Part 3. Numerical variables

Natalia Levshina © 2017

University of Mainz, June 2017

Outline

1. Measures of central tendency
2. Measures of dispersion
3. Graphical representations
4. Testing normality

Data frame ldt

```
> library(Rling) # loads a package that has been installed
> data(ldt) # loads the data
> head(ldt) # returns the first 6 rows
```

	Length	Freq	Mean_RT
marveled	8	131	819.19
persuaders	10	82	977.63
midmost	7	0	908.22
crutch	6	592	766.30
resuspension	12	2	1125.42
efflorescent	12	9	948.33

Data frame structure

```
> str(ldt) # displays the structure
'data.frame': 100 obs. of 3 variables:
$ Length : int 8 10 7 6 12 12 3 11 11 5 ...
$ Freq : int 131 82 0 592 2 9 14013 15 48 290 ...
$ Mean_RT: num 819 978 908 766 1125 ...
```

Attach a data frame

```
> head(ldt$Length)
[1]  8 10  7  6 12 12
> head(Length)
Error in head(Length) : object 'Length' not found
> attach(ldt)
> head(Length) #now you can access all variables directly
[1]  8 10  7  6 12 12
```

Mean, median and mode

> mean(Length)

[1] 8.23

> median(Length)

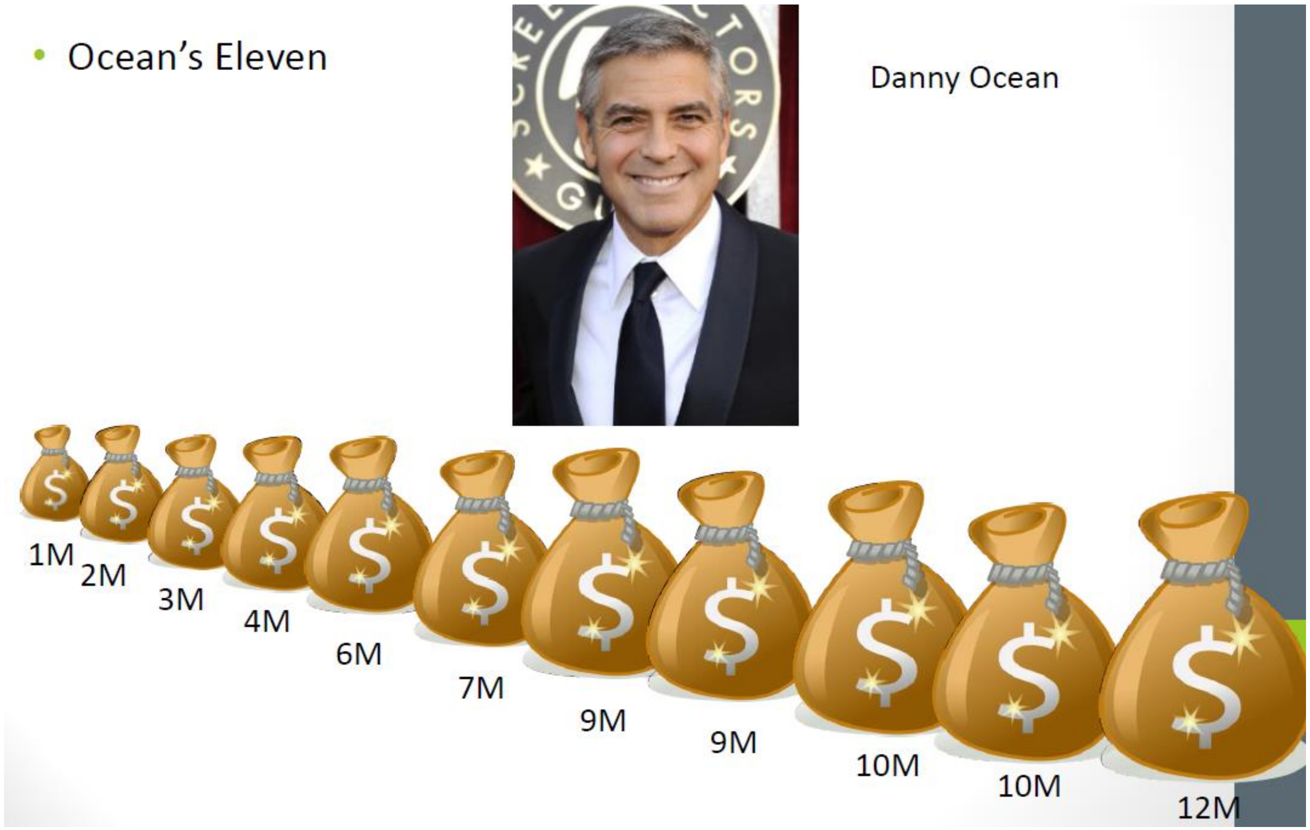
[1] 8

> table(Length) #shows how many times every value occurs; the most popular value is the mode

