University of St.Gallen

Methods: Statistics (4,120)

# 1. Data and Descriptive Statistics

Spring 2022

**Prof. Dr. Roland Füss**

Swiss Institute of Banking and Finance (s/bf)

# Contents

I. Data Sources and Data Collection

II. Scales of Measurement

III. Frequency Distributions

IV. Measures of Central Tendency

V. Measures of Dispersion

Appendix

# Learning Objectives

After the lecture, you know how:

- the **different data types can** be distinguished.
- the concepts of **descriptive statistics** are applied as well as interpreted in practice-oriented examples.
- data can be **prepared** and **graphically** displayed for interpretation.

# Literature

Levine, D.M., K. A. Szabat, and D.F. Stephan. (2016). *Business Statistics: A First Course*, 7th ed., United States: Pearson, **Chapter 1-3**.*

Stinerock, R. (2018). *Statistics with R*. United Kingdom: Sage. **Chapter 1-3**.*

Shira, Joseph (2012). *Statistische Methoden der VWL und BWL*, 4th ed., München et al.: Pearson, **Chapter 1-2**.

Weiers, R. M. (2011). *Introductory Business Statistics*, 7th ed., Canada: Thomson South-Western, **Chapter 1-4**.

*Mandatory literature

# I. Data Sources

1. **Primary data** is generated by the researcher for a specific problem.

2. **Secondary data** is collected by external research institutions or companies for some other purpose

| Internal Sources | Internet | Statistics | Other |
|---|---|---|---|
| • sales statistics<br>• order statistics<br>• customer statistics<br>• historical data<br>• data warehouse<br>• cost statistics<br>• etc. | • institute<br>• services<br>• databases<br>• portals<br>• etc. | • incomes<br>• employment<br>• interest rates<br>• money supply<br>• etc. | • publisher typology<br>• banking statistics<br>• consultant studies<br>• databases<br>• etc. |

# I. Data Collection

| Survey | Observation | Experiment |
|---|---|---|
| • standardized/unstandardized<br>• oral/written<br>• online/offline<br>• one-time/regular (panel)<br>• field/laboratory survey<br>• one-topic and multi-topic survey | • standardized/unstandardized<br>• personal/apparative<br>• participating/non-participating<br>• field/laboratory observation<br>• open/not open<br>• decision/behaviour covered | Experiments are not independent survey techniques, but variations of surveys and observations.<br>• laboratory/field experiment<br>• random/non-random experiment |

| Written Questionnaire | Oral Interview | Telephone Survey |
|---|---|---|
| • large area can be covered<br>• low costs<br>• no interviewer influence<br>• often bad return rate<br>• number of questions limited<br>• ordering effects | • high return rate<br>• low number of questions<br>• survey conditions can hardly be grasped<br>• additional information ascertainable<br>• high costs<br>• high interviewer impact | • manageable costs<br>• quickly implementable<br>• only certain survey topics possible<br>• no visual support possible |

# I. Data Collection (Example)

**Questionnaire**

***What do you think, how big is your knowledge in statistics?*** ____________
*(Rating according to school grading system at 1 for very good and 6 for very bad)*

***Are you interested in statistics?***
□ not at all □ little □ moderately strong □ very strong □ don’t know

***Sex:*** ***Family status:*** _______
□ female □ male ***Age:*** ______

***What amount of money is at your disposal each month?***
□ < 500 CHF □ 500-999 CHF
□ 1,000-1,499 CHF □ 1,500-1,999 CHF
□ 2,000-2,999 CHF □ 3,000 CHF and more

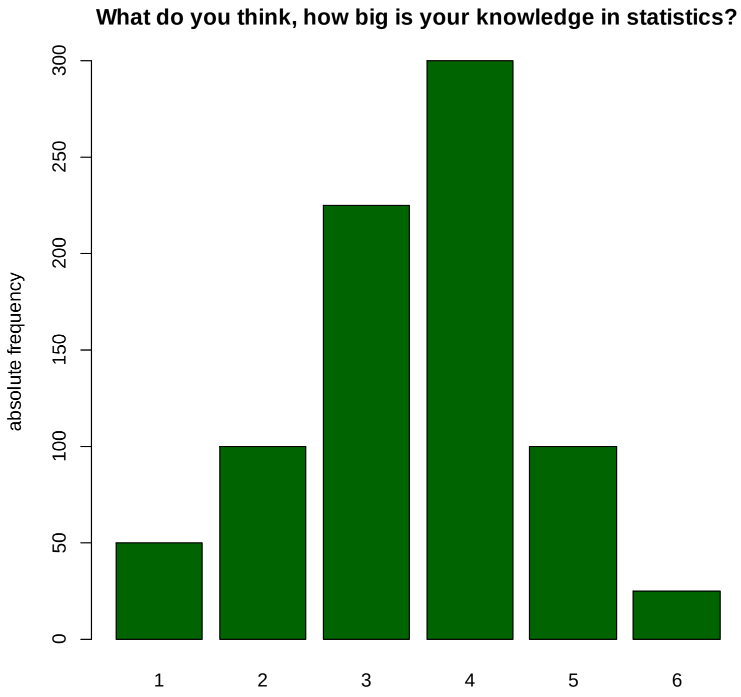
# I. Data Collection (Example)

**Collection:**

↓

**Data Mass:**

| ID | Knowledge in statistics | sex | Monthly amount of money |
|---|---|---|---|
| 1 | 6 | feminine | 499 CHF |
| 2 | 4 | masculine | 1.345 CHF |
| 3 | 2 | masculine | 2.050 CHF |
| 4 | 1 | feminine | 798 CHF |
| … | … | … | … |



The **mean value** $\mu = 3.47$ determines the position of the distribution.

The **standard deviation** $\sigma = 1.31$ determines the width of the distribution.

largeur

# II. Scales of Measurement

A **statistical population** consists of different units with identical **factual**, **temporal**, and **spatial** characteristics. The survey (or partial census) determines an **observation value** for them, i.e. one of the possible **characteristic values** is assigned to each unit**.**

The **scale of measurement** influences both the collectability of data and the choice of the appropriate sampling technique, and thus, the information content of the measurement results. We distinguish between the following **types of scales**:

# II. Scale of Measurement

| Type of Scale | Scale | Properties / Examples |
|---|---|---|
| **qualitative** | nominal | A = B or A ≠ B<br>distinctiveness, number represents category,<br>e.g. gender, profession or nationality |
| **comparative** | ordinal | A = B or A < B or A > B<br>possibility to build a rank order<br>e.g. ratings, grades, preferences |
| **quantitative** | interval | A = B or A < B or A > B **and**<br>B - A = C - B or B - A < C - B or B - A - A > C - B,<br>ranking and equal intervals between scale values,<br>e.g. year of birth, deviation from norm, Celsius |
|  | ratio | same as interval scale **and**<br>A = x * B<br>rationale zero point, equal intervals between adjacent scale values, calculation of measurement<br>e.g. prices, income, turnover, weight |

# II. Description of Data

1) **Qualitative variables** assign objects to a category:

Examples: gender, legal structure of company, nationality

2) **Quantitative variables** assign a value to an object:

- **Discrete:** Variables can take only certain values along an interval.

