Course of Study Bachelor Computer Science	Exercises Statistics WS 2023/24
Sheet I - Solutions	

1 Descriptive Statistics - Variables

1. Are the following variables qualitative or quantitative?

   (a) Body height
   (b) Hair color
   (c) Temperature in Celsius
   (d) Temperature in Kelvin
   (e) Number of bottles of wine in a student's flat
   (f) Birthday

   Answer:

   (a) Quantitative
   (b) Qualitative
   (c) Quantitative
   (d) Quantitative
   (e) Quantitative
   (f) Quantitative

2. Which scales should be used for the following variables?

   (a) Body height
   (b) Hair color
   (c) Temperature in Celsius
   (d) Temperature in Kelvin
   (e) Number of bottles of wine in a student's flat
   (f) Birthday

   Answer:

   (a) Ratio

(b) Nominal

(c) Interval

(d) Ratio

(e) Ratio

(f) Interval

3. It is possible to transform a variable "downwards", from a scale with more information contained, to a scale with less information contained. Give an example for the variable *Price for a bottle of wine* for the transformation from a ratio to an ordinal scale.

   Answer:

   - Ratio scale: Price measured in euro.
   - Ordinal scale: Price mesured as cheap, normal, expensive.

4. Is it possible to transform a variable "upwards", from a scale with less information contained, to a scale with more information contained? Give an example (showing if it is possible or not)!

   Answer: No, you cannot gain more information about a variable by transforming the variable from one scale to another. Example: Assume that the variable *Price for a bottle of wine* originally has an ordinal scale. This means that you know for each bottle if it is cheap, normal or expensive. Given **only** this information for each bottle, do you know the price in euro (ratio scale) for the bottles? No!

5. Consider the question of describing students attitudes towards to legalisation of marijuana, as well as what proportion of them wants to legalize the drug and whether this proportion differs by gender and age.

   (a) Which data collection method is most suitable here: survey or experiment?

   (b) How could you capture the attitudes towards legalisation in a single variable?

   (c) Which variables are needed to answer the questions? Describe the type and the scale of the variables.

   (d) How would an appropriate data set look? Try to describe the question in more details.

   Answer:

(a) Survey: The information would be obtained via a questionaire given to a sample of students.

(b) There are different options to find out students attitudes:

- simply ask: "What do you think about legalisation?"
  Problem: Capturing long answers in a variable attitutde may make it difficult to summarize and distil the information obtained.
- Common way: translate it into a score
  One could for example ask 5 "yes/no" questions which relate to attitudes towards legalisation like "Do you believe that legalisation would endanger the health of young people?", "Do you think legalization would encourage the entry into harder drugs?", ... The number of answers showing a positive attitude can be summed up. Thus the answers of each student can be summarized on a scale from 0 to 5.

(c) Needed variables are:

- Attitude: quantitative variable, ordinal scale
- Legalise: binary ("yes/no") variable capturing wether the student agrees to legalize Marihuana. This is qualititaive variable with nominal scale.
- Gender: qualititaive variable with nominal scale
- Age: quantitative (continous) variable with ratio scale.

(d) A data set might look as:

Student	Attitude	Legalize	Gender	Age
1	3	no	male	22
2	2	yes	female	25
⋮	⋮	⋮	⋮	⋮

Attitude refers to variables capturing attitudes towards legalisiation and "Legalise" is the score variable summarizing these questions.
More detailled questions:

- What is the median rate towards legalisation among students and how much does ist vary?
- What percentage of students answer "yes" when asked to legalise Marihuana?
- What is difference in the proportion calculated above when stratified by gender?

- What is average of those students who support the legalisation compared with the average age of those students who do not support the legalisation?