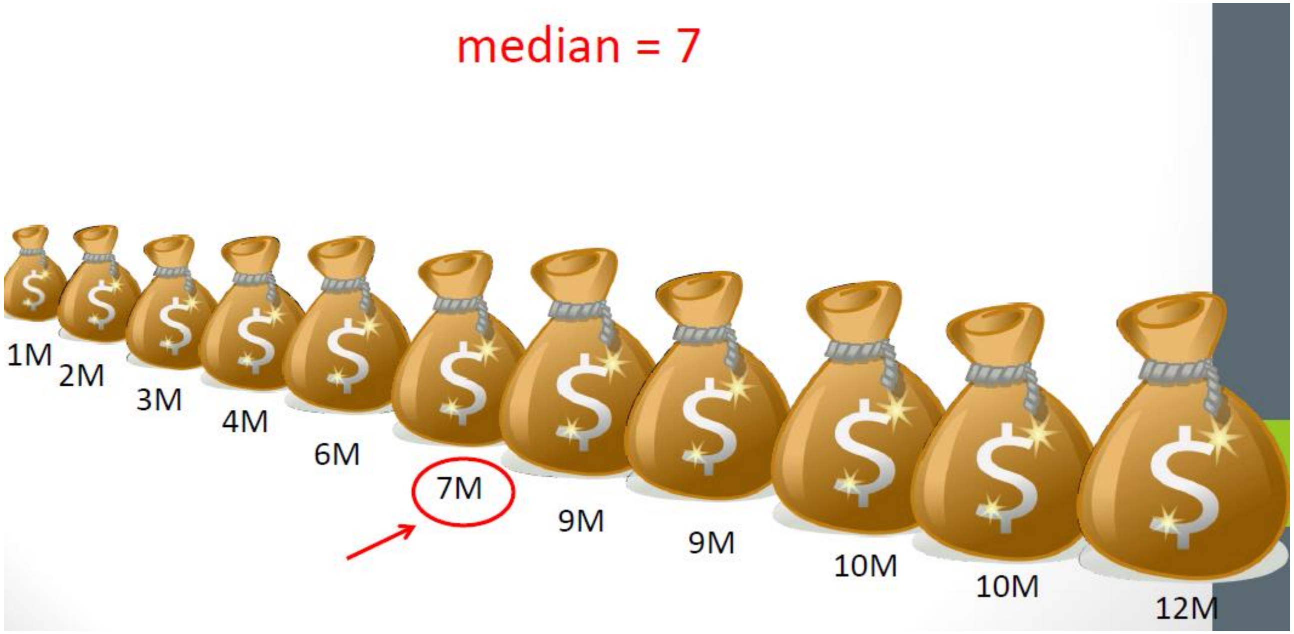
3 4 5 6 7 8 9 10 11 12 13 14 15

2 5 7 13 12 16 11 16 11 3 1 2 1

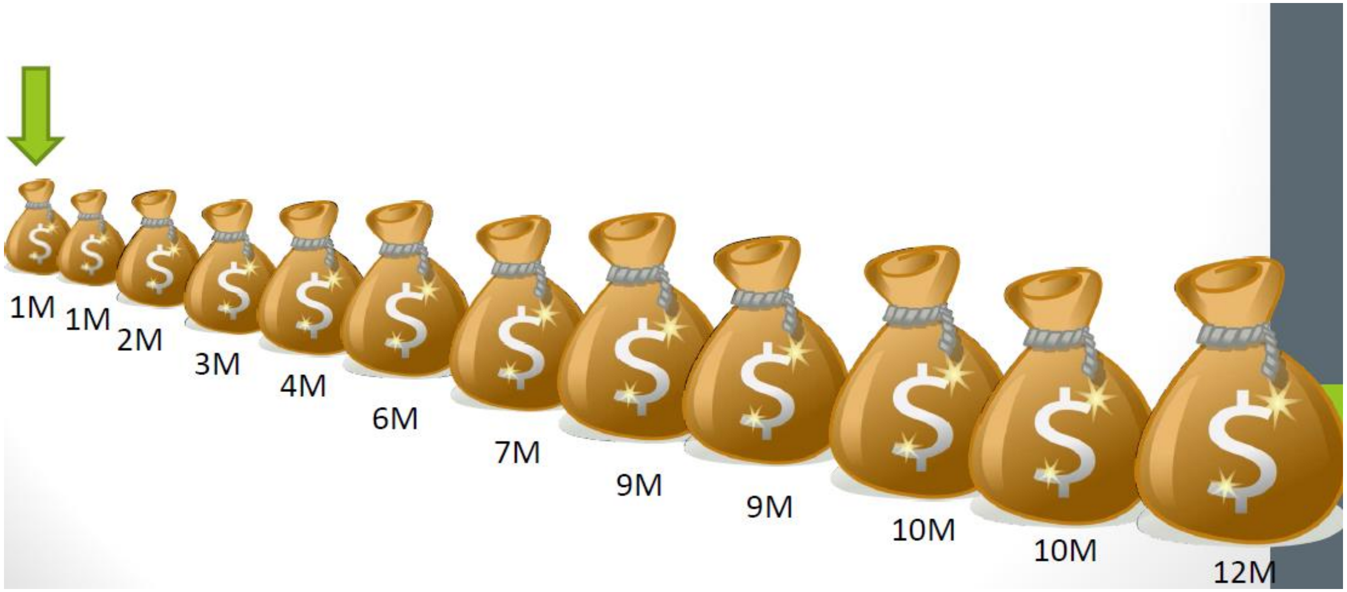
Understanding the median



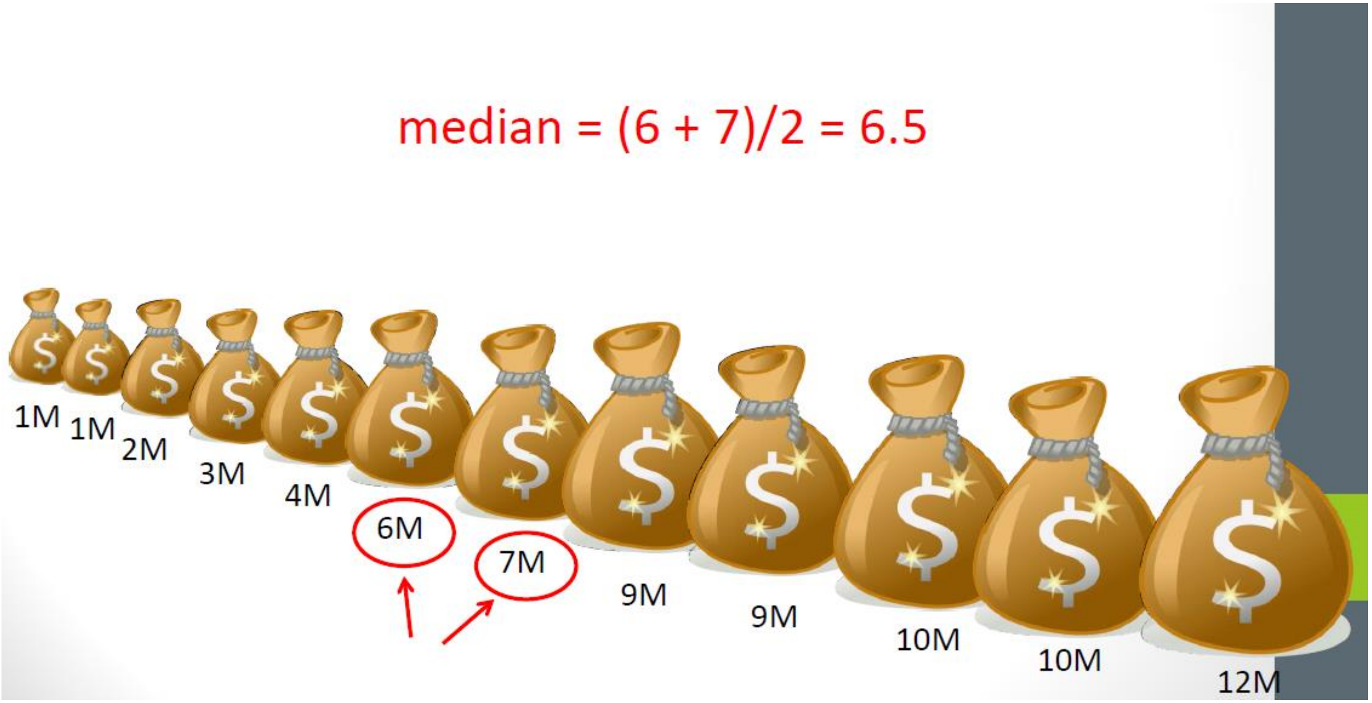
Ocean's 11: the median



Ocean’s 12



Ocean's 12: the median



Mean vs. median

- In some situations the median gives a better idea of the most typical value than the mean. The problem with the latter is that it is easily influenced by outliers, i.e. scores with unusually high or low values.
- For example, if twenty employees in a company have net salaries of €2000 a month, and the CEO's salary is €50000, the mean salary will be €4286, and the median will be €2000. The median gives a more realistic idea of the salaries in the company than the mean because the CEO's salary is exceptional.

