Statistical Computing

Master Data Science

Winter Semester 2017/18

BEUTH HOCHSCHULE FÜR TECHNIK BERLIN
University of Applied Sciences

Prof. Tim Downie

Workshop 6
Sampling, Measures of Location and of Dispersion

Section 2 includes exercises to do before next week's workshop.

1 In the Workshop

1.1 Standard Deviation and Variance in R

Last week you worked through an exercise using the Munich Accommodation Data. You will be using the same data this week.

(a) Read in the data as you did last week.

(b) Look at the variables `rent` in more detail.

```
> hist(Accom$rent)
> summary(Accom$rent)
```

(c) Use this information to calculate the range and inter quartile range for the rents.

(d) Use the example in the lecture notes to calculate the variance and standard deviation of `rent` in `R` *without* using the functions `var()` and `sd()`.

(e) Calculate the interval: $[\overline{x} - 2s_x; \overline{x} + 2s_x]$

(f) How many apartments have a rent within this interval, and what proportion of all apartments is this? Does this fit with the rule of thumb?

(g) Repeat the command from last week Use the following command to define a new variable called `Old`. The variable is a factor variable with `"yes"` if the build year is before 1950, and `"no"` otherwise.

```
> Accom$Old<-factor(Accom$Year<1950,labels=c("no","yes"))
```

(h) Use tapply to find the variance for old and new apartments. Which has the larger variance? Obtain a box-plot for `rent` in these two groups.

```
> boxplot(???~???,data=???)
```

(i) The length of each box in the plot indicates the interquartile mean for that group. Does this fit with your answer to part (g)?

1.2 Rescaling and the effect on mean, median, sd, variance and IQR.

The variable `Area` in the `Accom` data frame contains the official floor area of each apartment in square meters.
Suppose you are working for an American company, who usually deals with square feet. There are 10.76 square feet in a square meter.

(a) Create a variable called `Area_ft2` which contains the floor areas in square feet.

(b) Obtain the mean of `Area_ft2` and `Area` and calculate the Ration of the two. What do you notice about this ratio?

(c) Repeat this for the medians, standard deviations, variances and IQR?

To summarise: if the variable Y is a **linear transformation** of variable X, which means there relationship can be written using formula

$$y_i = ax_i + b,$$

with a and b constants, then the statistics can be directly transformed using the following:

Mean	$\overline{y} = a\overline{x} + b$
Median	$y_{0.5} = ax_{0.5} + b$
Standard deviation	$s_y = as_x$
Variance	$s_y^2 = a^2 s_x^2$
IQR	$y_{0.75} - y_{0.25} = ax_{0.75} + b - (y_{0.25} + b) = a(x_{0.75} - x_{0.25})$

1.3 Median

Write a few lines of R code to calculate the median of a numeric vector `x`.
Do not use the function `median()` except to check your result.
Try out your code using

```
> x<-Accom$rent[1:20]
```

Make sure your code works for odd and even sample sizes.

Tips:

1. To check if a value `number` is even use

   ```
   > (number %% 2) == 0
   ```

   The `%%` operator means *modulo* which means the remainder after integer division.

   $27/5 = 5.7$. With integer division the answer is 5 with a remainder of 2

   ```
   > 27%/%5
   [1] 5
   > 27%%5
   [1] 2
   ```

2. `> answer<-ifelse(lg,fun1(x),fun2(x))`
   If `lg` is `TRUE` return `fun1(x)`. If `FALSE` return `fun2(x)`.

   This also works if `lg` is a vector: `x` should also be a vector with the same length as `lg`. The if-else decision is run on each element of the `lg` and `answer` is also a vector of the same length.

2 Exercises to do at home.

Exercise 1: Variance

Three numeric variables a, b and c are given in the following table.

Variable	Data
a	1.0; 1.1; 1.2; 1.2; 1.3; 1.4; 1.5
b	20; 22; 25; 28; 30; 35; 40
c	-50; 20; 134; 219; 298; 504; 780; 1293

The three variances in random order are 50.62, 201900 and 0.02952.

Without using `R` assign the three variances to the three variables

Exercise 2: Correlation

For the following two (small!) samples x and y calculate the mean, variance and standard deviation of x and y and the covariance and korrelation coefficiente for x and y.

X	Y
10	20
60	140
70	130
20	30
30	60

You might find it helpful to compete the following table in order to calculate the statistics.

x	y	$(x_i - \overline{x})$	$(x_i - \overline{x})^2$	$(y_i - \overline{y})$	$(y_i - \overline{y})^2$	$(x_i - \overline{x})(y_i - \overline{y})$
10	20					
60	140					
70	130					
20	30					
30	60					
190	380	0		0		

Exercise 2: Linear Transformation

Monthly Temperature in Celsius data for Berlin in 2013 can be summarised with the following statistics.

```
> mean(temp)
[1] 9.85
> median(temp)
[1] 10.45
> sd(temp)
[1] 7.783607
> var(temp)
[1] 60.58455
> IQR(temp)
[1] 11.85
```

To convert the a temperature from Celsius (C) into Fahrenheit F use the formula:

$$F = 1.8C + 32$$

Calculate the above statistics in Fahrenheit