ECON 3818

Chapter 21

Kyle Butts

25 August 2021

Chapter 21: Comparing Two Means

Two-Sample Framework

Comparing two populations is one of the most common situations in statistics. These are called **two-sample problems**.

Can divide into groups, A and B

- *Example:* Women vs. Men; Econ Majors vs. Non-Econ Majors; Treated vs. Control

We want to know if they differ along some measurable margin

- *Example:* salary, hours of homework per week, health

This is different from the matched pairs set up because:

- We have a separate sample for each group and we cannot match the observations

Two-Sample Framework

Consider two groups, A and B. You have the following information for each group:

POPULATION/GROUP	SAMPLE MEAN	STANDARD DEVIATION
A	$\bar{X}_A$	σ_A
B	$\bar{X}_B$	σ_B

We use $\bar{X}_A$ and $\bar{X}_B$ to say something about the difference in population means, $\mu_A - \mu_B$

- Construct a confidence interval for $\mu_A - \mu_B$
- Test the hypothesis $H_0 : \mu_A - \mu_B = 0$

Conditions for Two-Sample Inference

We use $\bar{X}_A$ and $\bar{X}_B$ to say something about $\mu_A - \mu_B$

- We have two SRS's from two distinct populations
- The two samples are independent of one another
- We measure the same response variable for both samples
- Both populations are normally distributed
    - In practice, it is enough the distributions have similar shapes and that the data have no strong outliers.

Distribution of $\bar{X}_A$ and $\bar{X}_B$

If the sample mean, $\bar{X}_i \sim N\left(\mu_i, \frac{\sigma_i^2}{n}\right)$ for $i \in A, B$, then:

1. $\bar{X}_A$ and $\bar{X}_B$ is normally distributed
2. $E[\bar{X}_A - \bar{X}_B] = E[\bar{X}_A] - E[\bar{X}_B]$
3. $V[\bar{X}_A - \bar{X}_B] = V[\bar{X}_A] + V[\bar{X}_B]$ (by independence)

To summarize:

$$\bar{X}_A - \bar{X}_B \sim N(\mu_A - \mu_B, \frac{\sigma_A^2}{n} + \frac{\sigma_B^2}{n})$$

Distribution of $\bar{X}_A$ and $\bar{X}_B$

Therefore, when both σ^2 are known:

$$\frac{(\bar{X}_A - \bar{X}_B) - (\mu_A - \mu_B)}{\sqrt{\frac{\sigma_A^2}{n_A} + \frac{\sigma_B^2}{n_B}}} \sim N(0, 1)$$

Distribution of $\bar{X}_A$ and $\bar{X}_B$

As we mentioned, we don't always know the population variance, σ^2.

If we don't know these values, we can use the sample standard deviations s_A and s_B as estimators.

The standard error for the difference in sample means is:

$$SE_{\bar{X}_A - \bar{X}_B} = \sqrt{\frac{s_A^2}{n_A} + \frac{s_B^2}{n_B}}$$

Distribution of $\bar{X}_A$ and $\bar{X}_B$

Since we estimate the sample standard deviations, we should use the t-distribution

$$\frac{\bar{X}_A - \bar{X}_B - (\mu_A - \mu_B)}{\sqrt{\frac{s_A^2}{n_A} + \frac{s_B^2}{n_B}}}$$

can be approximated by the t-distribution, where the degrees of freedom is $min\{n_A, n_B\} - 1$

- Statistical software can be more exact, but the formulas get complicated

Two-Sample Confidence Interval

σ^2 Known

A confidence interval for $\mu_A - \mu_B$ with level of confidence C:

$$(\bar{X}_A - \bar{X}_B) \pm Z^{(1-C)/2} \cdot \sqrt{\frac{\sigma_A^2}{n_A} + \frac{\sigma_B^2}{n_B}}$$

Two-Sample Confidence Interval

σ^2 Known

Say we have two groups -- athletes and non-athletes and we're asked to construct a 95\% confidence interval for the difference in GPA $\mu_A - \mu_{NA}$

GROUP	SAMPLE MEAN	STANDARD DEVIATION	SAMPLE SIZE
Athletes	$\bar{X} = 2.8$	$\sigma = 0.4$	15
Non-athletes	$\bar{X} = 2.9$	$\sigma = 0.5$	25

$$CI = (2.8 - 2.9) \pm Z_{0.025} \cdot \sqrt{\frac{0.4^2}{15} + \frac{0.5^2}{25}} = [-0.38, 0.18]$$

Two-Sample Confidence Interval

σ^2 Unknown

Since we have to estimate σ^2 for both samples, we need to use the t-distribution to find the critical value:

$$(\bar{X}_A - \bar{X}_B) \pm t^{n-1, \frac{1-C}{2}} \cdot \sqrt{\frac{s_A^2}{n_A} + \frac{s_B^2}{n_B}}$$

Two-Sample Confidence Interval

σ^2 Unknown

We have 2 groups of students, and we're asked to construct 90\% confidence interval for difference in test scores, $\mu_A - \mu_B$

GROUP	SAMPLE MEAN	STANDARD DEVIATION	SAMPLE SIZE
Treated	$\bar{X} = 76$	$s = 9$	60
Control	$\bar{X} = 73$	$s = 5$	20

$$CI = (76 - 73) \pm t_{19}^{0.05} \cdot \sqrt{\frac{9^2}{60} + \frac{5^2}{20}} = [0.21, 5.79]$$

Two-Sample Hypothesis Testing

σ^2 Known

Researchers are asking college graduates how old they were when they had their first job. Researchers are curious to see if students who attended state schools got jobs earlier in life than those who attended private colleges.

