

# Module 5

## Hypotheses Tests: Comparing Two Groups

**Objective:** In medical research, we often compare the outcomes between two groups of patients, namely exposed and unexposed groups. At the completion of this module you will learn statistical methods to evaluate whether the observed difference between the two groups is due to chance or sampling variability – in other words, whether the observed difference between exposed and unexposed groups is statistically significant.

### 5.1 Introduction

In Module 4, we discussed the confidence intervals which give a range of likely values for a single parameter (e.g., true mean, true proportion) or for the difference between two parameters (e.g., difference in two true means, difference in two true proportions). In this module, we extend the idea of confidence intervals to hypotheses tests where we compare outcomes in two groups of patients. Using hypotheses tests, we investigate whether the data provides evidence that the treatment/exposure actually affects the outcome, or equivalently whether the observed difference between two groups has arisen by chance or sampling variability.

In fact, the objectives of confidence intervals and hypotheses tests are very similar. As we will see from this module, one may use confidence intervals and hypothesis testing procedures to arrive at the same conclusion about the research question. Let us consider the study described in Case 1 of Module 4 where the study objective is to investigate whether the means of body mass index (BMI) for male and female cardiac surgery patients in the population are different. Consider the BMI data in Table 4.2 – sample mean BMI for male and female patients are respectively  $29.21 \text{ kg/m}^2$  and  $27.04 \text{ kg/m}^2$  – the difference in sample means is  $2.17 \text{ kg/m}^2$ . What does the difference in sample means tells us about the difference in population means? In other words, what can we say about how much better (or worse) off are the male patients compared to female patients (note that a higher BMI is a risk factor for cardiac disease; the higher the BMI, the worse the patients' condition). This can be addressed by calculating a confidence interval for the difference in true means of BMI for male and female patients; procedures for confidence intervals have already been discussed in Module 4.

We may also wish to evaluate whether the data provides evidence that the exposure (e.g., sex in the above example) actually affects the outcome (e.g., BMI in the above example), or whether the observed difference between the sample means has arisen by chance or sampling variability. In other words, is the data consistent with there being no difference between true means of BMI for male and female patients. We address this issue by using hypotheses or significance tests.

In this module we discuss the following topics:

- Some preliminary concepts of hypotheses tests: Section 5.2
- Steps in hypothesis testing: Section 5.3.
- Comparing two groups: continuous data: Section 5.4
  - True standard deviations are assumed equal: Section 5.4.1
  - True standard deviations are assumed unequal: Section 5.4.2
- Comparing two groups: Paired data: Section 5.5
- Comparing two groups: Categorical data: Section 5.6
- Appendix

New notations:

- Null hypothesis:  $H_0$
- Alternative hypothesis:  $H_A$
- Observed significance level or  $p$ -value:  $p$
- Calculated  $t$ -score:  $T_{cal}$
- Calculated  $Z$ -score:  $Z_{cal}$

## 5.2 Basic Concepts of Hypotheses Testing

In this section we will discuss some basic concepts which are essential to an understanding of hypothesis tests, they are:

- Statistical hypotheses
- Types of statistical hypotheses
- Type I and II errors in hypothesis test
- Significance level
- Observed significance level or  $p$ -value
- Steps in hypotheses tests

### 5.2.1 Statistical Hypotheses

A hypothesis may be defined simply as a statement about one or more populations (or equivalently parameters). Statistical hypotheses are hypotheses that are stated in such a way that they may be evaluated by appropriate statistical techniques. For example, a hospital administration may hypothesize that the average length of stay of patients admitted to the Alfred hospital is higher than that of patients admitted to the Monash Medical Centre. A physician may hypothesize that a new drug will be more effective than an old drug for reducing pain for prostate cancer patients. A researcher may hypothesize that the BMI for diabetic patients is higher than that for non-diabetic patients. As we will see later in this module, all of these hypotheses can be evaluated using a statistical method called test of hypotheses or hypotheses tests.

### 5.2.2 Types of Statistical Hypotheses

There are two statistical hypotheses involved in hypotheses tests. They are:

- Null hypothesis and
- Alternative hypothesis.

#### Null Hypotheses

The null hypothesis is sometimes referred to as a hypothesis of no difference and is denoted by  $H_0$ . Let us consider the cardiac surgery study discussed earlier where we are interested in comparing means of BMI for male and female cardiac surgery patients in the population. The null hypothesis for this study can be stated as follows:

Null hypothesis: There is no difference between mean BMI for male and female patients in the population.

