

ADDIS ABABA MEDICAL AND BUSINESS COLLEGE
DEPARTEMENT OF ACCOUNTING
DEGREE WEEKEND 2013A & B
STATISTICS FOR MANAGEMENT II/business statistics II

CHAPTER 1: SAMPLING TECHNIQUE AND SAMPLING DISTRIBUTION

1. Introduction

Inferential statistics is a systematic method of inferring satisfactory conclusions about the population on the basis of examining a few representative units termed as sample. The process of selecting samples is called sampling. Generalization of the sample data results to the population, which is one of the characteristic features of research, needs scientific approach of searching for facts. Therefore, sampling must be scientific.

1.1. Definition of sample and census survey

A subset of the population selected for the study is known as sample. The group from which the samples are selected is called Universe or Population.

Sample survey: is a procedure, which makes one able to draw inferences about the population by observing or measuring few items.

Census survey: is a method of inquiry, which makes one able to draw inferences by observing each item constitute the population.

Objectives of sampling are:

To obtain maximum information about the characteristics of the population with less time, energy and expenditure

To obtain the best possible values of parameter.

Sampling refers to the method of selecting a sample from the universe. A proper procedure is to be adopted for evaluating the sample plan in order to select representative units of the universe. Sampling occupies a key role in the study and has acquired the status of a technical job.

The number of units in the sample is called Sample size. Not on a new line

* Sample size should never be too small nor too large but optimum. Optimum fulfills the needs of efficiency, representative ness, reliability and validity.

The size of sample for a study is determined on the basis of the following factors

- The size of the population
- The availability of resources

- The degree of accuracy
- The homogeneity or heterogeneity of the population
- The nature of the study
- The method of sampling technique adopted
- The nature of respondents

Advantages and disadvantages of sample survey

If the sample is drawn on scientific approach, the adopted sample design is good and the sample size is adequate. Sample method has some merits over the census method. That are:

- Sampling saves time and money.
- It is much convenient as it involves less personal staff.
- It is useful when population is infinitely large.
- It can be more accurately supervised and data can be carefully selected.
- It is useful in case of inspecting the quality of units, which we have to resort to sampling, such as testing the quality of bulbs, tubes, strength of stencils, testing explosives, etc.

Sampling method has its limitations and problems, which are:

It would give unreliable data if not designed and executed carefully. Samples are like medicines. They can be harmful if taken carelessly or without knowledge of their effect. The service of skilled, trained, qualified personnel for supervision; and sophisticated equipment and statistical techniques are required. In the absence of these, it may not be reliable.

Sample survey is not useful when information is needed about each and every unit of the population.

1.2. SAMPLING TECHNIQUES

Introduction

Statistical methods are especially appropriate for handling data (information), which are subject to variations, and for which we can observe only a fraction of the totality of observations, which may exist. Under this situation, techniques must be devised by

which we can make inferences about the nature of the totality of the universe from the particular observation we have.

1.3. Types of sampling techniques

Sampling technique refers to the method of selecting a sample from the universe (population). It occupies a key role in a study and has acquired the status of being a technical job. The right type of sampling technique is of paramount importance in the execution of a sample survey in accordance with the objectives and the scope of the inquiry. The sampling methods may broadly be classified as:

- Probability sampling (simple, stratified, & systematic)
- Non-probability sampling (judgment, convenient, quota, incidental, purposive, self-selected, etc.)
- Mixed sampling (cluster sampling)

Random (probability) sampling

Random sampling method is a method of selection of a sample such that each item within the population has equal chance of being selected.

In this method, there is no place for investigator's bias in sample selection since it depends on probability. It provides more accurate estimates in the sense of greater precision.

I. Simple Random Sampling Method (SRSM): involves very simple method of drawing a sample from a given population. The selection of samples is random in character.

The oldest method adopted in simple RS is the use of lottery system.

Suppose population size is 100 and sample size is 10. i.e. $N = 100$ and $n = 10$. Hundred chits would be prepared bearing the serial number of units in the universe. These chits would be put together and shuffled thoroughly, and then ten would be drawn one by one. The sampling units corresponding to the number on the selected chits will form a random sample. This method gives a sample, which is quite independent of the natures of universe. This method is commonly in practice even at present.

The other most practical and inexpensive method is the method of "Random Number Tables" (RNT). If we have to select a sample of size n from a universe of size N less than 9, then the numbers can be paired as 0 to 9. If we have to select a sample of size n

from a universe of size N less than 99, then the RNT will be from 00 to 99. If N is less than 999, then from 000 to 999 and so on.

Then, select any K from the RNT and if $K \leq N$, the k th unit will be selected as a sample and if $K > N$, divide K by N and take the remainder or the R th unit as a sample. This process continues till n number of samples are selected.

Example: From 40 big enterprises in Addis Ababa, we want to study the case of only 5 of them. Let 12, 59, 67, 81 and 97 be the numbers selected from the RNT. Then which of the items of the population are selected for the sample.

Solution: $N = 40$, $n = 5$, $K = 12, 59, 67, 81, 97$.

Since $N = 40$ is a two digit number, the RNT will be from 00 to 99.

$K = 12 < N \rightarrow$ The 12th item is selected for the sample.

$K = 59 > N \rightarrow 59 \div 40 = 1 + 19/40$ The 19th item is selected.

$K = 67 > N \rightarrow 67 \div 40 = 1 + 27/40$ The 27th item is selected

$K = 81 > N \rightarrow 81 \div 40 = 2 + 1/40$ The 1st item is selected

$K = 97 > N \rightarrow 97 \div 40 = 2 + 17/40$ The 17th item is selected

$\therefore$ The 1st, 12th, 17th, 19th and 27th items are selected for the 'study'.

II Stratified Random Sampling Method (STRSM)

Under this method, the whole population is divided into a number of homogeneous groups or strata. From each of these strata, random sample of size n is selected. Thus, stratified RS means selecting a number of random samples, one from each stratum of the universe. It is used when each group has small variation within itself but wide variation between the groups.

The sample may be either proportionate or disproportionate. Suppose the universe is divided into two groups consisting of 100 and 160 respectively and their respective sample sizes being 10 % of the universe. Meaning a sample of size $10 + 16 = 26$ is drawn in proportion to the total number of items. But in disproportionate stratified RSM, samples are taken from each stratum regardless of the number of units in the universe. Thus in the above example, an equal number of units i.e. 12 from each stratum may be drawn in which the total number of items in the sample is 24.

- The size of sample items which must be selected from the i th stratum is denoted by n_i and is given by

$$n_i = \frac{nN_i}{N}$$

Where n - Sample size
N - Population size
N_i - Size of the ith stratum

Example: In unity University College a survey is to be conducted on 120 students' tendency towards Mathematics. The total number of students in each field is as indicated below. Give the sample size of each field of study.

N _i	N ₁	N ₂	N ₃	N ₄	N ₅
Field of study	Accounting	Business	Law	Marketing	Architecture
No. of students	3000	2000	1500	2500	1000

Solution: n = 120

$$N_1 = 3000, \quad N_2 = 2000, \quad N_3 = 1500, \quad N_4 = 2500, \quad N_5 = 1000$$

$$N = N_1 + N_2 + N_3 + N_4 + N_5 = 10,000$$

Then,
$$n_1 = \frac{nN_1}{N} = \frac{120 \times 3000}{10000} = \underline{\underline{36}} \quad n_2 = \frac{nN_2}{N} = \frac{12}{1000} \times 2000 = \underline{\underline{24}}$$

Or
$$n_1 = \frac{n}{N} \times N_1 = \frac{120}{10000} \times 3000 = 36 \quad \text{i.e. } \frac{n}{N} = \underline{\underline{0.012}}$$

$$n_3 = \frac{12}{1000} \times 1500 = \underline{\underline{18}}, \quad n_4 = \frac{12}{1000} \times 2500 = \underline{\underline{30}}, \quad n_5 = \frac{12}{1000} \times 1000 = \underline{\underline{12}}$$

III. Systematic Random Sampling Method (SYRSM)

In this method, a random starting point is selected from the list representing the universe and the remaining units are automatically selected in a definite sequence at an equal spacing from one another. This method is recommended if the sample units are arranged in systematic order such as chronological, geographical, alphabetical, etc. and also if the sample units in the universe are uniquely identified. Systematic sampling is also called sampling by regular intervals or sampling by fixed intervals.

- To get a systematic sample of size n from a population of size N, draw a random

number i from 1 to K, where $K = \frac{N}{n}$, and then select i, i + K, i + 2K, i + 3K, ...

In general, the ith element of the sample is
$$n_i = \left[i + \frac{N}{n} w \right]^{th} \text{ item. where } 0 \leq w \leq n - 1$$

Or we can have an alternative method,

$A_i = A_1 + (i - 1) K$. Where A_1 - the random starting point or the first sample item.

