrt{\frac{1}{9} + \frac{1}{7}}$$

$$= 3 \pm (1.76 * 5.25)$$

$$(\mu_1 - \mu_2) = (6.24,\ 12.24)$$

Confidence Interval for the Population Proportion for Large Samples

The confidence interval for the population proportion P is

$$p \pm Z_{\alpha/2}\sqrt{\frac{pq}{n}}$$

where the sample proportion p is equal to the number of successes in the sample x, divided by the number of trails n (sample size)

$$i.e., \qquad p = \frac{x}{n}$$

Example: 23

A market research firm wants to estimate the share that foreign companies have in the U.S market for certain products. A random sample of 100 consumers is obtained, and 34 people in the sample are found to be users of foreign made products; the rest are users of domestic products. Give a 95% confidence interval for the share of foreign products in this product.

Solution:

$$Given\ \ x = 34\ \ and\ \ n = 100$$

$$\therefore\ \ our\ sample\ proportion\ is\ p = \frac{x}{n} = \frac{34}{100} = 0.34, \qquad q = 1 - p = 0.66$$

$$Also\ \ Z_{\alpha/2} = 1.96$$

$\therefore$ 95% confidence interval for p is

$$p \pm Z_{\alpha/2}\sqrt{\frac{pq}{n}}$$

$$= 0.34 \pm 1.96\sqrt{\frac{0.34 * 0.66}{100}}$$

$$= 0.34 \pm 0.0928$$

$$p = (0.2472,\ \ 0.4328)$$

Hence the firm may be 95% confident that foreign manufactures control any where from 24.72% to 43.28% of the market.

Example: 24

A survey of 748 randomly selected employees of dot.com companies showed that 35% feel about their jobs. Give 90% confidence interval for the proportion of dot.com company employees who feel secure about their jobs.

Solution:

$$Given\ \ n = 748\ \ and\ \ p = 0.35 \qquad q = 0.65$$

$$Here\ \ Z_{\alpha/2} = 1.645$$

$\therefore$ 90% confidence interval for P is

$$p \pm Z_{\alpha/2}\sqrt{\frac{pq}{n}}$$

$$= 0.35 \pm 1.645\sqrt{\frac{0.35 * 0.65}{748}}$$

$$= 0.35 \pm 0.02862$$

$$p = (0.3214,\ \ 0.3786)$$

Example: 25

A sample poll of 100 voters chosen at random from all voters in a given district indicated that 55% of them were in favour of a particular candidate. Find (i) 95% and (ii) 99% confidence limits for the proportion of all the voters in favour of this candidate.

Solution:

$$Given\ \ n = 100$$

p =sample proportion of voters favoring the candidate

$$p = 0.55\ \ and\ \ q = 0.45$$

(i) 95% confidence interval for $Z_{\alpha/2} = 1.96$ is

$$p \pm Z_{\alpha/2}\sqrt{\frac{pq}{n}}$$

$$= 0.55 \pm 1.96\sqrt{\frac{0.55 * 0.45}{100}}$$

$$= 0.55 \pm 0.0974$$

$$p = (0.4526,\ \ 0.6474)$$

(ii) 99% confidence interval for $Z_{\alpha/2} = 2.58$ is

$$p \pm Z_{\alpha/2}\sqrt{\frac{pq}{n}} = 0.55 \pm 2.58\sqrt{\frac{0.55 * 0.45}{100}}$$

$$= 0.55 \pm 0.1282$$

$$p = (0.4218,\ \ 0.6782)$$

Confidence Interval for the Difference between Two Population Proportions for Large Samples

Let p_1 & p_2 are the sample proportions obtained from large samples of sizes $n_1\ and\ n_2$ drawn from the respective populations having proportions P_1& P_2.

