Department of Inter Disciplinary Studies,
Faculty of Engineering,
University of Jaffna, Sri Lanka
MC 3020 - Lect_5 Model answers

1. (***Slide number 07***) It is claimed that in the 2008 Democratic Presidential Nomination Primaries in USA, Senator Barack Obama was preferred by the black voters. To test the claim, a research firm sampled 600 black democrats and found that 384 support the senator and in another sample of 720 non-black democrats 417 support the senator. Construct a 97% confidence interval for the difference between the two populations proportions.

   Solutions:

   97% confidence interval for the difference between the two populations proportions is computed as

$$\left(\hat{p_1} - \hat{p_2} - Z_{\alpha/2} * \sqrt{\frac{\hat{p_1}(1-\hat{p_1})}{n_1} + \frac{\hat{p_2}(1-\hat{p_2})}{n_2}}, \hat{p_1} - \hat{p_2} + Z_{\alpha/2} * \sqrt{\frac{\hat{p_1}(1-\hat{p_1})}{n_1} + \frac{\hat{p_2}(1-\hat{p_2})}{n_2}}\right)$$

   Here, $\hat{p_1} = \frac{384}{600} = 0.64$, $\hat{p_2} = \frac{417}{720} = 0.58$ and $Z_{\alpha/2} = 2.17$ from the statistical Table.

   Then, the 97% confidence interval for $p_1 – p_2$ is

$$\left(0.64 - 0.58 \mp 2.17 * \sqrt{\frac{0.64(1-0.64)}{600} + \frac{0.58(1-0.58)}{720}}\right)$$

$$(0.0017, 0.1183)$$

   Therefore, we can say with 97% confidence that the difference between the two population proportions is approximately 0.0017 and 0.1183.

2. (***Slide number 10***) It is claimed that in the 2008 Democratic Presidential Nomination Primaries in USA, Senator Barack Obama was preferred by the black voters. To test the claim, a research firm sampled 600 black democrats and found that 384 support the senator and in another sample of 720 non-black democrats 417 support the senator. Test the claim using 5% level of significance.

   Solutions:

   Step 1: Define the hypothesis

   $H_0 : p_1 \leq p_2$ (you may use equality sign here, ie $p_1 = p_2$)

   $H_1 : p_1 > p_2$

   Step 2: Test statistics

   The corresponding test statistic is

$$Z = \frac{\hat{p_1} - \hat{p_2} - (p_1 - p_2)}{\bar{p}(1-\bar{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$$

   where

$$\bar{p} = \frac{X_1 + X_2}{n_1 + n_2} = \frac{384 + 417}{600 + 720} = \frac{801}{1320} = 0.61$$

and

$$\hat{p_1} = \frac{384}{600} = 0.64, \quad \hat{p_2} = \frac{417}{720} = 0.58$$

Then

$$Z = \frac{0.64 - 0.58 - (0)}{0.61(1 - 0.61)\left(\frac{1}{600} + \frac{1}{720}\right)} = 1.822$$

Step 3: Critical values and identify critical region

The critical value is $Z_\alpha = Z_{0.05} = 1.645 < 1.822$, the calculated value falls in the critical region area.

Step 4: Statistical conclusion

So, we reject H_0

Step 5: General conclusion

Therefore, we can conclude that Senator Barack Obama is significantly more popular among the black Democrats.

3. (***Slide number 21***) Two different types of drugs 'A' and 'B' are tried on certain patients for increasing weight. Five randomly selected patients were given drug 'A' and 7 randomly selected patients were given drug 'B'. The increases in weight (in pounds) are given below:

$$\text{Drug 'A'} : 8, 12, 13, 9, 3$$

$$\text{Drug 'B'} : 10, 8, 12, 15, 6, 8, 11$$

Assume that the population distributions of the measurements are normal with equal variances. Construct a 95% confidence interval for the difference between the two means.

Solutions:

Since the two independent population distributions are normal, the population variances σ_1^2 and σ_2^2 are unknown but equal. The 95% confidence interval for two means $(\mu_1 - \mu_2)$ is computed as

$$\left(\bar{X} - \bar{Y} - t_{\alpha/2} s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}, \bar{X} - \bar{Y} + t_{\alpha/2} s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}\right)$$

where

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

is the pooled standard deviation from the two sample standard deviations, and the degrees of freedom for t is $n_1 + n_2 - 2$.

