STATISTICAL INFERENCE

1 INTRODUCTION

Statistical inference is that branch of Statistics in which one typically makes a statement about a population based upon the results of a sample. *In one-sample testing*, w*e essentially have to verify whether a population parameter could be equal to a certain proposed value*. Since nothing is known about the parameter, information has to be gathered from the population by selecting a sample which is as unbiased as possible. This sample should, as far as possible, contain most, if not all, the characteristics of its parent population. In other words, it must be the best possible representative of the population from which it comes. From sampling theory, we learn that the most appropriate method of sampling depends entirely on the structure of the population. We will assume that, to the best of our ability, we choose the most unbiased sample. It is worth mentioning, however, that it is very hard, if not impossible, to obtain the ideal sample because we can never eliminate sampling errors completely.

2 ONE-SAMPLE TESTING

If, for instance, we were required to test whether the population mean μ could be equal to a certain value μ_0, it would be quite natural to select a sample and determine the value of the point estimate of the population mean, that is, the sample mean $\bar{x}$, so as to have an approximate idea of the value of the population mean (read the chapter on *Estimation*). The whole problem then boils down to checking how 'far' $\bar{x}$ is from μ_0. 'Far' is subjective, that is, if μ_0 = 50, then someone may find that $\bar{x}$ = 47 is far from 50 but someone else may find that 47 is relatively close to 50. We have to remember that if $\bar{x}$ = 47, we cannot automatically conclude that the population mean can definitely not be equal to 50 just because 47 is not numerically equal to 50. There exist sampling errors which could have caused the sample mean to deviate from the true value of the population mean. The factor which determines how 'far' or 'near' $\bar{x}$ is from μ_0 is the significance level α of the test. Before going into the details of the problem, let us first become familiar with some terms and their respective notations.

When we have to test whether a parameter could be equal to a proposed value, the proposal is formulated as a hypothesis known as the *null hypothesis* denoted by H_0. For example, if we have to test whether the population mean is equal to 50, we would write $H_0 : \mu = 50$. Thus, a null hypothesis is just a formal *statement* where the parameter is *equated* to the proposed value.

In any testing procedure, the principle is to assume that the null hypothesis is true. On the basis of information obtained from a sample, we shall later on decide to accept or reject it. In the case of rejection, it is important that we give a more precise answer than 'the population mean is not equal to 50'. Statistically speaking, one will be more satisfied if the answer is 'the population mean is less than (or more than) 50' in the case when H_0 is rejected. This is the reason why every null hypothesis should always be accompanied by an *alternative hypothesis*, denoted by H_1. It must be ensured that H_0 and H_1 be *mutually exclusive* so that the acceptance of one implies the automatic rejection of the other.

To every null hypothesis, there exist *three* possible alternatives. For example, if $H_0 : \mu = 50$, then

1. $H_1 : \mu < 50$
2. $H_1 : \mu > 50$
3. $H_1 : \mu \neq 50$

are the possible alternatives. The choice of the correct alternative is made according to the formulation of the problem. The significance level α of the test is also known as the *critical region* or the *region of rejection* of H_0 and its location depends on the choice of H_1.

In most cases, we select large samples for the sake of accuracy in our estimation of the population parameter. Consequently, we may make an extensive use of the normal distribution theory as postulated by the *Central Limit Theorem*. In the above example, alternatives (1) and (2) are known as one-sided or one-tailed alternatives whereas the third one is called a two-sided or a two-tailed alternative. The term 'tailed' obviously comes from the lower and upper 'tails' of the normal distribution. We now show how the location of the critical region fluctuates with the choice of the alternative hypothesis H_1.

