Unit #7 (b): Hypothesis Testing

9.7.3-9.7.5, 9.9.4

At the end of this unit, students should be able to:

1. State and interpret hypotheses about differences across groups/populations.

2. Write and compute the test statistic for the a test comparing two population means with large samples (when the standard deviation is known and unknown). Use the test statistic to conduct the test and make a decision (using either a rejection region or a p-value).

3. Distinguish between lower-tail, upper-tail, and two-tailed tests. Reason about when each should be used.

4. Write and compute the test statistic for the two sample t-test. Recognize when this test applies. Use the test statistic to conduct the test and make a decision (using either a rejection region or a p-value).

5. Write and compute the test statistic for the test for proportion differences. Recognize when this test applies. Use the test statistic to conduct the test and make a decision (using either a rejection region or a p-value).

6. Describe the F distribution, including what it is parameterized by, how it arises, and what it is used to test.

Photo by Jan Huber on Unsplash

Often, researchers are interested in differences with respect to the mean of a continuous variable across different populations or groups.

Example: A coffee company has two different methods for brewing coffee, and the company is interested in knowing whether there is a difference in caffeine levels across brewing methods.

Photo by Jan Huber on Unsplash

First, some theory.

1 $x_1 = 80\text{mg}$
1 $x_2 = 82\text{mg}$
1 $\vdots$
2
2
2

Let $X_1, \ldots, X_{n_X}$ be a large random sample from a population with mean $\mu_X = E(X_i)$ and population variance $\sigma_X^2 = Var(X_i)$.

$H_0: \mu_X = \mu_Y \iff \mu_X - \mu_Y = 0$

Let $Y_1, \ldots, Y_{n_Y}$ be a large random sample from a different, independent population with mean $\mu_Y = E(Y_i)$ and population variance $\sigma_Y^2 = Var(Y_i)$.

- $E(\underbrace{\bar{X} - \bar{Y}}_{\text{unbiased est. of } \mu_X - \mu_Y}) = E(\bar{X}) - E(\bar{Y}) = \mu_X - \mu_Y$
- $Var(\bar{X} - \bar{Y}) = Var(\bar{X}) + Var(\bar{Y})$ [assuming X_i is ind. of Y_i]

$= \dfrac{\sigma_X^2}{n_X} + \dfrac{\sigma_Y^2}{n_Y}$

Consider testing: $H_0 : \mu_X - \mu_Y = \delta_0$ against one of the three alternatives (e.g., $H_1 : \mu_X - \mu_Y \neq \delta_0$).

(Either $n_x > 30$ and $n_y > 30$)

or

($X_i \sim N$ & $Y_i \sim N$)

The test statistic

$$Z = \frac{\bar{X} - \bar{Y} - \delta_0}{\sqrt{\left(\frac{\sigma_X^2}{n_X} + \frac{\sigma_Y^2}{n_Y}\right)}} \sim N(0,1).$$

If $Z = z$ falls within the rejection region (defined by the alternative hypothesis), then we reject H_0 (or have some evidence against H_0). Otherwise, we fail to reject H_0.

Suppose that we didn't know the population variances. *If they are assumed to be roughly equal*, then the test statistic

$$T = \frac{\bar{X} - \bar{Y} - \delta_0}{\sqrt{S_p^2\left(\frac{1}{n_X} + \frac{1}{n_Y}\right)}} \sim t_{n_X+n_Y-2}, \text{ where}$$

$$S_p^2 = \frac{(n_X - 1)S_X^2 + (n_Y - 1)S_Y^2}{n_x + n_Y - 2}$$ (the "pooled" estimator of $\sigma_X^2 = \sigma_Y^2$).

If $T = t$ falls within the rejection region (defined by the alternative hypothesis), then we reject H_0 (or have some evidence against H_0). Otherwise, we fail to reject H_0.

Photo by Jan Huber on Unsplash

Suppose that we didn't know the population variances. *If they are not assumed to be roughly equal*, then the test statistic

$$T = \frac{\bar{X} - \bar{Y} - \delta_0}{\sqrt{\left(\frac{S_X^2}{n_X} + \frac{S_Y^2}{n_Y}\right)}} \overset{approx}{\sim} t_v$$

Welch's t-test

where $v = \dfrac{\left(\frac{S_X^2}{n_X} + \frac{S_Y^2}{n_Y}\right)^2}{\left(\frac{\left(S_X^2/n_X\right)^2}{n_X - 1} + \frac{\left(S_Y^2/n_Y\right)^2}{n_Y - 1}\right)}$

If $T = t$ falls within the rejection region (defined by the alternative hypothesis), then we reject H_0 (or have some evidence against H_0). Otherwise, we fail to reject H_0.

In R:

```
t.test(x, y, alternative = “two.sided”,
var.equal = TRUE, conf.level = 0.9...)
```