A_i - the i th item in the sample

Example 1: - From the files of 24 cases of the federal high court, the cases of only 4 of these is to be seen. The fifth file was selected randomly. Indicate the remaining three elements of the sample.

Solution: - $N = 24$, $n = 4$, $A_1 = 5$

$$K = \frac{N}{n} = \frac{24}{4} = 6$$

Then $A_2 = A_1 + (2 - 1) K$

$= A_1 + K = 5 + 6 = 11$. The 11th file is the second element

$A_3 = A_1 + (3 - 1) K$

$= A_1 + 2K = 5 + 2(6) = 17$. The 17th file is the third element.

$A_4 = A_1 + (4 - 1) K$

$= A_1 + 3K = 5 + 3(6) = 23$. The 23rd file is the fourth element.

Example 2 : - If the 4th and 12th elements of a systematic sample are 70 and 126 (in the population) respectively, then which item of the population is the first element of this systematic sample.

Solution: - $A_4 = 70$, $A_{12} = 126$, $A_1 = ?$

$$\begin{aligned} A_4 &= 70 = A_1 + 3K \\ A_{12} &= 126 = A_1 + 11K \\ \hline 56 &= 8K \rightarrow K = 7 \end{aligned}$$

taking these two simultaneous equations,

Then $A_4 = A_1 + 3K$

$$70 = A_1 + 3(7)$$

$$\rightarrow A_1 = 70 - 21 = 49$$

$\therefore$ The 49th item of the population is the random starting point for the systematic samples.

Non-random (non-probability) sampling method

In this method, the chance of including any elementary unit of the population in the sample cannot be determined. It is simple to adopt and no complicated procedure is needed to draw a sample.

There are many non-random sampling techniques. Some of which are Judgment, Convenient and Quota sampling.

Judgment Sampling: - The exercise of good perception and appropriate strategy are taken into account. Samples are selected deliberately by the investigator. It is a personal view. So it becomes satisfactory with regards to one's research needs. For example, if a sample of 10 students is to be singled out from a class of 50 for analyzing the habits of students, the investigator would select ten students, who in his opinion are representative of the class.

Convenient Sampling: - Elements of the sample are selected by taking those elements of the population, which are readily available or convenient for the investigator.

Example: Asking people from one area.

Quota sampling: - In this technique, quota is set up according to given criteria, but the sample within prescribed quota is selected by personal judgment of the investigator. It is suitable in market and public opinion surveys where stratification is very difficult. However, it suffers from representativeness as the interviewer may select samples convenient for him with regards to location and sample unit.

It is the combination of judgment and stratified sampling methods. so it enjoys the merits of both

Example: - If we ask about Canada dry for a prescribed quota of 20 households, 15 students and 10 children, then this method is quota sampling.

Mixed sampling method (cluster sampling)

Groups of items (clusters), homogenous in character, are formed on location or class basis. Here a sample of cluster is selected and next within cluster, sub groups are identified for inclusion in the sample. It is also known as Area sampling as selection of units is made on the basis of place. The clusters may or may not be equal in size. The smaller the size of the cluster, the greater will be the accuracy. It is economical and much easier.

1.4. SAMPLING DISTRIBUTION

1.4.1. Sampling Distribution of the Mean

A large number of units chosen at random from a large population tends to possess the characteristics of that population. If the sample selected is representative of the population, the results of the sample mean and the population mean are to be equal.

One may get doubt that the results of sample data may be different from that of universe. But it is obvious that sample form part of the population.

Sampling Distribution

Sample mean is an estimate of population mean that can be more accurate by taking large samples.

Sampling distribution of the mean is probability distribution of all possible sample means of a given size, selected from a population.

Steps to construct sampling distribution of the mean:

Example: - Suppose a baby sitter has 5 children under her supervision with average age of 6 years. The age of each child is given as follows.

Childr en	X1	X2	X3	X4	X5
Age	2	4	6	8	10

These 5 children constitute our entire population, so that $N = 5$ and.

$$\text{The population mean} = \frac{\sum_{i=1}^5 X_i}{N} = \frac{2+4+6+8+10}{5} = \frac{30}{5} = 6$$

$$\text{The standard deviation is given by the formula } \delta = \sqrt{\frac{\sum (X - \mu)^2}{N}}$$

$$\therefore \delta = \sqrt{\frac{(2-6)^2 + (4-6)^2 + (6-6)^2 + (8-6)^2 + (10-6)^2}{5}} = \sqrt{\frac{40}{5}} = \sqrt{8} = 2.83$$

Now, Let us assume the sample size, $n = 2$ and take all the possible samples of size 2, from this population.

$$\text{Since } \left[\frac{5}{2} \right] = \frac{5!}{2!(3!)} = \frac{5 \times 4}{2} = 10 \text{ there are 10 possible samples}$$

These are

X1, X2----- (2, 4)----- $\bar{X}_1 = 3$	X2, X3----- (4, 6)----- $\bar{X}_5 = 5$
X1, X3----- (2, 6)----- $\bar{X}_2 = 4$	X2, X4----- (4, 8)----- $\bar{X}_6 = 6$
X1, X4----- (2, 8)----- $\bar{X}_3 = 5$	X2, X5----- (4, 10)----- $\bar{X}_7 = 7$

$$X1, X5----- (2, 10)----- \bar{X}_4 = 6$$

$$X3, X4----- (6, 8)----- \bar{X}_8 = 7$$

$$X3, X5----- (6, 10)----- \bar{X}_9 = 8$$

$$X4, X5----- (8, 10)----- \bar{X}_{10} = 9$$

If only the first sample was taken, the average would be 3, if the second was taken, the average would be 4 and so on. All these samples independently taken are totally unrepresentative of the population. However, if a grand mean, μ_x or $\bar{\bar{X}}$, of the distribution of these sample means is taken, i.e.,

$$\mu_x = \frac{\sum_{i=1}^{10} X_i}{10} = \frac{3+4+5+6+5+6+7+7+8+9}{10} = \frac{60}{10} = 6$$

The grand mean has the same value as the population mean. Let's organize this distribution of sample means into a frequency distribution and probability distribution.

Sample Mean	Frequency	Relative Frequency	Probability
3	1	1/10	0.1
4	1	1/10	0.1
5	2	2/10	0.2
6	2	2/10	0.2
7	2	2/10	0.2
8	1	1/10	0.1
9	1	1/10	0.1
			1.00

The probability distribution of the sample mean is referred to us "Sampling distribution of the mean."

Accordingly, the sampling distribution of the means of the ages of children as tabulated above has three predicate patterns.

These are: -

1-The mean of the sampling distribution and the means of the population are equal. This can be shown as follows:

Sample Mean ($\bar{X}$)	Probability P ($\bar{X}$)
3	0.1
4	0.1
5	0.2
6	0.2
7	0.2
8	0.1
9	0.1
1.00	

The mean of the sampling distribution is given by $\mu = \sum \bar{X}.P(\bar{X})$

$$\begin{aligned}
 \text{Thus, } \mu &= \sum \bar{X}.P(\bar{X}) = (3 \times 0.1) + (4 \times 0.1) + (5 \times 0.2) + (6 \times 0.2) + (7 \times 0.2) + (8 \times 0.1) \\
 &\quad + (9 \times 0.1) \\
 &= 0.3 + 0.4 + 1.0 + 1.2 + 1.4 + 0.8 + 0.9 \\
 &= 6.0
 \end{aligned}$$

This value is the same as the mean of population

2- The spread of the sample means in the distribution is smaller than in the population values. For instance, the spread in the distribution of sample means above is from 3 to 9, while the spread in the population was from 2 to 10.

3- The shape of the sampling distribution of the means tends to be “Bell-shaped” and approximates the normal probability distribution, even when the population is not normally distributed.

The Central Limit Theorem

The Central Limit Theorem States that:

“Regardless of the shape of population distribution, the distribution of the sample means approaches the normal probability distribution as the sample size increases.”

The question is how large should the sample size be in order for the distribution of sample means to approximate the normal distribution for any type of population. In practice, the sample sizes of 30 or larger are considered adequate for this purpose. This should be noted however, the sampling distribution would be normally distributed if the original population is normally distributed, no matter what the sample size.

The discussion on the sampling distribution is concerned with the proximity to “a” sample mean to the population mean. It can be seen that the possible values of sample means tend towards the population mean, and according to central limit theorem, the distribution of sample means tends to be normal for a sample size of n being larger than 30.