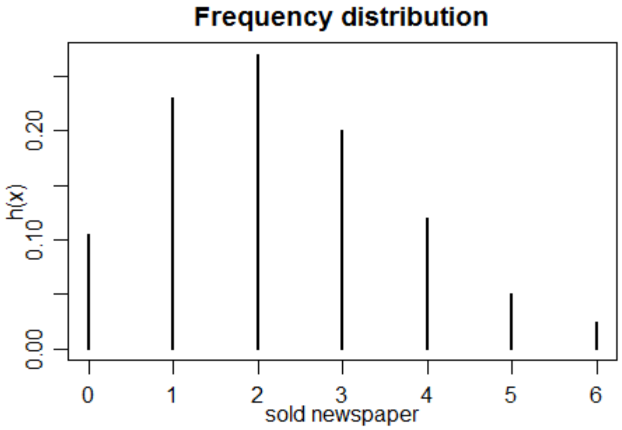
  Examples: Numbers in roulette from 0 - 36, numbers on dice.

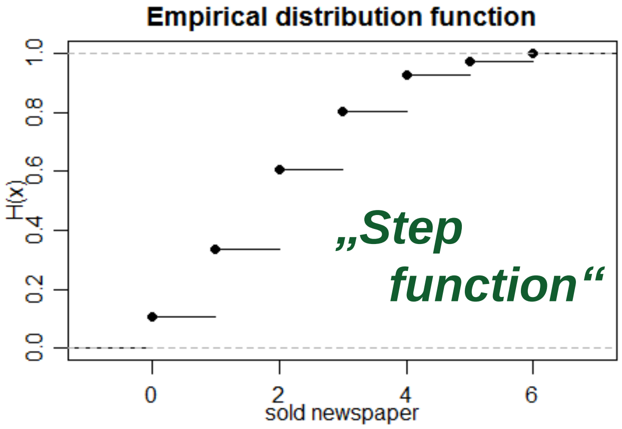
- **Continuous:** Variables can take any value in an interval.

  Examples: Speed, any value between 0 and 250 km/h

# III. Description of Data (R-Example 1)

Number of newspapers sold per day at a kiosk:



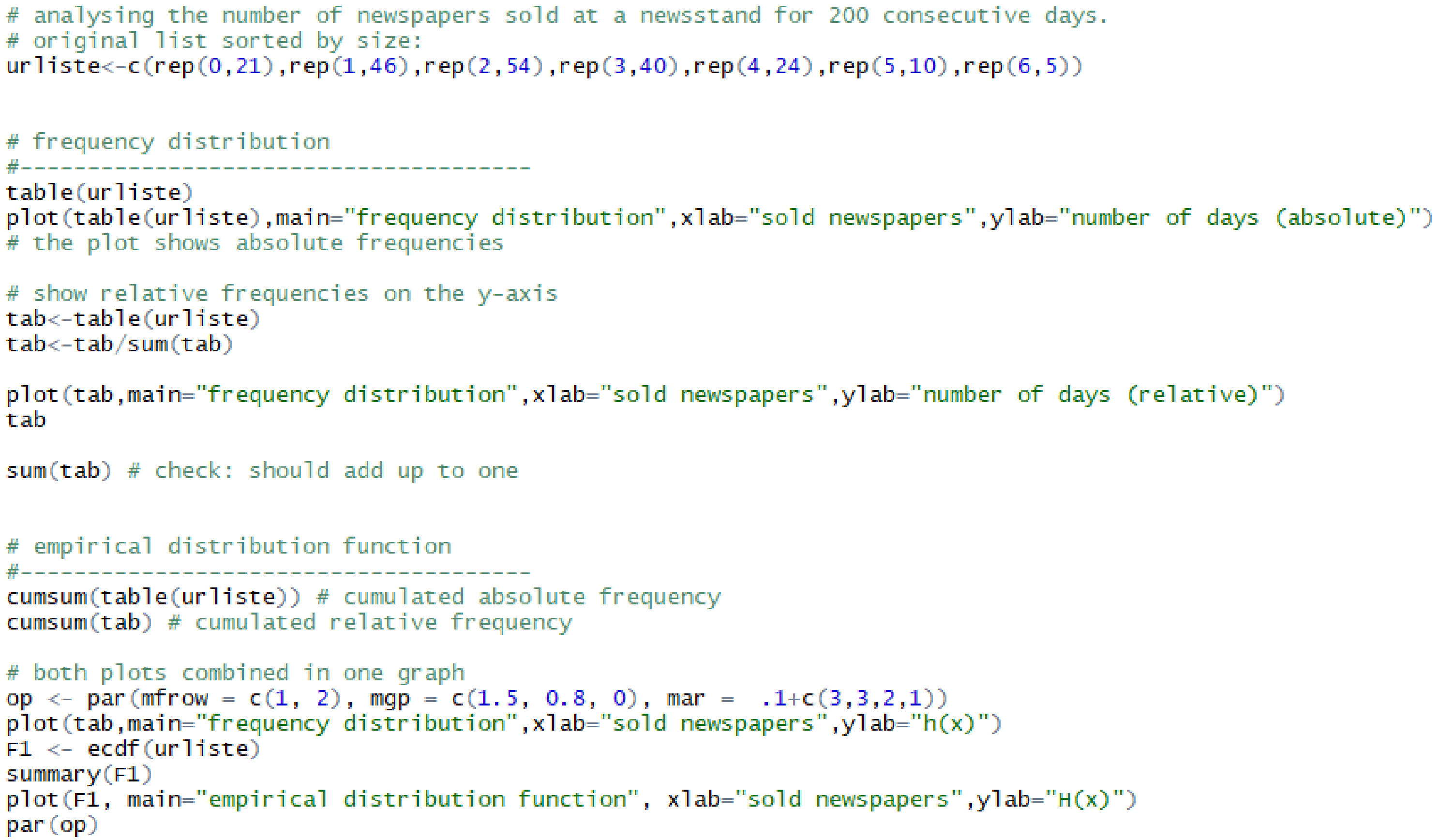


*„Step function“*

| $x_i$ | $n_i$ | $h_i$ | $H_i$ |
|---|---|---|---|
| 0 | 21 | 0.105 | 0.105 |
| 1 | 46 | 0.230 | 0.335 |
| 2 | 54 | 0.270 | 0.605 |
| 3 | 40 | 0.200 | 0.805 |
| 4 | 24 | 0.120 | 0.925 |
| 5 | 10 | 0.050 | 0.975 |
| 6 | 5 | 0.025 | 1 |
| | 200 | 1 | |

# III. Description of Data (R-Example 1)

Open the file "L1-Example_1.R" in R-Studio and reproduce the R-Code, which creates the previous graphs.



```
# analysing the number of newspapers sold at a newsstand for 200 consecutive days.
# original list sorted by size:
urliste<-c(rep(0,21),rep(1,46),rep(2,54),rep(3,40),rep(4,24),rep(5,10),rep(6,5))


# frequency distribution
#---------------------------------------
table(urliste)
plot(table(urliste),main="frequency distribution",xlab="sold newspapers",ylab="number of days (absolute)")
# the plot shows absolute frequencies

# show relative frequencies on the y-axis
tab<-table(urliste)
tab<-tab/sum(tab)

plot(tab,main="frequency distribution",xlab="sold newspapers",ylab="number of days (relative)")
tab

sum(tab) # check: should add up to one


# empirical distribution function
#---------------------------------------
cumsum(table(urliste)) # cumulated absolute frequency
cumsum(tab) # cumulated relative frequency

# both plots combined in one graph
op <- par(mfrow = c(1, 2), mgp = c(1.5, 0.8, 0), mar =  .1+c(3,3,2,1))
plot(tab,main="frequency distribution",xlab="sold newspapers",ylab="h(x)")
F1 <- ecdf(urliste)
summary(F1)
plot(F1, main="empirical distribution function", xlab="sold newspapers",ylab="H(x)")
par(op)
```

