Data Analysis and Visualization in R (IN2339)

Exercise Session 8 - Statistical Testing II

Hasan Celik, Christian Mertes, Vicente Yépez, Alexander Karollus

Quiz

1. Let p be the (true) probability that a coin lands on heads. You flip the coin n times. Using the resulting data, you compute an estimate for p, which we call $\hat{p}$, and a valid 95% confidence interval. This confidence interval is $[0.48, 0.52]$. Indicate which, if any, of the following statements are **correct**:

a) There is a 95% chance that $0.48 \leq p \leq 0.52$

b) Assuming the null hypothesis is **correct**, there is a 95% chance that $0.48 \leq p \leq 0.52$

c) Assuming the null hypothesis is **incorrect**, there is a 95% chance that $0.48 \leq p \leq 0.52$

d) If we were to repeat the experiment 100 times, and compute a confidence interval for each replicate, we expect that only around 5 of the computed confidence intervals will not contain p

2. Consider the following (extreme) 2x2 contingency table and assume we want to test the association of taking antiviral medicine with having symptoms from a viral disease. Indicate which, if any, of the following statements are **correct**:

```
##              has_symptoms
## got_antiviral FALSE TRUE
##         FALSE     1  100
##         TRUE     99    0
```

a) A Fisher's test, with `alternative = "two.sided"`, applied to this table will return a low P-value

b) A Fisher's test, with `alternative = "greater"`, applied to this table will return a low P-value

c) A Fisher's test, with `alternative = "less"`, applied to this table will return a low P-value

3. This question concerns the t-test and the Wilcoxon test. Indicate which, if any, of the following statements are **correct**:

a) The t-statistic is defined based on the difference in median between groups

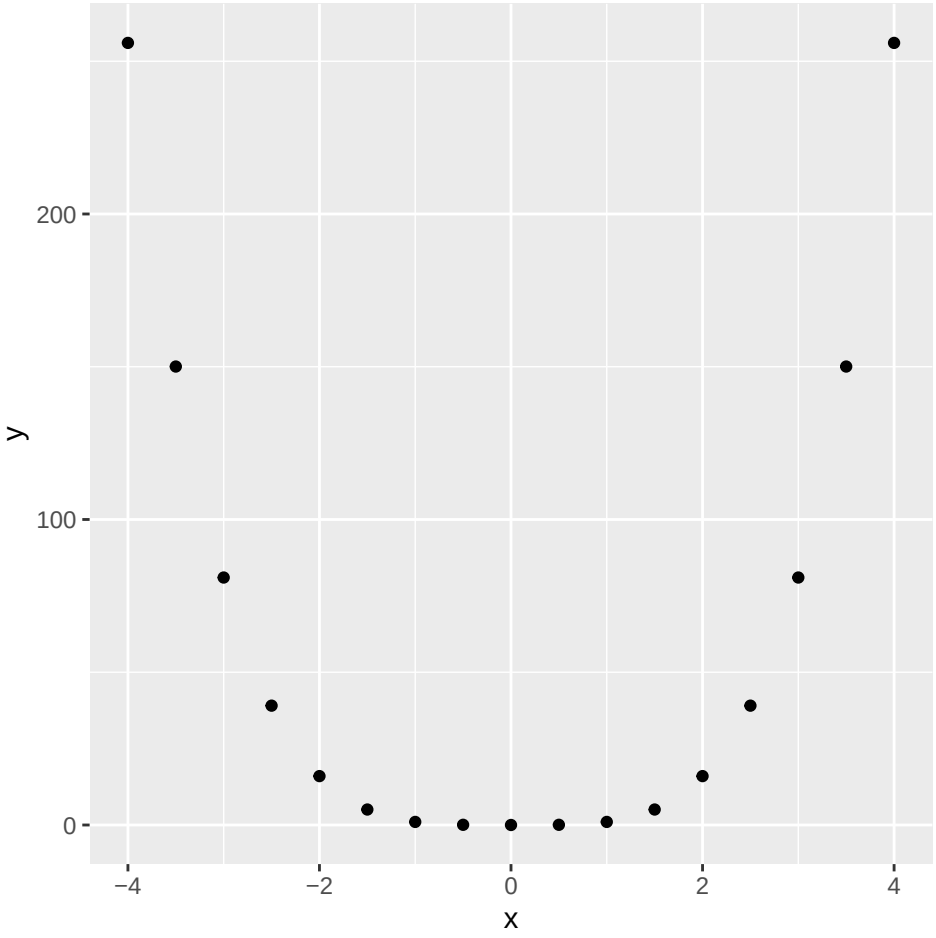
b) If one of the groups follows a bimodal distribution, then the t-test can still be applied without issue, because the t-statistic does not depend on the mode

c) If both groups follow a normal distribution, the t-test will be more powerful (i.e. more likely to detect deviations from the null) than the Wilcoxon

d) If the Wilcoxon test returns a very large P-value, e.g. $P > 0.9999$, then we can conclude that $P(X > Y) = P(Y > X)$, i.e. the ranks of our two groups follow the same distribution.

