CHAPTER ELEVEN:
INFERENTIAL STATISTICS

11.0 Introduction

Statistical inference is based on Estimation and Hypotheses.

11.1 ESTIMATION

Estimation is a technique used in summarizing the information contained in a sample (e.g. by calculating mean) to make inference about the related population. It is usually a process which deals with guessing (or estimating) the population characteristics (such as the population, mean, and standard deviation) from the information contained in the sample (such as sample mean and standard deviation).

Estimation, in practical situations provide an alternative to statistical applications and decision making where financial, or time constraints or involvement of destructive items could make investigation of the whole population not possible especially in biological and engineering sciences.

In general, everyone makes estimates in one aspect or the other. Managers use estimate because in all but the most trivial decisions, they must make rational decisions without complete information and with a great deal of uncertainty about the future.

11.1.1 Common Terms

Estimate: An estimate is a specific observed value of a statistics. It is actually a value calculated from the sample observations. When we observe a specific numerical value of an estimate, that value is called an estimator.

Estimator: An Estimator is any sample statistic (a random variable) used to predict (or guess) the population parameter. The sample mean can be an estimator of the population mean, and the sample proportion can be used as an estimator of the population proportion.

Statistic: This is a characteristic of a sample. The sample characteristic is also known as sample statistic.

Parameter: This is the characteristics of a population. The population characteristic is also known as population parameter.

11.1.2 Qualities of a good estimator

A good estimator must be:

1. **Un-biasedness:** An estimate is said to be unbiased if the mean of the sample distribution is equal to the population parameter. Therefore, the mean of the distribution of sample would be equal to the population mean. $\bar{X} = \mu$

2. **Efficiency:** Efficiency refers to the size of the standard error of the statistics. If two statistics from a sample of the same size are compared, the statistics that has smaller standard error, or standard deviation of the sample distribution is said to be efficient. An estimator is said to be more efficient than the other if in a repeated sampling, its variance is smaller.

3. **Consistency:** A statistic is consistent estimator of a population parameter if as the sample size increases, the value of the statistics become very close to the value of the population parameter.

4. **Sufficiency:** An estimator is sufficient if it makes so much use of the information in the sample than any other estimator. In other word, an estimator is said to be sufficient if it uses all the information in the sample in estimating the required population parameter.

11.2 TYPES OF ESTIMATION

Estimation is divided into two parts; point and interval estimation.

11.2.1 Point Estimation

This is the process of using a single-valued estimate (usually obtained from computation) to serve as an approximation to the population parameter of interest. It is a number obtained from computation of observed sample data which serves as an approximation to the population parameter. It should be noted that a point estimate is much more useful if it is accompanied by an estimate of the error that might be involved.

Point estimate is a single number that is used to estimate an unknown population parameter. For instance, sample mean ($\bar{x}$) could be used as an estimator of the population mean (μ) when the population mean (μ) is unknown i.e. the sample mean ($\bar{x}$) is a point estimate of the population mean (μ).

11.2.2 Point Estimate of the Population Mean

The sample mean is the best estimate of the population mean. It is unbiased, consistent, and the most efficient estimator. If a series of sample size $n \geq 30$ is taken from a population, it would be found that each of the sample mean is approximately equal to the population. The sample mean cluster closely around the population mean, the larger the sample, the more closely mean clusters around the population mean and the distribution of a sample mean follows a normal curve.

Example 11.1

Draw all the possible samples of size 2 from the population of four numbers 3, 5, 6, 8 with replacement and without replacement. Hence, calculate the population and sample means of the data and compare the results.

Solution

Population: 3, 5, 6, 8

Samples: **with replacement** **without replacement**

(3,3), (3,5), (3,6), (3,8) (3,5), (3,6), (3,8)

(5,3), (5,5), (5,6), (5,8) (5,6), (5,8) (6,8)

(6,3), (6,5), (6,6), (6,8)

(8,3), (8,5), (8,6), (8,8)

Population mean $= \frac{3+5+6+8}{4} = \frac{22}{4} = 5.5$

Sample means:

with replacement		without replacement	
3,3	3	3,5	4
3,5	4	3,6	4.5
3,6	4.5	3,8	5.5
3,8	5.5	5,6	5.5
5,3	4	5,8	6.5
5,5	5	6,8	7
5,6	5.5	**Total**	**33**
5,8	6.5		
6,3	4.5		
6,5	5.5		
6,6	6		
6,8	7		
8,3	5.5		
8,5	6.5		
8,6	7		
8,8	8		
Total	**88**		

With replacement: $\frac{\Sigma x}{n} = \frac{88}{16} = 5.5$

Without replacement: $\frac{\Sigma x}{n} = \frac{33}{6} = 5.5$

Conclusion: It is observed that in each case the sample mean is equal to the population. Thus, it shows that the sample mean ($\bar{x}$) is a point estimate of the population mean (μ).