# III. Description of Data

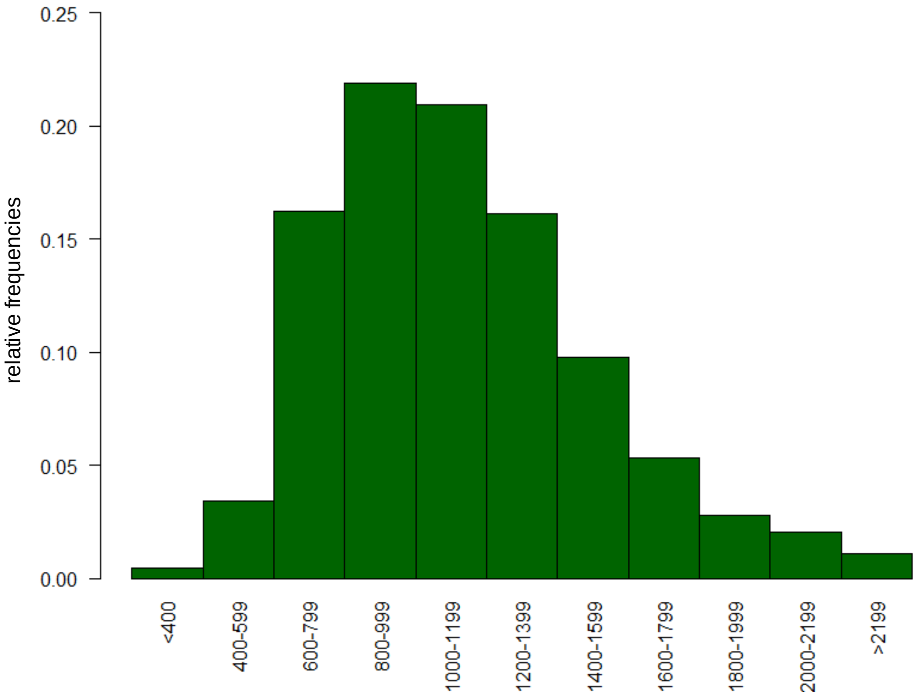
As a first step in statistical data analysis, the data should be **suitably prepared** so that a tabular or graphical representation provides a deeper **insight into the information content of** the data.

| Price Segments of Apartments in Canton Zurich | Number of Apartments in this Price Segment | | Empirical Distribution Function |
|---|---|---|---|
| | **Absolute Frequencies** | **Relative Frequencies** | |
| < 400 CHF | 615 | 0.46% | 0.46% |
| 400 - 599 CHF | 4504 | 3.40% | 3.86% |
| 600 - 799 CHF | 21492 | 16.21% | 20.07% |
| 800 - 999 CHF | 28997 | 21.87% | 41.07% |
| 1000 - 1199 CHF | 27777 | 20.95% | 62.89% |
| 1200 - 1399 CHF | 21372 | 16.12% | 79.01% |
| 1400 - 1599 CHF | 12970 | 9.78% | 88.79% |
| 1600 - 1799 CHF | 7033 | 5.30% | 94.04% |
| 1800 - 1999 CHF | 3695 | 2.79% | 96.88% |
| 2000 - 2199 CHF | 2711 | 2.04% | 98.92% |
| > 2199 CHF | 1451 | 1.09% | 100% |
| **Sum** | **132617** | **100%** | |

In the example of the **price ranges of apartments in the Canton of Zurich**, the following questions, among others, can be answered:

- What is the proportion of a single price segment?
- To what extent do the price segments differ with respect to their respective frequencies?

# III. Description of Data



Example: Price Segments of Apartments in the Canton of Zurich

In order to be able to compare several frequencies better - at a glance - it is advisable to display them graphically in the form of the **bar chart**.

This diagram consists of bars erected above the characteristic values and whose areas are **proportional (directly)** to the frequencies of the respective characteristics.

Therefore, the weightings of the different characteristic values can directly be compared. In this regard it is called a „frequency distribution“ resp. „**histogram**“.

# IV. Measures of Central Tendency

**Measures of central tendency** characterize the center of a distribution and reflect the "typical values". The most important parameters are :

- arithmetic mean
- weighted average
- geometric mean
- median
- mode
- quartiles

# IV. Measures of Central Tendency

- population mean $$\mu = \frac{\sum_{i=1}^{N} X_i}{N}$$

- sample mean $$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

where *N* = number of observations in the population

*n* = number of sample observations

*X* = Observations in the population

*x* = sample observations

# IV. Measures of Central Tendency

**Arithmetic Mean**

$$\mu = \frac{\sum_{i=1}^{N} X_i}{N}$$

**Weighted Average**

$$\mu = \frac{\sum_{i=1}^{N} w_i X_i}{\sum_{i=1}^{N} w_i}$$

**Price for a VW Golf in 5 markets (in €):**

| GB | F | D | J | CH |
|---|---|---|---|---|
| 21000 | 17000 | 19000 | 25000 | 28000 |

**Corresponding quantities sold:**

| GB | F | D | J | CH |
|---|---|---|---|---|
| 15000 | 13000 | 38000 | 18000 | 7000 |
| 0.16 | 0.14 | 0.42 | 0.20 | 0.08 |

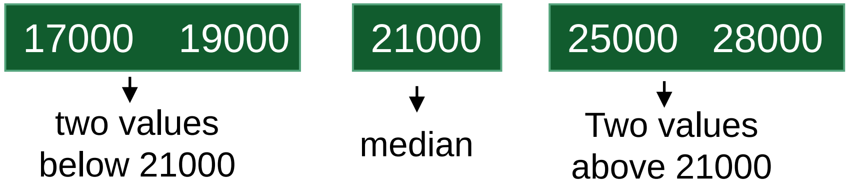
$$\mu = \frac{21000 + 17000 + 19000 + 25000 + 28000}{5} = 22000$$

$$\mu_w = 0.16 \times 21000 + 0.14 \times 17000 + 0.42 \times 19000$$
$$+ 0.2 \times 25000 + 0.08 \times 28000 = 20960$$

# IV. Measures of Central Tendency

50% of the values are smaller and 50% are larger than the **median**

Ordered Price Data:



17000 19000

21000

25000 28000

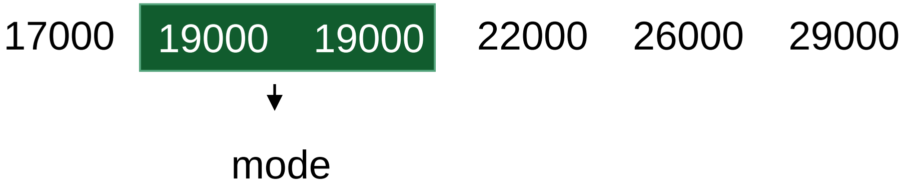
two values
below 21000

median

Two values
above 21000

The **mode** represents the value with the highest frequency in a distribution of observations (most typical value).

