Introduction to Statistical Methods in Political Science

Lesson Week 5: Sampling Distribution for the Difference of Two Means

Ignacio Urbina

Example: Job Satisfaction

One application of comparing two sample means is studying job satisfaction across different industries. For instance:

- Consider two different industries: information technology (IT) and finance.
- We gather a sample of workers from each industry and measure their job satisfaction scores.
- Let $\bar{x}_1$ be the mean job satisfaction score for IT workers and $\bar{x}_2$ for finance workers.

This analysis can help companies within these industries understand their employees' satisfaction and inform their management strategies.

Example: Physical Activity Study

In health science, the comparison of two sample means is key in evaluating the impact of different interventions:

- Imagine a study comparing the effect of two exercise regimens on cholesterol levels.
- One group follows a high-intensity training while the other a moderate intensity training.
- Let $\bar{x}_1$ be the mean decrease in cholesterol level in high-intensity group and $\bar{x}_2$ in moderate intensity group.

This comparison can give insights into which regimen is more effective, guiding health recommendations and future research.

Introduction to Sample Means

Consider two independent samples where:

- Sample 1: n_1 observations with mean $\bar{x}_1$
- Sample 2: n_2 observations with mean $\bar{x}_2$

We are interested in the statistic $\bar{x}_1 - \bar{x}_2$, the difference between two sample means.

Review of Expectations and Variances

Expectations:

- For any random variables X and Y, and constants a, b:

$$E(aX + bY) = aE(X) + bE(Y)$$

Variances:

- For independent random variables X and Y, and constants a, b:

$$\text{Var}(aX + bY) = a^2\text{Var}(X) + b^2\text{Var}(Y)$$

Expectation of the Statistic

Using the rules for expectations of linear combinations of random variables:

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

Where μ_1 and μ_2 are the true population means.

Variance of the Statistic

Using the rules for variances of linear combinations of *independent* random variables:

$$\text{Var}(\bar{x}_1 - \bar{x}_2) = \text{Var}(\bar{x}_1) + \text{Var}(\bar{x}_2)$$

Given $\bar{x}_1$ and $\bar{x}_2$ are independent,

$$\text{Var}(\bar{x}_1) = \frac{\sigma_1^2}{n_1}, \quad \text{Var}(\bar{x}_2) = \frac{\sigma_2^2}{n_2}$$

Thus,

$$\text{Var}(\bar{x}_1 - \bar{x}_2) = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}$$

Standard Error of the Statistic

$$SE(\bar{x}_1 - \bar{x}_2) = \sqrt{\text{Var}(\bar{x}_1 - \bar{x}_2)}$$

Substituting the variances,

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

Normal Approximation of the Sampling Distribution

If the sample sizes are large (Central Limit Theorem), for large sample sizes n_1 and n_2,

$$\bar{x}_1 - \bar{x}_2 \approx N\left(\mu_1 - \mu_2, \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}\right)$$

This normal approximation allows us to perform hypothesis testing and construct confidence intervals for $\mu_1 - \mu_2$.

The Plug-In Principle

In practice, we rarely know the true population variances σ_1^2 and σ_2^2. So, we often estimate them using the sample variances s_1^2 and s_2^2. This is known as the "plug-in principle".
Under this principle, we can substitute the population variances with the sample variances in our standard error formula:

$$SE(\bar{x}_1 - \bar{x}_2) = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

The unbiased estimator for the sample variance s^2 is given by:

$$s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2$$

Z Statistic under CLT and Plug-In Principle

Under the Central Limit Theorem (CLT) and when using the plug-in principle, we can compute the Z statistic for the difference in two sample means as follows:

$$Z = \frac{\bar{x}_1 - \bar{x}_2}{SE(\bar{x}_1 - \bar{x}_2)}$$

Where:
$\bar{x}_1$ and $\bar{x}_2$ = sample means,
$SE(\bar{x}_1 - \bar{x}_2)$ = standard error of the difference in sample means.
Replacing the standard error with sample variances, we get:

$$Z = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

Assumption: $\mu_1 - \mu_2 = 0$

Throughout the analysis, we have often assumed that $\mu_1 - \mu_2 = 0$. This is not an arbitrary assumption, it's a result from our null hypothesis.
In many comparative studies, we start with the null hypothesis (H_0) that there is no difference between the two population means, that is:

$$H_0 : \mu_1 - \mu_2 = 0$$

Because we assume the null hypothesis to be true until proven otherwise, we take $\mu_1 - \mu_2 = 0$ in our computation of the Z statistic:

$$Z = \frac{\bar{x}_1 - \bar{x}_2 - (\mu_1 - \mu_2)}{SE(\bar{x}_1 - \bar{x}_2)} = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

Small Sample Size: t-Distribution

If the sample sizes are small and the population is normally distributed, the distribution of the difference in means follows a t-distribution:

$$\frac{\bar{x}_1 - \bar{x}_2 - (\mu_1 - \mu_2)}{SE(\bar{x}_1 - \bar{x}_2)} \sim t_{df^*}$$

This allows us to still construct confidence intervals and perform hypothesis testing for $\mu_1 - \mu_2$.

Degrees of Freedom

The degrees of freedom (df^*) plays a crucial role in determining the t distribution. It is typically calculated based on the sample size. One straightforward approach is to simply use the smaller of the two degrees of freedom from each sample:

$$df^* = \min(n_1 - 1, n_2 - 1)$$

However, this method can be overly conservative and possibly lead to wider confidence intervals.

Degrees of Freedom

A more accurate and widely used approach is the Satterthwaite approximation:

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

The Satterthwaite approximation takes into consideration the variances of the two samples as well as their sizes, providing a more precise estimate of the degrees of freedom and, consequently, a more accurate t-distribution.

Pooled Variance

When we can assume that the variances of the two samples are equal ($\sigma_1^2 = \sigma_2^2$), we can use the pooled variance formula and the pooled degrees of freedom. This assumption simplifies the analysis and usually leads to narrower confidence intervals.
The pooled variance is a weighted average of the individual variances, given by the formula:

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

The degrees of freedom in this case is the sum of the degrees of freedom of each sample: $df = n_1 + n_2 - 2$.
Keep in mind, however, that the assumption of equal variances should be verified (via statistical tests or graphical methods) before proceeding with the pooled variance approach.

Important Notes for Quizzes and Testing

For the purposes of quizzes and testing regarding the t-test with two means, please note the following:

1. I will always tell you how to calculate the degrees of freedom.
2. I will always provide the formula for degrees of freedom.
3. I will be clear on whether you have to use separate variances or pooled variance. In this case, I will include in the quiz/exam document the formula for the pooled variance.