Introduction to R and RStudio

Some useful hints for the first steps

- Open a new script file via File→New→R Script.
- Save the script file at any time via File→Save.
- Comments to the code that are not evaluated can be made with the # icon.
- Send R code to the R console:
  - Click with the mouse on a line (any location). Then click on the Run-button. Only the selected line will be sent to the R console. The cursor will automatically move to the next line. You can now click again to send this line to the R console, and so on.
  - Mark the code you want to send to the R console. Then click the Run-button. So all the marked text will be sent to the R console.
  - $< Ctrl > + < Enter >$ on your keyboard instead of prerssing the Run-button

Useful shortcuts

- Assignment arrow $< -$: $< Alt > +-$
- Complete code:Tab key
- Comment in/out marked region:$< Ctrl > + < Shift > +C$
- Delete R console: $< Ctrl > +L$
- Cancel evaluation (if it takes too long): click or press $< Esc >$ in the R console.
- In the R console, retrieve previous code: Arrow keys (up and down).
- Switch to the editor with the cursor: $< Ctrl > +1$

- Move the cursor to the R console: $< Ctrl > +2$
- Save: $< Ctrl > +S$

Start now RStudio, open a new script file and solve the following tasks.

1. Calculate the following quantities:
   - the sum of 52.3, 74.8, 3.17
   - the square root of 144
   - the 10-based logarithm of 200 multiplied with sin of $\pi/4$
   - the cumulative sum of the numbers 1,3,18,20,2 (use the cumsum() command)
   - find 10 numbers between 0 and 20 rounded to the nearest integer value (hint use the command sample() or a combination of the commands round() and runif()).

   Hint: If you do not know command() use the ?command.

2. Assigning Variables
   - Assign the number 5 to x and the number 10 to y.
   - Calculate the product of x and y.
   - Store the result in a new variable z.
   - Inspect your workspace by clicking the "environment" tab in RStudio, and find the three objects.
   - Make a vector myvec of the objects x,y,z.
   - Find the minimum, the maximum and the mean of the vector.
   - Remove myvec from the workspace.

3. The numbers below are the first ten days of rainfall in a year

   ```
   0.1  0.5  2.3  1.1  11.3  14.7 23.4  15.7  0  0.9
   ```

   - Read them into a vector using the c() command.
   - Calulate the mean and the standard deviation.
   - Calculate the cumulative rainfall over these ten days. What is total sum of the rainfall?

- Which day saw the highest rainfall? Find an appropriate R command.
- Take a subset of the rainfall data where rain is larger than 10.
- What is mean rainfall for days where the rainfall was at least 5?
- Subset the vector where it is either exactly 0 or 1.1 and find the corresponding days.

4. The length of five cylinders are 2.5, 3.4, 4.8, 3.1, 1.7 and their diameters are 0.7, 0.4, 0.5, 0.5, 0.9.

   - Read these vectors into two vectors with appropriate names.
   - Calculate the volumes of each cylinder and store it in a new vector.
   - Assume the values are given in centimeter. Recalculate the volumes so that their units are cubic millimeter.

5. Inspect the R commands union(), setdiff() and intersect() implying set operations. Make two vectors

```
x <- c(1,2,3,4,5)
y <- c(3,5,7,9)
```

   - Find values that are contained in both x and y.
   - Find values that are in x but not y and vice versa.
   - Construct a vector that contains all values contained in either x or y. Compare the result with c(x,y).

6. Construct a matrix with 8 rows and 10 columns. The first row should contain the numbers 0, 2, 4, ..., 18 and the other rows should random integer numbers between 0 and 10. Use runif() to create the random numbers and as.integer() to transform to integers.

   - Calculate the row means of this matrix (use rowMeans()) and the standard deviation across the row means.
   - Store the rows 2,3,..,8 in a other matrix and calculate the column means (use colMeans()). Use the command hist() to create a histogram of the column means.

7. The R dataset mpg

   (a) Inspect the dataset mpg.

(b) Determine the types and the scales of measurement of all variables in the dataset mpg. Further more determine whether the variables are discret or continous.

Hint: The dataset mpg is part of the package ggplot2 and tibbles are part of the tidyverse package.

