Math 300 Lesson 29 Notes

Permutation Tests

YOUR NAME HERE

July, 2022

Contents

Objectives

1. Explain the permutation test and compare and contrast it with a bootstrap distribution.
2. Generate and interpret visualizations for data with two categorical variables.

Reading

Chapter 9 - 9.1

Lesson

There are no learning checks for this lesson.

- This is a key idea from the reading: *The same can be said for confidence intervals. There was one general framework that applies to all confidence intervals and the* `infer` *package was designed around this framework. While the specifics may change slightly for different types of confidence intervals, the general framework stays the same. We believe that this approach is much better for long-term learning than focusing on specific details for specific confidence intervals using theory-based approaches. As you'll now see, we prefer this general framework for hypothesis tests as well.*
- *Computer-based methods using randomization, simulation, and bootstrapping have much fewer restrictions.*
- We are going to bring ideas from earlier in the course to start our work on hypothesis testing.

Libraries

```
library(tidyverse)
library(infer)
library(moderndive)
library(nycflights13)
library(ggplot2movies)
```

Promotions problem

The reading discusses the research question "Does gender affect promotions at a bank?". Similar to the yawning experiment, we have two categorical variables and so many of the tools will be the same.

Summarizing data

Let's start by exploring the data.

```
promotions %>%
  group_by(decision, gender) %>%
  tally()
```

```
## # A tibble: 4 x 3
## # Groups:   decision [2]
##   decision gender     n
##   <fct>    <fct>  <int>
## 1 not      male       3
## 2 not      female    10
## 3 promoted male      21
## 4 promoted female    14
```

- What is the impact of changing the order of the variables in the `group_by()` function and which is better for a decision maker based on the research question?

Answer:

Visulaizing data

- Create a barchart to summarize the data.

Answer:

Bootstrap confidence interval

- Find a 90% bootstrap confidence interval the difference in proportions. Interpret this confidence interval.

Answer:

- Get bootstrap distribution

```
set.seed(612)
```

- Visualize the results.
- Get confidence interval

Permutation test

We want to answer the same research question but we approach it differently. The key idea is that we assume there is no difference between males and females and then find the sampling distribution under this assumption. For our problem this implies that the column `gender` is irrelevant. Thus we could arbitrarily change, shuffle, these labels. That is what we will do.

Here is the first six rows of the original data.

```
head(promotions)
```

```
## # A tibble: 6 x 3
##      id decision gender
##   <int> <fct>    <fct>
## 1     1 promoted male
## 2     2 promoted male
## 3     3 promoted male
## 4     4 promoted male
## 5     5 promoted male
## 6     6 promoted male
```

Let's shuffle the `gender` column.

```
# Complete the code
set.seed(272)
# promotions_mod <- promotions %>%
#   mutate(shuffled=sample(_____))
```

```
#head(promotions_mod)
```

Notice that the number of males and females has not changed since by default we sampled without replacement. However, the number promoted for each gender level has changed.

```
# Complete the code
# promotions_mod %>%
#   group_by(_________, decision) %>%
#   tally()  %>%
#   mutate(perc = round(_____ / sum(_______)*100,2))
```

This is the heart of the permutation test. We shuffle the gender labels repeatedly and find the test statistic. This allows us to determine if the observed test statistic from the original sample is likely under the assumption of no difference in promotion rates for males and females.

We will be using the `infer` package to conduct the permutation test later in this chapter. But before we learn how to do this, let's use data where students randomly shuffled the data.

Read in the data.

```
gender_promotions_shuffles <- read_csv("ch9_gender_promotions_shuffles.csv")
```

Look at the first 6 rows.

```
# Complete the code
```

We need to do some data wrangling to get it into **tidy** data. You are not accountable to these steps but we want to demonstrate data wrangling. The `infer` package is going to make this easier for us.

Get the data in tidy form.

```
# Complete the code
# shuffled_data_tidy  <- gender_promotions_shuffles %>%
#   pivot_longer(cols=-c("id", "decision"),
#                names_to = "team", values_to = "gender") %>%
#   mutate(replicate = rep(1:16, times = 48))
```

Check that we did this correctly.

```
#head(shuffled_data_tidy,n=17)
```

And another check.

```
#shuffled_data_tidy %>% group_by(replicate) %>% count(gender)
```

Let's find the difference in proportions between male and females.

```
# test_stats <- shuffled_data_tidy %>%
#   group_by(replicate) %>%
#   count(gender, decision) %>%
#   filter(decision == "promoted") %>%
#   mutate(prop = n / 24) %>%
#   select(replicate, gender, prop) %>%
#   pivot_wider(names_from="gender",values_from = prop) %>%
#   mutate(stat = m - f)
```

```
#head(test_stats)
```

```
# ggplot(data = test_stats, aes(x = stat)) +
#   geom_histogram(binwidth = 0.1, fill = "cyan", color = "black") +
#   geom_vline(xintercept = (21/24 -14/24 ), color = "red", size = 1) +
#   labs(x = "Difference in promotion rates (male - female)")
```

- Interpret this histogram

Answer:

Documenting software

- File creation date: 2022-07-07
- R version 4.1.3 (2022-03-10)
- `tidyverse` package version: 1.3.1
- `moderndive` package version: 0.5.4
- `infer` package version: 1.0.2
- `nycflights13` package version: 1.0.2
- `ggplot2movies` package version: 0.0.1