10.1: Z Tests and Confidence Intervals for a Difference Between Two Population Means

Taylor

University of Virginia

Introduction

Now we'll have two datasets $X_1, \ldots, X_m$ and $Y_1, \ldots, Y_n$. We'll do confidence intervals and hypothesis testing.

Recall we had three cases when we were dealing with hypothesis testing/confidence intervals for single samples. Here we will only have two. We don't get that same third case with two samples. This is because there are some complications when dealing with t rvs. That's why next chapter is devoted to that. So, just two cases here.

Introduction

Most often we want to test if $\mu_1 = \mu_2$. Or equivalently $\mu_1 - \mu_2 = 0$. We're going to use the second one more often because we can think about this difference in parameters as a single thing we can make hypotheses about. We can also test if the difference is some value other than zero.

Basic assumptions:

1. $X_1, \ldots, X_m$ is a random sample from a population with mean μ_1 and variance σ_1^2.
2. $Y_1, \ldots, Y_n$ is a random sample from a population with mean μ_2 and variance σ_2^2.
3. The X and Y samples are independent from each other.

Introduction

$\bar{X} - \bar{Y}$ is a natural estimator for $\mu_1 - \mu_2$.

1. $\bar{X} - \bar{Y}$ is unbiased
2. $SE_{actual}(\bar{X} - \bar{Y}) = \sqrt{\frac{\sigma_1^2}{m} + \frac{\sigma_2^2}{n}}$
3. $SE_{estimated}(\bar{X} - \bar{Y}) = \sqrt{\frac{s_1^2}{m} + \frac{s_2^2}{n}}$

Introduction

So

$$Z = \frac{\bar{X} - \bar{Y} - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{m} + \frac{\sigma_2^2}{n}}} \sim \mathcal{N}(0, 1).$$

We use this as our test statistic motivation when we know the actual standard error. It follows a normal distribution if our data are normally distributed. Or if that isn't true, it still works because we can use the CLT.

Synopsis

1. $H_0 : \mu_1 - \mu_2 = \Delta_0$
2. $z = \frac{\bar{x} - \bar{y} - \Delta_0}{\sqrt{\frac{\sigma_1^2}{m} + \frac{\sigma_2^2}{n}}}$
3. If $H_a : \mu_1 - \mu_2 > \Delta_0$ reject when $z \geq z_\alpha$
4. If $H_a : \mu_1 - \mu_2 < \Delta_0$ reject when $z \leq -z_\alpha$
5. If $H_a : \mu_1 - \mu_2 \neq \Delta_0$ reject when $z \geq z_{\alpha/2}$ OR when $z \leq -z_{\alpha/2}$

Correlation is not Causation

Side Note: has a few pages on how rejection of $H_0 : \mu_1 - \mu_2 = 0$ doesn't allow us to say anything about causality. It defines **observational** studies, **retrospective** studies, and **randomized controlled experiments**. We won't talk about this, but it might be worth looking at if you're interested in domains where this is a big issue.

Type 2 Error and Sample Size Determination

Let's say $H_0 : \mu_1 - \mu_2 = \Delta_0$ versus $H_a : \mu_1 - \mu_2 > \Delta_0$. Let Δ' denote the true difference in means. Also we write $SE(\cdot)$ instead of $SE_{actual}(\cdot)$. From now on we'll assume the context is clear.

$$\begin{aligned}
\beta(\Delta') &= P\left(\frac{\bar{X} - \bar{Y} - \Delta_0}{\text{SE}(\bar{X} - \bar{Y})} < z_\alpha | \Delta = \Delta'\right) \\
&= P\left(\bar{X} - \bar{Y} < z_\alpha \text{SE}(\bar{X} - \bar{Y}) + \Delta_0 | \Delta = \Delta'\right) \\
&= P\left(\frac{\bar{X} - \bar{Y} - \Delta'}{\text{SE}(\bar{X} - \bar{Y})} < z_\alpha + \frac{\Delta_0 - \Delta'}{\text{SE}(\bar{X} - \bar{Y})}\right) \\
&= \Phi\left(z_\alpha + \frac{\Delta_0 - \Delta'}{\text{SE}(\bar{X} - \bar{Y})}\right)
\end{aligned}$$

power calculations for other types of hypotheses are listed on page 489.

Large-Sample Tests

This is where we don't know $SE_{actual}(\bar{X} - \bar{Y})$ but we do know $SE_{estimated}(\bar{X} - \bar{Y})$.

Under H_0

$$Z = \frac{\bar{X} - \bar{Y} - \Delta_0}{\text{SE}_{estimated}(\bar{X} - \bar{Y})} \sim \mathcal{N}(0, 1)$$

The decision rules are the same as the ones on slide 6. Also, the power calculations are difficult for the same reason we discussed in the single sample case.

Confidence Intervals

Our confidence intervals are based on either

$$Z = \frac{\bar{X} - \bar{Y} - \Delta}{\text{SE}_{actual}(\bar{X} - \bar{Y})} \sim \mathcal{N}(0, 1)$$

or

$$Z = \frac{\bar{X} - \bar{Y} - \Delta}{\text{SE}_{estimated}(\bar{X} - \bar{Y})} \sim \mathcal{N}(0, 1)$$

Motivation

so we can get our confidence intervals based on this equation

$$P\left(-z_{\alpha/2} < \frac{\bar{X} - \bar{Y} - \Delta}{\mathrm{SE}_{actual}(\bar{X} - \bar{Y})} < z_{\alpha/2}\right) = 1 - \alpha$$

or this equation:

$$P\left(-z_{\alpha/2} < \frac{\bar{X} - \bar{Y} - \Delta}{\mathrm{SE}_{estimated}(\bar{X} - \bar{Y})} < z_{\alpha/2}\right) = 1 - \alpha.$$

I'll leave it to you to show that we can have confidence intervals

$$\bar{X} - \bar{Y} \pm z_{\alpha/2}\mathrm{SE}_{actual}(\bar{X} - \bar{Y})$$

or

$$\bar{X} - \bar{Y} \pm z_{\alpha/2}\mathrm{SE}_{estimated}(\bar{X} - \bar{Y})$$