Hypothesis Testing Procedures for Two-sample Data

Ryan Miller

Outline

1. One-sample vs. two-sample testing
2. Testing a difference in proportions
3. Testing a difference in means
4. Paired study designs

One-sample vs. two-sample testing

- **One-sample testing** uses a single sample in an attempt to falsify a hypothesis about a population (ie: $H_0 : p = p_0$ or $H_0 : \mu = \mu_0$)
  - Notice, the null hypothesis must specify a specific numeric value (ie: $p = 0.5$ or $\mu = 0$)

One-sample vs. two-sample testing

- **One-sample testing** uses a single sample in an attempt to falsify a hypothesis about a population (ie: $H_0 : p = p_0$ or $H_0 : \mu = \mu_0$)
  - Notice, the null hypothesis must specify a specific numeric value (ie: $p = 0.5$ or $\mu = 0$)
- **Two-sample testing** looks to compare two subgroups within a population (ie: $H_0 : p_1 - p_2 = 0$ or $H_0 : \mu_1 - \mu_2 = 0$)
  - Here, the null hypothesis is relational and be satisfied in many different ways (ie: p_1 and p_1 could both be 0.1, or could both be 0.6)
  - This will require us to make some minor adjustments in order to utilize Z and T tests

Motivating example - surgical site infections

- In the 1860's, surgeries often led to infections that resulted in death
- At the time, many experts believed these infections were due to "bad air"
  - Hospitals had policies that required their wards open their windows at midday to air out

Motivating example - surgical site infections

- In the 1860's, surgeries often led to infections that resulted in death
- At the time, many experts believed these infections were due to "bad air"
    - Hospitals had policies that required their wards open their windows at midday to air out
- It was customary for surgeons to move quickly from patient to patient with out any sort of special precautions
    - In fact, many took pride the accumulated stains on their surgical gowns as a measure of experience

Motivating example - surgical site infections

- In 1862, Louis Pasteur discovered food spoilage was caused by the proliferation of harmful micro-organisms
- Pasteur identified three methods for eliminating these micro-organisms: heat, filtration, and chemical disinfectants
  - His heating method became known as *pasteurization* and is widely applied to milk, beer, and many other food products

Motivating example - surgical site infections

- In 1862, Louis Pasteur discovered food spoilage was caused by the proliferation of harmful micro-organisms
- Pasteur identified three methods for eliminating these micro-organisms: heat, filtration, and chemical disinfectants
  - His heating method became known as *pasteurization* and is widely applied to milk, beer, and many other food products
- Joseph Lister, a Professor of Surgery at the Glasgow Royal Infirmary, became aware of Pasteur's work and hypothesized that it might explain the infections that frequently occurred following surgery
  - How would you recommend Lister evaluate his hypothesis?

Lister’s experiment

- Lister proposed a “sterile” protocol that required surgeons to wash their hands, wear clean gloves, and disinfect their instruments with a carbolic acid solution
  - He randomly assigned 75 patients to the “sterile” procedure or a control group
  - He then tracked how many patients survived until their discharage from the hospital

	Died	Survived
Control	16	19
Sterile	6	34

Analyzing Lister's experiment

The big picture goal of an experiment like Joseph Lister's is to *systematically rule out* possible explanations for an improvement in survival, thereby establishing the "sterile" protocol as the cause of the improvement. Explanations that need to be ruled out include:

1) Bias?

Analyzing Lister's experiment

The big picture goal of an experiment like Joseph Lister's is to *systematically rule out* possible explanations for an improvement in survival, thereby establishing the "sterile" protocol as the cause of the improvement. Explanations that need to be ruled out include:

1) Bias? Unlikely, even though double-blinding wasn't possible, it's unlikely the measurement of the outcome (survival) was biased. It's also unlikely that this is a non-representative group of patients (sampling bias)
2) Confounding variables?