In notation, the above hypothesis can also be written as:

$$H_0 : \mu_1 - \mu_2 = 0 \text{ or, } H_0 : \mu_1 = \mu_2$$

Here,  $\mu_1$  and  $\mu_2$  are the true mean BMI for the male and female patients respectively. Note that in reality we do not know whether the null hypothesis stated for a study is true or false, we just evaluate it on the basis of sample data.

#### Alternative Hypothesis

The alternative hypothesis is a second statement that contradicts the null hypothesis. It is a statement of what we will believe is true if our sample data causes us to reject the null hypothesis. We denote the alternative hypothesis by  $H_A$ . The alternative hypothesis for the BMI study is given by:

Alternative hypothesis: The mean BMI for male and female patients in the population are different.

In notation, this hypothesis is:  $H_A : \mu_1 - \mu_2 \neq 0$  or,  $H_A : \mu_1 \neq \mu_2$

This hypothesis is also called the two-sided hypothesis. In medical research usually a two-sided hypothesis is used because the effect of a group of patients (e.g., patients treated by treatment) could be in either direction. That is a treatment could have positive effect or negative effect on the outcome interests in a study. A two-sided alternative is a convention to use a tow-sided alternative hypothesis because in this way either a protective or detrimental effect of a treatment could be demonstrated.

The null and alternative hypotheses can be evaluated using hypothesis testing procedures. As we will see later in this module that in the testing process either we reject or do not reject the null hypothesis. If the null hypothesis is not rejected, we say that the data on which the test is based does not provide sufficient evidence to cause rejection and the difference between the mean BMI for male and female patients in the population is not statistically significant. However, if the testing procedure leads to rejection of the null hypothesis, we say that the data at hand is not compatible with the null hypothesis and the difference is statistically significant. Detailed discussion on evaluation of hypotheses will be discussed later in this module.

**Note:** Do not say “accept the null hypothesis”. The correct wording is “do not reject the null hypothesis” – i.e., the double negative.

### 5.2.3 Types of Errors

As mentioned earlier we do not know the condition of the null hypothesis, that is, whether the hypothesis is true or false. In fact, in the testing procedure we assume that the null hypothesis is true and then we evaluate this hypothesis on the basis of sample data. However, conclusions derived on the basis of sample data could have been affected by the sampling variation. Thus we may reject a null hypothesis when in fact it was true. Similarly, we may not reject the null hypothesis when it was false. Thus, in the hypothesis testing procedure there are two main kinds of errors that can occur, they are:

- Type I Error
- Type II Error

#### Type I Error

A type I error, also known as a rejection error or false positive, is made if we incorrectly reject the true null hypothesis. Let us assume that the null hypothesis – the population means of BMI for male and female patients are the same – is true. However, the test result rejects the null hypothesis. Thus, we have committed an error by rejecting a true null hypothesis. This error is known as type I error.

## Type II Error

The second kind of error that one can commit during a hypothesis testing procedure is a *type II error* or false negative. A type II error is made if we fail to reject a false null hypothesis. Suppose the null hypothesis – the population mean BMI of male and female patients are different – is false. However, the test results do not reject the null hypothesis. Thus, by supporting a false null hypothesis we commit an error which is called type II error.

The null hypothesis stated for a study is either true or false and a hypothesis testing procedure either rejects or does not reject the null hypothesis. Thus at the end of hypothesis testing procedures we must end up with one of the following decisions:

- Do not reject the null hypothesis when the null hypothesis is true – CORRECT decision.
- Do not reject the null hypothesis when the null hypothesis is false – WRONG decision (type II error).
- Reject the null hypothesis when the null hypothesis is true – WRONG decision (type I error)
- Reject the null hypothesis when the null hypothesis is false – CORRECT decision.

The above four situations can also be summarized in the following table (students are requested to understand this table, please!).

Test Result	Condition of the Null Hypothesis $H_0$	
	$H_0$ is True	$H_0$ is False
Do not reject the Null $H_0$	Correct decision	<i>Type II Error</i>
Reject the Null $H_0$	<i>Type I Error</i>	Correct decision

### 5.2.4 Significance Level and Power of a Test

**Significance Level:** A type I error occurs if the test results reject a true null hypothesis. The probability of a type I error is called the significance level of the test and is denoted by the Greek letter  $\alpha$  (alpha) – you will be learning this later.

