Lab 3: Introduction to Hypothesis Testing

0. Intro

Welcome to Lab 3!

Today's lab will explore the idea of hypothesis testing using random permutations. This technique is described in the chapter 4 of the textbook, in John Rauser's keynote address at Strata + Hadoop 2014, and is used often in practice.

0.1 Administrative details

Lab submissions are due by Monday, October 10 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.

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

1. Comparing Two Groups

Many studies focus on the comparison of groups. In observational studies, this comparison focuses on the how an attribute (response variable) differs between naturally occurring groups in a sample from the population of interest. In experiments, this comparison focuses on how an attribute differs across treatment groups. Both observational studies and experiments rely on samples from a population of interest, so there is a chance that the association we observe is due to the composition of the sample we drew rather than truly existing in the population. Stated another way, the response variable may appear associated with the explanatory variable because there is in fact an association in the population, or simply because the sample happened to come out that way. The purpose of statistical hypothesis testing is to quantitatively account for the possibility that two attributes may appear related in a sample even though they are not related in the population.

In this lab, we will explore the permutation test discussed in John Rauser's keynote address at Strata + Hadoop 2014, investigating how to formalize this test and implement it in R.

1.1. The data

Does consuming beer make you more attractive to mosquitoes? This is the question Levre et al. (2010) strove to answer. In this study conducted in Burkina Faso, Africa, 42 volunteers were recruited and randomly assigned to a treatment group: 25 volunteers consumed a liter of beer and 18 volunteers consumed a liter of water. The attractiveness of the volunteers to mosquitoes was then tested. Mosquitoes were released and caught in traps as they approached the volunteers. The resulting data are recorded in the file `mosquitoes.csv`.

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

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

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

```
##        id       treatment      count      
##  Min.   : 1.0   beer :25   Min.   :12.00  
##  1st Qu.:11.5   water:18   1st Qu.:19.00  
##  Median :22.0              Median :21.00  
##  Mean   :22.0              Mean   :21.77  
##  3rd Qu.:32.5              3rd Qu.:24.00  
##  Max.   :43.0              Max.   :31.00  
```

The rows describe the individual participants. The columns describe:

- **id**: An identifier for each participant
- **treatment**: Consumption of a liter of `beer` or a liter of `water`
- **count**: The number of mosquitoes caught in traps as they approached the volunteer

1.2. Exploring the association

Before conducting a statistical test for whether the association between beer consumption and attractiveness to mosquitoes is likely to be due to chance (as opposed to an association in the population), we will explore the possible association.

Question 2. Create side-by-side boxplots of the number of mosquitoes captured by group.

Question 3. Create overlaid density plots of the number of mosquitoes captured by group.

Question 4. What do the plots created in the previous two questions reveal about the association between beer consumption and one's attractiveness to mosquitoes?

Question 5. Calculate the following summary statistics for each treatment group: the five number summary, the mean, the standard deviation. Do these summary statistics support what you observed in the previous question?

2. Using Random Permutations

From the above summary statistics you should find that the difference in the mean number of mosquitoes attracted between the two groups is $\overline{x}_{\text{beer}} - \overline{x}_{\text{water}} = 4.4$. Is this a large enough difference to suggest that there are significantly more mosquitoes attracted to the beer group, on average?

There is no universal threshold for how large a difference in means must be to be "significantly different from zero." Instead, we compare the observed test statistic ($\overline{x}_{\text{beer}} - \overline{x}_{\text{water}}$) to a distribution that is generated empirically by randomly permuting the values in the data set.

The purpose of this comparison is to help us choose between two hypotheses:

Null hypothesis: The observed difference between distributions for the two treatments is due to chance because we randomly assigned the volunteers to a treatment group.

Alternative hypothesis: The observed difference is **not** due to random chance, but instead due to the treatment.

A **permutation test** (your book calls it a randomization test) generates groups that look like the groups we would see if the null hypothesis were true.

2.1. Hypothesis Tests

Each hypothesis test has four steps.

Step 1 is to **state the null and alternative hypotheses**. For example, if we are interested in whether beer increases the average number of mosquitoes attracted, we would state the following:

Null hypothesis: The true average number of mosquitoes attracted between the treatment groups is the same, so the observed difference is due to randomness introduced by random assignment.

The alternative is an explanation that contradicts the null hypothesis: there is a systematic difference that cannot be blamed on random assignment.

Alternative hypothesis: The difference in the average number of mosquitoes attracted is not due to chance, but instead due to the treatment.

Step 2 is to **select and compute a test statistic** that summarizes the data relevant to the null hypothesis. The test statistic should be a statistic that generally looks one way if the null hypothesis is true and another way if the alternative hypothesis is true.

Test statistic: The difference in the average number of mosquitoes attracted between the two treatment groups: $\overline{x}_{\text{beer}} - \overline{x}_{\text{water}}$.

