Hypothesis Testing

Part 4 - two-sample tests

Ryan Miller

Grinnell College
Statistics

Introduction

So far, we've used hypothesis tests to answer research questions like:

1. Do infants pick randomly between friendly and antagonistic characters?
2. Is the mean blood pressure of ICU patients different from 120 mg Hg?

These questions treated all of the data as a single sample, which was used as evidence in the test. Today we'll look at questions where the sample is divided into two groups.

One-Sample vs. Two-Sample Tests

One-Sample Tests	Two-Sample Tests
Does a population differ from a known benchmark value?	Do two populations (groups) differ from each other?
One sample of observations	Two groups of observations
$H_0 : p = 0.5$	$H_0 : p_1 = p_2$
$H_0 : \mu = 120$	$H_0 : \mu_1 = \mu_2$
Example: Is the mean SBP 120 mm Hg?	Example: Do users click more often with version A or B?

Importantly, two-sample tests do not hypothesize any specific parameter values, instead they hypothesize about how two groups compare.

The Two-Sample Z-test

A/B testing is an example of a two-sample hypothesis test, with the following scenario being a classic example:

- An online retailer wants to see if changing the color of the "buy now" button on their website will increase sales.
- They randomly assign half of their next 2000 visitors to see either the webpage with the current button (control), or an alternate version with the new color (treatment).
- They record how many visitors made purchases in each group as the outcome of interest.

The Two-Sample Z-test - Example (part 1)

Suppose 71 of 1000 visitors who were routed to the page with the new button color made a purchase, compared to only 54 of 1000 visitors who saw the old webpage.

1. Why shouldn't the developers simply conclude the new button color is more effective because 7.1% is greater than 5.4%?
2. State the null hypothesis and provide the corresponding sample statistics using proper statistical notation.

The Two-Sample Z-test

Similar to the one-sample tests we've seen, the two-sample Z-test will use a test statistic that looks like a Z-score:

$$Z = \frac{\text{observed} - \text{expected}}{SE} = \frac{(\hat{p}_1 - \hat{p}_2) - 0}{SE}$$

- For a single proportion, $SE(\hat{p}) = \sqrt{\frac{p(1-p)}{n}}$
- Now two proportions could vary from sample to sample, which the SE must account for: $SE(\hat{p}_1 - \hat{p}_2) = \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}$

Two-Sample Z-test

- For a one-sample test, we'd plug the hypothesized proportion into the SE formula, but for a two-sample test we aren't hypothesizing any specific values.
- Consequently, we'll use a *pooled proportion*, $\hat{p}_{\text{pool}}$, found by treating all of the data as a single sample (ie: ignoring the observed groups) in place of *both* p_1 and p_2.

$$\hat{p}_{\text{pool}} = \frac{\text{count}_1 + \text{count}_2}{n_1 + n_2}$$

- After using the observed sample proportions, $\hat{p}_1$ and $\hat{p}_2$, in the numerator, and the pooled proportion, $\hat{p}_{\text{pool}}$, in denominator, we compare our Z value with the $N(0,1)$ distribution to calculate the p-value and make a conclusion.

Two-Sample Z-test - Example (part 2)

Recall in our A/B test example 71 of 1000 visitors made a purchase in the treatment group, compared to 54 of 1000 visitors in the control group.

1. Calculate the *pooled proportion* of all visitors who made a purchase, regardless of group.
2. Use the pooled proportion and observed sample proportions to calculate a Z-value for the two-sample Z-test.
3. Use the "Normal distribution" menu of StatKey to find the p-value corresponding to your Z-value
4. Write a proper one-sentence conclusion that includes: context, evidence, and direction of the effect (if one was found)

Two-sample Z-test - Example (solution)

1. $\hat{p}_{\text{pool}} = 125/2000 = 0.0625$
2. $Z = \frac{0.071 - 0.054}{\sqrt{0.0625(1-0.0625)/1000 + 0.0625(1-0.0625)/1000}} = 1.57$
3. The two-sided p-value is 0.116
4. We find insufficient statistical evidence (p=0.116) that the purchase rate differs between visitors shown the two button colors.

The 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 usually rely upon R to find them, but a conservative approach uses $\min(n_1 - 1, n_2 - 1)$

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?

- With the suit, 12 swimmers had an average velocity of $\overline{x}_1 = 1.507$ m/s, with a standard deviation of $s_1 = 0.136$ m/s
- Without the suit, 12 swimmers had an average velocity of $\overline{x}_2 = 1.429$ m/s, with a standard deviation of $s_2 = 0.141$ 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 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*
  - $n_1\hat{p}_{\text{pool}} \geq 10$, $n_1(1-\hat{p}_{\text{pool}}) \geq 10$, $n_2\hat{p}_{\text{pool}} \geq 10$, and $n_2(1-\hat{p}_{\text{pool}}) \geq 10$
- 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 $SE = \sqrt{\frac{\hat{p}_{\text{pool}}(1-\hat{p}_{\text{pool}})}{n_1} + \frac{\hat{p}_{\text{pool}}(1-\hat{p}_{\text{pool}})}{n_2}}$ with $\hat{p}_{\text{pool}}$ being the *pooled proportion*
- Quantitative data: T-test
  - One-sample data: $H_0 : \mu = \_\_\_$ and $SE = \frac{s}{\sqrt{n}}$ and $df = n - 1$
  - Two-sample data: $H_0 : \mu_1 = \mu_2$ and $SE = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$ with df found using R
  - Paired data: just a one-sample test on the paired differences