**Power:** A type II error occurs if the test results do not reject the null hypothesis when the null hypothesis is false. The probability of occurring a type II error is denoted by the Greek letter  $\beta$  (beta).

One important issue in hypothesis testing is to calculate the power of a test. **The power of a test** is the probability of rejecting the null hypothesis when the null hypothesis is false. Recall the additive property of probability discussed in Module 2, the rule is: sum of total probabilities of mutually exclusive events is equal to one. Using this rule we can write, “probability of rejecting the null hypothesis when it is false + probability of not rejecting the null hypothesis when it is false” = 1. So, the probability of rejecting the null hypothesis when it is false = (1 – probability of type II error), which is the power of the test. In notation this can be written as follows:

$$\text{Power} = 1 - \beta = 1 - \text{Probability of type II error}$$

In hypothesis testing procedures, we try to keep both the probability of a type I error ( $\alpha$ ) and the probability of a type II error ( $\beta$ ) as small as possible. This attempt is a tradeoff, because making the  $\alpha$  small requires rejecting the null hypothesis less often. However making  $\beta$  small involves not rejecting the null hypothesis less often. Thus the aims of keeping both  $\alpha$  and  $\beta$  small are contradictory. In fact,  $\beta$  decreases, as  $\alpha$  increases and as  $\alpha$  decreases,  $\beta$  increases. So the general strategy of a test is to fix  $\alpha$  at some specific level and minimize  $\beta$ . A common practice is to fix the  $\alpha$  at some threshold value (e.g. at 0.05), the significance level of hypotheses tests.

### 5.2.5 Observed Significance Level or $p$ -value

Consider the BMI example discussed earlier. Let us assume that  $\mu_1$  and  $\mu_2$  are the true means of BMI for male and female cardiac surgery patients and we want to test the null hypothesis that they are the same, that is,  $H_0 : \mu_1 = \mu_2$ . The alternative hypothesis is that the population means are different, i.e.,  $H_A : \mu_1 \neq \mu_2$ . A hypothesis test begins by postulating that the true means are the same, so that any observed difference between sample means is due to sampling variation. Assuming the null hypothesis is true we calculate the probability of getting a difference between sample means as extreme as or more extreme than the observed difference. This probability is known as the  $p$ -value or observed significance level.

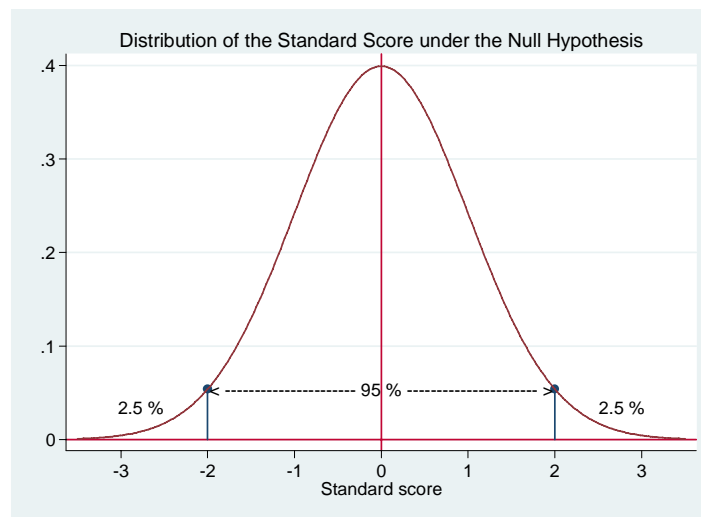
#### How do we calculate the $p$ -value?

Calculation of the  $p$ -value involves a little statistical theory. Most statistical computer packages give the exact  $p$ -value. However, the  $p$ -value can also be calculated manually using appropriate statistical tables. More details on how to calculate the  $p$ -value is discussed below.

