CHAPTER 8

Testing Hypothesis
Two Sample Tests

In the beginning of Chapter 7, we were referring to the problem of testing whether the shares of large cement companies give a better return than those of small cement companies. This question boils down to the question of testing whether two given samples come from the same population. In this chapter, we will study a method of testing the difference between the means of two populations using large and small samples. In some cases dependent samples are more appropriate for testing the difference of means and we get more precise results while using dependent samples. Finally, we give a method of testing the difference between two proportions.

8.1 HYPOTHESIS TESTING FOR DIFFERENCE BETWEEN MEANS: LARGE SAMPLES

An IT company wants to know whether the programmers and project leaders are equally satisfied with their jobs. A doctor wants to test whether two drugs prescribed by him for diabetes are equally effective. An HR agency wants to test whether there is difference between the salary packages offered to MBAs and MCAs. In all these examples, we are comparing the parameters of two populations. In this section, we will be testing the difference between the means of two populations using large samples.

Let X_1 and X_2 be two populations with mean m_1 and m_2 and standard deviations s_1 and s_2 respectively. For testing the difference between the population means we need to consider the sampling distribution of $\bar{x}_1 - \bar{x}_2$, $\bar{x}_1$ and $\bar{x}_2$ being the means of samples of sizes n_1 and n_2 ($n_1, n_2 \geq 30$) from these populations. As we consider only large samples in this section, $\bar{x}_1$ and $\bar{x}_2$ being means of large samples, follow normal distribution with means $\mu_{\bar{x}_1} = \mu_1$ and $\mu_{\bar{x}_2} = \mu_2$ and standard deviations $\sigma_{\bar{x}_1}$ and $\sigma_{\bar{x}_2}$, respectively. By additive property of normal distributions $\bar{x}_1 - \bar{x}_2$ is normal with mean

$$\boxed{\mu_{\bar{x}_1 - \bar{x}_2} = \mu_1 - \mu_2} \tag{8.1}$$

(In Section (4.8) we have seen that αX_1 and $X_1 + X_2$ are normal distributions when X_1 and X_2 are normal and α is a real number. Also, $\alpha\mu_{x_1} = \alpha\mu_{x_1}$. So, $X_1 - X_2 = X_1 + (-1)\,X_2$ is a normal distribution with $\mu_{\bar{x}_1 - x_2} = \mu_{\bar{x}_1} - \mu_{\bar{x}_2}$. As Variance $(-X_2)$ = Variance (X_2), Standard deviation $(X_1 - X_2)$ = Standard deviation (X_1) + Standard deviation (X_2)

$$\text{Variance}(\bar{x}_1 - \bar{x}_2) = \text{Variance}(\bar{x}_1) + \text{Variance}(\bar{x}_2)$$

$$= \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2} \quad \text{(see Section 5.3.2)}$$

So,

$$\boxed{\sigma_{\bar{x}_1 - \bar{x}_2} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \tag{8.2}$$

The test statistic:

$$\boxed{t = \frac{\bar{x}_1 - \bar{x}_2 - (\mu_1 - \mu_2)}{\sqrt{\dfrac{\sigma_1^2}{n_1} + \dfrac{\sigma_2^2}{n_2}}}} \tag{8.3}$$

Once the test statistic is known, the steps of hypothesis testing can be given as in Chapter 7.

Note: This test is applicable only when the samples are independent. (Two samples are independent if the selection of each sample is independent of the selection of the other sample.)

When σ_1 and σ_2 are not known, we can take the estimates s_1 and s_2 in place of σ_1 and σ_2 (as in Chapter 7).

EXAMPLE 8.1 Two independent samples were selected. The first sample had 40 elements with a mean of 40 and a standard deviation of 5. The second sample had 30 elements with a mean of 50 and a standard deviation of 7.

(a) Compute the standard error of the difference between the sample means.
(b) Test whether the samples come from populations having the same mean at a significance level of 0.05.