Most frequent value in the series of numbers

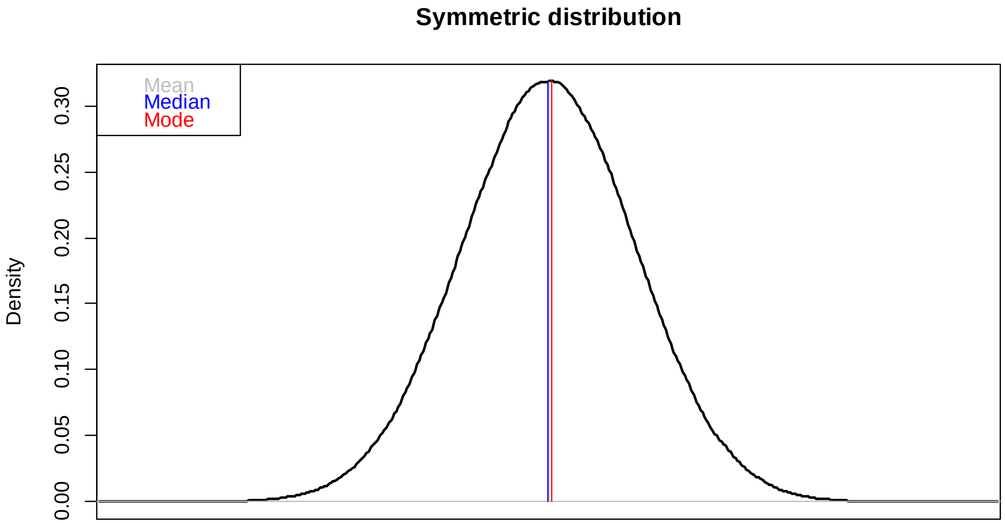


17000 19000 19000 22000 26000 29000

mode

# IV. Measures of Central Tendency

In the case of a **symmetrical distribution**, the mean value, median and mode are the same:



A symmetrical distribution has a **skewness** of 0. The skewness describes the **tendency of the distribution** to assume extreme values at the left or right at the end of the distribution function with above-average probabilities.

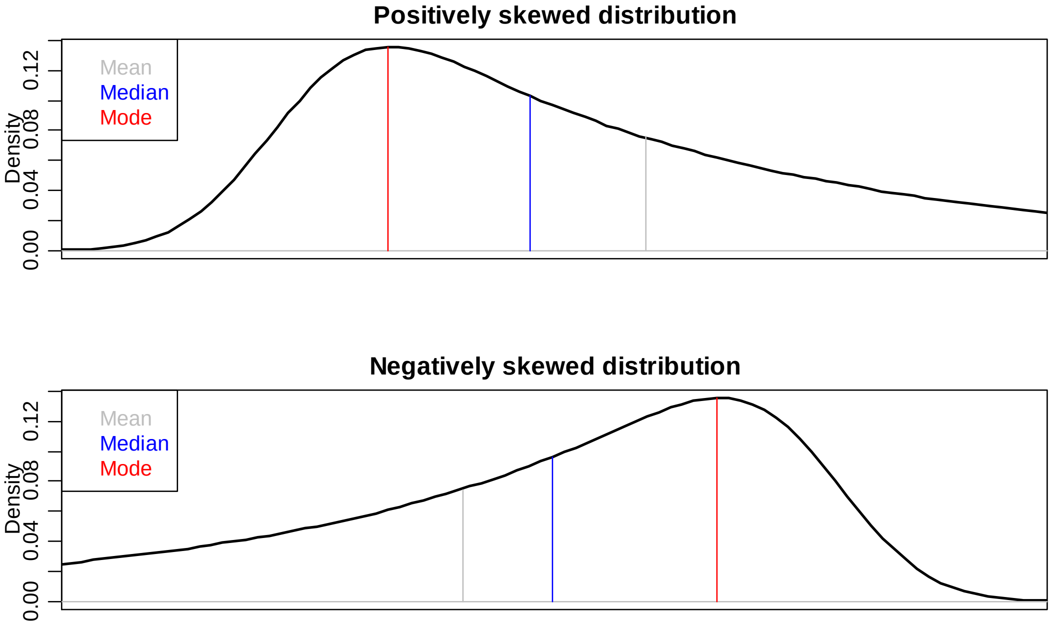
# IV. Measures of Central Tendency

**Positive skewness:** positive end longer, skewed to the right

mean > median > mode

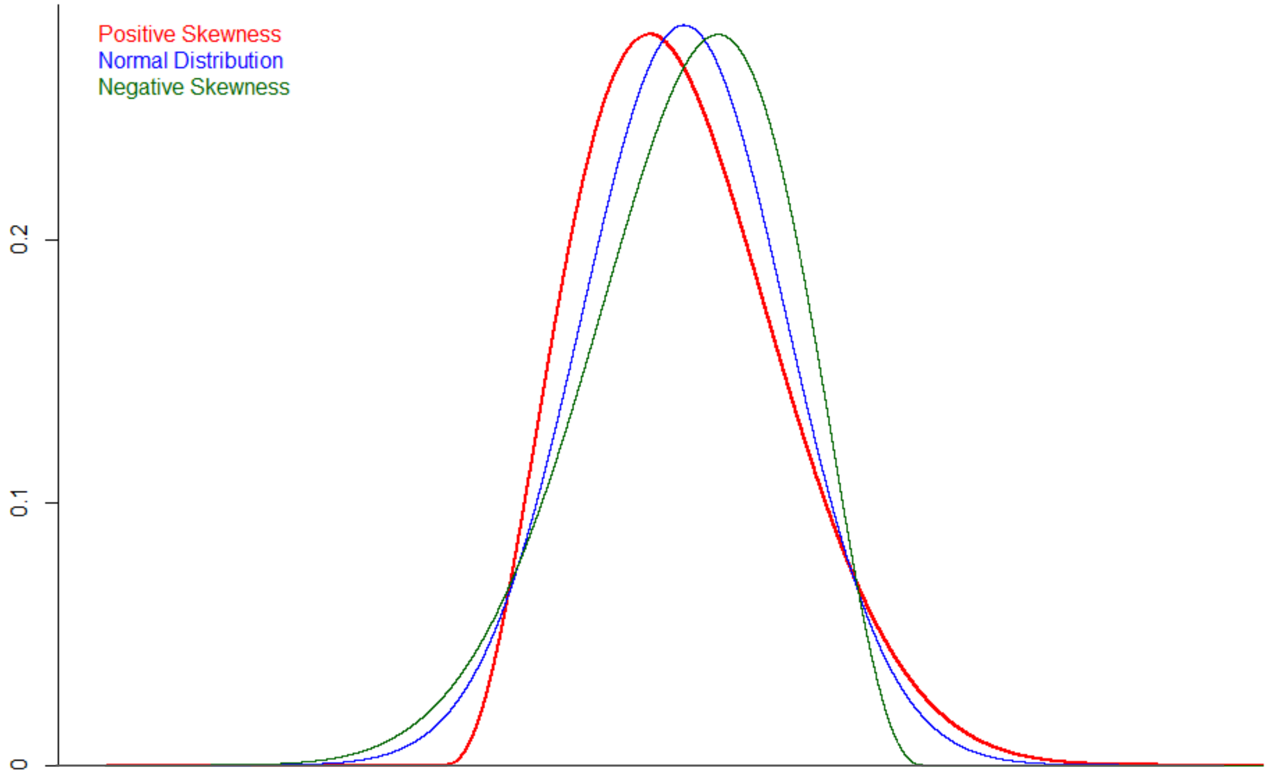
**Negative skewness:** negative end longer, skewed to the left