Consider the BMI data presented in Table 4.2. The sample mean BMI for male and female patients are respectively 29.21 and 27.04kg/m<sup>2</sup> and their difference is 2.17 g/m<sup>2</sup> – the observed difference. If the observed difference falls within  $\pm 2\text{SE}$  of the population mean difference under the null hypothesis (which is zero), we say that the event is

common and the difference is due to sampling variability – we do not reject the null hypothesis, this means we conclude in favor of the null hypothesis. However, if the observed difference falls outside  $\pm 2SE$  of the population mean difference, the event is rare and the observed difference is not due to sampling variability – we reject the null hypothesis and conclude in favor of the alternative hypothesis. However, there is no shortcut method to check whether the observed difference is rare event or not. In this regard, it is convention to transform the observed difference to a standard score and then use the normal probability table, Table 3.2 to justify if the observed difference is really rare event – the details discussion is follows.

**Figure 5.1 Distribution of the Standard Score under the Null Hypothesis**



**How do we determine whether the observed difference is a rare event?**

We transfer the observed difference to a standard score (SS) by dividing it by its standard error (see Module 4 for the formula for standard error). Note that for large samples, the distribution of the standard score for the difference in two sample means under the null hypothesis is normally distributed with mean zero and standard deviation one – the distribution is shown in Figure 5.1. Further, under the null hypothesis, the sampling distribution for the difference in two sample means and the sampling distribution for its SS are equivalent. Thus, clearly for large sample under the null hypothesis:

- If the observed difference falls within two standard error of zero (difference in population means), the SS score will fall within the interval (-2, 2).
- Alternatively, if the observed difference falls outside two standard error of zero, the SS score will fall outside the interval (-2, 2) – the probabilities outside the interval (-2, 2) is less than 0.05.

Thus, in general (irrespective of sample size) if the sum of the tail probabilities outside the interval of  $(-SS, +SS)$  is smaller than 0.05, the observed difference is a rare event. Otherwise, the observed difference is a common event.

**Note:** Since we are interested in the distance of the observed difference in either direction from the population mean difference under the null hypothesis, we calculate the tail probabilities above  $+SS$  and below  $-SS$  and add them up (see Figure 5.1) – i.e., we calculate probabilities in both tails. This probability is the required  $p$ -value. Thus, for a particular data set the  $p$ -value is the probabilities outside of the interval  $(-SS, +SS)$  in the sampling distribution of the standard score under the null hypothesis.

### **How do we evaluate the hypotheses?**

The idea is that the smaller the  $p$ -value, the stronger is the evidence against the null hypothesis of no difference. In other words, the smaller the  $p$ -value, the lower the probability of getting a difference between the sample means as extreme as or more extreme than the observed difference due to sampling variability. We interpret a  $p$ -value by examining whether it is smaller than a particular threshold value. A common practice is to consider a value of 0.05 as this threshold. This value of 0.05 is called the significance level of hypothesis tests. A  $p$ -value less than or equal to 0.05 is often reported as *statistically significant* and explained as being small enough to justify rejection of the null hypothesis, hence, the hypothesis tests are often been called *significance tests*.

### **In summary:**

- If the  $p$ -value is less than or equal to 0.05, reject the null hypothesis and conclude that the difference is statistically significant. Thus we reject chance (sampling variability) as an explanation for the difference.
- If the  $p$ -value is greater than 0.05, do not reject the null hypothesis and conclude that the difference is not statistically significant. Thus we do not reject chance (sampling variability) as an explanation for the difference.
- If the  $p$ -value is
  - very small (e.g. less than 0.001), there is strong evidence against the null hypothesis.
  - moderately small (e.g. less than or equal to 0.05 but greater than 0.001), increasing evidence against the null hypothesis.
  - Note: consider the above two statements as only conventions!!!

## 5.3 Steps in Hypothesis Testing

A hypothesis test is all about whether to reject or not to reject the null hypothesis. In this section, we will present hypothesis testing discussed above as a four-step procedure. These steps will break the testing procedures down into a logical sequence of actions and decisions. The steps are as follows.

- **Step 1: Hypotheses:**

Null hypothesis: The true means are the same, for example, the true means of BMI for male and female patients are the same or, in notation  $H_0 : \mu_1 - \mu_2 = 0$ .

Alternative hypothesis: The true means are different, for example, the means of BMI for male and female patients in the population are different – in notation this can be written by  $H_A : \mu_1 - \mu_2 \neq 0$ .

Assumptions:

- Each sampled population is normal
- Samples are drawn randomly
- Samples are independent of one another and patients within in each sample are independent of each other.

