Chapter 10 - Hypothesis Testing with Two Populations

Contents

1. Comparing Two Population Means: Independent Samples

1.1. Hypotheses

Null Hypothesis:

$$H_0 : \mu_1 - \mu_2 = D_0 \qquad \text{or} \qquad H_0 : \mu_1 = \mu_2 \text{ if } D_0 = 0$$

D_0 is the hypothesized difference between the two means.

1.2. Conditions for the significance test for $\mu_1 - \mu_2$

- Quantatative varaible with μ_1 and μ_2 defined in context
- Data is obtained using randomization (like simple random sampling)
- Both population distributions are approximately normal (or $n_1 + n_2 \geq 20$)
- Populations are independent, which result in independent samples
- Population standard deviations are unknown.

1.3. Properties of Sampling Distribution of $\bar{X}_1 - \bar{X}_2$

- The mean of the sampling distribution of $\bar{X}_1 - \bar{X}_2$ is $\mu_1 - \mu_2$; that is

$$E(\bar{X}_1 - \bar{X}_2) = \mu_1 - \mu_2$$

- The standard deviation ("standard error") of the sampling distribution of $\bar{X}_1 - \bar{X}_2$ is

$$\sigma_{\bar{X}_1 - \bar{X}_2} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

1.4. Properties of Sampling Distribution $\bar{x}_1 - \bar{x}_2$

Because we don't know population standard deviations, they are estimated using the two sample standard deviations from our independent samples.
The two-sample **test statistic** is calculated as:

$$t = \frac{(\bar{X}_1 - \bar{X}_2) - D_0}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

where

- $\bar{x}_1$ and $\bar{x}_2$ are the sample means
- s_1 and s_2 are the sample standard deviations
- n_1 and n_2 are the sample sizes of each sample.

The Test statistic gives us how many standard errors $\bar{x}_1 - \bar{x}_2$ is away from D_0

1.5. Degrees of Freedom (df)

The t-table is used to find the p-value based on the degrees of freedom.
The test statistic is approximated using the t-distribution with df as follows:

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

1.6. p-value and Conclusion

Example:
The Kona Corporation produces coconut milk. They have both a day shift (called the B shift) and a night shift (call the G shift) to do the process. They would like to know if the day shift and the night shift are equally efficient in processing the coconuts. A study is done sampling 9 G shifts and 16 B shifts. The average number of hours to process 100 pounds of coconuts for G shift and B shift are 2 and 3.2 hours. The sample standard

deviation for the G and B shift are 0.866 and 1.00. The number of hours to process 100 pounds of coconuts for both shifts are normally distributed. Is there a difference in the mean amount of time for each shift to process 100 pounds of coconuts? Test at the 5% level of significance.

- first get all data from the question

$n_1 = 16 \qquad \bar{x} = 3.2 \qquad s_1 = 1$
$n_2 = 9 \qquad \bar{x} = 2 \qquad s_2 = 0.866$

- determine if it meets conditions

  Simple random sample, both populations quantitative, normal population

- get the t-statistic

  To get the t-statistic, plug data into formula:

$$t = \frac{(\bar{X}_1 - \bar{X}_2) - D_0}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} = \frac{(3.2 - 2) - 0}{\sqrt{\frac{(1)^2}{16} + \frac{(0.866)^2}{9}}} = 3.142$$

  now we need to find the degrees of freedom

$$df = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\left(\frac{1}{n_1 - 1}\right)\left(\frac{s_1^2}{n_1}\right)^2 + \left(\frac{1}{n_2 - 1}\right)\left(\frac{s_2^2}{n_2}\right)^2} = \frac{\left(\frac{(1)^2}{16} + \frac{(0.866)^2}{9}\right)^2}{\left(\frac{1}{16} - 1\right)\left(\frac{1^2}{16}\right)^2 + \left(\frac{1}{9} - 1\right)\left(\frac{0.866^2}{9}\right)^2} \approx 18.847$$

  always round df up meaning 19 degrees of freedom.
  We find our t-statistic on the t-table with a value of 2.861.

  Conclusion:
  $t = 3.142 > 2.861$ on t-table, so $\frac{p\text{-value}}{2} < 0.005$
  p-value $< 0.005 < \alpha = 0.05$
  $\therefore$ reject H_0

2. Comparing Two Population Means: Matched Pairs

2.1. Matched Pairs Design

Definition
A matched pairs design is an experimental design where researchers match pairs of participants by relevant characteristics.