11.2.3 Central Limit Theorem

The central limit theorem describes the characteristics of sample means of the population selected at random from finite population. The means of the samples will be normally distributed and the mean value of the sample mean will be the same as the mean of the population. The distribution of sample means will have its own standard deviation, the distribution of the expected sampling error. It is known as the standard error of the mean. The standard error of the mean is a standard deviation of the distribution of sample means. It is computed by the use of this formula $S_{\bar{x}} = \frac{S}{\sqrt{N}}$

Where $S_{\bar{x}}$ = Standard error of the mean

S = Estimation of standard deviation

N = Sample

It can be seen that as the size of the sample increases, the standard error of the mean decreases.

Example 11.2

A bank calculates that its individual savings accounts are normally distributed with a mean of N200 and a standard deviation of N600. If the bank takes a random sample of 100 accounts, what is the estimate of the individuals mean accounts and what is the standard error of the mean?

Solution:

The sample size = 100, and sample mean = 200

So, the estimate of the individuals account is also 200 i.e $\bar{x} = 200 = \mu = 200$.

$$S_{\bar{x}} = \frac{S}{\sqrt{N}} = \frac{600}{\sqrt{100}} = \frac{600}{10} = 60$$

11.2.4 Interval Estimation

An interval estimate is a range of values used to estimate a population parameter. It is the range of values obtained from computations on the observed sample values, and believed with some degree of confidence (assurance) used to estimate the population parameter i.e. it is a range (or an interval) within which we expected the true value of the parameter to lie. It is an interval

determined by two numbers obtained from the computation on observed sample values, and it is expected to contain the unknown true value of the parameter. It indicates the error in two ways: by the extent of its range and by the probability of the true population parameter lying within that range.

11.2.5 Confidence Level, Confidence Interval and Confidence Limit

Confidence level: In statistics, confidence level is the probability that is associated with an interval estimate. This probability shows how confident we are that the interval estimate will include the population parameter. A higher probability means more confidence. In estimation, the most commonly used confidence levels are 90%, 95% and 99%.

It is a measure of the probability of any estimate being correct.

Confidence interval: This is the range of values that MUST be given (or taken) in order to be able to make an estimate at given level of confidence. For example, if we are 90% confident that the mean of a population of income of Nasarawa community will lie between N8500 and N24500, then, the range 'N8500 – N24500' is the confidence interval.

Confidence limits: Confidence limits are the upper and lower limits of the confidence interval. The confidence limits for population means are based on the sample mean. The standard error of the mean and on known characteristics of Normal distribution as follows:

Mean ± 96 δ includes 95% of the population

Mean ± 2.58 δ which includes 99% of the population

These characteristics can be applied in finding confidence limits for population mean when sample mean and standard error are calculated.

11.2.5.1 CONFIDENCE INTERVAL FOR SINGLE MEAN (X)

When δ^2 is known or n ≥ 30

$$\mu = \bar{x} \pm Z_{\alpha/2}\delta_{\bar{x}} \text{ where } \delta_{\bar{x}} = \frac{\delta}{\sqrt{n}} \text{ or } \frac{\delta^2}{\sqrt{n}}$$

$$\bar{x} - Z_{\alpha/2}\delta_{\bar{x}} \leq \mu \leq \bar{x} + Z_{\alpha/2}\delta_{\bar{x}}$$

When δ^2 is unknown or n < 30

$$\mu = \bar{x} \pm t_{\alpha/2}\delta_{\bar{x}} \text{ given } \bar{x} - t_{\alpha/2}\delta_{\bar{x}} \leq \mu \leq \bar{x} + t_{\alpha/2}\delta_{\bar{x}}$$

$\delta_{\bar{x}} = \sqrt{\frac{\delta^2}{n}}$ for sampling with replacement; $\delta_{\bar{x}} = \sqrt{\frac{\delta^2}{n}\left[\frac{N-n}{N-1}\right]}$ for sampling without replacement

11.2.5.2 CONFIDENCE INTERVAL FOR TWO MEANS

When δ_1^2, δ_2^2 are known and $n_1, n_2 \geq 30$

$$\mu_1 - \mu_2 = \bar{X}_1 - \bar{X}_2 \pm Z_{\alpha/2}\sqrt{\left[\frac{\delta_1^2}{n_1} + \frac{\delta_2^2}{n_2}\right]}$$

When δ_1^2, δ_2^2 are unknown & $n_1\, n_2 < 30$

$$\mu_1 - \mu_2 = \bar{X}_1 - \bar{X}_2 \pm t_{\alpha/2}, VS_p\sqrt{\left[\frac{1}{n_1} + \frac{1}{n_2}\right]}$$

Where $S_p^2 = \frac{(n_1-1)S_1^2+(n_2-1)S_2^2}{n_1+n_2-2}$ and $V = n_A + n_B - 2$

Example 11.3:

What are the 95% confidence limits for the population mean given the data from previous example where $\bar{x}$ = 150grams and $S_{\bar{X}}$ = 1.2grams

At 95% confidence level μ (population mean) is between the two values: $\bar{x} \pm 1.96\ S_{\bar{X}}$

Thus, we have $150 \pm 1.96 S_{\bar{X}}$ but $S_{\bar{X}}$ = 1.2grams

$150 \pm 1.96\ (1.2) = 150 \pm 2.352 \Rightarrow 150 + 2.352 = 152.35$ or $150 - 2.352 = 147.65$

The above implies that we are 95% confident that the population mean lies within the confidence zone i.e. between 147.65 grams and 152.35 grams.

