SAMPLING AND SAMPLING DISTRIBUTIONS

INTRODUCTION

Sampling in statistics is a common and important as salt is in food. In homes, ladies take out one teaspoonful to detect the quality what she is cooking. In medical sciences, a few drops of blood are taken and tested microscopically or chemically to know whether the blood contains some abnormalities or not.

Nowadays, sampling methods are extensively used in socio-economic surveys to know the living condition, cost of living index etc. of a class of people. In biological studies, experiments are conducted on some units (persons, animals or plants) and inferences are drawn about the breed or variety to which the units belong. In the industries sampling procedures are predominantly used for quality control.

Sampling theory is the study of relationships existing between a population and samples drawn from the population.

SOME CONCEPTS ASSOCIATED WITH SAMPLING

Sampling: - May be defined as the selection of some parts of an aggregate or totality on the basis of which a judgment or inference about the aggregate or totality is made.

Statistic: - Statistical measurable value of the sample or a measurable characteristic value of the sample.

Parameter: - A measurable value of the population or a measurable characteristic value of the population. It is a population result.

Sampling design: - A sample design is a definite plan for obtaining a sample from the sampling frame. It refers to the technique or the procedure that one would adopt in selecting some sampling units from which inferences about the population is drawn. Sampling design is determined before any data are collected. Sampling techniques are divided in to two: *Random Sampling* and non-*Random Sampling*. These designs have already been explained in the previous course.

Sampling error: - Sample surveys do imply the study of a small portion of the population and as such there would naturally be a certain amount of inaccuracy in the information collected. This in accuracy may be termed as sampling error or error variance. The discrepancies between population parameters and estimates (statistics), which are derived from a random sample is also the error or the sampling bias. In short, sampling error is the difference between a sample statistic and its corresponding population parameter.

IMPORTANT PROBABILITY DISTRIBUTIONS

In the previous course Stat 192, you are familiar with the definition of probability distributions, random variables, expectations, some discrete and continuous probability distributions-for a matter of reminding you the concepts:

The probability distribution of a random variable is a listing of the values that the random variable can take on together with the corresponding probabilities. A probability distribution shows the expected outcomes of an experiment and the probability of each of these outcomes. Probability distribution is a listing of all the outcomes of an experiment and the probability associated with each outcome. It can be presented in the form of a table, a graph or a formula.

Depending on the nature of the random variables from which the probabilities are generated, probability distribution is divided in to two. Discrete probability distribution and continuous probability distribution. A probability distribution generated by the use of discrete random variables leads to the formulation of *Discrete Probability Distributions*. The variable in a discrete probability distribution can take only certain values, usually integers. It is most often the result of counting or enumeration.

A probability distribution generated by the use of continuous random variables leads to the formulation of *Continuous Probability Distribution*. The variable in a continuous probability

distribution can take any value within a given range. A continuous probability distribution is usually the result of measurement.

Some of important probability distributions, which are commonly used are: Binomial probability distribution, Poisson probability distribution and the normal probability distribution. The first two are discrete probability distribution and the last one is a continuous probability distribution. These are some of the well-established probability distributions, which have a wide variety of application to problems often encountered.

Binomial Distribution

It is one of the most popular discrete distributions. The origin of binomial distribution lies in Bernoulli's trials. A Bernoulli's trial is an experiment having only two possible outcomes, that is, success or failure. In other words, the results of the trial are always dichotomous. For example, if we toss a coin, it will show either head or tail on the upper face.

Since Bernoulli's trials has been considered, certain conditions for the application of binomial distributions are obvious. For clarity,

i) The probability of a success (or failure) remains same in each trial,

ii) Trials must be independent,

iii) Number of trials must be finite,

iv) The probabilities of success and failure is one,

v) The probability of success vis-à-vis failure is not very low. A dichotomous variable X, which has the probability function,

$$P_\chi(\chi) = \binom{n}{x} p^x q^{n-x} \text{ for } x = 0, 1, 2, \ldots\ldots\ldots n \quad (1.1)$$

$$= 0 \qquad \text{otherwise}$$

is said to have binomial distribution. In the binomial function (1·1), n = number of trials; p = probability of a success; q = probability of failure and $p_\chi(\chi)$ = probability of getting exactly x

success in n trials. The distribution given by (1·1) is called binomial distribution since it is the (x + 1)th term in the binomial expansion of $(p+q)^n$.

Binomial distribution has two parameters, n and p(or q). and the mean of the binomial distribution is **np** and variance is **npq**. p + q = 1.

The binomial distribution tends to normal distribution as n increases. The normal approximation is correct enough if the mean np is greater than 15 for p = ½

Example 1.1

Consider a simple trial of tossing a perfectly round and balanced coins six times. Then the probability of getting

i) E_1: exactly three heads,

ii) E_2: at least three heads and

iii) E_3: not more than two heads, can be calculated by binomial distribution as follows:

Solution: For the given example, n = 6, p = q = ½

i) $$P(E_1) = \binom{6}{3}\left(\frac{1}{2}\right)^3\left(\frac{1}{2}\right)^{6-3} = \frac{6!}{3!(6-3)!}\cdot\frac{1}{2^6} = \frac{5}{16}$$

ii) $$P(E_2) = \sum_{x=3}^{6} P_x(x)$$

$$= \binom{6}{3}\left(\frac{1}{2}\right)^3\left(\frac{1}{2}\right)^{6-3} + \binom{6}{4}\left(\frac{1}{2}\right)^4\left(\frac{1}{2}\right)^{6-4} + \binom{6}{5}\left(\frac{1}{2}\right)^5\left(\frac{1}{2}\right)^{6-5} + \binom{6}{6}\left(\frac{1}{2}\right)^6$$

$$= \frac{1}{2^6}\left\{\binom{6}{3} + \binom{6}{4} + \binom{6}{5} + \binom{6}{6}\right\}$$

$$= \frac{1}{64}\left\{\frac{6\times5\times4}{3\times2\times1} + \frac{6\times5}{2\times1} + 6 + 1\right\} = \underline{\underline{\frac{21}{32}}}$$

iii) $P(E_3) = \sum_{x=0}^{2} P_x(x)$

$$= \binom{6}{0}\left(\frac{1}{2}\right)^0\left(\frac{1}{2}\right)^6 + \binom{6}{1}\left(\frac{1}{2}\right)\left(\frac{1}{2}\right)^{6-1} + \binom{6}{2}\left(\frac{1}{2}\right)^2\left(\frac{1}{2}\right)^{6-2}$$

$$= \frac{1}{2^6}\left\{\binom{6}{0} + \binom{6}{1} + \binom{6}{2}\right\} = \underline{\underline{\frac{11}{32}}}$$

Example 1.2

Let the probability of an item, to be defective, produced by a factory be .10. A sample of 10 items has been inspected. Then the probability of an event E that the sample has two defective items is,

$$P(E) = \binom{10}{2}(.1)^2(.9)^{10-2} = \frac{10\times 9}{2\times 1}.\frac{1}{10^2}\left(\frac{9}{10}\right)^8 = \frac{1}{2}\left(\frac{9}{10}\right)^9$$

Given the probability of x successes in binomial distributions, one can find the probability of x + 1 successes using the formula,

$P(X = x+1) = P(X = x) \cdot \left(\frac{n-x}{x+1}\right).\frac{p}{q}$ (1.2) Relation (1.2) helps in calculating the term by terms probabilities.

Poisson Distribution

Before we know the distribution, it becomes necessary to understand what is a Poisson random variable. A variable, which can take only one discrete value in an interval of time, howsoever small, is known as Poisson variables. Some of the well-known examples of Poisson variable are:

i) Number of mistakes in a typed page;

ii) Number of cars parked at a place in an hour, say between 10:00 A.M and 11:00A.M;

iii) Number of suicides in a certain period in a city or town etc.

If we consider a Poisson's process for a unit length of interval (time, length, space etc), the number of occurrences are a random variable which follow Poisson distribution. It has been named after

its inventor, Simeon D. Poisson, a French probabilist of nineteenth century. It is one of the most important discrete distributions. Poisson distribution is a classical approximation to binomial distribution, in which case n, the number of trials, is comparatively large and p, the probability of an occurrence, is small.

Let X be the number of occurrences in a Poisson process and μ be the actual average number of occurrences of an event in a unit length of interval, the probability function for Poisson distribution is,

$$P_x(x) = \frac{e^{-\mu}\mu^x}{\chi!} \text{ for } x = 0, 1, 2, \ldots\ldots.(1.3)$$

$$= 0 \text{ otherwise}$$

If we consider the length of interval as d of unit length, the average number of occurrences in d length of interval is μ d. Thus, the probability function in this situation is

$$P_x(x) = \frac{e^{-\mu d}(\mu d)^x}{\chi!} \text{ for } x = 0, 1, 2, \ldots\ldots(1.4)$$

$$= 0 \text{ otherwise}$$

The Poisson distribution as given by (1.3) has mean μ and its variance is also μ. It is the only distribution so far, of which the mean and variance are equal. Poisson distribution possesses only one parameter (μ).

Note:

1) It has been said that for Poisson viriate the number of trials n is large and the probability of the so-called success is small. Now the question arises what value of n is to be considered as large and what value of p as small. There is no hard-and-fast rule for this, but as a tradition if $n \geq 20$, it may be taken as large and p = 0.05 may be taken as small. Otherwise, the discrete variable with only two possibilities is generally taken to follow binomial distribution.

2) The value of $e^{-\mu}$ should either be seen from a table or may be calculated with the help of logarithm.

Example 1.3

The number of mistakes counted in one hundred typed pages of a typist revealed that she made 2.8 mistakes on an average per page. The probability that in a page typed by her,

i) There is no mistake

ii) There are two or less mistakes, can be calculated as under,

Solution: Given that $\mu = 2.8$

i) The probability, $p(x = 0) = \frac{e^{-2.8}(2.8)^0}{0!} = 0.061$

ii) The probability, $p(x \leq 2) = \sum_{x=0}^{2} \frac{e^{-2.8}(2.8)^x}{\chi!}$

$$= e^{-2.8}\left\{\frac{(2.8)^0}{0!} + \frac{(2.8)^1}{1!} + \frac{(2.8)^2}{2!}\right\}$$

$P(x \leq 2) = 0.471$

For Poisson vitiate x, the relationship between the probabilities, P(X = x) and P(X = x+1) is given by

$P(X = x) = \frac{e^{-\mu}.\mu^x}{\chi!}$ and

$P(X = x + 1) = \frac{e^{-\mu}.\mu^{(x+1)}}{(x+1)!} = \frac{e^{-\mu}.\mu^x}{\chi!}.\frac{\mu}{\chi+1}$

Thus $P(X = x + 1) = P(X = x) . \frac{\mu}{(\chi+1)}$ …….(1.5)

Example 1.4

If a Poisson random variable X is such that P(X = 1) = P(X = 2), then the probability of P(X = 3) can be found as:

Using (1.5) above,

$P(x = 2) = p(x = 1) . \frac{\mu}{2}$

Since P(X = 1) = P(X = 2)

$\Rightarrow \frac{\mu}{2} = 1 \Rightarrow \mu = 2$

Hence $P(X = 3) = \frac{e^{-2}.2^3}{3!}$

P(X = 3) = 0.180

Check Your Progress –1

In a Poisson distribution μ = 0.4

a) What is the probability that X = 0?

b) What is the probability that X>0?

Normal Distribution

Of all theoretical distributions for continuous variables, the most popular and commonly used distribution is the so-called normal or Gaussian distribution. A random variable X ios said to follow normal distribution, if and only if, its probability density function (which is the probability distribution of a continuous random variable X) is given by:

$$f_x(x) = \frac{1}{\delta\sqrt{2\pi}} e^{-\frac{1}{2\delta^2}(x-\mu)^2} \quad \ldots\ldots(1.6)$$

Where x is the real value of X; i.e.$-\infty < x < \infty$

The variable x is said to be distributed normally with mean μ and variance δ^2. i.e. X~N(μ, δ^2) read as X follows a normal distribution with mean μ and standard deviation δ. The density function given by (1.6) has two parameters, namely μ and δ. Here μ can take any value in the range $-\infty$ to ∞, where δ is any positive real value, i.e. δ >0.

Since probability can never be negative,

$f_x(x) \geq 0$ for all x.

In case, $\mu = 0$, $\delta = 1$, the density function for X is $f_x(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2}$ where $-\infty < x < \infty$, …..(1.7)

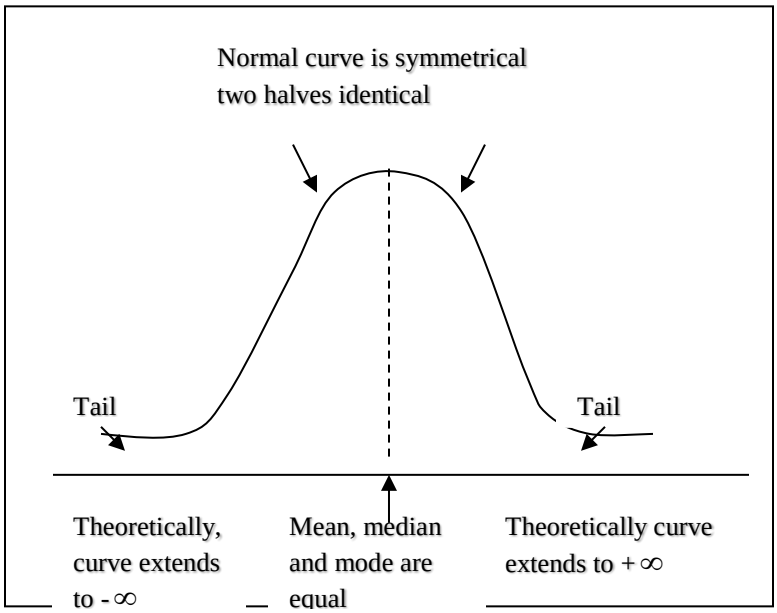
Notationally, X~N(0, 1)

In this situation the variable X is called the standardized normal variate and the distribution given by (1.7) is called the standardized normal distribution.