~~Standard error of the mean~~ ($\delta_{\bar{X}}$): is a measure of dispersion of the distribution of sample means and is similar to the standard deviation in a frequency distribution and it measures the likely deviation of a sample mean from the grand mean of the sampling distribution.

$$\delta_{\bar{X}} = \sqrt{\frac{\sum (\bar{X} - \mu_{\bar{X}})^2}{N}}, \text{ Where } N = \text{number of sample means.}$$

For the above example of the sampling distribution of the ages of 5 children, the standard error of the means is:

$$\begin{aligned} \delta_{\bar{X}} &= \sqrt{\frac{\sum_{i=1}^7 (\bar{X}_i - \mu_{\bar{X}_i})^2}{7}} \\ &= \sqrt{\frac{(3-6)^2}{7}} \\ &= \sqrt{\frac{28}{7}} = \sqrt{4} = 2 \end{aligned}$$

However, since it is not possible to take all possible samples from the population, we must use alternative method to compute $\delta_{\bar{X}}$

If mean is given for a finite population,

$$\delta_{\bar{X}} = \frac{\delta}{\sqrt{n}} \cdot \sqrt{\frac{(N-n)}{N-1}} \quad \begin{array}{l} \text{Where } \delta = \text{Population standard deviation} \\ N = \text{Population size} \\ n = \text{Sample size.} \end{array}$$

Since we generally deal with very large population, which can be considered infinite,

$\sqrt{\frac{(N-n)}{(N-1)}}$ would approach 1.

$$\text{Hence, } \delta_{\bar{X}} = \frac{\delta}{\sqrt{n}}$$

- The factor $\sqrt{\frac{N-n}{N-1}}$ is known as the finite correction factor and should be used when population size is finite.

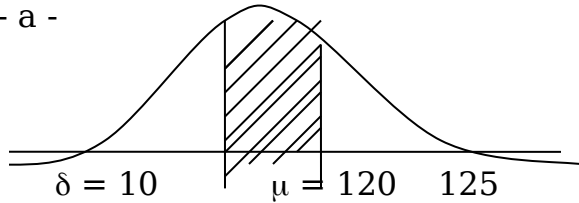
- Any single sample mean will become closer to the population mean, as the value of $\delta_{\bar{X}}$ decreases.

Example: The IQ scores of College students are normally distributed with the mean of 120 and standard deviation of 10.

a - what is the probability that the IQ score of any one student chosen at random is between 120 and 125?

b - If a random sample of 25 students is taken, what is the probability that the mean of this sample will be between 120 and 125?

Solution: - a -



Using the standardized normal distribution formula,

$$Z = \frac{(X - \mu)}{\delta} = \frac{125 - 120}{10} = \frac{5}{10} = 0.5$$

From the table, the area for $Z = 0.5$ is 0.915.

∴ There is a 19.15 % chance that a student picked up randomly will have an IQ score between 120 and 125.

b -
$$\delta_{\bar{x}} = \frac{\delta}{\sqrt{n}} = \frac{10}{\sqrt{25}} = \frac{10}{5} = 2$$

$$\text{Then } Z = \frac{X - \mu}{\delta_{\bar{x}}} = \frac{125 - 120}{2} = \frac{5}{2} = 2.5$$

The area for $Z = 2.5$ is 0.4938.

∴ There is a chance of 49.38 % that the sample mean will be between 120 and 125.

- As the sample size increases further, this chance will also increase.

- It can be noted that the probability of a sample mean being between 120 and 125 is much higher than the probability of an individual student having an IQ between 120 and 125.

CHAPTER 2: STATISTICAL ESTIMATION AND TEST OF HYPOTHESIS

Introduction to Statistical Estimation

Managers in business, education, social work, and other fields make decisions without complete information. Automobile manufacturers do not know exactly how many people will purchase new cars next year. The college registrar does not know exactly how many students will enroll next fall, but based on the past experience may lay down an estimate plan. Everyone makes estimates. When you get ready to cross a street, you estimate the speed of the car that is approaching, the distance between you and the car

and your own speed. Having made these quick estimates, you decide whether to wait, walk or run. In such decisions without complete information, there is a considerable uncertainty.

In statistical inference, one estimates about the population based on the result obtained from the sample selected from that population. Thus, estimation is a process by which we estimate various unknown population parameters from sample statistics.

Any sample statistic that is used to estimate a population parameter is called an estimator and an estimate is a numerical value of an estimator.

The sample mean is often used as an estimator of the population mean. Suppose that we calculate the mean daily revenue of a store for a random sample of 6 days and find it to be 1110 birr. If we use this value to estimate the daily revenue for the whole year, then the value 1110 birr would be an estimate.

There are two kinds of estimates, point estimates and interval estimates.

2.1. POINT AND INTERVAL ESTIMATION

2.1.1. POINT ESTIMATION

Point estimation is a statistical procedure in which we use a single value to estimate a population parameter. A point estimate is a single number that is used as an estimate of a population parameter, and is derived from a random sample taken from the population of interest.

Some of the most important point estimators are given below:

Population parameter	Point estimator
Mean, μ	$\bar{X} = \frac{\sum X_i}{n}$
Variance, σ^2	$S^2 = \frac{\sum (X_i - \bar{X})^2}{n-1}$
Standard deviation, σ	$S = \sqrt{S^2}$
Proportion, π	$P = \frac{X}{n}$

Example 19.1 To set the price of a product, one strategy is competition-oriented in which you fix the price of your product at the average level charged by other producers. Suppose you want to market a 200-gram bar of soap that you produce. The current wholesale prices charged by a random sample of 10 soap producers (in birr) are:

1.00 1.35 1.50 0.95 0.90 1.25 1.00 1.20 0.90 and 1.50

What is an estimate of the mean wholesale price charged by all soap producers? Find an estimate of the standard deviation in the wholesale prices of all the producers?

Solution: - The mean wholesale price or the population mean (μ) is estimated by the

sample mean $\bar{X}$, given by $\bar{X} = \sum_i x_i / n = (1.00 + 1.35 + \dots + 1.50) / 10 = 1.155$

Thus, an estimate of the mean wholesale price charged by all soap producers is 1.155 Birr. Based on this information, you might set the wholesale price per unit of your product at 1.155 Birr.

The standard deviation in the wholesale prices of all producers, what we call the population standard deviation (δ) and is estimated by the sample standard deviation.

$$= \sqrt{\frac{\sum_i (x_i - \bar{X})^2}{n-1}} = \sqrt{\frac{(1.00-1.155)^2 + (1.35-1.155)^2 + \dots + (1.50-1.155)^2}{9}} \\ = 0.237$$

Thus, the wholesale prices fluctuate below and above their mean by about 0.237 Birr, which is an estimate of the standard deviation in the wholesale prices of all producers.

Example 19.2: - suppose you are interested to know the proportion of fishes that are inedible as a result of chemical pollution of a certain lake. In a random sample of 400 fishes caught from this lake, 55 were found out to be inedible. Out of all fishes in this lake, what is an estimate of the proportion of inedible fishes?

Solution: -

The proportion of inedible fishes in the entire lake is what we call population proportion (Π). This is estimated by the sample proportion:

$$P = \frac{X}{n} = \frac{55}{400} = 0.1375 = 13.75 \text{ percent.}$$

Although point estimates are often useful, they do have one serious drawback: we do not know how close or far these values are from the population value they are supposed to estimate, and hence, we can not be certain of their reliability. In other words, a point estimate will be more useful if it is accompanied by an estimate of the error that might be involved. To this end, we use interval estimation.

2.1.2. INTERVAL ESTIMATION

Interval estimation is a statistical procedure in which we find a random interval with a specified probability of containing the parameter being estimated. An interval estimate is an interval that provides an upper bound and a lower bound for a specific population parameter whose value is unknown. This interval estimate has an associated degree of confidence of containing the population parameter. Such interval estimates are also called Confidence intervals and are calculated from random samples.

The interval estimate is an interval that includes the point estimate. For example, if the sample mean is say 0.28, one may report that the population mean is in the range of 0.25 and 0.31 with a probability of 0.95. i.e. the 95 percent confidence interval of the population mean is (0.25, 0.31). Clearly this interval contains the point estimated 0.28.

Confidence Interval for the Population MEAN (μ)

Case I. Sampling from a normally distributed population with known variance δ

Recall that Z_α denotes the value of Z for which the area under standard normal curve to its right is equal to α . Analogously, $Z_{\alpha/2}$ denotes value of Z for which the area to its right $\alpha/2$ and, $Z_{\alpha/2}$ denotes the value for which the area to its left is $\alpha/2$.