Analyzing Lister's experiment

The big picture goal of an experiment like Joseph Lister's is to *systematically rule out* possible explanations for an improvement in survival, thereby establishing the "sterile" protocol as the cause of the improvement. Explanations that need to be ruled out include:

1) Bias? Unlikely, even though double-blinding wasn't possible, it's unlikely the measurement of the outcome (survival) was biased. It's also unlikely that this is a non-representative group of patients (sampling bias)
2) Confounding variables? No, we'd expect any problematic variables to be balanced across the two groups due to random assignment
3) Random chance? . . . This is where hypothesis testing is useful

Hypothesis testing for two-sample categorical data

- In Lister's experiment, we're interested in $H_0 : p_1 - p_2 = 0$, which implies the survival rates for the treatment and control groups are the same
- In our introduction to hypothesis testing, we evaluated the null hypothesis by *simulating outcomes* that would be expected if the null hypothesis were true
  - In particular, we used "coin flips" to model the toy choices of the study's 16 infants
  - Can we apply a similar approach to Lister's experiment?

Pooled proportions

- A major challenge is that there are many different ways in which the treatment and control groups could have the same survival rate, and each would satisfy the null hypothesis
  - However, most realistic is to assume that a *pooled proportion* applies to each group
- In Lister's experiment, $\hat{p}_0 = \frac{19+34}{75} = 0.707$ is the overall survival rate, regardless of group

Evaluating $H_0 : p_1 - p_2 = 0$

We can use pooled proportion to simulate the survival outcomes we'd expect to see in each group if H_0 were true:

```r
## Set seed (for replication purposes)
set.seed(15)
nsim = 1000

## Simulate survival for the control group
control <- rbinom(nsim, size = 35, prob = 0.707)

## Simulate survival for the sterile group
sterile <- rbinom(nsim, size = 40, prob = 0.707)
```

Evaluating $H_0 : p_1 - p_2 = 0$

These simulated outcomes can be used to estimate the *p*-value, or the probability of a difference in survival that's at least as large as $19/35 - 34/40 = -0.307$

```r
## Simulated differences in proportions
diffs <- control/35 - sterile/40

## Estimate the two-sided p-value
2*sum(diffs <= (19/35 - 34/40))/nsim
```

```
## [1] 0.004
```

So, a difference in survival as large as the one seen in Lister's experiment would only happen 0.4% of the time if the "sterile" protocol made no difference.

Fisher's exact test

The simulation approach described on the previous few slides is an approximation of a method known as **Fisher's exact test**:

```
##
##  Fisher's Exact Test for Count Data
##
## data:  table
## p-value = 0.005018
## alternative hypothesis: true odds ratio is not equal to 1
## 95 percent confidence interval:
##   1.437621 17.166416
## sample estimates:
## odds ratio
##   4.666849
```

Fisher's exact test should be used to test for a difference in categorical outcomes across two groups. You can view it as a generalization of the exact binomial test.

The two-sample Z-test

Another way to use the pooled proportion is within the *standard error* suggested by the Central Limit theorem result:

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

This approach allows us to calculate a Z-value and perform a Z-test:

$$Z = \frac{\text{Observed}-\text{Null}}{SE} = \frac{(\hat{p}_1-\hat{p}_2)-0}{\sqrt{\frac{\hat{p}_0(1-\hat{p}_0)}{n_1}+\frac{\hat{p}_0(1-\hat{p}_0)}{n_2}}}$$

Remember that $\hat{p}_0$ is the *pooled proportion*, it represents the most likely survival rate when the null hypothesis is true

Example - two-sample Z-test

For Lister's experiment:

1) $H_0 : p_1 - p_2 = 0$ vs. $H_a : p_1 - p_2 \neq 0$
2) The pooled proportion that best reflects H_0 is: $\hat{p}_0 = \frac{19+34}{75} = 0.707$
3) $Z = \frac{\text{Observed}-\text{Null}}{SE} = \frac{(19/35-34/40)-0}{\sqrt{\frac{0.707(1-0.707)}{35}+\frac{0.707(1-0.707)}{40}}} = \text{-}2.916$

Example - two-sample Z-test

For Lister's experiment:

1) $H_0 : p_1 - p_2 = 0$ vs. $H_a : p_1 - p_2 \neq 0$
2) The pooled proportion that best reflects H_0 is: $\hat{p}_0 = \frac{19+34}{75} = 0.707$
3) $Z = \frac{\text{Observed}-\text{Null}}{SE} = \frac{(19/35-34/40)-0}{\sqrt{\frac{0.707(1-0.707)}{35}+\frac{0.707(1-0.707)}{40}}} =$ -2.916
4) The two-sided p-value is 0.0035 (see R code below). Thus, we can conclude that Lister's sterilization protocol *causes* an improvement in survival.

```
2*pnorm(-2.916, mean = 0, sd = 1, lower.tail = TRUE)
```

```
## [1] 0.003545505
```

Remarks - sample size and choice of statistical test

- Fisher's exact test is computationally expensive (especially for larger samples)
    - The two-sample Z-test is generally recommended when the "success-failure condition" is met for *both groups* (ie: the sample data contain at least 10 "successes" and 10 "failures" in each group)
    - However, modern computing has made it feasible to use Fisher's exact test in most circumstances
- Both tests produce a similar p-value for large samples, but the two-sample Z-test can be unreliable when the success-failure condition is not met

Practice

In 2015-16, the Golden State Warriors set an NBA record for most wins in a season. The table below shows a breakdown of the Warrior's wins and losses by whether the game was played on their home court, or on their opponent's court:

```
gsw = read.csv("https://remiller1450.github.io/data/GSWarriors.csv")
table(gsw$Location, gsw$Win)
```

```
##
##          L  W
##   Away   7 34
##   Home   2 39
```

1) Perform a two-sample Z-test to evaluate whether the observed difference in the Warrior's home vs. away success could be explained by random chance.
2) Briefly explain why a two-sample Z-test might be inappropriate, then analyze these data using Fisher's exact test (the preferred approach in this application)

Practice (solution)

1) $H_0 : p_1 - p_2 = 0$, where p_1 is the proportion of wins at home and p_2 is the proportion wins on the road. Then, $Z = \frac{(34/41-39/41)-0}{0.069} = 1.77$, where 0.069 is the standard error calculated using the pooled proportion. The two-sided p-value corresponding to this Z-value is 0.077, so there's borderline evidence of better performance at home.
2) The null hypothesis is still $H_0 : p_1 - p_2 = 0$, see the R code below for the p-value:

```r
gsw = read.csv("https://remiller1450.github.io/data/GSWarriors.csv")
fisher.test(table(gsw$Location, gsw$Win))$p.value
```

```
## [1] 0.1549418
```

Hypothesis testing for two-sample quantitative data

In order to test for a difference in means, we can begin with the same general approach as the two-sample Z-test:

- Propose $H_0 : \mu_1 - \mu_2 = 0$
- Find the corresponding sample outcome, $\bar{x}_1 - \bar{x}_2$
- Using CLT, estimate $SE = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$, where s_1 and s_2 are the sample standard deviations of each group

At this point we've estimated *two extra population parameters* using the sample data, so we must use the T-distribution:

$$T = \frac{\text{Observed} - \text{Null}}{SE} = \frac{(\bar{x}_1 - \bar{x}_2) - 0}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

Degrees of freedom are complicated, we'll either use `R` or take the smaller of $n_1 - 1$ and $n_2 - 1$ if forced to work "by hand"

Practice

- At the 2008 Beijing Olympics, 25 different swimming world records were broken, the most since 1976, when goggles were first used in competition

Practice

- At the 2008 Beijing Olympics, 25 different swimming world records were broken, the most since 1976, when goggles were first used in competition
  - Of these 25 new records, 23 were set by swimmers using a wetsuit known as the *LZR Racer*, a suit produced by Speedo whose design involved scientists at NASA
- But is this convincing evidence that LZR Racer provides an unfair advantage?
  - Are there any alternative explanations for 23 of 25 records being set by swimmers who wore LZR Racers?

