Course: Mathematical statistics

Week 8: Inference for two samples

Lecturer: Nagulama Moses

Kumi University

April 24, 2025

Outline

1. Test of difference between two means when variances σ_1^2, σ_2^2 are known
2. Test of difference between two means when variances σ_1^2, σ_2^2 are unknown
3. Test of difference between proportions

Intended learning outcomes

- Understand the assumptions of the Z-test for two means with known population variances.
- Calculate the test statistic using the appropriate t-distribution.
- Formulate hypotheses for testing the difference between two population proportions.
- Calculate the test statistic and critical value for given significance levels.

Test of difference between two means when variances σ_1^2, σ_2^2 are known

Interest is comparing the means of two different populations i,e are the means the same

$$H_0 : \mu_1 = \mu_2, H_0 : \mu_1 - \mu_2 = 0$$

$$H_1 : \mu_1 \neq \mu_2, H_1 : \mu_1 - \mu_2 \neq 0$$

Assumption

- The samples must be independent of each other
- The populations from which the samples are obtained are normally distributed with SD or
- The sample sizes must be both $n \geq 30$

- Recall $\bar{x_1} - \bar{x_2}$ is unbiased estimator $\mu_1 - \mu_2$
- sampling distribution of the $\bar{x_1} - \bar{x_2}$ is

$$\mu_{\bar{x_1} - \bar{x_2}} = \mu_1 - \mu_2$$

$$\sigma^2_{\bar{x_1} - \bar{x_2}} = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}$$

$$z = \frac{(\bar{x_1} - \bar{x_2}) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

is a standard normal random variable

- Test statistic is $z-$ statistic
- For a two tailed test

$$H_0 : \mu_1 = \mu_2$$

$$H_1 : \mu_1 \neq \mu_2$$

- Critical region is $z < -z_{\alpha/2}$ or $z > z_{\alpha/2}$ or $|z| > z_{\alpha/2}$
- For the right tailed test $z > z_\alpha$
- For left tailed test $z < -z_\alpha$

Example

A random sample of 50 hotels in mbale showed that the average hotel room rate is 60000 shillings with SD 5620 shillings. A random sample of

50 hotels from soroti showed that the average hotel room rate is 52200 shillings with SD of 4830 shillings. At 0.05 level of significance difference

in the hotel rate in the two towns, clearly state any assumptions made.

solution

parameter of interest is difference between means $\mu_1 - \mu_2$

$$H_0 : \mu_1 = \mu_2, H_1 : \mu_1 \neq \mu_2$$

level of significance $\alpha = 0.05$
critical region

$$n_1 = n_2 = 50$$

by CLT allows the approximation of hotel rates by a normal distribution samples are independent

$$|z| > z_{\alpha/2} = z_{0.025} = 1.96$$

$$\bar{x_1} = 60000, \bar{x_2} = 52200, \sigma_1 = 5620, \sigma_2 = 4830$$

$$z = \frac{(\bar{x_1} - \bar{x_2}) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

$$z = \frac{(60000 - 52200) - (0)}{\sqrt{\frac{5620^2}{50} + \frac{4830^2}{50}}} = 7.63$$

we reject the null hypothesis. there is a significant difference in hotel rate in the two towns

Example

A standardised mathematics test was given to 75 boys and 50 girls. The

boys had an average mark of 82 with a standard deviation of 8, the girls had a mean mark of 76 with a standard deviation of 6. Assuming that the

score in the test are normally distributed test at 0.05 level of significance that the average score of boys exceed that of girls by 10.

Given:

- $n_1 = 75, \quad \bar{x}_1 = 82, \quad \sigma_1 = 8$
- $n_2 = 50, \quad \bar{x}_2 = 76, \quad \sigma_2 = 6$
- Significance level: $\alpha = 0.05$
- Hypothesized difference: $D_0 = 10$

$$H_0 : \mu_1 - \mu_2 = 10 \quad \text{(Null hypothesis)}$$
$$H_1 : \mu_1 - \mu_2 > 10 \quad \text{(Alternative hypothesis)}$$

$$z = \frac{(\bar{x}_1 - \bar{x}_2) - D_0}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

$$= \frac{(82-76)-10}{\sqrt{\frac{8^2}{75} + \frac{6^2}{50}}} = \frac{6-10}{\sqrt{\frac{64}{75} + \frac{36}{50}}} = \frac{-4}{\sqrt{0.853 + 0.72}} = \frac{-4}{\sqrt{1.573}} \approx \frac{-4}{1.253} \approx -3.$$

Step 3: Critical Value

$$z_{0.05} = 1.645 \quad \text{(from standard normal table)}$$

Decision: Since $z = -3.19 < 1.645$, we **fail to reject** H_0.