Exercise

- Find the mean and the median of the reaction times in ldt.

A very useful function summary()

```
> summary(Mean_RT)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  564.2   713.1   784.9   808.3   905.2  1458.8
```

Outline

1. Measures of central tendency
2. Measures of dispersion
3. Graphical representation
4. Testing normality

Measures of dispersion

```
> range(Length) # the minimum and the maximum
[1]  3 15
> var(Length) #variance = sum of squared deviations
[1] 6.259697                  from the mean, divided by
                              the number of observations
                              - 1
> sd(Length) # standard deviation = the squared root
[1] 2.501939                              of variance
```

Why care about the dispersion?

- Consider two countries with a similar average income per capita. In one country the variance and standard deviation are relatively small because the finances are distributed fairly, whereas in the other they are very large because of several billionaires and many extremely poor people. Although the means are identical, life in the two countries will differ dramatically.

Statisticians make jokes, too ☺

- If your head is in the oven, and your feet are in the fridge, on average you're quite comfortable.

Image from moneymarketing.co.uk

Outline

1. Measures of central tendency
2. Measures of dispersion
3. Graphical representations
4. Testing normality

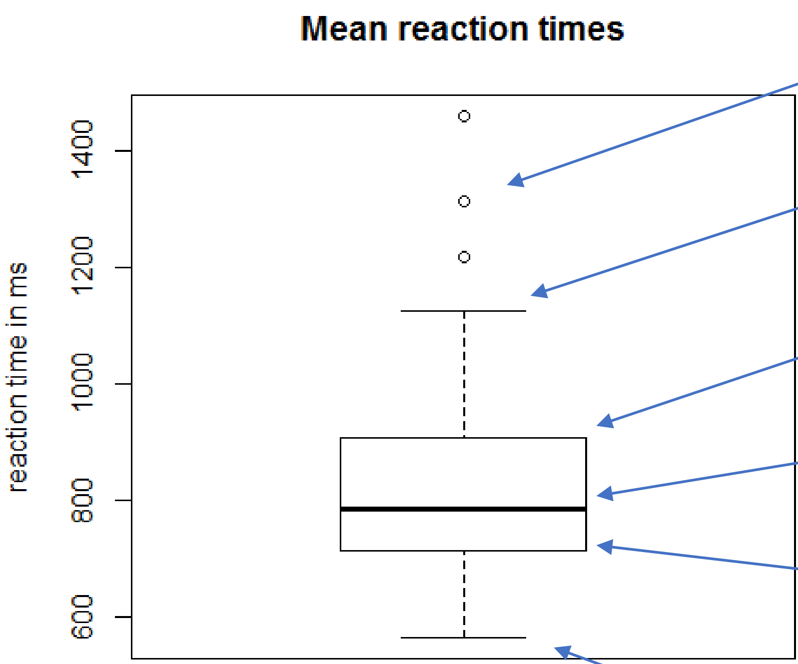
Boxplot

```
> boxplot(Mean_RT)
```

A bit more sophisticated:

```
> boxplot(Mean_RT, main = "Mean reaction times",
ylab = "reaction time in ms")
```