$H_0 : \mu = 50$

$H_1 : \mu < 50$

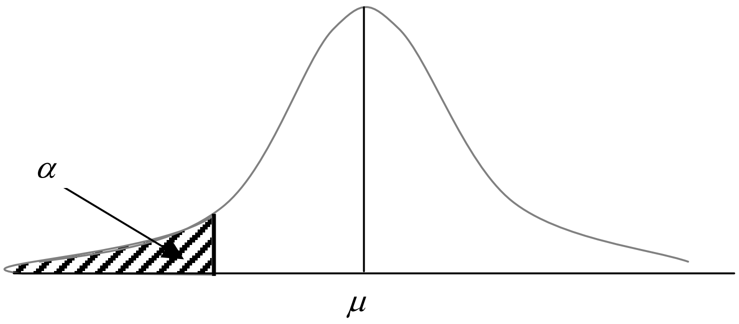

Fig. 2.1

$H_0 : \mu = 50$

$H_1 : \mu > 50$

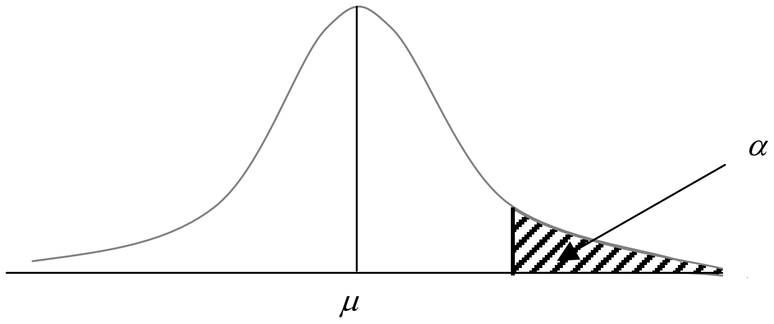

Fig. 2.2

$H_0 : \mu = 50$

$H_1 : \mu \neq 50$

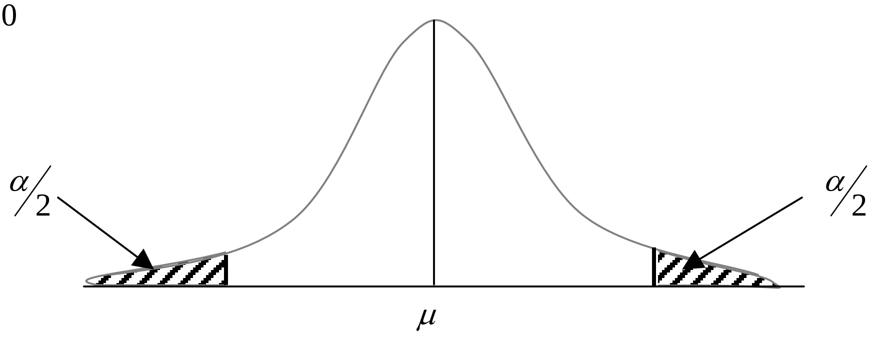

Fig. 2.3

In the following diagram, the *acceptance* and *rejection* regions are shown for a one-tailed alternative to the right. Note that the critical value is found at the boundary of these two regions. For a two-tailed alternative, there will be two critical regions and hence two critical values.

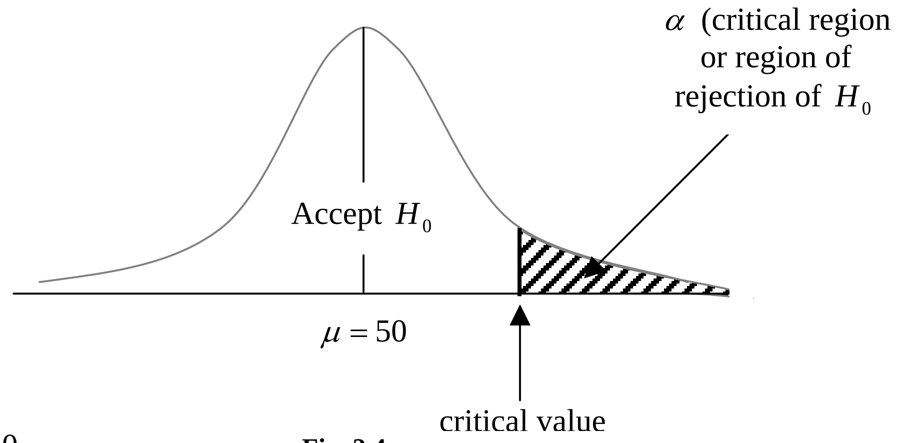

Fig. 2.4

$H_0 : \mu = 50$

$H_1 : \mu > 50$

2.1 Testing procedure

The following steps may be used as a guideline during any testing procedure. However, we have to bear in mind that different sample statistics are required when testing for different population parameters.