GROUP	SAMPLE MEAN	STANDARD DEVIATION	SAMPLE SIZE
State Colleges	$\bar{X} = 18.19$	$\sigma = 3.8$	20
Private Colleges	$\bar{X} = 20.98$	$\sigma = 4.2$	20

Test the following hypothesis at the $\alpha = 0.05$ significance level:

$$H_0 : \mu_A - \mu_B = 0$$

$$H_1 : \mu_A - \mu_B < 0$$

Two-Sample Hypothesis Testing

σ^2 *Known*

Calculate p-value using:

$$P(\bar{X}_A - \bar{X}_B \leq 18.19 - 20.98 \mid \mu_A - \mu_B = 0)$$

$$P\left(\frac{\bar{X}_A - \bar{X}_B - (\mu_A - \mu_B)}{\sqrt{\frac{\sigma_A^2}{n_A} + \frac{\sigma_B^2}{n_B}}} \leq \frac{18.19 - 20.98 - (0)}{\sqrt{\frac{3.8^2}{20} + \frac{4.2^2}{20}}}\right)$$

p-value $= P(Z \leq -2.2) = 0.014 \implies \text{reject } H_0$ because p-value $\leq \alpha = 0.05$

Two-Sample Hypothesis Testing

σ^2 Unknown

You want to test how attached individuals are to their friends, and whether that is different across people who volunteer for community service versus those who do not.

GROUP	SAMPLE MEAN	STANDARD DEVIATION	SAMPLE SIZE
Service	$\bar{X} = 105.32$	$s = 14.68$	57
No Service	$\bar{X} = 96.82$	$s = 14.26$	17

Test the following hypothesis at $\alpha = 0.01$ level:

$$H_0 : \mu_A - \mu_B = 0$$

$$H_1 : \mu_A - \mu_B \neq 0$$

Two-Sample Hypothesis Testing

σ^2 Unknown

$$t = \frac{\bar{X}_S - \bar{X}_N - (\mu_A - \mu_B)}{\sqrt{\frac{s_S^2}{n_S} + \frac{s_N^2}{n_N}}} = \frac{105.32 - 96.82 - (0)}{\sqrt{\frac{14.68^2}{57} + \frac{14.26^2}{17}}}$$

$$\implies t = \frac{8.5}{3.9677} = 2.142$$

Look at t-table, row with degrees freedom = 16.

$t_{16}^{0.025} = 2.12$ and $t_{16}^{0.01} = 2.58$, this means p-value is in between 0.025 and 0.01, **BUT** it's a two-tailed test so we need to multiply these probabilities by 2:

$0.02 < p\text{-value} < 0.05 \implies$ Do not reject null at $\alpha = 0.01$

Review of Chapter 21

In this chapter, we focus on making inferences about the relationship between the means of two different samples

- Confidence intervals around the difference in means $\mu_A - \mu_B \pm$ margin of error
- Generally testing $H_0 : \mu_A - \mu_B = 0$
- You'll be given sample means ($\bar{X}$), standard deviations (σ or s) and population size (n) of each sample.

If you're given σ, use Z-distribution

If you're given s, use t-distribution (unless **both** samples are large enough)

Calculating Margin of Error with Two Samples

Variances known:

$$(\bar{X}_A - \bar{X}_B) \pm Z^{\frac{1-C}{2}} \cdot \sqrt{\frac{\sigma_A^2}{n_A} + \frac{\sigma_B^2}{n_B}}$$

Variances unknown:

$$(\bar{X}_A - \bar{X}_B) \pm t^{n-1,\frac{1-C}{2}} \cdot \sqrt{\frac{s_A^2}{n_A} + \frac{s_B^2}{n_B}}$$

Calculating test-Statistic

Variances known:

$$Z = \frac{\bar{X}_A - \bar{X}_B - (\mu_A - \mu_B)}{\sqrt{\frac{\sigma_A^2}{n_A} + \frac{\sigma_B^2}{n_B}}}$$

Varainces unknown:

$$t = \frac{\bar{X}_A - \bar{X}_B - (\mu_A - \mu_B)}{\sqrt{\frac{s_A^2}{n_A} + \frac{s_B^2}{n_B}}}$$

Clicker Question

You're given the following information about average length of careers in NFL versus MLB.

GROUP	SAMPLE MEAN	STANDARD DEVIATION	SAMPLE SIZE
NFL	$\bar{X} = 3.3$	$s = 2.1$	20
MLB	$\bar{X} = 5.6$	$s = 3.5$	17

You want to construct a 90% confidence interval. Given this information, calculate the margin of error:

a. 1.6

b. 1.645

c. 1.96

Midterm Example

New research has developed a new drug designed to reduce blood pressure. In an experiment, 21 subjects were assigned randomly to the treatment group and receive the experimental drug. The other 23 subjects were assigned to the control group and received a placebo treatment. A summary of these data is:

GROUP	SAMPLE MEAN	STANDARD DEVIATION	SAMPLE SIZE
Treatment	$\bar{X} = 23.48$	$s = 8.01$	21
Placebo	$\bar{X} = 18.52$	$s = 7.15$	13

We want to test whether there was any difference in means across these two groups:

a. State the null and alternative hypothesis
b. Calculate p-value or range of p-values
c. Do you reject at $\alpha = 0.05$ level?
d. If you were incorrect in part c, what kind of error did you make?