mean < median < mode



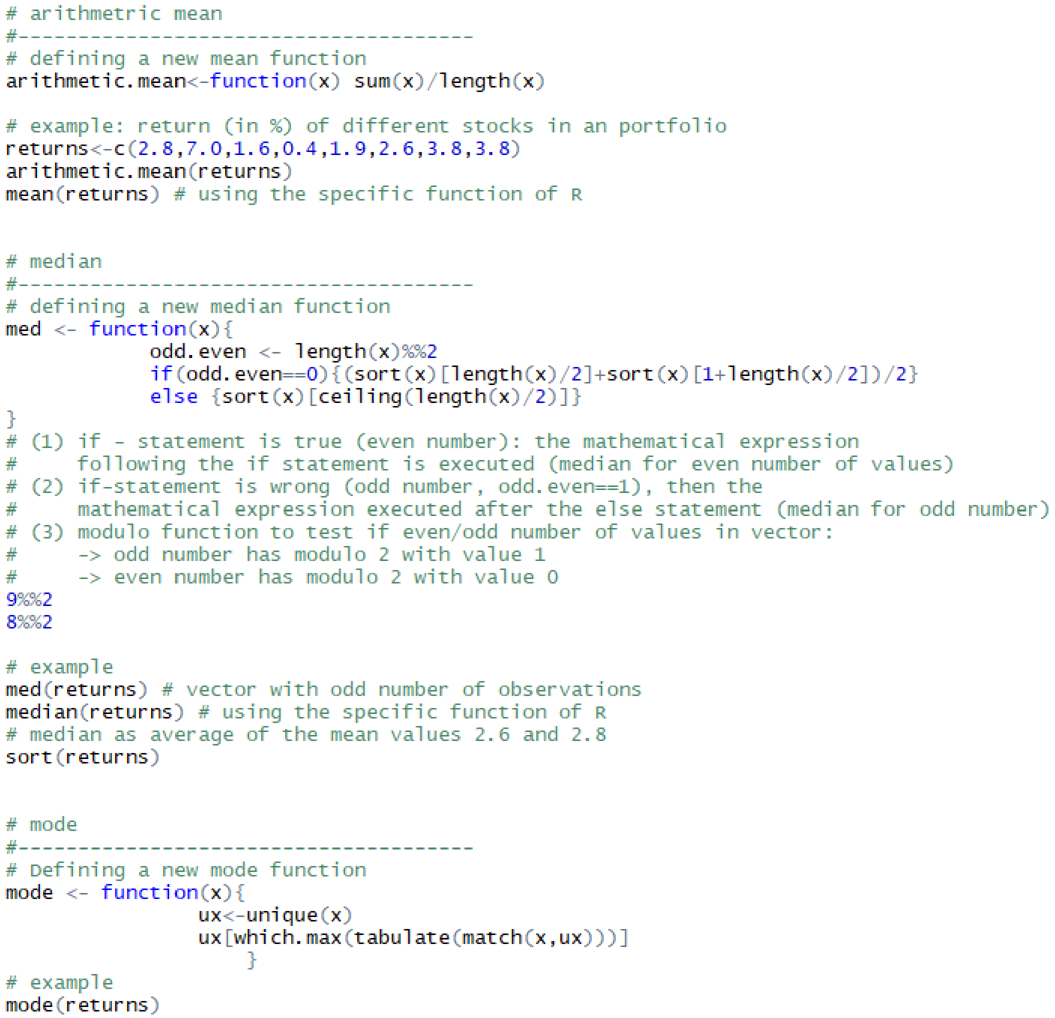
# IV. Measures of Central Tendency

**The comparison of** distributions with positive and negative skewness and a normal distribution shows different frequencies at the tails!



# IV. Position parameter (R-Example 2)

Open the file "VL1-Example_2.R" in R-Studio and reproduce the R-Code.



```
# arithmetic mean
#---------------------------------------
# defining a new mean function
arithmetic.mean<-function(x) sum(x)/length(x)

# example: return (in %) of different stocks in an portfolio
returns<-c(2.8,7.0,1.6,0.4,1.9,2.6,3.8,3.8)
arithmetic.mean(returns)
mean(returns) # using the specific function of R


# median
#---------------------------------------
# defining a new median function
med <- function(x){
            odd.even <- length(x)%%2
            if(odd.even==0){(sort(x)[length(x)/2]+sort(x)[1+length(x)/2])/2}
            else {sort(x)[ceiling(length(x)/2)]}
}
# (1) if - statement is true (even number): the mathematical expression
#     following the if statement is executed (median for even number of values)
# (2) if-statement is wrong (odd number, odd.even==1), then the
#     mathematical expression executed after the else statement (median for odd number)
# (3) modulo function to test if even/odd number of values in vector:
#     -> odd number has modulo 2 with value 1
#     -> even number has modulo 2 with value 0
9%%2
8%%2

# example
med(returns) # vector with odd number of observations
median(returns) # using the specific function of R
# median as average of the mean values 2.6 and 2.8
sort(returns)


# mode
#---------------------------------------
# Defining a new mode function
mode <- function(x){
                ux<-unique(x)
                ux[which.max(tabulate(match(x,ux)))]
                    }
# example
mode(returns)
```

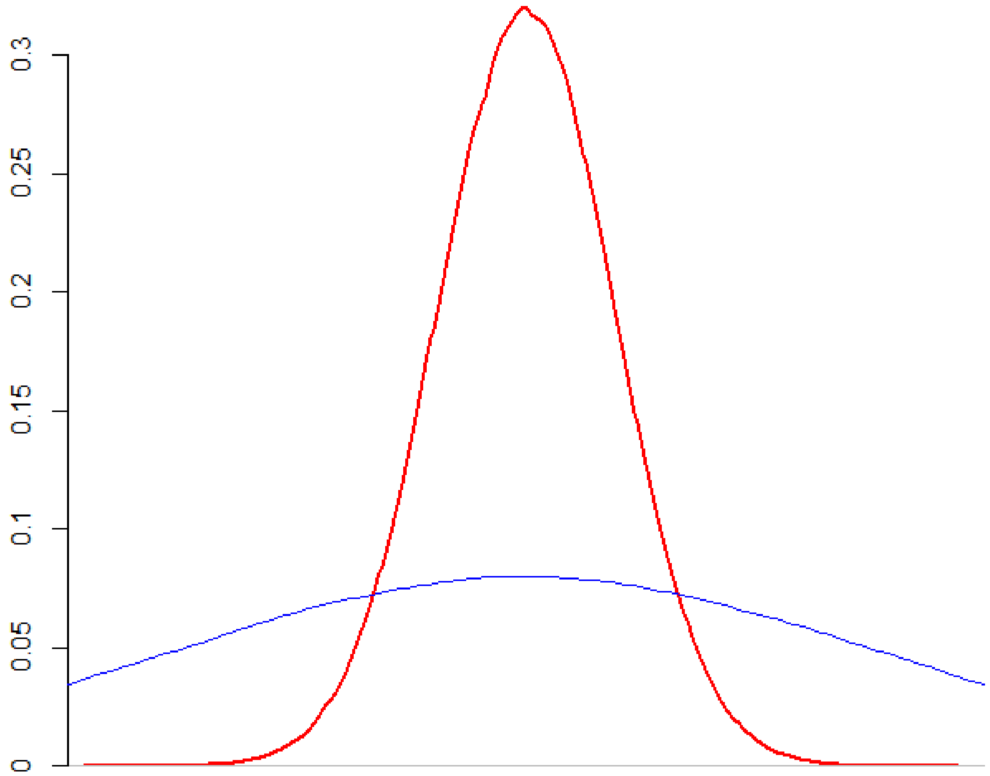
# V. Measures of Dispersion

**Measures of dispersion** characterize the spread of individual values around the center of the distribution. They play an essential role in many statistical tests. Important parameters include among others:

- variance
- standard deviation (square root of the variance)
- mean absolute deviation
- interquartile range
- range

# V. Measures of Dispersion

The distributions below have the same expected value. However, they show a **different dispersion**.



