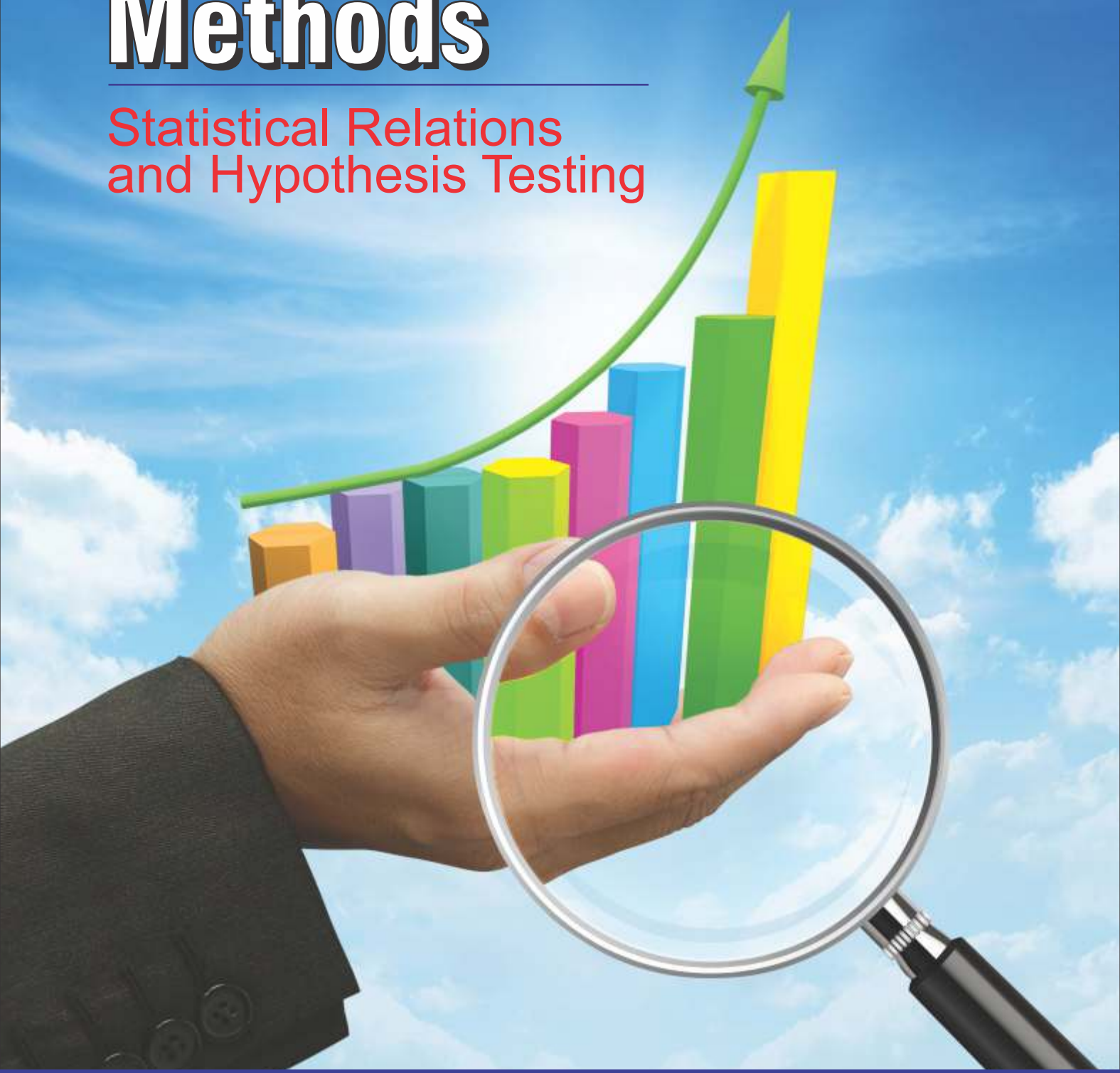
Block-II

Quantitative Methods

Statistical Relations and Hypothesis Testing



Block

II

STATISTICAL RELATIONS AND HYPOTHESIS TESTING

Ref. No. QM SLM 102021B2

For any clarification regarding this book, the students may please write to The ICFAI Foundation for Higher Education (IFHE), Hyderabad specifying the unit and page number.

While every possible care has been taken in type-setting and printing this book, The ICFAI Foundation for Higher Education (IFHE), Hyderabad welcomes suggestions from students for improvement in future editions.

Our E-mail id: cwfeedback@icfaiuniversity.in

Center for Distance and Online Education (CDOE)
The ICFAI Foundation for Higher Education
(Deemed-to-be-University Under Section 3 of UGC Act, 1956)
Donthanapally, Shankarapalli Road, Hyderabad- 501203.

BLOCK II: STATISTICAL RELATIONS AND HYPOTHESIS TESTING

Any management or business decision is based on data and its analysis. Any management decision cannot be based on unproven hypothesis. Hence, hypothesis testing is very much required for management decision making and implementation. Statistical techniques are used in hypothesis testing in management or business problems. Usually the hypotheses prove or disprove the existence of relationships, impacts and effects between different variables. To find out the relationship between two different variables correlation is used. To find out the impact of one variable on other, linear regression is used. In this block, statistical inference and hypotheses testing, correlation and linear regression are discussed.

Unit-5 Statistical Inference and Hypothesis Testing discusses about the various types of samples, sampling distribution, central limit theorem, and testing of hypothesis. Drawing conclusions on population based on study of a sample taken from population is known as statistical inference. A population means a collection of all the data points under study. Different statistical indices can be used to test the hypothesis.

Unit-6 Correlation and Linear Regression deals with the study of linear regression and coefficient of correlation. Linear regression constructs linear models while correlation coefficient gives a measure of the relevance of the model. The correlation analysis determines the degree to which the variables are related, regression analysis develops the relationship between the variables, and coefficient of correlation indicates the strength of a linear relationship between them. Regression analysis will bring cause and effect relationships between two or more variables.

Unit 5

Statistical Inference and Hypothesis Testing

Structure

5.1 Introduction

In the last unit of previous block, you studies different types of probability distributions and how to make decisions under certainty as well as under uncertainty. In this unit, you will study, sampling, standard error and test of hypothesis. Many a time, it is not possible to study each and every component of the entire data points under study due to time and cost constraints. In such cases, a few components are observed and statistical measures are applied. This process of inferring something about a large group called ‘population’, by studying a small part (sample) of it is called ‘sampling’. Population can be studied by enumeration and sampling. Due to constraints such as time, money and energy, most managers prefer the sampling method when the population is too large. This method may lead to errors due to the element of chance. Often, managers believe on some result called ‘hypothesis statement’. But sample data sometimes does not support the result obtained. The process that tests the possibility of the difference occurring due to chance elements is known as ‘hypothesis testing’.

5.2 Objectives

After going through the unit, you should be able to:

- List the advantages and disadvantages of sampling;
- Recall various types of sampling;
- Define sampling distribution;
- Explain central limit theorem;
- Define point estimates and interval estimates;
- Generalize testing of hypothesis; and
- Explain test of significance for difference between two means.

5.3 Population and Sample

The population is the set of data to which findings are to be generalized. For example, if we want to study about people living below poverty line in India, then the data point which represents such people is population. The basic idea of statistics is to extrapolate from the data you have collected to make general conclusions about the larger population from which the data sample was derived. For example, it is very difficult to measure the average life of the tubelights manufactured by a company. But, we can collect some pieces of tubelights (called 'sample') and based on the average life of this sample we can take the decision about population.

So, sample is a collection of data from the total population as: (a) All people below poverty line in Hyderabad, (b) All people below poverty line in South India.

Samples are a subset of the population and always lesser in size. Characteristics of population are known as 'parameters', whereas characteristics of sample are known as 'statistics'.

The following exhibit will facilitate us to understand the population and sample.

Exhibit 5.1: Post-poll Surveys by CSDS

Post-poll surveys were conducted by many research institutions. One such study was by the Centre for the Study of Developing Societies (CSDS), an Indian research institute through its research programme Lokniti along with The Hindu, in 4 different states - Assam, Kerala, Tamil Nadu and West Bengal in 2021 where the Assembly elections were conducted. The surveys were conducted from March 28, 2021 through May 1, 2021 in different phases in the states in the local languages.

Contd….

The survey was conducted with 3473 voters in Assam, 3424 voters in Kerala, 4354 voters in Tamil Nadu and 4223 voters in West Bengal. The study report presented many tables and charts based on the responses to various questions. The study concluded that the people verdicts in these states were based on the respective local issues in each state.

Source: https://www.thehindu.com/opinion/op-ed/local-factors-determine-electoral-outcomes-in-states/article34475075.ece and https://www.lokniti.org/POST_POLL_ANALYSIS_2021

As given in the Exhibit 5.1, sample data was collected by CSDS, then the data was classified and analysed.

Population: All eligible voters.

Population parameter: The proportion of all eligible voters who favor a particular candidate.

Sample: The eligible voters surveyed - 3473 voters (Assam), 3424 voters (Kerala), 4354 voters (Tamil Nadu) and 4223 (West Bengal).

Sample statistics: The proportion of eligible surveyed voters in each state who favor a particular candidate's win.

The following notations are used to denote population parameters and sample statistics in this unit.

	Population	Sample
Size	N	n
Mean	μ	$\bar{X}$
Standard Deviation	σ	s
Proportion	p	$\bar{p}$

The two methods of enumeration are – compete enumeration method or the census method and selective method or sample method. Both methods have their own advantages and disadvantages.

5.3.1 Advantages and Limitations of Sampling

Advantages

Sampling has the following advantages over a census (study of the entire population):

- **Cost:** The total cost of a sample will be much less than that of the population.
- **Time-lag:** With smaller number of observations it is possible to provide results much faster as compared to the total number of observations.

- **Less Work:** Sampling involves less work as compared to a census. Hence, the chances of errors while processing the data are less.
- **Destructive Testing:** Sometimes we cannot select the whole population for a test. For example, if a manufacturer wants to find out the average life of the tubelights he cannot use every tubelight for testing, as no tubelights would be left for the customers. Hence, sampling is used determine the average life of a tubelight in such a case.
- **Greater Scope:** Sampling has a greater scope regarding the variety of information by virtue of its flexibility and adaptability.

Finally, if population is too large it is not possible to test the entire population except using sampling.

Limitation of Sampling

- Possibility errors due to sampling may be high for small administrative areas.
- It may not be feasible for problems that require very high accuracy.

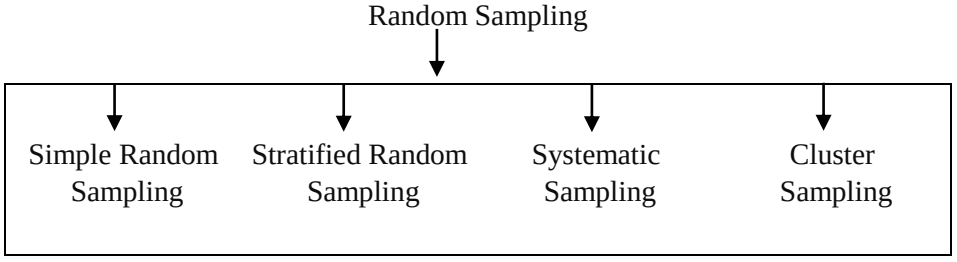
5.4 Types of Sampling

There are various methods of sampling. Based on the selection technique, it may be: (i) Non-random or Judgmental Sampling, (ii) Random or Probability Sampling.

Non-random or Judgmental Sampling: In this case, sample is normally selected on the personal judgment or convenience of the enumerator. Judgment sampling avoids the statistical analysis that is necessary to make probability samples. For example, the population under study for interview may be people living in Hyderabad. The investigator may choose as a sample, a certain number of his friends living in Hyderabad. He may exercise all care in attempting to make the sample representative of the population.

Random or Probability Sampling: Every element of the population has a known chance of being included in the sample. Random sampling is widely used in statistics as it is a more convenient and valid method of sampling. It may be of different types as illustrated by Figure 5.1.

Figure 5.1: Random Sampling



5.4.1 Simple Random Sampling

In simple random sampling, we select a certain number of elements randomly. Here each possible sample has an equal chance of being selected and each item in the entire population also has an equal chance of being selected. The selection is usually made with the help of random numbers.

Example 1

Suppose there are N = 500 students in a school from which a sample of n = 10 students is to be taken. The students are numbered from 1 to 100. Since our data runs into three digits we use random numbers that contain three digits. All numbers exceeding 500 are ignored because they do not correspond to any serial numbers in the data. No number is repeated twice. In this way, each possible sample has an equal chance of being selected. Further, each item in the entire population also has an equal chance of being selected.

5.4.2 Systematic Sampling

Systematic sampling is the most commonly used method of sampling. It is also called simple random sampling. Elements are selected from the population at a fixed interval (at Kth interval). Each element has an equal chance of being selected but each sample does not have an equal chance of being selected.

Example 2

Suppose the HAL colony is divided into N = 450 blocks numbered consecutively. A 10 percent sample of blocks is to be taken, which gives a sampling interval of k = 10. If the random number between 1 and 10 is 5, the blocks with the numbers 05, 15, 20, 25, 30... 450 may be selected as the sample.

