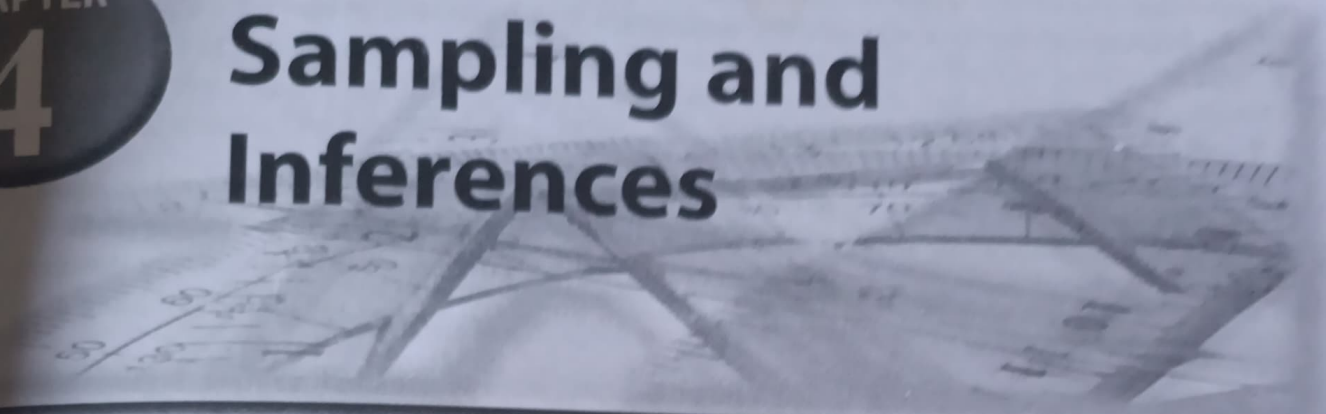
CHAPTER 4

Sampling and Inferences



4.1 INTRODUCTION

In our daily life, most of our decisions depend very much upon the inspection or examination of only a few objects out of the total lot. For example, in a shop, we assess the quality of rice, wheat or any other commodity by taking a handful of it from the bag and then decide whether to purchase it or not. Such examples are related to the concept of sampling. In this chapter, we will learn about random sampling and large sample tests. Then, we will also discuss sampling distribution and small sample tests.

4.2 POPULATION AND SAMPLING

We know that population is the set of all the observations under study, and a sample is a subset of the population drawn for analysis. The process of selecting such samples from the population is called **sampling**. The samples must be selected at random to exclude the possibility of any biasedness. This is called **random sampling**. Thus, in random sampling, each member of the population has equal chance of being included in the sample. A special type of random sampling, is **simple sampling**. In this sampling, each event has the same probability of success and the chance of success of different events is independent whether previous trials have been made or not.

The main objective of sampling is to draw as much inferences as possible of the population by examining only the sample. At the same time, sampling helps in minimising the effort, cost and time. The logic of sampling theory is similar to that of the logic of induction where we pass from particular (sample) to general (population). Such generalisation of inferences from the sample to the population is called **statistical inference**.

The statistical measures or constants of the population such as mean (μ) and variance (σ^2) are called the **parameters** of the population, whereas the statistical measures computed from the samples such as mean ($\bar{x}$) and variance (s^2) are called **statistics**. Population parameters are denoted using Greek letters or capital letters, and sample statistics are denoted using Roman letters. Most often, the population parameters are not

known and their estimates given by the corresponding samples (sample statistics) are used for the analysis of population. However, sample statistics based on different samples can vary from one sample to another. Sampling determines the reliability of these estimates.

4.3 SAMPLING DISTRIBUTION

Let all possible samples of size n be drawn from a population at random. Then, we compute some statistics, say mean $(\bar{x})$, for each of the samples. The means of different samples will not be the same.

If these different means are grouped according to their frequencies, the frequency distribution then obtained is called the **sampling distribution of mean**. Similarly, we can obtain the **sampling distribution of variance**, etc.

A sample having size $n \geq 30$ is called a **large sample,** otherwise a **small sample**. If the sample is large, then the sampling distribution of a statistic approaches a normal distribution. We observe that population may or may not be normal.

If we draw a sample from the population and make some measurement on it and put it back in the population before drawing another sample so that the parent population remains unchanged, then this is called **sampling with replacement**. Henceforth, we will only use the concept of sampling with replacement.