1. Formulate the null and alternative hypotheses
2. Depending on the alternative hypothesis and the significance level, define the
3. Perform the test-statistic, including any other relevant calculations
4. Compare the test-statistic value with the critical value(s) in order to decide whether to accept or reject the null hypothesis. Write down a conclusion in the context of the problem.

The test-statistic varies according to the parameter for which we are testing. In general, it is interesting to know that, whenever we use the *z*-test, that is, the normal distribution, the test-statistic is of the form

$$z = \frac{X - E[X]}{\text{var}[X]}$$

where X is the *unbiased point estimator* of the parameter under investigation.

2.2 Testing for the population mean – the *z*-test

When testing for the population mean μ, we first have to check whether the population variance is known. This is because the test-statistic depends on this vital factor. If σ^2 is known, then, no matter how large the sample size is, we use the *z*-test.

If σ^2 is unknown, we have to take the sample size into consideration. If *n* is large (greater than 30), we still use the *z*-test. The flowchart in **Fig. 2.2.1** below illustrates the procedure on how to use the appropriate test-statistic for a given situation in one-sample mean testing.

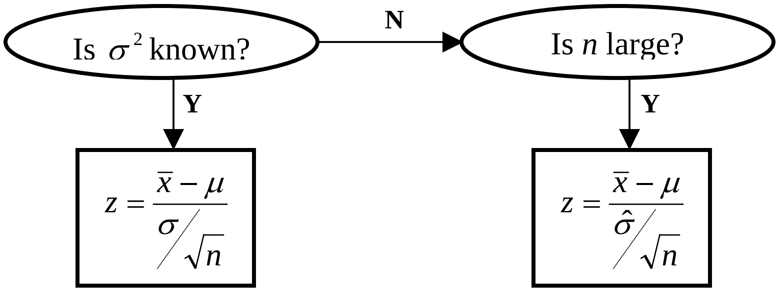
Is σ^2 known?
N
Is n large?
Y
Y
$z = \frac{\bar{x} - \mu}{\sigma / \sqrt{n}}$
$z = \frac{\bar{x} - \mu}{\hat{\sigma} / \sqrt{n}}$

Fig. 2.2.1

Example 1 (Population variance known)

The length of strings in the balls of string made by a particular manufacturer has mean μ m and variance 27.4 m^2. The manufacturer claims that μ = 300. A random sample of 100 balls of string is taken and the sample mean is found to be 299.2 m. Test whether this provides significant evidence, at the 3% level, that the manufacturer's claim overstates the value of μ.

Solution

The claim is that the population mean is 300. Thus, we start by formulating our null hypothesis according to the manufacturer's claim. Now, we have to check whether this is an *overstatement*, that is, whether the true value of the mean is in fact *less than* what is stated in the claim.

Hence, we are testing

$$H_0 : \mu = 300$$

$$\text{against } H_1 : \mu < 300$$

This is a one-tailed alternative to the left and the significance level is 3%. Since the population variance is *known*, we will use the *z*-test.

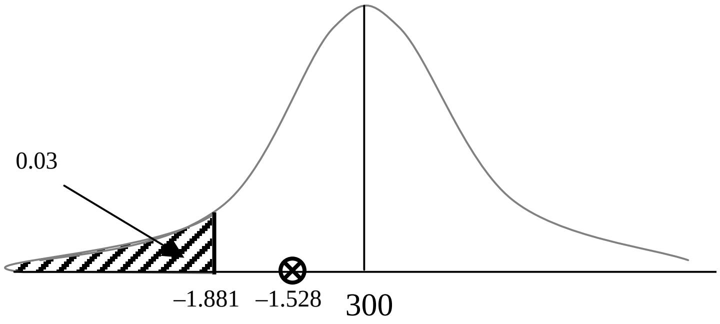

Fig. 2.2.2

The diagram above shows the critical region (shaded) under the null hypothesis that the population mean is 300. The critical *z*-value is –1.881 as obtained from the standard normal table. We know that the sample size *n* is 100 and that the sample mean $\bar{x}$ is 299.2 m.

By using the test-statistic $z = \dfrac{\bar{x} - \mu}{\sigma / \sqrt{n}}$, we obtain $z = \dfrac{299.2 - 300}{\sqrt{27.4 / 100}} = -1.528$.

Since –1.528 > –1.881 (indicated by an encircled cross in **Fig. 2.2.2**), we accept H_0 and conclude that the manufacturer was not overstating the value of the population mean length of string.