5.4.3 Stratified Sampling

Stratified sampling is more complex than systematic sampling. The whole population is first divided into homogenous mutually exclusive groups called strata. The segments are based on some predetermined criteria such as geographic location, size or demographic characteristic. After forming the strata, we can either select at random from each stratum a specified number of elements corresponding to the proportion of that stratum in the population as a whole, or draw an equal number of elements from each stratum and give weight to the results according to the stratum’s proportion of total population.

Example 3

The Population of Hyderabad may be divided into different strata such as students, commuters, senior citizens, etc.

5.4.4 Cluster Sampling

Cluster sampling divides the population into groups or clusters, and then selects a random sample from these clusters. This approach overcomes the constraints

of costs and time associated with a much dispersed population. Cluster sampling differs from sratified sampling as in the latter, the elements of each stratum are homogeneous (there are relatively minor variations within them). In cluster sampling, the elements of each cluster are not homogeneous but representative of the population.

Example 4

We may divide the city of Hyderabad into 4-5 sectors say Zone I, Zone II, Zone III, Zone IV and Zone V. From the addresses on the bills we could classify the customers in the following clusters:

Zone I	people
Zone II	people
Zone III	people
Zone IV	people
Zone V	people

5.5 Statistical Errors

Estimation based on sample or even population sometimes leads to error. These errors may be due to the error in collection process, or in process of analysis. For example, if we want to estimate the average savings of workers for the year 2004-2005, we would have to question every worker in our population or some sample of such workers. In either case, the average savings calculated may not be the true average savings for the population. Based on the types of errors, they may be classified into – non-sampling errors and sampling errors.

5.5.1 Non-sampling Errors

Non-sampling error occurs at the time of collection, and editing and is common to both sampling and census. Such errors may occur in a sample or in a census. For example, missing some people and double counting others, respondents giving incorrect answers or not answering some questions, imprecise questions, interviewers leading the respondent's answer or giving incorrect information, interviewing the wrong unit, and not capturing or coding the responses correctly.

5.5.2 Sampling Errors

Sampling error is the difference between the value of sample statistics and the value of the corresponding population parameter. For example, in case of mean sample error it can be written as $\overline{X} - \mu$. Sampling errors occurs because of chance. Normally, sampling error originates at the time of collecting samples.

5.6 Sampling Distribution

As discussed earlier, the value of population parameter (population mean, population standard deviation, population proportion, etc.,) is always constant, but the same is not true for sample statistics (sample mean, sample standard deviation, sample proportion of defectives, etc). Different samples from a population may have different sample statistics; they depend on the element selected in the sample. Further, sampling is applicable to random samples only, whose mean can express probability distribution. This distribution is called sampling distribution. Sampling theory helps us to determine whether the differences between two samples are actually due to chance variation or whether they are really significant.

> In general, probability distribution of sample statistics is called sampling distribution.

5.6.1 Population and Sample Proportion

Population proportion is the ratio of number of elements in a population with the specific characteristics. In statistics it is denoted by p. The sample proportion is the ratio of number of elements in a sample with the specific characteristics. So,

$$p = \frac{X}{N} \text{ and } \hat{p} = \frac{x}{n}$$

Where,

N = Total elements in the population,

n = Total elements in the sample,

X = Total elements in the population that possess that specific character,

x = Total elements in the sample that possess that specific character.

5.6.2 Standard Error of Statistics

Standard error of the statistics in the sampling distribution is standard deviation of the collected sample. For example, standard error of mean is equal to the standard deviation of sample distribution. Mean of all collected samples and mean of population is always equal but standard deviation of a sample mean may not be equal to the standard deviation of population, because variability of sample mean may not be same as the variability of population. The standard deviation or standard error of sample mean can be calculated by using formula:

$$\sigma_{\bar{x}} = \sigma/\sqrt{n}$$

This formula is useful when: (i) Sample is selected from finite population and without replacement. (ii) Sample is selected from infinite population with or

without replacement. If sample is selected from finite population and with replacement, the formula will be as follows:

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}},$$

Where,

N is the size of population and n is the size of sample.

Sampling is always done without replacement. While sampling n different items, if n is small with respect to N, we may regard "sampling without replacement" to be the same as "sampling with replacement". This is similar to substituting the Hypergeometric Distribution with the Binomial Distribution if (n/N) is small enough. Here, this will be the case when $(n/N) \leq 0.05$.

Following are the two important observations regarding sampling distribution: (i) Standard deviation of sampling distribution decreases if the sample size increases. (ii) Population standard deviation is more than the sample standard deviation. The formulae for the Standard Errors (S.E.) of some well-known statistics for random samples are given below:

a. Sample mean $\bar{X}$: $\sigma_{\bar{x}} = \sigma/\sqrt{n}$

b. Sample proportion $\bar{p}$: $\sigma_{\bar{p}} = \sqrt{(p\,q)/n}$, where, $q = 1 - p$

c. Difference of two sample means $\bar{X}_1$ and $\bar{X}_2$: i.e.,

$$\sigma_{\bar{X}_1 - \bar{X}_2} = \sqrt{(\sigma_1^2/n_1) + (\sigma_2^2/n_2)}$$

Where, $\bar{X}_1$ and $\bar{X}_2$ are the means of two random samples of size n_1 and n_2 drawn from two population with standard deviation σ_1 and σ_2 respectively.

d. Difference of two proportions:

$$\sigma_{(\bar{p}_1 - \bar{p}_2)} = \sqrt{(\hat{p}\hat{q}/n_1) + (\hat{p}\hat{q}/n_2)}$$

Where, $\bar{p}_1$ and $\bar{p}_2$ are the proportions of two random samples of size n_1 and n_2 drawn from two population and $\hat{p} = (n_1\,\bar{p}_1 + n_2\,\bar{p}_2)/(n_1 + n_2)$.

e. For a finite population of size N, when a sample is drawn without replacement, $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}}$

If the value of the sampling fraction n/N is less than 0.05, then finite population correction factor $\sqrt{(N-n)/(N-1)}$ need not be used.

5.6.3 Applications of Standard Error

Standard error is used in various applications of statistics. Standard error is a useful tool when it is used in tests of hypotheses or tests of significance. It gives an idea about the reliability and precision of a sample. It helps in determining the limits (confidence limits) within which the parameters are expected to lie.

Example 5

A simple random sample of size 81 is drawn from a finite population consisting of 225 units. If the population Standard Deviation is 14, find the standard error of sample mean when the sample is drawn (a) with replacement, (b) without replacement.

a. If the sample is drawn with replacement, then Standard Error of sample mean: $\sigma_{\bar{X}} = \sigma/\sqrt{n} = 14/\sqrt{81} = 14/9 = 1.55$

b. If the sample is drawn without replacement, then Standard Error of sample mean: $\sigma_{\bar{x}} = \sigma/\sqrt{n}.\sqrt{(N-n)/(N-1)}$

Here, σ = 14, n = 81, N = 225

$\sigma_{\bar{x}}$ = (14/9). $\sqrt{(225-81)/225-1)}$ = 1.55 x 0.801 = 1.24.

Example 6

The following details are available with regard to a hypothesis test on difference between means of two populations:

$n_1 = 16$ $n_2 = 9$

The samples are independently collected and it is assumed that the two populations have the same variance. The pooled estimate of the underlying common population variance is 217.04. What is the estimated standard error of difference between the means?

Estimated standard error of difference between the means:

$$= s_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}} = \sqrt{217.04\left(\frac{1}{16}+\frac{1}{9}\right)} = 6.138$$

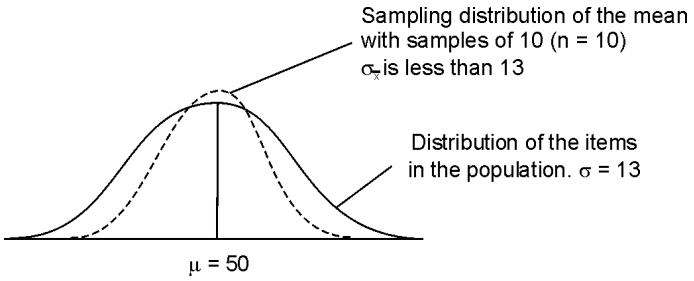
5.6.4 Shape of the Sampling Distribution

Shape of the sampling distribution depends on the shape of the population from which the sample has been taken. If the population is normally distributed, then the sample will also be normally distributed. In this case,

– Mean of sample $\mu_{\bar{x}} = \mu$ population mean

– Standard deviation of $\sigma_{\bar{x}} = \sigma/\sqrt{n}$

Figure 5.2: Population Distribution and Sampling Distribution of the Mean



In the normal situation most of the distribution is not exactly normal and the shape of distribution of the sample mean from which population is taken is also different. But for simplicity of decision, we consider most of the distribution as approximate to normal distribution. The approximation of normal distribution can be inferred by the central limit theorem.

5.7 Central Limit Theorem

According to the central limit theorem, sampling distribution of the sample mean $\overline{X}$ is approximately a normal distribution with mean μ and standard deviation which is equal to standard error $\overline{X}$, provided that the sample size n is sufficiently large. The mean and standard deviation of the sampling mean $\overline{X}$ are:

- Mean of sample mean $\mu_{\bar{x}} = \mu$ population mean
- Standard deviation of $\sigma_{\bar{x}} = \sigma/\sqrt{n}$

Further, if p be the proportion of defectives in a random sample of size n drawn from a population having the proportion of defectives p, then the sampling distribution of the sample proportion of defectives $\bar{p}$ is approximately a normal distribution with mean p and standard deviation is equal to standard error of $\bar{p}$, provided the sample size n is sufficiently large. The sample size is said to be large if $n \geq 30$, again remember that for $\sigma_{\bar{x}} = \sigma/\sqrt{n}$, n/N must be less than or equal to 0.05.

According to the central limit theorem if n approaches a large number, the distribution of sample mean approaches normality. Therefore, for large sample sizes ($n \geq 30$), regardless of the original population distribution, the normal distribution is a good approximation to the sampling distribution of x when σ is known, and this can be used as practical application of central limit theorem.

Sampling distribution of sample proportion $\hat{p}$ is approximately normal, if the sample size is sufficiently large. Sample size is considered to be large if $np > 5$ and $nq > 5$.

Example 7

The time that it takes to find a taxi when leaving a restaurant follows a left skewed distribution with a mean of 12 minutes and a standard deviation of 3 minutes. If 100 restaurant patrons are randomly sampled and the average time that it takes for them to find a taxi is calculated, then what type of sampling distribution will be obtained?

By central limit theorem for large samples, the sample mean is approximately normally distributed with mean = population mean, and standard deviation $\sigma/\sqrt{n}$

$\therefore E(\bar{x}) = \mu = 12$ min

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{3}{\sqrt{100}} = 0.30 \text{ min}$$

$\therefore$ Approximately normal distribution with $E(\bar{x}) = 12$ min and $\sigma_{\bar{x}} = 0.30$ min.

Example 8

A sample of size 96 has been taken from a population and the estimated standard error of proportion is found to be 0.05. What is the sample proportion?

$$\hat{\sigma}_{\bar{p}} = \sqrt{\frac{\hat{p}\hat{q}}{n}} = 0.05$$

or $$\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = 0.05$$

$$\therefore \quad \hat{p}(1-\hat{p}) = 96\,(0.05)^2$$

or $$\hat{p} - \hat{p}^2 = 0.24$$

or $$\hat{p}^2 - \hat{p} + 0.24 = 0$$

$$\therefore \ \hat{p} = \frac{-(-1) \pm \sqrt{(-1)^2 - (4 \times 1 \times 0.24)}}{2 \times 1} = \frac{1 \pm 0.20}{2} = 0.40.$$

5.8 Estimation

The value assigned to population parameter based on sample is called estimate. Estimate can be either point estimate or an interval estimate. Point estimate is a single number that estimates the exact value of the population parameter of interest whereas, an interval estimate includes a range of possible values that are likely to include the actual population parameter.

5.8.1 Point Estimates

A manager is sometimes interested in single value estimator than in interval. In point estimates normally we select a single value to a population parameter. The basic characteristic in choosing a sample statistic as a point estimate of the value of the population parameter is that the sample statistic be an unbiased estimator. Normally a statistician is interested in selecting unbiased and small variance.

5.8.2 Characteristics of a Good Estimator

The sample statistics (mean, median etc.), used for estimating a population parameter is called an estimator. The unbiased estimator should have the following characteristics:

Unbiased: In unbiased estimates of a population the expected value is equal to that of the parameter.

Consistency: Unbiased estimators are called as consistent if the difference between the estimator and parameter grows smaller as the sample size grows larger.

Efficiency: If there are two estimators then the estimator whose variance is less is said to be relatively efficient. We can show the standard error of the sample median as greater than $\sigma_{\bar{X}}$. Therefore $\bar{X}$ is more efficient.

Sufficiency: This is the degree to which all possible information in the sample is used, to estimate the corresponding parameter.

5.8.3 Point Estimator of Population Mean

The best estimator of population mean is the sample mean as it possesses the characteristics of a good estimator. Suppose, we want to find out the average starting salary offered by a software finance company to its employees in Hyderabad. As the data in the population may be too large, we may decide to study a random sample of such companies and thereby draw conclusions about the population mean. In this case, the sample mean $\bar{X}$ is a point estimator of the population mean μ.