Since for large samples, the sampling distribution of differences in proportions $(p_1 - p_2)$ is normally distributed, the confidence limits for the difference between two population proportions are

$$(p_1 - p_2) \pm Z_{\alpha/2}\sqrt{\frac{p_1q_1}{n_1} + \frac{p_2q_2}{n_2}}$$

$$where \quad p_1 = \frac{x_1}{n_1}, \quad q_1 = 1 - p_1$$

$$p_2 = \frac{x_2}{n_2}, \quad q_2 = 1 - p_2$$

Example 26

In a random sample of 100 men are taken from a village A, 60 were found to be consuming alcohol. In other sample of 200 men are taken from village B, 100 were found to be consuming alcohol. Construct 95% confidence interval in respect of difference in the proportions of men who consume alcohol.

Solution:

$$Given\ \ n_1 = 100 \quad n_2 = 200$$

$$and \quad p_1 = \frac{x_1}{n_1} = \frac{60}{100} = 0.6, \quad q_1 = 0.4$$

$$p_2 = \frac{x_2}{n_2} = \frac{100}{200} = 0.5, \quad q_2 = 0.5$$

Here $Z_{\alpha/2} = 1.96$

$\therefore$ 95% confidence interval for $(p_1 - p_2)$ is

$$(p_1 - p_2) \pm Z_{\alpha/2} \sqrt{\frac{p_1 q_1}{n_1} + \frac{p_2 q_2}{n_2}}$$

$$= (0.6 - 0.5) \pm 1.96 \sqrt{\frac{0.6 * 0.4}{100} + \frac{0.5 * 0.5}{200}}$$

$$= 0.1 \pm (1.96 * 0.06)$$

$$= 0.1 \pm 0.1176$$

$\therefore \quad 0.018 < P_1 - P_2 < 0.218$

Example: 27

Two operators perform the same operation of applying a plastic coating to a part. A random sample of 100 parts from the first operator shows that 6 are non-conforming. A random sample of 200 parts from the second operator shows that 8 are non-conforming. Find a 90% of confidence interval for the difference in the proportion of non-conforming parts produced by the two operators.

Solution:

$$Given \quad n_1 = 100 \quad n_2 = 200$$

$$and \quad p_1 = \frac{x_1}{n_1} = \frac{6}{100} = 0.06, \quad q_1 = 0.94$$

$$p_2 = \frac{x_2}{n_2} = \frac{8}{200} = 0.04, \quad q_2 = 0.96$$

Here $Z_{\alpha/2} = 1.645$

$\therefore$ 90% confidence interval for $(p_1 - p_2)$ is

$$(p_1 - p_2) \pm Z_{\alpha/2} \sqrt{\frac{p_1 q_1}{n_1} + \frac{p_2 q_2}{n_2}}$$

$$= (0.06 - 0.04) \pm 1.645 \sqrt{\frac{0.06 * 0.94}{100} + \frac{0.04 * 0.96}{200}}$$

$$= 0.02 \pm (1.645 * 0.0275)$$

$$= 0.02 \pm 0.0452$$

$$\therefore \quad 0.0275 < P_1 - P_2 < 0.0652$$

Determining the Sample Size (Using Confidence Interval)

The confidence interval for estimating a population mean is given by

$$\bar{x} \pm Z_{\propto} * \frac{\sigma}{\sqrt{n}}$$

$$or \quad \bar{x} \pm E \quad where \quad E = Z_{\propto} * \frac{\sigma}{\sqrt{n}}$$

i.e., the minimum allowable error for the differences between the population mean and the sample mean

$$\therefore \quad \sqrt{n} = \frac{Z_{\propto}\ \sigma}{E} \quad or \quad n = \frac{(Z_{\propto})^2\ \sigma^2}{E^2}$$

Here both the values of $Z_{\propto}$ and E must be specified. The value of the population σ may be actual or estimated.

Example: 28

Given a population with a standard deviation of 8.6, what sample size is needed to estimate the mean of the population within ± 0.5 with 99% of confidence?

Solution:

Given $E = 0.5 \quad and \quad \sigma = 8.6 \quad also \quad Z_{\propto} = 2.58$

$$\therefore \quad n = \frac{(Z_{\propto})^2 \sigma^2}{E^2} = \frac{(2.58)^2 (8.6)^2}{0.5^2}$$

$$n = 1970$$

Example: 29

A cigarette manufacturer wishes to use a random sample to estimate the average nicotine content. The sampling error should not be more than 1 m.g above or below the true mean with a 99% confidence. The population S.D is 4 m.g. What sample size should company use to satisfy these requirements.

