Module 12
Two-Sample t-Tests

Module Learning Outcomes

- Understand the use or the role of the two variables.
- Define a proper parameter and set up the null hypothesis and alternative hypothesis for two-sample t-tests.
- Check the equal variance assumption.
- Calculate the test statistic and find p-value.

12.1 Use of Variables - Revisit

- In Module 1, we introduced the use (or role) of variables.
- Now that we have two variables, the importance of the use/role should be more apparent.
- When there are two variables in a study, we will need to tell them apart: one is called the **independent variable,** and the other is called the **dependent variable**.
- Another way of calling them are the **response variables** and the **explanatory variables**.
- MLO: Understand the use or the role of the two variables.

12.2 Overview

- In the past two modules, we have looked at hypothesis testing for a single variable (either numerical variable or categorical variable).
- In this and the next modules, we are applying the similar procedure for two variables.
- In fact, there are four main areas we will focus on and they are summarized below.

Response Variable	Exploratory Variable	Procedure Name
Numerical Variable	Categorical Variable	Two-Sample t-Tests Or Matched-Pair t-Tests
Categorical Variable	Categorical Variable	Two-Sample Proportion Tests
Numerical Variable	Numerical Variable(s)	Correlation & Regression (in DANA 4810)
Categorical Variable	Numerical Variable(s)	Logistic Regression (in DANA 4820)

12.3 Assumptions and Review of Notation/Symbols

- Assumption #1: All parameters are unknown to us.
- Assumption #2: The variable of interest (X) is assumed to have a Normal distribution.
- Assumption #3: The variability (measured by the standard deviation) is assumed to be the same between the two groups. This is what most people call the **equal variance assumption**.
- Here are some key notations or symbols used in this procedure.

Symbols	Description
μ_1, μ_2	Average of the populations 1 and 2
$\bar{X}_1, \bar{X}_2$	Average of the samples 1 and 2
σ_1, σ_2	Standard deviation of the populations 1 and 2
s_1, s_2	Standard deviation of samples 1 and 2
s_1^2, s_2^2	Variance of samples 1 and 2
n_1, n_2	Sample Size of samples 1 and 2

- Note: Statistics are used to estimate unknown parameters.

12.4 Two-Sample t-Tests

- The two-sample t-test is used to see how the average (of the numerical response variable) would vary in the two classes of the (categorical) explanatory variable (i.e. two populations).
- As its name implies, we will use the t-distributions to find p-values.
- There is one response variable that is a numerical variable.
  And there is one explanatory variable that is a categorical variable with two classes only.
- FYI: For 3 or more classes in explanatory variable, the procedure is called the ANOVA (or the ANalysis Of VAriance). More of this in DANA 4810.

12.5 Null Hypothesis and Alternative Hypothesis

- Since the response variable is a numerical variable, a natural parameter is the average.
- And because we are making comparison of the explanatory variable between two groups, the most appropriate parameter is the difference between the two population averages, i.e. $\mu_1 - \mu_2$.
- Note: The definition of group 1 and group 2 is arbitrary. It means that with a different definition, the test statistic will be different (one positive and the other one negative). That being said, the p-value and the conclusion should be consistent.
- Otherwise, the setup is exactly the same as Z-tests or t-tests.
- The three sets of hypotheses are listed here for your reference.

	Null Hypothesis	**Alternative Hypothesis**
Lower-tailed test	$H_0: \mu_1 - \mu_2 \geq \Delta_0$	$H_a: \mu_1 - \mu_2 < \Delta_0$
Upper-tailed test	$H_0: \mu_1 - \mu_2 \leq \Delta_0$	$H_a: \mu_1 - \mu_2 > \Delta_0$
Two-tailed test	$H_0: \mu_1 - \mu_2 = \Delta_0$	$H_a: \mu_1 - \mu_2 \neq \Delta_0$

- The symbol (Δ_0) is the **hypothesized difference** and it is pronounced as “delta sub zero”.
- In general, if we only want to compare the two means, then $\Delta_0 = 0$.
- MLO: Define a proper parameter and set up the null hypothesis and alternative hypothesis for two-sample t-tests.

12.6 Test Statistic and p-Values

Scenario #1

- When the equal variance assumption is true, the test statistic is:

$$TS1 = \frac{(\bar{X}_1 - \bar{X}_2) - \Delta_0}{\sqrt{s_p^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

  where the **pooled variance** (s_p^2) is a weighted average of the two sample variances.
- The pooled variance can be calculated by:

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

- As mentioned above, the hypothesized difference is typically zero. Hence, the test statistic can be simplified as:

$$TS1 = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{s_p^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

- There is no single correct way to verify if the two variances are indeed "equal".
- That being said, here is one "quick and dirty" way to check the equal variance assumption. When one sample variance is no more than two times the other sample variance, we can safely believe the equal variance assumption is valid. In other words, we simply check if the ratio of the two variances is between 0.5 and 2.

$$\frac{1}{2} < \frac{s_i^2}{s_j^2} < 2$$

- This test statistic follows the t-distribution with $\nu = n_1 + n_2 - 2$ degrees of freedom.
- In other words, you can find p-value accordingly.

Scenario #2

- If the equal variance assumption is not reasonable, then we will have to use the individual sample variances to compute the test statistic:

$$TS2 = \frac{(\bar{X}_1 - \bar{X}_2) - \Delta_0}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \text{ or } TS2 = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \text{ when } \Delta_0 = 0.$$

- This test statistic follows the t-distribution with a complicated degree of freedom, given as:

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

- As you can imagine, manual calculation of this degree of freedom is not simple. Also note that the value is not likely a whole number (or counting number).
- In other words, as recently as about 30-40 years ago, we did not have computers or software to find the degree of freedom. Hence, finding p-value without the equal variance assumption is very painful.
- As a result, most textbooks or scholars at the time would push for the equal variance assumption even if it is not too reasonable to do so.
- Nowadays, with computers, we can find any degree of freedom (whole number or not) and the corresponding p-value within seconds.
- MLO: Check the equal variance assumption.
- MLO: Calculate the test statistic and find p-value.

12.7 Conclusion

- The conclusion is made exactly the same way as before.