At the 99% confidence level, the population mean (μ) will be between the two values:

$150 \pm 2.58\ S_{\bar{X}}$ but $S_{\bar{X}}$ = 1.2grams

$150 \pm 2.58\ (1.2) = 150 \pm 3.096 \Rightarrow$ Either; $150 + 3.096 = 153.096$

or $150 - 3.096 = 146.904$

Also, this means that we are 99% confident that the population mean lies between 146.904 grams and 153.096 grams (called the confidence zone).

11.2.5.3 CONFIDENCE INTERVAL FOR SINGLE PROPORTION

When $n \geq 30$; $P \neq Z_{\alpha/2}\, \delta_P$ and When $n < 30$; $P \pm t_{\alpha/2}, V\delta_P$

$\delta_p^2 = \frac{Pq}{n}$; sampling with replacement

$\delta_p^2 = \frac{Pq}{n}\left[\frac{N-n}{N-1}\right]$; sampling without replacement

Example 11.4

In a random sample of 500 students selected from FPN, it was found that 340 have a fan in their rooms. Construction 95% C.I. for the entire students that own a fan.

Solution:

$P = \frac{n}{N} = \frac{340}{500} = 0.68\ ;\ q = 1 - P = 1 - 0.68 = 0.32$

But $Z_{0.025} = 1.96$

Thus, $P \pm Z_{\alpha/2}\sqrt{\frac{(Pq)}{n}} = 0.68 \pm 1.96\sqrt{\frac{(0.68\times0.32)}{500}}$

$= 0.68 \pm 1.96\sqrt{\frac{0.2176}{500}} = 0.68 \pm 1.96\sqrt{0.0004352}$

$= 0.68 \pm 1.96(0.02) = 0.68 \pm 0.04$

$\Rightarrow 0.68 - 0.04 < P < 0.68 + 0.04$

Hence, $0.64 < P < 0.72$

11.2.5.4 CONFIDENCE INTERVAL FOR TWO PROPORTIONS

When n_1, n_2 are known or $n_1, n_2 \geq 30$

$$P_1 - P_2 \pm Z_{\alpha/2}\sqrt{\frac{P_1q_1}{n_1} + \frac{P_2q_2}{n_2}}$$

When $n_1\ n_2 < 30$

$$\bar{X}_1 - \bar{X}_2 \pm t_{\alpha/2}, V\delta_{P_1-P_2}\sqrt{\left[\frac{1}{n_1} + \frac{1}{n_2}\right]}$$

$\delta_{P_1-P_2} = \bar{P}(1-\bar{P})$ where $\bar{P} = \frac{n_1P_1+n_2P_2}{n_1+n_2}$

Example 11.5

Some students claimed that the proportion of females' students in the school of business and sciences is 0.67 to verify their claim a random sample of 25 and 27 students were taken from school of business and science respectively with the following results 0.56 & 0.43 of the proportion of female student in the respective school. Can we conclude that their claim is true at 95% C.I.?

Solution

$n_1 = 25, n_2\ 27, P_1 = 0.56, P_2 = 0.43$

$$\bar{P} = \frac{n_1P_1 + n_2P_2}{n_1 + n_2} = \frac{25(0.56) + 27(0.43)}{25 + 27}$$

$$= \frac{14 + 11.61}{52} = \frac{25.61}{52} = 0.4925 \approx 0.49$$

Thus, $1 - \bar{P} = 1 - 0.49 = 0.51$

So, $\delta_{P_1 - P_2} = \bar{P}(1 - \bar{P}) = 0.49(0.51) = 0.2499 \approx 0.25$

$V = n_1 + n_2 - 2 = 25 + 27 - 2 = 50$

$t_{\alpha/2}, V = t_{0.025}, 50 = 2.01$

$$\bar{X}_1 - \bar{X}_2 \pm t_{\alpha/2}, V\delta_{P_1 - P_2}\sqrt{\left[\frac{1}{n_1} + \frac{1}{n_2}\right]} = 0.56 - 0.43 \pm 2.01(0.25)\sqrt{\left[\frac{1}{25} + \frac{1}{27}\right]}$$

$$= 0.13 \pm 0.5025\sqrt{[0.04 + 0.037]} = 0.13 \pm 0.5025\sqrt{0.077}$$

$$= 0.13 \pm 0.5025(0.2775) = 0.13 \pm 0.14 \Rightarrow 0.13 - 0.14 < P_1 - P_2 < 0.13 + 0.14$$

$$\Rightarrow -0.01 < P_1 - P_2 < 0.27$$

The claim is not true at 95% C.I since 0.67 is not within the interval.

11.2.6 Estimation for Large and Small Samples

In situation where the samples are large i.e. (n > 30), the sample standard deviation (s) is used as an estimate of the population standard deviation (δ).