Step 3 is to estimate the distribution of the test statistic assuming that the null hypothesis is true. Given only a sample, we cannot draw new samples directly from the population. Instead, we randomly permute the pairings of conditions and values to see how the test statistic would vary for a sample of the given size, the split between conditions, and the observed difference in means.

We could do this manually using code similar to the below, though we would need to repeat it a large number of times, say 9,999.

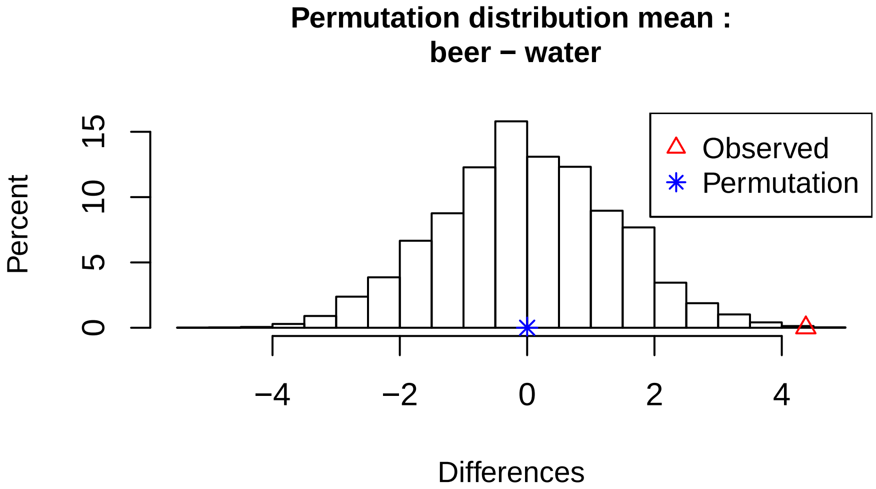
```r
## We could manually reshuffle the treatment column
## and recompute the difference in means
shuffled <- mutate(mosquitos, treatment = sample(treatment))

mean_shuffled <- shuffled %>%
  group_by(treatment) %>%
  summarize(avg = mean(count))

diff(mean_shuffled$avg)
```

```
## [1] -0.9377778
```

Alternatively, we can use the `permTest` function in R carryout the entire permutation test. Note that the below code prints the results of the test on the screen and also produces a histogram of the permutation distribution. `permTest` requires the following arguments:

- a formula of the form `response_variable ~ explanatory_variable`,
- a `data` set,
- and the `alternative` hypothesis. We use `greater` or `less` if the alternative hypothesis specifies that the difference is likely to be positive or negative; otherwise, we use `two.sided` (i.e. the difference could fall in either tail).

```r
permutations <- permTest(count ~ treatment, data = mosquitos, alternative = "greater")
```



```
##
##  ** Permutation test **
##
##  Permutation test with alternative: greater
##  Observed  mean
##   beer :  23.6    water :  19.22222
##  Observed difference: 4.37778
##
##  Mean of permutation distribution: 0.00014
##  Standard error of permutation distribution: 1.37397
##  P-value:  7e-04
##
##  *-------------*
```

Question 6. What is the mean of the permutation distribution? (Note: this will change each time you run the code.)

Step 4 is to interpret the result and draw a conclusion. If the observed test statistic is "too far out" in the tails of the permutation distribution, then it would rarely happen by chance alone, providing evidence that something else is likely happening.

Question 7. Do you think that the observed test statistic is too far out in the tail of the permutation distribution to be due to chance alone? Why or why not?

Instead of visually comparing the observed test statistic to the permutation distribution, we can also calculate the proportion of simulated test statistics in the permutation distribution that are at least as far out in the tail of the permutation distribution as the observed test statistic. This proportion is called the **p-value**, and is reported in the output of the `permTest` function.

Question 8. What p-value was reported for your permutation test?

A p-value of 0.05 or below is conventionally called "statistically significant" and a p-value of 0.01 or below is conventionally called "highly statistically significant", although these thresholds are arbitrary.

Question 9. Is the observed difference in average mosquito attractiveness statistically significant? Is it highly statistically significant?

Question 10. If the results are statistically significant, what does this indicate about beer's impact on mosquito attractiveness?

3. Wrapping Up

Congratulations, you just carried out your first hypothesis test using random permutations! Be sure to review the four steps of hypothesis tests and the underlying logic of permutation tests before our next class. We will see more examples of permutation tests and discuss some important issues surrounding hypothesis testing, but the underlying logic won't change.

References

John Rauser's keynote address at Strata + Hadoop 2014 serves as great motivation to learn about permutation tests, and reference bootstrapping. (We will do that too!)

Lefèvre, Thierry, et al. "Beer consumption increases human attractiveness to malaria mosquitoes." *PloS one* 5.3 (2010): e9546. http://dx.doi.org/10.1371/journal.pone.0009546