4.3.1 Standard Error

The standard deviation of the sampling distribution of a statistic is called the **standard error** (SE) of the statistic. Thus, the standard error of means is the standard deviation of the sampling distribution of means. Standard error is used to evaluate the difference between the sample statistic and the corresponding population parameter and between two sample statistics. The reciprocal of standard error is called **precision**.

For large samples, the standard errors of some well-known statistics are given in Table 4.1.

Table 4.1 Standard errors of some well-known statistics

Statistic	Standard error
Difference between sample mean ($\bar{x}$) and population mean (μ)	$\sigma/\sqrt{n}$
Difference between sample standard deviation (s) and population standard deviation (σ)	$\sigma/\sqrt{2n}$
Difference between sample proportion (p) and population proportion (P)	$\sqrt{PQ/n}$
Difference between two sample means ($\bar{x_1} - \bar{x_2}$)	$\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$
Difference between two sample standard deviations ($s_1 - s_2$)	$\sqrt{\frac{\sigma_1^2}{2n_1} + \frac{\sigma_2^2}{2n_2}}$
Difference between two sample proportions ($p_1 - p_2$)	$\sqrt{\frac{P_1Q_1}{n_1} + \frac{P_2Q_2}{n_2}}$

In Table 4.1, n is the sample size, σ^2 the population variance, s the sample standard deviation, P the population proportion, p the sample proportion, $Q = 1 - P$.

4.4 TEST OF SIGNIFICANCE

In sampling theory, we are mainly concerned with the study of test of significance. This test enables us to decide, on the basis of sample results, whether the difference between the observed sample statistic and hypothetical parameter value or the difference between two independent sample statistics is significant or might be attributed due to chance or fluctuations of sampling.

4.4.1 Null Hypothesis and Alternative Hypothesis

For applying the test of significance, we set up a hypothesis which is tested for possible rejection under the assumption that it is true. Such hypothesis is called the **null hypothesis** and is denoted by H_0.

The hypothesis complementary to null hypothesis is called the **alternative hypothesis** and is denoted by H_1.

Suppose we want to test the null hypothesis that the population has an assumed value of mean μ_0. Then, we have $H_0: \mu = \mu_0$. The three possible alternative hypotheses will be

(i) $H_1: \mu \neq \mu_0$ $(\mu > \mu_0 \text{ or } \mu < \mu_0)$ —This alternative hypothesis is called the **two-tailed alternative hypothesis** or the **two-tailed test**.

(ii) $H_1: \mu > \mu_0$ —This alternative hypothesis is called the **right-tailed alternative hypothesis** or the **single-tailed test**.

(iii) $H_1: \mu < \mu_0$ —This alternative hypothesis is called the **left-tailed alternative hypothesis** or the **single-tailed test**.

4.4.2 Critical Region and Level of Significance

A region R, corresponding to a sample statistic t, in the sample space which amounts to the rejection of the null hypothesis H_0 is called the **critical region** or the **region of rejection**. The complementary region $\bar{R}$ which amounts to the acceptance of H_0 is called the **acceptance region**.

The probability of the value of the variate falling in the critical region is called the **level of significance**. If t is the value of statistics obtained using a random sample of size n and R is the critical region, then the probability α that a random value of statistic t belongs to the critical region is the level of significance and is given by $P(t \in R \mid H_0) = \alpha$.

The level of significance is always fixed in advance before studying the characteristics of random sample. It is usually expressed as a percentage, and the total area of the critical region is written as $\alpha\%$ level of significance. The two popular values of the level of significance which are usually employed in testing the hypothesis are 5% and 1%.

4.4.3 Single-Tailed and Two-Tailed Tests

A test of any statistical hypothesis in which the alternative hypothesis is single tailed (right tailed or left tailed) is called a **single-tailed test**. A test of any statistical hypothesis in which the alternative hypothesis is two tailed is called a **two-tailed test**.

4.4.4 Critical Values

In case of large samples, if t is any statistics and $E(t)$ is the corresponding population mean, then the variable $z = \dfrac{t - E(t)}{\text{SE}(t)}$ is normally distributed with mean 0 and variance unity. Value of test statistics z which separates the critical region and the accepted region is called the **critical value** or the **significant value** of z. We denote this value by z_α, where α is the level of significance. Critical value depends on

1. the prescribed level of significance and
2. the alternative hypothesis, whether it is a two-tailed test or a single-tailed test.