Consider the following figure

From the above figure we have:

$$P(-Z_{\alpha/2} < Z < Z_{\alpha/2}) = 1 - \alpha$$

But we know that $Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$ follows standard normal distribution. Thus

$$\begin{aligned} P\left(-Z_{\alpha/2} < \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} < Z_{\alpha/2}\right) &= 1 - \alpha \\ \Rightarrow P(-Z_{\alpha/2} \cdot \sigma/\sqrt{n} < \bar{X} - \mu < Z_{\alpha/2} \cdot \sigma/\sqrt{n}) &= 1 - \alpha \\ \Rightarrow P(\bar{X} - Z_{\alpha/2} \cdot \sigma/\sqrt{n} < \mu < \bar{X} + Z_{\alpha/2} \cdot \sigma/\sqrt{n}) &= 1 - \alpha \end{aligned}$$

Thus, a $(1 - \alpha)$ 100% confidence interval for the population mean μ is given by:

$$\bar{X} \pm Z_{\alpha/2} \sigma/\sqrt{n}$$

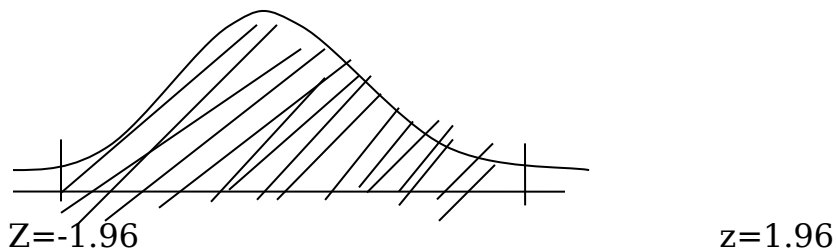
Where $\bar{X}$ is the sample mean, $Z_{\alpha/2}$ is the value of Z for which the area to its right is $\frac{\alpha}{2}$. Common confidence intervals are the 95 percent and the 99 percent confidence

intervals. The 95 percent confidence interval means that about 95 percent of the similarly constructed intervals will contain the parameter being estimated. If we use the 99 percent level of confidence, then we expect about 99 percent of the intervals to contain the parameter being estimated.

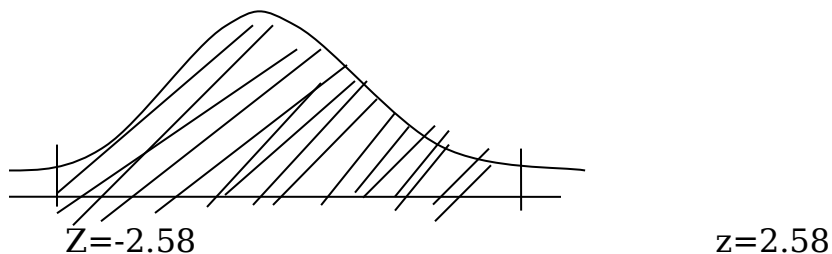
Another interpretation of the 95 percent confidence interval is that 95 percent of the sample means for a specified sample size will be within 1.96 standard deviations of the hypothesized population mean. Similarly, for a 99 percent confidence interval, 99 percent of the sample means will lie within 2.58 standard deviations of the hypothesized population mean.

If $\alpha = 0.05$, then the $(1 - \alpha)$ 100 percent confidence interval, which is the $(1 - 0.05)$ 100 = 95 percent confidence interval and if $\alpha = 0.01$, then the $(1 - \alpha)$ 100 percent confidence interval will be the $(1 - 0.01)$ 100 % which is the 99 % confidence interval. Where α is called the confidence coefficient.

If $\alpha = 0.05$, then $Z_{\alpha/2} = Z_{0.025} = 1.96$ and



If $\alpha = 0.01$, then $Z_{\alpha/2} = Z_{0.005} = 2.58$



* The total area under the normal curve is 1. or one can report as,

95 % of the area under the standard normal curve is between Z value - 1.96 and 1.96 and similarly 99 % of the area under the standard normal curve is between Z value - 2.58 and 2.58.

Thus, the 95 percent confidence interval of the mean for known standard deviation σ is given by,

$$\bar{X} \pm 1.96 \frac{\sigma}{\sqrt{n}} \quad \text{and the 99 \% confidence interval is given by} \quad \bar{X} \pm 2.58 \frac{\sigma}{\sqrt{n}}$$

If the population standard deviation is not known, then we approximate the population standard deviation by the sample standard deviation S given by:

$$S = \frac{\sqrt{\sum (X_i - \bar{X})^2}}{n-1}$$

Then the 95 % confidence interval is given by

$$\bar{X} \pm 1.96 \frac{S}{\sqrt{n}}$$

And the 99 % confidence interval is given by

$$\bar{X} \pm 2.58 \frac{S}{\sqrt{n}} \quad \text{Where}$$

$\bar{X}$ - Sample mean,

S - sample standard deviation

2.58 is $Z_{\alpha/2}$

$\sqrt{n}$ - the root of the sample size.

Example 19.3 In a certain small city, to estimate the mean monthly expenditure for food, a random sample of 25 households was randomly selected yielding a mean of 200 birr. From experience, it is known that such expenditures are normally distributed with a standard deviation of 50 Birr.

What is the point estimate of the mean monthly expenditures for food of all households in the city?

Find a 95 percent confidence interval for the mean monthly expenditures for food of all households in the city.

Given

$$\bar{X} = 200 \text{ Birr}$$

$$\sigma = 50 \text{ Birr}$$

$$n = 25$$

answer = (180.40 Birr, 219.60 Birr)

i.e. we are 95 percent confident that the true mean monthly expenditure for food (μ) is between 180.40 Birr and 219.60 Birr.

Example 19.4

A manufacturer claims that his tyre lasts 20,000 miles on average. A consumer organization tests a random sample of 64 tyres and reported an average of 19,200 miles with a standard deviation of 2,000 miles. Does a 99 % confidence interval for the mean life of all tyres produced by the manufacturer supports the claim?

Given: $n = 64$, $\bar{X} = 19,200$ miles, $S = 2000$ miles. Though we have no information about the normality of the population by central limit theorem, for large n , say $n \geq 30$. We assume that the distribution is normal. In our case as $n = 64 \geq 30$ then we consider the normality.

Then for 99 % confidence interval, $\alpha = 0.01$ and $\alpha/2 = 0.005$

And from the table of standard normal,

$$Z_{\alpha/2} = Z_{0.005} = 2.58$$

Thus, A 99 % confidence interval for the mean (μ) will be:

$$\begin{aligned}\bar{X} \pm Z_{\alpha/2} S/\sqrt{n} \\&= 19,200 \pm (2.58) (2000 / \sqrt{64}) \\&= 19,200 \pm 645 \\&= (18,555 \text{ miles}, 19,845 \text{ miles})\end{aligned}$$

Hence, we are 99 percent confident that the true mean mileage is at most 19,845. Which is less than the claimed mean 20,000 miles. Therefore, the claim is not true.

Case II. Small sample confidence interval for the population mean: Sampling from a normally distributed population with σ^2 unknown and $n < 30$.

If the population variance σ^2 is not known, then it must be estimated by the sample variance S^2 as,

$$S^2 = \frac{\sum (X_i - \bar{X})^2}{n-1}$$

Under this situation, since σ^2 is estimated by S^2 , the sampling distribution of the mean deviates from the Normal distribution for small size, or we say the sampling distribution of $\bar{X}$ follows the students t distribution with $n - 1$ degrees of freedom.

For $n > 30$, the student t distribution can be approximated by the Normal distribution.

Like the Normal distribution, the t-distribution is symmetrical about the mean = 0. But it is flatter as compared to the Normal distribution. However, as the sample size increases the t-distribution loses its flatness and becomes approximately Normal.

The shape of the t-distribution is determined by the degrees of freedom. Degrees of freedom can be defined as the number of values we can choose freely. Suppose we are dealing with a sample of size $n = 6$, and we know the mean of these 6 numbers is 5. Symbolically, we have:

$$\frac{a+b+c+d+e+f}{6} = 5$$

Now, we are free to assign any value to a, b, c, d and e,

Say $a = 3$, $b = 2$, $c = 4$, $d = 5$ and $e = 3$. But, we are no more free to assign a value to f since:

$$\begin{aligned} \frac{a+b+c+d+e+f}{6} = 5 &\Rightarrow \frac{17+f}{6} = 5 \Rightarrow 17+f = 30 \\ &\Rightarrow f = \underline{\underline{13}} \end{aligned}$$

That is, in order for the mean of these 6 numbers to be 5, f must be 13. If we assign another number for f, then the mean will not be equal to 5. Thus, we are free to choose only 5 values and the 6th one is determined automatically.

Hence, the degrees of freedom is:

$$n - 1 = 6 - 1 = 5$$

Generally, for a sample of size n, the degree of freedom is $n - 1$. The values of t for different degrees of freedom and different values of α are tabulated. $t_{\alpha}(n - 1)$ denotes the value of t for which the area under the curve to its right is equal to α with $(n - 1)$ degrees of freedom.

