Lab 4: Comparing Two Proportions

Prof. Adam Loy

Wednesday, October 12, 2016

0. Intro

Welcome to Lab 4!

Today's lab will continue to explore the idea of hypothesis testing using permutation tests, but this lab will focus on the situation where we have two categorical variables.

0.1 Administrative details

Lab submissions are due by Friday, October 14 by 4:00 p.m. To submit your assignment, please upload both the .Rmd and .html files on Moodle.

0.2 Setup

You can download the .Rmd file for this lab and the data from the course webpage as a zip file. If you are using a Mac, then you will need to use a browser other than Safari for the download.

If you are using the RStudio server, then you can upload the entire zip folder directly onto the server.

We will use the following packages during this lab. Make sure that you have downloaded all of them before running the commands.

```
library(ggplot2)
library(dplyr)
library(CarletonStats)
```

1. Study details

Is yawning contagious? Conventional wisdom says yes: When we see someone else yawn, we are prone to let out a yawn ourselves, but will data support this claim if we put it to a scientific test?

The folks at *MythBusters*, a popular television program on the Discovery Channel, investigated this issue by using a hidden camera. Fifty people attending a local flea market were recruited to participate. Subjects were ushered, one at a time, into one of three rooms by co-host Kari. She yawned (planting a yawn "seed") as she ushered subjects into two of the rooms, and for the other room she did not yawn. The researchers decided in advance, with a random mechanism, which subjects went to which room. As time passed, the researchers watched to see which subjects yawned.

The data for this study are found in the file `yawning.csv`.

Question 1. Is this an observational study or an experiment? Briefly justify your answer.

Question 2. Why did the researchers randomly assign subjects to the "yawn seed" group?

Now that we understand the study design, we can load the data set.

```
yawning <- read.csv("data/yawning.csv")
summary(yawning)
```

```
##     YawnSeed    Response
##  Control:16   NoYawn:36
##  Seeded :34   Yawn  :14
```

```
str(yawning)
```

```
## 'data.frame':    50 obs. of  2 variables:
##  $ YawnSeed: Factor w/ 2 levels "Control","Seeded": 2 2 2 2 2 2 2 2 2 2 ...
##  $ Response: Factor w/ 2 levels "NoYawn","Yawn": 2 2 2 2 2 2 2 2 2 2 ...
```

The rows describe the individual subjects. The columns describe:

- **YawnSeed**: Whether or not the "yawn seed" was planted (`Control` or `Seeded`)
- **Response**: Whether or not the subject yawned (`NoYawn` or `Yawn`)

The researchers goal is to determine whether or not yawning is contagious.

Question 3. State the null hypothesis using words and notation.

Question 4. State the alternative hypothesis using words and notation.

2. Exploratory data analysis

Before conducting a hypothesis test to determine whether or not yawning is contagious, we will explore the association present in the data set.

Question 5. Create a bar chart of whether or not the subject yawned (`Response`) segmented by whether or not the "yawn seed" was planted (`YawnSeed`).

We can summarize the data numerically in a two-way table since both variables are categorical. Using `dplyr` we can calculate all of the necessary counts using the `group_by` and `summarize` functions, as shown below. Note that this table does not look like a traditional two-way table because it is in tidy format.

```
yawn_by_seed <- group_by(yawning, YawnSeed, Response)
summarize(yawn_by_seed, counts = n())
```

```
## Source: local data frame [4 x 3]
## Groups: YawnSeed [?]
##
##   YawnSeed Response counts
##     <fctr>   <fctr>  <int>
## 1  Control   NoYawn     13
## 2  Control     Yawn      3
## 3   Seeded   NoYawn     23
## 4   Seeded     Yawn     11
```

Question 6. Calculate the proportion of subjects in the "yawn seed" group that yawned.

Question 7. Calculate the proportion of subjects in the control group that yawned.

Question 8. Describe the apparent association between the variables revealed by the plot and summary statistics.