Box-and-whisker plot



Outliers (more that 1.5 times IQR)

Upper whisker: max.1.5 times IQR from the upper hinge

Upper hinge: 3rd quartile, or 75%

Median (or the 2nd quartile, or 50%)

Lower hinge: the 1st quartile, or first 25%

Lower whisker: max. 1.5 times IQR from the lower hinge

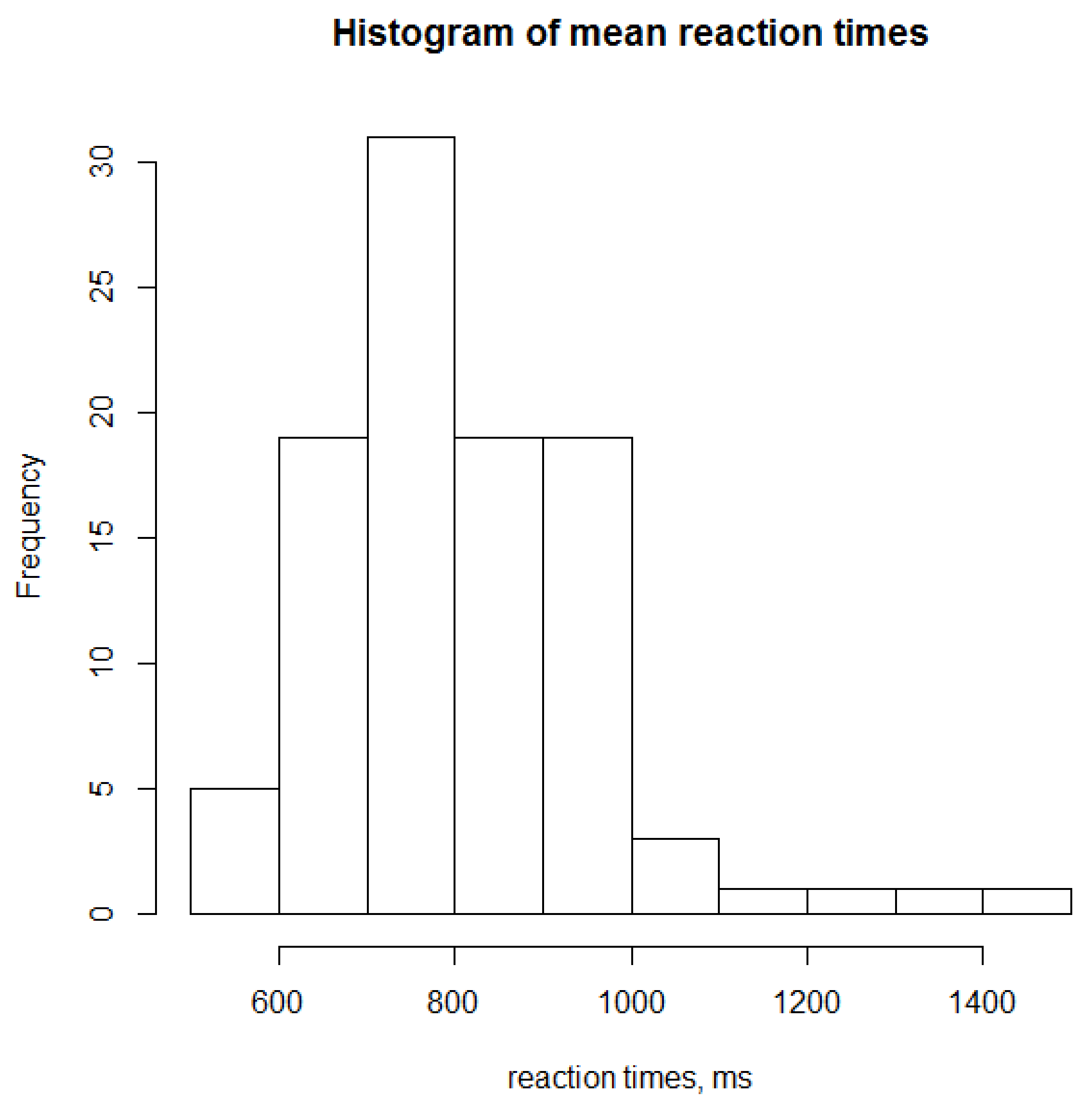
Boxplot stats

```
> boxplot.stats(Mean_RT)
$stats # l. whisker, l. notch, median, u. notch, u. whisker
[1]  564.180  712.285  784.940  905.930 1125.420

$n #number of observations
[1] 100

$conf # ignore for the time being
[1] 754.3441 815.5359

$out # outliers
[1] 1314.33 1216.81 1458.75
```

