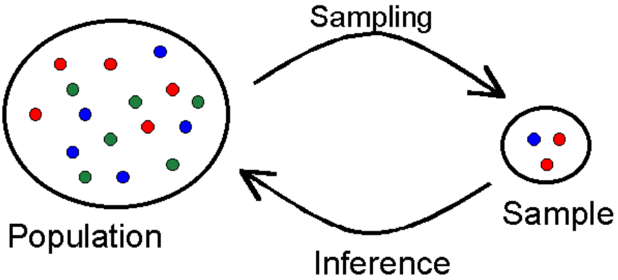
Two-Sample Hypothesis Tests

Ryan Miller

Grinnell College
Statistics

Introduction

Recall that statisticians use hypothesis testing to make inferences about a *population*:



In our toy choice example, we saw that a majority of the sample favored the “helper”, but we really wanted to know if this finding could be generalized to a broader population

One-Sample vs. Two-Sample Testing

One-sample tests hypothesize something about the entire population:

$$H_0 : p = 0.5$$

$$\text{or } H_0 : \mu = 120$$

The entire sample is then used to as evidence against the null hypothesis via the *p*-value:

$$Pr(\hat{p} \geq 14/16 | p = 0.5)$$

$$\text{or } Pr(\overline{x} \geq 132.7 | \mu = 120)$$

Note: These numbers come from previous examples (infant toy choice, and ICU patient blood pressures)

One-Sample vs. Two-Sample Testing

Two-sample tests hypothesize something about groups within the population:

$$H_0 : p_1 = p_2 \iff p_1 - p_2 = 0$$

$$\text{or } H_0 : \mu_1 = \mu_2 \iff \mu_1 - \mu_2 = 0$$

- We do not hypothesize specific values for the population parameters (p_1, p_2 or μ_1, μ_2)
  - We view our available data as two-samples, as cases in group 1 only provide information about p_1 (or μ_1) and cases in group 2 only provide information about p_2 (or μ_2)
- The *p*-value is now based how each sample group differs
  - For example: $Pr(\overline{x}_1 - \overline{x}_2 \geq 10 | \mu_1 = \mu_2)$ if we observed a difference in means of 10-units

Grinnell College
Statistics

Two-Sample Z-test

We will use the **two-sample Z-test** for *two-sample categorical data*, or scenarios where we want to compare proportions observed in two different groups:

- Typically, we use $H_0 : p_1 = p_2$ and $H_a : p_1 \neq p_2$
- We won't cover the details, but CLT gives us:
  $SE = \sqrt{\frac{p_0(1-p_0)}{n_1} + \frac{p_0(1-p_0)}{n_2}}$
- Because there are many different ways to satisfy H_0, we will use a *pooled proportion*, p_0, found by treating all of the data as a single sample (ie: ignoring the observed groups)
- We then use $Z = \frac{\hat{p}_1 - \hat{p}_2}{SE}$ and compare to a $N(0,1)$ distribution to get the p-value (just like we did for the one-sample Z-test)

Grinnell College
Statistics

Two-Sample Z-test Example

Until 2002, hormone replacement therapy (HRT) was commonly prescribed to postmenopausal women. This changed in 2002, when a large clinical trial was stopped early for safety concerns.

In the trial, 8506 women were randomized to take HRT and 8102 were randomized to take a placebo. Researchers observed 164 cases of cardiovascular disease (CVD) in the HRT group, but only 122 CVD cases in the placebo group.

