Hypothesis Testing

Part 2 - mathematical models and test statistics

Ryan Miller

Grinnell College
Statistics

Introduction

So far we've used the following procedure to perform hypothesis testing:

1) Propose the null and alternative hypotheses
2) Use StatKey to simulate outcomes under the null hypothesis (ie: the null distribution)
3) Compare the outcome observed in the real data against the null distribution to find the p-value

This procedure can be made more general using *Normal models* and *test statistics*, or Z-scores, that describe how many standard errors (SE) the observed outcome is above or below what we'd expect under H_0

Proceduralized hypothesis testing

Randomization test:

- Propose H_0

The Z-test:

- Propose H_0

Proceduralized hypothesis testing

Randomization test:

- Propose H_0
- Use StatKey to simulate outcomes under the null hypothesis

The Z-test:

- Propose H_0
- Use a CLT formula to calculate SE, then find $Z = \frac{\text{Observed} - \text{Null}}{SE}$

Proceduralized hypothesis testing

Randomization test:

- Propose H_0
- Use StatKey to simulate outcomes under the null hypothesis
- Locate the observed outcome in the null distribution and count the simulated outcomes that are at least as extreme

The Z-test:

- Propose H_0
- Use a CLT formula to calculate SE, then find $Z = \frac{\text{Observed} - \text{Null}}{SE}$
- Locate the test statistic, Z, in the Standard Normal curve and find the area outside it

Example (single proportion)

For a single proportion, Central Limit theorem suggests:

$$SE = \sqrt{\frac{p(1-p)}{n}}$$

- In the infant toy choice study, $H_0 : p = 0.5$, which reflected infants choosing randomly between the two toys
  - The observed outcome was $\hat{p} = 14/16 = 0.875$, or 14 of 16 infants chose the "helper"

Example (single proportion)

For a single proportion, Central Limit theorem suggests:

$$SE = \sqrt{\frac{p(1-p)}{n}}$$

- In the infant toy choice study, $H_0 : p = 0.5$, which reflected infants choosing randomly between the two toys
  - The observed outcome was $\hat{p} = 14/16 = 0.875$, or 14 of 16 infants chose the "helper"

1) Under H_0, $SE = \sqrt{\frac{p_0(1-p_0)}{n}} = \sqrt{\frac{0.5(1-0.5)}{16}} = 0.125$

Example (single proportion)

For a single proportion, Central Limit theorem suggests:

$$SE = \sqrt{\frac{p(1-p)}{n}}$$

- In the infant toy choice study, $H_0 : p = 0.5$, which reflected infants choosing randomly between the two toys
  - The observed outcome was $\hat{p} = 14/16 = 0.875$, or 14 of 16 infants chose the "helper"

1) Under H_0, $SE = \sqrt{\frac{p_0(1-p_0)}{n}} = \sqrt{\frac{0.5(1-0.5)}{16}} = 0.125$
2) Then, $Z = \frac{\hat{p}-p_0}{SE} = \frac{0.875-0.5}{0.125} = 3$

Example (single proportion)

For a single proportion, Central Limit theorem suggests:

$$SE = \sqrt{\frac{p(1-p)}{n}}$$

- In the infant toy choice study, $H_0 : p = 0.5$, which reflected infants choosing randomly between the two toys
  - The observed outcome was $\hat{p} = 14/16 = 0.875$, or 14 of 16 infants chose the "helper"

1) Under H_0, $SE = \sqrt{\frac{p_0(1-p_0)}{n}} = \sqrt{\frac{0.5(1-0.5)}{16}} = 0.125$
2) Then, $Z = \frac{\hat{p}-p_0}{SE} = \frac{0.875-0.5}{0.125} = 3$
3) Comparing $Z = 3$ against a Standard Normal curve, the one-sided p-value is 0.0013 (two-sided is 0.0026)

Practice #1

We've previously discussed a study conducted by Johns Hopkins University that found 31 of 39 babies born 15 weeks early went on to survive. According to Wikipedia, the survival rate for babies born this early is 70%. Does the Johns Hopkins University study provide compelling evidence to refute Wikipedia's claim?