Histogram

```
> hist(Mean_RT, main = "Histogram of mean reaction times", xlab = "reaction times, ms")
```

A histogram shows the frequencies of different values aggregated in bins.

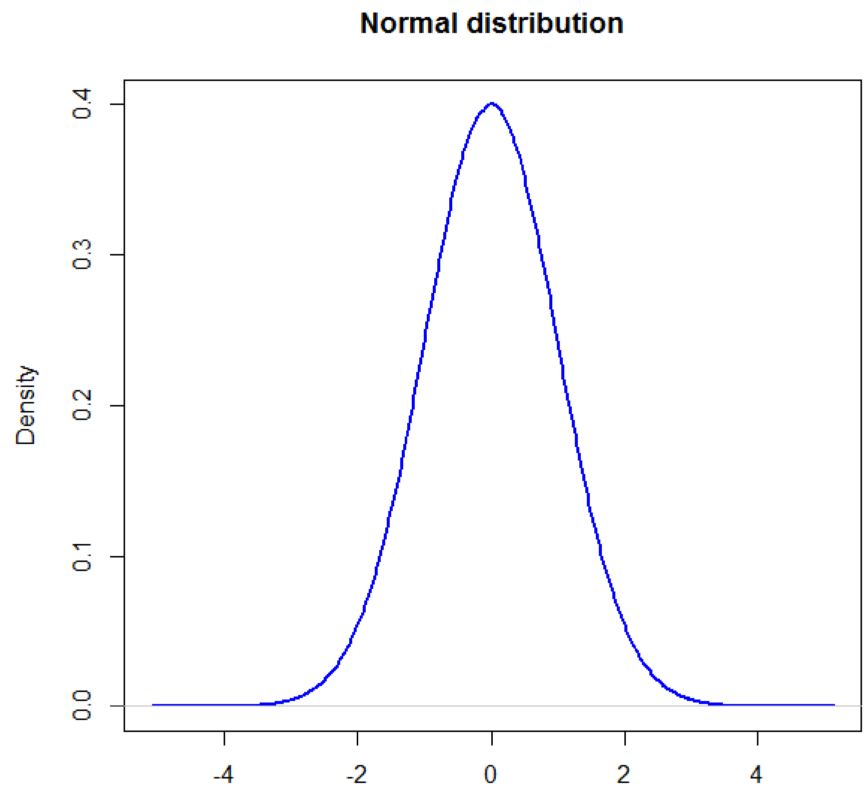

Histogram of mean reaction times
Frequency
0
5
10
15
20
25
30
600
800
1000
1200
1400
reaction times, ms

Outline

1. Measures of central tendency
2. Measures of dispersion
3. Graphical representations
4. Testing normality

Normal distribution

- A bell-shaped curve
- Mean = median = mode



Why is it important?

- Some tests require that the data should be normally distributed.
- This is why one should learn how to test if the data are normally distributed.

