# Part 3
# Basic descriptive statistics and plots

Natalia Levshina © 2015

Part of the course taught at the University of Mainz, Germany
26-28 May 2015

# Data set 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 set ldt

```
> 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 ...
```

# Outline

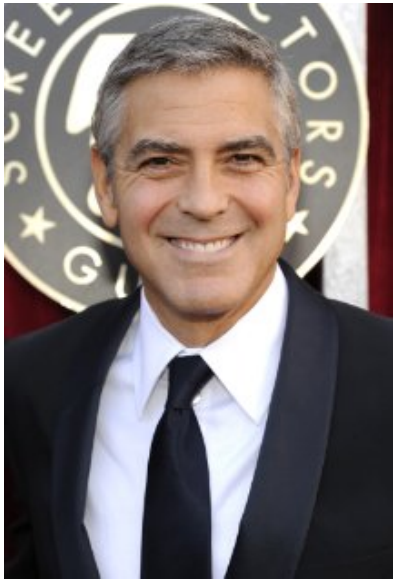
1. Measures of central tendency
2. Measures of dispersion
3. Boxplot and histogram

# Mean, median and mode

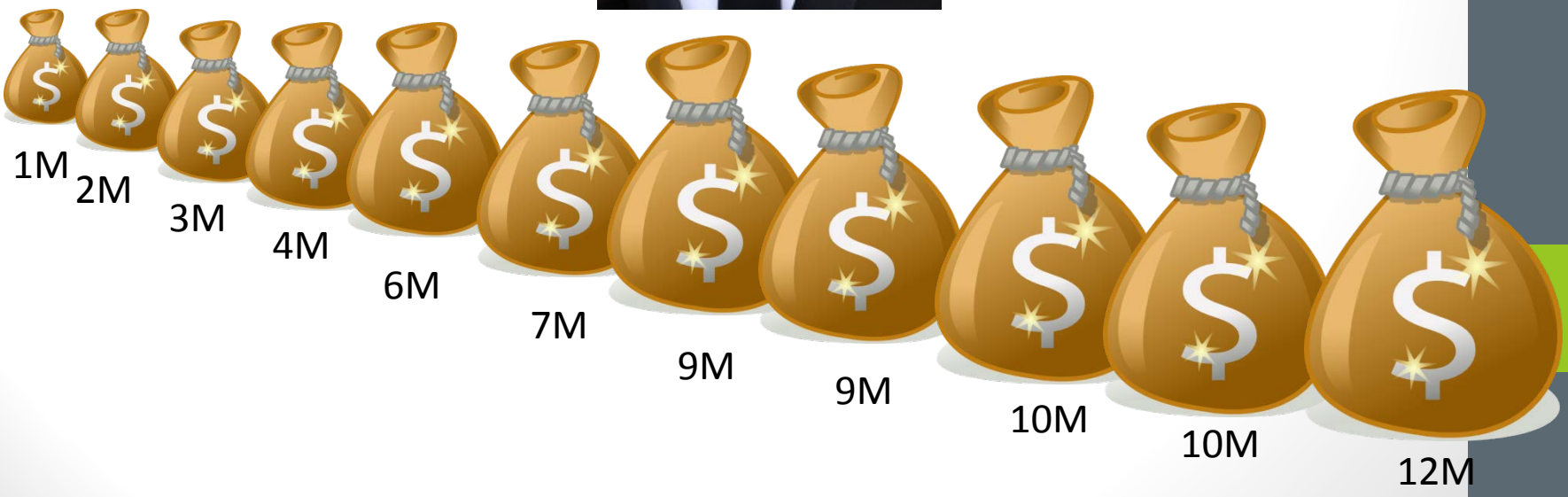
```
> mean(ldt$Length)
[1] 8.23

