Module 13
Two-Sample Proportion Tests

<u>Module Learning Outcomes</u>

- Define a proper parameter and set up the null hypothesis and alternative hypothesis for two-sample proportion tests.
- Calculate the test statistic.
- Find the p-value and check conditions and make an appropriate conclusion.

13.1 Two-Sample Proportion Tests

- In this module, both the response variable and explanatory variable are categorical variables.
- In both cases, there will only be two classes (or categories) in each categorical variable.
- When either one has more than two classes, the extended version of the procedure is called the χ^2-test of homogeneity, which will be covered in DANA 4820.

13.2 Assumptions and Review of Notation/Symbols

- There typically are not many assumptions when working with categorical variables.
- The only one is that all parameters are unknown to us.
- Here are some key notations or symbols used in this procedure.

Symbols	Description
p_1, p_2	Proportion of the populations 1 and 2
X_1, X_2	Number of Successes of samples 1 and 2
n_1, n_2	Sample Size of samples 1 and 2
$\bar{p}_1, \bar{p}_2$	Proportion of the samples 1 and 2

- Note: Statistics are used to estimate unknown parameters.

13.3 Null Hypothesis and Alternative Hypothesis

- Since the response variable is a categorical variable, a natural parameter is the proportion.
- And because we are making comparison of the explanatory variable between two groups, the most appropriate parameter is the difference between the two population proportions, i.e. $p_1 - p_2$.
- Note: The definition of group 1 and group 2 is arbitrary.
- Otherwise, the setup is very similar to the previous module.
- The three sets of hypotheses are listed here for your reference.

	Null Hypothesis	**Alternative Hypothesis**
Lower-tailed test	$H_0: p_1 - p_2 \geq \Delta_0$	$H_a: p_1 - p_2 < \Delta_0$
Upper-tailed test	$H_0: p_1 - p_2 \leq \Delta_0$	$H_a: p_1 - p_2 > \Delta_0$
Two-tailed test	$H_0: p_1 - p_2 = \Delta_0$	$H_a: p_1 - p_2 \neq \Delta_0$

- 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 proportions, then $\Delta_0 = 0$.
- MLO: Define a proper parameter and set up the null hypothesis and alternative hypothesis for two-sample proportion tests.

13.4 Test Statistic and p-Values

Scenario #1

- If $\Delta_0 = 0$, then it means $p_1 = p_2$ is assumed to be true, under the null hypothesis.
- If so, it is reasonable for us to "group" the two sample proportions and create a "better" estimate of the population proportion (p).
- Hence, the test statistic is:

$$TS = \frac{(\bar{p}_1 - \bar{p}_2) - 0}{\sqrt{\bar{p}(1-\bar{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} = \frac{\bar{p}_1 - \bar{p}_2}{\sqrt{\bar{p}(1-\bar{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

  where the **pooled proportion** ($\bar{p}$) is one sample proportion, after combining the two samples into a single one.
- The pooled proportion can be calculated by:

$$\bar{p} = \frac{X_1 + X_2}{n_1 + n_2}$$

Scenario #2

- If $\Delta_0 \neq 0$, the above pooling of sample proportions is not possible.
- In other words, we will have to use the individual "variances" to proceed.

$$TS = \frac{(\bar{p}_1 - \bar{p}_2) - \Delta_0}{\sqrt{\frac{\bar{p}_1(1-\bar{p}_1)}{n_1} + \frac{\bar{p}_2(1-\bar{p}_2)}{n_2}}}$$

  where $\bar{p}_1 = \frac{X_1}{n_1}$ and $\bar{p}_2 = \frac{X_2}{n_2}$.
- MLO: Calculate the test statistic.

13.5 Finding p-Values and Making Conclusion

- You can find the p-value using the Z-table.
- Note that the p-values are only approximations.
- The approximation is good (enough) when the following 6 conditions are all met (three from group 1 and three from group 2).

1) $n_1 \geq 30$	2) $n_1 \times \bar{p}_1 \geq 5$	3) $n_1 \times (1-\bar{p}_1) \geq 5$
4) $n_2 \geq 30$	5) $n_2 \times \bar{p}_2 \geq 5$	6) $n_2 \times (1-\bar{p}_2) \geq 5$

- Note: Unlike the one-sample proportion tests when we used the hypothesized value (p_0) to check the conditions, we can only use the sample proportions to check the conditions.
- The conclusion is made exactly the same way as before.
- MLO: Find the p-value and check conditions and make an appropriate conclusion.