The critical value z_α of the test statistic for a two-tailed test is given by

$$P(|z| > z_\alpha) = \alpha \tag{4.1}$$

i.e., the total area of the critical region lying at both the ends under the probability curve is α (see Fig. 4.1). Since a normal curve is symmetrical, $P(z > z_\alpha) = P(z < -z_\alpha)$. Now, from (4.1), we have $P(z > z_\alpha) + P(z < -z_\alpha) = \alpha$ or $2P(z > z_\alpha) = \alpha$ or $P(z > z_\alpha) = \alpha/2$, i.e., the area under each tail is $\alpha/2$. The value $z = z_\alpha$ is called the **upper critical value**, and the value $z = -z_\alpha$ is called the **lower critical value**. The acceptance region is given by $(-z_\alpha, z_\alpha)$.

The critical value z_α of the test statistic for a right-tailed test is given by

$$P(z > z_\alpha) = \alpha \tag{4.2}$$

i.e., the total area of the critical region α is the area of the right tail under the probability curve.

The critical value z_α of the test statistic for a left-tailed test is given by

$$P(z < -z_\alpha) = \alpha \tag{4.3}$$

i.e., the total area of the critical region α is the area of the left tail under the probability curve.

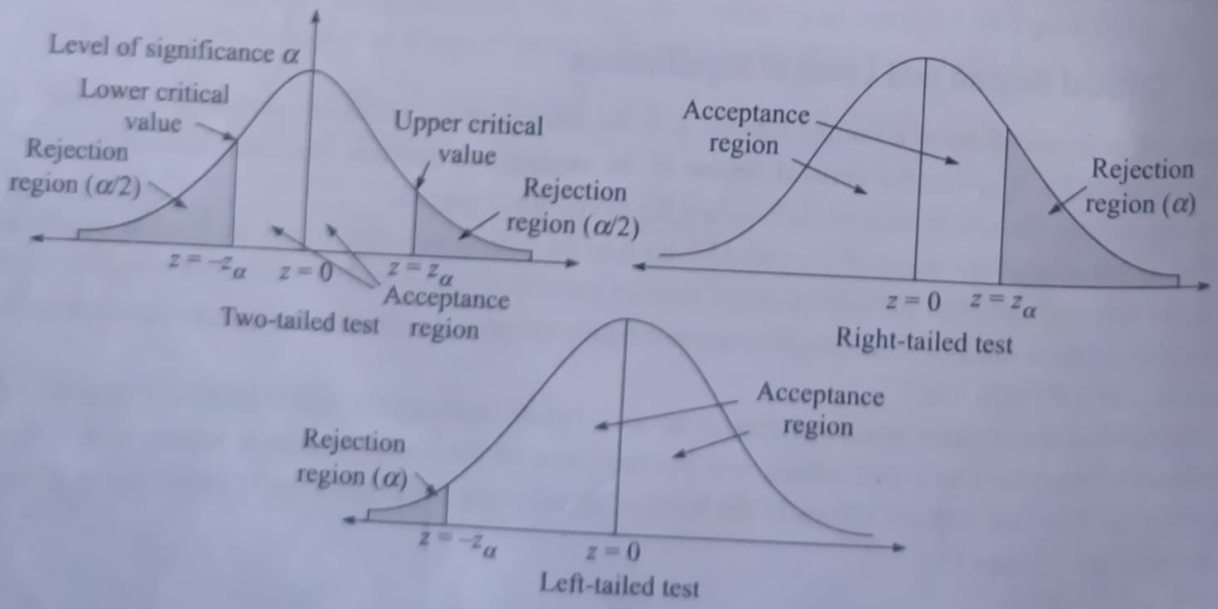


Fig. 4.1

Due to symmetry, we have

$$P(|z| > z_\alpha) = P(z > z_\alpha) + P(z < -z_\alpha) = P(z > z_\alpha) + P(z > z_\alpha) = 2\alpha \qquad \text{[Using (4.2)]}$$

Thus, the critical value of z for a single-tailed test at the level of significance α is the same as the critical value of z for the two-tailed test at the level of significance 2α.

The critical values of z at commonly used level of significance for these tests are listed in Table 4.2.

Table 4.2

Test	Critical value	Level of significance		
		1%	5%	10%
Two tailed	$\lvert z_\alpha \rvert$	2.58	1.96	1.645
Right tailed	z_α	2.33	1.645	1.28
Left tailed	z_α	–2.33	–1.645	–1.28

4.4.5 Confidence Limits

The interval in which a population parameter is supposed to lie is called the **confidence interval** for that population parameter. The end points of this interval are called the **confidence limits** or the **fiducial limits**. The probability that is associated with the confidence interval is called the **confidence level**. It is usually written as $1 - \alpha$, where α is the level of significance. Confidence level indicates the confidence that the experimenter has that the population parameter actually lies within the confidence interval.