> median(ldt$Length)
[1] 8

> table(ldt$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 Eleven



Danny Ocean



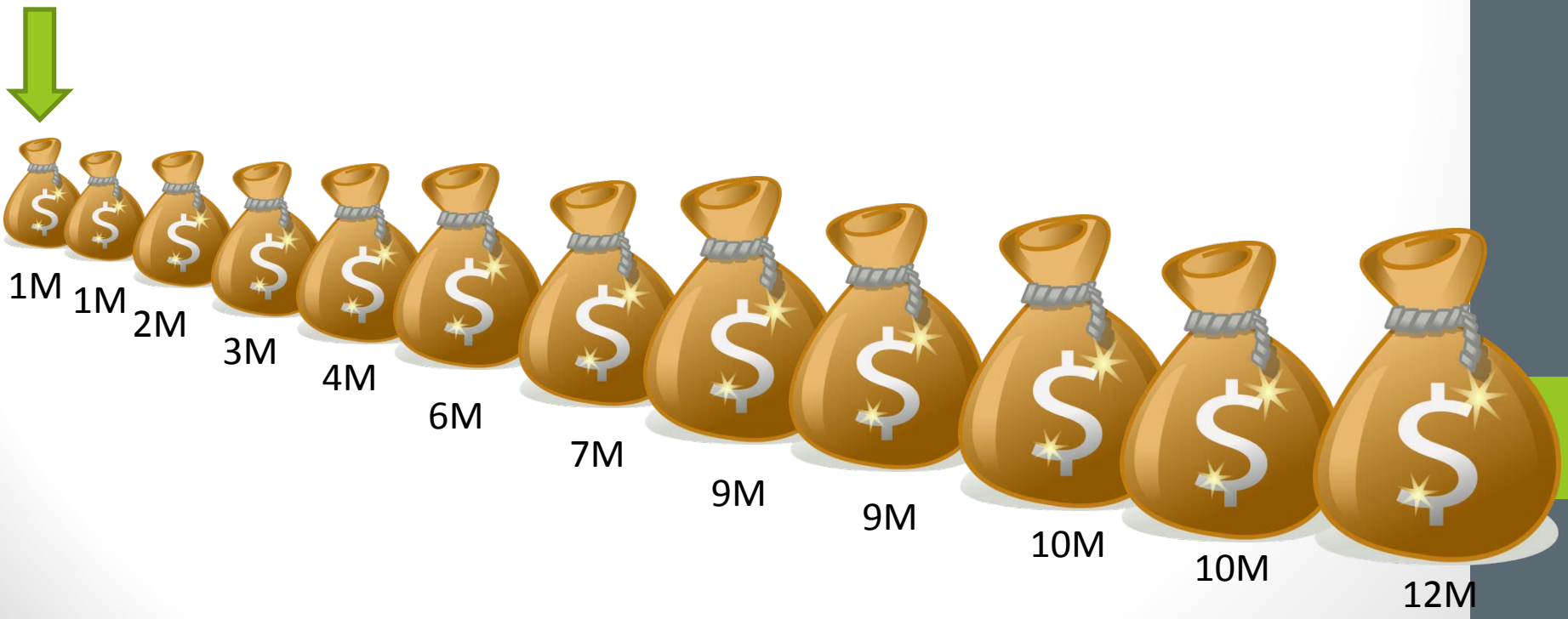
# Understanding the median

- Ocean's Eleven

median = 7

# Understanding the median

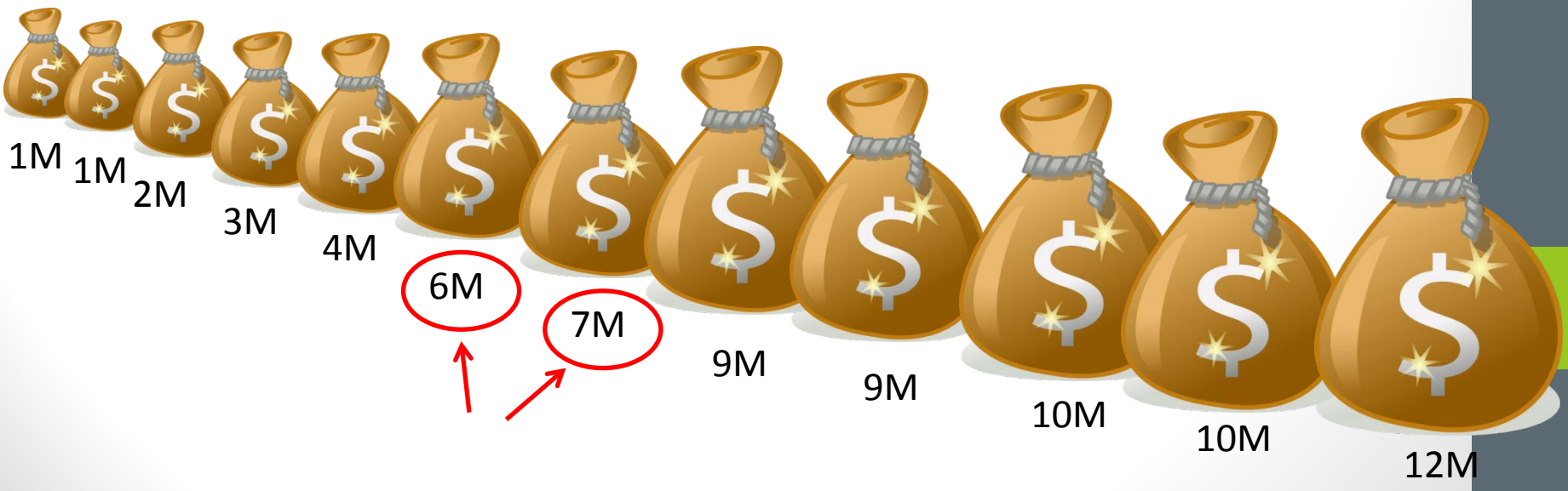
- Ocean's Twelve



# Understanding the median

- Ocean's Twelve

median = (6 + 7)/2 = 6.5



# 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 minimum, maximum, mean and median values of the reaction times in ldt.

# summary() function