The normal probability distribution and its accompanying normal curve have the following characteristics:

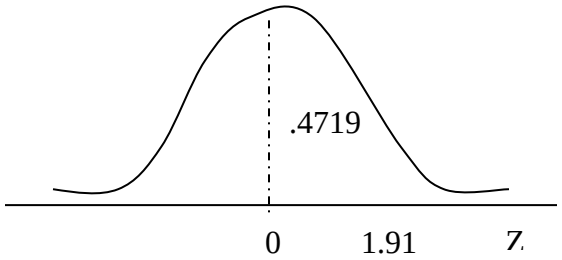
1. The normal curve is bell-shaped and has a single peak at the exact center of the distribution. The mean, median and mode of the distribution are equal and located at the peak. Thus, half the area under the curve is above this center point, and the other half is below it.
2. The normal probability distribution is symmetrical about its mean. If we cut the normal curve vertically at this central value, the two halves will be mirror images.
3. The normal curve falls off smoothly in either direction from the central value. It is asymptotic, meaning that the curve gets closer and closer to the x-axis but never actually touches it. That is, the "tails" of the curve extend indefinitely in both directions. In real-world problems, however, this is somewhat unrealistic. The life of a bulb, for example, could not be 100 years.

These characteristics are summarized in the chart below.



There is not just one normal probability distribution, but rather a "family" of them. For a random variable X which follows normal curve with mean μ and standard deviation δ, i.e X~N(μ, δ^2), the location and shape of the normal curve depends on μ and δ where μ and δ can take any value with in their range. Hence, no master table for the area under the curve can be prepared. This problem is very well overcome by consideration of a variable Z where Z = $\frac{\chi_i - \mu}{\delta}$. The variable Z is always distributed with mean zero and variance unity, i.e. Z~N(0, 1). The variable Z is known as standard normal variate. It is called standard as whatever be the parameters of the normal distribution of X, the transformed variable Z has always the normal distribution whose parameters are 0 and 1. Hence, only one table for area or ordinates of the normal curve is sufficient. The area as a matter of fact gives the probability for an event that Z takes certain values. These probabilities (areas) can always be found with the help of table. We hope that in your previous courses, you are familiarized to this table and able to read from this table. Just to remind you using one example, suppose the Z value is computed to be 1.91. To read the area under the normal curve between the

mean and X, go down the column of the table headed by letter Z to 1.9 and then move horizontally to the right and read the probability under column headed 0.01. It is .4719. This means that 47.19 percent of the area under the curve is between the mean and the X value 1.91 standard deviations above the mean. This is also interpreted as the probability that an observation is between 0 and 1.91 standard deviations above the mean. This is also interpreted as the probability that an observation is between 0 and 1.91 standard deviations of the mean.



1.4 SAMPLING DISTRIBUTION OF THE SAMPLE MEAN, PROPORTION AND VARIANCE

Sampling distribution describes the way in which a statistic or a function of statistics, which is/are the function(s) of the random variables x_1, x_2, ……., x_n, will vary form one sample to another sample of the same size. Such sampling distributions have given a filling to the number of test statistics for hypotheses testing. We are often concerned with sampling distribution in sampling analysis. If we take certain number of samples and for each sample compute various statistical measures such as mean, standard deviation, etc., then we can find that each sample may give its own value for the statistic under consideration. All such values of a particular statistic, say mean, together with their relative frequencies will constitute the sampling distribution of the particular statistic, say mean. Accordingly, we can have sampling distribution of mean, or the sampling distribution of standard deviation or the sampling distribution of any other statistical measure. It may be noted that each item in a sampling distribution is a particular statistic of a sample. The sampling distribution tends quite closer to the normal distribution if the number of samples (sample

size) is large. The significance of sampling distribution follows from the fact that the mean of a sampling distribution is the same as the mean of the universe.

Some important sampling distributions, which are commonly used are: (1) sampling distribution of mean; (2) sampling distribution of proportion; (3) Student's 't' distribution
4) F distribution; and (5) chi-square distribution. Some of this sampling distributions are mentioned in brief as below.

Sampling distribution of the sample means:

Sampling distribution of mean refers to the probability distribution of all the possible means of random samples of a given size that we take from a population. If samples are taken from a normal population, N(μ, δ), the sampling distribution of mean would also be normal with mean $\mu_{\bar{x}}$ = μ and standard deviation = $\delta/\sqrt{n}$, where μ is the mean of the population, δ is the standard deviation of the population and n means the number of items in a sample. But when sampling is from a population, which is not normal (may be positively or negatively skewed), even then, as per the central limit theorem, the sampling distribution of mean tends quite closer to the normal distribution, provided the number of sample items is large i.e., more than 30. In case we want to reduce the sampling distribution of mean to unit normal distribution i.e., N(0, 1), we can write the normal variate Z = $\frac{\bar{\chi} - \mu}{\delta / \sqrt{n}}$ for the sampling distribution of mean. This characteristic of the sampling distribution of mean is very useful in several decision situations for accepting or rejecting of hypotheses.

Sampling distribution of proportion:

Like sampling distribution of mean, we can as well have a sampling distribution of proportion. This happens in case of statistics of attributes. Assume that we have worked out the proportion of defective parts in large number of samples, each with say 100 items, that have been taken from an infinite population and plot a probability distribution of the said proportions, we obtain what is

known as the sampling distribution of the said proportions. Usually the statistics of attributes correspond to the conditions of a binomial distribution that tends to become normal distribution as n becomes larger and larger. If 'p' represents the proportion of detectives. i.e., of successes and 'q' represents the proportion of non-defectives i.e., of failures (or q = 1-p) and if 'p' is treated as a random variable, then the sampling distribution of proportion of successes has mean = p with standard deviation = $\sqrt{\frac{p.q}{n}}$ where n is the sample size. Persuming the binomial distribution approximating the normal distribution for large n, the normal variate of sampling distribution of proportion

$z = \frac{\hat{p} - p}{\sqrt{\frac{p.q}{n}}}$, where $\hat{p}$ (pronounced a p-hat) is the sample proportion of successes, can be used for testing of hypotheses.

i.e $Z = \frac{\hat{p} - p}{\sqrt{\frac{p.q}{n}}} \sim N(0, 1)$

In the above discussion, the assumption made is if n is very large. If the sampled population is not infinite, or not even very large, we will make some adjustments in standard error of the sample means and to the standard error of the sample proportions as follows.

For a finite population, if N is the population size and n is the sample size then:

Standard error of the sample means:

$$\delta_{\bar{x}} = \frac{\delta}{\sqrt{n}} \sqrt{\frac{N-n}{N-1}}$$

and standard error of the sample proportions;

$\delta_{\hat{p}} = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \cdot \sqrt{\frac{N-n}{N-1}}$ Where the term $\sqrt{\frac{N-n}{N-1}}$ is called the finite-population correction factor.

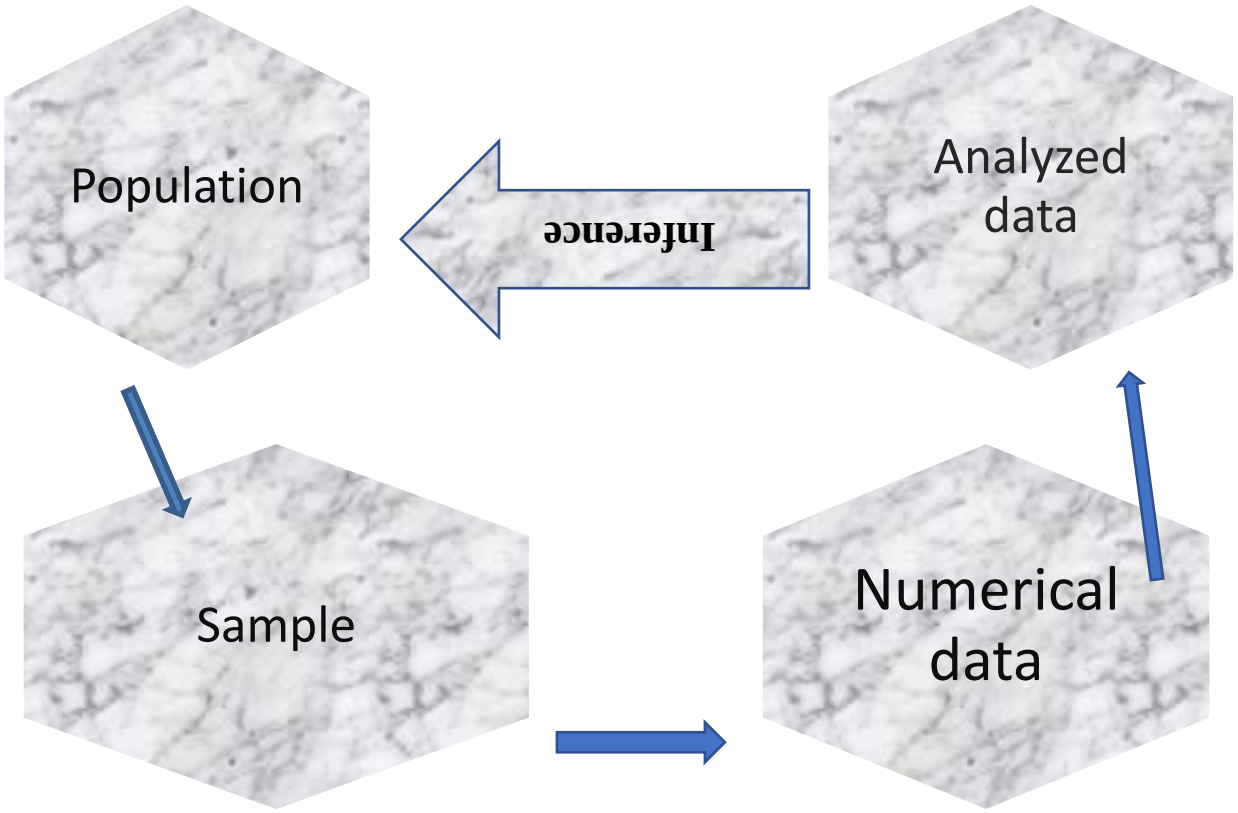
Check Your Progress –2

A consignment of 2000 transitors was received by a manufacturing concern. The concerned person drew a random sample of 60 transistors from the lot and tested them to find the number of defective pieces. It was found that 7 transistors out of the 60 were out of order. Calculate the estimated standard error of the sample proportion.

Sampling distribution of the sample variance: - As we have seen $\bar{x}$ is distributed normally with mean μ and variance δ^2 / n and $\sqrt{n}\frac{(\bar{x}-\mu)}{\delta}$ is distributed as standard normal variate. In a similar way, if δ^2 is the population variance and S^2 is the sample variance, for known δ^2, $\frac{ns^2}{\delta^2}$ is distributed as chi-square with (n-1) degree of freedom. i.e., $\frac{ns^2}{\delta^2} \sim \chi^2_{n-1}$ where n is the sample size, S^2 is sample variance, δ^2 is population variance and χ^2 is a chi-square distribution which is the sum of squares of independent standard normal variate and n-1 is the degree of freedom. (Degree of freedom are the number of independent observations in a set of observations).

CHAPTER TWO
STATISTICAL ESTIMATION

• Inference is the process of making interpretations or conclusions from sample data for the totality of the population.
• It is only the sample data that is ready for inference.
• In statistics, there are two ways through which inference can be made:
  - Statistical estimation
  - Statistical hypothesis testing.



Data analysis is the process of extracting relevant information from the summarized data.
This happens the reason being in many cases values for a population parameter are unknown. If parameters are unknown it is generally not sufficient to make some convenient assumption about their values, rather those unknown parameters should be estimated.
In business, many decision are made without complete information.
A firm does not know exactly what will be its sales volume next year or next month. A college does not know exactly how many students will enroll next year. Both must estimate to make decision about the future.

Types of Estimates

1 Point estimate

A number or a simple number is used to estimate a population parameter. A random sample of observations is taken from the population of interest and the observed values are used to obtain a point estimate of the relevant parameter.

a. The sample mean, $\bar{x}$, is the best estimator of the population mean μ. Different samples from a population yield different point estimates of μ,
In general:

The statistic $\bar{x}$ estimates μ

S estimates σ

S^2 estimates σ^2

Estimators and their properties / Goodness of an estimator/

The properties of good estimators are

a) Unbiasedness

b) Efficiency
c) Consistency and
d) Sufficiency

a) An estimator is said to be unbiased if its expected value is equal to the population parameter it estimates.
$E(\bar{x}) = \mu$ The sample mean , $\bar{x}$, is therefore, an unbiased estimator of the population mean. Any systematic deviation of the estimator away from the parameter of interest is called Bias.
b) An estimator is efficient if it has a relatively small variance (as standard deviation). The sample means have a variance of σ/n value which is less than σ. So the sample mean is an efficient estimator of the population mean.
c) An estimator is said to be consistent if its probability of being close to the parameter it estimates increases as the sample size increases. The sample mean is a consistent estimator of μ.
This is so because the standard deviation of $\bar{x}$ is $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$. As the sample size n increases, the standard deviation of $\bar{x}$ decreases and hence the probability that $\bar{x}$ will be closes to its expected value, μ, increases.
d) An estimator is said to be sufficient if it contains all the information in the data about the parameter it estimates. The sample mean is sufficient estimator of μ. Other estimators like the median and mode do not consider all values. But the mean considers all values (added and divided by the sample size).

