Data Analysis and Visualization in R (IN2339)

Exercise Session 7 - Statistical Testing I

Felix Brechtmann, Jun Cheng, Vicente Yepez, Julien Gagneur

Quiz

1. The null hypothesis of a study states 'Both genders of US Olympic competitors are equally likely to win gold medals at the Olympics'. What are possible one-tailed alternative hypotheses? More than one answer can be correct.

a) There are gender differences in US Olympic competitors in the number of gold medals won at the Olympics.
b) Female US Olympic competitors win more gold medals at the Olympics.
c) Male US Olympic competitors win more gold medals at the Olympics.

2. The p-value of a certain hypothesis test is 0.007. What can the researcher conclude? More than one answer can be correct.

a) If the study is repeated 1000 times, 7 times will fail to produce a significant result.
b) We can be only 0.7% confident that the null hypothesis is true.
c) The effect size of the study is large.
d) Assuming the null hypothesis is true, the probability to make this extreme or more extreme observation is 0.7%.

3. Let X be a vector that contains some collected data. Which of the following lines of code produces one sample of a case resampling bootstrap? More than one answer can be correct.

a) sample(X, size = length(X), replace = T)
b) sample(X, size = length(X), replace = F)
c) sample(X, size = length(X), replace = T, prob = 1/length(X))
d) sample(X, size = length(X), replace = T, prob = rep(c(0.2, 0.5), each = length(X) / 2))

Tutorial

The following exercises will be solved during the tutorial sessions.

Section 00 - Getting Ready

1. Make sure you have already installed and loaded the following libraries:

```
library(ggplot2)
library(data.table)
library(magrittr)
library(tidyr)
library(dplyr)
library(patchwork) # optional, makes plots nicer
```

2. Load the yeast data

```
genotype <- fread("./extdata/eqtl/genotype.txt")
genotype <- melt(genotype, id.vars = "strain", variable.name = "marker",
                 value.name = "genotype")
growth <- fread("./extdata/eqtl/growth.txt")
growth <- melt(growth, id.vars = "strain", variable.name = "media",
               value.name = "growth_rate")
marker <- fread("./extdata/eqtl/marker.txt")
```

Section 01 - Permutation test of growth rate difference

1. The following code recreates the example shown in the lecture to test the association of the genotype at marker 5211 with the growth rate difference in Maltose medium. Note that the code is written using functions, meaning that it will work for any marker, not just marker 5211. Read it carefully to understand what happens in each function. Then execute the code. The Lecture example and a description of the dataset can be found here: https://gagneurlab.github.io/dataviz/resampling-stat.html#yeast-dataset.

```
# Plotting the growth rate difference
getMaltoseDt = function(mrk){
  growth_mrk <-  merge(growth, genotype[marker %in% mrk, .(strain, genotype, marker)],
                       by = 'strain', allow.cartesian = TRUE)
  growth_mrk[media == "YPMalt"]
}

# boxplot
plot_growth_one_mk <- function(mk){
    ggplot(getMaltoseDt(mk), aes(genotype, growth_rate)) +
    geom_boxplot() +
    labs(title = mk) + theme_bw(base_size = 16)
}
plot_growth_one_mk("mrk_5211")

# Function to calculate the difference of the medians of two genotypes
median_diff <- function(dt){
  dt[genotype == 'Wild isolate', median(growth_rate, na.rm=T)] -
    dt[genotype == 'Lab strain', median(growth_rate, na.rm=T)]
}

# Function to permute the table, plot the resulting histogram
# and compute a p-value
p_val_medians <- function(dt, N_permu = 1000){
  # It will return both a pvalue and plot a histogram of T_star
  T_ref <- median_diff(dt)
  T_star <- sapply(1:N_permu, function(x){
```

```
      median_diff(dt[, genotype := sample(genotype)]) })
  # Plot
  g <- ggplot(data = data.table(T_star = T_star), aes(T_star)) + geom_histogram() +
    geom_vline(aes(xintercept=T_ref, color="T_ref")) + xlim(-3,3)
  print(g) # Needed to render plot inside function call

  # Compute and return the p value

  # First compute each tail seperately
  p_val_right <- (sum(T_star >= T_ref) + 1) / (N_permu + 1)
  p_val_left <- (sum(T_star <= T_ref) + 1) / (N_permu + 1)
  # Then combine the above to obtain the double sided p-value.
  p_val <- 2 * min(p_val_right, p_val_left)
  return(p_val)
}

# Calling the function:
p_val_medians(getMaltoseDt("mrk_5211"))
```

2. Using the code above, plot and test whether markers 1653 and 5091 associate with growth. Interpret your results.

Section 02 - Permutation test of marker association

1. We just concluded that both markers 5211 and 5091 are significantly associated with growth. However, this could be confounded. A common source of confounding in genomics is due to "linkage", which describes the phenomenon of markers being inherited together. A biological explanation for linkage is provided here: https://www.khanacademy.org/science/biology/classical-genetics/chromosomal-basis-of-genetics/a/linkage-mapping

To investigate the issue of linkage in our dataset, test if marker 5091 significantly associates with marker 5211. Define a null hypothesis, a statistic and use permutation testing to answer the question. Strengthen your answer with a relevant plot.

Hint: start from the table below and think about how this can be permuted.

```
mks_geno <- genotype[marker %in% c("mrk_5091", "mrk_5211")] %>%
  spread(marker, genotype)
head(mks_geno)
```

```
##     strain     mrk_5091     mrk_5211
## 1: seg_01B   Lab strain   Lab strain
## 2: seg_01C   Lab strain   Lab strain
## 3: seg_01D Wild isolate Wild isolate
## 4: seg_02B Wild isolate   Lab strain
## 5: seg_02C   Lab strain   Lab strain
## 6: seg_02D   Lab strain Wild isolate
```

Homework

Please solve the exercises below at home. The solutions will be discussed in the central exercise.

Section 03 - Controlling for a 3rd variable

1. We found that marker 5211 and marker 5091 are associated with growth. However, we also found that both markers are associated with each other. Thus, the association of one of these markers with growth could be explained away by the association of the other one with growth.

Now, we would like to know if marker 5091 still associates with growth in maltose (YPMalt) when conditioned on marker 5211. Define a null hypothesis, a statistic and use permutation testing to answer the question. Strengthen your answer with a relevant plot.

2. Now, test if marker 5211 associates with growth in maltose when conditioned on marker 5091. Are the results the same? Discuss.

Section 04 - Confidence Intervals

1. Estimate 95% equi-tailed confidence intervals for the difference of the medians of growth in maltose for each genotype at marker mrk_5211. Use the case resampling bootstrap scheme and report bootstrap percentile intervals. Propose a visualization of the results. Try it also with markers 5091 and 1653.