Solution: The given data are:

$$n_1 = 40 \quad \bar{x}_1 = 40 \quad s_1 = 5 \quad n_2 = 30 \quad \bar{x}_2 = 44 \quad s_2 = 7$$

(a) $\sigma_{\bar{x}_1 - \bar{x}_2} = \sqrt{\dfrac{s_1^2}{n_1} + \dfrac{s_2^2}{n_2}} = \sqrt{\dfrac{25}{40} + \dfrac{49}{30}} = \sqrt{2.258} = 1.503$

(b) As we want to test whether the sample means are significantly different, we frame H_0 and H_1 as follows:

Step 1 $H_0 : \mu_1 = \mu_2 \quad H_1 : \mu_1 \neq \mu_2$
Step 2 $\alpha = 0.05$ (Given)

Step 3 $t = \dfrac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\text{S.E.}}$

Step 4 As $\alpha = 0.05$ and the test is two-tailed, the critical values are $\pm$ 1.96

Rejection rule: Reject H_0 if the calculated test statistic is greater than 1.96 or less than -1.96

Step 5 Calculated test statistic $= \dfrac{40 - 44 - 0}{1.503} = -2.661$

As the calculated test statistic is less than -1.96, we reject H_0.

So, the samples come from populations having different means.

Note: The method we used in Example 8.1 can be used for testing whether two samples come from two populations whose means differ by a given value. In this case, $H_0 : \mu_1 - \mu_2 = k$, k being a given value.

EXAMPLE 8.2 A hotel in a city was having an occupancy rate of 70% per day with a standard deviation of 18.2% for 50 days. In order to increase the occupancy rate the hotel erected flex boards in prominent locations of the city and found that the occupancy rate rose to 82.7% per day with a standard deviation of 19.7% in the next 75 days. Do you have enough statistical evidence to conclude that the occupancy rate has increased by 5% due to the display of flex boards. Test at a significance level of 0.05.

Solution: Let X_1 and X_2 denote the occupancy rates after and before the erection of flex boards. Let μ_1 and μ_2 denote the average daily occupancy rates per day.

The given data are:

$$\bar{x}_1 = 82.7 \quad s_1 = 19.7 \quad n_1 = 75 \quad \bar{x}_2 = 70 \quad s_2 = 18.2 \quad n_2 = 50$$

As we want to test whether the occupancy rate has increased, we frame H_0 and H_1 as follows:

Step 1 $H_0 : \mu_1 - \mu_2 = 5$ $H_1 : \mu_1 - \mu_2 > 5$

Step 2 $\alpha = 0.05$ (Given)

Step 3 The test statistic is $t = \dfrac{\bar{x}_1 - \bar{x}_2 - (\mu_1 - \mu_2)}{\sqrt{\dfrac{s_1^2}{n_1} + \dfrac{s_2^2}{n_2}}}$

Step 4 As H_1 is $\mu_1 - \mu_2 > 5$, the test is a right-tailed test. So, the critical value is 1.645.

Rejection rule: Reject H_0 if calculated test statistic is greater than 1.645.

Step 5 Calculated test statistic $= \dfrac{82.7 - 70.5}{\sqrt{\dfrac{19.7^2}{75} + \dfrac{18.2^2}{50}}} = \dfrac{7.7}{\sqrt{5.175 + 6.625}}$

$$= \frac{7.7}{\sqrt{11.8}} = \frac{7.7}{3.435} = 2.242$$

As 2.242 > 1.645 we reject H_0. So, we have enough statistical evidence to conclude that the occupancy rate has increased by 5% after the erection of flex boards.

EXAMPLE 8.3 A large engineering company in Chennai purchases a particular component from two suppliers, one in Maharashtra and the other in Haryana. The data regarding 30 orders placed with each of the suppliers are as follows:

The supplier from Maharashtra supplies the components in 12 days on an average (after receiving the order) with a standard deviation of 3 days. The supplier from Haryana takes 14 days on an average for delivering the components with a standard deviation of 2 days.

Test whether the supplier from Haryana is less prompt in delivering the components at a significance level of 0.05.

Solution: Let μ_1 and μ_2 be the average time taken by the suppliers from Maharashtra and Haryana for delivering the components.

The given data are:

$$\bar{x}_1 = 12 \; s_1 = 3 \; n_1 = 30 \text{ (Maharashtra)}$$
$$\bar{x}_2 = 14 \; s_2 = 2 \; n_2 = 30 \text{ (Haryana)}$$

A supplier is less prompt in delivery when he takes more time for delivery. So, we frame H_0 and H_1 as follows:

$$H_0 : \mu_1 = \mu_2 \; H_1 : \mu_1 < \mu_2$$
$$\alpha = 0.05 \text{ (Given)}$$

The test statistic is

$$t = \frac{\bar{x}_1 - \bar{x}_2 - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

As $\alpha = 0.05$ and the test is a left-tailed. The critical value is −1.645.