Example 2 (Population variance unknown – large sample)

A supermarket manager investigated the lengths of time that customers spent shopping in the store. The time, *x* minutes, spent by each of a random sample of 150 customers was measured, and it was found that $\sum x = 2871$ and $\sum x^2 = 60029$. Test, at the 5% level of significance, the hypothesis that the mean time spent shopping by customers is 20 minutes against the alternative that it is less than this.

Solution

The statement is that the population mean is 20 minutes and we have been given the alternative hypothesis that the mean is less than 20 minutes.

Hence, we are testing

$$H_0 : \mu = 20$$

$$\text{against } H_1 : \mu < 20$$

This is a one-tailed alternative to the left and the significance level is 5%. Since the population variance is *unknown*, we have to check the size of the sample. The value of *n* is 150, which is statistically considered to be large (*n* > 30), so that will use the *z*-test.

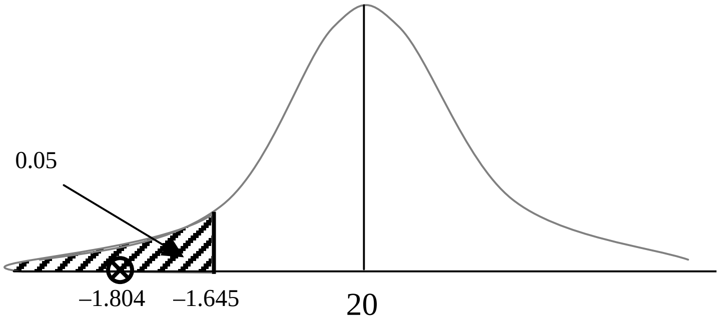

Fig. 2.2.3

The diagram above shows the critical region (shaded) under the null hypothesis that the population mean is 20. The critical *z*-value is –1.645 as obtained from the standard normal table.

We start by calculating the values of $\bar{x}$ and $\hat{\sigma}$ as required in the test-statistic since, this time, the population variance is unknown. From formulae,

$$\bar{x} = \frac{2871}{150} = 19.14 \text{ and } \hat{\sigma}^2 = \left(\frac{150}{149}\right)\left(\frac{60029}{150} - (19.14)^2\right) = 34.081.$$

Using the test-statistic $z = \dfrac{\bar{x} - \mu}{\hat{\sigma}/\sqrt{n}}$, we obtain $z = \dfrac{19.14 - 20}{\sqrt{34.081/150}} = -1.804$.

Since –1.804 < –1.645, (indicated by an encircled cross in **Fig. 2.2.3**), we reject H_0 and conclude that the mean time spent shopping by customers is less than 20 minutes.

2.3 Testing for the population proportion

We often want to know the proportion of individuals in a population which satisfy a certain characteristic. For example, it would be interesting to know the percentage of *left-handed people* in Mauritius or the proportion of books in a library *which contain more than 500 pages*. As usual, it will be assumed that the population is infinite so that information may only be obtained by selecting a sample. The population proportion is denoted by *p*.

In general, when we select individuals, they either satisfy or do not satisfy the characteristic under investigation. If it ever happens that an individual falls in both categories simultaneously (for example, someone *ambidextrous*), then that individual is automatically discarded for the sake of calculations. It is thus quite natural to use the binomial distribution because each individual will either be labelled as '*success*' or '*failure*', depending on whether it satisfies the characteristic or not. If we want to have an idea of the value of *p*, we select a sample of size *n* and count the number, *x*, of individuals satisfying the required characteristic.

We have learnt from the chapter on *Estimation* that the sample proportion $\dfrac{x}{n}$ is an unbiased estimator of the population *p* and has variance $\dfrac{p(1-p)}{n}$. By the *Central Limit Theorem*, for large samples, $\hat{p} \sim N\left(p, \dfrac{p(1-p)}{n}\right)$.

In this course, we will not consider testing for one-sample proportions by means of the *binomial* or *Poisson* distributions but rather their approximations by the normal distribution (without *continuity correction*).

The test-statistic to be used will therefore be

$$z = \frac{\hat{p} - p}{\sqrt{\dfrac{p(1-p)}{n}}}$$

The testing procedure will be identical to that used for testing for a one-sample mean except, precisely, for the test-statistic.

