# Learning Objectives Exam 01: Two Categorical Variables

**LO 1.** Distinguish between the **explanatory variable(s)** and the **response variable** in an experiment.

- **explanatory variable**: a variable that is either controlled by the experimenter or assumed to be the cause of the outcome of the experiment.
- **response variable**: the primary variable of interest that is believed to be affected by the values of the explanatory variables.

The designation of which variable is response or explanatory is often not fixed, and may be a decision of the investigator.

**LO 2.** Distinguish between **observational** and **experimental** studies and the conclusions that can be derived from them.

- **observational study**: draws inferences from a sample to a population where the explanatory variable is not under the control of the researcher. This may be because of ethical concerns, logistical constraints, or taking a *sample of convenience*.
- **experimental study**: draws inferences from a sample in which the explanatory variable of interest is selected at random.

Casual relationships can only be identified through the use of experimental studies.

**LO 3.** Understand the role of the **null hypothesis** and **alternative hypothesis** in inferential statistics.

- **Null hypothesis**: A statement about an unknown parameter, usually that there is no relationship between two measured phenomena, or no association among groups. Denoted by $H_0$.
- **Alternative hypothesis**: A statement that directly contradicts the null hypothesis. Denoted by $H_A$

**LO 4.** Define a **test statistic** and its **sampling distribution** under the null hypothesis ($H_0$). Identify the value of a test statistic within R's output and as seen in scientific literature.

**LO 5.** Understand how simulation can be used to produce a estimation of a sampling distribution.

**LO 6.** Describe the **p-value** in terms of a inferential statistics

- The **p-value** is the probability of observing a test statistics that is *at least as extreme* as the one observed.
- A small p-value indicates strong support for the alternative hypothesis ($H_A$) in favor of the null hypothesis.
- A large p-value does not indicate that the null hypothesis is likely to be true. Instead, it just indicates that there is no strong evidence against it. Think of a court trial where the defendant is found innocent.

**LO 7.** Apply and interpret the use of the Z-test for differences in proportions within **R**. Memorize code for producing the test output from a dataset. Identify the null and alternative hypotheses, test statistic, p-value, and observed odds ratio.

**LO 8.** Understand the assumptions behind the Z-test for difference in proportions and common alternatives:

- **Z-test for proportions** (`tmod_z_test_prop`): assumes that independent variables are selected beforehand and samples are independent

- **Fisher's Exact Test** (`tmod_chi_squared_test`): assumes that the marginal sums of the contingency table are fixed and known
- **Chi-squared Test** (`tmod_fisher_test`): assumes that both categorical variables are randomly determined by the experiment

You should also understand that science and social science research commonly conflates these tests and you should not put much faith on their 'correct' usage. Also, the results are generally not too far off between each of the tests.

**LO 9.** Install R, RStudio, and required packages on your laptop. (This will not be on the exam)

**LO 10.** Understand how to start a new R notebook, load libraries, and execute code.

**LO 11.** Organize tabular data using the **unit of observation**.

**LO 12.** Produce a comma separated values (CSV) or Excel file with tabular data.

**LO 13.** Memorize code for reading a dataset into R, including the relevant package (either `readxl` or `readr`).

**LO 14.** Memorize code for running an hypothesis test in R with the `tmodels` package.

**LO 15.** Follow general naming conventions when constructing variable names.

- only include lowercase letters, underscores, and numbers
- do not start a variable name with a number of underscore
- do not use spaces; replace with underscores when needed
- keep names short and concise; do not add superfluous information (for example, use `age` instead of `patient_age`, `age_years`, or `recorded_age` unless you need to distinguish two types of age variables)