Shapiro test

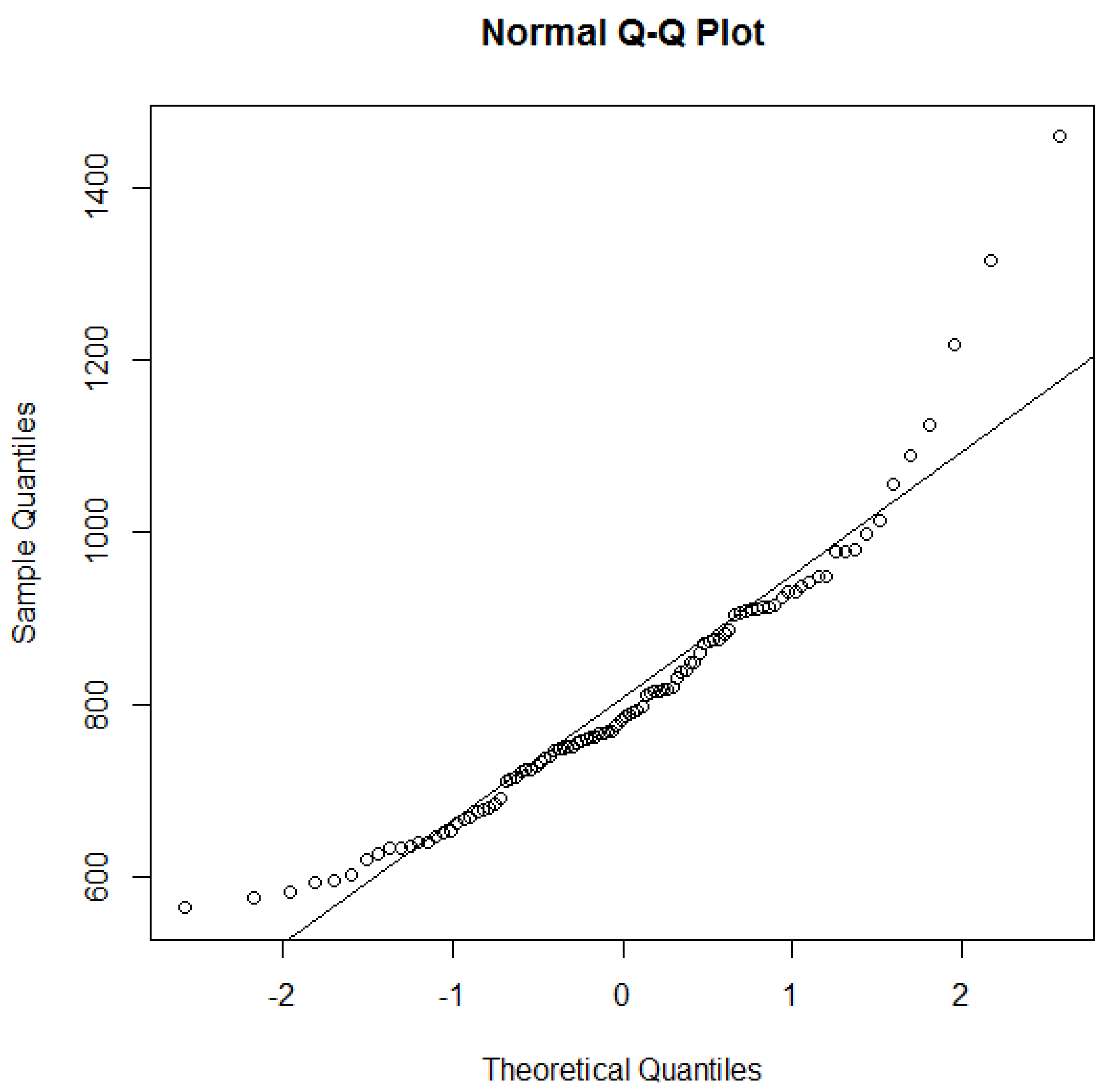
```
> shapiro.test(Mean_RT)

        Shapiro-Wilk normality test

data:  Mean_RT
W = 0.92006, p-value = 1.418e-05
```

A small *p*-value (less than 0.05) means that the assumption that the data are normally distributed can be rejected.