Solution:

Given $E = 1\ m.g \quad and \quad \sigma = 4\ m.g \quad also \quad Z_{\propto} = 2.58$

$$\therefore \quad n = \frac{(Z_{\propto})^2 \sigma^2}{E^2} = \frac{(2.58)^2 (4)^2}{1^2}$$

$$n = 106.5 \quad or \ \ 107\ app.$$

Example: 30

From previous studies the population standard deviation for a placement test has been determined to be 12.4. The test is scored on a scale of $0 - 100$. A placement agency wants to be 90% confident that the average test score of a sample falls within plus or minus 3 points of the population average score. How large a sample should be selected?

Solution:

Given $E = 3 \quad \sigma = 12.4 \quad also \quad Z_\propto = 1.645$

$$\therefore \quad n = \frac{(Z_\propto)^2 \sigma^2}{E^2} = \frac{(1.645)^2 (12.4)^2}{3^2}$$

$$n = 46.23 \quad or \;\; 46\, app$$

Example: 31

Suppose the marketing manager wishes to estimate the population means annual usage of home heating oil to within ± 50 gallons of the true value and he wants to be 95% confidence of correctly estimating the true mean. On the basis of a study taken from the previous year, he belives that the standard deviation can be estimated as 325 gallons. Find the sample size needed.

Solution:

Given $E = 50 \quad \sigma = 325 \quad also \quad Z_\propto = 1.96$

$$\therefore \quad n = \frac{(Z_\propto)^2 \sigma^2}{E^2} = \frac{(1.96)^2 (325)^2}{50^2}$$

$$n = 162.31 \quad or \;\; 162 \;\; app$$

Sample Size for Estimating a Population Proportion

The confidence interval for estimating population proportion is given by

$$p \pm Z_\propto \sqrt{\frac{pq}{n}} \quad or \quad p \pm E \qquad where \;\; E = Z_\propto \sqrt{\frac{pq}{n}}$$

$$\therefore \quad \sqrt{n} = \frac{Z_\propto \sqrt{pq}}{E} \quad \Longrightarrow \quad n = \frac{(Z_\propto)^2 \; pq}{E^2}$$

where the values of $Z_\propto$ and E are predetermined. The value of proportion p may be actual or estimated from the past experience.

Example: 32

A firm wishes to estimate with a maximum allowable error of 0.05 and 95% level of confidence, the proportion of consumers who prefer its product. How large a sample will be required in order to make such an estimate if the preliminary sales report indicates that 25% of all consumers perform the firm's product?

Solution:

Given $p = 0.25 \quad and \quad q = 0.75$

$$E = 0.05 \quad also \quad Z_{\propto} = 1.96$$

$$\therefore \quad n = \frac{(Z_{\propto})^2 pq}{E^2} = \frac{(1.96)^2(0.25 * 0.75)}{0.05^2}$$

$$n = 288$$

Example: 33

For a test market find the sample size needed to estimate the true proportion of consumers satisfied with a certain new product within ± 0.04 at 90% confidence.

Solution:

If proportion is not given, take $p = q = 0.5$

Given $E = 0.04 \quad also \quad Z_{\propto} = 1.645$

$$\therefore \quad n = \frac{(Z_{\propto})^2 p * q}{E^2} = \frac{(1.645)^2(0.5 * 0.5)}{0.04^2}$$

$$n = 423$$

Example: 34

The operations manager wants to have 90% confidence of estimating the proportion of non-conforming newspapers to within ± 0.05 of its true value. In addition, because the publisher of the newspaper has not previously undertaken such a study, no information is available from the past data. Determine the sample size needed.