e) (Bonus) If you specify `exact=FALSE`, the R function `wilcox.test` assumes that the Mann-Whitney U test-statistic follows a normal distribution. Thus, the `wilcox.test` with `exact=FALSE` should not be used if you expect your data to deviate from normality

4. Look at the following plot, where $y = x^4$. Indicate which, if any, of the following statements about correlation are **correct**:



a) If we only consider points with $x > 0$, then the Pearson correlation of x with y will be higher than the Spearman correlation.

b) If we consider all points, both the Pearson correlation and the Spearman correlation of x with y will be zero.

c) If we only consider points with $x < 0$, both the Pearson correlation and the Spearman correlation of x with y will be negative.

d) If the Pearson and Spearman correlations between two variables are both zero, then the two variables are independent (i.e. knowing x gives you no information about y and vice-versa).

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(datasets)
```

2. Load yeast data and required packages using the following code:

```
gene <- fread("./extdata/eqtl/gene.txt")
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 - Warm Up: Choosing the right test

1. You are a data science consultant helping researchers pick the right tests to evaluate their hypotheses. For each hypothesis, indicate **which test** from the ones you have seen in the lecture would be most appropriate:

a) A researcher collects data on the height (measured in cm) and weight (measured in g) of Germans. She hypothesizes that there is a significant association between how tall Germans are and how much they weigh. She would like to test this hypothesis without making any distributional assumptions.

b) A researcher collects data on the weight (measured in g) of Bavarians before and after the Oktoberfest. She would like to know whether there is a significant difference in average weight after the Oktoberfest as compared to before it. Prior research indicates that the weight of Bavarians is approximately normally distributed.

c) A researcher is evaluating a rapid antigen test. The company manufacturing the test claims that if someone is infected, the test will correctly return a positive result 99% of the time. The researcher hypothesizes that, in practice, the test is often improperly administered and therefore significantly less sensitive. She asks 1000 individuals, which have all been confirmed to be infected by a PCR test, to self-administer the antigen test. She records how often the antigen test correctly returns a positive result.

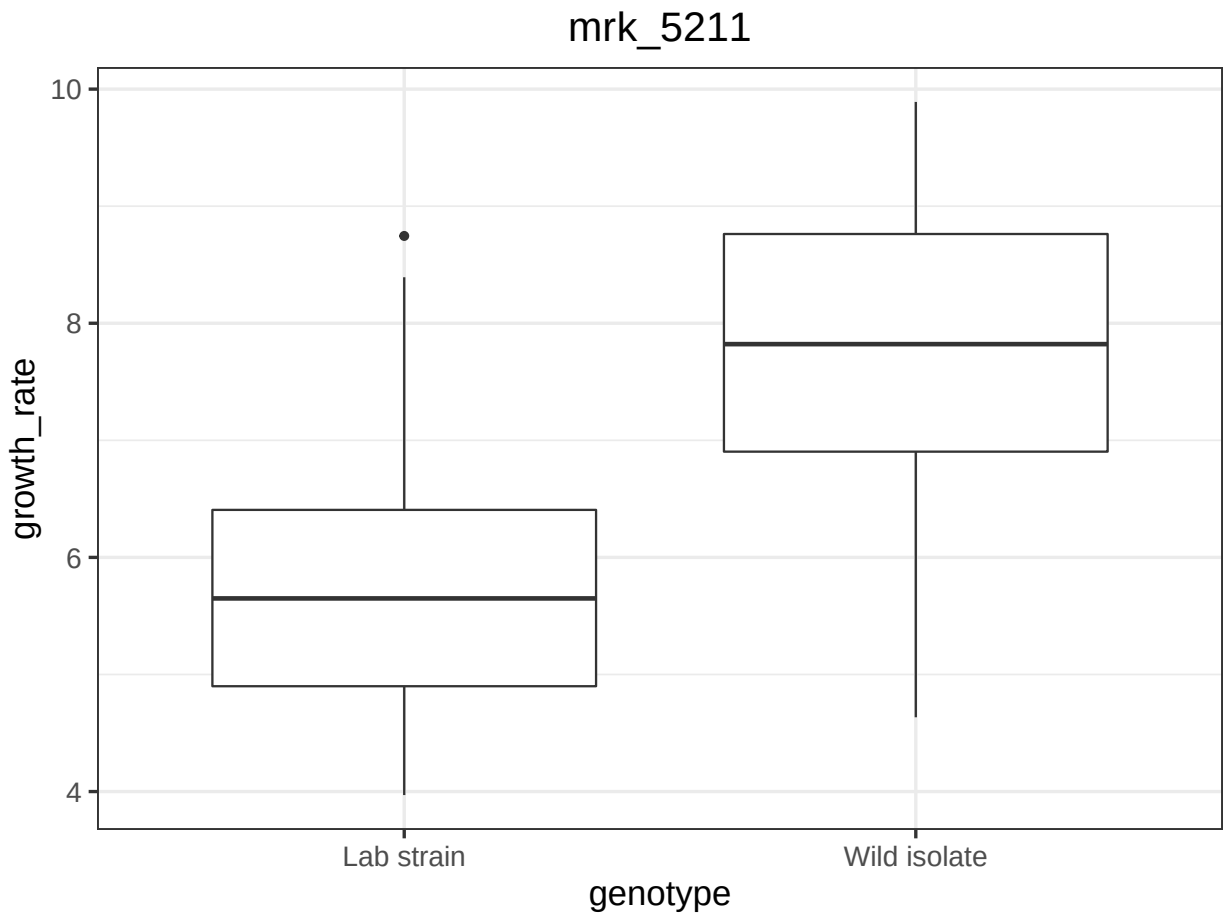
d) The company manufacturing the test has collected a bigger dataset, comprising both infected and non-infected individuals. For each individual, they record two datapoints: the result of a PCR test (infected/not-infected), which is taken as ground truth, and the result of a self-administered antigen test (positive/negative). They would like to show that, even if self-administered, the test still gives *some* information about infection status and thus is better than nothing.