5.8.4 Point Estimator of Population Proportion

Sample proportion possesses desirable characteristics of the estimator; therefore the sample proportion $\bar{p}$ is an estimator of the population proportion. Often we are interested in population proportions i.e., the ratio of observations in a population which can either have or do not have a certain characteristic. For example, (i) The proportion of students in class V whose age is below 10. (ii) The proportion of companies which have profit margin ratio of more than 20. In this case, we can study a random sample. If a certain percentage in a sample has a specific character we can use this as a rough estimate for whole proportion.

5.8.5 Point Estimator of Population Variance

The sample variance can be used as a point estimator of the population variance, and the sample standard deviation is a point estimator of the population standard deviation.

5.8.6 Interval Estimates

Point estimator is often insufficient and gives wrong idea about population. To overcome this problem we normally use interval estimate, which can give more clear idea about the population.

Terminology for Interval Estimate

An interval estimate is a range of values within which the actual value of the population parameter may fall.

Interval limits are the lower and upper values of the interval estimate.

A confidence interval is an interval estimate for which there is a specified degree of confidence that the actual value of the population parameter will fall within the interval. These are normally probability statements.

A confidence coefficient is the proportion of intervals that would include the population value in the long run if the process leading to the creation of the interval were repeated many times. It can be expressed as a fraction or a decimal.

A confidence level like the confidence coefficient, expresses the degree of certainty that the interval will include the actual value of the population parameter. It is expressed as a percentage for example, a 95 confidence coefficient is the equivalent of a 95 percent confidence level.

Accuracy is the difference between the observed sample statistic and the actual value of the population parameter being estimated. This may also be referred to as estimation error or sampling error.

Confidence Interval for a Population Mean (σ known)

When the sample size is more than 30, i.e., $n \geq 30$ as per central limit theorem, we can assume that underlying population is normally distributed and interval can be estimated using the formula given below: $\overline{X} \pm Z\sigma$, where, Z = the Z value corresponding to the level of confidence.

5.8.7 The Student t Distribution

When the sample size is small the normal distribution cannot give the appropriate result. So we can use t distribution to find the interval when: (a) Sample size is less than 30 ($n \leq 30$) (b) Population standard deviation is not known.

To find the t value in the table, perform the following steps: (i) Determine the confidence level. (For this example, assume 95%). (ii) Subtract this decimal from 1.00 to calculate the significance level (α). (For this example α = 1.00 – 0.95 = 0.05). (iii) Divide the result in half. (For this example, we would get 0.05/2 = 0.025). (iv) Determine the number of degrees of freedom, d.f = n – 1. (For example, for a sample size of 20, the d.f = 20 – 1 = 19). (v) Now we would look in the column headed for = 0.025 and row for df = 19 and find the respective value 2.0930.

Percentiles of the Student t Distribution Single Tail Probabilities of a Type I Error					
df/ α	0.100	0.050	0.025	0.010	0.005
19	1.3277	1.7291	2.0930	2.5395	2.8609

In this case, the confidence interval is $\bar{X} \pm t(s/\sqrt{n})$.

Example 9

A random sample of 64 vegetable vendors in a city was taken. The average price of potatoes was found to be Rs.7.50 per kg with a standard deviation of 48 paise per kg. You are required to find out the interval within which the true mean price per kg of potatoes for the entire city will fall 95% of the time.

Given: n = 64, $\bar{X}$ = 7.50 (Rs./kg.), s = 48 (Paise/kg) = 0.48 (Re/kg)

Since the standard deviation of population is not known, the sample standard deviation will be used as an estimate of the population standard deviation.

∴ Estimated standard error of mean,

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{s}{\sqrt{n}} = \frac{0.48}{\sqrt{64}} = 0.06$$

95% confidence interval is required.

∴ Z (from normal tables) = ± 1.96

Lower confidence limit = $\bar{X} - 1.96\ \sigma_{\bar{x}}$ = 7.50 – 1.96(0.06) = Rs.7.38/kg.

Upper confidence limit = $\bar{X} + 1.96\ \sigma_{\bar{x}}$ = 7.50 + 1.96(0.06) = Rs.7.62/kg.

Check Your Progress - 1

1. According to ----------if n approaches a large number, the distribution of sample mean approaches normality
2. A simple random sample of size 100 is drawn from a finite population consisting of 400 units. If the population Standard Deviation is 18, find the standard error of sample mean when the sample is drawn without replacement.
   a. 1.65
   b. 1.42
   c. 1.68
   d. 1.56
   e. 1.98
3. If the population variance and the sample size are 26 and 5 respectively, then the standard error of the sample mean will be
   a. 0.192
   b. 0.438
   c. 0.980
   d. 2.280
   e. 11.62.

4. A random sample of 100 vegetable vendors in a city was taken. The average price of potatoes was found to be Rs.36.50 per kg with a standard deviation of 90 paise per kg. You are required to find out the interval within which the true mean price per kg of potatoes for the entire city will fall 95% of the time.
   a. 36.32,36.67
   b. 36,38
   c. 36.5,37.4
   d. 35.8,37.4
   e. 36.32,37.
5. The set of data for which findings are to be generated is called……..
6. Characteristics of sample are called
   a. Parameters
   b. Statistics
   c. Probability
   d. Hypothesis
   e. Inference
7. Which of the following is not an advantage of sampling?
   a. Less cost
   b. Faster data
   c. Lesser work
   d. Destructive testing
   e. Highly accurate data
8. Which of the following is not a type of random sampling?
   a. Simple Random sampling
   b. Stratified random sampling
   c. Systematic sampling
   d. Cluster sampling
   e. Judgmental sampling
9. Fill in the blank in the following statement.

   Non sampling error occurs at the time of collection and …….

5.9 Testing of Hypothesis

Testing of hypothesis is a new concept for a new reader, but it's quite a familiar aspect of a real life problem. For example, a placement consultant recruiting fresh graduates from colleges for various companies observed that the mean

annual salary offered to the candidates placed by him in the retailing firms was Rs.480000 in 2019. From January 2020 to March 2020, the consultant placed as many as 150 fresh graduates in the retailing firms. The consultant is of the opinion that the mean annual salary offered by the retailing firms to the candidates placed by him from January 2020 to March 2020, is more than the mean salary offered by the same employers in 2019. He has collected a sample of 36 candidates placed in 2020 in the retailing firms and has observed that the mean annual salary offered to them is Rs.504000 with a standard deviation of Rs.18000. It is to be determined whether the candidates placed by the consultant from January 2020 to March 2020, in the retailing firms, have been offered a mean annual salary greater than the mean annual salary offered to them in 2019.

How can we solve this type of problem? Normally, the following steps are required in testing hypothesis. (a) State null and alternative hypothesis; (b) State decision rule; (c) Compute test statistic; (d) Draw a picture showing acceptance region(s), rejection region(s), and position of the decision rule; (e) Make the decision; and (f) Relate the decision to the context of the problem.

The first step in hypothesis testing is to make an assumption about the population. In this example, the consultant opinion is the assumption about the population. The next step is to gather sample and determine the sample statistic. There are actually two hypotheses that are tested. One is called null hypothesis and other is called alternative hypothesis. The null hypothesis asserts that there is no (significant) difference between the statistic and the population parameter. Alternative hypothesis contradicts the null hypothesis H_0, and is denoted by the symbol H_1. In general, the Null Hypothesis is H_0: $\mu = \mu_0$, H_0: $\mu \geq \mu_0$ or H_0: $\mu \leq \mu_0$ and the Alternative Hypothesis may be:

i. H_1: $\mu \neq \mu_0$ ($\mu > \mu_0$ or $\mu < \mu_0$) or

ii. H_1: $\mu > \mu_0$, or

iii. H_1: $\mu < \mu_0$.

The corresponding tests of hypotheses are called two-tailed (or two-sided), right-tailed (one-sided) and left-tailed (one-sided) tests respectively. The Alternative Hypothesis in (i) is known as a two-tailed alternative and in (ii) and (iii) is known as right-tailed and left-tailed alternatives respectively.

In our example,

Null hypothesis H_0 : population mean $\mu = 480000$

Alternative hypothesis H_1 : population mean $\mu > 480000$.

This is one tail right hand test. Normally, to test the validity of our hypothesis the difference between the hypothesized value and the actual value of the sample statistic will be determined. If the difference between the hypothesized population parameter and the actual value is large then we automatically reject our hypothesis. If it is small, we accept it.

5.9.1 Level of Significance: α

Level of significance defines unlikely values of sample statistic if null hypothesis is true i.e., the percentage cases that would be rejecting the hypothesis when it is actually true and is symbolized as α. For example, if the significant level is 5%, it means out of 100 we would reject about 5 cases when they should be actually accepted. $(1 - \alpha)$ is the probability of accepting a true null hypothesis, and is referred to as the confidence level. If the significant level is 5%, we can say that we are 95% confident that we have made the right decision. Acceptance and rejection region can be drawn in two ways:

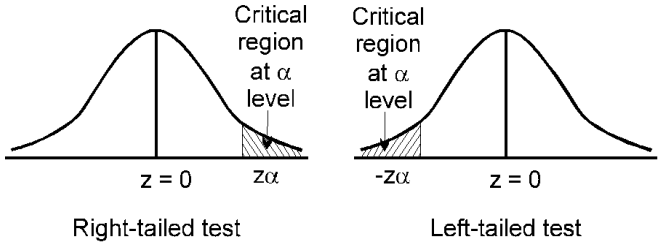
Scale of the Original Variable Method: In this method, we can compute limit(s) and determine where the sample statistic lies in relation to the limit(s).

Standardized Scale Method: In this method, we can compute a standardized statistic based on the sample data and compare where the standardized statistic lies in relation to the critical value as determined by the test's significance level.

5.9.2 Two-Tailed and One-Tailed Tests

As we discussed earlier, hypothesis test may be one-tail test or two-tail test. In one-tail test, rejection region is located in only one-tail of the sampling distribution whereas in two-tail test rejection region is located in both the tails of the sampling distribution. If the test is two-tailed, then we can write H_1: $\mu \neq \mu_0$ and if it is one-tailed (i.e., either right-tailed or left-tailed), then we can write $\mu \geq \mu_0$ or $\mu \leq \mu_0$.

Figure 5.3: A Picture Showing Acceptance Region, Rejection Region(s)



Finding Critical 'z' Values for Two-Tailed Test

If the significant level is given as 5%, how can we find the value of Z in a one-tail test?

- Calculate the value of $\alpha/2$ and deduct from 0.50 in the above example; it would be 0.50 – 0.025 = 0.475
- Look for the value of 0.475 in the Z-table.

Z	0.05	0.06	0.07
1.8	0.4678	0.4686	0.4693
1.7	0.4744	0.4750	0.4756

5.10 Large Sample Test for Mean (one-tailed test)

Example 10

Let us solve the following example:

A placement consultant recruiting fresh graduates from colleges for various companies observed that the mean annual salary offered to the candidates placed by him in the retailing firms was Rs.480000 in 2019. From January 2020 to March 2020, the consultant placed as many as 150 fresh graduates in the retailing firms. The consultant is of the opinion that the mean annual salary offered by the retailing firms to the candidates placed by him from January 2020 to March 2020, is more than the mean salary offered by the same employers in 2019. He has collected a sample of 36 candidates placed in 2020 in the retailing firms and has observed that the mean annual salary offered to them is Rs.504000 with a standard deviation of Rs.18000. It is to be determined whether the candidates placed by the consultant from January 2020 to March 2020, in the retailing firms, have been offered a mean annual salary greater than the mean annual salary offered to them in 2019.

First Method: Standardized Scale Method

Here,

H_0 : $\mu = 480000$

H_1 : $\mu > 480000$

Sampling fraction = $\frac{n}{N} = \frac{36}{800} = 0.045 < 0.05$

Hence, this should be treated as sampling from infinite population. Standard error of mean,

$$\hat{\sigma}_{\bar{x}} = \frac{S}{\sqrt{n}} = \frac{18000}{\sqrt{36}} = 3000$$

Since the sample size is 36 > 30 the normal distribution can be used,

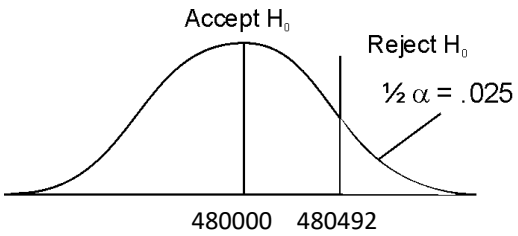
∴ Standardized value of sample mean,

$$Z = \frac{\bar{x} - \mu}{\hat{\sigma}_{\bar{x}}} = \frac{504000 - 480000}{3000} = \frac{24000}{3000} = 8$$

At the significance level of 5%, the critical value for the test = 1.64

Standardized value (8) > Critical value (1.64).

Figure 5.4: Accept/Reject H_0? Examples using Critical Values



Hence, the sample statistic falls in the rejection region. So, the null hypothesis is rejected and alternative hypothesis is accepted. Hence, at 5% significance level, it can be concluded that the mean annual salary offered to the candidates placed in the retail firms from January 2020 up to March 2020, is more than Rs.48000.

Second Method: Scale of the Original Variable Method

Here,

H_0 : $\mu = 480000$

H_1 : $\mu > 480000$

Sampling fraction = $\frac{n}{N} = \frac{36}{800} = 0.045 < 0.05$

Hence, this should be treated as sampling from infinite population.