Here,

$$\bar{X} = \frac{\sum X}{n_1} = \frac{45}{5} = 9, s_1^2 = \frac{n_1 \sum X^2 - (\sum x)^2}{n_1(n_1 - 1)} = \frac{5 * (467) - (45)^2}{5(5 - 1)} = 15.5$$

$$\bar{Y} = \frac{\sum Y}{n_2} = \frac{70}{7} = 10, s_2^2 = \frac{n_2 \sum Y^2 - (\sum Y)^2}{n_2(n_2-1)} = \frac{7*(754)-(70)^2}{7(7-1)} = 9$$

$$s_p = \sqrt{\frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2}} = \sqrt{\frac{(5-1)15.5+(7-1)9}{5+7-2}} = 3.41$$

$t_{\alpha/2,10} = t_{0.025,10} = 2.2281$ from Table for 10 degrees of freedom. Then, the 95% confidence interval for $\mu_1 - \mu_2$ is

$$\left(\bar{X} - \bar{Y} - t_{\alpha/2} s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}, \bar{X} - \bar{Y} + t_{\alpha/2} s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}\right)$$

$$\left(9 - 10 - 2.2281 * 3.41\sqrt{\frac{1}{5} + \frac{1}{7}}, 9 - 10 + 2.2281 * 3.41\sqrt{\frac{1}{5} + \frac{1}{7}}\right)$$

$$(-5.45, 3.45)$$

Thus, the 95% confidence interval for the difference between the two means is approximately $(-5.45, 3.45)$.

4. (***Slide number 22***) To test effect of a fertilizer on rice production, 64 plots of land having equal areas were chosen. Half of these plots were treated with fertilizer and the other half were untreated. Other conditions were the same. The mean yield of rice on the untreated plots was 4.8 quintals with a standard deviation of 0.4 quintal, while the mean yield on the treated plots was 5.1 quintals with a standard deviation of 0.36 quintal. Construct a 94% confidence interval estimate for the mean difference between the untreated plots and treated plots.

   Solutions:

   Since the two independent population distributions are not assumed to be normal, the population variances σ_1^2 and σ_2^2 are unknown, and the sample sizes n_1 and n_2 are large, the 94% confidence interval for $\mu_1 - \mu_2$ is computed as

$$\left(\bar{X} - \bar{Y} - Z_{\alpha/2}\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}, \bar{X} - \bar{Y} + Z_{\alpha/2}\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}\right)$$

   Here, $n_1 = 32, \bar{X} = 4.8, s_1 = 0.4, n_2 = 32, \bar{Y} = 5.1, s_2 = 0.36$ and $Z_{\alpha/2} = Z_{0.03} = 1.88$, from Table.

   Then the 94% confidence interval estimate for $\mu_1 - \mu_2$ is

$$\left(\bar{X} - \bar{Y} - Z_{\alpha/2}\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}, \bar{X} - \bar{Y} + Z_{\alpha/2}\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}\right)$$

$$\left(4.8 - 5.1 - 1.88\sqrt{\frac{0.4^2}{32} + \frac{0.36^2}{32}}, 4.8 - 5.1 + 1.88\sqrt{\frac{0.4^2}{32} + \frac{0.36^2}{32}}\right)$$

$$(-0.4788, -0.1212)$$

   Thus, the 94% confidence interval estimate for the mean difference between the untreated plots and treated plots is approximately $(-0.4788, -0.1212)$.

5. (**Slide number 30**) Two different types of drugs 'A' and 'B' are tried on certain patients for increasing weight. Five randomly selected patients were given drug 'A' and 7 randomly selected patients were given drug 'B'. The increases in weight (in pounds) are given below:

$$\text{Drug 'A'} : 8, 12, 13, 9, 3$$

$$\text{Drug 'B'} : 10, 8, 12, 15, 6, 8, 11$$

Assume that the population distributions of the measurements are normal with equal variances. Do the two drugs differ significantly with regard to their effect in increasing weight? Use 0.05 significance level.