4.9.2 Interval Estimates
Interval estimate states the range within which a population parameter probably lies. The interval with in which a population parameter is expected to lie is usually referred to as the *confidence interval*.
The confidence interval for the population mean is the interval that has a high probability of containing the population mean, μ
Two confidence intervals are used extensively.

1. 95% confidence interval and
2. 99% confidence interval

A 95% confidence interval means that about 95% of the similarly constructed intervals will contain the parameter being estimated. If we use the 99% confidence interval we expect about 99% of the intervals to contain the parameter being estimated.
Another interpretation of the 95 % confidence interval is that 95 % of the sample means for a specified sample size will lie with in 1.96standred deviations of the hypothesized population mean. For 99% the sample means will lie, with in 2.58 standard deviations of the hypothesized population mean.

Where do the values 1.96 and 2.58 come from?
The middle 95% of the sample mean lie equally on either side of the mean. And logically 0.95/2=0.4750 or 47.5% of the area is to the right of the mean and the area to the left of the mean is 0.4750.

The Z value for this probability is 1.96.
The Z to the right of the mean is + 1.96 and Z to the left is – 1.96.

4.9.2.1 Constructing Confidence Interval

a) Compute the standard error of the mean

Standard error of the mean is the standard deviation of the sample means.

$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$ σ = population standard deviation

If the population standa n = sample size ıe standard deviation of the sample s, is used to approximate the population standard deviation. $S_{\bar{x}} = \frac{S}{\sqrt{n}}$

This indicates that the error in estimating the population mean decreases as the sample size increases.

b) The 95% and 99% confidence intervals are constructed as follows when n ≥ 30.

95% confidence interval $\bar{x} \pm 1.96 \frac{S}{\sqrt{n}}$

99% confidence interval $\bar{x} \pm 2.58 \frac{S}{\sqrt{n}}$

1.96 and 2.58 indicate the Z values corresponding to the middle 95% or 99% of the observation respectively.

In general a confidence interval for the mean is computed by $\bar{x} \pm Z \frac{S}{\sqrt{n}}$, Z reflects the selected level of confidence.

Example. An experiment involves selecting a random sample of 256 middle managers for studying their annual income. The sample mean is computed to be Br. 35,420 and the sample standard deviation is Br. 2,050.

a. What is the estimated mean income of all middle managers (the population) ?
b. What is the 95% confidence interval c(rounded to the nearest 10)
c. What are the 95% confidence limits?
d. Interpret the finding.

Solution

a. Sample mean is 35 420 so this will approximate the population mean so μ = 35420. It is estimated from the sample mean.
b. The confidence interval is between 35170 and 35670 found by

$\bar{X} \pm 1.96 \frac{S}{\sqrt{n}} = 35420 \pm 1.96 \left(\frac{2050}{\sqrt{256}} \right) = 35168.87$ and 35671.13

c. The end points of the confidence interval are called the confidence limits. In this case they are rounded to 35170 and 35670. 35170 is the lower limit and 35070 is the upper limit.
d. Interpretation

If we select 100 samples of size 256 form the population of all middle managers and compute the sample means and confidence intervals, the population mean annual income would be found in about 95 out of the 100 confidence intervals. About 5 out of the 100 confidence intervals would not contain the population mean annual income.

Check Your Progress –2

A research firm conducted a survey to determine the mean amount smokers spend on cigarette during a week. A sample of 49 smokers revealed that the sample mean is Br. 20 with standard deviation of Br. 5. Construct 95% confidence interval for the mean amount spent.

Confidence interval for a population proportion

The confidence interval for a population proportion is estimated

$\bar{p} \pm Z\sigma_p$

Where σ_p is the standard error of the proportion and

$$\sigma_{\bar{p}} = \sqrt{\frac{\bar{p}(1 \quad \bar{p})}{n}}$$

Therefore the confidence interval for population proportion is constructed by

$$\bar{p} \pm Z \sqrt{\frac{\bar{p}(1 \quad \bar{p})}{n}}$$

Example. Suppose 1600 of 2000 union members sampled said they plan to vote for the proposal to merge with a national union. Union by laws state that at least 75% of all members must approve for the merger to be enacted. Using the 0.95 degree of confidence, what is the interval estimate for the population proportion? Based on the confidence interval, what conclusion can be drawn? $\bar{p} = \frac{1600}{2000} = 0.8$. The sample proportion is 80%

The interval is computed as follows. $\bar{p} \pm Z \sqrt{\frac{\bar{p}(1 \quad \bar{p})}{n}} = 0.80 \pm 1.96 \sqrt{\frac{0.80(1 \quad 0.8)}{2000}} = 0.08 \pm 1.96 \sqrt{0.00008}$

= 0.78247 and 0 – 81753 rounded to 0.782 and 0.818.

Based on the sample results when all union members vote, the proposal will probably pass because 0.75 lie below the interval between 0.782 and 0.818.

Check Your Progress –3

A sample of 200 people were assumed to identify their major source of news information; 110 stated that their major source was television news coverage. Construct a 90% confidence interval for the proportion of people in the population who consider television their major source of news information.

Confidence interval for small sample (Student's Distribution)

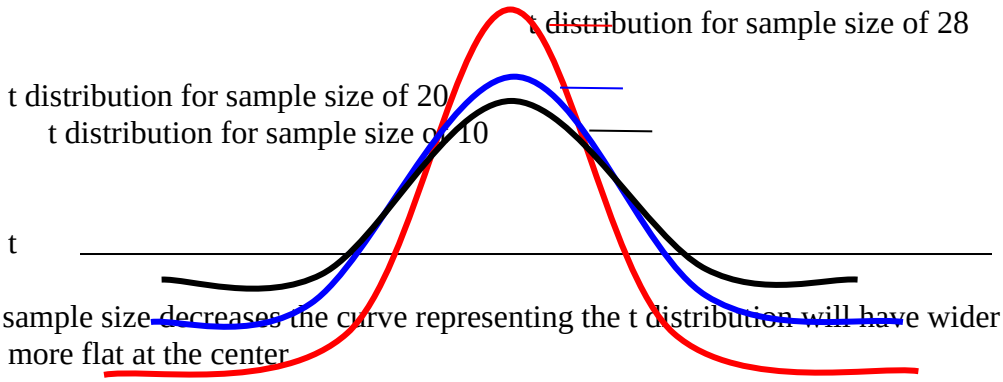
When the population is large and normal and the standard deviation is known the standard normal distribution is employed to construct the confidence interval for the mean and proportion. If the sample size is at least 30, the sample standard deviation can substitute the population standard deviation and the results are deemed satisfactory.

If the sample size is less than 30 and population standard deviation is unknown, the standard normal distribution, Z, is not appropriate. The student's t or the t distribution is used.

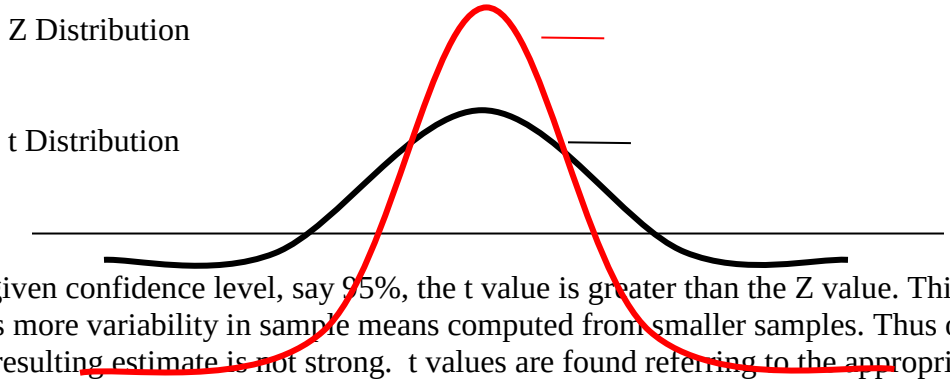
Characteristics of the Student's t Distribution

Assuming that the population of interest is normal or approximately normal, the following are the characteristics of the t distribution

1. It is a continuous distribution
2. It is bell-shaped and symmetrical
3. There is not one t distribution, but rather a 'family' of t distribution. All have the same mean of zero but their standard deviation differ according to the sample size, n. The t distribution differs for different sample size.
4. It is more spread out and flatter at the center than is the Z. However as the sample size increases the curve representing t distribution approaches the Z distribution.



As the sample size decreases the curve representing the t distribution will have wider tails and will be more flat at the center



For a given confidence level, say 95%, the t value is greater than the Z value. This is so because there is more variability in sample means computed from smaller samples. Thus our confidence in the resulting estimate is not strong. t values are found referring to the appropriate degrees of freedom in the t table. Degrees of freedom means the freedom to freely move data points or the freedom to freely assign values arbitrarily.

Degrees of freedom (df) = n – 1 where n is the sample size.

This implies that we can freely move or assign values for all data points except the last n^{th} value. If the mean of the distribution is specified there is a freedom to assign any value for all data points except the lost point.

Example - the mean of five data points is 12. Then it follows that the sum of all the five points is 60 = (5 x 12). Thus if five points are constrained to have a sum of 60 or a mean of 12, we have 5 – 1 = 4 degrees of freedom. If all the five data points are missing we are free to assign any value as long as their sum is 60 say 14, 12, 10, 9, 15. If 4 are missing we are free to assign any value since 60 minus the known value of a data point is known.

If two are un know, 14, 16, 10, x_3, x_4 since $14 + 16 + 10 + x_3 + x_4 = 60$
Then $x_3 + x_4 = 60 - 40 = 20$
$x_3 + x_4 = 20$. We can assign any value as long as their sum is 20. 10, 10 or 9.11 or 15.5 etc…
But if the four data points are known, (10, 14, 16, 12), the 5^{th} data point will have a predetermined value i.e. 60 – 52 = 8. Now we are not free to assign arbitrary value for this data point. Degrees of freedom can be obtained from the deviation based on the assumption that sum of the differences (d) between the mean and all values of the random variable (x) is zero. I.e., if we subtract the mean from all values of x the sum of the difference will be zero consider the above five data points. Their mean is 12 and their sum 60. Thus $(x_1 - 12) + (x_2 + 12) + (x_3 - 12) + (x_4 - 12) + (x_5 - 12) = 0 = d_1 + d_2 + d_3 + d_4 + d_5 = 0$
Now we are free to assign any value for only four missing differences as long as this sum is zero. So we have still n – 1 degrees of freedom.

Computing t value
The t variable representing the student's t distribution is defined as

$t = \dfrac{\bar{x} - \mu}{s/\sqrt{n}}$ where: $\bar{x}$ is the sample mean of n measurements, μ is the population mean and s is the sample standard deviation

Note that t is just like $Z = \dfrac{\bar{x} - \mu}{\sigma/\sqrt{n}}$ except that we replace σ with s. unlike our methods of large samples, σ cannot be approximated by s when the sample size is less than 30 and we can not use the normal distribution. The table for the t distribution is constructed for selected levels of confidence for degree of freedom up to 30. To use the table we need to know two numbers, the tail area, (1 minus confidence level selected), and the degree of freedom.
(1 – confidence level selected) is α, the Greek letter alpha. This is the error we committee in estimating.

The confidence interval for the sample mean is $\bar{x} \pm \alpha/2\,(n-1)\,S/\sqrt{n}$
Example. A traffic department in town is planning to determine mean number of accidents at a high-risk intersection. Only a random sample of 10 days measurements were obtained.
Number of accidents per day were
8, 7 10 15 11 6 8 5 13 12
Construct a 95% confidence interval for the mean number of accident per day.
a) Compute $\bar{x}$ and s

$\bar{x} = \dfrac{95}{10} = 9.5$ per day

$S_{\bar{x}} = \sqrt{\dfrac{\sum(x-\bar{x})^2}{n-1}} = \sqrt{\dfrac{94.5}{9}} = 3.24$ per day

The confidence level is 95% so

$\alpha = 1 - 0.95 = 0.05$

$$\frac{\alpha}{2} = \frac{0.05}{2} = 0.025$$

The degree of freedom, df = n – 1 = 10 – 1 = 9 from the t table $t_{0.025}$, df 9 = 2.76

The confidence interval is

$$\bar{x} \pm t_{.0025}\ df(9)\ \frac{s}{\sqrt{n}}$$

$$9.5 \pm (2.26)\ \frac{3.24}{\sqrt{10}}$$

$$9.5 \pm 2.3$$

7.2 to 11.80

With 95% confidence, the mean number of accident at this particular intersection is between 7.2 and 11.8.

Worksheet Questions

1. An electrical firm manufactures light bulbs that have a length of life that is approximately normally distributed with a standard deviation of 40 hours. If a random sample of 30 bulbs has an average life of 780 hours, find a 99% confidence interval for the population mean of all bulbs produced by this firm.
2. A random sample of 400 households was drawn from a town and a survey generated data on weekly earning. The mean in the sample was Birr 250 with a standard deviation Birr 80. Construct a 95% confidence interval for the population mean earning.