Practice

- At the 2008 Beijing Olympics, 25 different swimming world records were broken, the most since 1976, when goggles were first used in competition
  - Of these 25 new records, 23 were set by swimmers using a wetsuit known as the *LZR Racer*, a suit produced by Speedo whose design involved scientists at NASA
- But is this convincing evidence that LZR Racer provides an unfair advantage?
  - Are there any alternative explanations for 23 of 25 records being set by swimmers who wore LZR Racers?
- Recognize that these data are *observational*, so it could be that all of the best swimmers were wearing this suit. Therefore, an *experimental* study should be performed

Practice

- The `wetsuits` data contains the results of an experiment involving 12 competitive swimmers
  - Each swam 1500m for time under two conditions: wearing a high-tech wetsuit, or wearing a placebo suit identical in appearance
  - It was randomly determined which condition the participant experienced first
- The columns `Wetsuit` and `NoWetsuit` record the respective velocities (in m/s) over the 1500m swim

```r
wet <- read.csv("https://remiller1450.github.io/data/Wetsuits2.csv")
```

Use `R` to find the sample mean and standard deviation of each group, then perform a two-sample T-test "by hand"

Practice (solution)

- Consider $H_0 : \mu_1 - \mu_2 = 0$, where μ_1 is the average velocity when wearing a wetsuit and μ_2 is the average velocity when wearing a normal swimsuit.
- We observed $\bar{x}_1 = 1.507$, $\bar{x}_2 = 1.429$, $s_1 = 0.136$, and $s_2 = 0.141$
- Thus, the T-value relating the sample data to the null hypothesis is $T = \frac{(1.507 - 1.429) - 0}{\sqrt{0.136^2/12 + 0.141^2/12}} = 1.379$
- Comparing this against a t-distribution with $df = 11$, the two-sided p-value is 0.195, indicating insufficient evidence of any difference in velocity

The t.test function

We can use the `t.test` function in R to perform this test using the precise degrees of freedom:

```
t.test(x = wet$Wetsuit, y = wet$NoWetsuit, alternative = "two.sided")
```

```
##
##  Welch Two Sample t-test
##
## data:  wet$Wetsuit and wet$NoWetsuit
## t = 1.3688, df = 21.974, p-value = 0.1849
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.03992937  0.19492937
## sample estimates:
## mean of x mean of y
##  1.506667  1.429167
```

Notice the actual degrees of freedom are slightly below $n_1 + n_2 - 2 = 22$, which why the "by hand" approach uses the lower bound $\min(n_1 - 1, n_2 - 1)$

Non-parametric alternatives

The two-sample *t*-test is designed to work in two settings:

1) Small, Normally distributed samples
2) Large samples of any distributional shape (ie: $n_1 \geq 30$ and $n_2 \geq 30$)

Outside of these settings, the Wilcoxon Rank-Sum test can be used to test whether the medians of each group are equal:

```
## Warning in wilcox.test.default(x = wet$Wetsuit, y = wet$NoWetsuit, alternative =
## "two.sided"): cannot compute exact p-value with ties

##
##  Wilcoxon rank sum test with continuity correction
##
## data:  wet$Wetsuit and wet$NoWetsuit
## W = 95.5, p-value = 0.1838
## alternative hypothesis: true location shift is not equal to 0
```

Comments - paired designs

The “Wetsuits” study used a **paired design** where each subject served as their own control. Therefore, we should treat it as *one-sample data* and analyze the *paired differences*:

```
t.test(wet$Difference, mu = 0)$p.value
```

```
## [1] 8.885414e-08
```

```
t.test(x = wet$Wetsuit, y = wet$NoWetsuit, mu = 0)$p.value
```

```
## [1] 0.1848961
```

Paired designs can provide a tremendous statistical advantage (variability within individuals tends to be lower than variability between individuals), and they also help control for confounding variables!

Summary

This presentation covered two new hypothesis testing scenarios:

1) Two-sample categorical data, where we evaluate $H_0 : p_1 - p_2 = 0$ using either Fisher's exact test or a two-sample Z-test
2) Two-sample quantitative data, where we evaluate $H_0 : \mu_1 - \mu_2 = 0$ using either a two-sample T-test or the Wilcoxon Rank-Sum test

All of the fundamental concepts we've previously covered apply to these new situations, but we must be aware of when and how to implement these new statistical tests.