Practical Exercises for **Day 3**

Sonja Hartnack, Terence Odoch & Muriel Buri

October 2017

Exercise 8

Apply the summary statistics to the `perulung_ems` and `ToothGrowth` data set.

Exercise 9A: Plausibility Checks

- What can go wrong?
- Identify different strategies for spotting these potential errors.
  - Logical errors
  - Spelling mistakes
- Import the data set `bacteria_plausibility_check.csv` to R.
- Detect the **six** errors in the imported data set `bacteria_plausibility_check.csv` in R.
- Find possible solutions in R how to handle these challenges.
- Do all variables have the correct data type (numeric, integer, factor)? - If not, do correct / define them.

Exercise 9B: Missing Values

- Check out the difference between the different missing values

```r
y1 <- c(2, 4, 3, NA, 6, 1)
y2 <- c("diseased", "healthy", NA, "NA")
y3 <- c(1, "NA", 0, 1, NaN)

is.na(y1)
which(is.na(y1))
is.na(y2)
```

```r
which(is.na(y2))
is.na(y3)
which(is.na(y3))
is.nan(y3)
```

- Create a vector with missing values and determine the mean and median

```r
myvector <- c(1:3,NA,NA,1:3)
mean(myvector)
mean(myvector,na.rm=TRUE) # calculates c(1, 2, 3, 1, 2, 3)
median(myvector,na.rm=TRUE)
```

- If `x = c (22,3,7,NA,NA,67)` what will be the output for the R statement `length(x)`?

```r
x <- c (22,3,7,NA,NA,67)
length(x)
```

- If `x = c(NA,3,14,NA,33,17,NA,41)` which line of R code removes all occurrences of NA in x.

```r
x <- c(NA,3,14,NA,33,17,NA,41)
x[!is.na(x)]
x[is.na(x)]
x[which(is.na(x))] <- 0
```

- If `y = c(1,3,12,NA,33,7,NA,21)` what R statement will replace all occurrences of NA with 11?

```r
y <- c(1,3,12,NA,33,7,NA,21)
y[y=="NA"] <- 11
y[is.na(y)] <- 11
y[y==11] <- NA
```

- If `x = c(34,33,65,37,89,NA,43,NA,11,NA,23,NA)` then what will count the number of occurrences of NA in `x`?

```r
x <- c(34,33,65,37,89,NA,43,NA,11,NA,23,NA)
sum(x=="NA")
sum(x == "NA", is.na(x))
sum(is.na(x))
```

- Create a vector and find the number of missing values and their position

```
x1 <- c(rnorm(10,5,2),NA,5:12,NA,6,7.5,NA)
is.na(x1)
summary(x1)
sum(is.na(x1))
which(is.na(x1))
```

- Now, create the vector x2 and assess the difference to x1

```
x2 <- c(rnorm(10,5,2),NA,5:12,NA,6,7.5,NA,log(-2))
x2
```

- What is the meaning of "NA" versus "NaN"?
- Replace the missing values in x1 with a 0, and check that no NAs are present try two different commands to coerce the NAs into 0

```
x1[is.na(x1)] <- 0
is.na(x1)
# or
ifelse(is.na(x1),0,x1)
```

Exercise 10

- Import the data set `water_errors.csv` to R: A data frame with 61 observations on the following 6 variables.
  - **location**: a factor with levels `North` and `South` indicating whether the town is as north as Derby.
  - **town**: the name of the town.
  - **mortality**: averaged annual mortality per 100.000 male inhabitants.
  - **hardness**: calcium concentration (in parts per million).
  - **smoker**: If there are any smokers living in town.
  - **num.of.cig**: In case, smokers live in town, what number of cigarettes do they smoke per day.
- Detect the errors in the imported data set `water_errors.csv` in R.
- Find possible solutions in R how to handle these challenges.
- Do all variables have the correct data type (numeric, integer, factor)? - If not, do correct / define them.