3. A major truck has kept extensive records on various transactions with its customers. If a random sample of 16 of these records shows average sales of 290 liters of diesel fuel with a standard deviation of 12 liters, construct a 95% confidence interval for the mean of the population sampled.

UNIT 3 - HYPOTHESIS TESTING

3.1 Introduction

A research worker or an experimenter has always some fixed ideas about certain population(s) vis-à-vis population parameter(s) based on prior experiments, surveys or experience. Sometimes these ideas might have been fixed in the mind vicariously. There is a need to ascertain whether these ideas or claims are correct or not by collecting information in the form of data. In this way, we come across two types of problems, first is to draw inferences about the population on the basis of sample data and the other is to decide whether our sample observations have come from a postulated population or not. The first type of problem, the problem of estimation, has almost been covered in the previous unit. In this unit, we would be dealing with the second type of problem, the problem of hypothesis testing.

Generally, a hypothesis is established before hand. By hypothesis we mean to give postulated or stipulated value(s) of a parameter. Also, instead of giving values, some relationship between parameters is postulated in the case of two or more populations. On the basis of observational data, a test is performed to decide whether the postulated hypothesis be accepted or not and this involves certain amount of risk. This amount of risk is termed as a *level of significance*. When the hypothesis is accepted, we consider it a non-significant result and if the revenue situation occurs, it is called a significant result.

A test is defined as a statistical procedure governed by certain rules, which leads to take a decision about the hypothesis for its acceptance or rejection on the basis of sample values.

Statistical tests of hypothesis play an important role in industry, biological sciences, social sciences, economics, etc. The use of tests has been made clear through a number of practical problems.

Basic Concepts in hypothesis testing

Basic concepts in the context of testing of hypothesis need to be explained.

a) Null Hypothesis and Alternative Hypothesis: -

A hypothesis is an assertion or conjecture about the parameter(s) of population distribution(s). Alternative hypothesis is usually the one which one wishes to prove and the null hypothesis is the one which one wished to disprove. Thus, a null hypothesis represents the hypothesis we are trying to reject, and alternative hypothesis represents all other possibilities.

A hypothesis which is to be actually tested for acceptance or rejection is termed as null hypothesis. It is denoted by Ho and alternative hypothesis which is denoted by H_1 or H_A is a statement about the population parameter or parameters, which gives an alternative to the null hypothesis within the range of pertinent values of the parameter.

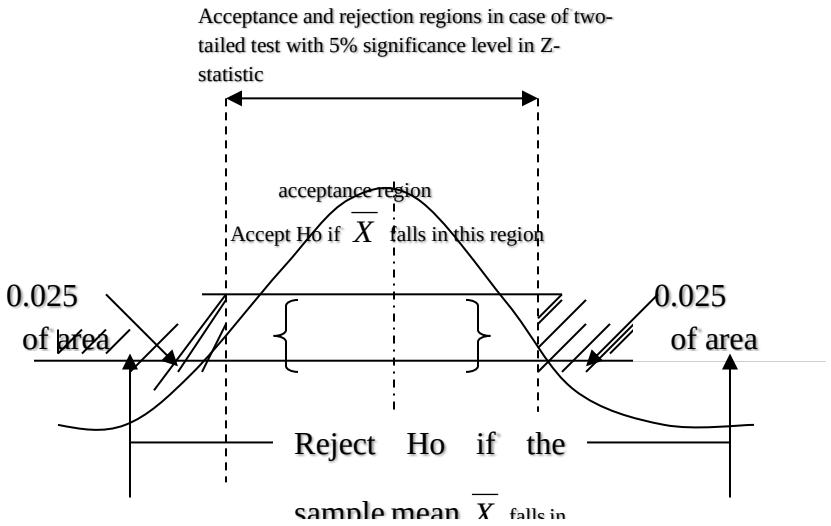
b) **Level of Significance**: - Is the maximum value of the probability of rejecting Ho when it is true and is usually determined in advance before testing the hypothesis or level of significance is the quantity of risk of the type I error which we are ready to tolerate in making a decision about Ho. In other words, it is the probability of type I error, which is tolerable.

The level of significance is denoted by α and is conventionally chosen as 0.05 or 0.01. α = 0.01 is used for high precision and α = 0.05 for moderate precision.

c) **Decision rule or test of hypothesis: -** Given a hypothesis Ho and an alternative hypothesis Ha, we make a rule which is known as decision rule according to which we accept Ho (i.e., reject Ha) or reject Ho (i.e., accept Ha). For instance, if Ho is that a certain lot is good (there are very few defective items in it) against Ha that the lot is not good (there are too many defective items in it), then we must decide the number of items to be tested and the criterion for accepting or rejecting the hypothesis. We might test 10 items in the lot and plan our decision saying that if there are none or only 1 defective item among the 10, we will accept Ho otherwise we will reject Ho (or accept Ha). This sort of basis is known as decision rule.

d) **Type I and Type II errors: -** In the context of testing of hypothesis, there are basically two types of errors we can make. We may reject Ho when Ho is actually true and we may accept Ho when in fact Ho is false. The former is known as Type I error and the latter as type II error. In other words, Type I error means rejection of hypothesis which should have been accepted and Type II error means accepting the hypothesis which should have been rejected. Type I error is denoted by α (alpha) known as α error, also called the level of significance of the test; and Type II error is denoted by β (beta) known as β error. The probability of Type I error is usually determined in advance and is understood as the level of significance of testing the hypothesis.

e) **Two-tailed and one-tailed tests: -** These two terms are quite important and must be clearly understood. A two-tailed test rejects the null hypothesis if, say, the sample mean is significantly higher or lower than the hypothesized value of the mean of the population. Such a test is appropriate when the null hypothesis is some specified value and the alternative hypothesis is a value not equal to the specified value of the null hypothesis. Symbolically, the two-tailed test is appropriate when we have Ho: $\mu = \mu o$ and Ha: $\mu \neq \mu o$ which mean $\mu > \mu o$ or $\mu < \mu o$. Thus, in a two-tailed test, there are two rejection regions, one on each tail of the curve which can be illustrated as under:

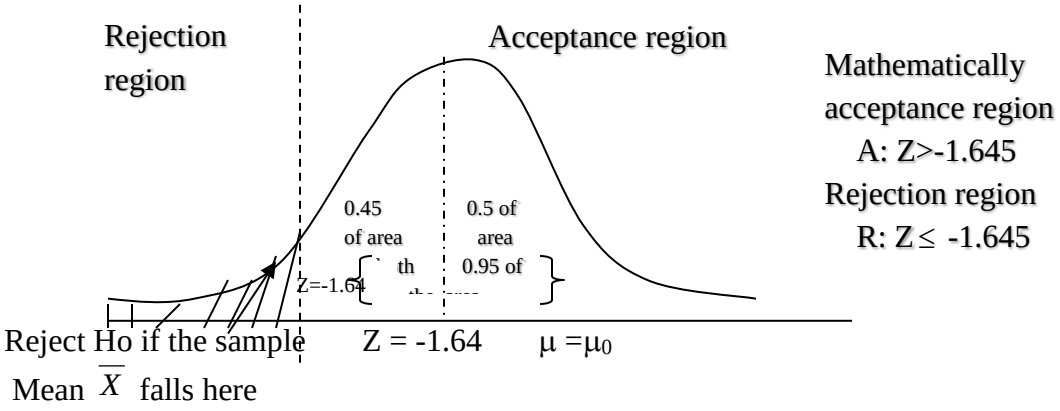


Mathematically we can state:

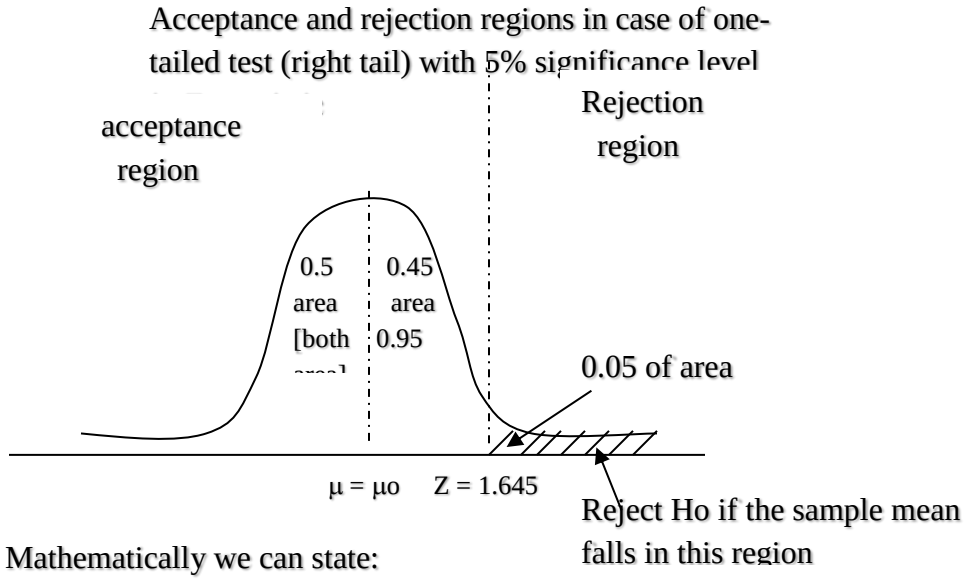
Acceptance Region A: $|z| \leq 1.96$

Rejection Region R: $|z| > 1.96$

A one-tailed test would be used when we are to test, say, whether the population mean is either lower than or higher than some hypothesized value. For instance, if our Ho: $\mu = \mu o$ and Ha: $\mu < \mu o$, then we are interested in what is known as Left-tailed test (wherein there is one rejection rejoin only on the left tail) which can be illustrated as below



In case our Ho: $\mu = \mu o$ and Ha: $\mu > \mu o$, we are then interested in what is known as one-tailed test (right-tail) and the rejection region will be on the right tail of the curve as shown below.



Mathematically we can state:

Acceptance Region A: $Z \leq 1.645$

Rejection Region R: $Z > 1.645$

Testing a hypothesis about a single population mean

Mean of the population can be tested presuming different situation such as the population may be normal or other than normal, it may be finite or infinite, sample size may be large or small, variance of the population may be known or unknown and the alternative hypothesis may be two-sided or one sided. Our testing technique will differ in different situations. We may consider some of the important situations.

1.Population normal, population sample size may be large or small but variance of the population is known, Ha may be one-sided or two-sided:

In such a situation, the test statistic is taken to be the Z-test, which is worked out as under:

$$Z = \frac{\overline{X} - \mu_0}{\delta_p / \sqrt{n}} \sim N(0,1)$$

Where $\overline{X}$ - is the sample mean

μo – is the hypothesized population mean under Ho.

δ_p – is the population variance

n – is the sample size.

Example: - A sample of 400 male students is found to have a mean height 67.47 inches. Can it be reasonably regarded as a sample from a large population with mean height 67.39 inches and standard deviation 1.30 inches? Test at 5% level of significance.

Solution: - Taking the null hypothesis that the mean height of the population is equal to 67.39 inches, we can write:

Ho: $\mu = 67.39$"

Ha: $\mu \neq 67.39$"

And the given information as $\overline{X} = 67.47$", $\delta_p = 1.30$", n = 400. Assuming the population to be normal, we can work out the test statistic Z as under:

$$Z = \frac{\overline{X} - \mu_0}{\delta_p / \sqrt{n}} = \frac{67.47 - 67.39}{1.3/\sqrt{400}} \quad \frac{0.08}{0.065} = 1.231$$

As Ha is two-sided in the given question, we shall be applying a two-tailed test for determining the rejection regions at 5% level of significance, which comes to as under. Using normal curve area table:

R: $|z| > 1.96$

The observed value of $Z_{\alpha/2}$ is 1.231 which is the acceptance region since R: $|z| > 1.96$ and thus Ho is accepted. We may conclude that the given sample (with mean height = 67.47") can be regarded to have been taken from a population with mean height 67.39″ and standard deviation 1.30″ at 5% level of significance.

If the alternative hypothesis is

Ha: $\mu < \mu o = 67.39$"

Then R: $Z < -Z_\alpha$

Zcal < -1.645

i.e compare 1.231 with –1.645 as Zcal > -1.645 then Ho is accepted

If the alternative hypothesis is

Ha: $\mu > \mu o = 67.39$" then compare Zcal = 1.231 with $Z_\alpha = 1.645$ in this case the rejection region R: $Z > Z_\alpha$ but as 1.231 < 1.645, again Ho remains accepted.

Check Your Progress –1

1. The following information is available

Ho: $\mu = 50$
$H_1 = \mu \neq 50$
The sample mean is 49, and the sample size is 36. The population standard deviation is 5. Use the .05 significance level.

a) is this a one-tailed or a two-tailed test?
b) State a decision rule.
c) Compute the value of the test statistic
d) What is your decision regarding Ho?

2. *Population normal, sample size small and variance of the population unknown, Ha may be one-sided or two-sided.*

In such a situation, t-test is used and the test statistic t is worked out as under:

$t = \dfrac{\overline{X} - \mu_0}{\sigma_s / \sqrt{n}}$ with degree of freedom = (n – l)

and σ_s = sample variance = $\sqrt{\dfrac{\sum_i (X_i - \overline{\chi})^2}{n-1}}$

Example: - From past records it is known that the mean life of a battery used in a digital clock is 305 days. The lives of the batteries are normally distributed. The battery was recently modified to last longer. A sample of 20 modified batteries were tested. It was discovered that the mean life was 311 days. And the sample standard deviation was 12 days. At the 0.05 level of significance, did the modification increase the mean life of the battery?