It is also clear that the distribution of sample mean is approximately normal such that the attribute of normal distribution can be used to calculate the confidence limits using the standard error of the mean.

The above is not the same for small sample i.e. (n < 30) because the arithmetic mean of small samples is not normally distributed. Therefore, students' t - distribution table is required in this situation.

For Large Samples

Population mean (μ) is between the two values: $\bar{x} \pm Z_{\frac{\alpha}{2}} S_{\bar{x}} = \bar{x} \pm Z_{\frac{\alpha}{2}} \frac{S}{\sqrt{N}}$

At 95% confidence level Z = 1.96

At 99% confidence level Z = 2.58

(These are values of Z from the normal area table)

Population mean (μ) is between the two values: $\bar{x} \pm t_{\frac{\alpha}{2},V}\ S_{\bar{X}} = \bar{x} \pm t_{\frac{\alpha}{2},V}\ \frac{S}{\sqrt{N}}$

The value of t is obtained from the t-distribution table for the required confidence level.

Example 11.6

The chief accountant of a company believe that his company monthly sales is town A is 70, 000 more than that of town B. The statistics department of his company try to investigate his believes by taken a random sample from each town. The following results were obtained:

From town 1, $n_1 = 100$, $\bar{X}_1 = 590, 000$, $S_1^2 = 9050$,

From town 2, $n_2 = 150$ $\bar{X}_2 = 585, 000$, $S_2^2 = 8700$,

i. Can you help this company construct 95% C.I. to ascertain his believe?

ii. Suppose the sample size are reduce to 20, 25, respectively in each town with the sample mean and variance remain same, can we still arrive at the same conclusion at 95% confidence.

Solution:

i. $\bar{X}_1 - \bar{X}_2 \pm Z_{\alpha/2}\sqrt{\frac{\delta_1^2}{n_1} + \frac{\delta_2^2}{n_2}}$ But $Z_{0.05/2} = Z_{0.025} = 1.96$

$$590000 - 585000 \pm 1.96\sqrt{\frac{9050}{100} + \frac{8700}{150}} = 5000 \pm 1.96\sqrt{90.5 + 58}$$

$$= 5000 \pm 1.96\sqrt{148.5} = 5000 \pm 1.96(12.1861) = 5000 \pm 23.8848$$

$$\Rightarrow 4976.12 < \mu_A - \mu_B < 5023.88$$

ii. For sample sizes 20, 25 i.e. n

$$\mu_1 - \mu_2 = \bar{X}_1 - \bar{X}_2 \pm t_{\alpha/2}, V S_p\sqrt{\left[\frac{1}{n_1} + \frac{1}{n_2}\right]}$$

Where $S_p^2 = \frac{(n_1-1)S_1^2+(n_2-1)S_2^2}{n_1+n_2-2}$ and $V = n_1 + n_2 - 2$

$$S_p^2 = \frac{(20-1)9050 + (25-1)8700}{20+25-2} = \frac{(19)9050 + (24)8700}{45-2} = \frac{171950 + 208800}{43} = \frac{380750}{43} = 8854.65$$

Now, $t_{\alpha/2}, V = t_{0.05/2}, 43 = t_{0.025}, 43 = 2.02$, and $S_p = \sqrt{8854.65} = 94.10$

Thus, $\bar{X}_1 - \bar{X}_2 \pm t_{\alpha/2}, V S_p\sqrt{\left[\frac{1}{n_1} + \frac{1}{n_2}\right]} \Rightarrow 590000 - 585000 \pm 2.02(94.10)\sqrt{\left[\frac{1}{100} + \frac{1}{150}\right]}$

$$= 5000 \pm 190.082\sqrt{[0.01 + 0.0067]} = 5000 \pm 190.082\sqrt{[0.0167]}$$

$$= 5000 \pm 190.082(0.1292) = 5000 \pm 24.56$$

$$\Rightarrow 5000 - 24.56 < \mu_A - \mu_B < 5000 + 24.56 \Rightarrow 4975.44 < \mu_A - \mu_B < 5024.56$$

Conclusion

Since 70, 000 does not fall within the interval, we therefore conclude that the chief accountant believe is not true at 95% C.I. when the sample size were reduce to 20825 in both town A & town B.

11.3 TESTING OF HYPOTHESIS

Hypothesis testing begins with an assumption, called a hypothesis that we make about a population parameter. Hypothesis is an assertion, investigation or claim made which can be tested about a population. It is an investigation to determine the difference between sample statistic and hypothesized population parameter.

Hypothesis Formulated

1. The Null hypothesis (H_0)
2. Alternative hypothesis (H_i)

The Null hypothesis (H_0) is a statement about the population to be tested. It is usually a hypothesis of no difference and as such contains equality sign. It is stated with the hope of rejecting it.

The alternative hypothesis (H_i) represents the rest of the population values other than those contained in the null hypothesis. Rejecting Ho: leads to accepting H_i:

Result of Hypothesis Testing

There are only four possible outcomes when testing the null hypothesis (H_0):

1. A true hypothesis is accepted – a right decision
2. A false hypothesis is rejected – a right decision
3. A true hypothesis rejected – a wrong decision or type I error
4. A false hypothesis accepted – a wrong decision or type II error.