Solutions:

Step 1: Define the hypothesis

$H_0 : \mu_1 = \mu_2$

$H_1 : \mu_1 \neq \mu_2$

Step 2: Test statistics

Since the two independent population distributions are normal, the population variances σ_1^2 and σ_2^2 are unknown and it is mentioned that the variances are equal, the test statistic is

$$T = \frac{\bar{X} - \bar{Y} - 0}{s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} = \frac{9 - 10}{3.41\sqrt{\frac{1}{5} + \frac{1}{7}}} = -0.5008$$

where

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

is the pooled standard deviation from the two sample standard deviations. Here,

$$\bar{X} = \frac{\sum X}{n_1} = \frac{45}{5} = 9, s_1^2 = \frac{n_1 \sum X^2 - (\sum x)^2}{n_1(n_1 - 1)} = \frac{5 * (467) - (45)^2}{5(5 - 1)} = 15.5$$

$$\bar{Y} = \frac{\sum Y}{n_2} = \frac{70}{7} = 10, s_2^2 = \frac{n_2 \sum Y^2 - (\sum Y)^2}{n_2(n_2 - 1)} = \frac{7 * (754) - (70)^2}{7(7 - 1)} = 9$$

$$s_p = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}} = \sqrt{\frac{(5 - 1)15.5 + (7 - 1)9}{5 + 7 - 2}} = 3.41$$

Step 3: Critical values and identify critical region

The critical values are $\pm t_{\alpha/2,df} = \pm t_{0.025,10} = \pm 2.2881$, the calculated value falls in the non critical region area.

Step 4: Statistical conclusion

we do not reject H_0

Step 5: General conclusion

Therefore, we can conclude that the means are not significantly different.

6. (***Slide number 31***) To test effect of a fertilizer on rice production, 64 plots of land having equal areas were chosen. Half of these plots were treated with fertilizer and the other half were untreated. Other conditions were the same. The mean yield of rice on the untreated plots was 4.8 quintals with a standard deviation of 0.4 quintal, while the mean yield on the treated plots was 5.1 quintals with a standard deviation of 0.36 quintal. Can we conclude that there is a significant improvement in rice production because of the fertilizer at 4% level of significance?

   Solutions:

   Step 1: Define the hypothesis

   $H_0 : \mu_1 \geq \mu_2$

   $H_1 : \mu_1 < \mu_2$

   Step 2: Test statistics

   Since the two independent population distributions are not assumed to be normal, the population variances σ_1^2 and σ_2^2 are unknown, and the sample sizes n_1 and n_2 are large, the test statistic is

$$Z = \frac{\bar{X} - \bar{Y} - 0}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} = \frac{4.8 - 5.1}{\sqrt{\frac{0.4^2}{32} + \frac{0.36^2}{32}}} = -3.15$$

   Step 3: Critical values and identify critical region

   The critical value is $-Z_\alpha = -Z_{0.04} = -1.75 > -3.15$, the calculated value falls in the critical region area.

   Step 4: Statistical conclusion

   We reject H_0

   Step 5: General conclusion

   Therefore, we can conclude that there is a significant improvement in rice production because of the fertilizer.

7. (***Slide number 32***) An urban economist wanted to determine whether the mean price of a home in Lemont is less than the mean price of a home in Naperville. A random sample of homes sold in each neighborhood results in the following statistics, where the means and standard deviations are in thousands of dollars:

$$\text{Lemont} : n_1 = 50, \bar{x_1} = 200, s_1 = 45$$

$$\text{Naperville} : n_2 = 50, \bar{x_2} = 300, s_2 = 75$$

   Test the claim that housing is less expensive in Lemont than in Naperville at the 5% level of significance.