Solution:

Given $E = 0.05 \quad and \quad Z_{\propto} = 1.645 \quad assume \ p = q = 0.5$

$$\therefore \quad n = \frac{(Z_{\propto})^2 \ p * q}{E^2} = \frac{(1.645)^2(0.5 * 0.5)}{0.05^2}$$

$$n = 270.6 \quad or \ \ 271 \ app$$

Example: 35

The university is conducting raising tuition fees to improve school facilities, and they want to determine what percentage of students favour the increase. The university needs to be 90% confident that the percentage has been estimated to within 2% of the true value. How large a sample is needed to guarantee this accuracy regardless of the true percentage?

Solution:

Given $E = 0.02$ and $Z_\propto = 1.645$ $assume$ $p = q = 0.5$

$$\therefore \quad n = \frac{(Z_\propto)^2 \ p * q}{E^2} = \frac{(1.645)^2 \ (0.5 * 0.5)}{0.02^2}$$

$$n = 1701.5 \quad or \ 1702 \ students \ app$$

Example:

An article in the journal of testing and evaluation presents the following 20 measurements on residual flame time (in seconds) of treated specimen of children's night wear:

9.85	9.93	9.75	9.77	9.67
9.87	9.67	9.94	9.85	9.75
9.83	9.62	9.74	9.99	9.88
9.85	9.95	9.93	9.92	9.89

Assume that residual flame follows normal distribution find 95% confidence interval on the mean residual flame time.

Solution:

Here $n = 20$, $1 - \alpha = 0.95$, $\alpha = 0.05$ $(\sigma \ unknown)$

x	$x - \overline{x}$	$(x - \overline{x})^2$
9.85	0	0
9.87	-0.02	0.0004
9.83	0.02	0.0004
9.95	-0.1	0.01
9.93	-0.08	0.0064
9.67	0.18	0.0324
9.92	-0.07	0.0049
9.95	-0.1	0.01
9.75	0.1	0.01
9.94	-0.09	0.0081
9.74	0.11	0.0121
9.93	-0.08	0.0064
9.77	0.08	0.0064
9.85	0	0

9.99	-0.14	0.0196
9.92	-0.07	0.0049
9.67	0.18	0.0324
9.75	0.1	0.01
9.88	-0.03	0.0009
9.89	-0.04	0.0016
197.05		0.1769

$$\sum x_i = 197.05 \qquad \sum (x_i - \overline{x})^2 = 0.1769$$

$$W.k.t \quad \overline{x} = \frac{\sum x_i}{n} = \frac{197.05}{20} = 9.85$$

$$s^2 = \frac{\sum (x_i - \overline{x})^2}{n-1} = \frac{0.1769}{19} = 0.0093$$

Confidence interval is given by

$$\overline{x} \pm t_{\frac{\alpha}{2},\ n-1} \frac{s}{\sqrt{n}}$$

$$= 9.85 \pm 2.309 \frac{\sqrt{0.0093}}{\sqrt{20}}$$

$$= 9.85 \pm 0.0498$$

$$= 9.8002, \ \ 9.8998$$

Therefore the 95% confidence interval is $9.8002, \ 9.8998$.

Example:

An insurance company policy sells foreign travel policy to those going abroad. The company in reputed to settle the claims within a period of 2 months. However, the new CEO of the company came to know about the delay in settling claims. He, therefore, ordered the concerned official to take a sample of 100 claims, and report the proportion of cases which were settled within 2 months. The CEO received the proportion as 0.6. What are the 95% confidence limits for such proportion?

Solution:

$$Here \quad n = 100, \quad p = 0.6$$

$$p \pm Z_{\frac{\alpha}{2}} \sqrt{\frac{pq}{n}} \quad where \quad q = 1 - p$$

$Z_{\frac{\alpha}{2}} = 1.96 \ at \ 95\%$ Confidence level

$$0.6 \pm 1.96\sqrt{\frac{0.6*0.4}{100}} = 0.6 \pm 0.096$$

$$= 0.504,\ \ 0.696$$

Thus, the CEO could be 95% confident that about 50 t0 70% of claims are settled within two months.

Example:

A national television network samples 1400 voters after each has cast a vote in a state gubernatorial election. Of these 1400 voters, 742 claim to have voted for the Democratic candidate and 658 for the Republican candidate. There are only two candidates in the election.

i) Assuming that each sampled voter actually voted as claimed and that the sample is a random sample from the population of all voters is there enough evidence to predict the winner of the election? Base your decision on a 95% confidence interval.

ii) Base your decision on a 99% confidence interval. Explain why it requires greater evidence to make a prediction when we require greater confidence of being correct.