Test Statistic

The test statistics is the decision variable whose magnitude is used to decide whether to reject the null hypothesis (Ho) or not. Its value is computed from the relation below:

$$\text{Test Statistics} = \frac{\text{Sample statistics - hypothesiz ed parameter}}{\text{Standard error of statistics}}$$

$$Z = \bar{x} - \mu \Big/ \left(\frac{S}{\sqrt{N}}\right) \text{ where } n > 30 \text{ and } t = \bar{x} - \mu \Big/ \left(\frac{S}{\sqrt{N}}\right) \text{ where } n < 30$$

Degree of Freedom

Degree of freedom is the number of values we are can choose freely. We will use degree of freedom when we select a t-distribution to estimate a population mean. It is denoted by $V = n - 1$ where n equals the sample size.

E.g. find the t-value for a 95% confidence level, with a sample size of 20.

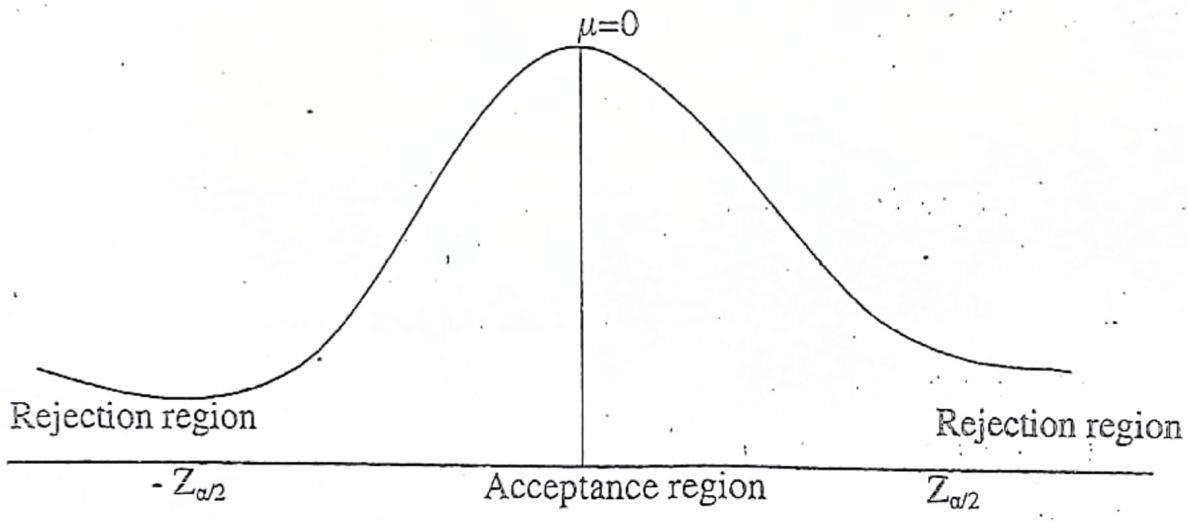
Now, V = n-1 where n = 20. Thus, V = 20 – 1 = 19

From the t-table in the appendix headed 0.05, we shall find the value aligned to 19 (which is 2.093).

Decision Rule

The value of the test statistics falls within one of the two regions namely:

1. The rejection or critical region
2. The acceptance region



Significance Level

The purpose of hypothesis testing is not to question the computed value of the sample statistic but to make a judgment about the *difference* between the sample statistic and a hypothesized population parameter. There is no single standard or universal level of significance for testing hypothesis. In some instances, a 5% level of significance is used and at other time 1% level of significance is used. The higher the significance level we use for testing hypothesis, the higher the probability of rejecting a null hypothesis when it is true.

11.3.1 Procedures for Test of Hypothesis

1. State the null and alternative hypothesis
2. Set up a suitable significance level
3. Determine the appropriate test statistic
4. Carry out the computation
5. State the Decision rule
6. Using the decision rule and the computed value, conclude whether to accept or reject the null hypothesis.

Type I and Type II Error

A type I error is the error of rejecting a hypothesis where it is in fact true.

A type II error is the error of accepting a hypothesis where it is in fact false. The errors are mutually exclusive; it can be type I or type II error or both.

The risks associated with the two types of error are represented by α and β, thus:

Type I error = α

Type II error = β

The α risk is the level of significance chosen for the hypothesis test which are usually (commonly) 1% or 5%

11.3.2 Choosing the Distribution to use in Hypothesis Testing

After deciding the level of significance to use i.e. either 95% or 99%, the next step is to determine the appropriate distribution to use. This could be between The Normal distribution (Z-table) or t-distribution (t-table)

11.3.2.1 Normal Distribution

This is to be used under the following conditions.