Solution: - Sample size n = 20 is small.
Ho: $\mu \leq \mu o = 305$
H_1: $\mu > \mu o$
Sample variance σ_s =12, sample mean $\overline{X}$ = 311

Test statistics: $t = \dfrac{\overline{X} - \mu_0}{\sigma_s / \sqrt{n}}$

$= \dfrac{311 - 305}{12 / \sqrt{20}}$

$= 2.236$

and since the test is one-sided (right-tailed) the rejection region R: $t > t_\alpha$ ($\alpha = 0.05$)
t-tabulated for (n – 1) = 19 degree of freedom from the t -table is $t_{0.05}$ (1 9) = 1.729
as t = 2.236 > 1.729 is in the rejection region, then reject Ho and accept H_1, that the mean is greater than 305 days. It is concluded that the modification increased battery life.

Similarly: - for small sample with unknown population the two-tailed test is obtained by comparing |t| with $t_{\alpha/2}(n-1)$. i.e. R: $|t| > t_{\alpha/2}(n-1)$ and for the alternative hypothesis $\mu < \mu o$, (left tailed) test one should compare t calculated with $-t_\alpha(n-1)$ i.e. the rejection region is R: $t < -t_\alpha(n-1)$

3. ***Population may not be normal but sample size is large, variance of the population may be known or unknown, and Ha may be one – sided or two-sided:***

In such a situation, by central limit theorem, we can assume the normal distribution and use Z-test and workout the test statistic Z as:

$Z = \frac{\bar{\chi} - \mu_0}{\sigma p \sqrt{n}}$ in case of infinite

Population with known variance and replace by σ_s (sample variance) if the population variance σ_p is not known.

The relationship between interval estimation and hypothesis testing

Hypothesis testing is the second method of statistical inference is closely associated with the confidence interval. While the confidence interval shows the lower and upper values of the parameter under consideration, hypothesis testing will help us to determine if a statement concerning specific values of the parameter is strongly supported by information obtained from the sample data. One can relate confidence intervals as acceptance regions for two tailed tests, the upper limits of the confidence intervals with the critical point in the right-tailed tests and the lower limits of the confidence intervals can also be related with the critical point of the left-tailed tests. One can also relate the critical values and the p-values (the area to the right of or left) of the computed test statistics.

CHAPTER FOUR

CHI-SQUARE DISTRIBUTIONS (X^2)

The **chi-square distribution** is *the sum of the squares of **k** independent random variables* and therefore can never be **less than zero**; it extends indefinitely in the **positive direction**. Actually the chi-square distributions constitute a family, with each distribution defined by the degrees of freedom (*df*) associated with it. For small *df* values the chi-square distribution is skewed considerably to the right (positive values). As the *df* increase, the chi-square distribution begins to approach the normal curve. Because of space limitations, chi-square values are listed only for certain probabilities.

Chi-square distribution involves the frequency of events and it helps to understand the relationship between two categorical variables; grade level, sex, age group, year, and others. Chi-square distribution helps to compare what we actually observed with what we expected oftentimes using population data or theoretical data. It assists us in determining the role of random chance variation between our categorical variable. We use Chi-square distribution and critical value to accept or reject our hypothesis. Binomial distribution in which only two possible outcomes could occur on a single trial in an experiment. An extension of the binomial distribution is a multinomial distribution in which more than two possible outcomes can occur in a single trial. **The chi-square goodness-of-fit test** is *used to analyze probabilities of multinomial distribution trials along a single dimension.* For example, if the variable being studied is economic class with three possible outcomes of lower income class, middle income class, and upper income class, the single dimension is economic class and the three possible outcomes are the three classes. On each trial, one and only one of the outcomes can occur. In other words, a family unit must be classified either as lower income class, middle income class, or upper income class and cannot be in more than one class.

Chi-square goodness-of-fit tests are one tailed because a chi-square of zero indicates perfect agreement between distributions. Any deviation from zero difference occurs in the positive direction only because chi-square is determined by a sum of squared values and can never be negative. With four categories in this example (excellent, pretty good, only fair, and poor), $k = 4$. The degrees of freedom are $k - 1$ because the expected distribution is given: $k - 1 = 4 - 1 = 3$. For = .05 and $df = 3$, the critical chi-square value is After the data are analyzed, an observed chi-square greater than 7.8147 must be computed in order to reject the null hypothesis.

The chi-square goodness-of-fit test compares the *expected,* or theoretical, *frequencies* of categories from a population distribution to the *observed,* or actual, *frequencies* from a distribution to determine whether there is a difference between what was expected and what was observed. On several occasions a decision maker needs to understand whether an actual sample distribution matches with a known theoretical probability distribution such as binomial, Poisson, normal and so on. The chi-square goodness-of-fit test enables as to determine the extent to which theoretical distribution coincides with empirical sample distribution. To apply this test, a particular theoretical distribution is first hypothesized for a given population and then the test is carried out to determine whether or not sample data could have come from the population of interest with the hypothesized theoretical distribution. The observed or values come from the sample and the expected frequencies or values come from the theoretical hypothesized probability distribution. The goodness-of-fit test now focuses on the difference between the observed values and the expected values. Large differences between the two distributions throw doubt on the assumption that the hypothesized theoretical distribution is correct. On the other hand, small difference between the two distributions may be assumed to be resulting from sampling error. For example, airline industry officials might theorize that the ages of airline ticket purchasers are distributed in a particular way. To validate or reject this expected distribution, an actual sample of ticket purchaser ages can be gathered randomly, and the observed results can be compared to the expected results with the chi-square goodness-of-fit test. This test also can be used to determine whether the observed arrivals at teller windows at a bank are Poisson distributed, as might be expected. In the paper industry, manufacturers can use the chi-square goodness-of-fit test to determine whether the demand for paper follows a uniform distribution throughout the year.

Formula 4.1 is used to compute a chi-square goodness-of-fit test.

CHI-SQUARE GOODNESS-OF-FIT TEST 4.1

$$\chi^2 = \sum \frac{(f_o - f_e)^2}{f_e}$$

$$df = k - 1 - c$$

where

f_o = frequency of observed values
f_e = frequency of expected values
k = number of categories
c = number of parameters being estimated from the sample data

This formula compares the frequency of observed values to the frequency of the expected values across the distribution. The test loses one degree of freedom because the total number of expected frequencies must equal the number of observed frequencies; that is, the observed total taken from the sample is used as the total for the expected frequencies.

In addition, in some instances a population parameter, such as λ, μ, or δ is estimated from the sample data to determine the frequency distribution of expected values. Each time this estimation occurs, an additional degree of freedom is lost. As a rule, if a uniform distribution is being used as the expected distribution or if an expected distribution of values is given, *k* - 1 (degrees of freedom) are used in the test. In testing to determine whether an observed distribution is Poisson, the degrees of freedom are *k* - 2 because an additional degree of freedom is lost in estimating. In testing to determine whether an observed distribution is normal, the degrees of freedom are *k* - 3 because two additional degrees of freedom are lost in estimating both and from the observed sample data. Karl Pearson introduced the chi-square test in 1900.

The general steps to conduct goodness – of-fit test for any hypothesized population distribution are:-

Step 1: State the null hypotheses and alternative hypothesis

H_o: The observed distribution is the same as the expected distribution.

H_1: The observed distribution is not the same as the expected distribution.

Step 2: Select a random sample and record the observed frequencies (f_o values) for each category.

Step 3: calculate expected frequencies (f_e values)

Step 4: compute the value of test statistic. The statistical test being used is

$$X^2 = \sum \frac{(f_o - f_e)^2}{f_e}$$

Step 5: using level of significance α and *df=n-1* provided that the number of expected frequencies are 5 or more for all categories, find the critical value of X^2.

Step 6: compare the value of test statistic Vs the critical value of X^2, and the following decision rule:-

- If X^2 calculated is greater than its critical value reject H_o
- Otherwise do not reject H_o

Example 1: How can the chi-square goodness-of-fit test be applied to business situations? One survey of U.S. consumers conducted by *The Wall Street Journal* and NBC News asked the question: "In general, how would you rate the level of service that American businesses provide?" The distribution of responses to this question was as follows:

Excellent	8%
Pretty good	47%
Only fair	34%
Poor	11%

Suppose a store manager wants to find out whether the results of this consumer survey apply to customers of supermarkets in her city. To do so, she interviews 207 randomly selected consumers as they leave supermarkets in various parts of the city. She asks the customers how they would rate the level of service at the supermarket from which they had just exited. The response categories are excellent, pretty good, only fair, and poor. The observed responses from this study are given as follows:-

Results of a Local Survey of Consumer Satisfaction

Response	Frequency (f_o)
Excellent	21
Pretty good	109
Only fair	62
Poor	15

Now the manager can use a chi-square goodness-of-fit test to determine whether the observed frequencies of responses from this survey are the same as the frequencies that would be expected on the basis of the national survey. α=0.05

Step 1: State the null hypotheses and alternative hypothesis

Ho: The observed distribution is the same as the expected distribution.

Ha: The observed distribution is not the same as the expected distribution.

Step 2: Select a random sample and record the observed frequencies (f_o values) for each category.

Results of a Local Survey of Consumer Satisfaction

Response	Frequency (f_o)
Excellent	21
Pretty good	109
Only fair	62
Poor	15

Step 3: calculate expected frequencies (f_e values)

Response	Expected Proportion	Expected Frequency (f_e) (proportion × sample total)
Excellent	.08	(.08)(207) = 16.56
Pretty good	.47	(.47)(207) = 97.29
Only fair	.34	(.34)(207) = 70.38
Poor	.11	(.11)(207) = 22.77
		207.00

Step 4: compute the value of test statistic. The statistical test being used is

$$\chi^2 = \sum \frac{(f_o - f_e)^2}{f_e}$$

Response	f_o	f_e	$\frac{(f_o - f_e)^2}{f_e}$
Excellent	21	16.56	1.19
Pretty good	109	97.29	1.41
Only fair	62	70.38	1.00
Poor	15	22.77	2.65
	207	207.00	6.25

Step 5: using level of significance α and *df=n-1* provided that the number of expected frequencies are 5 or more for all categories, find the critical value of X^2.

With four categories in this example (excellent, pretty good, only fair, and poor), $k = 4$. The degrees of freedom are k - 1 because the expected distribution is given:

k - 1 = 4 - 1 = 3. For = .05 and df = 3, the critical chi-square value is 7.8147.

Step 6: compare the value of test statistic Vs the critical value of X^2. The critical value of X^2 is 7.8147 and the observed value of chi-square of or test statics is 6.25. Because the observed value of chi-square of 6.25 is not greater than the critical table value of 7.8147, the store manager will not reject the null hypothesis.

Business Implications:-

Thus the data gathered in the sample of 207 supermarket shoppers indicate that the distribution of responses of supermarket shoppers in the manager's city is not significantly different from the distribution of responses to the national survey. The store manager may conclude that her customers do not appear to have attitudes different from those people who took the survey.

Example 2: Dairies would like to know whether the sales of milk are distributed uniformly over a year so they can plan for milk production and storage. A uniform distribution means that the frequencies are the same in all categories. In this situation, the producers are attempting to determine whether the amounts of milk sold are the same for each month of the year. They ascertain the number of gallons of milk sold by sampling one large supermarket each month during a year, obtaining the following data. Use $\alpha = .01$ to test whether the data fit a uniform distribution.

Solution

Step 1: State the null hypotheses and alternative hypothesis

H_0: The monthly figures for milk sales are uniformly distributed.

H_1: The monthly figures for milk sales are not uniformly distributed.

Step 2: Select a random sample and record the observed frequencies (f_o values) for each category.