Standard error of mean,

$$\hat{\sigma}_{\bar{x}} = \frac{S}{\sqrt{n}} = \frac{18000}{\sqrt{36}} = 3000$$

Since the sample size is 36 > 30 the normal distribution can be used.

At the significance level of 5%, the critical value for the test = 1.64

Standardized value (8) > Critical value (1.64).

$$\frac{\bar{x} - \mu}{\hat{\sigma}_{\bar{x}}} > 1.64$$

$$\bar{x} - 480000 > 492$$

$$\bar{x} > 480492$$

Since, $\bar{x} = 500000$

Hence, the sample statistic falls in the rejection region. So, the null hypothesis is rejected and alternative hypothesis is accepted. Hence, at 5% significance level, it can be concluded that the mean annual salary offered to the candidates placed in the retail firms from January 2020 up to March 2020, is more than Rs.480000.

Example 11

The mean lifetime of 100 electric bulbs produced by a manufacturing company is estimated to be 1570 hours with a standard deviation of 120 hours. If μ is the mean lifetime of all the bulbs produced by the company, test the hypothesis μ = 1600 hrs, using a level of significance of 0.05. This is a two-tailed test as we want to test if the hypothesis for mean lifetime is exactly 1600 hrs.

N = 100; $\bar{x}$ = 1570; S = 120; $\mu_0 = 1600$

$H_0 : (\mu = 1600)$; $H_1 : (\mu \neq 1600)$

The test statistic is: $Z = \frac{\bar{x} - \mu_0}{S/\sqrt{n}} = \frac{1570 - 1600}{120/\sqrt{100}} = -2.5$

The Critical Region at 5% level is $[|Z| \geq 1.96]$; calculated value falls in C.R. Hence, the null hypothesis H_0 is rejected. Hence, it can be concluded that the mean lifetime is not 1600 Hrs.

5.11 Small Sample Test for Mean

For a small sample, where the size is less than 30 and population standard deviation is not known, we can use, t-distribution as test statistics.

Example 12

A commodity merchant knows that the mean retail price of a specific variety of rice three months ago was Rs.14.50 per kg. In the current month the merchant has collected the information on the price charged for the same variety of rice by 16 randomly selected merchants in the same city. It was found from the sample that the mean retail price was Rs.15.00 per kg and the standard deviation was Rs.1.25 per kg. It is to be tested whether the mean retail price of the rice in the current month is more than Rs.14.50 per kg. At a significance level of 5 percent, what is the conclusion?

$H_0 : \mu = 14.50$; $H_1 : \mu > 14.50$; $\alpha = 0.05$

Estimated standard error of mean: $\hat{\sigma}_{\bar{X}} = \frac{s}{\sqrt{n}} = \frac{1.25}{\sqrt{16}} = 0.3125$

Since the population standard deviation is not known and the sample size is less than 30, the t-distribution will be used.

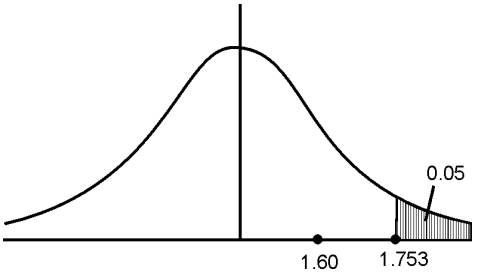
Degrees of freedom = n – 1 = 16 – 1 = 15

Since this is a right tailed test with $\alpha = 0.05$, the rejection area is 0.05 under the right tail. Hence, we should look under the 0.10 column in the t-distribution (0.10 area in both tails combined).

$\therefore$ Critical t-value = 1.753

Standardized sample statistic, $t = \frac{\bar{X} - \mu_{H_0}}{\hat{\sigma}_{\bar{X}}} = \frac{15.00 - 14.50}{0.3125} = 1.60$

Figure 5.5: The Graph of pdf for Standard Normal Distribution



We find that the sample statistic falls in the acceptance region. Hence, we accept the null hypothesis, H_0. It is concluded that the mean price of the rice is not more than Rs.14.50 per kg.

5.12 Hypotheses Involving Means of Two Populations (Large Sample)

Two samples are said to be independent if they are drawn from two different populations. This section discusses how the hypothesis test can be conducted to calculate the difference between two means of large and independent populations. Suppose for two independent populations, population means are μ_1 and μ_2 respectively. Two samples are taken, one from each population, of sizes n_1 and n_2 respectively. We need to test whether $\mu_1 = \mu_2$.

$\bar{X}_1$ = The mean of the sample drawn from population 1

$\bar{X}_2$ = The mean of the sample drawn from population 2.

From the central limit theorem, $\bar{X}_1$ and $\bar{X}_2$ is approximately normally distributed with standard deviation of $\left(\sigma_1/\sqrt{n_1}\right)$ and $\left(\sigma_2/\sqrt{n_2}\right)$. Using this, we can calculate the difference between the two means of large and independent populations. Standard error:

$$\sigma_{\bar{x}_1-\bar{x}_2} = \sqrt{\frac{\sigma^2_1}{n_1} + \frac{\sigma^2_2}{n_2}}$$

Null and alternative hypothesis are any of the following:

$$\begin{matrix} H_0 : \mu_1 = \mu_2 \\ H_a : \mu_1 \neq \mu_2 \end{matrix} \text{ (two tail)} \quad \begin{matrix} H_0 : \mu_1 \geq \mu_2 \\ H_a : \mu_1 < \mu_2 \end{matrix} \text{ (right tail)} \quad \begin{matrix} H_0 : \mu_1 \leq \mu_2 \\ H_a : \mu_1 > \mu_2 \end{matrix} \text{ (left tail)}$$

This can also be written as follows:

$$\begin{matrix} H_0 : \mu_1 - \mu_2 = 0 \\ H_a : \mu_1 - \mu_2 \neq 0 \end{matrix} \text{ (two tail)} \quad \begin{matrix} H_0 : \mu_1 - \mu_2 \geq 0 \\ H_a : \mu_1 - \mu_2 < 0 \end{matrix} \text{ (right tail)} \quad \begin{matrix} H_0 : \mu_1 - \mu_2 \leq 0 \\ H_a : \mu_1 - \mu_2 > 0 \end{matrix} \text{ (left tail)}$$

$$\text{Test statistic} = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\text{Standard Error}}$$

Because the difference between the normal distributes and random variables is also normal, we can take sample standard deviation in place of population standard deviation if the sample is sufficiently large; in this case, it is,

$$\sigma_{\bar{x}_1-\bar{x}_2} = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{s^2_1}{n_1} + \frac{s^2_2}{n_2}}}$$

Example 13

Two medical research laboratories, Alpha Pharma Ltd. (APL) and Orient Laboratories Ltd. (OLL), have been independently conducting research for producing drugs that will provide relief to those suffering from asthma. Very recently they have produced drugs for providing relief to the asthma patients. The drug produced by APL was tested on a group of 80 asthma patients and it provided an average of 10.5 hours of relief, with a standard deviation of 1.5 hours. The drug produced by OLL was tested on a group of 64 asthma patients and it provided an average of 9.6 hours of relief, with a standard deviation of 2.5 hours. It is to be tested whether the mean hours of relief provided by the drug produced by APL is more than the mean hours of relief provided by the drug produced by OLL. Let the following notations be used:

μ_1 = Population mean hours of relief provided by the drug of APL.

μ_2 = Population mean hours of relief provided by the drug of OLL.

$H_0 : \mu_1 = \mu_2$; $H_1 : \mu_1 > \mu_2$

Estimated standard error of the difference between the two means:

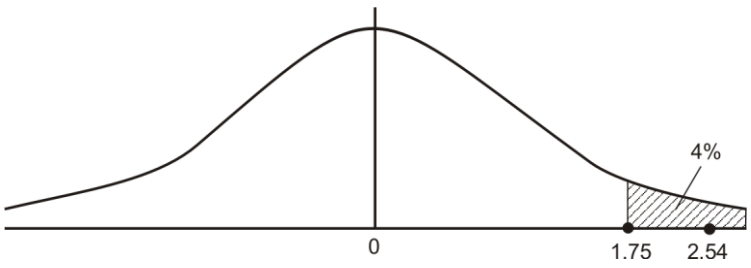
$$\hat{\sigma}_{\bar{x}_1-\bar{x}_2} = \sqrt{\frac{\hat{\sigma}_1^2}{n_1}+\frac{\hat{\sigma}_2^2}{n_2}} = \sqrt{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}} = \sqrt{\frac{(1.5)^2}{80}+\frac{(2.5)^2}{64}} = 0.3547 \text{ hours}$$

Standardization of difference between sample means:

$$Z = \frac{(\bar{x}_1-\bar{x}_2)-(\mu_1-\mu_2)_{H_0}}{\hat{\sigma}_{\bar{x}_1-\bar{x}_2}} = \frac{(10.5-9.6)-0}{0.3547} = 2.54$$

Critical value at 4 percent level of significance = 1.75

Figure 5.6: Critical Resin for Z Test



So, we can see that the difference between the sample mean falls in the rejection region. Hence, null hypothesis is rejected and alternative hypothesis is accepted. Therefore, at a significance level of 4%, we can conclude that the mean number of hours of relief provided by the drug produced by APL is significantly higher than OLL.

5.13 Hypotheses Involving Means of Two Populations (Small Sample)

For small sample, the pooled estimate of σ^2 and $\sigma_{\bar{x}_1-\bar{x}_2}$ can be calculated using the following equation:

$$s^2_p = \frac{(n_1-1)S_1^2+(n_2-1)S_2^2}{n_1+n_2-2}, \quad \hat{\sigma}_{\bar{x}_1-\bar{x}_2} = s_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}$$

Example 14

The following details are available with regard to a hypothesis test on difference between means of two populations:

$n_1 = 16$ $s_1 = 16$

$n_2 = 9$ $s_2 = 12$

The samples are independently collected and it is assumed that the two populations have the same variance. What is the estimated standard error of difference between means?

Pooled estimate of the underlying common population variance,

$$s_p^2 = \frac{(n_1-1)s_1^2+(n_2-1)s_2^2}{n_1+n_2-2} = \frac{(16-1)16^2+(9-1)12^2}{16+9-2} = 217.04$$

Estimated standard error of difference between means:

$$= s_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}} = \sqrt{217.04\left(\frac{1}{16}+\frac{1}{9}\right)} = 6.138$$

5.14 Type I and Type II Errors

In the testing of hypothesis, the null hypothesis is rejected or accepted on the basis of the sample statistic. Thus, the decision is always error prone. The possible errors in the testing of hypothesis are:

- Type I Errors – The probability of Type I error is also called the significance level of the test. If we use higher significance level for testing hypothesis, the higher is the probability of committing Type I error.
- Type II Errors occurs in Hypothesis testing when the Null Hypothesis is false and accepted. It is also represented by β or power of test.

Actual	Decision	
	Accept H_0	Reject H_0
H_0 is True	Correct Decision	Wrong Decision (Type I error)
H_0 is False	Wrong Decision (Type II error)	Correct Decision

There exists a tradeoff between the two types of errors. The probability of making one type of error can be reduced only if we are willing to increase the probability of making the other type of error.

Example 15

An electronic equipment manufacturer uses hypothesis testing for the inspection of incoming raw material. The null hypothesis is that the number of defective items in the lot is equal to the acceptable limit. The alternative hypothesis is that the number of defective items in the lot is more than the acceptable limit. If the manufacturer performs Type I error (i.e., Reject null hypothesis when it is true), then he will reject the good lot. If he performs Type II error (i.e., Accept null hypothesis when it is false), then he will accept the lot with number of defectives more than the acceptable limit. In this particular test, Type II error is more serious, so the β of the test should be very low.

5.15 Large Sample Test for Proportion

Often we conduct the test of hypothesis about a population proportion. For example, 40% of the employees included among them those working effectively for less than 8 hours a day. The Managing Director of the company wants to know whether this percentage holds good. To test the hypothesis that the proportion p in the population has a specified value p_0, we have to first check the size of sample.

The procedure includes these steps:

For a large sample size n (n > 30) sampling distribution of $\bar{p}$ is approximately normal. The Null Hypothesis is H_0: $p = p_0$. The Alternative Hypothesis may be (i) H_1: $p \neq p_0$ or (ii) H_1: $p < p_0$, or (iii) H_1: $p > p_0$.

The test statistic $z = \dfrac{\bar{p} - p_0}{\text{Standard Error of } p}$ where, Standard error of $p = \sigma_{\bar{p}} = \sqrt{pq/n}$.

Example 16

Western Electricals Company manufactures heavy duty voltage stabilizers. The General Manager of the company wants to compare the performance of the stabilizers manufactured by the company with the industry standards. He knows that 18% of all the stabilizers sold in the industry require repairs during the first year of sale. The company sampled 144 customers and found that in case of 36 customers the stabilizers required repairs in the first year of sale. Using normal distribution and a significance level of 5% you are required to determine if the performance of the stabilizers manufactured by the company is different from the industry standards.