Answer:

variable	type	level	discrete/continous
manufacturer	qualitative	nominal	discrete
model	qualitative	nominal	discrete
displ	quantitative	ratio	continous
year	quantitative	interval	discrete
cyl	quantitative	ratio	discrete
trans	qualitative	nominal	discrete
drv	qualitative	nominal	discrete
cty	quantitative	ratio	continous
hwy	quantitative	ratio	continous
fl	qualitative	nominal	discrete
class	qualitative	nominal	discrete

8. Lists

(a) Create a list containing the string “John”, the string “Mary” and the vector (4,6,10). The list describes a family with names of father, mother and the ages of the children.

(b) Create a second list containing the names Bob, Cate and Susan of the children.

(c) Concatenate the lists. Use the c() and the list() function. What are the differences?

(d) Access the concatenated list to create directly a list containing for every child name and age.

```
# Sheet 1: Introduction to R, RStudio
# WS 2022/23

# Task 1:  Calculate the following quantities:
# sum of 52.3, 74.8, 3.17
52.3+74.8+3.17
# the square root of 144
144**0.5
# the 10-based logarithm of 200 multiplied with sin of $\pi/4$
log10(200)*sin(pi/4)
# the cumulative sum of the numbers 1,3,18,20,2
cumsum(c(1,3,18,20,2))
# find 10 numbers between 0 and 20 rounded to the nearest
sample(x = 0:20, size = 10, replace = FALSE)
# or
round(runif(n = 10, min = 0, max = 20))

# Task 2: Assigning Variables
```

```
# Assign the number 5 to x and the number 10 to y.
x <- 5
y <- 10
# Calculate the product of x and y.
x * y
# Store the result in a new variable z.
z <- x * y
# Inspect your workspace by clicking the ''environment'' tab in RStudio, and find the three objects.

# Make a vector myvec of the objects x,y,z.
myvec <- c(x,y,z)
# Find the minimum, the maximum and the mean of the vector.
min(myvec)
max(myvec)
mean(myvec)
# Remove myvec from the workspace.
rm(myvec)

# Task 3 rainfall in the first ten days in a year
# Read them into a vector using the c() command.
rainfall <- c(0.1,0.5,2.3,1.1,11.3,14.7,23.4,15.7,0,0.9)
# Calulate the mean and the standard deviation.
mean(rainfall)
sd(rainfall)
# Calculate the cumulative rainfall over these ten days. What is total sum of the rainfall?
cumsum(rainfall)
sum(rainfall)
# Which day saw the highest rainfall? Find an appropriate R command.
which.max(rainfall)
which(rainfall == max(rainfall))
# Take a subset of the rainfall data where rain is larger than 10.
rainfall[rainfall >10]
# What is mean rainfall for days where the rainfall was at least 5?
mean(rainfall[rainfall >=5])
# Subset the vector where it is either exactly 0 or 1.1 and find the corresponding
# days where the rainfall is 0 or 1.1
rainfall[rainfall == 0 | rainfall == 1.1] # alternative solution: rainfall[rainfall %in% c(0,1.1)]
which(rainfall %in% c(0,1.1))

# Task 4: The length of five cylinders and their diameters
# Read these vectors into two vectors with appropriate names.
len <- c(2.5, 3.4, 4.8, 3.1, 1.7)
diam <- c(0.7, 0.4, 0.5, 0.5, 0.9)
# Calculate the volumes of each cylinder and store it in a new vector.
vol <- len * (0.5*diam)**2 * pi
vol
# Assume the values are given in centimeter. Recalculate the volumes so that
# their units are cubic millimeter.
vol.cm <- 10*len * (10*diam/2)**2 * pi
vol.cm

# Task 5: Inspect the R commands union(), setdiff() and intersect()
x <- c(1,2,3,4,5)
y <- c(3,5,7,9)
# Find values that are contained in both x and y.
intersect(x,y)
# Find values that are in x but not y and vice versa.
setdiff(x,y) # x without y
setdiff(y,x) # y without x
# Construct a vector that contains all values contained in either x
# or y. Compare the result with c(x,y).
union(x,y)
# c(x,y) only concatenates x and y
c(x,y)