Month	f_o
January	1610
February	1585
March	1649
April	1590
May	1540
June	1397
July	1410
August	1350
September	1495
October	1564
November	1602
December	1655
Total	18,447

Step 3: calculate expected frequencies (f_e values). If the frequencies are uniformly distributed, the same number of gallons of milk is expected to be sold each month. The expected monthly figure is $\frac{18,447}{12} = 1537.25$ gallons

Month	f_e
January	1537.25
February	1537.25
March	1537.25
April	1537.25
May	1537.25
June	1537.25
July	1537.25
August	1537.25
September	1537.25
October	1537.25
November	1537.25
December	1537.25
Total	18,447.00

Step 4: compute the value of test statistic. The statistical test being used is

$$\chi^2 = \sum \frac{(f_o - f_e)^2}{f_e}$$

Month	f_o	f_e	$\frac{(f_o - f_e)^2}{f_e}$
January	1610	1537.25	3.44
February	1585	1537.25	1.48
March	1649	1537.25	8.12
April	1590	1537.25	1.81
May	1540	1537.25	0.00
June	1397	1537.25	12.80
July	1410	1537.25	10.53
August	1350	1537.25	22.81
September	1495	1537.25	1.16
October	1564	1537.25	0.47
November	1602	1537.25	2.73
December	1655	1537.25	9.02
Total	18,447	18,447.00	$\chi^2 = 74.37$

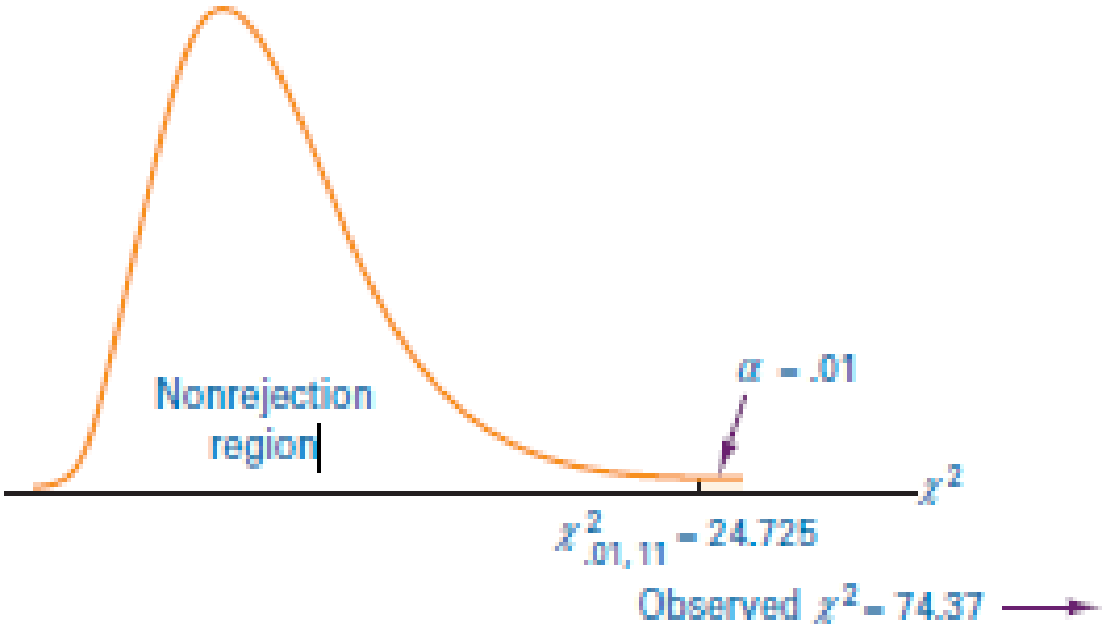
Step 5: using level of significance α and *df = n-1* provided that the number of expected frequencies are 5 or more for all categories, find the critical value of X^2.

There are 12 categories and a uniform distribution is the expected distribution, so the degrees of freedom are k - 1 = 12 - 1 = 11. For = .01, the critical value is 24.725.

Step 6: compare the value of test statistic Vs the critical value of X^2. The critical value of X^2 is 24.725 and the observed value of chi-square of or test statics is 74.37. The observed value of 74.37 is greater than the critical table value of 24.725, so the decision is to reject the null hypothesis. This problem provides enough evidence to indicate that the distribution of milk sales is not uniform.

Business Implications:

Because retail milk demand is not uniformly distributed, sales and production managers need to generate a production plan to cope with uneven demand. In times of heavy demand, more milk will need to be processed or on reserve; in times of less demand, provision for milk storage or for a reduction in the purchase of milk from dairy farmers will be necessary. The following Minitab graph depicts the chi-square distribution, critical chi-square value, and observed chi-square value.



Example 3: Suppose a teller supervisor believes the distribution of random arrivals at a **Dashen** bank is Poisson and sets out to test this hypothesis by gathering information. The following data represent a distribution of frequency of arrivals during 1-minute intervals at the bank. Use .05 to test these data in an effort to determine whether they are Poisson distributed.

Number of Arrivals	Observed Frequencies
0	7
1	18
2	25
3	17
4	12
≥5	5

Solution

Step 1: State the null hypotheses and alternative hypothesis

H_0: The frequency distribution is Poisson.

H_1: The frequency distribution is not Poisson.

Step 2: Select a random sample and record the observed frequencies (f_o values) for each category.

Number of Arrivals	Observed Frequencies
0	7
1	18
2	25
3	17
4	12
≥5	5

Step 3: Calculate expected frequencies (f_e values). To determine the expected frequencies, the supervisor must obtain the probability of each category of arrivals and then multiply each by the total of the observed frequencies. These probabilities are obtained by determining lambda and then using the Poisson formula. As it is the mean of a Poisson distribution, lambda can be determined from the observed data by computing the mean of the data. In this case, the supervisor computes a weighted average by summing the product of number

of arrivals and frequency of those arrivals and dividing that sum by the total number of observed frequencies.

Number of Arrivals	Observed Frequencies	Arrival × Observed
0	7	0
1	18	18
2	25	50
3	17	51
4	12	48
≥5	5	25
	84	192
	$\lambda = \frac{192}{84} = 2.3$	

With this value of lambda and the Poisson distribution, the supervisor can determine the probabilities of the number of arrivals in each category. The expected probabilities are determined using these probabilities and the total of 84 from the observed data, the supervisor computes the expected frequencies by multiplying each expected probability by the total (84) as follows:-

Arrivals	Expected Probabilities	Expected Frequencies
0	.1003	8.42
1	.2306	19.37
2	.2652	22.28
3	.2033	17.08
4	.1169	9.82
≥5	.0837	7.03
		84.00

Step 4: Compute the value of test statistic. The supervisor uses the above expected frequencies and the observed frequencies to compute the observed value of chi-square/test statistics.

Arrivals	Observed Frequencies	Expected Frequencies	$\frac{(f_o - f_e)^2}{f_e}$
0	7	8.42	.24
1	18	19.37	.10
2	25	22.28	.33
3	17	17.08	.00
4	12	9.82	.48
≥5	5	7.03	.59
	84	84.00	$\chi^2 = 1.74$

Step 5: Using level of significance α and *df = n-1* provided that the number of expected frequencies are 5 or more for all categories, find the critical value of X^2. The degrees of freedom are k - 1 = 6 - 1 =5. For = .05, the critical table value is 11.070.

Step 6: Compare the value of test statistic Vs the critical value of X^2. The critical value of X^2 is 11.070 and the observed value of chi-square of or test statics is 1.74. The observed value of 1.74 is not greater than the critical chi-square value of 11.070, so the supervisor's decision is to not reject the null hypothesis. In other words, he fails to reject the hypothesis that the distribution of bank arrivals is Poisson.

Chapter Five

Analysis of Variance (ANOVA)

Introduction

ANOVA has a lot to do in statistics. In the case of regression analysis, **ANOVA** is the technique of decomposing the total variation in the dependent variable into its components (regression and error). The regression component of the total variation is the variation in the dependent variable due to the independent variable under consideration, whereas, the error component of the total variation is the variation in the dependent variable due to other factors, which are not included in the regression model.

Student learning objectives

After the completion of this lesson, students must be able to:

- ✓ Understand the application of ANOVA
- ✓ Use ANOVA as a hypothesis testing instrument

(See appendix)

6.1 The meaning of ANOVA

Analysis of variance (ANOVA) is a statistical instrument which is used for testing as to whether the means of more than two quantitative populations are equal. It helps in identifying whether that two different sample data classified in terms of a single variable are meaningful, provides a meaningful comparison between sample data which are classified according two or more variables etc.

The sample statistic which we use here is the sample F-distribution

6.2 Assumptions in ANOVA

ANOVA can only hold true when the following assumptions are met:

- Observations are drawn from normally distributed populations
- Observations represent random samples from a population
- Variance or standard deviations of the population are equal

As a hypothesis testing tool, ANOVA involves the steps that were used to undertake hypothesis testing.

You can also refer to the following table as a general process of ANOVA; ANOVA table

Source of variation	Degrees of freedom, df	Sum of squares, ss	Mean sum of squares, mss	F-statistic
Regression	$df_1 = k$	$Rss = \sum(\hat{y} - \bar{y})^2$	$mRss = Rss/df1$	$F_{cal} = \frac{mRss}{mEss}$
Error	$df_2 = n-(1+k)$	$Ess = \sum(y - \hat{y})^2$	$mEss = Ess/df2$	
Total	$df_1 + df_2 = n-1$	$Tss = \sum(y - \bar{y})^2$		

Key: k = number of independent variables under consideration.

N = sample size (no. of pairs of X and Y)

Rss = Regression sum of squares.

Ess = Error sum of squares

Tss = Total sum of squares.

mRss = mean Regression sum of squares

mEss = mean Error sum of squares

Tss = Rss+Ess

F_{cal} is used to test the overall significance of a regression model. Testing the overall significance of the model means checking whether or not X and Y are related. To do so, we simply test if β_1 is different from zero or not. That is, we test the following hypotheses:

$H_0 : \beta_1 = 0$

$H_a : \beta_1 \neq 0$

- If H_0 is rejected, the model is significant (X and Y are related).
- If H_0 is not rejected, the model is not significant (X and Y are not related).

N.B: H_0 is rejected if $|F_{cal}| > F_{\alpha}(df1, df2)$

$F_{\alpha}(df1, df2)$ is a value to be read from F-table. A sample of the F-table is given below.

α =0.05				
df2 \ df1	1	2	3	4……
1	161.4			
2	18.51			
3	10.13			
4	7.71			
5	.			
.	.			
.	.			
.	.			
10 →	.			
11	.			
12	4.96 ←			
.				
.				
.				
.				

Example

The following data are collected on the supply and price of a certain product.

Price (**X**)	2	4	6	8	10	12	14	16	18	20
Supply (**Y**)	10	20	50	40	50	60	80	90	90	120

a) Construct a regression equation of **Y** on **X**

b) Test the significance of the regression model for the price-supply data (use α =0.05).

Solution: $H_0 : \beta_1 = 0$
$H_a : \beta_1 \neq 0$

ANOVA table:

Source of variation	Df	ss	mss	F_{cal}
Regression	df1=1	Rss=507	507	$\frac{507}{38.7} = 13.1$
Error	df2=10	Ess=387	38.7	
Total	df1+ df2=11	Tss=894		

Decision: Reject H_0,

Because [F_{cal}=13.1]> [$F_{\alpha}(df1, df2) = F_{0.05}(1,10) = 4.96$, see the F-table in the appendix].

N.B: The coefficient of determination (**r^2**) can also be expressed in terms of Rss and Ess.

That is, $r^2 = \frac{Rss}{Tss}$

For example for the given price-supply data, $r^2 = \frac{Rss}{Tss} = \frac{507}{894} = 0.567$ which is the same as the value previously obtained by simply squaring the coefficient of correlation (**r**).

The other approach of computing the F-value is shown below:

$$F = \frac{n(variance\ of\ samples)}{mean\ of\ sample\ variance}$$

Where, n is the sample size

The F-distribution table has two unique degrees of freedom $(df = v)$ values:

Degree of freedom in the numerator, (v_1) and degree of freedom in the denominator, (v_2)

The degree of freedom $df = v$ is computed as:

$v_1 = numerator\ degree\ of\ freedom = k - 1$ and

$$v_2 = denominator\ degree\ of\ freedom = k(n-1)$$

Where k is the number of sample and n is the sample size.

Before trying to use the F-test and its distribution as a sample statistic/or critical values, it is important to compute the following:

$$variance\ of\ samples = S^2 = \frac{\sum(x-\bar{x})^2}{n-1}$$

$$mean\ of\ sample\ means = \frac{\sum s^2}{k} = \frac{s_1^2 + s_2^2 + \cdots + s_k^2}{k}$$

$$grand\ mean = \bar{\bar{x}} = \frac{\sum \bar{x}}{k}$$

$$variance\ of\ sample\ means = s_k^2 = \frac{\sum(\bar{x}-\bar{\bar{x}})}{k-1}$$

Once we calculate these values we may proceed for the analysis of variance.

Example:

AGIP Oil Company wanted to determine if the amount of oil delivered by its truck to customers is the same in its three sales districts. The company obtained a random data as given below:

Gallons delivered in one delivery

Districts		
1	**2**	**3**
81	100	295
179	158	82
142	272	155
199	248	271
124	62	212

Perform an ANOVA at a 0.05 level of significance to determine if the mean amounts per delivery in the sales districts are equal.

Step-1: state the hypothesis

Ho: mean delivery amounts for the three districts equal

Ho: mean delivery amounts for the three districts is not equal

Step-2: state the decision rule

$v_1 = k - 1 = 3 - 1 = 2$

$v_2 = k(n-1) = 3(5-1) = 12$

$$F\alpha, v_1, v_2 = F0.05, 2.12, = 3.89$$

Then, reject *Ho*: if sample $F > 3.89$

Step-3: compute sample F

District 1		**District 2**		**District 3**	
x_1	$(x_1 - \bar{x}_1)^2$	x_2	$(x_2 - \bar{x}_2)^2$	x_3	$(x_3 - \bar{x}_3)^2$
81	4096	100	4624	295	8464
179	1156	158	100	82	14641
142	9	272	10816	155	2304
199	2916	248	6400	271	4624
124	441	62	11236	212	81
725	**8618**	**840**	**33176**	**1015**	**30114**

$$\bar{x}_1 = \frac{\sum x_1}{n} = \frac{725}{5} = 145$$

$$\bar{x}_2 = \frac{\sum x_2}{n} = \frac{840}{5} = 168$$

$$\bar{x}_3 = \frac{\sum x_3}{n} = \frac{1015}{5} = 203$$

$$\bar{\bar{x}} = \frac{145 + 168 + 203}{3} = 172$$

$S^2 = \frac{(145-172)^2+(168-172)^2+(203-172)^2}{3-1} = 853$, is the variance of sample means

Again, compute for the sample variances:

$$S_1^2 = \frac{8618}{4} = 2154.5$$

$$S_2^2 = \frac{33176}{4} = 8294$$

$$S_3^2 = \frac{30114}{4} = 7528.5$$

Then, the mean of sample variances will be given as:

$$\frac{\sum S^2 \bar{x}}{k} = \frac{2154.5 + 8294 + 7528.5}{3} = 5992.3$$

$$F = \frac{n(variance\ of\ samples)}{mean\ of\ sample\ variance} = \frac{3(853)}{5992.3} = 0.427$$

Step-4: accept or reject Ho:

As $F_{\text{calculated}}$ is less than the table F table value, then accept Ho; that is 0.0427 > 3.89 which is false.