```
> summary(ldt$Length)
```

| Min. | 1st Qu. | Median | Mean | 3rd Qu. | Max. |
|---|---|---|---|---|---|
| 3.00 | 6.00 | 8.00 | 8.23 | 10.00 | 15.00 |

# Outline

1. Measures of central tendency
2. Measures of dispersion
3. Boxplot and histogram

# Measures of dispersion

```
> range(ldt$Length) # range = 15 – 3 = 12
[1]  3 15

> var(ldt$Length) # variance = sum of squared deviations
[1] 6.259697                from the mean divided by the
                            number of observations minus 1

> sd(ldt$Length) # standard deviation = square root of
 [1] 2.501939               variance
```

# Why care about measures of 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.

# Statistical joke

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

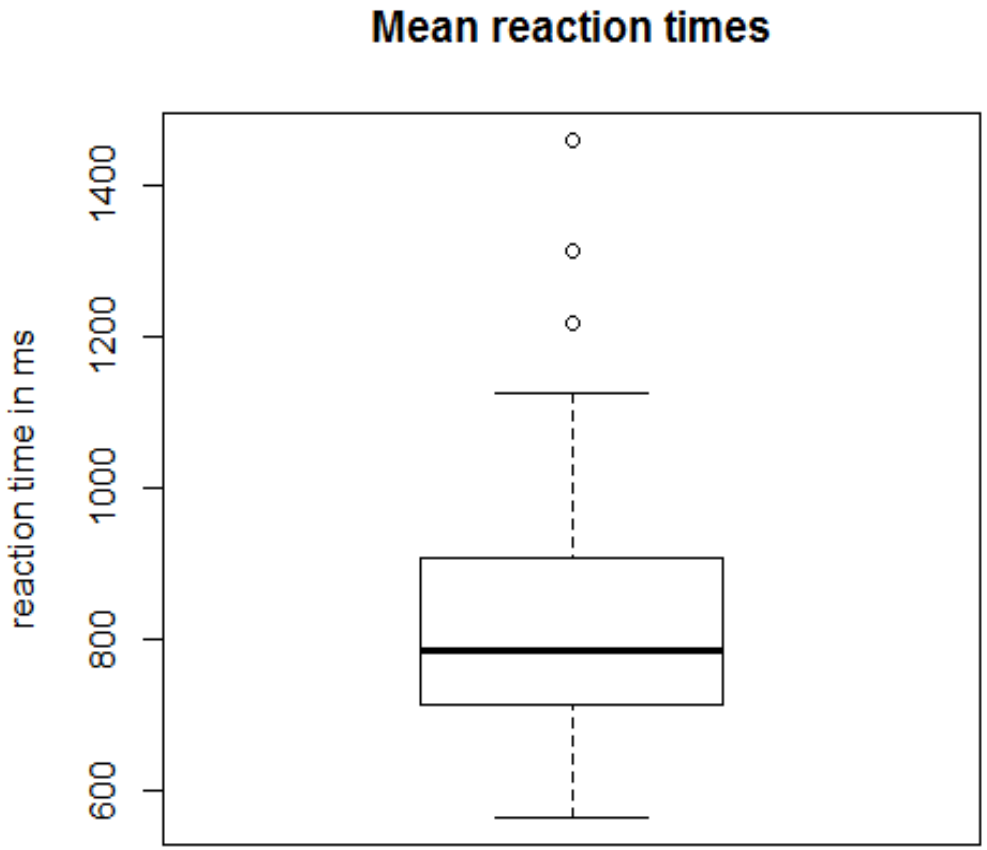

Image from moneymarketing.co.uk