Consider that the sampling distribution of statistic t is normal with mean μ and standard deviation σ. The sample statistic t can be expected to lie in the interval $(\mu - 1.96\sigma, \mu + 1.96)$ for 95% times, i.e., we can be confident of finding μ in the interval $(t - 1.96\sigma, t + 1.96\sigma)$ in 95% cases. Thus, we call $(t - 1.96\sigma, t + 1.96\sigma)$ the 95% confidence interval for the estimation of μ. The ends of this interval (i.e., $t \pm 1.96\sigma$) are called 95% confidence limits (or fiducial limits) for t. Similarly, $S \pm 2.58\sigma$ are 99% confidence limits. The number 1.96 is called the confidence coefficient. The values of confidence coefficients corresponding to various levels of significance can be found from the normal curve area table given in the Appendix.

4.4.6 Errors in Testing a Hypothesis

The main aim of sampling theory is to draw valid inferences about the population parameters on the basis of sample results. Because of these reasons, we are liable to commit the following two types of errors:

Type I error: We reject H_0, when it is true. If we write $P[\text{Reject } H_0 \mid H_0] = \alpha$, then α (level of significance) is called the **size of type I error**. It is also referred to as **producer's risk**.

Type II error: We accept H_0, when it is not true, i.e., accept H_0 when H_1 is true. If we write $P[\text{Accept } H_0 \mid H_1] = \beta$, then β is called the **size of type II error**. It is also referred to as **consumer's risk**.

The statistical testing of hypothesis aims to limit the type I error to preassigned values (say, 5% or 1%) and to minimise the type II error. Both these errors can be reduced by increasing the size of the sample (if possible).

4.4.7 Steps for Testing a Hypothesis

1. **Null hypothesis:** Define the null hypothesis H_0.

2. **Alternative hypothesis:** Define the alternative hypothesis H_1 so as to decide the test to be used two-tailed test or single-tailed test.
3. **Level of significance:** Depending on the problem, fix the appropriate level of significance α in advance. This level of significance is fixed before drawing the random sample.
4. **Critical value:** Obtain the value of z_α at the level of significance α.
5. **Test statistics:** Compute the test statistic $z = \dfrac{t - E(t)}{\text{SE}(t)}$ under the null hypothesis.
6. **Conclusion:** Compare $|z|$ with z_α. If $|z| > z_\alpha$, then we reject H_0 and accept H_1 for the level of significance α. This implies that the difference $|t - E(t)|$ is significant for the level of significance α. If $|z| < z_\alpha$, then we accept H_0 and reject H_1 for the level of significance α. This implies that the difference $|t - E(t)|$ is due to some fluctuations in sampling, and hence, the difference is not significant for the level of significance α.

4.5 TEST OF SIGNIFICANCE FOR LARGE SAMPLES

We know that the sample having size $n \geq 30$ is called a **large sample**. For such samples, distributions such as binomial, Poisson, negative binomial and hypergeometric, approach normal distribution assuming the population is normal.

4.5.1 Test for Difference Between Sample Proportion and Population Proportion

Let P and p be population proportion and sample proportion, respectively. Let a sample of size n be drawn from the population. Clearly, this is the same as a series of n independent trials with constant probability P of success.

If X is the number of successes in n independent trials with constant probability P of success for each trial, then $X \sim B(n, P)$. Therefore,

$$E(X) = nP,\ V(X) = nPQ,\ Q = 1 - P$$

Since $p = X/n$ is the observed proportion of success in the sample, we have

$$E(p) = E\left(\frac{X}{n}\right) = \frac{1}{n}E(X) = \frac{nP}{n} = P,\ V(p) = V\left(\frac{X}{n}\right) = \frac{1}{n^2}V(X) = \frac{n(PQ)}{n^2} = \frac{PQ}{n}$$

Hence, $\text{SE}(p) = \sqrt{\dfrac{PQ}{n}}$ and $z = \dfrac{p - E(p)}{\text{SE}(p)} = \dfrac{p - P}{\sqrt{PQ/n}} \sim N(0, 1)$

This z is used to test the significant difference between the population proportion and the sample proportion.

Note:

(i) The probable limits for the observed proportion p of successes are $P \pm z_\alpha\sqrt{PQ/n}$.

(ii) If P is not known, then take $P \approx p$. Then and the approximate limits for the proportion of population are $p \pm z_\alpha\sqrt{pq/n}$, where $q = 1 - p$.