Try this!

Suppose that a typewriter manufacturer has prepared three different study manuals for use by typists learning to operate an electronic word processing typewriter. Each manual was studied using a simple random sample of 5 typists. The time to achieve proficiency was recorded for each typist, and the sample mean learning times for manuals A, B, and C were hours. The manufacturer wants to know whether the variation in sample means is large enough to show that: population mean learning times for the manuals are different. Perform an ANOVA test at a 0.05 level of significance to determine whether mean learning times are equal.

Manual-I	**Manual-II**	**Manual-III**
21	17	31
27	25	28
29	20	22
23	15	30
25	23	24

Self-test questions:

1. A research firm has tested four random samples of types of light bulbs by keeping bulbs lit unit they burned out. The following table gives burning times in tens of hours:

Type			
1	2	3	4
78	65	77	76
78	73	69	83
72	75	68	77
71	71	75	83
77	67	77	82
80	69	72	85

   Perform ANOVA (use $\alpha = 0.05$) to determine if mean burning times for the bulb types are equal

2. A company has three manufacturing plants and company officials want to determine whether is a difference in the average age of workers at the three locations. The following data is the age of five randomly selected workers at each plant:

Plant (age of employees)		
1	2	3
29	32	25

27	33	24
30	31	24
27	34	25
28	30	26

Perform ANOVA test (use $\alpha = 0.01$) to determine whether there is a significant difference in the mean ages of the workers at the three palnts

3. Hogos wanted to build a new store in three different kebeles in Mekelle town. In an attempt to study whether the numbers of potential customers are different for the three kebelles, passerby counts were made for a random sample of 11 periods at each kebelle. The sample means and sample variances are given below:

Kebelle	Sample mean	Sample variance
16	125	288
17	141	248
19	124	304

Using $\alpha = 0.01$, perform ANOVA test to determine if the mean numbers of passerby at the three kebelles is equal.

Chapter Six

Regression and Correlation Analysis

Introduction

Do you want to know what will be the effect of advertisement expenditures on the sales volume of your company's product? May be yes! But how are you going to do it? Perhaps you need to know industry experiences relating to the amount of expenditure, frequency of the advertisement, timing of the advertisement, type of media used and volume of reach by each media type, consumers' information elasticity etc. Based on this information, you can predict the effect on the volume of sales.

5.1. Regression Analysis

Definition- Regression analysis is one of the statistical analysis tools used in the perdition of the value of one variable (dependent variable) given the value of another variable-(independent variable), when the two variables are related to each other. In regression analysis we shall develop on estimating equation. For example- in the above introductory note the industry experiences on amount of expenditure, frequency of advertisement type of media used and the volume of reach

by each media etc are the known variable values (independent variable) based on which we can project what the volume of sales (dependent variables) will be in future period. In fact regression analysis enables to study and measure the statistical relationship between two or more variables; however, we will focus only to relationships involving two variables.

Example

If you want to study and measure the relationship between price and quality demanded,

- collect data
- present the data in an order array [as pairs of (x, y)]
- compute (determine) the functional linear relationship between the variables

Methods to determine the regression line

1. The Scatter Diagram

A scatter diagram is a graph of observed points where each point represents the two coordinate values. So, simply by looking at the chart it is possible to determine the extent of association between the two variables.

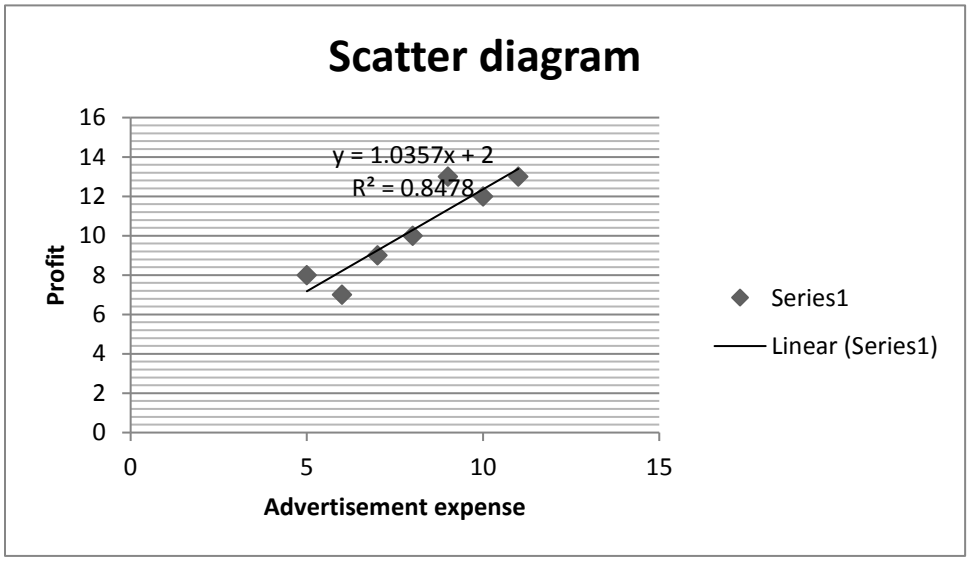
The wider the scatter in the chart, the less close is the relationship. The closer the points and the closer they came to falling on a line passing through them, the higher the degree of relationship.

Example The following data represents the money spent on advertising of a product and the consequent profits achieved from each advertising period for a given product

Advertising	Profit
5	8
6	7
7	9
8	10
9	13
10	12
11	13

Required draw the scatter diagram

The scatter diagram is drawn by locating the X-Y points or values on the graph as shown in the graph below (the dotted points):



From the trend in the relationship, you can see that it is increasing even though the relationship is not perfect. In other words, profit increases with an increase in advertisement expenditure.

Exercise

A teacher wants to study the number of students absent on a given day is related to the mean temperature on that day. A random sample of 10 days is used for the study. The following table shows data on the number of students absent from class and average mean temperature.

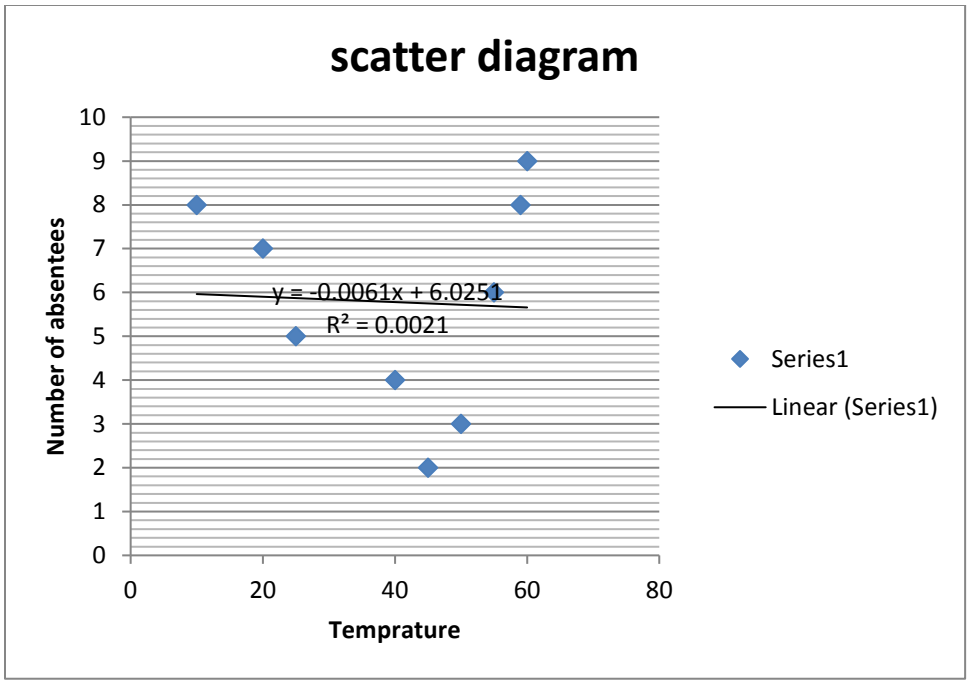
Absent students 8 7 5 4 2 3 6 8 9

Temperature (C^o) 10 20 25 40 45 50 55 59 60

a. Determine which variable is dependent and which is independent

b) Draw a scatter diagram of these data

   i. From the data we can understand that the number of absentee students is affected by the change in temperature. That is temperature is independent variable and absenteeism is a dependent variable



The dots represent the scatter diagram. From the above diagram, however, we see that temperature and number of absenteeism have little relationship as indicated by the regression line in the diagram.

Activity

The Mekelle University Environmental Health department wants to determine the statistical relationships between many different variables and the common cold. The following table contains the data on the use of facial tissues and the number of days that the common cold symptoms were exhibited by seven people

Facial tissues	2000	1500	500	750	600	900	1000
Number of days	60	60	10	15	5	25	30

a) Determine the dependent and independent variables
b) Draw the scatter diagram
c) What is the type of the relationship
d) Interpret your graph

The Least Square Method

With this method we find the line of best fit that involves representativeness, i.e., the distance between the line and the points is minimal. Least Square method is a mathematical procedure to

find the equation for the straight line that minimizes the sum of the square distances between the line and the data points, as measured in the vertical (or Y) direction.

The derivation of the equations needed to compute the Y-intercept and slope of a regression line using the method of least squares requires the use of calculus. For simplicity, we present here the equations only;

$$b_1 = \frac{n\sum xy - [(\sum x)(\sum y)]}{n\sum x^2 - (\sum x)^2}$$

Where $\sum_x$ = sum of the x values

$\sum_y$ = sum of the y values

$\left(\sum x\right)^2$ = sum of x values squared

$\sum xy$ = sum of the product of x and y for each period observation.

n = Number of x-y observations

$$b_0 = \frac{\sum y}{n} - b_1 \frac{\sum x}{n} \qquad \text{Or } b_0 = \bar{y} - b_1 \bar{x}$$

Where $\sum_x$ = sum of the x values

$\sum_y$ = sum of the y values

b_1 = slope of the line computed using equation

n = Number of x-y observations

The equation of the line is given by $\hat{Y} = b_0 + b_1(x)$

Remember $b_1 = \frac{n\sum xy - [(\sum x)(\sum y)]}{n\sum x^2 - (\sum x)^2}$

Let's take the following example which was used to draw a scatter diagram above:

Advertising(x)	Sales (y)	XY	X^2
5	8	40	25
6	7	42	36
7	9	63	49
8	10	80	64
9	13	117	81
10	12	120	100
11	13	143	121
Total 56	72	605	476

$$b_1 = \frac{7(605) - [56x72]}{7(476) - (56)^2}, \quad b_1 = \frac{203}{196} = 1.036$$

And $b_0 = \bar{y} - b_1\bar{x}$ but $\bar{y} = \frac{72}{7} = 10.29$ and $\bar{x} = \frac{56}{7} = 8$

$$\therefore b_0 = 10.29 - 1.036(8) = 2.002$$

$\hat{Y} = 2.002 + 1.036(x)$ is the equation of the regression line.

Interpretation; from the equation of the line we can see that for unit increase in advertisement expense, sales increases by 1.036 birr.

b. If the advertisement expenses were 7 units, sales will be computed as

$\hat{Y} = 2.002 + 1.036(7) = 9.254 units$

Example

The Maintenance Head of IVECO (Ethiopia) wants to know whether or not there is a positive relationship between the annual maintenance cost of their new bus assemblies and their age. He collects the following data:

Bus	Maintenance cost (birr) (y)	Age (yrs) (x)	XY	X^2	Y^2
1	859	8	6,872	64	737,881
2	682	5	3,410	25	465,124
3	471	3	1413	9	221,841
4	708	9	6,372	81	501,264
5	1,049	11	12,034	121	1,100,401
6	224	2	448	4	50,176
7	320	1	320	1	102,400
8	651	8	5,208	64	423,801
9	1094	12	12,588	144	1,196,836
	6058	59	48,665	513	4.799,724

Required

a. Plot the scatter diagram

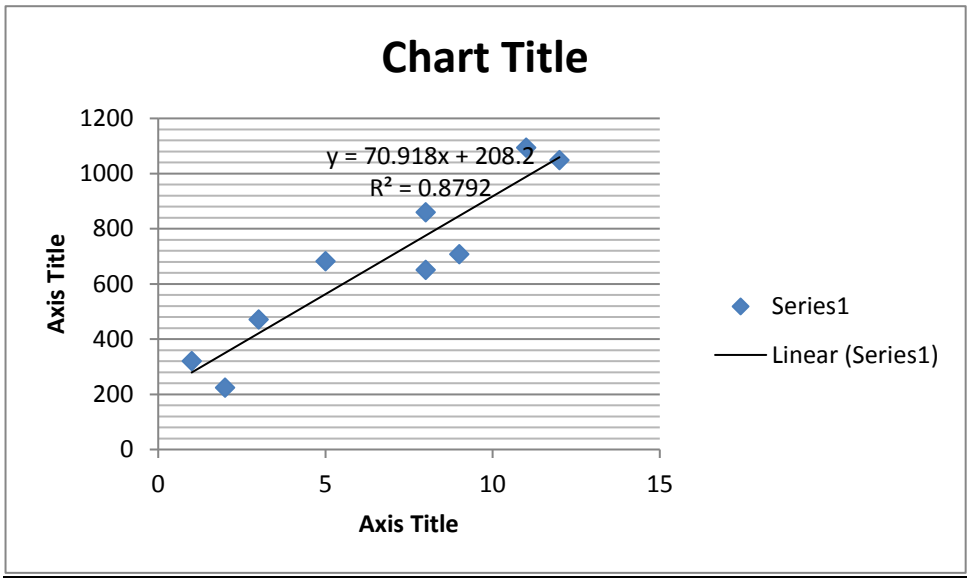
b. What kind of relationship exists between these two variables?

