Computational Statistics
Summer 2023
Exercise Sheet 2

Fabio Miranda - Fabio.MalcherMiranda@hpi.de
Hendrik Raetz - Hendrik.Raetz@hpi.de
Pauline Hiort - Pauline.Hiort@hpi.de

May 4th 2023

This week's exercise focuses on summarizing data in R / Python, visualization, and hypothesis testing. The problems 8 and 9 aim at computing and comparing univariate statistical tests.

Problem 7 - Visualization
Summarizing and visualizing data is a central part of any profound statistical analysis[1]. Since you have already built a reputation by enrolling in the CompStats class, a friend asks you to analyze the average scoring statistics of two players and provide a recommendation on which player he should buy for his fantasy statsball league. Download the corresponding data set fantasy.csv from moodle.

A Please compute the mean and standard deviation for the two players (player1_home, player2_home). Which player would you recommend to buy (higher scores are better)?

B Visualize the data using a histogram - compare the visualization with your computed summary statistics.

C Visualize the data using a boxplot - would you update your recommendation?

D Lastly, visually compare the relationship between the home and away scores for each player. Do you think visualizing the data is helpful?

[1] examples: `https://en.wikipedia.org/wiki/Exploratory_data_analysis`

Problem 8 - Testing by hand
Please use pen and paper to do a step-by-step computation for an 1) unpaired two-sample t-test (you may assume equal variance) and 2) a Mann-Whitney-test for comparing the following two populations (do not derive critical values by hand! Critical values for Mann-Whitney[2] and the t-test[3].

population$_1$: 1, 2, 3
population$_2$: 6, 9, 10

Problem 9
You had your thrill at Berghain, and now you just want a chill Saturday night. You are torn between watching Babylon Berlin and Dark. You ask a bunch of friends to rate the two shows on a scale of 1 to 5. The following are the gathered data.

Babylon: 3, 4.5, 5, 5, 3, 3, 3, 3, 4.5, 4.5, 4.5, 4, 3, 3,5
Dark: 2, 4.5, 3, 1, 2, 3, 5, 5, 5, 5, 3, 1, 1, 1, 1

You clearly see that Dark has a bunch of 1's and Babylon Berlin has no 1's! However, you remember that numbers can be deceiving. So you decide to run a statistical test to check the assumption that Babylon has higher ratings than Dark. Please compare the ratings of the two TV shows to determine which one to watch. State the central assumptions and perform the appropriate plots to justify which test to perform (t-test or Mann-Whiteney) at the significance level of 5%. We recommend using R commands: `t.test` , `wilcox.test`. The goal is to determine which TV show has the significantly higher rating.

Problem 10 - Multiple Testing (pen & paper)

Multiple testing describes the process of applying a statistical test multiple times. To realize the implications, please perform the following computations.

A If you conduct 100 hypothesis tests with $\alpha = 0.05$, how many significant results do you expect by chance if there is no difference in the data?

B Compute the probability of falsely rejecting at least one null hypothesis, if 20 independent tests are performed at $\alpha = 0.05$.

[2] http://www.real-statistics.com/statistics-tables/mann-whitney-table/
[3] http://www.stat.ufl.edu/~winner/tables/stat_tables.html

Problem 11 - Multiple Testing

You are now an expert on conducting hypothesis tests! This new achievement has also led you to become more fascinated with the consumption and the science behind beer. After leaving Berghain, you remember the example Professor Renard mentioned in class. The idea of saving Guinness time and money has suddenly consumed you. What if you could save Guinness time and money by recommending them to use the Mann-Whitney test instead of wasting time checking for t-test assumptions? Time is money, so let's design an experiment with many tests to explore the implications on false positives (type I error) and false negative (type II error) test results. Let's assume that most of Guiness' data is indeed normally distributed. Is the Mann-Whitney test as good as the t-test for finding true differences? **Hint:** In R, the following commands might become useful `rnorm`, `sample`, `replicate`.

Experimental Design:

- set $\alpha = 0.05$
- simulate three populations, each with $n = 1000$ observations. Use $\mu_1 = \mu_2 = 2$, $\mu_3 = 4$ and $\sigma_1 = \sigma_2 = \sigma_3 = 1$
- generate 4 different samples (see below) from the above described populations and sample 5 observations from each population 1000 times
- sample 1: draw 5 samples, 1000 times from $\mathcal{N}(\mu_1, \sigma_1)$
- sample 2: draw 5 samples, 1000 times from $\mathcal{N}(\mu_2, \sigma_2)$
- sample 3: draw 5 samples, 1000 times from $\mathcal{N}(\mu_3, \sigma_3)$
- sample 4: draw 5 samples, 500 times from $\mathcal{N}(\mu_2, \sigma_2)$ and 5 samples, 500 times from $\mathcal{N}(\mu_3, \sigma_3)$

Tasks:

A Compute two-sided t-tests and Mann-Whitney tests for the scenarios: *no difference* (sample1, sample2), *difference* (sample1, sample3), *mixed* (sample1, sample4) and store the p-values. Compare the number of significant test results and discuss your findings.

B Please, visualize the p-values with histograms and discuss the distributions of p-values. Did you expect the outcome for the t-tests between the *no difference set-up* (sample1, sample2)?

C Use the simulation ground truth data (no difference, difference, mix) to compute the type I and type II error for both tests (most interesting for the mix scenario). Would you recommend Guinness to ban the t-test from their analysis?

D Lastly, apply Bonferroni correction by adjusting the α level for your tests. How many significant test results do you get? Discuss the application of Bonferroni correction and its implications for type I and type II errors in a scenario with many tests.

These exciting real world problems are not graded, however, we believe these will enrich your experience in the Computational Statistics course and therefore, it is strongly recommended that you attempt to complete these problems. We will discuss the problems on Monday, **May 15, 2023**. Each student is *required* to present at least one solution during the semester. To ensure an interactive exercise experience, solutions will be presented each week by either volunteers or randomly generated volunteers via R.