Example 4.1 A coin was tossed 400 times and the head turned up 216 times. Test the hypothesis that the coin is unbiased.

Solution: Here, $n = 400$, X = Number of success = 216

p = Proportion of success in the sample = $X/n = 216/400 = 0.54$

P = Population proportion (i.e., getting head or tail) = 0.5 and $Q = 1 - P = 1 - 0.5 = 0.5$

H_0: The coin is unbiased, i.e., $P = 0.5$

H_1: The coin is not unbiased, i.e., $P \neq 0.5$ (two tailed)

Under H_0, test statistic $z = \dfrac{p - P}{\sqrt{PQ/n}} = \dfrac{0.54 - 0.5}{\sqrt{0.5 \times 0.5/400}} = 1.6$

Since $|z| < 1.96$, the hypothesis is accepted at 5% level of significance and we may conclude that the coin is unbiased at 5% level of significance.

Example 4.2 A cubical die is thrown 9000 times and a throw of 3 or 4 is observed 3240 times. Show that the die cannot be regarded as an unbiased one and find the extreme limits between which the probability of a throw of 3 or 4 lies.

Solution: Here, $n = 9000$, X = Number of success = 3240

p = Probability of success (i.e., getting 3 or 4 on die) = $2/6 = 1/3$, $Q = 1 - 1/3 = 2/3$

P = Probability of success in sample $X/n = 2340/9000 = 0.36$

H_0: The die is unbiased, i.e., $P = 1/3$

and H_1: $P \neq 1/3$ (two-tailed test)

Under H_0, test statistic $z = \dfrac{p - P}{\sqrt{PQ/n}} = \dfrac{0.36 - 0.33}{\sqrt{(1/3) \times (2/3) \times (1/9000)}} = 0.03496$

Since $|z| = 0.03496 < 1.96$, the hypothesis is accepted at 5% level of significance and we may conclude that the die is unbiased at 5% level of significance.

We have to find 95% confidence limits of the proportion. It is given by

$$P \pm z_\alpha \sqrt{\frac{PQ}{n}} = 0.33 \pm 1.96 \sqrt{\frac{0.33 \times 0.67}{9000}} = 0.33 \pm 0.0097 = 0.3203 \text{ and } 0.3397$$

Example 4.3 A manufacturer claimed that at least 95% of the equipments which he supplied to a factory conformed to the specifications. An examination of a sample of 200 pieces of the equipment revealed that 18 were faulty. Test this claim at a significant level of **(i)** 0.05 and **(ii)** 0.01.

Solution: Here, $n = 200$, X = Number of success = $200 - 18 = 182$, p = Proportion of success $= X/n = 182/200 = 0.91$, P = Probability of success = 0.95, $Q = 1 - 0.95 = 0.05$

H_0: The proportion of success in the lot is 95% at least, i.e., $P \geq 0.95$

and H_1: $P < 0.95$ (left-tailed test)

Under H_0, test statistic $z = \dfrac{p - E(p)}{\text{SE}(p)} = \dfrac{p - P}{\sqrt{PQ/n}} = \dfrac{0.91 - 0.95}{\sqrt{(0.95 \times .05)/200}} = \dfrac{-0.04}{\sqrt{0.95 \times 0.05/200}}$

$$= \frac{-0.04}{\sqrt{0.0002375}} = \frac{-0.04}{0.0154} = -2.6$$

(i) Since $z = -2.6 < -1.645$, the hypothesis is rejected for left-tailed test at 5% level of significance and we can conclude that manufacturer's claim is rejected at 5% level of significance.

(ii) Since $z = -2.6 < -2.33$, the hypothesis is rejected for left-tailed test at 1% level of significance and we can conclude that manufacturer's claim is rejected at 1% level of significance.

Example 4.4 Twenty people were attacked by a disease and only 18 survived. Will you reject the hypothesis that the survival rate of those attacked by this disease is 85%?

Solution: Here, $n = 20$, X = Number of persons who survived = 18, p = Proportion of people who survived = 18/20 = 0.9, P = Probability of people who survived = 0.85, $Q = 1 - 0.85 = 0.15$

H_0: The proportion of people who survived the attack is 85%, i.e., $P = 0.85$

and H_1: $P > 0.85$ (right-tailed test)

Under H_0, the test statistic $z = \dfrac{p - P}{\sqrt{PQ/n}} = \dfrac{0.90 - 0.85}{\sqrt{(0.85 \times 0.15)/20}} = \dfrac{0.05}{0.08} = 0.625$

Since $|z| = 0.625 < 1.656$, i.e., the calculated value $|z|$ is less than the tabulated value of z, i.e., z_α at the level of significance for right-tailed test, the hypothesis is accepted.