Sample size, n = 144

Sample proportion, $\bar{p} = 36/144 = 0.25$

H0 : p = 0.18 (= p0); H1 : $p \neq 0.18$; $\alpha = 0.05$

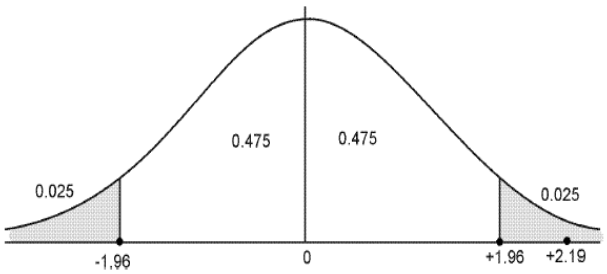
Standard error of proportion:

$$\sigma_{\bar{p}} = \sqrt{pq/n} = \sqrt{\frac{0.18(1-0.18)}{144}} = 0.032$$

Standardizing the sample proportion:

$$\frac{\bar{p} - p_0}{\text{Standard Error of } p} = \frac{0.25 - 0.18}{0.032} = 2.1875 \quad \text{or } 2.19$$

Figure 5.7: Critical Resin for Test of Proportion



From the normal distribution curve, we can find out that the sample proportion falls in the rejection region. Hence, we infer that the performance of the stabilizers manufactured by the company significantly differs from the industry standards.

Example 17

Consider the above example, and determine if the performance of the stabilizers manufactured by the company is less than the industry standards.

H0 : p = 0.18 (= p_0); H1 : p < 0.18

This is left hand test so $\alpha = 0.05$ and the rejection level will be Z < –1.96, but Z = 2.19, so it falls outside the rejection region. Hence, we infer that the performance of the stabilizers manufactured by the company is not less than the industry standards.

5.16 Hypotheses Involving Proportions of Two Populations

For any of the following hypotheses:

$H_0 : p = p$, $H_a : p \neq p$ (Two tail) $\quad H_0 : p \geq p$, $H_a : p < p$ (Right tail) $\quad H_0 : p \leq p$, $H_a : p > p$ (Left tail)

Or

$H_0 : p - p = 0$, $H_a : p - p \neq 0$ (Two tail) $\quad H_0 : p - p \geq 0$, $H_a : p - p < 0$ (Right tail) $\quad H_0 : p - p \leq 0$, $H_a : p - p > 0$ (Left tail)

To estimate the overall proportion of successes, we have to calculate the combined proportions from both the samples.

$$\hat{p} = \frac{n_1\bar{p}_1 + n_2\bar{p}_2}{n_1 + n_2}$$

Standard error = $\hat{\sigma}_{\bar{p}_1-\bar{p}_2} = \sqrt{\frac{\hat{p}\hat{q}}{n_1} + \frac{\hat{p}\hat{q}}{n_2}}$

The test statistic is:

$$z = \frac{(\bar{p}_1 - \bar{p}_2) - (p_1 - p_2)}{\sqrt{\frac{\hat{p}_1\hat{q}_1}{n_1} + \frac{\hat{p}_1\hat{q}_1}{n_2}}}$$ which has the standard normal distribution.

Example 18

A survey firm conducts door-to-door surveys on a variety of issues. Some individuals cooperate with the interviewer and complete the interview questionnaire while others do not. The following data are available.

Respondents	Sample Size	Number Not Cooperating
Men	200	110
Women	300	210

It is to be tested at a significance level of 0.05, whether the proportion of men and women who cooperate with the interviewer are equal. Which of the following can be inferred from the test?

p_1 : proportion of men cooperating

p_2 : proportion of women cooperating.

Null hypothesis: $p_1 = p_2$

Alternative hypothesis: $p_1 \neq p_2$

$$\bar{p}_1 = \frac{90}{200} = 0.45\,;\ \bar{p}_2 = \frac{90}{300} = 0.30$$

Estimated value of the population proportion,

$$p = \frac{(n_1\bar{p}_1) + (n_1\bar{p}_2)}{n_1 + n_2} = \frac{(200x0.45) + (300x0.30)}{200 + 300} = \frac{180}{500} = 0.36$$

$$\sigma_{\bar{p}_1-\bar{p}_2} \text{ (estimated)} = \sqrt{\left(\frac{p(1-\bar{p})}{n_1}\right) + \left(\frac{p(1-\bar{p})}{n_2}\right)}$$

$$= \sqrt{\left(\frac{0.36(1-0.36)}{200}\right) + \left(\frac{0.36(1-0.36)}{300}\right)}$$

Test statistic: $\frac{(0.45-0.30)-0}{0.04382} = 3.42$

Critical values: ± 1.96

The test statistic goes beyond the critical value in the right tail. Hence, we reject null hypothesis. We can conclude that the proportion of men and women who cooperate with the interviewer are not equal.

Check Your Progress - 2

10. Type II Errors occurs in Hypothesis testing when the Null Hypothesis is ----------

11. A sample of size 100 has been taken from a population and the estimated standard error of proportion is found to be 0.04. What is the sample proportion?
    a. 0.4
    b. 0.3
    c. 0.1
    d. 0.2
    e. 0.6

12. The mean lifetime of 1000 electric bulbs produced by a manufacturing company is estimated to be 1450 hours with a standard deviation of 80 hours. If μ is the mean lifetime of all the bulbs produced by the company, and we wish to test the hypothesis μ = 1500 hrs, compute the Z statistic.
    a. 19.76
    b. -19.76
    c. 21.12
    d. -21.12
    e. 26.76

13. The drug produced by CIPLA was tested on a group of 80 asthma patients and it provided an average of 12 hours of relief, with a standard deviation of 0.5 hours. The drug produced by ABOTT was tested on a group of 64 asthma patients and it provided an average of 10 hours of relief, with a standard deviation of 0.4 hours. It is to be tested whether the mean hours of relief provided by the drug produced by CIPLA is more than the mean hours of relief provided by the drug produced by ABOTT. Compute estimated standard error of the difference between the two means
    a. 0.25
    b. 0.15
    c. 0.075
    d. 0.75
    e. 0.015

5.17 Summary

- Sampling methods are widely used in statistics for estimating character of population. They are used due to various characteristics like low cost and time etc.
- Non-random and Random sampling are the two main types of sampling used in estimation. Simple random sampling, Stratified random sampling, Systematic sampling, Cluster sampling are the different types of random sampling. Estimation based on sample or even population sometimes leads to error. Sampling distribution helps to determine whether the differences between two samples are actually due to chance variation or whether they are really significant. In general, probability distribution of sample statistics is called sampling distribution.
- The standard deviation of the collected sample is normally termed as standard error of the statistics, which is the most useful tool used in tests of hypotheses. It helps in determining the limits (confidence limits) within which the parameters are expected to lie. The most common test is to test the significance of the mean and proportion. Some other advanced tests are test of difference of mean and test of difference of proportion.

5.18 Glossary

Alpha: The probability of a Type I error.

Alternate Hypothesis: The conclusion we accept when the data fails to support the null hypothesis.

Beta: The probability of a Type II error.

Central Limit Theorem: A result assuring that the sampling distribution of the mean approaches normality as the sample size increases, regardless of the shape of the population distribution from which the sample is selected.

Cluster Sampling: The population is divided into groups, or clusters, to select a random sample of these clusters.

Confidence Limits: The upper and lower boundaries of a confidence interval.

Estimate: A specific observed value of an estimator.

Estimator: A sample statistic used to estimate a population parameter.

Hypothesis: An assumption made about population parameter.

Null Hypothesis: The hypothesis, or assumption, about a population parameter we wish to test, usually an assumption of the *status quo*.

One-Tailed Test: A hypothesis test in which there is only one rejection region.

Population: The set of data to which findings are to be generalized.

Sample: A collection of data from population.

Stratified Sampling: The population is divided into relatively homogeneous groups, called strata. Then either at random from each stratum a specified number of elements corresponding to the proportion of that stratum in the population as a whole are selected or an equal number of elements from each stratum are drawn and given weight to the results according to the stratum's proportion of total population.

Systematic Sampling: Elements are selected from the population at a uniform interval that is measured in time, order, or space.

Two-Tailed Test: A hypothesis test in which the null hypothesis is rejected if the sample value is significantly higher or lower than the hypothesized value of the population parameter.

Type I Error: Rejecting a null hypothesis when it is true.

Type II Error: Accepting a null hypothesis when it is false.

5.19 Suggested Readings/Reference Material

1. Gupta, S. P. Statistical Methods. 46th Revised ed. New Delhi: Sultan Chand & Sons. 2021.
2. I. Levin Richard, H. Siddiqui Masood, S. Rubin David, Rastogi Sanjay. Statistics for Management. Pearson Education; Eighth edition, 2017.
3. Gerald Keller. Statistics for Management and Economics. Cengage, 2017.
4. Arora, P. N., and Arora, S. CA Foundation Course Statistics. 6th ed. S Chand Publishing, 2018.
5. Mario F Triola. Elementary Statistics. 13th ed., 2018.
6. David R. Anderson, Dennis J. Sweeney, Thomas A. Williams, Jeffrey D. Camm, James J. Cochran. Statistics for Business and Economics. 13th Edition, Cengage Learning India Pvt. Ltd., 2019.
7. S D Sharma. Operations Research. Kedar Nath Ram Nath, 2018.
8. Hamdy A. Taha. Operations Research: An Introduction. 10th ed., Pearson, 2016.
9. Malhotra, N. (2012), Marketing Research: An Applied Orientation, 7th ed., Pearson, 2019.
10. Cooper, D.R. and Schindler, P.S. and J. K. Sharma (2018), Business Research Methods, 12th edition, McGraw-Hill Education.

5.20 Self-Assessment Questions

1. Describe the process of hypothesis testing.

2. Describe the null and alternate hypothesis typically represented in the hypothesis testing process.

3. Define the significant level.

4. State the condition under which it is appropriate to use one-talied test and two-tailed test.

5. What do you mean by power of test?

5.21 Answers to Check Your Progress Questions

1. The central limit theorem

2. (d) 1.56

3. (c). 980

4. (a) 36.32, 36.67

5. Population

6. (b) Statistics

7. (e) Highly accurate data is not possible in sampling

8. (e) Judgmental sampling

9. Editing.

10. False and accepted.

11. (d) 0.2

12. (b) -19.76

13. (c) 0.075

Unit 6

Correlation and Linear Regression

Structure

6.1 Introduction

In the previous unit, you learnt different types of sampling methods, inferences and tests of hypothesis. In this unit, you will learn the correlation between two variables and its significance. In business and finance, it may be observed that many variables are related, change in one variable may bring about change in other variables; or change in some variables may bring change in one variable. For example, the change in manufacturing cost of the product will show its effect on selling price of the product, sales, and profitability of the business firm. The study of the relationship in variables is very important for decision-making.

6.2 Objectives

After going through the unit, you should be able to:

- Define correlation between two variables;
- Explain how to Use Scatter diagrams to visualize the relationship between two variables;
- Outline the use of regression analysis to estimate the relationship between two variables;
- Explain Least squares estimating equation;

- Analyze coefficient of determination as a measure of strength of the relationship between two variables;
- Indicate limitations of regression and correlation analyzes and caveats about their use; and
- Explain the use of regression and correlation in finance.

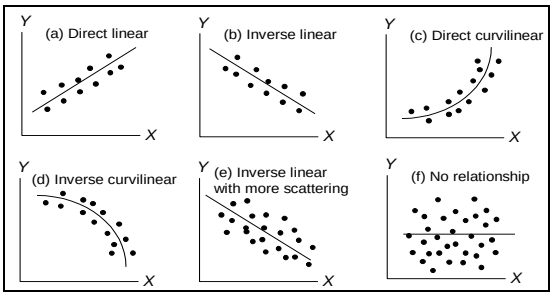
6.3 Correlation

There are many ways to measure the relationship between two or more variables. Two useful methods often used are scatter diagram and correlation.

6.3.1 The Scatter Diagram

The scatter diagram indicates the nature of the potential relationship between the variables in two ways. First, it indicates whether the variables are related and if they are related then the kind of line, or estimating equation that describes this relationship. Possible relationship between two variables X and Y is shown in figure 6.1.

Figure 6.1: Possible Relationships between Two Variables X and Y in Scatter Diagrams



6.3.2 Correlation Analysis

As seen earlier, a scatter diagram shows the types of relationship using the graphical representation. The same result can be arrived at with a single number called correlation coefficient, which measures the degree of **linear relationship** among the variables. The direction of change is indicated by + or – signs; the former refers to the movement in the same direction and the latter in the opposite direction; an absence of correlation is indicated by zero. This coefficient ranges between –1 and 1. Further note that in figure 1 (a) the slope of line is positive and in 1 (e) the slope of line is negative.

6.3.3 Calculating Coefficient of Correlation

Pearson's Coefficient of Correlation is the most widely used method for calculating coefficient of correlation and is expressed as:

$$r = \frac{Cov(X, Y)}{\sigma_X, \sigma_Y}$$

Where, σ_X, σ_Y are the standard deviation of X and Y.

The covariance denoted by Cov (X, Y) of two random variables X and Y is defined by:

$$\frac{\sum(X-\bar{X})(Y-\bar{Y})}{n-1} \text{ or, } r=\frac{\sum(X-\bar{X})(Y-\bar{Y})}{\sqrt{\sum(X-\bar{X})^2\ \sum(Y-\bar{Y})^2}}$$

Sometimes, coefficient of determination is also used to measure the strength of a relationship between two variables. It is explained by r^2 i.e., the square of the correlation coefficient. It explains the extent of the variation of a dependent variable and is expressed by the independent variable. A high value of r^2 shows a good linear relationship between the two variables. If r = 1 and r^2 = 1, it indicates a perfect relationship between the variables. The detailed discussion about coefficient of determination is given in succeeding units.