Section 02 - Test the association between markers and growth

1. Reproduce the growth boxplot for marker 5211 using the following code:

```r
getMaltoseDt <- function(mrk){
  growth_mrk <-  merge(growth, genotype[marker == mrk, .(strain, genotype)],
                       by = 'strain')
  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) +
    theme(plot.title = element_text(hjust = 0.5))
}
plot_growth_one_mk("mrk_5211")
```



2. Last week, using permutation, we saw that some markers associated with growth. Which of the statistical tests from the lecture would you use to test this association? For each marker, we are not certain if the genotype will cause a positive or negative effect on growth; therefore, which kind of alternative hypothesis would you choose: double-sided, right or left? Apply the test to marker 5211 to obtain a p-value and see if the association that we found last week still holds. If more than one of the tests are appropriate, use them all and compare the results.

3. Given a marker and the name of a statistical test, make a function that returns the p-value of the association of that marker with respect to growth. You can use the template below to construct this

function. Then test your function on the markers 1653 and 5091. Since we used the same markers last week: do you get similar results as last week?

```
test_growth <- function(mk, test){

  m_dt <- getMaltoseDt(mk)

  if(test == 'wilcoxon') {
          pval <- # Your code here
    } else {
          pval <- # Your code here
  }
  return(pval)
}
```

Section 03 - Correlations and correlation tests

1. Investigate the correlation between `Sepal.Length` and `Sepal.Width` within the iris dataset. Calculate the correlation and plot your results. Are there any issues with your results? Discuss.

2. Repeat the previous analysis for each species independently. How does the correlation between `Sepal.Length` and `Sepal.Width` change?

Homework

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

Section 04 - Pitfalls when data deviates from the assumption of normality

1. The file **stats-pitfalls.csv** contains simulated data that will help us analyze how the t-test can fail. Load and visualize the data. Apply both the t-test and Wilcoxon test on it. What do you observe? Which test is the better choice here?

Hint: use stat_summary() or geom_vline() to add points/lines to the ggplot object.

2. In dataviz land, we want to know whether there is correlation between attendance to the exercise sessions and the points achieved in the final exam. We provide simulated data below. Load the data from **exam_correlation.tsv**. Calculate the correlation between attendance and points using `Pearson` and `Spearman` methods and visualize it. Some students will drop out of the distribution since they were planning to take the retake exam and skipped the first exam, thus obtaining a grade of zero. Which correlation method should be preferred in this context and why?

Section 05 - Let's do a test

1. Consider the dataset `mtcars`. Which statistical test that we studied do you suggest to test the association between the variable `cylinder > 4` and the variable `gear > 3`? Justify the choice of the test and provide the two-sided p-value rounded to two significant digits using `signif(...,digits=2)`

2. Assume that $\alpha = 0.05$ is our threshold of significance. What did we show in part (1) of this exercise?

3. If (1) had asked us to "test if there is a *positive* association between the variable `cylinder > 4` and the variable `gear > 3`", how would our answer change? What do we conclude?

4. If (1) had additionally specified "do not make any assumption of normality", how would our answer change?

Section 06 - Test the association between markers

In this exercise, we will explore the impact of genetic linkage. See https://www.khanacademy.org/science/ap-biology/heredity/non-mendelian-genetics/a/linkage-mapping for an introduction.

1. Make a function that given any two markers, returns the P-value of the appropriate statistical test to evaluate the association between the markers. Test your function for, e.g., `mrk_1` vs `mrk_13314`.

2. Test the association of every other marker with marker `mrk_1`.

3. Plot the P-values versus the genomic position of the associated marker. Use `-log10(pval)`, as this accentuates small P-values. What do you observe? (Hint: note that marker `mrk_1` 1 is located at the very beginning of Chromosome 1)

4. Plot the histogram of P-values, for (a) markers situated on chromosome 1 and (b) markers situated on any other chromosome.

5. Compute the fraction of P-values which are smaller or equal to 0.05, for (a) markers situated on chromosome 1 and (b) markers situated on any other chromosome. What do you observe?