   Solutions:

   Step 1: Define the hypothesis

   $H_0 : \mu_1 \geq \mu_2$

   $H_1 : \mu_1 < \mu_2$

   Step 2: Test statistics

   Since the two independent population distributions are not assumed to be normal, the population variances σ_1^2 and σ_2^2 are unknown, and the sample sizes n_1 and n_2 are large, the test statistic is

$$Z = \frac{\bar{X} - \bar{Y} - 0}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} = \frac{200 - 300}{\sqrt{\frac{45^2}{50} + \frac{75^2}{50}}} = -8.0845$$

Step 3: Critical values and identify critical region The critical value is $-Z_\alpha = -Z_{0.05} = -1.645$, the calculated value falls in the critical region area.

Step 4: Statistical conclusion

The test statistic value falls in critical region area, so we reject the null hypothesis.

Step 5: General conclusion

We can conclude that, housing is less expensive in Lemont than in Naperville at the 5% level of significance.

8. (***Slide number 49***) To compare the demand for two different entrees, the manager of a cafeteria recorded the number of purchases for each entrée on seven consecutive days. The data are shown in the next table: Since the cafeteria demands depend on days of the week, the data are considered to be related. Assume that the differences of the measurements follow a normal distribution.

	Monday	Tuesday	Wednesday	Thursday	Friday	Saturday	Sunday
A	420	374	434	395	637	594	679
B	391	343	469	412	538	521	625

(a) Construct a 95% confidence interval for the mean difference.

(b) Test the hypothesis that the demand for item A is higher than the demand for item B. Use 5% level of significance.

Solutions:

(a) 95% confidence interval for the mean difference is given by

$$\left(\bar{D} - t_{\alpha/2,n-1} * \frac{s_D}{\sqrt{n}}, \bar{D} + t_{\alpha/2,n-1} * \frac{s_D}{\sqrt{n}}\right)$$

where the degrees of freedom for t is n–1 and

$$s_D^2 = \frac{n \sum D^2 - (\sum D)^2}{n(n-1)}$$

the sample standard deviation for the differences.

Days	*Mon*	*Tue*	*Wed*	*Thu*	*Fri*	*Sat*	*Sun*
A	420	374	434	395	637	594	679
B	391	343	469	412	538	521	625
D = A – B	29	31	–35	–17	99	73	54

$$\bar{D} = \frac{\sum D}{n} = \frac{234}{7} = 33.43, s_D = \sqrt{\frac{n \sum D^2 - (\sum x)^2}{n(n-1)}} = \frac{7 * (21362) - (234)^2}{7(7-1)} = 47.5$$

95% confidence interval for the mean difference is given by

$$\left(\bar{D} - t_{\alpha/2,n-1} * \frac{s_D}{\sqrt{n}}, \bar{D} + t_{\alpha/2,n-1} * \frac{s_D}{\sqrt{n}}\right)$$

$$\left(33.43 - 2.4469 * \frac{47.5}{7}, 33.43 + 2.4469 * \frac{47.5}{7}\right)$$

$$(-10.5, 77.36)$$

(b) **Step 1: Define the hypothesis**

$H_0 : \mu_A \leq \mu_B \equiv \mu_D \leq 0$

$H_1 : \mu_A > \mu_B \quad\quad \mu_D > 0$

Step 2: Test statistics

The applicable test statistic is:

$$T = \frac{\bar{D} - \mu_D}{s_D/\sqrt{n}} = \frac{33.43 - 0}{47.5/\sqrt{7}} = 1.86$$

Step 3: Critical values and identify critical region

The critical value is $t_{\alpha,df} = t_{0.05,6} = 1.9432 > 1.86$, the calculated value falls in the non critical region area.

Step 4: Statistical conclusion

we do not reject H_0,

Step 5: General conclusion

Hence, we conclude that the demand for item A is not significantly higher than the demand for item B.

9. (***Slide number 52***) A test preparation company claims that its SAT preparation course improves SAT math scores. The company administers the SAT to 9 randomly selected students and determines their scores. The same students then participate in the course. Upon completion, they retake the SAT. The results are presented below:

$$\text{Before} : 436, 431, 270, 463, 528, 377, 397, 413, 525$$

$$\text{After} : 443, 429, 287, 501, 522, 380, 402, 450, 548$$

Test the claim that the preparatory course improves SAT math scores at the 10% level of significance. (Assume that the differences between the scores have an approximate normal distribution.)

Solutions:

Try it yourself. I will upload the solution later on!