Example

In a public opinion poll, 1000 randomly chosen electors were asked whether they would vote for the '*Purple Party*' at the next election and 357 replied '*Yes*'. The leader of the '*Purple Party*' believes that the true percentage of electors who would vote for his party is 0.4. Test at the 8% level whether he is overestimating his support.

Solution

The leader's belief is formulated as the null hypothesis $H_0 : p = 0.4$. Since we want to know whether he is exaggerating, we have to check if, in fact, the percentage of the population supporting his party is less than 40%. Thus, the alternative hypothesis is $H_1 : p < 0.4$.

Since the sample size is large (n = 1000), we use the normal approximation to the binomial distribution with $np = 1000 \times 0.4 = 400$ and variance $np(1-p) = 1000 \times 0.4 \times 0.6 = 240$. Furthermore, the *sample proportion* $\hat{p} = \dfrac{357}{1000} = 0.357$.

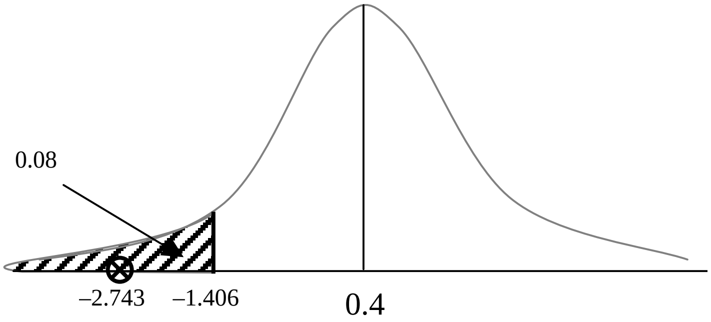

Fig. 2.3.1

The statistic value is $z = \dfrac{\hat{p} - p}{\sqrt{\dfrac{p(1-p)}{n}}} = \dfrac{0.357 - 0.4}{\sqrt{\dfrac{(0.4)(0.6)}{1000}}} = -2.743$.

Since –2.743 < –1.406, reject H_0 and conclude that the '*Purple Party*' leader was indeed overestimating his support.

3 TWO-SAMPLE TESTING

The approach to *two-sample* hypothesis testing is identical to its one-sample counterpart except that, in this case, there is no ‘proposed value’. Comparison is made directly between the values of two population parameters - in general, we test for equality between these parameters (*means* or *proportions*).

One very important aspect to be considered from the statistician’s point of view is that, whenever we test for the difference between two population means, *it is compulsory to test for equality of the population variances first*. This is because the choice of the test-statistic depends on the fact that the variances are equal or not. This condition, however, is not applicable when we test for equality of population proportions.

3.1 TESTING FOR EQUALITY BETWEEN MEANS

It could sometimes prove to be essential to compare the means of two distributions before making an important decision. For example, we might wish to verify whether the mean lifespan of women is longer than that of men in general. Otherwise, we may be tempted to check whether there has been an improvement in the number of marks of students who have been through an intensive training for a certain period. As will be seen later, different types of testing, and hence, test-statistics, would be used for these two cases. But first, let us examine each case in detail and then illustrate it by means of an example.

Large independent samples

When testing for equality of means for two *independent* populations, we start by selecting a sample from each population. For the purpose of this course, *we will assume that the variances of the two populations are equal*. It is then just a matter of testing whether the difference between the two sample means is statistically significant.

If the samples are *large* ($n > 30$), then, according to the *Central Limit Theorem*, we may use the normal distribution theory once more in determining the test statistic to be used.

If X_1 and X_2 are two independent *normal* variables such that $X_1 \sim N(\mu_1, \sigma_1^2)$ and $X_2 \sim N(\mu_2, \sigma_2^2)$, then, using the laws of expectation and variance, the difference between these variables, $X_1 - X_2$, will also follow a *normal* distribution with expectation and variance calculated as follows:

$$E[X_1 - X_2] = E[X_1] - E[X_2] = \mu_1 + \mu_2$$
$$\text{var}[X_1 - X_2] = \text{var}[X_1] + \text{var}[X_2] = \sigma_1^2 + \sigma_2^2\,.$$