Rejection rule: Reject H_0 if calculated test statistic is less than −1.645.

$$\text{Calculated test statistic} = \frac{12-14}{\sqrt{\frac{3^2}{30}+\frac{2^2}{30}}} = \frac{-2\sqrt{30}}{\sqrt{13}} = \frac{-2(5.477)}{3.606} = -3.038$$

As −3.038 < −1.645 we reject H_0.

So, the supplier from Haryana is less prompt in delivery at a significance level of 0.05.

8.2 HYPOTHESIS TESTING FOR DIFFERENCE BETWEEN MEANS: SMALL SAMPLES

In this section, we start with two samples of sizes n_1 and n_2 taken from two normal populations with m_1 and m_2 as means and σ_1^2, σ_2^2 as variances.

Then,

$$\mu_{\bar{x}_1} = \mu_1 \text{ and } \mu_{\bar{x}_2} = \mu_2$$

We used the formula $\sigma_{\bar{x}_1-\bar{x}_2} = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}$ for calculating the standard error of $\bar{x}_1 - \bar{x}_2$ in the case of large samples. This was possible since Variance $(\bar{x}_1 - \bar{x}_2)$ = Variance $(\bar{x}_1)$ + Variance $(\bar{x}_2) = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}$.

In the case of small samples, $\bar{x}_1$ and $\bar{x}_2$ follow t-distribution with $n_1 - 1$ and $n_2 - 1$ as degrees of freedom respectively. So, $\bar{x}_1 - \bar{x}_2$ follows a t-distribution only when $\sigma_1^2 = \sigma_2^2$ (σ_1^2 and σ_2^2 are variances of the populations from which the samples are taken). So, the method we are going to discuss is applicable only when $\sigma_1^2 = \sigma_2^2$. (The method of testing whenever $\sigma_1^2 \neq \sigma_2^2$ is given in Chapter 10).

Thus, we are going to test whether there is significant difference between means of two populations whose variances are equal.

Let s^2 be the common value of σ_1^2 and σ_2^2. Then, $\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$ becomes $\sqrt{\frac{\sigma^2}{n_1} + \frac{\sigma^2}{n_2}} = \sigma\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$. When σ^2 is not known, we need to replace σ^2 by s^2. We take a weighted average of s_1^2 and s_2^2 with their degrees of freedom as weights and denote the weighted average by s_p^2 (s_p stands for 'pooled estimate'). So,

$$s_p^2 = \frac{(n_1 - 1)\, s_1^2 + (n_2 - 1)\, s_2^2}{(n_1 - 1) + (n_2 - 1)} \quad \text{i.e.,}$$

$$\boxed{s_p^2 = \frac{(n_1 - 1)\, s_1^2 + (n_2 - 1)\, s_2^2}{n_1 + n_2 - 2}} \tag{8.4}$$

So, in the case of two small samples of sizes n_1 and n_2

$$\boxed{\hat{\sigma}_{\bar{x}_1 - \bar{x}_2} = s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \tag{8.5}$$

and the test statistic is

$$\boxed{t = \frac{\bar{x}_1 - \bar{x}_2 - (\mu_1 - \mu_2)}{s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}} \tag{8.6}$$

Once we know the test statistic, the other steps are as in single small sample test (discussed in Chapter 7).

EXAMPLE 8.4 A business researcher wants to test the fuel efficiency of two cars A and B and collects the following information from 12 users of car A and 9 users of car B:

$\bar{x}_A$ = 19 km/litre $\quad s_A$ = 3.8 km/litre

$\bar{x}_B$ = 24 km/litre $\quad s_B$ = 4.3 km/litre

Test whether the average mileage offered by car B is better than the average mileage offered by car A at a significance level of 0.01.