Example 19.5 a) for $n = 20$ and $\alpha = 0.025$, find

$$t_{\alpha}(n - 1)$$

Solution:

From the t-distribution table, $t_{0.025}(19) = 2.093$ (shaded area = 0.025)

$$\text{b) If } n = 26, \quad \alpha = 0.005$$

$$\text{then } t_{\alpha}(n - 1) = t_{0.005}(25) = 2.787$$

(from the table of t-distribution)

Under such situations, a $(1 - \alpha) 100 \%$ Confidence interval for the population mean μ is given by:

$$\bar{X} \pm t_{\alpha/2} (n-1) S/\sqrt{n}$$

Example 19.6

One measure of a company's financial health is its debt-to equity ratio. This quantity is defined to be the ration of the company's corporate debt to the company's equity. If this ratio is too high, it is one indication of financial instability. For obvious reasons, banks often monitor the financial health of companies to which they have extended commercial loans. Suppose that, in order to reduce risk, a large bank has decided to initiate a policy limiting the mean debt-to- equity ratio for its portfolio of commercial loans to 1.5. In order to estimate the mean debt-to-equity ratio of its loan portfolio, the bank randomly select a sample of 15 of its commercial loan accounts. Audits of these companies result in the following debt-to-equity ratios:

1.31	1.05	1.45	1.21	1.19
1.78	1.37	1.41	1.22	1.11
1.48	1.33	1.29	1.32	1.65

A stem-and-leaf display of these ratios is reasonably mound shaped. Furthermore, the sample mean and standard deviation of these ratios can be calculated to be $\bar{X} = 1.343$ and $S = 0.192$

Suppose that the bank wishes to calculate a 95% confidence interval for a loan portfolio's mean debt-to-equity ratio, μ . Since the bank has taken a small sample of size 15, it is appropriate to calculate an interval based on the t distribution. We have $n - 1 = 15 - 1 = 14$ degrees of freedom, and the level of confidence $100 (1 - \alpha)$ percent = 95 percent implies that $\alpha = 0.05$. Therefore, we use the t point $t_{\alpha/2} = t_{0.05 / 2} = t_{0.025} = 2.145$ (from, the table). It follows that the 95 percent confidence interval for μ is

$$\begin{aligned} \left(\bar{X} \pm t_{.025} \frac{S}{\sqrt{n}} \right) &= \left[1.343 \pm 2.145 \left(\frac{.192}{\sqrt{15}} \right) \right] \\ &= [1.343 \pm 0.106] \\ &= [1.237, 1.449] \end{aligned}$$

This interval says that the bank is 95 percent confident that the mean debt-to-equity ratio for its portfolio of commercial loan accounts is between 1.237 and 1449. Based on

this interval, the bank has strong evidence that the portfolio's mean ratio is less than 1.5 (or that the bank is in compliance with its new policy).

2.2. DETERMINING THE SAMPLE SIZE IN ESTIMATION

Whenever we take a sample for inferential purposes, there is always a sampling error. This sampling error is controlled by selecting a sample that is adequate in size. If the sample size is small, then we may fail to achieve the objective of our analysis, and if it is too large, then we waste the resources when we gather the sample.

When we estimate the population mean μ by the sample mean $\bar{X}$, with probability $(1 - \alpha)$ the maximum error E will be:

$$E = Z_{\alpha/2} \sigma / \sqrt{n} \quad \text{if } \sigma \text{ is known}$$

$$E = Z_{\alpha/2} S / \sqrt{n} \quad \text{if } \sigma \text{ is not known}$$

With probability $(1 - \alpha)$, the sampling error will not exceed some prescribed quantity E if the sample size is at least:

$$n = \left[\frac{Z_{\alpha/2} \sigma}{E} \right]^2$$

If n comes out fractional, round up to the next integer.

Example 19.7. The owner of a chain of hotels wants to determine the mean number of rooms occupied per day (so that he can have an estimate of the average daily revenue obtained by renting rooms). From past records, the standard deviation of the daily occupancy is known to be 9 rooms.

How large a sample of days should be taken so that the true mean number of rooms occupied per day will not differ from the sample mean by more than 3 rooms at the 95 percent confidence level?

At the 99 percent confidence level, what is the maximum error committed in estimating the true mean by the sample mean if a random sample of 64 days is taken?

Solution: -

Given $\sigma = 9$ rooms

$E = 3$ rooms, $(1 - \alpha) 100 \% = 95 \% \Rightarrow \alpha = 0.05$

$$\Rightarrow Z_{\alpha/2} = Z_{0.025} = 1.96$$

$$\therefore n = \left(\frac{Z_{\alpha/2} \sigma}{E} \right)^2 = \left(\frac{1.96 \times 9}{3} \right)^2 = 345.744$$

Therefore, a sample of size at least 35 days is required.

$$n = 64, (1 - \alpha) 100\% = 99\% \Rightarrow \alpha = 0.01$$

$$\Rightarrow Z_{\alpha/2} = Z_{0.005} = 2.58$$

$$E = Z_{\alpha/2} \sigma / \sqrt{n} = (2.58) \left(\frac{9}{\sqrt{64}} \right) = 2.9$$

Therefore, if we use a random sample of 64 days, then we are 99% certain that the error in estimation will not exceed 2.9 rooms. I.e. the difference between the average daily occupancy computed from the sample and the true average daily occupancy will not exceed 2.9 rooms.

Interval estimate - The interval, within which a population parameter probably lies, based on sample information.

Point estimate - A single number computed from a sample and used to estimate a population parameter.

Sampling error - The difference between a sample statistic and its corresponding population parameter of containing the population parameter is called Confidence Interval.

CHAPTER 3: HYPOTHESIS TESTING

3.0. 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.

3.1. 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.

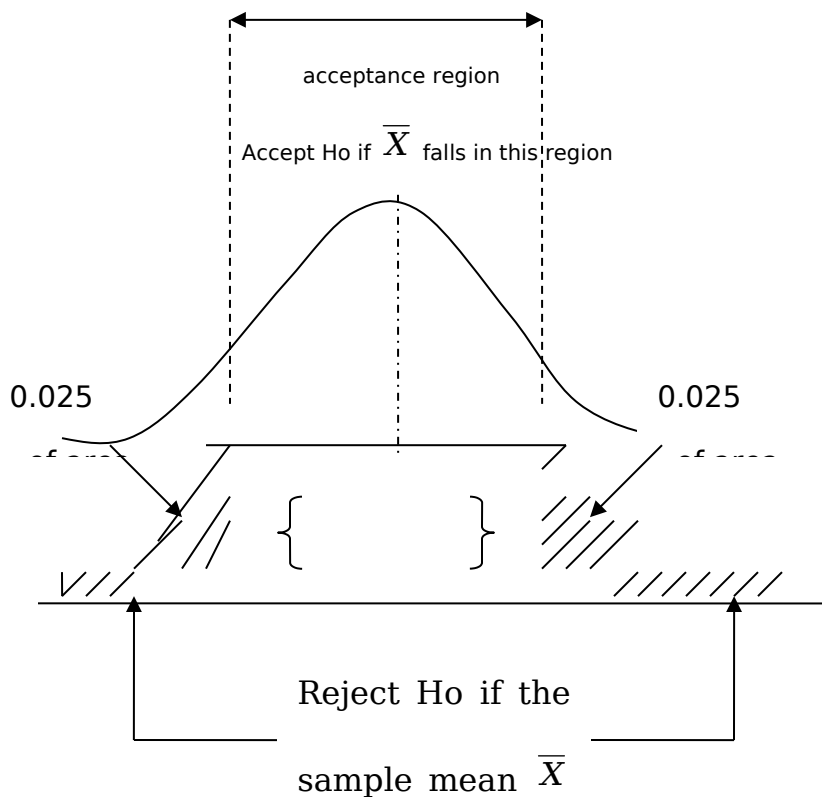
A hypothesis which is to be actually tested for acceptance or rejection is termed as null hypothesis. It is denoted by H_0 and alternative hypothesis which is denoted by H_1 or H_A is a statement about the population parameter or

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 H_0 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 H_0 . 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. $\alpha = 0.01$ is used for high precision and $\alpha = 0.05$ for moderate precision.
- c) Decision rule or test of hypothesis: - Given a hypothesis H_0 and an alternative hypothesis H_a , we make a rule which is known as decision rule according to which we accept H_0 (i.e., reject H_a) or reject H_0 (i.e., accept H_a). For instance, if H_0 is that a certain lot is good (there are very few defective items in it) against H_a 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 H_0 otherwise we will reject H_0 (or accept H_a). 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 H_0 when H_0 is actually true and we may accept H_0 when in fact H_0 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 $H_0: \mu = \mu_0$ and $H_a: \mu \neq \mu_0$ which mean $\mu > \mu_0$ or $\mu < \mu_0$. Thus, in a two-tailed test, there are two rejection regions, one on each tail of the curve which can be illustrated as under:

Acceptance and rejection regions in case of two-tailed test with 5% significance level in Z- statistic