1) Propose a null hypothesis and an alternative hypothesis
2) Use CLT to find a standard error, then calculate a Z-value measuring how standard errors the observed outcome is from the value specified in H_0
3) Compare Z against the Standard Normal distribution to find the p-value

Practice #1 (solution)

1) $H_0 : p = 0.7$ vs. $H_a : p \neq 0.7$
2) $SE = \sqrt{\frac{0.7(1-0.7)}{39}} = 0.073$, then notice we observed $\hat{p} = 31/39 = 0.795$, so $Z = \frac{0.795-0.7}{0.073} = 1.29$
3) The two-sided p-value corresponding to $Z = 1.29$ is 0.198, indicating that these data do not provide sufficient evidence to refute Wikipedia's claim.

Modifications for testing a single mean

For a single mean, recall that CLT suggests:

$$\overline{x} \sim N(\mu, \frac{\sigma}{\sqrt{n}})$$

However, estimating σ (an unknown population parameter) via s (the sample standard deviation) introduces additional uncertainty that necessitates the t-distribution

Modifications for testing a single mean

Using $SE = \frac{s}{\sqrt{n}}$, the one-sample T-test calculates a T-value:

$$T = \frac{\bar{x} - \mu}{SE}$$

Then compares this T-value against a t-distribution with $df = n - 1$ to find the p-value

Practice #2

In 2010, new international rules were created to regulate swimsuit coverage and material after an inordinate amount of records were set at the 2008 Olympics by swimmers wearing a suit known as the LZR Racer. The data below are the 1500m swim velocities of 12 competitive swimmers with and without a scientifically designed wet suit.

https://remiller1450.github.io/data/Wetsuits2.csv

1) Find the sample mean and sample standard deviation of the variable "difference"
2) Propose null and alternative hypotheses involving the variable "difference"
3) Find the T pertaining to these data, then compare it against the proper t-distribution to find the p-value.

Practice #2 (solution)

1) Using the Descriptive Statistics and Graphs section of StatKey, $\overline{x} = 0.077$ and $s = 0.022$
2) $H_0 : \mu = 0$, or the average improvement in velocity when wearing the wetsuit is zero, vs. $H_a : \mu \neq 0$
3) $T = \frac{0.077 - 0}{0.022/\sqrt{12}} = 12.12$, the p-value is nearly zero, indicating overwhelming evidence that the wetsuit improves swim velocity

Comments (paired designs)

- The wetsuits study is an example of a **paired design**, a type of design where each subject serves as their own control
- Paired designs have a number of statistical advantages over other designs, with an important one being the elimination of confounding variables

Two-sample data

- So far, we've used the Z-test to evaluate hypotheses involving a *single proportion*, and the T-test to evaluate hypotheses involving a *single mean*
  - These are *one-sample tests*, as they treat all of the data as a single sample (group)

Two-sample data

- So far, we've used the Z-test to evaluate hypotheses involving a *single proportion*, and the T-test to evaluate hypotheses involving a *single mean*
  - These are *one-sample tests*, as they treat all of the data as a single sample (group)
- The Z-test can also test hypotheses involving a *difference in proportions* (ie: $H_0 : p_1 - p_2 = 0$)
  - Similarly, the T-test can also test hypotheses involving a *difference in means* (ie: $H_0 : \mu_1 - \mu_2 = 0$)
- These applications are called *two-sample tests*, as they involve splitting the data into two groups

Differences in proportions

Recall that the CLT suggests the following normal approximation for a difference in proportions:

$$\hat{p}_1 - \hat{p}_2 \sim N\left(p_1 - p_2, \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}\right)$$

- How might we use this approximation to determine the null distribution for $H_0 : p_1 - p_2 = 0$?
- Are there any difficulties in using the same approach we did for a single proportion to derive a *two-proportion z-test*?

Null Distribution for a Difference in Proportions

- For a difference in proportions, the null hypothesis doesn't explicitly specify the values of p_1 and p_2
  - This means we can't simply plug-in the "null value" in the CLT result and get the null distribution

Null Distribution for a Difference in Proportions

- For a difference in proportions, the null hypothesis doesn't explicitly specify the values of p_1 and p_2
  - This means we can't simply plug-in the "null value" in the CLT result and get the null distribution
- In fact, there are infinitely many values of p_1 and p_2 that will satisfy $H_0 : p_1 - p_2 = 0$
  - The proper choice will both satisfy the null hypothesis, and be consistent with our data