Solution: Let μ_1 and μ_2 be the average mileage yielded by cars A and B. We are given that:

$$\bar{x}_A = 19 \qquad s_A = 3.8 \qquad n_1 = 12$$
$$\bar{x}_B = 24 \qquad s_B = 4.3 \qquad n_2 = 9$$

So, $n_1 + n_2 - 2 = 12 + 9 - 2 = 19$

We want to test whether the mileage offered by car B is better than the mileage offered by car A. So, we frame H_0 and H_1 as follows:

$$H_0 : \mu_1 = \mu_2 \; H_1 : \mu_1 < \mu_2$$
$$\alpha = 0.05 \text{ (Given)}$$

The test statistic is
$$t = \frac{\bar{x}_A - \bar{x}_B - (\mu_1 - \mu_2)}{s_p\sqrt{\dfrac{1}{n_1} + \dfrac{1}{n_2}}}$$

It is a t-distribution with 19 degrees of freedom.

$$s_p^2 = \frac{(n_1 - 1)\, s_A^2 + (n_2 - 1)\, s_B^2}{n_1 + n_2 - 2}$$
$$= \frac{11(3.8)^2 + 8(4.3)^2}{19}$$
$$= \frac{306.76}{19} = 16.1453$$

So, $s_p = 4.0181$

As $\alpha = 0.01$ and the test is left-tailed, we find the table value under the column corresponding to 0.02 for 19 degrees of freedom. It is 2.539.

Rejection rule: If calculated test statistic is less than –2.539 we reject H_0.

$$\text{Calculated test statistic} = \frac{19 - 24 - 0}{(4.0181)\sqrt{\dfrac{1}{12} + \dfrac{1}{9}}} = \frac{-5}{(4.0181)\sqrt{0.194}}$$
$$= \frac{-5}{(4.0181)(0.44)} = \frac{-5}{1.768} = -2.828$$

As $-2.828 < -2.539$, we reject H_0.

So, car B offers more mileage than car A.

EXAMPLE 8.5 The average daily wages of 15 labourers engaged in construction sector in Tamil Nadu is ₹ 300 with the standard deviation of ₹ 25. The average daily wages of 10 labourers engaged in constructing sector in Karnataka is ₹ 325 with a standard deviation of

₹ 35. Test whether the daily wages in construction sector in Tamil Nadu is different from the daily wages in construction sector in Karnataka at a significance level of 0.05.

Solution: Let μ_1 and μ_2 be the means of daily wages of labourers in Tamil Nadu and Karnataka. The given data are:

$$\bar{x}_1 = 300 \qquad s_1 = 25 \qquad n_1 = 15$$
$$\bar{x}_2 = 325 \qquad s_2 = 35 \qquad n_2 = 10$$

So, $n_1 + n_2 - 2 = 15 + 10 - 2 = 23$

As we want to test whether the wages in Tamil Nadu and Karnataka are different, we frame H_0 and H_1 as follows:

$$H_0 : \mu_1 = \mu_2 \; H_1 : \mu_1 \neq \mu_2$$
$$\alpha = 0.05 \text{ (Given)}$$

The test statistic is
$$t = \frac{\bar{x}_1 - \bar{x}_2 - (\mu_1 - \mu_2)}{s_p\sqrt{\dfrac{1}{n_1} + \dfrac{1}{n_2}}}$$

It is a t-distribution with 23 degrees of freedom.

$$s_p^2 = \frac{(n_1 - 1)\, s_1^2 + (n_2 - 1)\, s_2^2}{n_1 + n_2 - 2}$$
$$= \frac{14(25)^2 + 9(35)^2}{23}$$
$$= \frac{19775}{23} = 859.783$$

$\therefore$
$$s_p = 29.322$$

As $\alpha = 0.05$ and the test is two-tailed, we find the table value under the column corresponding to 0.05 for 23 degrees of freedom. It is 2.069.

Rejection rule: If calculated test statistic is greater than 2.069 or less than –2.069, we reject H_0.

$$\text{Calculated test statistic} = \frac{300 - 325}{(29.322)\sqrt{\dfrac{1}{15} + \dfrac{1}{10}}} = \frac{-25}{(29.322)(0.408)} = -2.0896$$

As $-2.0896 < -2.069$, we reject H_0.

So, the daily wages in Tamil Nadu and Karnataka and different.

EXAMPLE 8.6 Two samples each of size 10 are drawn from companies belonging to the sectors of Computer Software and Pharmaceuticals and their earnings per share (as on 17th November, 2013) are given in Tables 8.1 and 8.2.