Test of difference between two means when variances σ_1^2, σ_2^2 are unknown

- n_1, n_2 are small sample sizes. samples are taken from independent populations that are normally distributed

- case I:variance are unequal
- Test statistic

$$t = \frac{(\bar{x_1} - \bar{x_2}) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

with degree of freedom $n_1 + n_2 - 2$

case II: The variance are unknown but equal.

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

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

Example

It is hypothesised that there is no difference in salary of lecturers in private and those in government universities. A sample of lecturers from each type

of university is selected and the means and standard deviations of their net annual salaries in dollars are obtained the results are shown blow

private $\bar{x}_1 = 26800, s_1 = 600, n_1 = 10$ and public $\bar{x}_2 = 25400, s_2 = 400, n_2 = 8$.

Assume that the samples are taken from population with equal variance and that the salaries are normally distributed is there sufficient evidence at 0.05 level to reject the claim.

solution

parameter of interest $\mu_1 - \mu_2$

$$H_0 : \mu_1 - \mu_2 = 0$$

$$H_1 : \mu_1 - \mu_2 \neq 0$$

$$\alpha = 0.05$$

critical region $|t| > t_{\alpha/2, n_1+n_2-2} = 2.120$

$$t_{0.025,16} = 2.120$$

$$|t| > 2.12$$

variances are equal but unknown

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

$$s_p = \sqrt{\frac{9(360000) + 7(160000)}{16}} = 539.5$$

$$t = \frac{(\bar{x_1} - \bar{x_2}) - (\mu_1 - \mu_2)}{s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

$$t = \frac{(26.800 - 25.400) - 0}{539.5\sqrt{\frac{1}{10} + \frac{1}{8}}} = 5.47$$

reject the null hypothesis and conclude that there is a significant difference in the salaries of lecturers in the two universities.

Example

A standardised test was given to two classes independent samples of sizes $n_1 = 10, n_2 = 10$ we taken from the two classes, the mean score and standard deviation for the two samples are $\bar{x_1} = 80, \bar{x_2} = 78, s_1 = 7.5, s_2 = 6.8.$ assuming that the scores are distributed, is there sufficient evidence at 0.01 level of significance based on this data to believe that students in the first class did better than those in the second class if

(a) the variance of the two population are unequal

(b) if the samples are from two population with equal variances

solution

$$n_1 = 10, n_2 = 10, \bar{x_1} = 80, \bar{x_2} = 78, s_1 = 7.5, s_2 = 6.8$$

parameter of interest is $\mu_1 - \mu_2$

$$H_0 : \mu_1 - \mu_2 = 0, H_1 : \mu_1 - \mu_2 > 0 \alpha = 0.01$$

$$t > t_{\alpha, n_1+n_2-2}, t > t_{0.01,18} = 2.552$$

$$t = \frac{(\bar{x_1} - \bar{x_2}) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} = 0.62$$

we fail to reject the null hypothesis because $0.62 < 2.552$

(b)

$$s_p = \sqrt{\frac{9(56.25) + 9(46.24)}{10 + 10 - 2}} = 7.2$$

$$t = \frac{2}{7.2\sqrt{\frac{1}{10} + \frac{1}{10}}} = 0.62$$

we fail to reject the null hypothesis

Test of difference between proportions

- $\hat{p}_1 = \frac{x_1}{n_1}$ a point estimate of p_1
- $\hat{p}_2 = \frac{x_2}{n_2}$ a point estimate of p_2
- $\hat{p}_1 - \hat{p}_2$ is unbiased estimate of $p_1 - p_2$
- sampling distribution of $\hat{p}_1 - \hat{p}_2$

$$\mu_{\hat{p}_1 - \hat{p}_2} = p_1 - p_2$$

$$\sigma^2_{\hat{p}_1 - \hat{p}_2} = \frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}$$

since p_1 and p_2 are unknown, we use

$$\hat{p} = \frac{x_1 + x_n}{n_1 + n_2}$$

$$= \frac{n_1 \hat{p}_1 + n_2 \hat{p}_2}{n_1 + n_2}$$

variance can be estimated by

$$\sigma^2_{\hat{p}_1 - \hat{p}_2} = \frac{\hat{p}(1-\hat{p})}{n_1} + \frac{\hat{p}(1-\hat{p})}{n_2}$$

$$\sigma_{\hat{p}_1 - \hat{p}_2} = \sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$$

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

Example

A pole is taken to determine if there is a significant difference in the proportion of voters in two sub counties who support a candidate in a parliamentary election. if 120 out of 200 voters in sub county A support

the candidate and a sample of 240 out of 500 from sub county B support the candidate would you agree that the proportion of voters in sub county A who support the candidate is higher than the proportion of voters in sub county B. use a 0.025 level of significance .