# V. Measures of Dispersion

## Variance and standard deviation

- Population variance $$\sigma^2 = \frac{\sum_{i=1}^{N} (X_i - \mu)^2}{N}$$

- Sample variance $$s^2 = \frac{\sum_{i=1}^{n} (x_i - \overline{x})^2}{n-1}$$

where *N* = number of observations of the population

*n* = number of sample observations

*X* = observations in the population

*x* = sample observation

- Standard deviation $\sigma = \sqrt{\sigma^2}$, or $s = \sqrt{s^2}$

# V. Measures of Dispersion (Example)

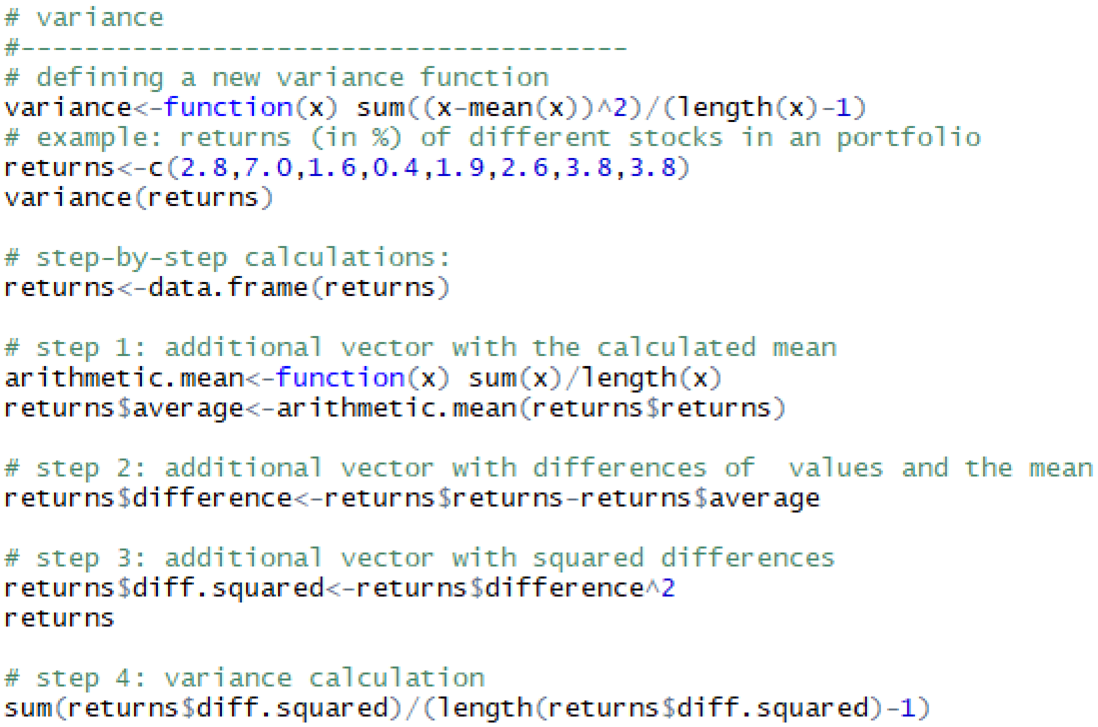
| ***Market i*** | **Price = $X_i$** | $\mu$ | $X_i - \mu$ | $(X_i - \mu)^2$ |
|---|---:|---:|---:|---:|
| GB | 21000 | 22000 | -1000 | 1 000 000 |
| F | 17000 | 22000 | -5000 | 25 000 000 |
| D | 19000 | 22000 | -3000 | 9 000 000 |
| J | 25000 | 22000 | 3000 | 9 000 000 |
| CH | 28000 | 22000 | 6000 | 36 000 000 |
| | | | | 80 000 000 |

Variance: $$\sigma^2 = \frac{80000000}{5} = 16000000$$

Standard deviation: $$\sigma = \sqrt{16000000} = 4000$$

# V. Measures of Dispersion (R-Example 3)

Open the file "L1-Example_3.R" in R-Studio and reproduce the R-Code.



```r
# variance
#--------------------------------------
# defining a new variance function
variance<-function(x) sum((x-mean(x))^2)/(length(x)-1)
# example: returns (in %) of different stocks in an portfolio
returns<-c(2.8,7.0,1.6,0.4,1.9,2.6,3.8,3.8)
variance(returns)

# step-by-step calculations:
returns<-data.frame(returns)

# step 1: additional vector with the calculated mean
arithmetic.mean<-function(x) sum(x)/length(x)
returns$average<-arithmetic.mean(returns$returns)

# step 2: additional vector with differences of  values and the mean
returns$difference<-returns$returns-returns$average

# step 3: additional vector with squared differences
returns$diff.squared<-returns$difference^2
returns

# step 4: variance calculation
sum(returns$diff.squared)/(length(returns$diff.squared)-1)
```

# V. Measures of Dispersion

**Chebyshev theorem:** The percentage of observations that lie **within *k* standard deviations symmetrically around the mean** will be at least:

$$\left(1-\frac{1}{k^{2}}\right)\cdot 100\%$$

**Intuition:** The smaller the standard deviation of the distribution, the greater the probability that individual observations are close to the mean value. The theorem specifies the minimum percentage of observations that fall within a given number of standard deviations from the mean.

# V. Measures of Dispersion (Example)

The mean and standard deviation of the distribution of a population (or sample) are $\mu$ = 34.6 and $\sigma$ = 49.3. The percentage of observations within *k* = *2* standard deviations around the mean is then 75%.

$$\left(1-\frac{1}{2^2}\right)\cdot 100\% = 75\%$$

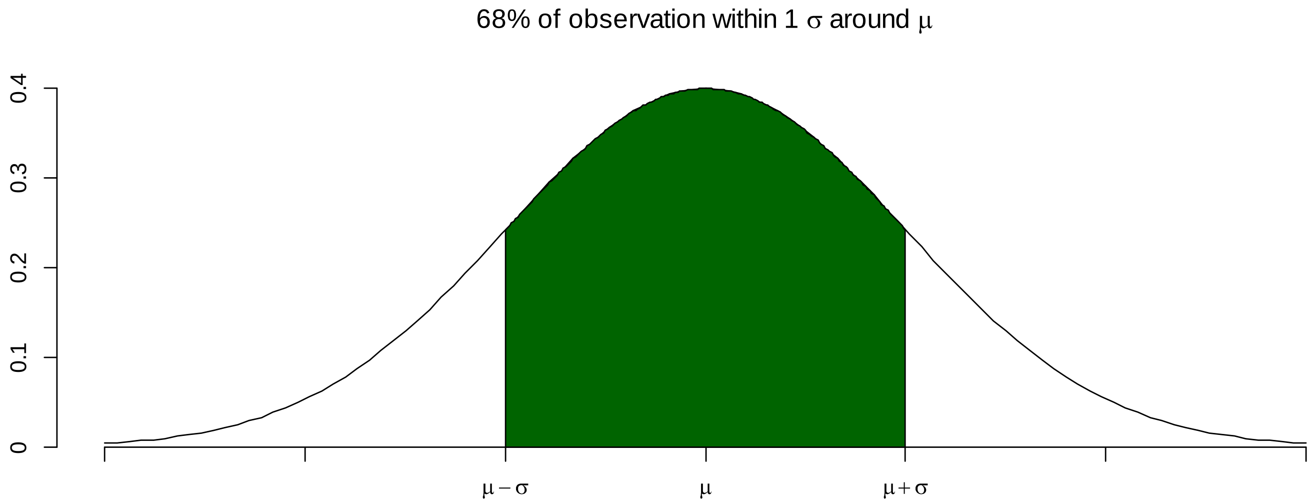
At least 75% of the observations are in the following interval:

$$34.6 \pm 2 \cdot 49.3$$

# V. Measures of Dispersion

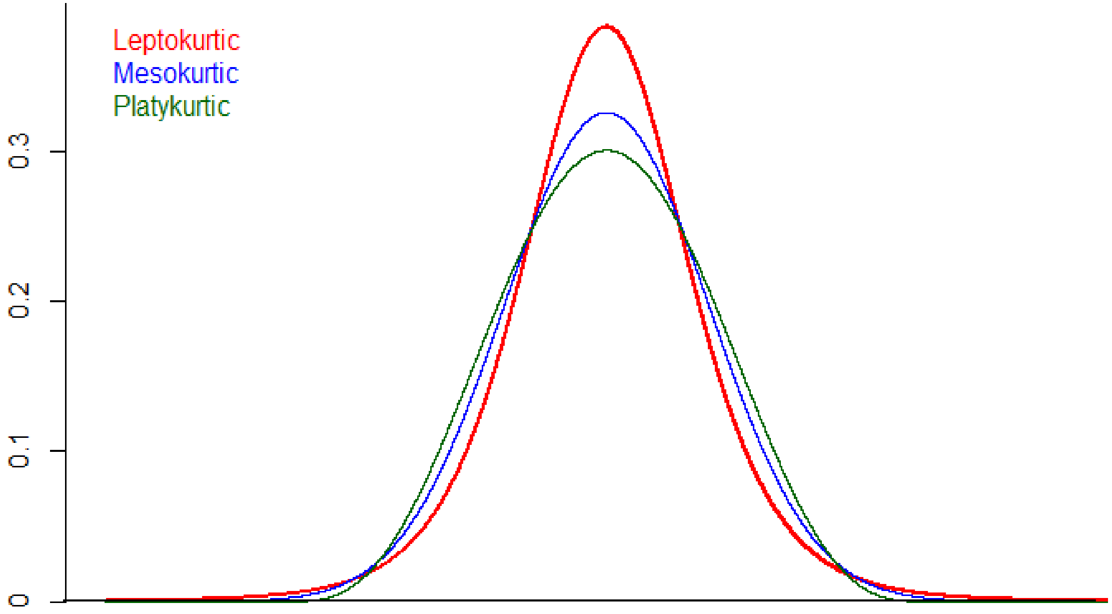
The Chebyshev theorem is valid for any distribution. However, the following rules of thumb apply only to **bell-shaped, symmetrical distributions***:*

- about **68%** of the observations fall within one SD around the mean value
- about **95%** of the observations fall within two SD around the mean value
- **almost all (99.7%)** observations fall within three SD around the mean value



# V. Measures of Dispersion

The **Kurtosis** expresses whether the distribution is rather narrow-pointed or broad-pointed compared to a normal distribution. **With the same standard deviation,** distributions can therefore concentrate differently around the mean.



Distribution 1: Mean=0, Variance=1.5, Skewness=0, Excess Kurtosis=3
Distribution 2: Mean=0, variance=1.5, Skewness=0, Excess Kurtosis=0
Distribution 3: Mean=0, Variance=1.5, Skewness=0, Excess Kurtosis=-0.5

# V. Measures of Dispersion

## Mean Absolute Deviation (MAD)

- MAD of the population

$$MAD = \frac{\sum_{i=1}^{N} \left| X_i - \mu \right|}{N}$$

- MAD of the sample

$$MAD = \frac{\sum_{i=1}^{n} \left| x_i - \overline{x} \right|}{n}$$

where *N* = number of observations of the population

*n* = number of sample observations

*X* = observations in the population

*x* = sample observations

# V. Measures of Dispersion (Example)

| ***Market i*** | **Price = $X_i$** | $\mu$ | $X_i - \mu$ | $\lvert X_i - \mu \rvert$ |
|---|---:|---:|---:|---:|
| GB | 21000 | 22000 | -1000 | 1000 |
| F | 17000 | 22000 | -5000 | 5000 |
| D | 19000 | 22000 | -3000 | 3000 |
| J | 25000 | 22000 | 3000 | 3000 |
| CH | 28000 | 22000 | 6000 | 6000 |
| | | | | 18000 |

Mean Absolute Deviation:

$$MAD = \frac{18000}{5} = 3600$$

# V. Measures of Dispersion

**Quartiles** $Q_1$, $Q_2$, $Q_3$ divide the (sorted) population **into four equally sized sub-populations** (i.e., the first 25% of the values are smaller than $\mathrm{Q}_1$, $Q_2$ corresponds to the median, 75% of the values are smaller than $Q_3$).

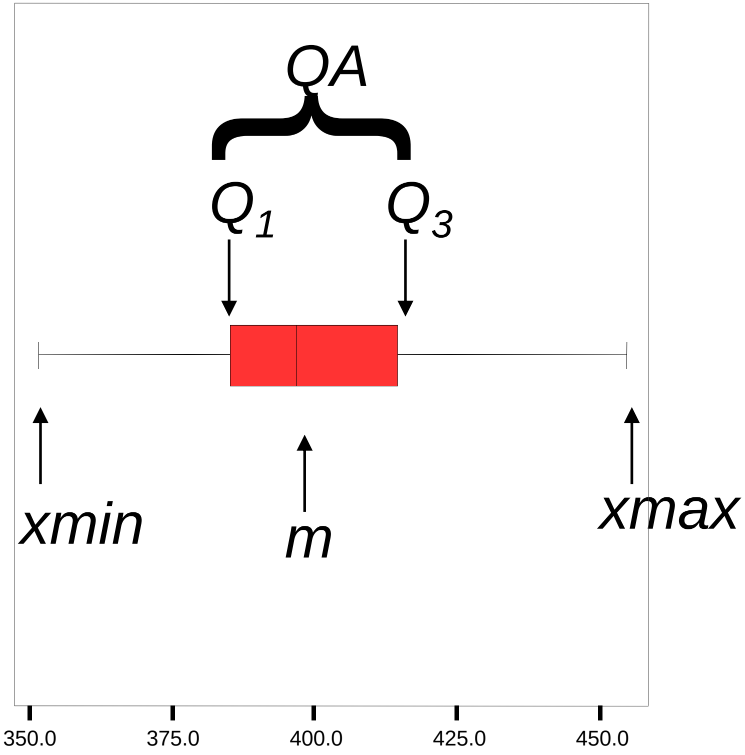
$$Q_1 = x_{\left[\frac{n+1}{4}\right]} \qquad Q_2 = x_{\left[\frac{2(n+1)}{4}\right]} \qquad Q_3 = x_{\left[\frac{3(n+1)}{4}\right]}$$