We thus write $(X_1 - X_2) \sim N(\mu_1 - \mu_2, \sigma_1^2 + \sigma_2^2)$. Since the respective sample means $\overline{x}_1$ and $\overline{x}_2$, according to the *Central Limit Theorem*, are normally distributed such that $X_1 \sim N\left(\mu_1, \dfrac{\sigma_1^2}{n_1}\right)$ and $X_2 \sim N\left(\mu_2, \dfrac{\sigma_2^2}{n_2}\right)$, following the same procedure as in one-sample testing, we deduce that the test-statistic should be

$$z = \frac{(\overline{x}_1 - \mu_1) - (\overline{x}_2 - \mu_2)}{\sqrt{\dfrac{\sigma_1^2}{n_1} + \dfrac{\sigma_2^2}{n_2}}} = \frac{(\overline{x}_1 - \overline{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\dfrac{\sigma_1^2}{n_1} + \dfrac{\sigma_2^2}{n_2}}}$$

where n_1 and n_2 are the sample sizes (*not necessarily of the same size*) from each population respectively. If the population variances σ_1^2 and σ_2^2 are unknown, we replace them by their respective unbiased estimates $\hat{\sigma}_1^2$ and $\hat{\sigma}_2^2$.

Example

A company has two regional head offices in *Manchester* and *Glasgow*. Workers in the *Glasgow* office claim that they are paid less than the workers in the *Manchester* office. To test this claim, a researcher takes a random sample of 100 workers from each office. The following set of data is recorded:

	Manchester	*Glasgow*
Sample size	100	100
Mean salary	£ 25 700	£ 25 000
Standard deviation	£ 2 000	£ 21000

Table 3.1.1

Using a 5% level of significance, test the claim that the *Glasgow* workers are paid lower salaries on average.

Solution

Let us denote *Manchester* and *Glasgow* as populations 1 and 2 respectively, hence the subscripts for means, variances and sample sizes. We formulate the null and alternative hypotheses as follows:

$$\begin{aligned} H_0 &: \mu_1 = \mu_2 \quad (\mu_1 - \mu_2 = 0) \\ \text{against } H_1 &: \mu_1 > \mu_2 \quad (\mu_1 - \mu_2 > 0) \end{aligned}$$

Since the sample sizes (100) are statistically considered as *large*, we use the normal distribution. The *critical value* corresponding to a significance level of 0.05 is 1.645 from the standard normal table.

****
Note that, since the population variances are unknown, we shall replace them by their unbiased estimators. It is interesting to note that $\hat{\sigma}^2 = \frac{ns^2}{n-1}$ is equivalent to $\frac{\hat{\sigma}^2}{n} = \frac{s^2}{n-1}$. This relationship is very useful in the sense that we no more have to compute the unbiased estimates but can use the sample standard deviations themselves directly in the test-statistic.

****

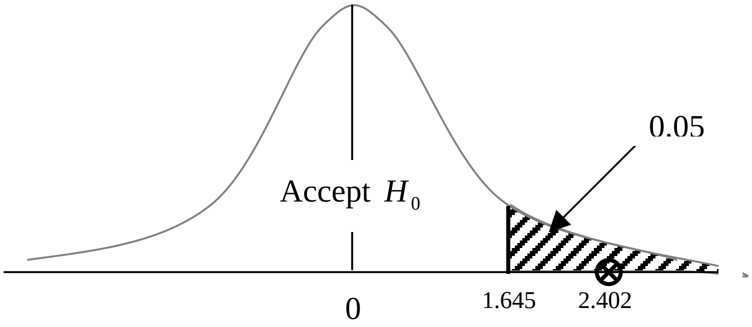

Fig. 3.1.2

Using the information given in **Table 3.1.1**, the test-statistic value is

$$z = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{(n_1 - 1)} + \frac{s_2^2}{(n_2 - 1)}}} = \frac{(25700 - 25000) - (0)}{\sqrt{\frac{(2000)^2}{99} + \frac{(2100)^2}{99}}} = 2.402$$

Since 2.402 > 1.645 (see **Fig. 3.1.2**), we reject H_0 and conclude that *Manchester* workers indeed get paid higher salaries than their *Glasgow* counterparts.

3.2 TESTING FOR EQUALITY BETWEEN PROPORTIONS

We recall that, for a *one-sample* test for the population proportion *p*, its unbiased estimator (sample proportion) followed a *normal* distribution with mean *p* and variance $\frac{p(1-p)}{n}$ according to the *Central Limit Theorem*.