- **Step 2: Test Statistic**

The test statistic can be calculated by dividing the difference in sample means by the standard error of the difference in sample means. The decision to reject or not to reject the null hypothesis depends on the magnitude of the value of the test statistic. A general formula for test statistic is as follows.

$$\text{Test Statistic} = \frac{\text{Difference in Sample Means}}{\text{SE of Difference in Sample Means}}$$

The value of the test statistic tells us how many standard errors above or below is the difference of sample means from the difference in population means provided the null hypothesis is true. For large as well as small samples we consider t-distribution as the distribution of the Test Statistic, see Module 4. The calculated value of this test statistic is denoted by  $T_{cal}$ .

- **Step 3: The  $p$ -value**

Calculate the  $p$ -value manually following the instructions discussed in Section 5.2.5 or use any statistical packages you are familiar with. Microsoft Excel can also be used to calculate the  $p$ -value for some simple statistical tests.

- **Step 4: Conclusion**

Compare the p-value with the significance level of 0.05. If the p-value is less than the significance level, reject the null hypothesis and conclude that the difference in population means is statistically significant, i.e., the sample data supports a difference in population means.

Alternatively, if the p-value is greater than the significance level (0.05), we do not reject the null hypothesis and conclude that we do not have sufficient evidence in favor of the alternative hypothesis – thus the difference in population means is not statistically significant.

## **5.4 Comparing Two Groups: Continuous data**

Similar to confidence intervals, we also encounter three cases when testing statistical significance for the difference in two true means or simply the difference in two groups. The cases are as follows:

- **Case 1:** True standard deviations for the two groups are unknown, but assumed equal.
- **Case 2:** True standard deviations for the two groups are unknown, but assumed unequal.
- **Case 3:** True standard deviations for the two groups are assumed known.

As was discussed in Module 4, in real life the use of Case 3 is very limited, therefore, we will not discuss this case in this module.

### **5.4.1 Case 1: True SDs are Assumed Equal**

The true standard deviations are usually unknown, but we can have their sample estimates. If the standard deviations in the sample data for the exposed and non-exposed groups are similar, we recommend assuming equality for true standard deviations. Let us consider the following study.

Consider the BMI example where the objective of the study is to investigate whether the means of BMI for male and female cardiac surgery patients in the population are the same. We have random samples of 25 male and 19 female patients from the respective populations; the data was shown in Table 4.2, Module 4. The sample means and standard deviations were calculated in Section 4.3.1, Case 1.

The steps for testing the equality of the means of BMI for male and female patients are as follows:

- Step 1: Null and alternative hypotheses

Null hypothesis: The means of BMI for male and female cardiac surgery patients in Victoria are the same, i.e.,  $H_0 : \mu_1 - \mu_2 = 0$

Alternative hypothesis: The populations means of BMI for this two groups of patients are different, i.e.,  $H_A : \mu_1 - \mu_2 \neq 0$

- Step 2: Calculation of the Test Statistic

$$\text{Test Statistic} = \frac{\text{Difference in Sample Means}}{\text{SE of Difference in Sample Means}}$$

From 4.3.1, Case 1 we have: difference in sample means is 2.17 kg/m<sup>2</sup>, standard error for the difference in sample means is 1.48. Thus the calculated value of the test statistic is:

$$T_{cal} = \frac{2.17}{1.48} = 1.47$$

- Step 3: Calculation of the  $p$ -value

The degree of freedom for the t-distribution is 42. Using t-distribution table, Table 4.7 in Module 4, the  $p$ -value lies between 0.1 and 0.2 (detailed discussion on how to calculate  $p$ -value using statistical table is given at the end of this section). Note that when using a statistical package we can calculate an exact  $p$ -value.

- Step 4: Conclusion:

Since the  $p$ -value is larger than the significance level of 0.05, we do not have sufficient evidence to reject the null hypothesis, that is, the difference between the means of BMI for male and female patients in the population is not statistically significant.

The 95% confidence interval calculated in Module 4 is (-0.833, 5.165 kg/m<sup>2</sup>), which includes the value of zero. Thus we arrive at the same conclusion from the confidence interval and the hypothesis tests.

### How to use Table 4.7 to calculate $p$ -value manually?

Follow the following steps:

- Note that the  $p$ -value in the first row of Table 4.7 is the sum of probabilities in both tails.

