ADS2 Practical 4: Simulating Sample Data

Chaochen Wang

Semester 1, 2019/20

Work through this guide alone or in groups. Facilitators are here to help. The time it takes to complete this practical can vary between individuals - this is OK. Do not worry if you do not finish within the session.

Learning Objectives

- Explain reasons for using synthetic datasets, including ethical reasons
- Create synthetic data sets in R
- Use synthetic data sets to test a data analysis workflow

Blood glucose test

Let us assume that a company recently developed a new device to quickly test the concentration of blood glucose level in human. To examine the performance of the device, the company is planning to recruit 10,000 volunteers and measure their blood glucose levels with the new device (ND) as well as in a traditional way (TW). Before the recruitment, you are asked to simulate the data and prepare a pipeline to analyze the data.

1. Please simulate two datasets, each with three variables: volunteer number, glucose test results using ND and glucose test results using TW. How would you simulate the data? Please try to simulate two datasets with 100 people and 10,000 people. In the first dataset, set the means and standard derivations to 143,48 in both methods; in the second one, set the means and standard derivations to 143, 48 and 156, 55 respectively. Check the means and standard derivations.

2. When the data are collected, we want to know if there is significant difference between the two methods on examining blood glucose test. Please do a t-test on the data, pay attention to the parameter "paired" in your test. Perform the test on both datasets. If you sample 100 people out of the second dataset, what is the result now, why? Repeat sampling 5 times. You will find the results are different. Try set.seed() command and repeat sampling again.

3. Let us assume that we have collected the data and want to further take advantage of the data to see if age or gender is a factor that affects the level of blood glucose. Please put more information to the second dataset, assign each person an age (in the range of 18-80) and a gender. You may find cbind command useful.

   Afterwards, please assess the influence of age and gender on glucose level. You can use plot, boxplot, t-test etc. You may use formula expression in your t-test.

   Option: Try ggplot2 package to generate boxplots and plots.

   What would you expect in your plot if age or gender plays a role in glucose metabolism? How would you simulate the data if so.

4. Now you can export your first dataset to a csv file. You can use write.csv command to do it. Then swap your dataset with your partner's and test your pipeline in 3 again.