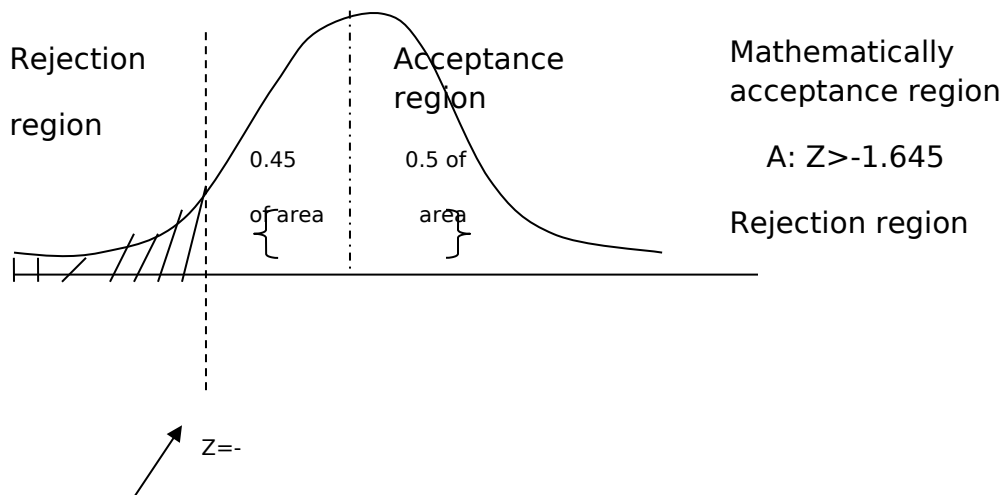


Mathematically we can state:

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

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 $H_0: \mu$

$= \mu_0$ and $H_a: \mu < \mu_0$, then we are interested in what is known as Left-tailed test (wherein there is one rejection region only on the left tail) which can be illustrated as below

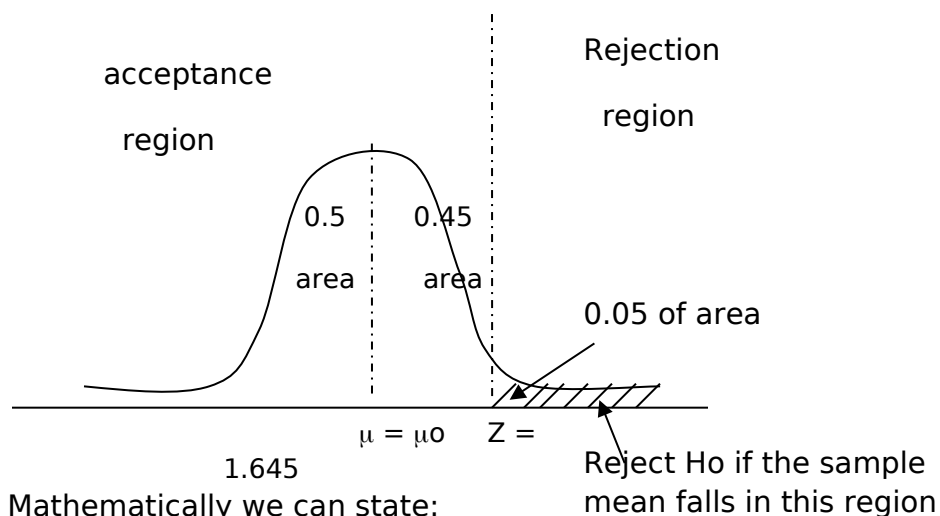


Reject H_0 if the sample $Z = -1.64$ $\mu = \mu_0$

Mean $\bar{X}$ falls here

In case our $H_0: \mu = \mu_0$ and $H_a: \mu > \mu_0$, 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.

Acceptance and rejection regions in case of one-tailed test (right tail) with 5% significance level in Z -statistic



Mathematically we can state:

Acceptance Region A: $Z \leq 1.645$

3.2. 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 infinite, sample size may be large or small but variance of the population is known, H_a 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:

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

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

μ_0 - is the hypothesized population mean under H_0 .

σ_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:

$H_0: \mu = 67.39''$

$H_a: \mu \neq 67.39''$

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

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

As H_a 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 H_0 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

$$H_a: \mu < \mu_0 = 67.39''$$

$$\text{Then } R: Z < -Z_{\alpha}$$

$$Z_{cal} < -1.645$$

i.e compare 1.231 with -1.645 as $Z_{cal} > -1.645$ then H_0 is accepted

If the alternative hypothesis is

$H_a: \mu > \mu_0 = 67.39''$ then compare $Z_{cal} = 1.231$ with $Z_{\alpha} = 1.645$ in this case the rejection region $R: Z > Z_{\alpha}$ but as $1.231 < 1.645$, again H_0 remains accepted.

Check Your Progress -1

1. The following information is available

$$H_0: \mu = 50$$

$$H_1: \mu \neq 50$$

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 H_0 ?

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

In such a situation Z-test is used and the test statistic Z is worked out as under (using a finite population correction):

$$Z = \frac{\bar{X} - \mu_0}{(\sigma_p / \sqrt{n}) \sqrt{\frac{N-n}{N-1}}}$$

Example: - Suppose we are interested in a population of 20 industrial units of the same size, all of which are experiencing excessive labor turnover problems. The past records show that the mean of the distribution of annual turnover is 320 employees, with a standard deviation of 75 employees. A sample of 5 of these industrial units is taken at random which gives a mean of annual turnover as 300 employees. Is the sample mean consistent with the population mean? Test at 5% level.

Solution: - Taking the null hypothesis that the population mean is 320 employees, we can write:

$H_0: \mu = 320$ employees

$H_a: \mu \neq 320$ employees

And the given information as under:

$\bar{X} = 300$ employees, $\sigma_p = 75$ employees

$$n = 5; \quad N = 20$$

Assuming the population to be normal, we can workout the test statistic Z as under:

$$Z = \frac{\bar{X} - \mu_0}{(\sigma_p / \sqrt{n}) \sqrt{\frac{N-n}{N-1}}} \\ = \frac{300 - 320}{75 / \sqrt{5} \sqrt{\frac{20-5}{20-1}}} = -0.67.$$

As H_a is two-sided in the given question, we shall apply 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 is -0.67 which is in the acceptance region since $R: |z| > 1.96$ and thus, H_0 is accepted and we may conclude that the sample mean is consistent with population mean i.e., the population mean 320 is supported by sample results.

3. Population normal, population infinite, sample size small and variance of the population unknown, H_a 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:

$$t = \frac{\bar{X} - \mu_0}{\sigma_s / \sqrt{n}} \text{ with degree of freedom } = (n - 1)$$

$$\text{and } \sigma_s = \text{sample variance} = \sqrt{\frac{\sum_i (X_i - \bar{x})^2}{n - 1}}$$

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.

$$H_0: \mu \leq \mu_0 = 305$$

$$H_1: \mu > \mu_0$$

Sample variance $\sigma_s = 12$, sample mean $\bar{X} = 311$

$$\text{Test statistics: } t = \frac{\bar{X} - \mu_0}{\sigma_s / \sqrt{n}}$$

$$\begin{aligned} &= \frac{311 - 305}{12 / \sqrt{20}} \\ &= 2.236 \end{aligned}$$

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 $t_{0.05} (19) = 1.729$

as $t = 2.236 > 1.729$ is in the rejection region, then reject H_0 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 $\mu < \mu_0$, (left tailed) test one should compare t calculated with $-t_{\alpha}(n-1)$ i.e. the rejection region is $R: t < -t_{\alpha}(n-1)$

4. Population normal, population finite, sample size small and variance of the population unknown, and H_a may be one-sided or two-sided:

In this case, t- test is used and the test statistic 't' is worked out as under (using finite population multiplier):

$$t = \frac{\bar{X} - \mu_0}{(\sigma_s / \sqrt{n}) \sqrt{\frac{N-n}{N-1}}} \text{ with d.f} = (n-1) \text{ and}$$

$$\sigma_s = \sqrt{\frac{\sum (x_i - \bar{x})^2}{n-1}}$$

5. Population may not be normal but sample size is large, variance of the population may be known or unknown, and H_a 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:

$$Z = \frac{\bar{x} - \mu_0}{\sigma_p \sqrt{n}} \text{ in case of infinite}$$

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

And for a finite population, using the finite population correction for the standard error, we use the modified statistic Z as:

$$Z = \frac{\bar{x} - \mu_0}{(\sigma_p / \sqrt{n}) \sqrt{\frac{N-n}{N-1}}} \text{ for known finite population variance } \sigma_p \text{ and replacing } \sigma_p \text{ by } \sigma_s \text{ for if the population variance } \sigma_p \text{ is not known.}$$

The decision rule is similar to those examples given earlier for one-tailed and two-tailed tests using area under standard normal distribution.

