SCHOOL OF ELECTRICAL ENGINEERING AND COMPUTING

# BIG DATA AND DATA ANALYTICS
# LAB PROJECT 1

This lab project is based on a dataset from the National Institute of Diabetes and Digestive and Kidney Disease, which is available from the UCI Machine Learning Repository (Lichman, 2013):

https://archive.ics.uci.edu/ml/datasets/Pima+Indians+Diabetes

## EXERCISE 1 (1 MARKS) *[R-CODE]*

Use R to load the dataset "pimadata.csv". Determine the number of lines and columns in the dataset. Safe these values into two separate variables called "numberoflines" and "numberofcolumns". Use the *cat*-command to display these two variables on the R console.

## EXERCISE 2 (1 MARKS) *[R-CODE]*

Use R to compute the mean BMI value (variable: BMI) for subjects with diabetes. Then, use R to compute the mean BMI value (variable: BMI) for subjects without diabetes. Display the difference in mean BMI between these two groups on the R console using the *cat*-command.

## EXERCISE 3 (2 MARKS) *[R-CODE]*

Use R to determine the standard deviation and the variance of the variable TSFT for subjects with diabetes. Display these values on the screen using the *cat*-command. Note: Only take the observations of the subjects with TSFT greater than zero into account. How can a TSFT value of 0 be interpreted?

## EXERCISE 4 (1 MARK)

In your own words, describe the difference between a standard error (of the mean) and a standard deviation using the variable TSFT in the "pimadata.csv" dataset as an example.

## EXERCISE 5 (2 MARKS) *[R-CODE]*

Based on the dataset "pimadata", create a new column "agecat" in the dataframe that describes the age category of a person. Distinguish between the following categories: "21 to 35", "36 to 55", and "56 to 85". Convert the column into a factor variable using the as.factor() command. For each of these age categories, calculate the median BMI (i.e., the median BMI for subjects aged 21 to 35, the median BMI for subjects aged 36 to 55, and the median BMI for subjects aged 56 to 85).

Use R to determine the median BMI across the three different categories “21 to 35”, “36 to 55”, and “56 to 85”) and combine them into a vector using the *c()*-command. Note: Please only take the observations into account where the BMI is greater than zero. Safe the vector into a variable called “medianBMIs”.

Then, use the *cat*-command to display the minimum median BMI (i.e., the lowest of the three median BMIs) and the maximum median BMI (i.e., the highest of the three median BMIs) on the R console. The minimum and maximum values should be determined based on the newly created variable “medianBMIs”.

## EXERCISE 6 (1 MARK) *[R-CODE]*

Use if() to compare the median BMIs in the “21 to 35” and the “36 to 55” setting and display a textual statement on the screen that describes which of the two median BMIs is higher.

## EXERCISE 7 (2 MARKS) *[R-CODE]*

Write a function that determines the 99% CI of the mean for a given vector *x*. Call this function “calc99CI”. The function should return a vector of two values: (i) the lower bound of the 99% CI and (ii) the upper bound of the 99% CI. Use the *c*() to combine the two values into a vector.

Use this function to display the 99% CI of the mean for the BMI for subjects aged “36 to 55”. Note: Please only take the observations into account where the BMI is greater than zero.

[Important: CI stands for confidence interval. In the lecture, we discussed the multiplier for the 95% CI, which is 1.96. The multiplier for the 99% CI is 2.575]

## REFERENCES

Lichman, M. (2013). UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science.

## DATASET

| Pimadata | *Pima Indians Diabetes Database* |
|---|---|

**Description**

A diabetes dataset. All patients here are females at least 21 years old of Pima Indian heritage. Note: Even though the dataset donors made no such statement, it seems very likely that several values zero values encode missing data for several variables.

**Usage**

Pimadata

**Format**

A data frame with 768 observations on the following 9 variables.

| | |
|---|---|
| `timesPregnant` | Number of times pregnant |
| `PCG` | Plasma glucose concentration a 2 hours in an oral glucose tolerance test |
| `DBP` | Diastolic blood pressure (mm Hg) |
| `TSFT` | Triceps skin fold thickness (mm) |
| `insulin` | 2-Hour serum insulin (mu U/ml) |
| `BMI` | Body mass index (weight in kg/(height in m)^2) |
| `DPF` | Diabetes pedigree function. It provides some data on diabetes mellitus history in relatives and the genetic relationship of those relatives to the patient. This measure of genetic influence gives an idea of the hereditary risk one might have with the onset of diabetes mellitus. |
| `age` | Age (Years) |
| `diabetes` | 1 tested positive for diabetes<br>0 tested negative for diabetes |

**Source**

Lichman, M. (2013). UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science.