Solution:

$$Here\ \ n = 1400,\ \ p\ = \frac{x}{n} = \frac{742}{1400} = 0.53$$

$$p \pm Z_{\frac{\propto}{2}}\sqrt{\frac{pq}{n}}\quad where\quad q = 1 - p$$

$$(\mathbf{i}).\quad \mathbf{Z_{\frac{\alpha}{2}}\ = 1.96\ \ at\ \ 95\%\ \ confidence\ interval}$$

$$= 0.53\ \pm 1.96\sqrt{\frac{0.53*0.47}{1400}}$$

$$= 0.53\ \pm 0.026\ = 0.504,\ \ 0.556$$

Since our interval is over 0.50, we can be reasonably confident that the Democratic candidate is the winner.

$$(\boldsymbol{ii}).\quad \boldsymbol{Z_{\frac{\alpha}{2}} = 2.58\quad at\ \ 99\%\ \ confidence\ interval}$$

$$= 0.53\ \pm 2.58\sqrt{\frac{0.53*0.47}{1400}}$$

$$= 0.53\ \pm 0.034\ = 0.496,\ \ 0.564$$

Since our interval includes .5 (and goes slightly below), we cannot confidently determine a winner based on this sample.

Example:

According to an automobile agency, average lives of two premium brands, 'A' and 'B' of car tyres are 45,000 kms and 42,000 kms. Suppose that these mean lives are based on random samples of 50 brand 'A' and 40 brand 'B', and that the standard deviations of these two brands were 3000 kms and 2000 kms, respectively.

i) What is the point estimate of the difference in mean lives of the two tyres?

ii) Construct a 95% confidence interval for difference between the two means.

Solution:

Let m_1 and m_2 are the population means of two premier brands of tyres 'A' and 'B' respectively.

$$Here\ \ \overline{x_1} = 45000,\ \ \overline{x_2} = 42000,\ \ n_1 = 50,\ \ n_2 = 40,\ \ \sigma_1 = 3000,\ \ \sigma_2 = 2000,$$

$$Z_{\frac{\propto}{2}}\ \ at\ 95\%\ \ confidence\ l\ \ evel = 1.96$$

95% confidence interval for the difference between two means is given by

$$(\overline{x_1} - \overline{x_2})\ \pm\ Z_{\frac{\propto}{2}}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

$$= (45000 - 42000) \pm 1.96\sqrt{\frac{3000^2}{50} + \frac{2000^2}{40}}$$

$$= 3000\ \pm\ 1.96\ 529.15\ = 3000\ \pm\ 1037.13$$

$$= 1962.87\ ,4037.13$$

Therefore 95% confidence interval for the difference between two premier brands of tyres is 1962.87 to 4037.13.

Example:

A soap manufacturing company wanted to estimate the difference between the proportions of loyal users of its soap in urban and rural areas. In a sample of 1200 users from urban areas, 300 users were found to be loyal users, and in the sample of 1500 from rural areas, 300 were found to be loyal. What is the point estimate of the difference of proportion of loyal users of the soap in urban and rural areas; Construct a 95% confidence interval for the proportion of the same.

Solution:

Let p_1 and p_2 be the proportion of loyal users in urban and rural areas respectively.

Here $X_1 = 300,\ X_2 = 300,\ n_1 = 1200,\ n_2 = 1500\ ,\ Z_{\frac{\propto}{2}}$ at 95% $confidence\ level$=1.96

$$p_1 = \frac{X_1}{n_1} = \frac{300}{1200} = 0.25$$

$$p_2 = \frac{X_2}{n_2} = \frac{300}{1500} = 0.20$$

$$q_1 = 1 - p_1 = 1 - 0.25 = 0.75$$

$$q_2 = 1 - p_2 = 1 - 0.20 = 0.8$$

The point estimate of the difference of proportion of loyal users of the soap in urban and rural areas is