Example 1

The following data shows the sales and operating costs of a firm over a period of eight years:

Year	2013	2014	2015	2016	2017	2018	2019	2020
Sales (Rs. in lakh)	106	140	144	155	136	150	146	143
Operating costs (Rs. in lakh)	100	114	120	125	105	124	122	118

What is the coefficient of correlation between sales and operating costs of the firm?

Solution

Let the following notations be used:

X : Sales (Rs. in lakh); Y : Operating costs (Rs. in lakh)

A	X	106	140	144	155	136	150	146	143	ΣX = 1,120
B	Y	100	114	120	125	105	124	122	118	ΣY = 928
C	$X-\bar{X}$	–34	0	4	15	–4	10	6	3	
D	$Y-\bar{Y}$	–16	–2	4	9	–11	8	6	2	
E	$(X-\bar{X})^2$	1156	0	16	225	16	100	36	9	$\Sigma(X-\bar{X})^2$=1,558
F	$(Y-\bar{Y})^2$	256	4	16	81	121	64	36	4	$\Sigma(Y-\bar{Y})^2$ = 582
G	C × D	544	0	16	135	44	80	36	6	$\Sigma(X-\bar{X})$ $(Y-\bar{Y})$ = 861

$$\bar{X}=\frac{\sum X}{n}=\frac{1,120}{8}=140\ ;\ \bar{Y}=\frac{\sum Y}{n}=\frac{928}{8}=116$$

$$r=\frac{\sum(X-\bar{X})\,(Y-\bar{Y})}{\sqrt{\sum(X-\bar{X})^2\ \Sigma(Y-\bar{Y})^2}}=\frac{861}{\sqrt{1,558\ x\ 582}}=0.9042$$

Interpretation from the value of r = 0.9042

Positive value of r represents positive correlation between X and Y.

6.4 Rank Correlation

Rank correlation is also a technique used to test the direction and strength of the relationship between two variables. It is used when both the underlying variables are ordinal or when the data are available in the ordinal form irrespective of the type of variable. In other words, it's a tool to show whether any one set of numbers has an effect on another set of numbers.

Spearman's rank correlation coefficient can be calculated using the following formula:

$$R = 1 - \frac{6\sum_{i=1}^{N} D_i^2}{N^3 - N}$$

Where,

N = Number of individuals ranked

D_i = Difference in the ranks of the i^{th} individual.

Example 2

Ten competitors in a singing contest are ranked by three judges in the following order:

1st Judge	1	6	5	10	3	2	4	9	7	8
2nd Judge	3	5	8	4	7	10	2	1	6	9
3rd Judge	6	4	9	8	1	2	3	10	5	7

Use the rank correlation coefficient to determine which pair of judges has the nearest approach to common tastes in vocal or singing.

Solution

In order to find out which pair of judges has the nearest approach to common tastes in singing, we compare Rank Correlation between the judgments of:

(i) 1st judge and 2nd judge; (ii) 2nd judge and 3rd judge; (iii) 1st judge and 3rd judge.

Rank by 1st Judge R_1	Rank by 2nd Judge R_2	Rank by 3rd Judge R_3	$(R_1 - R_2)^2$ D^2	$(R_2 - R_3)^2$ D^2	$(R_1 - R_3)^2$ D^2
1	3	6	4	9	25
6	5	4	1	1	4
5	8	9	9	1	16
10	4	8	36	16	4

3	7	1	16	36	4
2	10	2	64	64	0
4	2	3	4	1	1
9	1	10	64	81	1
7	6	5	1	1	4
8	9	7	1	4	1
N = 10	N = 10	N = 10	$\Sigma D^2 = 200$	$\Sigma D^2 = 214$	$\Sigma D^2 = 60$

Rank correlation between the judgment of 1st and 2nd judges:

$$R = 1 - \frac{6\Sigma D^2}{N^3 - N}$$

$$\therefore R_{(1 \text{ and } 2)} = 1 - \frac{6 \times 200}{10^3 - 10} = 1 - \frac{1{,}200}{990} = 1 - 1.212 = -0.212$$

Rank correlation between the judgments of 2nd and 3rd judges:

$$\therefore R_{(2 \text{ and } 3)} = 1 - \frac{6 \times 214}{10^3 - 10} = 1 - 1.297 = -0.297$$

Rank correlation between judgments of 1st and 3rd judges:

$$R_{(1 \text{ and } 3)} = 1 - \frac{6 \times 60}{10^3 - 10} = 1 - \frac{360}{990} = 0.636.$$

Since coefficient of correlation $R_{(1 \text{ and } 3)} = 0.636$ is maximum in the judgments of the first and third judges, we conclude that 1st judge and 3rd judge have the nearest approach to common tastes in singing.

Check Your Progress - 1

1. A relationship model was developed between two variables X and Y, with X as the independent variable.

   After computations,

   $\sum(X-X^{\wedge})(Y-Y^{\wedge}) = 682$

   $\sum(X-X^{\wedge})^2 = 484$

   $\sum(Y-Y^{\wedge})^2 = 1454$

   The coefficient of correlation between X and Y is

   a. 0.2717
   b. 0.8129
   c. 0.3731
   d. 0.2563
   e. 0.7863.

2. Two judges ranked 10 people in a contestant as given below:

   Judge 1: 6 4 3 1 2 7 9 8 10 5

   Judge 2: 4 1 6 7 5 8 10 9 3 2

   Calculate the coefficient of correlation.

   a. 0.987

   b. 0

   c. 1

   d. 0.5

   e. 0.685

3. Which of the following technique shows the negative relationship between two variables in a diagram?

   a. Correlation

   b. Scatter Diagram

   c. Frequency diagram

   d. Rank correlation

   e. Regression

4. A statistician plotted the relationship between two variables from the collected data, which was shown as a horizontal line in the scatter diagram. Which of the following relationship exists between them?

   a. Direct linear

   b. Inverse linear

   c. Direct curvilinear

   d. Inverse curvilinear

   e. No relationship

5. If the correlation coefficient is above 0.9. The strength of relationship between the variables is _____

6.5 Simple Linear Regression

As a financial analyst one often tries to find out the relationship between the variables to draw conclusions about the future. For example, the financial analyst may be interested in finding how stock price will move with the movement in economic indication like, GDP, Inflation, export, etc. In the earlier section, we have measured only the strength of linear relationship. Using the simple linear regression, we can establish the linear relationship between two variables.

Regression analysis is related with study of the dependent of one variable, the dependent variable, on one or more variables, the independent variable or

explanatory variable, with a view to estimating and/or predicting the (population) mean or average value of the dependent variable in terms of the known or fixed (in repeated sampling) values of the independent variable.

In the two variables X and Y, the variable we are seeking to explain (to find) is called Dependent Variable and the other variable is called Independent Variable.

Consider, a total population of 100 families in a hypothetical community and their weekly income (X) and weekly expenditure (Y) both in dollars. The 100 families are divided into some income groups. The conditional expected value of the weekly expenditure with respect to weekly income i.e. E(Y|X) depends on the given values of the variable X. E (Y| X_i) is a function of X_i or E (Y| X_i) = f (X_i). Now the question is how we define the function f (X_i). Because in real situation we don't have a clear idea about the entire population. We assume that E (Y| X_i) is a linear function of X_i i.e.

$$E(Y|X_i) = f(X_i) = A + B_1X_i$$

The above equation represents the Population Regression Function (PRF), where the A and B_1 are unknown and fixed parameters, also known as the intercept and slope respectively. In regression analysis our objective is to estimate the PRF i.e. estimating the value of unknown parameters (A and B_1) on the basis of observation on Y and X.

If we look into the above example, the next question that comes to our mind is what about the consumption expenditure of an individual family related to a fixed weekly income? The individual family weekly expenditure does not necessarily increase or decrease as the income level increase or decrease. Given an income level X_i, an individual family income expenditure is clustered around the conditional expected value i.e. E (Y| X_i). Thus, the deviation of the individual Y_i from the E (Y| X_i) is represented as

$$E_i = Y_i - E(Y|X_i)$$

$$\text{or, } Y_i = E(Y|X_i) + E_i$$

$$\text{or, } Y_i = A + B_1X_i + E_i$$

E_i is defined as the error term and it's a random component. Our objective is to estimate the PRF on the basis of sample information. Sample Regression Function (SRF) is used to sample regression line.

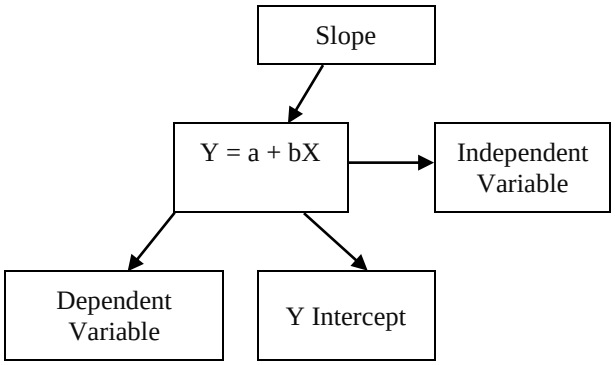
$$\hat{y} = a + b_1x$$

Where, $\hat{y}$ = Estimator of E (Y| X_i), a = Estimator of A, and b_1 = Estimator of B_1

The following regression equation explains that relation:

$$Y = a + bX + \varepsilon_x$$

or



Where, ε_x are random variables with mean 0. b is called slope or regression coefficient of the equation.

For example, the regression equation Y = 10 + 5X, slope is b = 5. a = 10 is the Y intercept or constant term.

6.5.1 Assumptions in Regression

If we consider the population Regression Function Yi = A + B1Xi + Ei. it shows that the value of Yi depends on the value of Xi and Ei. If we don't specific about how Xi and Ei are generated, then we can't make any statistical inference about Yi, A, and B1.

The following assumptions in regression are very important to find out the statistical inferences about the regression parameters:

i. The relationship between the variables X and Y is linear, which implies the formula E(Y|X = x) = a + bx at any given value of X = x.

ii. X values are fixed in repeated sampling. Values taken by the regressor *X* are considered fixed in repeated samples.

iii. Zero mean value of error E_i. Given the value of *X*, the mean or expected value of the error term E_i is zero. E (E_i |X_i) = 0.

iv. Equal variance of error E_i. Given the value of *X*, the variance of the error term E_i is σ^2. Var (E_i |X_i) = σ^2.

v. Given any two *X* values such as X_i and X_j (i and j are not same), the correlation between the corresponding error terms E_i and E_j is zero.

vi. Zero covariance between E_i and X_i.

vii. At each X, the distribution of Y_x is normal, and the variances σ_X^2 are equal. This implies that ε_x 's have the same variance σ^2.

viii. The Y-values are independent of each other.

ix. The independent variable X is not random.

x. The expected value of error term is zero.

xi. No assumption is made regarding the distribution of X.

The above assumptions related to error terms E_i can be stated as $E_i \sim N(0, \sigma^2)$. Also, under the normality assumptions, the error terms are not only correlated but also independently distributed i.e. $E_i \sim NID(0, \sigma^2)$.

Direct and Indirect Relation

If a causal relationship is established, the relationship may be direct or inverse too. In case of direct relationship the slope of the line will be positive, whereas in the case of indirect relationship slope of the line will be negative.

6.5.2 Calculating Regression Constants

The main objective of regression analysis is to find a line that provides the best possible description of the relationship of dependent and independent variables. Alternatively, we can say how the SRF should be constructed so that a is "as close as" possible to the true A and b_1 is "as close as" as possible to true B_1. The method of Ordinary Least square (OLS) is one of the popular methods of finding unknown parameters. According to the Method of Least Squares, the regression should be established in a manner such that when we take sum of squares values of the deviation between predicted values and observed values.

$$\sum e_i^2 = \sum (y_i - \hat{y}_i)^2 = \sum (y_i - a - b_1 x)^2$$

Differentiate the above equation with respect to a and b_1, we get the two normal equations (showing below).

$$na + b\sum x = \sum y \; ; \; a\sum x + b\sum x^2 = \sum xy$$

We can obtain the following formulae to calculate the coefficients 'a' and 'b' in the regression line: $\hat{Y} = a + bX$

$$b = \frac{\Sigma XY - n\overline{X}\,\overline{Y}}{\Sigma X^2 - n\overline{X}^2} = \frac{Cov(X,Y)}{V(X)} \text{ and } a = \overline{Y} - b\overline{X}$$

Regression coefficient can also be found with the help of correlation coefficient. The relation between two is summarized below:

Regression coefficient of Y on X: $b_{yx} = r\frac{\sigma_y}{\sigma_x}$

Regression coefficient of X on Y: $b_{xy} = r\frac{\sigma_x}{\sigma_y}$

Where,

'r' is the coefficient of correlation between X and Y

σ_x is the population standard deviation of X

σ_y is the population standard deviation of Y.