The **quartile distance** *(interquartile range) indicates* the range of the 50% mean values and is defined as follows: $IQR = Q_3 - Q_1$ . **The interquartile range is a measure of dispersion, but the individual quartiles represent measures of central tendency.**

# V. Box-Plots

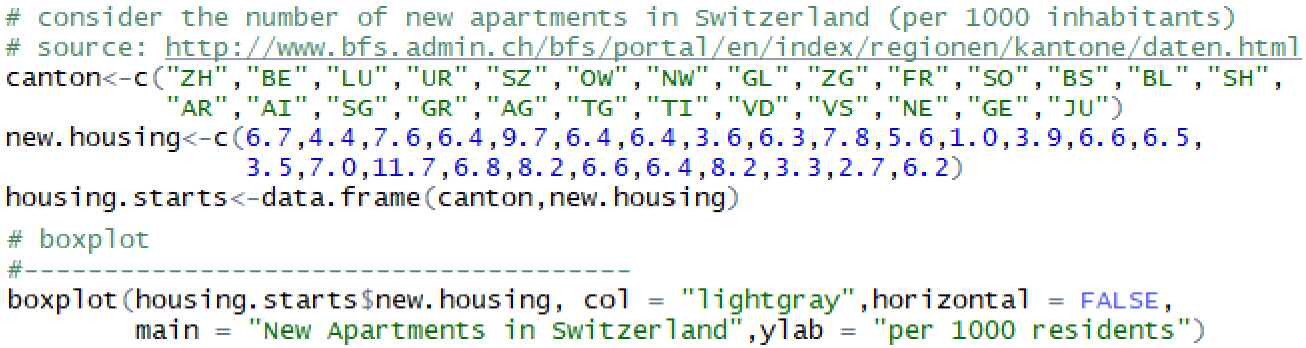
Alternatively, data sets can also be characterized with the help of ***box-plots.*** These representations are a graphical conversion of a **pentagram** (5-number measures) into a graphical distribution symbol. The components of a box plot are:

- A **scale** that is parallel to main axis of the plot.
- A **rectangle** (box) ranging from the 1st to the 3rd quartile.
- **Vertical dashes** indicating the median (*m*) and the extreme values ($x_{min}$, $x_{max}$).
- A **connection line** from the middle of the box to the vertical dashes of the extreme values

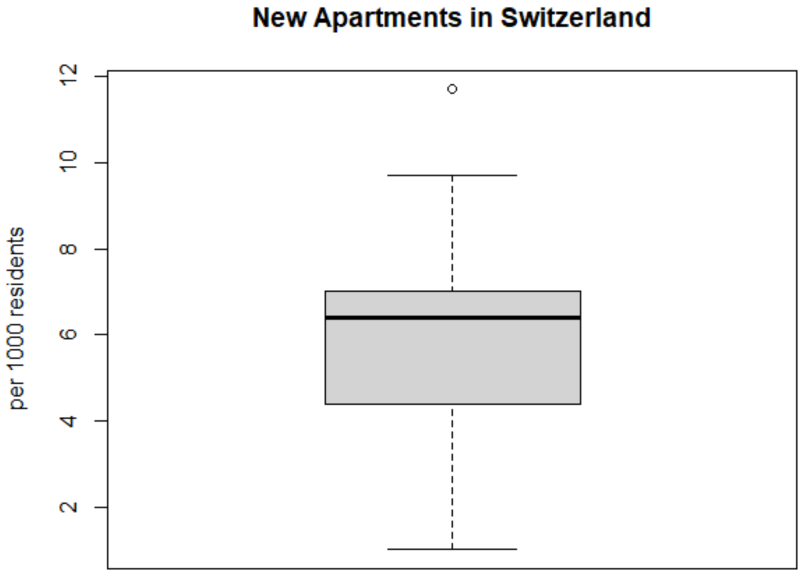


# V. Boxplots (R-Example 4)

Open the file "L1-Example_4.R" in R-Studio and reproduce the R-Code.



```
# consider the number of new apartments in Switzerland (per 1000 inhabitants)
# source: http://www.bfs.admin.ch/bfs/portal/en/index/regionen/kantone/daten.html
canton<-c("ZH","BE","LU","UR","SZ","OW","NW","GL","ZG","FR","SO","BS","BL","SH",
          "AR","AI","SG","GR","AG","TG","TI","VD","VS","NE","GE","JU")
new.housing<-c(6.7,4.4,7.6,6.4,9.7,6.4,6.4,3.6,6.3,7.8,5.6,1.0,3.9,6.6,6.5,
               3.5,7.0,11.7,6.8,8.2,6.6,6.4,8.2,3.3,2.7,6.2)
housing.starts<-data.frame(canton,new.housing)
# boxplot
#---------------------------------------
boxplot(housing.starts$new.housing, col = "lightgray",horizontal = FALSE,
        main = "New Apartments in Switzerland",ylab = "per 1000 residents")
```



# Appendix: Sampling Variance (1/2)

Given is a population with $\mu$ and variance $\sigma^2$ from which the observations $x_1$, …, $x_n$ are drawn randomly (with replacement). It follows that the sample variance is an **unbiased estimator** of the variance in the population (this implies E($s^2$) = $\sigma^2$).

$$
\begin{aligned}
\mathrm{E}(s^2) &= \mathrm{E}\left(\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\overline{x})^2\right) = \frac{1}{n-1}\mathrm{E}\left(\sum_{i=1}^{n}(x_i-\mu+\mu-\overline{x})^2\right) \\
&= \frac{1}{n-1}\mathrm{E}\left(\sum_{i=1}^{n}\left((x_i-\mu)^2-2(x_i-\mu)(\overline{x}-\mu)+(\overline{x}-\mu)^2\right)\right) \\
&= \frac{1}{n-1}\mathrm{E}\left(\sum_{i=1}^{n}(x_i-\mu)^2-2(\overline{x}-\mu)\sum_{i=1}^{n}(x_i-\mu)+n(\overline{x}-\mu)^2\right) \\
&= \frac{1}{n-1}\mathrm{E}\left(\sum_{i=1}^{n}(x_i-\mu)^2-2n(\overline{x}-\mu)(\overline{x}-\mu)+n(\overline{x}-\mu)^2\right) \\
&= \frac{1}{n-1}\mathrm{E}\left(\sum_{i=1}^{n}(x_i-\mu)^2-n(\overline{x}-\mu)^2\right) \\
&= \frac{1}{n-1}\left(\sum_{i=1}^{n}\mathrm{E}\left[(x_i-\mu)^2\right]-n\,\mathrm{E}\left[(\overline{x}-\mu)^2\right]\right)
\end{aligned}
$$

# Appendix: Sample Variance (2/2)

The terms in the last row can be substituted by the formulas for the definition of the variance and for the calculation of the standard error of the sample mean from the population variance and the sample size:

$$\mathrm{E}\left[(x_i-\mu)^2\right] = \sigma^2$$

$$\mathrm{E}\left[(\bar{x}-\mu)^2\right] = \sigma_{\bar{x}}^2 = \frac{\sigma^2}{n}$$

It follows:

$$\begin{aligned}\mathrm{E}(s^2) &= \frac{1}{n-1}\left(n\sigma^2 - n\frac{\sigma^2}{n}\right)\\ &= \frac{1}{n-1}\sigma^2(n-1) = \sigma^2\end{aligned}$$

Likewise, the sample standard deviation *s is* an **unbiased estimator** for $\sigma$.