3.3. Testing a hypothesis about a population proportion

Some assumptions must be made and conditions met before testing a population proportion. To test a hypothesis about a population proportion, a random sample is chosen from the population. This is called the experiment. It is assumed that the binomial assumptions are met: (1) the sample data collected are result of counts; (2) an outcome of an experiment is classified into one of two mutually exclusive categories - a

“success” or a “failure”; (3) the probability of a success stays the same for each trial; and (4) the trials are independent, meaning that the outcome of one trial does not affect the outcome of any other trial.

The test we will conduct shortly is appropriate when both np and $n(1-p)$ are greater than 5. n is the sample size, and p is the population proportion. In addition, n should be large. Researchers disagree on how large n should be; some say 30, some say 50, and others say 100. For the purpose of this book, the sample size should be at least 50. The distribution of sample proportions is approximately Normally distributed by the central limit theorem for sufficiently large sample size. Thus, we can use the Normal distribution table to find the critical values for testing hypothesis about the population proportion (true proportion). This test is a good example of the case where in the normal probability distribution is applied to approximate a binomial probability distribution with a great deal of accuracy.

If P_o is the assumed or hypothesized population proportion, then the test statistic will be:

$$Z_{cal} = \frac{p - p_o}{\sqrt{\frac{p_o(1 - p_o)}{n}}}$$

Where P is the sample proportion, and n is the sample size

Example: - Suppose prior elections in a state indicated that it is necessary for a candidate for governor to receive at least 50 percent of the vote in the northern section of the state to be elected. The incumbent governor is interested in assessing his chances of returning to office and plans to conduct a survey consisting of 2000 registered voters in the northern section of the state. If the sample survey of the 2000 potential voters in the northern part of the state revealed that 1, 500 planned to vote for the incumbent governor, will the incumbent governor be re elected? Test at 5% level of significance.

Solution: - as $n.p$ and $n(1-p)$ exceed 5 and n is greater than 50, the following test of hypothesis can be conducted.

$$H_0: P \geq .80$$

$$H_1: P < .80$$

As H_1 states a direction, the test is one-tailed.

Z is an appropriate statistic, found by

$$Z = \frac{p - p_o}{\sqrt{\frac{p_o(1 - p_o)}{n}}} \quad \text{where } P \text{ is the sample proportion } P = \frac{1500}{2000} = \frac{3}{4} = 0.75$$

and P_o is the population proportion = 0.8

$$\text{Thus } Z_{cal} = \frac{0.75 - 0.8}{\sqrt{\frac{0.8(1 - 0.8)}{2000}}} = \frac{-0.05}{0.00894} = -5.5$$

and $-Z_{\alpha}$ tabulated, $-Z_{0.05} = -1.645$

the rejection region $R: Z < -Z_{\alpha}$ and since $Z_{cal} = -5.5 < -1.645$ which is in the rejection region, the null hypothesis is rejected at 0.05 level. i.e. The difference of 5% points between the sample percent (75%) and the hypothesized population percent 80% is statistically significant.

It is probably not due to sampling variation. To put it another way, the evidence at this point does not support the claim that the incumbent governor will return to the governor's mansion for another five years.

The two-tailed and right-tailed tests can be conducted accordingly for population proportions.

CHAPTER 4: CHI-SQUARE TESTS

4.1 Introduction

In earlier units, we considered tests for parametric hypothesis. In developing such tests, we made the assumption that the parent distribution is of specified nature (in most cases it was assumed that the distribution is of the normal type). We shall now consider hypothesis of more fundamental nature, where these assumptions themselves are questioned and one seeks to verify them on the basis of sample observations.

Chi-square is an important non-parametric test and as such no rigid assumptions are necessary in respect of the type of population. We require only the degrees of freedom (implicitly of course the size of the sample) for using this test.

4.2 Test of goodness of fit

Chi-square test of goodness of fit enables us to see how well does the assumed theoretical distribution (such as Binomial distribution, Poisson distribution or normal distribution) fit to the observed data. When some theoretical distribution is fitted to the given data, we are always interested in knowing as to how well this distribution fits with the observed data. The chi-square test can give answer to this. If the calculated value of χ^2 is less than the table value at a certain level of significance, the fit is considered to be a good one, which means that the divergence between the observed and expected frequencies is attributable to fluctuations of sampling. But if the calculated value of χ^2 is greater than its table value, the fit is not considered to be a good one.

The test statistics for test of goodness of fit is given by:

$$\chi^2 = \sum \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

Where,

O_{ij} = observed frequency of the cell in the i^{th} row and j^{th} column.

E_{ij} = Expected frequency of the cell in i^{th} row and j^{th} column.

If two distributions (observed and theoretical) are exactly alike, $\chi^2 = 0$; but due to sampling errors, χ^2 is not equal to zero and as such we must know the sampling distribution of χ^2 so that we may find the probability of an observed χ^2 being given by a random sample from the hypothetical universe. Instead of working out the probabilities, we can use ready table, which gives probabilities for given values of χ^2 . Whether or not a calculated value of χ^2 is significant can be ascertained by looking at the tabulated values of χ^2 for given degrees of freedom at a certain level of significance. If the calculated value of χ^2 is equal to or exceeds the table value, the difference between the observed and expected frequencies is taken as significant, but if the table value is more than the calculated value of χ^2 , then the difference is considered as insignificant i.e., considered to have arisen as a result of chance and as such can be ignored.

Degrees of freedom play an important part in using the chi-square distribution and the test based on it. Thus, one must correctly determine the degrees of freedom. If there are 10 frequency classes and there is one independent constraint, then there $(10 - 1) = 9$ degrees of freedom. Thus, if 'n' is the number of groups and one constraint is placed by making the totals of observed and expected frequencies equal, the d.f. would be equal to $(n - 1)$. In the case of contingency table (i.e., a table with 2 columns and 2 rows or a table with two columns and more than two rows or a table with two rows but more than two columns or a table with more than two rows and more than two columns), the d.f. is worked out as: $df = (c - 1)(r - 1)$ where 'c' means the number of columns and 'r' means the number of rows. Now to test the goodness of fit using the statistic

$$\chi^2 = \sum \left[\frac{(f_o - f_e)^2}{f_e} \right] \text{ With } K - 1 \text{ degrees of freedom where } K \text{ is the number of categories, and}$$

f_o is an observed frequency in a particular category,

f_e is an expected frequency in a particular category,

One can follow the following steps:

1. Calculate the difference between each paired f_o and f_e .
2. The differences are squared
3. Each squared difference is divided by the appropriate expected frequency, f_e . These quotients are then summed. The sum is the computed value of chi-square.

Before starting the application of chi-square distribution, it is better to see conditions for the application of χ^2 - test.

If there is an unusually small-expected frequency in a cell, chi-square (if applied) might result in an erroneous conclusion. This can happen because f_e appears in the denominator, and dividing by a very small number makes the quotient quite large!. Two generally accepted rules regarding small cell frequencies are:

- 1) If there are only two cells, the expected frequency in each cell should be 5 or more.
- 2) For more than two cells, χ^2 should not be applied if more than 20% of the f_e cells have expected frequencies less than 5.

These two problems can be overcome by combining some of the categories.

Example: - A die is thrown 132 times with the following results:

Number turned up	1	2	3	4	5	6
frequency	16	20	25	14	29	28

Is the die unbiased?

Solution: - Let us take the hypothesis that the die is unbiased. If that is so, the probability of obtaining any one of the six numbers is $1/6$ and as such the expected frequency of any one number coming upward is $132 \times \frac{1}{6} = 22$. Now we can write the observed frequencies along with expected frequencies and work out the value of χ^2 as follows:

No	of	observed	expected		
Turned up	frequency	frequency	$(f_o - f_e)$	$(f_o - f_e)^2$	$(f_o - f_e)^2 / f_e$
	(f_o)	(f_e)			
1	16	22	-6	36	36/22

2	20	22	-2	4	4/22
3	25	22	3	9	9/22
4	14	22	-8	64	64/22
5	29	22	7	49	49/22
6	28	22	6	36	36/22

$$\sum_i \frac{(f_o - f_e)^2}{f_e} = 9$$

Hence the calculated value of $\chi^2 = 9$

: - Degrees of freedom in the given problem is $(2-1)(6-1) = 1 \times 5 = 5$.

The table value of χ^2 for 5 degrees of freedom at 5% level of significance is 11.071. Comparing calculated and table values of χ^2 , we find that calculated value is less than the table value and as such could have arisen due to fluctuations of sampling. The result, thus, supports the hypothesis and it can be concluded that the die is unbiased.

Example: - Genetic theory states that children having one parent of blood type A and the other of blood type B will always be of one of the three types A, AB, B and that the proportion of the three types will on an average be as 1:2:1. A report states that out of 300 children having one A parent and B parent, 30% were found to be type A, 45% type AB and the remaining type B. Is the assumed proportion 1:2:1 true?