The coefficient of correlation can also be expressed as geometric mean of the two regression coefficients.

$$r = \sqrt{b_{yx} \times b_{xy}}$$

6.5.3. Statistical Property of the Estimators a and b_1

a. Both a and b_1 are the unbiased estimators of A and B_1. It means that E(a) = A and E(b_1) = B_1.

b. Both a and b_1 can be represented as the linear combination of y_i.

c. The variance of a and b_1 are as followed

$$Var(b_1) = \frac{\sigma^2}{\sum(x_i - \bar{x})^2}$$

$$Var(a) = \frac{\sum x_i^2}{n\sum(x_i - \bar{x})^2}\sigma^2$$

Example 3

To formulate the relationship between the quantity that can be sold and the price for one of its new milk products, the marketing manager of a company has collected the following information.

Price (Rs.)	Quantity that can be sold ('000 units)
28.0	130
26.0	145
25.0	170
24.0	190
21.0	215
19.0	225

i. To formulate the relationship between the price and the quantity that can be sold with the help of the least square method the following table will be useful.

ii. Use this relationship to predict the quantity that can be sold if the price is Rs.22.

							Total
X	28	26	25	24	21	19	143
Y	130	145	170	190	215	225	1,075
XY	3,640	3,770	4,250	4,560	4,515	4,275	25,010
X^2	784	676	625	576	441	361	3,463

In order to determine the regression line for the plot of main exam points vs model exam points, we need to calculate the coefficients a and b.

$$\sum XY = 25{,}010;\ \sum X^2 = 3{,}463;\ \sum X = 143;\ \sum Y = 1{,}075$$

$$\bar{X} = \frac{\Sigma X}{n} = \frac{143}{6} = 23.83\ ;\ \bar{Y} = \frac{\Sigma Y}{n} = \frac{1{,}075}{6} = 179.17$$

$$b = \frac{25{,}010 - 6(23.83)(179.17)}{3{,}463 - 6(23.83)(23.83)} = -10.89$$

$$a = 179.17 - (-10.89)(23.83) = 438.68$$

The relationship between price (X) and quantity sold (Y) is:

$$Y = 438.68 - 10.89X$$

From the slope of this regression (–10.89) it can be stated that for every unit increase in price, it is estimated that the average demand will decrease by 10.89 points.

Suppose we want to use this relationship to predict the quantity that can be sold. If the price is Rs.22, it becomes

For price (X) = 22, quantity sold (Y) = 438.68 – 10.89 x 22 = 199.10

Since the quantity is given in 1000s $\rightarrow$ actual quantity is 199,100 units.

6.6 Coefficient of Determination

Coefficient of determination (R^2) is measured by the formula:

$$R^2 = \frac{\sum\left(\hat{Y} - \bar{Y}\right)^2}{\sum\left(Y - \bar{Y}\right)^2}$$

The numerator $\sum\left(\hat{Y} - \bar{Y}\right)^2$ is termed as **total variation** in Y and the denominator is called variation explained by the model. In other words, we can say that the *coefficient of determination* is the ratio of the explained variation to the total variation.

The *coefficient of determination* is a measure of how well the regression line represents the data. If the regression line passes exactly through every point on the scatter plot, it would be able to explain all of the variation.

The coefficient of determination can be written in terms of correlation also.

$\sqrt{R^2}$ = r, the coefficient of correlation measures the strength of the linear association between x and y.

Therefore $0 \le R^2 \le 1$, and greater the value of R^2, better it fits the data on line.

6.6.1 Regression Sum of Square and Error Sum of Square

Regression Sum of Square (RSS) is the portion of total sum square that is explained by the regression line; and ESS is the portion of sum square that is not explained by the use of the regression model.

RSS (Regression Sum of Squares) = $\sum \left(\hat{Y} - \overline{Y}\right)^2$

ESS (Error Sum of Squares) = $\sum \left(Y - \hat{Y}\right)^2$

We can show that TSS = RSS + ESS

TSS (Total Sum of Squares) = $\sum \left(Y - \overline{Y}\right)^2$

And coefficient of determination = $\frac{RSS}{TSS}$

Coefficient of determination is also calculated by $R^2 = \frac{a\sum Y + b\sum XY - n\overline{Y}^2}{\sum Y^2 - n\overline{Y}^2}$

6.7 The Standard Error

Standard error measures the reliability of the linear equation. It tells about the width of the error and hence the value of Y spread for given X. It's a measure of dispersion like standard deviation and measures the variability or scatter of the observed values around the regression line.

The **standard error of the estimate** for a regression equation is given by

$$s_e = \frac{\sqrt{\sum \left(Y - \hat{Y}\right)^2}}{n-2}$$

Where,

Y = Values of the dependent variable

$\hat{Y}$ = Estimated values from the estimating equation that corresponds to each Y value

n = Number of data points used to fit the regression line.

In the above equation, you can observe that the sum of the squared deviations is divided by n – 2 and not by n. This is because we have lost 2 degrees of freedom in estimating the regression line. We used the sample to compute a and b.

6.7.1 Shortcut Method to Compute Standard Error (s_e)

Equation for calculating standard error of estimate is also given by the formula:

$$s_e = \frac{\sqrt{\sum Y^2 - a\sum Y - b\sum XY}}{n-2}$$

Where,

X = Values of the independent variable

Y = Values of the dependent variable

a = Y-intercept of the estimating equation

b = Slope of the estimating equation

n = Number of data points.

6.8 Making Inference about Population Parameter

The sample regression line can be used as a true estimator of population regression line. Let the population regression line be Y = A + BX and sample regression equation be Y = a + bx. Now we can make inferences about the slope B of the true regression line, with the given sample regression line slope b.

6.8.1 Prediction Interval

Standard error can be used to predict interval around $\hat{Y}$ which would contain the actual Y.

For a large sample where $n \geq 30$,

$\hat{Y} \pm Zs_e$ would be the interval, where Z is the appropriate Standard Normal Value.

For a small sample where $n < 30$,

$\hat{Y} \pm ts_e$ would be the interval, where t is the appropriate t value (with n – 2 degrees of freedom).

6.8.2 Accepting the Model

We know b is an estimator of B. The random variable b is a normal distribution with mean B. Now, we can calculate the test statistics t with n – 2 degrees of freedom using the formula given below:

$$t = \frac{b - B_{H_0}}{s_b} \quad \ldots (1)$$

Where, standard error of regression coefficient (s_b) can be calculated using the following formula:

$$s_b = \frac{s_e}{\sqrt{\sum X^2 - n\overline{X}^2}} \quad \ldots (2)$$

Remember that $s_e = \sqrt{\frac{\sum Y^2 - a\sum Y - b\sum XY}{n-2}}$

Finally, we can find the value of t for the given significant level and for n – 2 degree of freedom and compare with the calculated t statistics. In equation (2), if t statistics falls within the acceptance region, we can accept the null hypothesis; and if t statistics fall within the rejection region, we reject the null hypothesis.

Example 4

Consider our previous example of sample linear equation Y = 60 + 5X.

Suppose we want to check whether the computed model is spurious or not at 5% significant level, we can set up the hypothesis as given here.

H_0: B = 0 (There is no linear relationship);

H_1: B > 0 (There is a positive relationship)

$$s_e = 13.83;\ s_b = \frac{13.83}{\sqrt{2528 - 10x\ 14\ x\ 14}} = \frac{13.83}{23.83} = 0.58$$

Therefore, $t = \frac{b - B}{0.58} = \frac{5 - 0}{0.58} = 8.62$

At α = 5%, the rejection region is t > 2.132

But the calculated t value = 8.62; As 8.62 $\geq$ 2.132

Therefore we reject H_0; that means there is a linear relationship.

6.8.3. F Test

An F-test can also be used to test for significance in regression. In linear regression, we have only one independent variable. Thus, the F-test will provide the same result as the t-test. The reason behind the use of the F test for determining whether the regression relationship is statistically significant and is based on the development of two independent estimates of σ^2. The next question is to perform the F-test, for which we need to find out the F statistic. To find the F statistic, we require the concept of mean square. The mean square can be evaluated dividing the sum of square by the corresponding degree of freedom. In simple linear regression, the variation occurred by the regression line has degree of freedom equals to the number of independent variables i.e. 1, whereas the degree of freedom of the variation due to error is equal to (n-2). If the null hypothesis is true, MSR and MSE are two independent estimates of σ^2 and the sampling distribution of MSR/MSE follows an F distribution with numerator df equal to one and denominator df equal to n-2.

The summary of F-test is as follows:

$$H_0: B_1 = 0$$

$$H_a: B_1 \neq 0$$

Test statistics: F=MSR/MSE

Rejection Rule: Reject H_0 if p value $\leq \alpha$

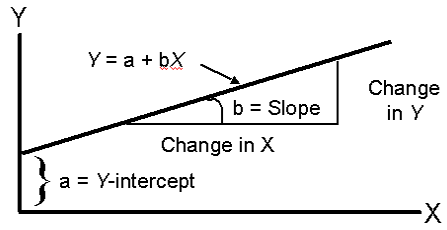
Reject H_0 if $F_{cal} \geq F_\alpha$

F_α is based on F-distribution with 1 degree of freedom in the numerator and (n-2) degrees of freedom in the denominator.

6.9 Geometry of Regression

In this section, we will interpret the shape of regression line for different values of a and b. As discussed earlier, a is the intercept and b is the slope of the line.

Figure 6.2: Linear Equations



6.9.1 Slope of Regression Line

Slope is basically the inclination of a line and it represents the change in the dependent variable for every unit change in the independent variable. Suppose there are two points on the straight line (X_1, Y_1) and (X_2, Y_2), then slope of the line will be

$$b = \frac{Y_2 - Y_1}{X_2 - X_1}$$

We have seen that the slope of a line may be positive, negative, zero or undefined. The sign of the slope indicates whether it is rising or falling. The magnitude (absolute value) of the slope indicates the relative steepness of the line.

6.10 Misuses and Caveats

Misuses

- Estimating line can be applied beyond the range of data (extrapolation). But estimating equation is valid only over the same range as the one from which the sample was taken initially.
- Usually regression and correlation are used to determine a cause and effect relationship, which is another mistake, because regression and correlation do not determine the cause and effect. In other words, if there is correlation between X and Y it does not mean one causes the other.
- Sometimes the regression line fails to recognize some variables that can be dependent on time.
- Misrepresenting r^2 and r values is another mistake. For example, r = 0.4 does not mean that 40% variation is explained by the line. But it means that only 16% of variation is explained by the line, because the value of $r^2 = 0.16$.
- Regression analysis mistakenly calculates relationships when they do not exist.

Caveats

- Before using the regression and correlation know the limitations of the technique.
- Use common intelligence while applying the regression and correlation.

Check Your Progress - 2

6. Two variables X and Y have the regression relationship $\hat{Y} = 5 + 4X$. Then which of the following is correct?
   a. For a given value of Y = 22, the estimated value of X is 4
   b. For a given value of Y = 25, the estimated value of X is 5
   c. The slope of the line is 5
   d. The intercept of the above line is 4
   e. For a given value of Y = 20, the estimated value of X is 2
7. If the variances of X and Y are 2.3 and 4.1, and the covariance between them is 2.1 then the coefficient of correlation between X and Y is
   a. 0.098
   b. 0.222
   c. 0.408
   d. 0.684
   e. 0.748.
8. The slope and the intercept on Y-axis for a simple regression line are 3 and 5 respectively. If the independent variable has a value of 5 then the estimated value of the dependent variable is
   a. 3
   b. 5
   c. 20
   d. 27
   e. 28.

6.11 Applications in Finance

Correlation and regression have various applications in finance as discussed below.

6.11.1 Risk of a Portfolio

Correlation is used to calculate risk of portfolio. Suppose an investor has a portfolio consisting of two stocks A and B whose return and risk are as follows:

		Value (Rs. Thousand)	Return	Standard Deviation
Security	A	30	30%	1.5
Security	B	70	20%	2.5
Portfolio	Value	100		

The return on the portfolio of the above securities will be the weighted average return.

$R_p = W_1R_1 + W_2R_2$

i.e., return of the portfolio will be 0.3 x 30 + 0.7 x 0.20 = 22 %

But the risk of portfolio will not necessarily be the same as weighted average risk of securities, because the risk of the portfolio is dependant on the coefficient of correlation between them. For –1 correlation coefficient, the two returns move exactly opposite to each other and the portfolio risk will be minimum and for +1 correlation coefficient, return of two securities will move in the same direction, so the risk of portfolio would be maximum. For a 2 asset portfolio, the standard deviation σ_p is given by

$$\sigma_p = \sqrt{W_1^2 \sigma_1^2 + W_2^2 \sigma_2^2 + 2W_1 W_2 \rho_{1,2} \sigma_1 \sigma_2}$$

Where,

W_1 and W_2 are the weights for the assets in the portfolio.

$\rho_{1,2}$ is the coefficient of correlation between the returns of the two assets.

σ_1 and σ_2 are the standard deviations of the returns of the two assets respectively.

Note: $\rho_{1,2}\sigma_1\sigma_2$ = Covariance between returns of the two assets.

6.11.2 Characteristic Line

Financial analysts often talk about beta of share which measures the sensitivity of the stock price with the market index. Bombay Stock Exchange (BSE), Sensitive Index (Sensex), BSE National Index, NSE (nifty) are some popular market indices. As they represent the movement of the whole market, they are called as market index. Beta equals to 1.2 means if the market index moves by 10%, the securities of beta 1.2 will move by 10 x 1.2 or 12%. Beta of securities more than one represents aggressive securities and beta of less than one represents defensive securities.