Null Distribution for a Difference in Proportions

- For a difference in proportions, the null hypothesis doesn't explicitly specify the values of p_1 and p_2
  - This means we can't simply plug-in the "null value" in the CLT result and get the null distribution
- In fact, there are infinitely many values of p_1 and p_2 that will satisfy $H_0 : p_1 - p_2 = 0$
  - The proper choice will both satisfy the null hypothesis, and be consistent with our data
- Using a **pooled proportion**, $\hat{p}_{1+2}$, in place of *both* p_1 and p_2 accomplishes this
  - $\hat{p}_{1+2}$ is calculated by *ignoring the grouping variable* that defines the two proportions
  - For example, if $\hat{p}_1 = 12/20$ and $\hat{p}_2 = 7/15$, then $\hat{p}_{1+2} = \frac{7+12}{20+15}$

Practice #3

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 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) Perform a two-sample Z-test

Practice #3 (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

The two-sample T-test

When testing a difference in means (rather than proportions), we must make two major changes:

1) $SE = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$, a result derived from the Central Limit theorem
2) Because the SE relies upon s_1 and s_2 as estimates of σ_1 and σ_2 (population parameters), we now need to calculate a T-value and compare it to a t-distribution.

Since we're analyzing two groups (ie: two samples of data), the degrees of freedom are complicated. For "by hand" calculations we'll use the smaller of $n_1 - 1$ and $n_2 - 1$ (a conservative approach), but we'll prefer to use R to find the exact degrees of freedom whenever possible.

Practice #1

We've previously analyzed data from an experiment where 12 swimmers participated in a 1500m time trial with an without a scientifically designed wetsuit. In this example, we'll see what happens when we *ignore the paired study design*.

- When swimming with the wetsuit, the average velocity was $\overline{x}_1 = 1.507$ m/s, with a standard deviation of $s = 0.136$ m/s
- When swimming without the wetsuit, the average velocity was $\overline{x}_2 = 1.429$ m/s, with a standard deviation of $s = 0.141$ m/s

1) For $H_0 : \mu_1 - \mu_2 = 0$ (wetsuit - no wetsuit), report the observed sample difference in means and its standard error.
2) Perform a two-sample T-test "by hand".

Practice #1 (solution)

1) The observed difference in means is $\overline{x}_1 - \overline{x}_2 = 1.507 - 1.429 = 0.078$, the standard error is
$$SE = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} = \sqrt{\frac{0.136^2}{12} + \frac{0.141^2}{12}} = 0.057$$
2) The T-value is $T = \frac{0.078-0}{0.057} = 1.37$, we need to use $df = 12 - 1 = 11$, so the two-sided p-value is 0.198. This seems to suggest insufficient evidence of a difference in velocity, but we need to remember that it's ignoring the paired design of the study!

Loose ends - sample size conditions

Both of these two-sample hypothesis testing approaches are built upon Central Limit theorem results:

1) For proportions, the Z-test requires 10 "successes" and 10 "failures" in each sample/group (ie: $n_1 p_1 \geq 10 \ldots$)
2) For means, the T-test requires either Normally distributed data (if the sample/group sizes are small), or sufficiently large samples of $n_1 \geq 30$ and $n_2 \geq 30$ (regardless of how the data are distributed)

If these conditions are not met, simulation-based tests (StatKey) are a reasonable alternative.

Loose ends - confidence intervals and hypothesis testing

Recall that we've found P confidence interval estimates using the formula Point Estimate $\pm c * SE$, which is based upon a Central Limit theorem result suggesting:

$$\overline{x} \sim N(\mu.\sigma/n)$$

- This gave rise to the formula $SE = s/n$ when estimating a population's mean

Hypothesis testing uses the exact same result, but it acknowledges that we do not know μ and would like to hypothesize it's value:

$$\overline{x} \stackrel{?}{\sim} N(\mu_0.\sigma/n)$$

Loose ends - confidence intervals and hypothesis testing

- Because both methods of statistical inference estimate variability in the same way, they will yield *compatible results*
  - That is, if a hypothesis test produces a two-sided p-value less than α, then the corresponding $(1-\alpha)\%$ confidence interval will *not contain* the hypothesized value (under H_0) μ_0
  - Similarly, if a p-value is larger than α, the hypothesized value μ_0 will be contained in the $(1-\alpha)\%$ confidence interval
- As a concrete example, suppose we have the null hypothesis $H_0 : \mu = 0$ and we observe $\overline{x} = 3.5$ yielding a two-sided p-value of 0.06
  - The 95% CI estimate of μ would contain 0, but the 90% CI would not