c. Determine the simple regression equation

d. Estimate the annual maintenance cost for a five-year-old bus

Solution

a.



b. As shown in the diagram, there is a positive and direct relationship which is equal to 87.9% (R^2 = 0.879). You will see how the R^2 is computed as well as its meaning.

c) $b_1 = \dfrac{n\sum xy - [(\sum x)(\sum y)]}{n\sum x^2 - (\sum x)^2}$

$$b_1 = \frac{9x48{,}665 - (59x6058)}{9x513 - (59)^2} = \frac{437{,}985 - 357{,}422}{4617 - 3481} = \frac{80{,}563}{1{,}136} = 70.92$$

$b_0 = \bar{y} - b_1\bar{x}$

$$= \frac{\sum y}{n} - 70.92\left(\frac{\sum x}{n}\right)$$

$$= \frac{6058}{9} - 70.92\left(\frac{59}{9}\right) = 673.11 - 464.92 = 208.19$$

Then $y_r = 70.92 + 208.19x$,

d) When the bus is only five years

$y_r = 70.92 + 208.19(5)$

1, 111.87 birr is the maintenance cost at age five

5.2. Correlation Analysis

It is desirable to measure the extent of the relationship between x and y as well as observe it in a scatter diagram. The measurement used for this purpose is the correlation coefficient. This is a numerical value ranging-1 to +1 that measures the strength of the linear relationship between two quantitative variables. Correlation coefficient (ρ= rho) exist for a population of date values and for each sample selected from it.

Correlation coefficient characteristics

Data Collection	Correlation coefficient	Range of Values
Population	P	$-1 \leq p \leq +1$
Sample	r	$-1 \leq r \leq +1$

For both **p** and **r**

-1: prefect negative relationship

0: No linear relationship

+1: Perfect positive relationships

These values are rarely encountered in real world situations, but they are good benchmarks for evaluating the correlation coefficient of any data collection. Karl Pearson's Coefficient of Correlation (Pearson Product Moment Correlation Coefficient) (r)

$$r = \frac{n\sum xy - (\sum x)(\sum y)}{\sqrt{n\sum x^2 - (\sum x)^2}.\sqrt{n\sum y^2 - (\sum y)^2}}$$

Where $\sum_x$ = sum of the x values

$\sum_y$ = sum of the y values

$\sum x^2$ = sum of squared x values

$\sum y^2$ = sum of squared y values

$(\sum x)^2$ = the sum of the x values squared

$(\sum y)^2$ = sum of y values squared

$\sum xy$ = the sum of the product of x and y for each period observation.

n = number of x-y observations

Example

The following table comprises that data on the weight of a cars and miles covered for the sample of 5 cars.

Weight of a car	Miles
2,743	21.4
3,518	15.2

1,855	38.9
5,214	12.7
4,341	17.8

Required

a. Compute the Pearson product moment correlation coefficient

b. Interpret your answer.

Solution

First thing you do is that find the square, sum and product values of the sample data as below:

Weight of a car(x)	Mileage (y)	XY	X^2	Y^2
2,743	21.4	58,700.2	7,524,049	457.96
3,518	15.2	53,473.6	12,376,324	231.04
1,855	38.9	72,159.5	3,441,025	1,513.21
5,214	12.7	66217.8	27,185,796	161.29
4,341	17.8	77,269.8	18,844,281	316.84
Total 17,671	106.0	327,820.9	69,371,475	2,680.34

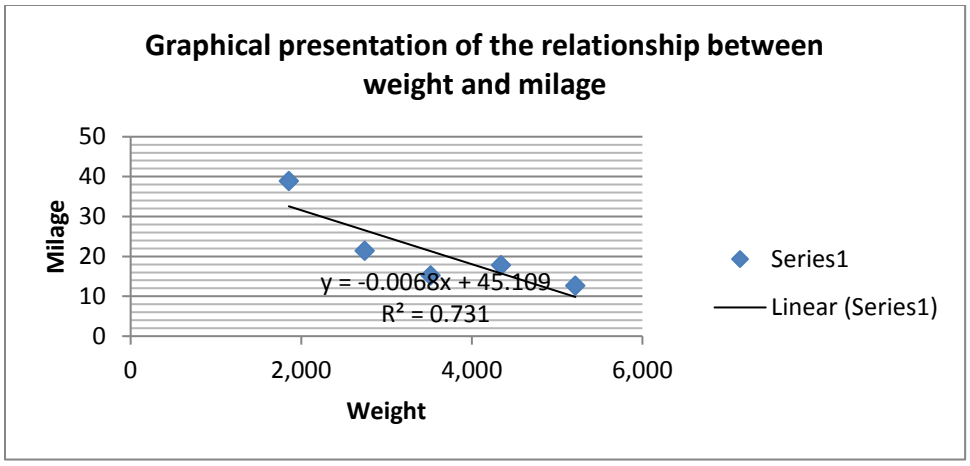
Then use the numbers in the table and insert them in to the formula

$$r = \frac{n\sum xy - (\sum x)(\sum y)}{\sqrt{n\sum x^2 - (\sum x)^2}\sqrt{n\sum y^2 - (\sum y)^2}}$$

$$= \frac{5(327,820.9) - (106x17,71)}{\sqrt{5(69,371,475) - (17,671)^2}\sqrt{5(2,680.34) - (106)^2}} = \frac{-234,021.5}{273,494.4} = -0.855 \cong -0.86$$

c) The correlation coefficient r = -0.86 indicates a rather strong negative linear relationship between car weight and miles per gallon in to the sample. That is, cars that weight more seem to get fewer miles per gallon and vice versa.

You may also see this same relationship in the following diagram with the R^2 value being 0.731:



***The coefficient of Determination**

Another measure of goodness of fit of the regression line is the Coefficient of Determination, which is the square of the correlation coefficient, that is

Coefficient of Determination = $\mathbf{r^2}$

The value of $\mathbf{r^2}$ lies between 0 and 1, inclusive.

r^2 measures close to 1indicates a strong correlation between the variables

r^2 measure close to 0 indicates little or no correlation

The total change or variation in the dependent variable can be divided in to two:

a. Explained variation- is the change in the dependent variable(Y) explained by changes in the independent variable(X). The proportion of variation is given by: $\mathbf{r^2.100\%}$

b. Unexplained variation- is the variation in the dependent variable(Y) due to chance, excluded variables etc.

The proportion of unexplained variation is given by: $\mathbf{(1-r^2).100\%}$

For example, find the proportion of explained and unexplained variations for the above example,

Solution

$\mathbf{r^2}$ is computed to be -0.86; then the proportion of explained variation is given as 0.86.100 = 86% and the proportion of unexplained variation is (1-0.86).100 = 14%

Activity

AFRICA Insurance Share Company feels that the amount of time a sales person spends with clients should be positively related to the size of that clients account.

The company gathers the following information so as to see whether the relationship is positive:

Client	Accounts Size, y	Minutes spent x
1	1056	108
2	825	123
3	651	62
4	748	95
5	894	58
6	1,242	134
7	1,058	87
8	1,112	78
9	1,259	120

Required

a. Compute the correlation coefficient
b. What would be the interpretation

Acct size y	Minutes spent x	XY	X^2	Y^2
1056	108			
825	123			
651	62			
748	95			
894	58			
1,242	134			
1,058	87			
1,112	78			
1,259	120			
8,845	865			

The following data are collected on the supply and price of a certain product.

Price (**X**)	2	4	6	8	10	12	14	16	18	20
Supply (**Y)**	10	20	50	40	50	60	80	90	90	120

a) Construct a regression equation of **Y** on **X.**
b) Find the value of **Y** (supply) when the price (**X)** is 11.
c) How much should the price be in order for the supply to be 100?
d) Find the correlation coefficient b/n **X**&**Y**.
e) Find the percentage variation in supply which is not explained by price.

5.3. Rank Correlation Method

The method assumes that data units can be ordered (ranked so that one can measure the degree of correlation between the series of ranks (often two series). The method is called rank correlation coefficient *R*

R is given by: $R = 1 - \frac{6\sum D^2}{N(N^2 - 1)}$, **Where** N = the number of individuals in each series and

D = difference between the ranks of the two series

To perform the computation the number of individuals (N) in the series must be assigned ranks.

Example

In the 2000 Miss Millennium Ethiopia Beauty contest two judges ranked eight candidates A, B, C, D, E, F, G, H, in order of their performance, as is shown below.

	A	B	C	D	E	F	G	H
$Judge_1$	5	2	8	1	4	6	3	7
$Judge_2$	4	5	7	3	2	8	1	6

	Judge $_1$(X)	$Judge_2$(Y)	D(X-Y)	D^2
A	5	4	1	1
B	2	5	-3	9
C	8	7	1	1
D	1	3	-2	4
E	4	2	2	4
F	6	8	-2	4
G	3	1	2	4
H	7	6	1	1
				$\sum D^2$ 28

Find the rank correlation coefficient

Solution

Find the rank correlation coefficient

$$R = 1 - \frac{6\sum D^2}{N(N^2 - 1)} = 1 - \frac{6(28)}{8(8^2 - 1)} = 1 - \frac{168}{8(64 - 1)} = 1 - \frac{168}{8(63)} = 1 - \frac{168}{504} =$$

1-0.33 = 0.67 = 67%

Example

The following table presents the scores of students in New Millennium College 3rd year Management Students

Marks in:	1	2	3	4	5	6	7	8	9	10
Mathematics	55	74	40	50	65	74	69	80	40	43
Statistics	62	60	55	70	72	67	80	79	52	40

Compute the rank correlation coefficient

Solution

Students	Maths (X)	Rank	Statistics	Rank	D=X-Y	D^2
1	55	6	62	7	-1	1
2	75	2	68	5	-3	9
3	40	10	55	8	2	4
4	50	7	70	4	3	9
5	65	5	72	3	2	4
6	74	3	67	6	-3	9
7	69	4	80	1	3	9
8	80	1	79	2	-1	1
9	41	9	52	9	0	0
10	43	8	40	10	-2	4
						$\sum D^2 = 50$

R= $1-\frac{6\sum D^2}{N(N^2-1)} = 1-\frac{6(50)}{1000-10} = 1-\frac{300}{990}$ =1-030330 = 0.697=69.7%

Example

A company hired six computer technicians. The technicians were given a test designed to measure their basic knowledge. After a year of service, their boss was asked to rank to each technician's job performance. Test scores and performance ranking are given below:

Technician	Test score	Performance ranking
1	82	3
2	60	6
3	80	2
4	67	5
5	94	1
6	89	4

Is there any relationship between test score and job performance?

Solution

Test score	Rank test score	Performance score	D	D^2
82	3	3	0	0
60	6	6	0	0

80	4	2	2	4
67	5	5	0	0
94	1	1	0	0
89	2	4	-2	4
			$\sum d = 0$	$\sum d^2 = 8$

Then the rank correlation coefficient is calculated as

$$R= 1-\frac{6\sum D^2}{N(N^2-1)} = 1-\frac{6(8)}{6(6^2-1)} = 1-\frac{48}{6(35)} 1-\frac{48}{210} = 1-0.2285714 = 0.77$$

Then we can say there is a good positive relationship between the two variables

Self-Test Questions

1.A specialist in a hospital claims that the number of full-time employees in the hospital can be estimated by counting the number of beds in the hospital. Now, a researcher wants to establish a regression model so as to predict the number of full-time employees by the number of beds. After a survey in 12 hospitals, the researcher obtained the following data:

Hospital ID	Beds	FTEs
1	23	69
2	29	95
3	29	102
4	35	118
5	42	126
6	46	125
7	50	138
8	54	178
9	64	156
10	66	184
11	76	176
12	78	225

a. Plot the scatter diagram
b. Find the estimated regression line
c. Find *r*
d. What type of relationship do they have?

2. The following is a data on the number of students and the annual sales turnover of fast food restaurants around major universities;

Restaurant ID	Number of students (X)	Sales volume (Y)
1	2	58

2	6	105
3	8	88
4	8	118
5	12	117
6	16	137
7	20	157
8	20	169
9	22	149
10	26	202

a. Does there seem to be any relationship between the sales volume and the number of students in these universities?
b. Find the estimated regression equation
c. What type of relationship do they have
d. Compute the Karl Pearson's Co-efficient of Correlation

3. The coefficient of correlation b/n tying speed and typing error was found to 0.4. The percentage variation in typing errors due to inattention is 4 times as great as the percentage variation due to speed. Find **r** between typing errors and inattention.

4. The coefficient of rank correlation of the marks of 10 students in statistics and accounting was found to be 0.8. It was later discovered that the difference in ranks in the two subjects of one student was wrongly taken as 7 instead of 9. Find the correct **r'**.