If we extend this theory to *two-sample* testing, that is, when testing for equality of two population proportions p_1 and p_2, a sample will be selected from each population and its sample proportion determined. Assuming that samples of sizes n_1 and n_2 are chosen from populations 1 and 2 and that x_1 and x_2 observations are found that satisfy the characteristic under investigation from the respective populations, then the sample proportions will be

$$\hat{p}_1 = \frac{x_1}{n_1} \text{ and } \hat{p}_2 = \frac{x_2}{n_2}.$$

Thus, $\hat{p}_1 \sim N\left(p_1, \frac{p_1(1-p_1)}{n_1}\right)$ and $\hat{p}_2 \sim N\left(p_2, \frac{p_2(1-p_2)}{n_2}\right)$ according to the *Central Limit Theorem*.

Since the *linear com*bination of two normal distributions is also a normal distribution, we determine the distribution of $(\hat{p}_1 - \hat{p}_2)$ as follows:

$$E[\hat{p}_1 - \hat{p}_2] = p_1 - p_2 \text{ and}$$

$$\text{var}[\hat{p}_1 - \hat{p}_2] = \frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}$$

Therefore, $(\hat{p}_1 - \hat{p}_2) \sim N\left(p_1 - p_2, \frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}\right)$

The test-statistic for two-sample testing for equality for proportions is

$$z = \frac{(\hat{p}_1 - \hat{p}_2) - (p_1 - p_2)}{\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}} = \frac{(\hat{p}_1 - \hat{p}_2)}{\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}}$$

since

1. The null hypothesis is given by $H_0 : p_1 = p_2 \ (p_1 - p_2 = 0)$
2. The values of p_1 and p_2 are unknown, hence estimations being used for the variances.

Example

To verify the percentages of men and women who are *HIV positive* in a certain community, two samples were selected and the following information was recorded:

	Men	*Women*
Sample size	550	720
Number of *HIV positive*	286	396

Table 3.2.1

Can we conclude, at the 5% significance level, that the proportions of *HIV positive* men and women in the community are equal?

Solution

Since we are only testing whether the proportions are equal, there is no specific direction, that is, we use a *two-tailed* alternative hypothesis.

$$H_0 : p_1 = p_2 \quad (p_1 - p_2 = 0)$$
$$H_1 : p_1 \neq p_2 \quad (p_1 - p_2 \neq 0)$$

Denoting the men and women populations by 1 and 2 respectively, the recorded data may be summarised as

$$n_1 = 550 \qquad x_1 = 286 \qquad \hat{p}_1 = \frac{286}{550} = 0.52$$

$$n_1 = 720 \qquad n_1 = 396 \qquad \hat{p}_1 = \frac{396}{720} = 0.55$$

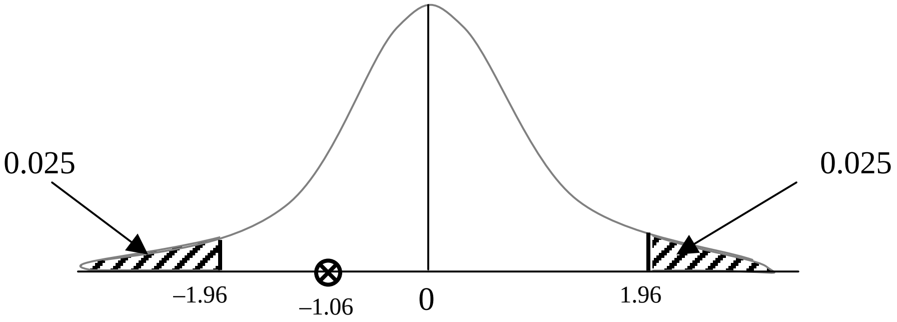

Fig. 3.2.2

The test-statistic value is

$$z = \frac{(\hat{p}_1 - \hat{p}_2)}{\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}} = \frac{(0.52 - 0.55)}{\sqrt{\frac{(0.52)(0.48)}{550} + \frac{(0.55)(0.45)}{720}}} = -1.06$$

Since –1.96 < –1.06 < 1.96 (see Fig. **3.2.2** above), we do not have enough evidence to reject H_0. We conclude that the proportions of *HIV positive* men and women in that community are equal.