Week3

July 2025

## BACKGROUND

In this activity you will consider the usefulness of some suggested R code, identify data issues, and produce descriptive statistics.

A researcher is interested in the heights and weights of a population of adults, and whether these differ according to whether the adult has an active lifestyle or not and whether adults are left or right-handed. In general, they are interested in any interesting findings that can be found by exploring the dataset.

## SUPPORT

There is a supporting video that introduces this exercise, that uses the same data structure but different sample values. This may help you understand the context.

[Watch this video on YouTube](https://www.youtube.com/watch?v=Rkcm9FRFTB8)

## DATA STRUCTURE

Load the dataset here! Call the dataset ‘content’.

content <- read.csv("../Data/weight\_height.csv")

The variables collected for a number of adults are: weight – in kilograms height – in centimetres handed - if adult is ‘left’ or right’ handed active - is adult considered to be active? ‘yes’ or ‘no’

## Helpful Code

Some useful code is given below. Try it out and see what it does.

# Summary of numeric columns  
summary(content[, sapply(content, is.numeric)], na.rm = TRUE)

## weight height   
## Min. :11.00 Min. : 0.0   
## 1st Qu.:77.25 1st Qu.: 175.6   
## Median :79.82 Median : 178.5   
## Mean :79.68 Mean : 179.8   
## 3rd Qu.:82.21 3rd Qu.: 181.6   
## Max. :92.89 Max. :1792.4

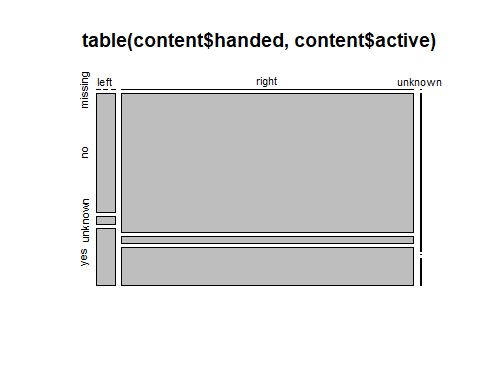
# Frequency tables  
table(content$handed)

##   
## left right unknown   
## 85 1320 7

table(content$handed, content$active)

##   
## missing no unknown yes  
## left 0 55 4 26  
## right 2 997 51 270  
## unknown 0 6 0 1

# Mosaic plot  
mosaicplot(table(content$handed, content$active))



# Summary statistics by handedness  
by(content, factor(content$handed), summary)

## factor(content$handed): left  
## weight height active handed   
## Min. :71.10 Min. :168.1 Length:85 Length:85   
## 1st Qu.:76.91 1st Qu.:176.3 Class :character Class :character   
## Median :79.14 Median :178.5 Mode :character Mode :character   
## Mean :79.36 Mean :178.3   
## 3rd Qu.:82.00 3rd Qu.:181.2   
## Max. :92.49 Max. :186.4   
## ------------------------------------------------------------   
## factor(content$handed): right  
## weight height active handed   
## Min. :11.00 Min. : 0.0 Length:1320 Length:1320   
## 1st Qu.:77.28 1st Qu.: 175.6 Class :character Class :character   
## Median :79.86 Median : 178.6 Mode :character Mode :character   
## Mean :79.70 Mean : 180.0   
## 3rd Qu.:82.22 3rd Qu.: 181.6   
## Max. :92.89 Max. :1792.4   
## ------------------------------------------------------------   
## factor(content$handed): unknown  
## weight height active handed   
## Min. :76.78 Min. : 2.0 Length:7 Length:7   
## 1st Qu.:77.19 1st Qu.:170.0 Class :character Class :character   
## Median :80.21 Median :173.1 Mode :character Mode :character   
## Mean :79.88 Mean :149.0   
## 3rd Qu.:81.94 3rd Qu.:175.6   
## Max. :83.91 Max. :176.8

# Proportions of handedness  
prop.table(table(content$handed))

##   
## left right unknown   
## 0.060198300 0.934844193 0.004957507

# Replace "unknown" with NA in 'handed'  
content$handed[content$handed == 'unknown'] <- NA  
  
# Replace extreme weight values with NA  
content$weight[content$weight > 250] <- NA  
  
# Count of non-missing weight values  
length(na.omit(content$weight))

## [1] 1412

# Aggregated height statistics by handedness and activity  
agg<- aggregate(  
 height ~ handed + active,  
 data = content,  
 FUN = function(x) c(mean = mean(x), sd = sd(x))  
)

## Preliminary exploratory data analyses.

Adapting the above code, or otherwise, prepare some summary statistics for the attached data.

Consider what this reveals about any issues with the raw data, and further data preparation that may be required prior to statistical analyses. List those issues here.

## Data Cleaning

Apply any further data preparation that may be required prior to statistical analyses, to ensure that only valid observations will be included in future analyses.

## Exploratory data analyses using only the valid observations.

Produce summary statistics using only the valid observations. You may wish to consider for example:

How many adults in the sample are active?

What is the best estimate for the proportion of adults in the population that are active?

How many active adults in the sample are right-handed?

How many valid observations are there for weight?

What is the sample minimum weight?

What is the sample maximum weight?

What is the sample mean weight?

What is the sample standard deviation for weight?

What is the sample minimum weight of the active adults?

What is the sample maximum weight of the active adults?

What is the sample mean weight of the active adults?

What is the sample standard deviation weight of the active adults?

## Conclusions

Offer some interpretation. Some examples of research questions you may want to answer are:

How do active adults compare to non-active adults in terms of weight and height?

How does the average and variability of weight and height differ by handedness?