1. When sample size is larger than 30 i.e $n>30$
2. When population standard deviation is known
3. When population standard deviation is not known

11.3.2.2 The t-distribution

The use of t-distribution for estimating is required under the following conditions:

1. When sample size is less than or equal to 30 i.e $n \leq 30$
2. When the population standard deviation is unknown.

Furthermore, in using the t-distribution, we assume that the population is normal or approximately normal. Thus: $t = \bar{x} - \mu \Big/ \left(\frac{S}{\sqrt{N}}\right)$

Where S is the estimate of population standard deviation (σ); $S = \sqrt{\frac{\sum(x-\bar{x})^2}{n-1}}$

Attributes of t-distribution

a. It is symmetrical

b. It is flatter than the normal distribution

c. There is different t-distribution for every possible sample

d. As the sample size gets larger, the shape of the distribution losses it flatness and become approximately equal to the normal distribution.

11.3.3 Two Tailed and One Tailed Test of Hypothesis

A two tailed test of hypothesis of population mean will reject the null hypothesis if the sample mean is significantly higher than or lower than the population the hypothesized population. Thus, there are two rejection zones in two tailed test.

The two tailed test is appropriate when the null hypothesis (H_0) is equal to the population mean (μ) i.e. $H_0 = \mu$ and the alternative hypothesis is not equal to population mean (μ) i.e. $Hi \neq \mu$.

A one tailed test is used when a two tailed test is not adequate. A one tailed test can be a left tailed test or a right tailed test depending on the condition.

A left-tailed test is used if the hypothesis are $H_0 = \mu$ and $H_i < H_0$. In such situation the sample mean is significantly below the hypothesized population mean which will lead to rejecting the null hypothesis in favour of the alternative hypothesis. In other word, the rejection region is the lower tail (left tail) of the distribution of the sample mean.

A right tailed test is used when the hypothesis are $H_0 = \mu$ and $H_i > H_0$. Only value of the sample mean that are significantly above the hypothesized population mean will make us to reject the null hypothesis in favour of the alternative hypothesis.

For example, a marketing manager asked her marketing representative to observe a limit on marketing expenses. The manager hopes to keep expenses to an average of N200 per marketing representative per day. Two months after, the limit is imposed, a sample of submitted daily expenses was taken to see if the limit is been observed.

The null hypothesis is $H_0 = \mu = 200$. The manager is concerned only with excessive expenses, thus the alternative hypothesis is $Hi = \mu > 200$. An upper tailed test is being used. The null hypothesis is rejected only if the sample mean is significantly higher than 200.

Example 11.7:

A fisherman decides that he needs a line that will be more than 10kg if he is to catch the size of fish he desired. He tested 100 pieces of brand of lines and find $\bar{x} = 10.4$kg with a standard deviation of 0.08kg based on the sample data, does this requirement meet the test at $\alpha = 0.05$?

Solution

Since $N > 30$, we use $Z = \bar{x} - \mu \Big/ \left(\frac{S}{\sqrt{N}}\right)$

Where $H_0: \mu = 10$, $H_i: \mu > 10$, $N = 100$, $\bar{x} = 10.4$, $S = 0.08$, $\alpha = 0.05$

$$Z = 10.4 - 10 \Big/ \left(\frac{0.08}{\sqrt{100}}\right) = 0.4 \Big/ \left(\frac{0.08}{10}\right) = \frac{0.4}{0.008} = 50$$

Decision rule: Reject H_0 if $Z_{cal} > Z_{tab}$ but $Z_{cal} = 50$ and $Z_{tab} = 1.96$

Conclusion: Since the Z_{cal} is greater than the Z_{tab} (i.e. $50 > 1.96$), we reject the null hypothesis (H_0) and accept the alternative hypothesis (H_i). This implies that his decision is made.

Example 11.8:

A press company hypothesized that the life of its largest week press is 14500hrs with a known standard deviation of 2100hrs from a sample of 25 press, the company finds a sample mean of 1300hrs. at 0.01 significance level, should the company conclude that the average life of the press is less than hypothesized parameter ?

Solution

Since $N < 30$, we use $P = \dfrac{\mu - \bar{x}}{\left(\frac{S}{\sqrt{N}}\right)}$

Where $H_0: \mu = 14500$, $H_i: \mu > 14500$, $N = 25$, $\bar{x} = 1300$, $S = 2100$, $\alpha = 0.01$

$$P = \frac{\mu - \bar{x}}{\left(\frac{S}{\sqrt{N}}\right)} = \frac{14500 - 1300}{\left(\frac{2100}{\sqrt{25}}\right)} = \frac{13200}{\left(\frac{2100}{5}\right)} = \frac{13200}{420} = 31.42$$

Decision rule: Reject H_0 if $Z_{cal} > Z_{tab}$. From computation and table, $Z_{cal} = 31.42$ and $Z_{tab} = 1.96$

Conclusion:

Since the Z_{cal} is greater than the Z_{tab} (i.e. $50 > 1.96$), we reject the null hypothesis (H_0) and accept the alternative hypothesis (H_i). This implies that his decision is made.