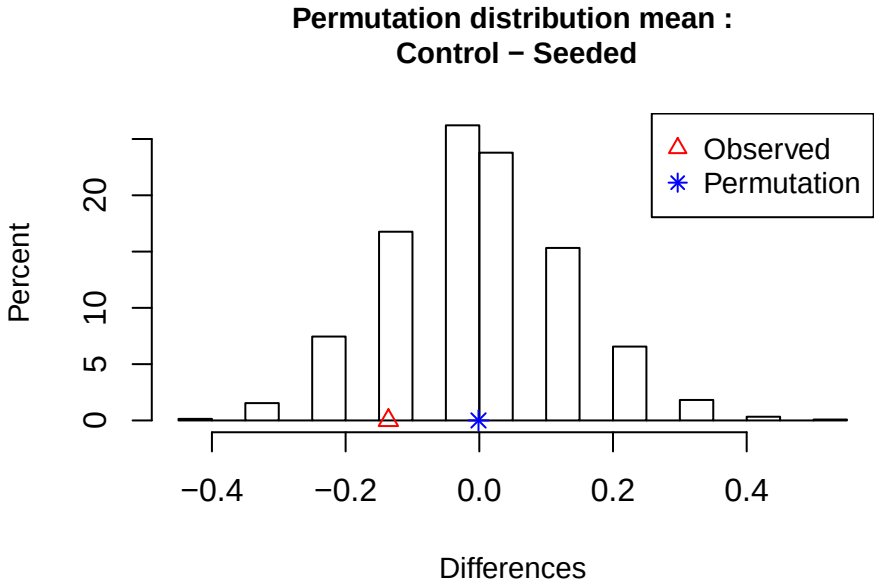
2. Testing the hypotheses

Recall that we use the `permTest` function to create a permutation distribution consisting of 9,999 simulations. `permTest` expects a formula specifying `response_variable ~ explanatory_variable`. In this study the formula would be `Response ~ YawnSeed`; however, `Response` is a categorical variable and the `permTest` function expects a quantitative variable. To force R to create a variable that appears to be quantitative we

will use `as.numeric(Response == "Yawn")` in place of `Yawn`. This results in the value `Yawn` being changed to a `1` and the value `NoYawn` being changed to a `0`.

Now that we know how to deal with the case of two categorical variables we can run a permutation test.

```
test <- permTest(as.numeric(Response == "Yawn") ~ YawnSeed, data = yawning, alternative = "less")
```




```
##
##  ** Permutation test **
##
##  Permutation test with alternative: less
##  Observed  mean
##   Control :  0.1875   Seeded :  0.32353
##  Observed difference: -0.13603
##
##  Mean of permutation distribution: -0.00122
##  Standard error of permutation distribution: 0.13795
##  P-value:  0.259
##
##  *-------------*
```

Using the above results answer the following:

Question 9. Is the permutation distribution centered around 0? (Note: the exact center will change each time you run the code, so generalize here.) Explain why this makes sense.

Question 10. Is the observed value of the statistic out in the tail of the permutation distribution or not so much? In other words, does the observed result appear to be typical or surprising when the null hypothesis is true?

Question 11. What is the p-value reported in the output of the `permTest` function? How is this calculated?

Question 12. Using the guidelines for assessing the *strength of evidence* from p-values, would you conclude that the *MythBusters* result provides much evidence that yawning is contagious?

Question 13. If you have decided that the two groups differ significantly, would you be justified in drawing a cause-and-effect conclusion between the yawn seed and an increased probability of yawning? Explain, based on how this study was conducted.

Question 14. Based on how the sample was selected, to what larger population would you feel comfortable generalizing the results of this study, if any? Justify your answer.

3. Effect of sample size

What if the the yawning study involved 500 people, 10 times as many as the actual study, but that the proportions of subjects who yawned in each group are unchanged? The file `yawning500.csv` contains these hypothetical data.

```
# Loading the new data set
yawning500 <- read.csv("data/yawning500.csv")
```

Question 15. Use the `permTest` function to create a permutation distribution consisting of 9,999 simulations. Create a histogram of the permutation distribution with a superimposed marker representing the observed difference in the proportion of yawns between the yawn seed and control groups.

Question 16. How did the permutation distribution change?

Question 17. What is the new p-value?

What if the the yawning study involved 5,000 people, 100 times as many as the actual study, but that the proportions of subjects who yawned in each group are unchanged? The file `yawning5000.csv` contains these hypothetical data.

```
# Loading the new data set
yawning5000 <- read.csv("data/yawning5000.csv")
```

Question 18. Use the `permTest` function to create a permutation distribution consisting of 9,999 simulations. Create a histogram of the permutation distribution with a superimposed marker representing the observed difference in the proportion of yawns between the yawn seed and control groups.

Question 19. How did the permutation distribution change?

Question 20. What is the new p-value?

Question 21. What is happening to the p-value as the sample size is increasing? Why might this be?

Acknowledgements

This lab was adapted from exploration 5.2 from *Introduction to Statistical Investigations* by Tintle et al.