$$p_1 - p_2 = 0.25 - 0.20 = 0.05$$

95% confidence interval for the difference between two proportions is given by

$$(p_1 - p_2) \pm Z_{\frac{\propto}{2}} \sqrt{\frac{p_1\ p_2}{n_1} + \frac{q_1 q_2}{n_2}} \quad where \quad q_1 = 1 - p_1 \ \ and \ \ q_2 = 1 - p_2$$

$$= 0.05 \pm 1.96 \sqrt{\frac{0.25 * 0.75}{1200} + \frac{0.20 * 0.80}{1500}}$$

$$= 0.05 \pm 1.96\,(0.0162) = 0.05 \pm 0.032$$

$$= 0.018, \quad 0.082$$

Therefore 95% confidence interval of the difference of proportion of loyal users of the soap in urban and rural areas is $0.018, \ 0.082$.

Example:

A company wants to determine the average time to complete a certain job. The past records show that the s.d of the completion times for all the workers in the company has been 10 days, and there is no reason to believe that this would have changed. However, the company feels that because of the procedural changes, the mean would have changed. Determine the sample size so that the company may be 95% confident that the sample average remains within ± 2 days of the population mean.

Solution:

Given that the company may be 95% confident that the sample average remains within ± 2 days of the population mean and $\sigma = 10$.

$$i.e \ \ |\overline{x} - \mu| < 2$$

$$p(|\overline{x} - \mu| < 2) = 0.95$$

$$p\left(\frac{|\overline{x} - \mu|}{\frac{\sigma}{\sqrt{n}}} < \frac{2}{\frac{\sigma}{\sqrt{n}}}\right) = 0.95$$

$$p\left(|z| < \frac{2}{\frac{\sigma}{\sqrt{n}}}\right) = 0.95 \qquad \ldots.(1) \ \ where \ \ |z| = \frac{|\overline{x} - \mu|}{\frac{\sigma}{\sqrt{n}}}$$

We know that $P\,(|z| < 1.96) = 0.95 \qquad \ldots\ldots(2)$

From (1) and (2), we get

$$\frac{2}{\sigma/\sqrt{n}} = 1.96 \implies \frac{2\sqrt{n}}{\sigma} = 1.96$$

$$\frac{2\sqrt{n}}{10} = 1.96 \implies \sqrt{n} = 1.96 * 5 = 9.8$$

$$\therefore \; n = 9.8^2 = 96.04$$

Approximately the sample size is 97.

Example:

20% of the population of a town is supposed to be rice eaters. At 95% level of confidence, what should be the sample size, so that the sampling error is not more than 5% above or below the true proportion of rice eaters?

Solution:

Let p_0 be the proportion of rice eaters in a town.

$Given \; p_0 = 20\% = 0.2, \; \alpha = 5\%, \; p - p_0 \leq 0.05$

$$q_0 = 1 - p_0 = 1 - 0.2 = 0.8$$

$$P\{|x - m| \leq 0.05\} = 0.95$$

$$P\left(\frac{|\hat{p} - p_0|}{\sqrt{\frac{p_0 q_0}{n}}} \leq \frac{0.05}{\sqrt{\frac{p_0 q_0}{n}}}\right) = 0.95$$

$$P\left(|z| \leq \frac{0.05}{\sqrt{\frac{p_0 q_0}{n}}}\right) = 0.95 \quad where \quad z = \frac{|\hat{p} - p_0|}{\sqrt{\frac{p_0 q_0}{n}}}$$

$$\frac{0.05}{\sqrt{\frac{p_0 q_0}{n}}} = 1.96 \qquad Since \quad P(|z| \leq 1.96) = 0.95$$

$$\frac{0.05}{\sqrt{\frac{0.2 * 0.8}{n}}} = 1.96 \implies \frac{0.05\sqrt{n}}{\sqrt{0.16}} = 1.96$$

$$\sqrt{n} = \frac{1.96 * \sqrt{0.16}}{0.05} = 15.68$$

$$\therefore \; n = (15.68)^2 = 245.86$$

Therefore the sample size is approximately 246.