Lecture 16

Part 3 Estimation and Hypothesis Test

Comparing Two Populations

Comparing Two Populations

- Thus far, our statistical inference has focused on a single population parameter from a single population.
- We will now look at comparing population parameters from two populations.
- Specifically, we will make inferences about:
  - The difference between two population means, $\mu_1 - \mu_2$.
  - The difference between two population proportions, $p_1 - p_2$.

Independent Samples

- Suppose we ask the question: “Are men taller than women?”
- We need two benchmarks - population mean height for men and population mean height for women.
- Select a random sample of men and a random sample of women.
- Compare the sample mean height of men to the sample mean height of women.
- These are **independent samples** - the samples are collected independently of each other.

Paired Samples

- Let's ask another question: "Are brothers taller than their sisters?"
- We can no longer select a random sample of men and a random sample of women.
- We need to sample brother and sister *pairs*.
- For each pair, we could calculate the difference in heights between the brother and sister, then analyze these differences.
- These are **paired samples** - each observation in one sample is matched or paired to an observation in the other sample.

Independent or Paired Samples

- Whether you should collect independent samples or paired samples will be determined by your question.
- It is important to establish this at the beginning of your study or experiment, as the way you perform your inference (e.g., confidence intervals, hypothesis tests) will change depending on how the samples are collected.

Flow Chart

We'll go back to this later!

- Do we have *independent* samples or *paired* samples?
  1. Independent samples.
     - Are the population variances, σ_1^2 and σ_2^2, known?
       - Known: Test using Z-statistic.
       - Unknown: Are the population variances equal, i.e., does $\sigma_1^2 = \sigma_2^2$?
       - Equal: Test using T-statistic.
       - Unequal: Can't do test by hand.
  2. Paired samples.
     - Test using T-statistic.

Population Variances are Known

Testing $\mu_1 - \mu_2$

- Suppose we want to test hypotheses regarding the difference between two population means using independent samples drawn from each population.
- Hypotheses:

$$H_0 : \mu_1 - \mu_2 = D_0$$
$$H_1 : \mu_1 - \mu_2 (\neq, <, >) D_0$$

- What is a reasonable estimator of $\mu_1 - \mu_2$?

Population Variances are Known

Testing $\mu_1 - \mu_2$

- We can use $\bar{X}_1 - \bar{X}_2$ as an estimator of $\mu_1 - \mu_2$.
- What is the sampling distribution of $\bar{X}_1 - \bar{X}_2$?
- For independent samples and large enough sample sizes, the sampling distribution of $\bar{X}_1 - \bar{X}_2$ is approximately normal with mean and variance given by:

$$E(\bar{X}_1 - \bar{X}_2) = \mu_1 - \mu_2$$

$$V(\bar{X}_1 - \bar{X}_2) = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}$$

Population Variances are Known

Testing $\mu_1 - \mu_2$

- Test statistic:
  - We can calculate a standardized Z-statistic to use as our test statistic!

$$Z = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

$$= \frac{(\bar{X}_1 - \bar{X}_2) - D_0}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

Population Variances are Known

Testing $\mu_1 - \mu_2$

- Decision rule:
  - We compare the Z-statistic to a standard normal distribution and either determine rejection region(s) or calculate a p-value to decide whether or not to reject the null hypothesis.
  - For example, at significance level α reject H_0 if:
    - $Z > z_{\frac{\alpha}{2}}$ or $Z < -z_{\frac{\alpha}{2}}$ (two-tailed $\neq$).
    - $Z > z_{\alpha}$ (upper-tailed $>$).
    - $Z < -z_{\alpha}$ (lower-tailed $<$).

Population Variances are Known

Confidence Interval for $\mu_1 - \mu_2$

- Recall the equivalence between a two-tailed hypothesis test and a confidence interval.
- A $100(1-\alpha)\%$ confidence interval for $\mu_1 - \mu_2$ when the population variances are known is given by:

$$(\bar{X}_1 - \bar{X}_2) \pm z_{\frac{\alpha}{2}} \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

Population Variances are Unknown

Testing $\mu_1 - \mu_2$

- Generally, the population variances, σ_1^2 and σ_2^2, are rarely known.
- If they are unknown, how we proceed in our test depends on whether we can assume the unknown population variances are equal.
  - If variances are equal, test can be performed by hand.
  - If variances are not equal, test is complicated and cannot be done by hand.

Population Variances are Unknown

Testing Equality of Variances

- But how do we test for equality of variances?
- By testing the following hypotheses:

$$H_0 : \sigma_1^2 = \sigma_2^2 \quad \left(\text{or } \frac{\sigma_1^2}{\sigma_2^2} = 1\right)$$

$$H_1 : \sigma_1^2 \neq \sigma_2^2 \quad \left(\text{or } \frac{\sigma_1^2}{\sigma_2^2} \neq 1\right)$$

Population Variances are Unknown

Testing Equality of Variances

- Test statistic:
  - We use the following test statistic, called the F-statistic, to test the equality of variances:

$$F = \frac{s_1^2}{s_2^2}$$

  - We would then reject H_0 if the F-statistic is too large ($\gg 1$) or too small ($\ll 1$).
  - But what is the sampling distribution of this F-statistic under H_0?

F-distribution

- The null distribution of this F-statistic is an F-distribution with $n_1 - 1$ numerator degrees of freedom and $n_2 - 1$ denominator degrees of freedom.
- The F-distribution is a special continuous distribution:
    - It has two parameters called the numerator degrees of freedom and the denominator degrees of freedom.
    - F-tables give critical values that cut off probability A in the upper tail.
    - There is a different table for each value of A.
    - Rows and columns display the degrees of freedom.