1) State the null and alternative hypotheses used to test whether the risk of CVD is higher in women taking HRT
2) Find the *pooled proportion*, and the SE for this application
3) Apply the Z-score transform to find the Z-value, then find the p-value and make a conclusion

Two-Sample Z-test Example (solution)

1) $H_0 : p_1 - p_2 = 0$, where p_1 is the proportion of cases of cardiovascular disease in the HRT group, and p_2 is the equivalent proportion for the placebo group.
2) $\hat{p}_0 = \frac{164+122}{8506+8102} = 0.017$, so $SE = \sqrt{\frac{0.017(1-0.017)}{8506} + \frac{0.017(1-0.017)}{8102}}$ $= 0.002$
3) $Z = \frac{(164/8506-122/8102)-0}{0.002} = 2.11$, the corresponding p-value (two-sided) is 0.034, which is strong evidence of a higher rate of cardiovascular disease in the HRT group

Two-Sample T-test

- The **two-sample T-test** is used for *two-sample quantitative data*
- Typically, we use $H_0 : \mu_1 = \mu_2$ and $H_a : \mu_1 \neq \mu_2$
- CLT gives us $SE = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$
- From here we apply the Z-score transformation to calculate a T-value, which is used to find the p-value
  - Degrees of freedom are complicated because n_1 and n_2 typically aren't equal, we'll rely upon R to find them

Two-Sample T-test Example

In the 2008 Olympics an unprecedented number of swimming world records were set by athletes using Speedo's LZR Racer, a uniquely engineered full-body swimsuit. But does the suit really impact a swimmer's speed?

- Without the suit, 12 swimmers had an average velocity of $\overline{x}_1 = 1.507$ m/s, with a standard deviation of $s = 0.136$ m/s
- With the suit, 12 swimmers had an average velocity of $\overline{x}_2 = 1.429$ m/s, with a standard deviation of $s = 0.141$ m/s

Calculate the SE and T-value, then compare to a t-distribution with $df = 11$ to find the two-sided p-value

Paired Samples

- In the wetsuit example, it was actually the *same 12 swimmers* that swam with and without the suit
  - Thus, we didn't actually have two independent samples, but rather one sample that we measured in two different ways
- This is known as a **paired design**, and it comes with the advantage of controlling for the variability *between swimmers*
  - The *paired T-test* uses the average difference observed within swimmers and $H_0: \mu_{diff} = 0$ in a one-sample T-test

Paired T-test Example

```r
## Load Data
swim_data = read.csv("https://remiller1450.github.io/data/Wetsuits.csv")

## Find the paired differences and give them to t.test()
paired_difference = swim_data$Wetsuit - swim_data$NoWetsuit
t.test(x = paired_difference, mu = 0)
```

```
##
##  One Sample t-test
##
## data:  paired_difference
## t = 12.318, df = 11, p-value = 8.885e-08
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  0.06365244 0.09134756
## sample estimates:
## mean of x
##    0.0775
```

Sample Size Considerations

Just like the one-sample Z and T tests, the tests we saw today are based upon probability models that will only accurately approximate the null distribution under certain conditions:

- The two-sample Z-test is appropriate when at least 10 of each outcome are expected in both groups
  - $n_1 p_0 \geq 10$, $n_1(1-p_0) \geq 10$, $n_2 p_0 \geq 10$, and $n_2(1-p_0) \geq 10$
- The two-sample T-test is appropriate in either of the following situations:
  - Both groups came from Normally distributed populations
  - $n_1 \geq 30$ and $n_2 \geq 30$, regardless of how the data are distributed

Note that these are common rules of thumb, there aren't any definitive cutoffs for when a procedure does/doesn't work

Conclusion

We've now covered Z and T tests for both one-sample and two-sample data. You should know how the following:

- Categorical data: Z-test
  - One-sample data: $H_0 : p = \_\_\_$ and $SE = \sqrt{\frac{p(1-p)}{n}}$
  - Two-sample data: $H_0 : p_1 = p_2$ and $SE = \sqrt{\frac{p_0(1-p_0)}{n_1} + \frac{p_0(1-p_0)}{n_2}}$ with p_0 being the *pooled proportion*
- Quantitative data: T-test
  - One-sample data: $H_0 : \mu = \_\_\_$ and $SE = \frac{\sigma}{\sqrt{n}}$ and $df = n - 1$
  - Two-sample data: $H_0 : \mu_1 = \mu_2$ and $SE = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$ with df found using R
  - Paired data: just a one-sample test on the paired differences

Grinnell College
Statistics