- For the above example the calculated value of the test statistic is 1.47 and the d.f. is 42.
- Open the t-distribution table, Table 4.7, the top row in this table shows the  $p$ -values and the first column are degrees of freedom.
- Find the value of  $T_{cal} = 1.47$  along the row with d.f. of 42 or its nearest value which is 40. Along the row with d.f. of 40 there is no value exactly equal to the calculated value of the test statistic – this value lies between 1.303 and 1.684. Therefore, go along to the columns for 1.303 and 1.684 and record the values at the top of these columns which are 0.20 and 0.10. Thus, the required  $p$ -value lies between 0.10 and 0.20.

**Meaning of the above p-value:** If we repeat similar study 100 times (say), between 10 to 20 times the observed difference in sample means will be due to sampling variability or by chance – thus the sampling variability could be an explanation for the observed difference.

## 5.4.2 Case 2: True SDs are Assumed Unequal

We consider this case when standard deviations for each group in the population are unknown but assumed unequal. Let us consider the study discussed in Module 4, Case 2 (Section 4.3.1) where the study objective was to compare the serum creatinine level of the cardiac surgery patients with and without diabetes. The sample data was shown in Table 4.3, Module 4. From Section 4.3.1 (Case 2): (a) The standard error is 0.0162 mmol/L, (b) degree of freedom is 59 and (c) difference in sample means is 0.0235 mmol/L .

The steps for hypothesis testing are as follows:

- Step 1: Null and alternative hypotheses

Null hypothesis: The population mean creatinine levels for diabetic and non-diabetic cardiac patients are the same.

Alternative hypothesis: The mean creatine levels for these two groups of patients are different in the population.

- Step 2: Calculation of the Test Statistic

$$Test\ Statistic = \frac{Difference\ in\ Sample\ Means}{SE\ of\ Difference\ in\ Sample\ Means}$$

The calculated value of the test statistic is

$$T_{cal} = \frac{0.0235}{0.0162} = 1.45$$

- Step 3: Calculation of  $p$ -value

The calculated value for the test statistic is 1.45 and the degree of freedom for the  $t$ -distribution is 59. Using Table 4.7, the  $p$ -value lies between 0.1 and 0.2.

- Step 4: Conclusion

The difference between the mean creatinine level for diabetic and non-diabetic cardiac patients in the population is not statistically significant because the  $p$ -value is greater than 0.05 – the data does not support the rejection of the null hypothesis.

The 95% confidence interval for the difference in population mean creatinine level is (-0.0089, 0.0560 mmol/L) which includes zero. Thus, from the confidence interval and the hypothesis tests we arrive at the same conclusion that the difference in population means is not statistically significant.

## 5.5 Comparing Two Groups: Paired Data

In Section 5.4, we discussed hypothesis testing for the difference between means of two independent populations. However, in this section we discuss the test of significance for the difference in two means of data involving **paired samples**. The null hypothesis to be tested here is – there is no difference between the two population means. The test statistic for testing this hypothesis is as follows:

$$Test\ Statistic = \frac{Mean\ of\ Differences}{SE}$$

Here SE is calculated for the mean of differences. The distribution of this test statistic is also a  $t$ -distribution with  $n - 1$  degrees of freedom, here  $n$  is the number of pairs of patients. The formula for the standard error has already been discussed in Section 4.5 of Module 4.

Consider the study discussed in Section 4.5 of Module 4 where the data was collected on 9 women to investigate whether a low calorie diet helps weight loss for obese women. The data for this study has been presented in Table 4.6. The effectiveness of the prescribed diet will be evaluated using the hypothesis testing procedure carried out below.

- **Step 1: Hypotheses**

Null hypothesis: There is no difference between the mean of weights for the women before and after VLCD.

Alternative hypothesis: The VLCD is effective, i.e., the before and after diet population mean of weights are different.

Assumptions: The differences between weights computed from each pair of data constitutes a simple random sample from a normally distributed population.

Since the data was recorded for each patient in two different occasions, the weights are paired and we use a paired t-test instead of an independent two sample t-test.

- **Step 2: Test Statistic**

The calculated value of the test statistic is given by

$$T_{cal} = \frac{\bar{d}}{SE} = \frac{-22.59}{\frac{5.32}{\sqrt{9}}} = -12.74$$

- **Step 3: The *p*-value**

The calculated *p*-value using Table 4.7 is less than 0.001. The exact *p*-value is ‘0.0000’ (using STATA); we can also calculate exact *p*-value using Microsoft Excel data analysis option.