Examples

- A study evaluates the effectiveness of a new drug for treating hypertension. The researchers match participants on their age, gender, BMI, and baseline blood pressure and then randomly assign the members of each pair to recieve the drug or a placebo.
- A study evaluates the effectiveness of a new teaching method for slow learners. The researcher matches participants based on their reading IQs, gender, and age. Then they randomly assign the members of each pair to recieve the teaching method.

2.2. Hypothesis

Null Hypothesis:

$$H_0 : \mu_d = D_0 \qquad \text{where} \qquad \mu_d = \mu_1 - \mu_2$$

2.3. Conditions for the Significance Test for μ_d

- A random sample of differences is selected from the target population of differences.
- Two samples are of the same sample size n.
- The differences of the two populations are approximately normally distributed.
- Two samples are paired (dependent).

2.4. Test statistic

The test statistic for matched pair samples is calculated as:

$$t = \frac{\bar{x}_d - \mu_d}{s_d/\sqrt{n}}$$

where

- $\bar{x}_d$ is the mean difference between two samples.
- s_d is the standard deviation of the difference between two samples.
- n is the size of the two samples.

2.5. p-value and Conclusion

- If p-value $< \alpha$, reject the null hypothesis.
- If p-value $> \alpha$, fail to reject the null hypothesis.
- Based on your decision, write a conclusion in terms of the original research question

---

2.6. Example

An experiment is conducted to compare the starting salaries of male and female college graduates who find jobs. Pairs are formed by choosing a male and a female with the same major and similar grade point averages (GPAs). Assume the salaries are Normally distributed. Suppose a random sample of 10 pairs is formed in this manner and starting annual salary of each person is recorded. The results are shown in the following table. Test if the mean starting salary μ_1 for males is higher than the mean starting salary μ_2 for females, using a 0.05 significance level.

Data on Annual Salaries for Matched Pairs of College Graduates

Pair	Male	Female	Difference Male − Female
1	$29,300	$28,800	$ 500
2	41,500	41,600	−100
3	40,400	39,800	600
4	38,500	38,500	0
5	43,500	42,600	900
6	37,800	38,000	−200
7	69,500	69,200	300
8	41,200	40,100	1,100
9	38,400	38,200	200
10	59,200	58,500	700

- First we get all the data from the question

$n_1 = n_2 = 10$
$\bar{x}_d = 400$
$s_d = 434.61$

- Calculate t-statictic

$$t = \frac{\bar{x}_d - \mu_d}{s_d/\sqrt{n}} = \frac{400 - 0}{434.61/\sqrt{10}} \approx 2.91$$

$$df = 9 \quad (n - 1)$$

Cont. next page.

- By t-table:

$t = 2.91$ is between $(2.821, 3.250)$
$\therefore$ p-value is between $(0.005,\ 0.01) < \alpha = 0.05$
$\therefore$ Reject H_0
"There's sufficient evidence (from the sample) that the mean starting salaries for males are higher than the starting salaries for females."

3. Comparing Two Population Proportions: Independent Sampling

3.1. Hypotheses

Nul Hypothesis

$H_0 : p_1 - p_2 = D_0$ or $H_0 : p_1 = p_2$ if $D_0 = 0$

3.2. Conditions for the Significance Test for $p_1 - p_2$

- The two independent samples are randomly selected
- $n_1\hat{p}_1 \geq 5, \quad n_1(1-\hat{p}_1) \geq 5$
- $n_2\hat{p}_2 \geq 5, \quad n_2(1-\hat{p}_2) \geq 5$

3.3. Properties of Sampling Distribution of $\bar{X}_1 - \bar{X}_2$

- The mean of the sampling distribution of $(\hat{p}_1 - \hat{p}_2)$ is $(p_1 - p_2)$; that is

$$E(\hat{p}_1 - \hat{p}_2) = p_1 - p_2$$

- The standard deviation of the sampling distribution $(\hat{p}_1 - \hat{p}_2)$ is

$$\sigma_{p_1 - p_2} = \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}$$

- If the sample sizes n_1 and n_2 are large, the sampling distribution of $(\hat{p}_1 - \hat{p}_2)$ is approximately normal.

By CLT (if sample is large enough):

$$\hat{p}_1 \sim \text{Normal}\left(p_1, \left(\frac{p_1(1-p_1)}{n_1}\right)^2\right)$$

$$\hat{p}_2 \sim \text{Normal}\left(p_2, \left(\frac{p_2(1-p_2)}{n_2}\right)^2\right)$$

$$\hat{p}_1 - \hat{p}_2 \sim \text{Normal}\left(p_1 - p_2, \left(\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}\right)^2\right)$$