# Outline

1. Measures of central tendency
2. Measures of dispersion
3. Boxplot and histogram

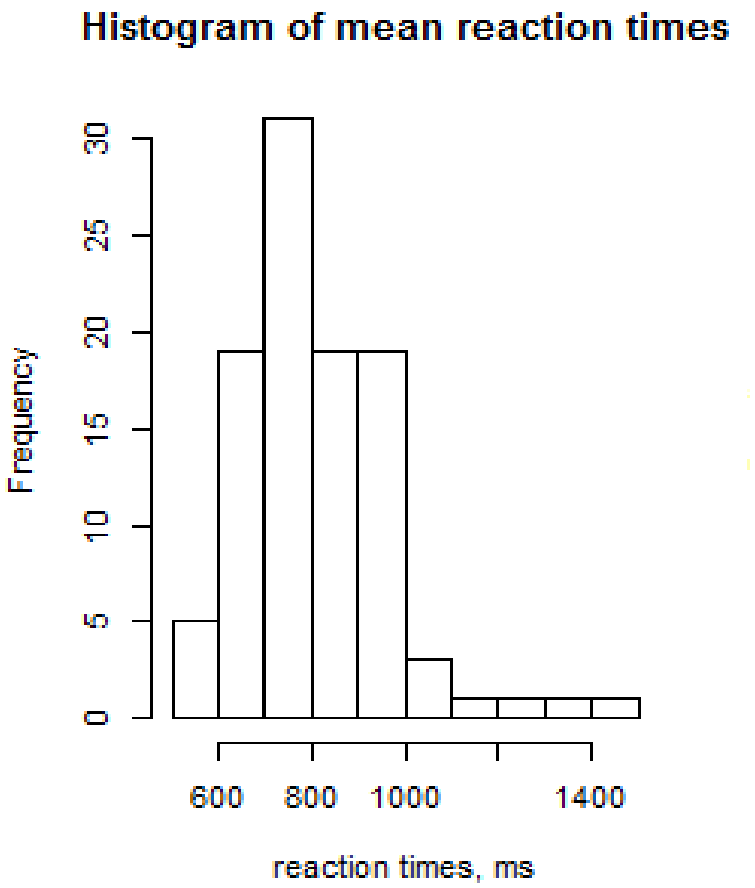
# Boxplot

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


Mean reaction times
reaction time in ms
600
800
1000
1200
1400

# Histogram

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


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

# Exercise

| | Subj 1 | Subj 2 | Subj 3 | Subj 4 | Subj 5 | Subj 6 | Subj 7 | Subj 8 | Subj 9 | Subj 10 |
|---|---|---|---|---|---|---|---|---|---|---|
| Reaction time, ms | 583 | 667 | 1149 | 827 | 488 | 452 | NA | 739 | 455 | 572 |

1. Create a vector with the reaction times from the table.
2. Compute the mean and the median (mind the NA value: add na.rm = TRUE).
3. Compute the range, variance and standard deviation
4. Create a box plot and a histogram