- **Step 4: Conclusion**

Since the *p*-value is close to zero, there is strong evidence against the null hypothesis of no difference between pre and post diet weights in the population. Hence, we conclude that the difference between the true means of weight for the obese women in the population is statistically significant.

The 95% confidence interval for (After – Before) is (–26.68, –18.50 kg), since the interval does not include zero, the difference is statistically significant. This means that the before and after weights using the VLCD may be different. This finding is also supported by the hypothesis testing procedures performed above. Note: The 95% confidence interval for (Before – After) is ( 18.50, 26.68kg).

## 5.6 Comparing Two Groups: Categorical Data

In previous sections we discussed comparison of two groups using two sample t-test and paired t-test. These test procedures are appropriate only for continuous data. However, in medical research often we compare two groups of patients where data has been recorded on a categorical scale.

Consider the BMI data in Table 4.5, Module 4 where we compare the proportion of male and female patients in the population with BMI > 25 kg/m<sup>2</sup>. We have random samples of 40 male and 50 female patients from the population. Let us assume that  $\pi_1$  and  $\pi_2$  are population proportions for male and female patients who have BMI higher than the above limit.

Here we test the following hypotheses:

Null hypothesis: The proportions of male and female patients with BMI above 25 kg/m<sup>2</sup> are the same in the population, i.e.,  $H_0 : \pi_1 - \pi_2 = 0$

Alternative hypothesis: The proportions of male and female patients in the population with BMI above 25 kg/m<sup>2</sup> are different. In notation,  $H_A : \pi_1 - \pi_2 \neq 0$

The test statistic for testing the above hypotheses is as follows:

$$Z = \frac{\text{Difference in sample proportions}}{\text{SE for the difference in sample proportions}}$$

Here we assume that the sample sizes are reasonably large. Calculated value of the test statistic is denoted by  $Z_{cal}$ .

The standard error for the difference between sample proportions is given by (under the null hypothesis):

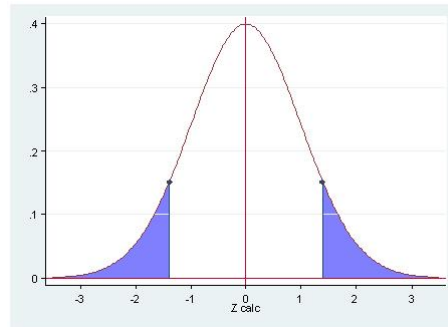
$$SE = \sqrt{\frac{p(1-p)}{n_M} + \frac{p(1-p)}{n_F}}$$

Here,

- $p = \frac{\text{Number of patients with BMI} > 25}{\text{Total number of patients in the study}} = \frac{31 + 32}{40 + 50} = 0.7$ , pooled proportion
- $n_1 = 40$ , number of male patients.
- $n_2 = 50$ , number of female patients.
- $SE = 0.0972$  (using the above formula)
- Difference in sample proportions = 0.135 (see Module 4)

- Test Statistic,  $Z_{cal} = 0.135/0.0972 = 1.39$

Using the normal distribution probability table, Table 3.2 in Module 3, the probability above  $Z_{cal} = 1.39$  is 0.0823, hence the probability below  $-Z_{cal} = -1.39$  is 0.0823 (because the normal distribution is symmetric). Thus the required p-value is 0.1646.



Since the p-value is larger than 0.05, we do not have sufficient evidence to reject the null hypothesis and conclude that the difference is not statistically significant. Thus we arrive at the same conclusion from the hypothesis tests and 95% confidence interval calculated in Module 4.

## Module Summary:

### Terms and notations:

- Null hypothesis & Alternative hypothesis
- Type I error & probability of type I error
- Type II error & probability of type II error
- Power of a test
- Calculated value of a Z-score:  $Z_{cal}$
- Calculated value of a t-score:  $T_{cal}$
- P-value
- Significance level

### Hypothesis test:

- **Continuous data**
  - Test for equality of two true means – assuming equal standard deviations
  - Test for equality of two true means – assuming unequal standard deviations
  - Test for equality of two true means – paired data
- **Categorical data**
  - Test for equality of two true proportions

**Note:** *For any data analyses, both hypothesis test (if any) and 95% confidence interval should be reported. We should arrive at the same conclusion from hypothesis test and 95% confidence interval.*