Example 4.5 A large consignment contains bad apples, the exact number of which is not known. A random sample of 500 apples was taken from the consignment and 60 were found to be bad. Obtain the 98% confidence limit for the percentage of bad apples in the consignment.

Solution: Here, $n = 500$, X = Number of bad apples in the sample = 60, p = Proportion of bad apples in the sample = 60/500 = 0.12

Since the critical value of z at 2% level of significance is 2.33, 98% confidence limits (2% level of significance) for population proportion are

$$p \pm 2.33\sqrt{(pq/n)} = 0.12 \pm 2.33\sqrt{(0.12 \times 0.88/500)}$$

$$= 0.12 \pm 2.33 \times \sqrt{0.000212} = 0.12 \pm 2.33 \times 0.01453$$

$$= 0.12 \pm 0.03386 = 0.0861, 0.015386$$

Hence, 98% confidence limit for bad apples in the consignment is (8.61,15.38).

EXERCISE 4.1

1. About 325 men out of 600 men chosen from a big city were found to be smokers. Does this information support the conclusion that the majority of men in the city are smokers?
2. A die is tossed 960 times and it falls with 5 upwards 184 times. Is the die biased?
3. A sample of 900 days is taken from meteorological records of a certain district and 100 of them are found to be foggy. What is the probable limit to the percentage of foggy days?
4. A random sample of 500 fuses was taken from a large consignment and 65 were found to be defective. Show that the percentage of defectives in the consignment almost certainly lies between 8.5 and 17.5

5. A manufacturer claims that only 4% of the products supplied by him are defective. A random sample of 600 products contained 36 defectives. Test the claim of the manufacturer.
6. In a random sample of 200 people in a city, 108 like to purchase imported watches and the remaining like to purchase local watches. Can we conclude that both the imported and the local watches are popular in the city? (Use 2% level of significance.)
7. A bag contains defective articles, the exact number of which is not known. A sample of 100 from the bag gives 10 defective articles. Find the limits for the proportion of defective articles in the bag.
8. From a large lot of mangoes, a random sample of 600 mangoes was drawn and 60 were found to be bad. Find the standard error of the proportion of bad mangoes in this sample. Hence, find the 3σ limits for the percentage of bad mangoes in this lot.
9. About 400 apples are taken at random from a large basket and 40 are found to be bad. Estimate the proportion of bad apples in the basket.
10. A sample of 600 persons selected at random from a large city shows that the percentage of males in the sample is 53. It is believed that the ratio of males to the total population in the city is 0.5. Test whether the belief is confirmed by the observation.
11. Balls are drawn from a bag containing equal number of black and white balls, each ball being replaced before drawing another. In 2250 drawings, 1018 black and 1232 white balls are drawn. Do you suspect some bias on the part of the drawer?
12. A manufacturer claims that only 10% of the articles produced are below the standard quality. Out of a random inspection of 300 articles, 37 are found to be of poor quality. Test the manufacturer's claim at 5% level of significance.

4.5.2 Test for Difference Between Two Sample Proportions

Let two samples of sizes n_1 and n_2 be drawn from two populations with population proportions P_1 and P_2, respectively. Also, let X_1 and X_2 be the number of successes and p_1 and p_2 be the observed proportions of successes in these samples, respectively. Then $p_1 = X_1/n_1$ and $p_2 = X_2/n_2$. Defining $Q_1 = 1 - P_1$ and $Q_2 = 1 - P_2$, we have

$$E(p_1) = P_1,\ E(p_2) = P_2,\ V(p_1) = \frac{P_1Q_1}{n_1} \text{ and } V(p_2) = \frac{P_2Q_2}{n_2}$$

Since for large samples, p_1 and p_2 are normally distributed, their difference is also normally distributed. Therefore,

$$E(p_1 - p_2) = E(p_1) - E(p_2) = P_1 - P_2$$

$$V(p_1 - p_2) = V(p_1) + V(p_2) = \frac{P_1Q_1}{n_1} + \frac{P_2Q_2}{n_2}$$

Hence, $$\text{SE}(p_1 - p_2) = \sqrt{\frac{P_1Q_1}{n_1} + \frac{P_2Q_2}{n_2}} \qquad (4.4)$$

and the standardised variable corresponding to $p_1 - p_2$ is given by