# Task 6 Construct a matrix with 8 rows and 10 columns. The first row should
# contain the numbers 0, 2, 4, ..., 18 and the other rows should random
# integer numbers between 0 and 10.
mat1 <- matrix(c(seq(0,18, by = 2),
                 as.integer(runif(70,0,1000))),
               nrow = 8, ncol = 10, byrow = TRUE)
mat1
# Calculate the row means of this matrix (use rowMeans()) and the standard
# deviation across the row means.
rm <- rowMeans(mat1)
rm
sd(rm)
# Store the rows 2,3,..,8 in a other matrix and calculate the column means.
# Use the command hist() to create a histogram of the column means.
# removing the first row of mat1
mat2 <- mat1[-1,]
mat2
# colum means of mat2
```

```
cm <- colMeans(mat2)
cm
# creating a histogram of cm
hist(cm)

# Task 7 The R dataset mpg
# Inspect the dataset mpg.
library(ggplot2)
library(tidyverse)
help(mpg)
names(mpg)
head(mpg)
# Create an empty tibble str_mpg with variables name, type,
# level and dc of type character().
str_mpg <- tibble(name = character(), type = character(), level = character(), dc = character())
# Add for every variable in the dataset mpg a row containing for every variable the
# name, the type, the level of measurement and discrete/continous.
str_mpg <- str_mpg %>%
  add_row(name = "manufacturer", type = "qualitative", level = "nominal", dc = "discrete") %>%
  add_row(name = "model", type = "qualitative", level = "nominal", dc = "discrete") %>%
  add_row(name = "displ", type = "quantitative", level = "ratio", dc = "continous") %>%
  add_row(name = "year", type = "quantitative", level = "interval", dc = "discrete") %>%
  add_row(name = "cyl", type = "quantitative", level = "ratio", dc = "discrete") %>%
  add_row(name = "trans", type = "qualitative",level = "nominal", dc = "discrete") %>%
  add_row(name = "drv", type = "qualitative", level = "nominal", dc = "discrete") %>%
  add_row(name = "cty", type = "quantitative", level = "ratio", dc = "continous") %>%
  add_row(name = "hwy", type = "quantitative", level = "ratio", dc = "continous") %>%
  add_row(name = "fl", type = "qualitative", level = "nominal", dc = "discrete") %>%
  add_row(name = "class", type = "qualitative", level = "nominal", dc = "discrete")
head(str_mpg)
# Display the structure
str(str_mpg)
# Use the tibble to display all variables which are quantitative and discrete
# applying the R function subset().
subset(str_mpg, subset = (type == "quantitative" & dc == "discrete"))
# alternative solution
str_mpg %>% filter(type == "quantitative", dc == "discrete")

# Task 8
# a) Create a list containing the string "John", the string "Mary" and the vector
#    (4,6,10). The list describes a family with names of father, mother and the ages
#    of the children.
family <- list(father = "John", mother = "Mary", child.ages = c(4,6,10))
# b) Create a second list containing the names Bob, Cate and Susan of the children.
children.names <- list("Bob", "Cate", "Susan")
# c) Concatenate the lists. Use the c() and the list() function. What are the differences?
family.1 <- c(family, children.names)     # 3 strings are added to the list family
family.2 <- list(family, children.names)  # Another list has been added
# d) Access the concatenated list to create directly a list containing for every
#    child name and age.
children.1 <- list(child1 = c(family.1[[4]], family.1[[3]][1]),
                   child2 = c(family.1[[5]], family.1[[3]][2]),
                   child3 = c(family.1[[6]], family.1[[3]][3]))
children.2 <- list(child1 = c(family.2[[2]][1], family.2[[1]][[3]][1]),
                   child2 = c(family.2[[2]][2], family.2[[1]][[3]][2]),
                   child3 = c(family.2[[2]][3], family.2[[1]][[3]][3]))
```