solution

$$H_0 : p_1 - p_2 = 0$$

$$H_1 : p_1 - p_2 > 0$$

$$\alpha = 0.025$$

critical region $z > z_{0.025} = 1.96$

$$\hat{p}_1 = \frac{120}{200} = 0.6$$

$$\hat{p}_2 = \frac{240}{500} = 0.48$$

$$\hat{p} = \frac{0.6(200) + 0.48(500)}{200 + 500} = 0.51$$

$$z = \frac{(0.6 - 0.48) - 0}{\sqrt{0.51(0.49)(\frac{1}{200} + \frac{1}{500})}} = 2.87$$

we reject H_0.
The proportion of voters in sub county A is higher

Example

A cigarette manufacturing farm produces two brands of cigarette. it is found that 56 out of 200 smokers prefer brand A and 29 out of 150 smoker prefer brand B

(i) construct a 90% CI for the difference in proportions of smokers for the two brands of cigarettes

(ii) can we conclude at 0.05 level of significance that brand A out sales brand B.

Given:

- Brand A: $x_1 = 56,\ n_1 = 200,\ \hat{p}_1 = \frac{56}{200} = 0.28$
- Brand B: $x_2 = 29,\ n_2 = 150,\ \hat{p}_2 = \frac{29}{150} \approx 0.1933$

Construct a 90% Confidence Interval for the difference in proportions.

The formula for the confidence interval is:

$$\hat{p}_1 - \hat{p}_2 \pm z_{\alpha/2} \cdot \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$$

$$
\begin{aligned}
SE &= \sqrt{\frac{0.28(1-0.28)}{200} + \frac{0.1933(1-0.1933)}{150}} \\
&= \sqrt{\frac{0.2016}{200} + \frac{0.1559}{150}} \\
&= \sqrt{0.001008 + 0.0010393} = \sqrt{0.0020473} \approx 0.0453
\end{aligned}
$$

For 90% CI, $z_{\alpha/2} = 1.645$. Margin of Error:

$$ME = 1.645 \times 0.0453 \approx 0.0745$$

Confidence interval:

$$0.28 - 0.1933 \pm 0.0745 = 0.0867 \pm 0.0745 = (0.0122,\ 0.1612)$$

Interpretation: We are 90% confident that the true difference in proportions lies between 1.22% and 16.12%.

Hypothesis Test at $\alpha = 0.05$ level of significance:

- $H_0 : p_1 = p_2$
- $H_1 : p_1 > p_2$

Compute pooled proportion:

$$\hat{p} = \frac{x_1 + x_2}{n_1 + n_2} = \frac{56 + 29}{200 + 150} = \frac{85}{350} \approx 0.2429$$

Standard error:

$$\begin{aligned} SE &= \sqrt{\hat{p}(1 - \hat{p}) \left(\frac{1}{n_1} + \frac{1}{n_2} \right)} \\ &= \sqrt{0.2429 \cdot 0.7571 \cdot \left(\frac{1}{200} + \frac{1}{150} \right)} \\ &= \sqrt{0.1838 \cdot 0.01167} = \sqrt{0.002174} \approx 0.0466 \end{aligned}$$

Compute test statistic:

$$z = \frac{0.28 - 0.1933}{0.0466} \approx \frac{0.0867}{0.0466} \approx 1.86$$

At $\alpha = 0.05$, the critical z-value is 1.645.
Since $z = 1.86 > 1.645$, we **reject the null hypothesis**.
Conclusion: There is sufficient evidence at the 0.05 level of significance to conclude that brand A outsells brand B.

Example

In a study to estimate the proportion of children who attend video shows when they are supposed to b in school, it was found that 63 out of 100

children in school A attend the show atleast once a month and 59 out of 125 from school B attended the shows atleast once a month. Is there a

significant difference between the proportion of children who attend video shows from the two schools. use 0.05 level of significance

solution

parameter of interest is $p_1 - p_2$

$$H_0 : p_1 - p_2 = 0$$

$$H_1 : p_1 - p_2 \neq 0$$

$$\alpha = 0.05$$

critical region $|z| > z_{\alpha/2} = z_{0.025} = 1.96$

$$\hat{p}_1 = \frac{63}{100} = 0.63$$

$$\hat{p}_2 = \frac{59}{125} = 0.472$$

$$\hat{p} = \frac{100(0.63) + (0.472)125}{100 + 125} = 0.54$$

$$z = \frac{0.16 - 0}{\sqrt{0.54(0.46)(\frac{1}{100} + \frac{1}{125})}}$$

$$z = 2.39$$

we reject the null hypothesis and conclude that there is a significant difference between children who attend the video show

References

- Hogg,R;Mckean,J;Craig,A(2012).Introduction to mathematical statistics, 7th edition, pearson Prentice Hall, 2012.
- Hastings K.J,(1997) Probability and statistics, Addison Wesley reading,massachusetts.

Thank You!

Next Lecture: Chi-square test