Solution: - the observed frequencies of type A, AB and B as given in the question are 90, 135 and 75 respectively which is obtained by calculating 30%, 45% and 25% of 300.

The expected frequencies of type A, AB and B (as per the genetic theory) should have been 75, 150 and 75 respectively. (1:2:1 of 300) we now calculate the value of χ^2 as follows:

Type	Observed frequency	Expected frequency E_i	$(O_i - E_i)$	$(O_i - E_i)^2$	$(O_i - E_i)^2 / E_i$

	O_i				
A	90	75	15	225	225/75
AB	135	150	-15	225	225/150
B	75	75	0	0	0/75

$$\therefore \chi^2 = \sum_i \frac{(O_i - E_i)^2}{E_i} = 225/75 + 225/150 = 4.5$$

$$\text{d.f} = (n-1) = 3-1 = 2$$

at 5% level $\chi^2_{0.05, 2} = 5.991$ and $\chi^2_{\text{cal}} = 4.5$

As $\chi^2_{\text{cal}} < \chi^2_{\text{tab}}$. Then do not reject H_0 . i.e. The genetic theory that the proportion of A:AB:B = 1:2:1 is accepted.

In most of hypothesis testing, the population under study has been taken to follow a known distribution such as normal, binomial or Poisson distribution. There is neither enough evidence nor enough logic, in the number of cases, to assume a particular distribution for the data. In such a situation one has every right to question the validity of such an assumption. To test the assertion of how closely the actual distribution approximates to a particular theoretical distribution, chi-square test is appropriate. To assume that the population is distributed normally is a common practice and hence we explain the test of goodness of fit of normal population.

Let there be K class intervals and the corresponding frequencies be $f_1, f_2, \dots, f_k$. The area of normal curve within each interval is found from the table of area under the normal curve. On multiplying this area by the total of frequencies, we get the expected (theoretical) frequencies. In this way, the expected frequency for each interval is obtained. Since the tables are provided for standard normal curve, we first change each limit of class interval into a standard normal deviate by using the formula,

$$Z = \frac{X_i - \bar{x}}{s}$$

Where $\bar{x}$ is the sample mean and s is the sample standard deviation.

Then the chi-square statistic is calculated as: $\chi^2 = \sum_i \frac{(f_o - f_e)^2}{f_e}$, $i = 1, 2, \dots, k$ and degrees of freedom for χ^2 in this case is d.f $(k - p - 1)$ where K is the number of class intervals and p is the number of parameters of the distribution estimated. For normal distribution two parameters population mean (μ) and population standard deviation (σ) are estimated by sample mean $\bar{x}$ and sample standard deviation S . thus $p = 2$. Hence in this case, chi-square has $(k - 3)$ degrees of freedom.

If the calculated value of χ^2 is greater than the table value of χ^2 for $(k - p - 1)$ d.f, and level of significance α , reject H_0 . Rejection of H_0 means, that the postulated theoretical distribution is not fit to the observed data, or in other words the data do not support the assertion about the theoretical distribution.

Note: - If d.f = 1, then to apply χ^2 distribution, make an adjustment known as *Yates' correction* which involves reducing each value of $|O_i - E_i|$ by $\frac{1}{2}$ and use the modified χ^2

given by $\chi^2 = \sum_{i=1}^n \frac{\left[|O_i - E_i| - \frac{1}{2}\right]^2}{E_i}$ which is χ^2 statistic used when d.f = 1

4.3 Test of independence

Before conducting test of independence, let us say few about *contingency table*. IN contingency table, data are often based on counting of objects or units. These numbers fall in various categories of attributes in a two-way classification and are very well displayed systematically by a table known as contingency table. We can express a contingency table as a rectangular array of order $(m \times p)$, having mp cells. Where m denotes the number of rows which are equal to the number of categories of an attribute or criterion A, and p denotes the number of columns equal to the number of categories of attribute or criterion B. The matrix notation for contingency table is as follows:

Contingency table

In the above table, cell frequency O_{ij} in $(i, j)^{th}$ cell represents the number of objects or individuals possessing the characteristics of A_i and B_j . Any of the row total R_i and column total C_j is called the *marginal total*, where $i = 1, 2, \dots, m$; $j = 1, 2, \dots, p$ and n is the grand total. Such a table eases out the process of calculations of corresponding expected frequencies and subsequently the value of chi-square statistics.

Example: - In a large concern, an opinion regarding three types Total employees were categories, namely- technical and obtained by way of presented in form of given below.

	Attribute B			Total
	B_1	$B_2 \dots B_j \dots B_p$		
Attribute A	A_1	O_{11}	$O_{12} \dots O_{1j} \dots O_{1p}$	R_1
	A_2	O_{21}	$O_{22} \dots O_{2j} \dots O_{2p}$	R_2
	$\vdots$	$\vdots$	$\vdots$	$\vdots$
	A_i	O_{i1}	$O_{i2} \dots O_{ij} \dots O_{ip}$	R_i
	$\vdots$	$\vdots$	$\vdots$	$\vdots$
	A_m	O_{m1}	$O_{m2} \dots O_{mj} \dots O_{mp}$	R_m
Employees category	Bonus Schemes			Total
	Type I	Type II	Type III	
Labourer	190	243	197	630
Clerical	82	44	44	170
Technical	23	78	34	135
Executive	Total	C_1	$C_2 \dots C_j \dots C_p$	n
Total	300	377	283	960

manufacturing survey was conducted of bonus schemes. divided into four labourers, clerical, executives. The results opinion survey are contingency table as

The table indicates that out of 960 employees, 190 belonging to labour class, preferred the type I scheme, and 243 preferred type II scheme. Similarly, any other cell frequency may be interpreted. 630, 170, 135 and 25 are the row marginal totals and 300, 377, and 283 are the column marginal totals.

Test of Independence of Factors:

It is apparent now, that a contingency table is a rectangular array having rows and columns associated with different factors. The hypothesis that one factor is independent of the other or not, can be tested by the chi-square test. The hypothesis, H_0 : Two factors are independent of each other against H_1 : Two factors are dependent, can be tested by the statistic,

$$\chi^2 = \sum_{i=1}^m \sum_{j=1}^p \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

Statistic χ^2 has $(m-1)(p-1)$ d.f., whereas the data are presented in the form of the contingency table of m rows and p columns as given earlier.

In the calculation of χ^2 , E_{ij} is the expected frequency corresponding to $(i, j)^{th}$ cell observed frequency O_{ij} . Under the null hypothesis H_0 , the expected frequency,

$$E_{ij} = \frac{i^{th} \text{ row total} \times j^{th} \text{ column total}}{\text{Sample size}}$$

$$E_{ij} = \frac{R_i \times C_j}{n}$$

Once the expected frequencies are calculated, the value of chi-square is obtained by the

formula $\chi^2 = \sum_{i=1}^m \sum_{j=1}^p \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$, To make the decision about H_0 , calculated value of chi-square is compared with the table value of chi-square for $(m-1)(p-1)$ d.f and α level of significance. If $\chi^2_{cal} > \chi^2_{\alpha, (m-1)(p-1)}$, reject H_0 , which means that two factors are not independent.

For a data given in the example of contingency table, if we test the hypothesis:

H_0 : opinion about bonus schemes is independent of the types of employees.

Against H_1 : H_0 is not true, as follows:

$$E_{ij} = \frac{R_i \times C_j}{n}$$

The expected values are calculated with the formula $E_{ij} = \frac{R_i \times C_j}{n}$ and presented in the table below, in parentheses, adjacent to the observed frequency and the value of chi-square then calculated.

Expected frequencies are calculated as,

$$E_{11} = \frac{630 \times 300}{960} = 196.88 = 197$$

$$E_{12} = \frac{630 \times 377}{960} = 247.41 = 247$$

$$\vdots$$

$$E_{43} = \frac{25 \times 283}{960} = 7.37 = 7$$

Employees category	Bonus Schemes			Total
	Type I	Type II	Type III	
Labourer	190	243(247	197(186)	630
Clerical	(197)	)	44(50)	170
Technical	82(53)	44(67)	34(40)	135
Executive	23(42)	78(53)	8(7)	25
	5(8)	12(10)		
Total	300	377	283	960

$$\chi^2 = 48.4028$$

The table value of chi-square for $(4-1)(3-1) = 6$ d.f and $\alpha = 0.05$ is 12.59. Since χ^2_{cal} is greater than 12.59, we reject H_0 . It means that the opinion about bonus scheme is influenced by the category of employees.

Exercise 1. 1, 072 school boys were classified according to intelligence, and at the same time their economic conditions were recorded. The results are shown in the following table. Judge at 5% level whether there is any association between intelligence and economic conditions.

Economic Conditions	Intelligence			
	Excellent	Good	Mediocre	Dull
Good	48	199	181	82
Not good	81	185	190	106