Beta is used to measure the Securities.

The total risk of a security is measured in terms of the variance (σ^2) [or standard deviation (σ)] of its returns. Total risk comprises both systematic and unsystematic components. How do we split the total risk, namely, the variance, into the systematic and unsystematic risk components? This is relatively simple.

Total Risk of a Security i = σ_i^2 ; Systematic Risk of a Security i = $\beta_i^2\sigma_m^2$

Where, β_i is the beta of security i and σ_m^2 is the variance of the market portfolio.

But, $\beta_i = \frac{\sigma_{i,m}}{\sigma_m^2}$; Substituting for β_i in above equation we have

Systematic Risk of Security i = $\frac{\sigma_{i,m}^2}{\sigma_m^2} = \frac{\rho_{im}^2.\sigma_i^2.\sigma_m^2}{\sigma_m^2} = \rho_{i,m}^2\,\sigma_i^2$;

since $\rho_{i\,m} = \frac{\sigma_{i,m}}{\sigma_i\sigma_m}$

where $\rho_{i,m}$ is the correlation coefficient between the security return and the market return.

Therefore, Unsystematic Risk of a Security I = $\sigma_i^2 - \beta_i^2\sigma_m^2 = \sigma_i^2\ (1 - \rho_{im}^2)$.

In terms of percentage of total risk (security variance):

Systematic Risk = $(\rho_{im}^2\,\sigma_i^2/\sigma_i^2)$ x 100 = $100\,\rho_{im}^2$

Unsystematic Risk = $(\sigma_i^2\,(1 - \rho_{im}^2)/\sigma_i^2)$ x 100 = 100 $(1 - \rho_{im}^2)$.

6.11.3 Cost-Volume-Profit Analysis

The Cost-Volume-Profit (CVP) analysis is used widely in management and cost account to determine the relationship governing sales, profit and cost. One of the basic assumptions in CVP analysis is that total cost can be segregated into fixed, variable cost and semi-variable cost. Simple linear regression is commonly used to segregate the fixed and variable components of semi-fixed or semi-variable costs.

6.11.4 Time Series and Demand Forecasting

Forecasting of future demand and sales is an important aspect of budgeting. Many techniques are used to forecast the future demand and sales. In the trend projection method, the time series of demand for a product over the past years is studied and the trend observed is projected into the future. A linear relationship (if it exists) is normally established using historical data.

6.11.5 Tests in Investments

Investors and analysts question if the price of stock is dependent on historical price, because if the market is not efficiently working then the investor having

more information can earn higher profit than the investor who has less information. Efficient Market Hypothesis holds that security prices fully reflect all available information any time. Statistical tests like auto correlation test and run tests have been applied to examine whether the market is efficient or not, and if it is efficient, then what is the level of information reflected in the market price of share. Based on the level of information market can be in weak efficiency, semi-strong efficiency or strong efficiency form. **Weak** form of efficiency says that current prices fully reflect all historical information. According to the semi-strong form, current stock prices reflect all publicly available information such as earnings, stock and cash dividends, splits, mergers and takeovers, interest rate changes, etc. According to the strong form, prices of securities fully reflect all available information – both public and private run tests and serial correlation test are used to test the weak form of efficiency.

6.11.6 Serial Correlation Tests

Serial correlation has been used to find the correlation coefficient between current changes and past changes in a share's price. The test selects a certain number of stocks. The changes in the prices of these stocks are observed for a particular period. The same is done for another period. Correlation analysis is conducted on these changes. If the correlation between these changes is near or equal to zero, it implies that the price changes are independent of each other.

6.11.7 Run Tests

A run is defined as "a sequence of identical occurrences preceded and followed by different occurrences or by none at all." For testing the randomness of share prices, we take a series of stock prices. Starting with the first price, each price change is denoted by a plus (+) or a minus (–) sign. Plus (+) sign indicates that the price under consideration has increased compared to its preceding price and a minus (–) sign indicates that the price under consideration has decreased compared to the preceding price. In case, the price under consideration is same as its preceding price, we indicate it by a zero. In case, the sign has changed from plus to minus or from minus to plus, a new run is counted to have begun. To test the independence of the prices, we require:

Total number of runs : r

Number of positive price changes : n_1

Number of negative price changes : n_2

Once we have this data, the mean and the standard deviation of the mean, are calculated by using the formulae given below:

$$\mu_r = \frac{2n_1n_2}{n_1+n_2}+1 \text{ and } \sigma_r = \sqrt{\frac{2n_1n_2(2n_1n_2-n_1-n_2)}{(n_1+n_2)^2(n_1+n_2-1)}}$$

At a given level of significance, we calculate the upper and lower limits and check whether the number of runs observed from the test falls within the limits or not. If it is between the limits, we conclude that the prices are random or independent of each other.

6.12 Regression using Microsoft-Excel

A company is in the process of deciding whether to purchase a maintenance contract for its new computer wheel alignment and balancing machine. Managers feel that maintenance expense should be related to usage, and they collected the following information on weekly usage (hours) and annual maintenance expense (in hundreds of dollars).

Weekly Usage (Hours)	Annual Maintenance Expense
13	17
10	22
20	30
28	37
32	47
17	30
24	32.5
31	39
40	51.5
38	40

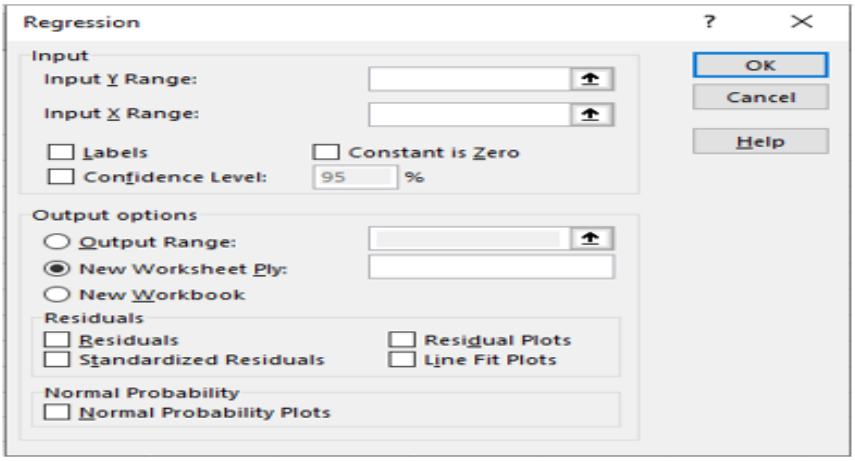
How to perform the regression analysis of the above example in EXCEL is explained below.

Step1: Go to **Data** tab in Excel.

Step 2: Under the Data tab, go to **Data Analysis** toolbox.

Step 3: Select the **Regression** from the Data Analysis Toolbox.

Step 4: After that a new window will open.



Step 5: Enter A1:A11 in the input X range

Step 6: Enter B1: B11 in the output Y range

Step 7: Select the labels

Step 8: Select the confidence level and make it 95%

Step 9: Select the output range and choose any cell reference in the box. In this case, we choose F3 cell

Step 10: Click OK

Weekly Usage (Hours)	Annual Maintenance Expense
13	17
10	22
20	30
28	37
32	47
17	30
24	32.5
31	39
40	51.5
38	40

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.927537
R Square	0.860326
Adjusted R Squa	0.842867
Standard Error	4.196982
Observations	10

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	867.9827	867.9827	49.27615	0.000110448
Residual	8	140.9173	17.61466		
Total	9	1008.9			

	Coefficients	*ndard Err*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*ower 95.0%*	*pper 95.0%*
Intercept	10.36698	3.698483	2.803036	0.02309	1.838262695	18.8957	1.838263	18.8957
Weekly Usage (h	0.957827	0.136448	7.019697	0.00011	0.643176162	1.272478	0.643176	1.272478

From the Excel Output, we can observe that the regression equation is $\hat{y}$ = 10.3669 + 0.9578 x. The value of coefficient of determination is 0.8603 i.e. 86.03 percent of the variability in the dependent variable can be explained by the linear relationship between annual maintenance and weekly usage. In the Anova table, the p value is 0.00011 which is less than α. So, we can reject the null hypothesis and conclude that the relationship between annual maintenance and weekly usage is significant.

Check Your Progress - 3

9. The coefficient of determination is _____of the coefficient of correlation

10. The shares advisory consultant was analyzing the volatility or risk of a portfolio (Y) compared to the market as a whole (X). Which will help him measure this value better?

    a. Rank correlation
    b. Coefficient of correlation
    c. Beta
    d. Standard deviation
    e. Standard error

6.13 Summary

- Correlation and regression are the statistical measures used to describe the nature and strength of the relationship between two or more variables. While correlation analysis determines the degree to which the variables are related, regression analysis develops the relationship between the variables.
- The coefficient of determination r^2 is used to analyze how good the fit is and gives the variation explained by the model. Regression and correlation methods have wide application in finance. They are used to measure risk in portfolio, establishing characteristic lines and measuring cost volume relationship.

6.14 Glossary

Coefficient of Correlation: It indicates the direction of the relationship between two variables, direct or inverse. It is also measured as the square root of the coefficient of determination.

Coefficient of Determination: A measure of the proportion of variation in Y, the dependent variable that is explained by the regression line, that is, by Y's relationship with the independent variable.

Correlation Analysis: A technique used to determine the degree of linear relationship between two variables.

Dependent Variable: The variable be predicted in the regression analysis.

Direct Relationship: If the value of the independent variable increases, so does the value of the dependent variable, and this relationship is known as direct relationship.

Independent Variables: The known variable, or variables, in regression analysis.

Inverse Relationship: A linear relationship between two variables in which, as the independent variable increases, the dependent variable decreases.

Regression Line: A line fitted to a set of data points to estimate the relationship between two variables.

Regression: The general process of predicting one variable from another by statistical means, using previous data.

Scatter Diagram: A graph of points on a rectangular grid; the X and Y co-ordinates of each point correspond to the two measurements made on some particular sample element, and the pattern of points illustrates the relationship between the two variables.

Standard Error: The standard deviation of the sampling distribution of a statistic.

6.15 Suggested Readings/Reference Material

1. Gupta, S. P. Statistical Methods. 46th Revised ed. New Delhi: Sultan Chand & Sons. 2021
2. I. Levin Richard, H. Siddiqui Masood, S. Rubin David, Rastogi Sanjay. Statistics for Management. Pearson Education; Eighth edition, 2017
3. Gerald Keller. Statistics for Management and Economics. Cengage, 2017.
4. Arora, P. N., and Arora, S. CA Foundation Course Statistics. 6th ed. S Chand Publishing, 2018.
5. Mario F Triola. Elementary Statistics. 13th ed., 2018.
6. David R. Anderson, Dennis J. Sweeney, Thomas A. Williams, Jeffrey D. Camm, James J. Cochran. Statistics for Business and Economics. 13th Edition, Cengage Learning India Pvt. Ltd., 2019.
7. S D Sharma. Operations Research. Kedar Nath Ram Nath, 2018.
8. Hamdy A. Taha. Operations Research: An Introduction. 10th ed., Pearson, 2016.
9. Malhotra, N. (2012), Marketing Research: An Applied Orientation, 7th ed., Pearson, 2019.
10. Cooper, D.R. and Schindler, P.S. and J. K. Sharma (2018), Business Research Methods, 12th edition, McGraw-Hill Education.

6.16 Self-Assessment Questions

1. Correlation and Regression are not same. Explain.

2. What are the merits and demerits of correlation?

3. Define direct and inverse relationships.

4. Explain the reason for construction scatter diagram.

5. What do you mean by curvilinear relationship? Explain the difference between liner and curvilinear relationship.

6.17 Answers to Check Your Progress Questions

1. (b) 8129

(682/sqrt(484*454)

2. (a) 987

3. (b) Scatter Diagram

4. (e) No relationship- A horizontal line in the scatter diagram indicates no relationship between the variables since dependent variable becomes constant for any value of independent variable.

5. Strong/very strong as the value is very nearer to ONE

6. (b) For a given value of Y = 25, the estimated value of X is 5

7. (b) 0.222

8. (c) 20

9. The coefficient of determination is ***square*** of the coefficient of correlation.

10. (c) Beta

It measures the volatility or risk of a portfolio compared to the market value. While the securities with beta value below 1 show less volatility but provide lower, but more stable returns.

Quantitative Methods

Course Structure

Block I: Introduction to Statistics and Probability	
1.	Arranging Data
2.	Central Tendency and Dispersion
3.	Probability
4.	Probability Distribution and Decision Theory
Block II: Statistical Relations and Hypothesis Testing	
5	Statistical Inference and Hypothesis Testing
6	Correlation and Linear Regression
Block III: Statistical Regression and Quality Control	
7	Multiple Regression
8	Time Series Analysis
9	Quality Control
Block V: Statistical Distributions, Variations and IT	
10	Chi-Square Test and Analysis of Variance
11	Role of IT in Modern Business Enterprise
12	Statistical Software Tools
Block V: Advanced Statistics	
13	Index Numbers
14	Simulation
15	Linear Programming
Block VI: Business Research	
16	Introduction to Business Research Methods
17	Questionnaire Design
18	Report Writing