11.3.4 Chi - Square (X^2) Distribution

This is another method used in the testing of hypothesis. It is used when the researcher wants to compare the observed distribution with the expected distribution. The formula to be used for chi-square (x^2) is stated below:

$$X^2 = \sum\left[\frac{(O-E)^2}{E}\right]$$ Where; O = Observed frequency, E = Expected frequency

$$\Rightarrow E = \frac{CT \times RT}{GT}$$; CT = Column total, RT = Row total, GT = Grant total

The calculated chi-square (X^2) calculated from the above is compared with the table value of X^2 for a given level of significance and the number of degree of freedom.

Example 11.9:

Suppose that in a four geographical location in Nigeria, the management of Durex company samples the attitudes of its employees towards job performance review. Respondents responses were shown as captured below.

Observed Frequency

	North-East	South-South	North-West	South- East	Total
Yes	68	75	57	79	279
No	32	45	33	31	141
Total	100	120	90	110	420

H_0: Job performance review should not be done in every two years interval but quarterly.

Hi: Job performance review should not be done quarterly but in every two years interval responses on job performance review timing

Solution

The expected frequency is obtained as $E_{ij} = \frac{CT \times RT}{GT} = \frac{ColumnTota\ l \times RowTotal}{GrandTotal}$

For YES column

North – East

$$e_{11} = \frac{279 \times 100}{420} = \frac{27900}{420} = 66.4$$

South – South

$$e_{12} = \frac{279 \times 120}{420} = \frac{27900}{420} = 79.7$$

North – West

$$e_{13} = \frac{279 \times 90}{420} = \frac{25110}{420} = 59.8$$

South – East

$$e_{14} = \frac{279 \times 110}{420} = \frac{30690}{420} = 73.1$$

For NO column

North – East

$$e_{23} = \frac{141 \times 100}{420} = \frac{14100}{420} = 33.6$$

South – South

$$e_{22} = \frac{141 \times 120}{420} = \frac{16920}{420} = 40.3$$

North – West

$$e_{13} = \frac{141 \times 90}{420} = \frac{12690}{420} = 30.2$$

South – East

$$e_{24} = \frac{141 \times 110}{420} = \frac{15510}{420} = 36.9$$

Expected Frequency

	North-East	South-South	North-West	South- East	Total
Yes	66.4	79.7	59.8	73.1	279
No	33.6	40.3	30.2	36.9	141
Total	100	120	90	110	420

Observed Freq. (O)	Expected Freq.(E)	O – E	$(O-E)^2$	$\frac{(O-E)^2}{E}$
68	66.4	1.6	2.56	0.039
32	33.6	-1.6	2.56	0.076
75	79.7	– 4.7	22.09	0.277
45	40.3	– 4.7	22.09	0.548
57	59.8	– 2.8	7.84	0.131
33	30.2	2.8	7.84	0.260
79	73.1	5.9	34.81	0.476
31	36.9	– 5.9	34.81	0.943
				2.750

Degree of freedom = V = (R-1) (C-1) = (2 – 1)(4 – 1) = 1 x 3 = 3

At this point, we check table of values of the X^2 (cut-off point) for degree of freedom, 3 at 5% significance level

$X^2_{3,(0.05)} = 7.815$ (also called tabulated value).

Decision rule: Since the tabulated value (X^2) = 7.815 is greater than the computed (or calculated) value ($X^2 = 2.750$), we accept the null hypothesis and reject the alternative hypothesis.

Conclusion: This means that attitudes of employees towards job performance review should not be done every two years but quarterly.

Example 11.9:

Mr Ishaku, a president of National Health Insurance Scheme (NHIS), is opposed to NHIS. He argues that it would be too costly to implement. He asked Mr Abdul, his staff statistician to check the matter out. Ishaku believes that the length of days of stays in the hospital is dependents on the

type of health insurance that people have. Abdul collected data on random sample of 660 hospital stays and summarized it below:

Observed Frequency table

		< 5	5-10	>10	Total
Fraction of cost	< 25%	40	75	65	180
Covered	25-50%	30	45	75	150
By Insurance	>50%	40	100	190	330
Total		110	220	330	660

H_0: Length of stay and insurance are independent

Hi: Length of stay depends on type of insurance

Test the hypothesis at 1% level of significance.

Expected Frequency table

		< 5	5-10	>10	Total
Fraction of cost	<25%	30	60	90	180
Covered	25-50	25	50	75	150
By Insurance	>50%	55	110	165	330
Total.		110	220	330	660

The chi-square (x^2) is thus calculated below:

Observed Frequency O	Exp. Frequency E	O – E	$(O-E)^2$	$\frac{(O-E)^2}{E}$
40	30	10	100	3.33
30	25	5	25	1.00
40	55	-15	225	4.09
75	60	15	225	3.75
45	50	-5	25	0.5
100	110	-10	100	0.91
65	90	-25	625	6.94
75	75	0	0	0.00
190	165	25	625	3.79
				24.31

Degree of freedom: $V = (R-1)(C-1) = (3-1)(3-1) = 2 \times 2 = 4$

Now, we check table of values of the X^2 (cut-off point) at degree of freedom, 4 at 1% significance level

$X^2_{4,(0.01)} = 13.277$ (also called tabulated value).

Decision rule: Since the computed X^2 value (24.31) is greater than the tabulated X^2 value (13.277), we reject the null hypothesis and accept the alternative hypothesis.