QQ-plot

```
> qqnorm(Mean_RT)
> qqline(Mean_RT)
```

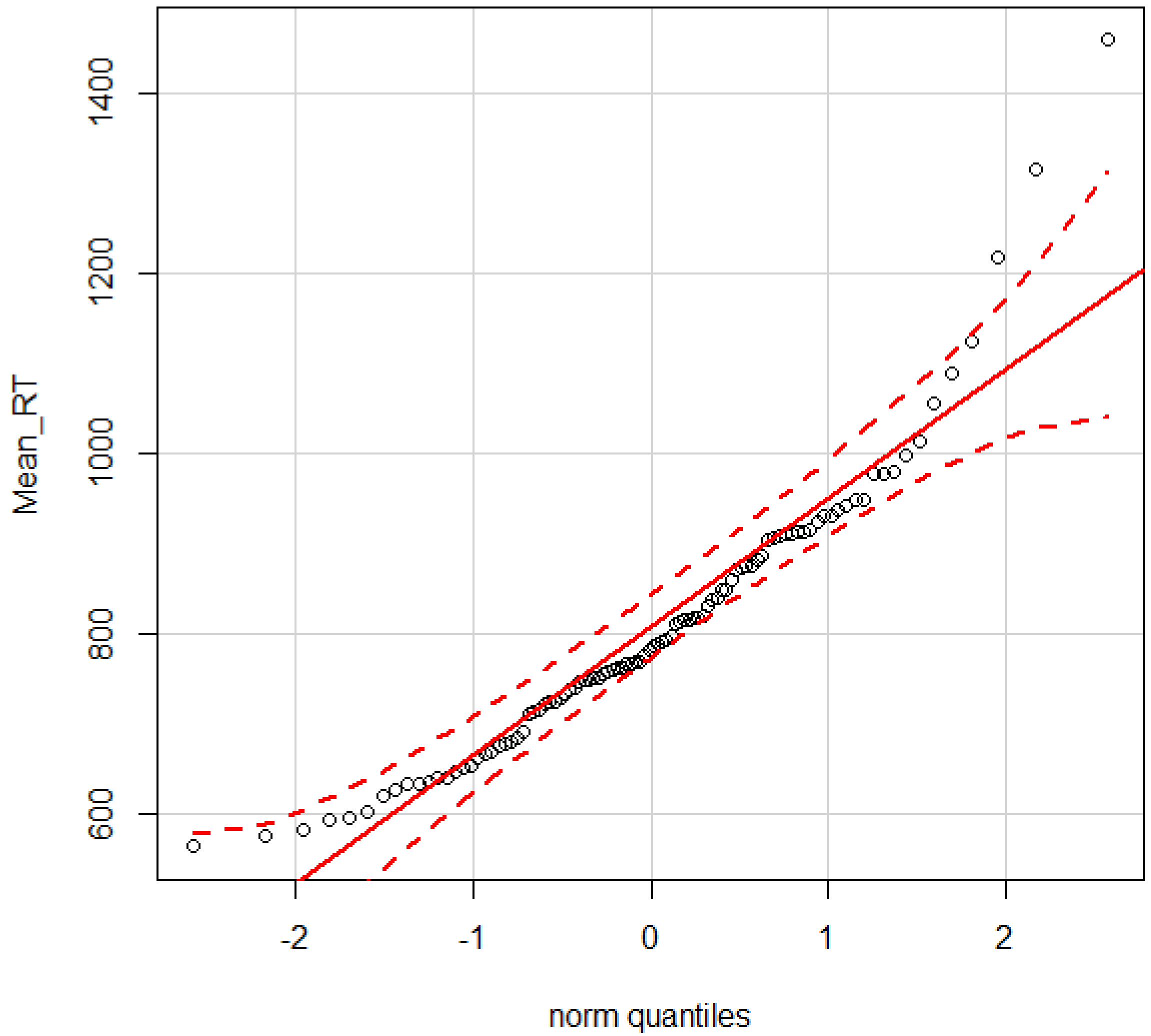
The points should lie close to the line, which is not the case.


Normal Q-Q Plot
Sample Quantiles
600
800
1000
1200
1400
Theoretical Quantiles
-2
-1
0
1
2

Another convenient function

```
> library(car)
> qqPlot(Mean_RT)
```

The plot displays a 95% confidence ‘envelope’ around the distribution. The points should be inside the envelope. If not, there’s a problem. The plot clearly shows that the three outliers with high scores are problematic.



Mean_RT
1400
1200
1000
800
600
-2
-1
0
1
2
norm quantiles

What to do with outliers?

- If they represent errors, remove them.
- If there are correct, one has several options:
  - Use a test which does not have a normality assumption (e.g. a non-parametric test).
  - Transform the variable (e.g. taking a logarithm).
  - Assign smaller values, e.g. based on the number of standard deviations from the mean.

Exercise

- Remove the outliers:

```
> Mean_RT_new <- Mean_RT[Mean_RT < 1200]
> Length(Mean_RT_new)
[1] 97
```

- Perform diagnostics: does the new sample look more normally distributed?