Conclusion: This means that length of stay in the hospital depends on the type of insurance scheme.

11.3.5 Goodness of Fit Test

The chi-square test can be used to decide whether a particular probability distribution, such as Binomial, Poisson, or Normal, is the appropriate distribution. The chi-square test enables us to ask this question and to test whether there is a significance difference between an observed frequency distribution and a theoretical frequency distribution. In this manner, we can determine the goodness of fit test of a theoretical distribution (i.e how well it fits the distribution of data that we have actually observed)

Example 11.10:

Suppose that Benaiah Company requires that High school senior who are seeking positions be interviewed by three different executives. This enables the company to have a consensus evaluation of each candidate. Each executive gives the candidates either a positive or negative rating as contained in the table below. 100 candidates were interviewed.

Possible positive rating from three executives	No. of candidates receiving each of this rating
0	18
1	47
2	24
3	11
Total	100

For manpower planning purposes, the Director of recruitment for this company thinks that the interview process can be approximated with P = 40%. If the Director wants to test the hypothesis at 5% level of significance. How should he proceed ?

H_0: A binomial distribution with p = 40 is good description

H_1: A binomial distribution with p = 40 is not a good decision

Binomial probability $= P(x) = {}^{n}c_{x}\,(p)^{x}\,(q)^{n-x}$ where; $n = 3, p = 0.4, q = 0.6, x = 0$

Thus, $P(0) = {}^{3}c_{0}\,(0.4)^{0}\,(0.6)^{3-0} = 1 \times 1 \times (0.6)^{3} = 1 \times 0.216 = 0.216$

When $n = 3, p = 0.4, q = 0.6, x = 1$

$\Rightarrow P(1) = {}^{3}c_{1}\,(0.4)^{1}\,(0.6)^{3-1} = 3 \times (0.4) \times (0.6)^{2} = 3 \times 0.4 \times 0.36 = 0.432.$

When $n = 3, p = 0.4, q = 0.6, x = 2$

$\Rightarrow P(2) = {}^{3}c_{2}\,(0.4)^{2}\,(0.6)^{3-2} = 3 \times 0.16 \times 0.6 = 0.288$

When $n = 3, p = 0.4, q = 0.6, x = 3$

$\Rightarrow P(3) = {}^{3}c_{3}\,(0.4)^{3}\,(0.6)^{3-3} = 1 \times 0.064 \times 1 = 0.064.$

Possible position rating from 3 interviews (A)	Observed frequency of candidates (B)	Binomial prob. Of possible outcomes (C)	No. of candidates interviewed (D)	Expected Frequency of candidates (E) = C x D
0	18	0.216	100	21.6
1	47	0.432	100	43.3
2	24	0.288	100	28.8
3	11	0.064	100	6.4
	100			100.00

The Chi - Square (X^2) is computed below

Observed Freq. (O)	Exp. Freq. (E)	O – E	$(O-E)^2$	$\frac{(O-E)^2}{E}$
18	21.6	-3.6	12.96	0.6
47	43.2	3.8	14.44	0.334
24	28.8	-4.8	23.04	0.8
11	6.4	4.6	21.16	3.3063
				5.0403

Degree of freedom for chi-square goodness of fit test is given by V= k -1

We must show the number of classes (symbolizes K) for which we have compared the observed and expected frequencies. The question above contains four classes i.e. 0, 1, 2, 3. Hence,

V = 4 – 1 = 3.

So, the degree of freedom is 3 and it is to be checked at 5% level of significance (which gives 7.815 by using the table)

Decision rule:

Since the computed chi-square goodness of fit test (5.0404) is less than the tabulated value (7.815), we accept the null hypothesis and reject the alternative hypothesis.

Conclusion:

This implies that the binomial distribution with p = 40 is a good description of our observed frequencies.

EXERCISES 11

1. Explain the term statistical estimation
2. Outline the qualities of good estimator
3. What is hypothesis testing?
4. Differentiate between type I and type II errors

5. A company manufacturing rope whose breaking strength has a mean of 26kg and standard deviation of 3kg with a sample of 26 construct a 95% confidence interval for the population mean.

6. In a random sample of 500 students selected from FPN, it was found that 340 have a fan in their rooms. Construction 95% C.I. for the entire students that own a fan.

7. A mail-order company charges flat for postages regardless of the weight of the package. The policy is based on the results of a study conducted several years ago which revealed that the mean weight of mail package was 17.5 gm with a standard deviation of 3.6gm. the head of the accounting department feels the mean weight of packages being mailed today may not be 17.5gm and that the flat rate charge perhaps should be changed. A random sample of 100 packages gave a mean of 18.4gm. at 5% level of significance, test the claim that the mean weight of the package is 17.5gm; assuming that the weights are approximately normally distributed.

8. the average length of time for new students to register their course has been found to be 150 minutes. A random sample of 16 students revealed an average registration time of 142 minutes with a standard deviation of 12 minutes. Test at 5% level of significance the hypothesis that the population average. Length of time is less than 150 minutes.