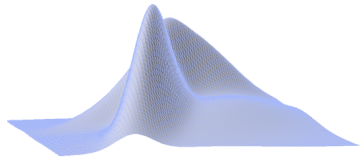


Wirtschafts- und Sozialstatistik
Universität Trier

Elements of Statistics

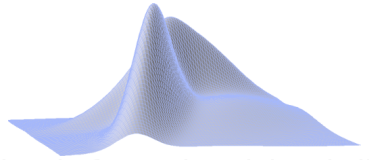
Chapter 3:
Measures of central tendency and measures of variation

Ralf Münnich, Jan Pablo Burgard and Florian Ertz

University of Trier
Faculty IV
Economic and Social Statistics Department

Winter term 2021/22

Statistical measures



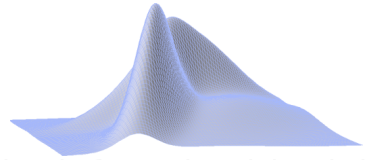
Wirtschafts- und Sozialstatistik
Universität Trier

Statistical measures are calculated in order to characterise distributions by means of suitable parameters. In this context, the scaling of the variables of interest plays a major role, as it determines the suitability of different measures.

A distinction is made between

- measures of central tendency,
- measures of variation and
- further measures to describe a distribution.

Properties of measures of central tendency



Wirtschafts- und Sozialstatistik
Universität Trier

Axiom of identity: If all values are identical, the measure of central tendency should adopt that same value.

Axiom of inclusion: The value of the measure of central tendency should be in the interval $[x_{\min}; x_{\max}]$.

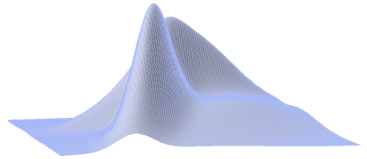
Axiom of translation: If all values are shifted by a common value, the value of the measure of central tendency should be shifted by this common value as well.

Axiom of homogeneity: If the frequencies of all m different values are (multiplicatively) changed by a common value in such a way that the relative frequencies stay constant, the value of the measure of central tendency should not change (homogeneity of degree zero).

In certain circumstances, restrictions regarding the scaling, like non-negativity, have to be respected as well.

(See **Assenmacher, W. (2010): Deskriptive Statistik, 4th edition, Springer**.)

The mode



Wirtschafts- und Sozialstatistik
Universität Trier

Let a variable of an arbitrary scaling be given. The mode x_M is the value which occurs most frequently.

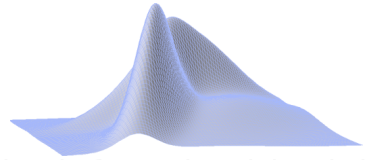
For $i : x_i = x_M$ the following holds:

$$n_i \geq n_j \qquad \forall\, j \neq i \quad .$$

Distributions which have exactly one mode are called unimodal.
In this case $n_i > n_j$ holds for all $j \neq i$.

In the context of continuous variables, we may call the mode the densest value.

The quantile



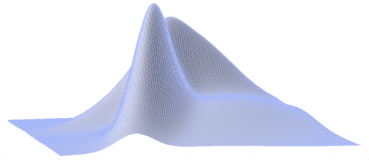
Wirtschafts- und Sozialstatistik
Universität Trier

The value x_p is called p-quantile if the following holds:

$$x_p = F^{-1}(p) := \inf\{x | F_n(x) \geq p\}$$

Remarks:

- $0 < p < 1$
- In this case, the quantile is defined by means of the inverse empirical distribution function.
- The definition may be broadened (symmetrical case; see Schaich and Münnich, 2001).
- The second *quartile* $x_{0.50}$ is called median (see below).
- $x_{0.25}$ is the first quartile and $x_{0.75}$ is the third quartile.



Example 3.1: Quantiles (1)

Given a sample of unsorted original values ($n = 10$):
5.7; 15.6; 12.6; 8.7; 11.9; 15.9; 1.6; 4.9; 19.8; 14.3

After reordering:
1.6; 4.9; 5.7; 8.7; 11.9; 12.6; 14.3; 15.6; 15.9; 19.8

Calculating the 0.2-quantile via the inverse distribution function:

$$x_{0.2} = x_{[2]} = 4.9 \ .$$

Calculating the first quartile (0.25-quantile) via the inverse distribution function:

$$x_{0.25} = x_{[3]} = 5.7 \ .$$

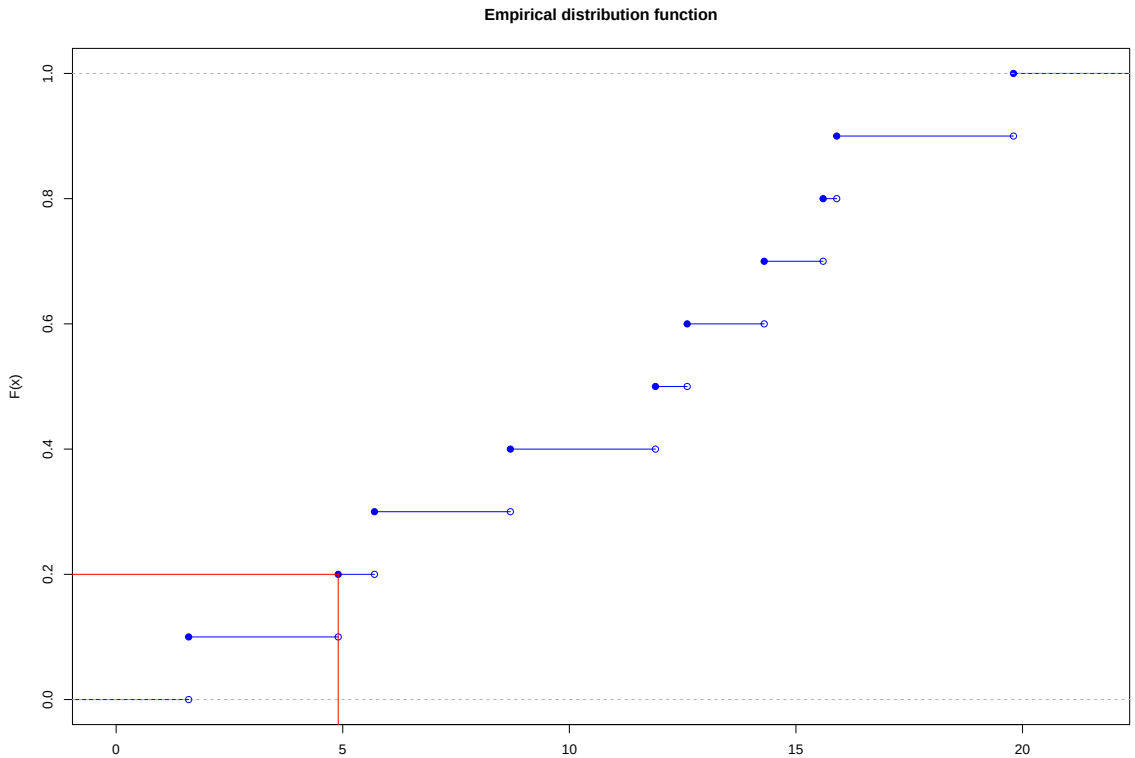
Data input in R:

```
x3_1 <- c(5.7, 15.6, 12.6, 8.7, 11.9, 15.9, 1.6,
          4.9, 19.8, 14.3)
```

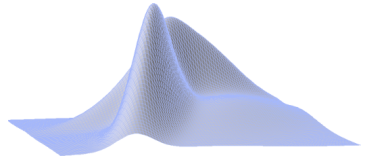
Example 3.1: Quantiles (2)

Graphical depiction of $x_{0,2}$:

Wirtschafts- und Sozialstatistik
Universität Trier

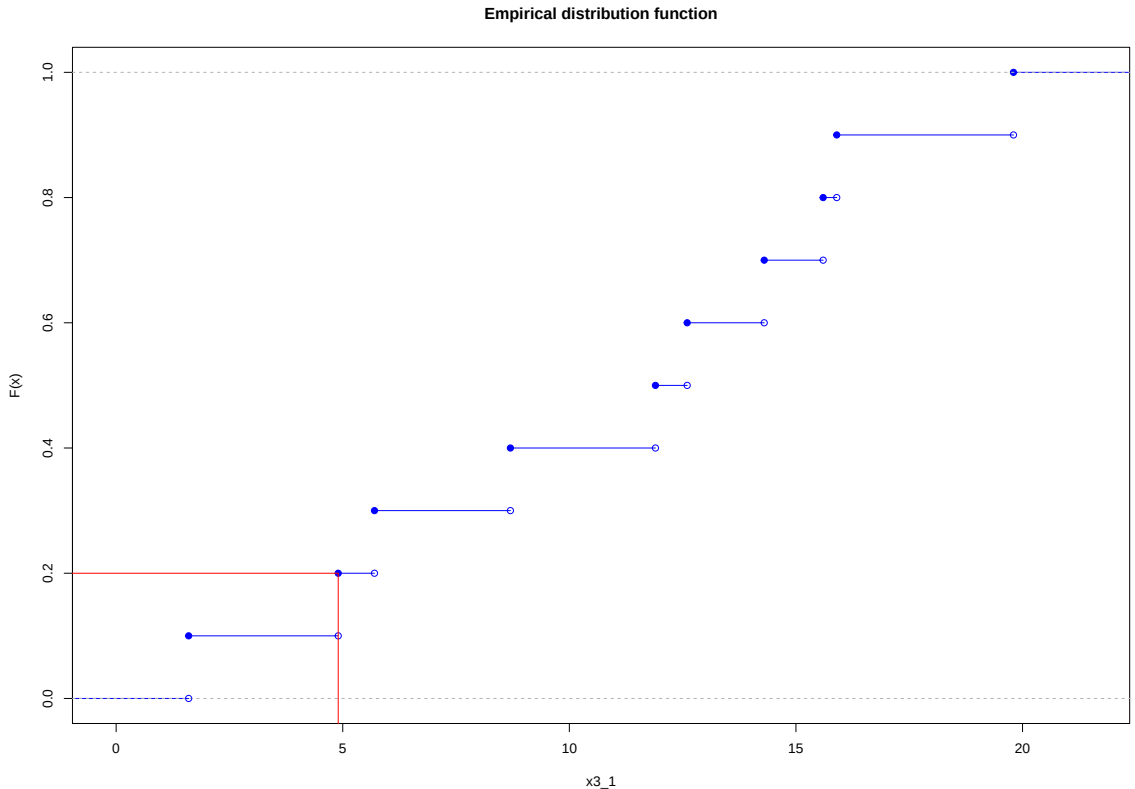



```
> plot(ecdf(sort(x3_1)), xlab = "x3_1", ylab = "F(x)",
+      col = "blue",
+      main = "Empirical distribution function")
> points(sort(x3_1), seq(from = 0,to = 0.9,by = 0.1),
+        col = "blue")
> lines(x = c(-1,4.9), y = c(0.2,0.2), col = "red")
> lines(x = c(4.9,4.9), y = c(0.2,-1), col = "red")
```



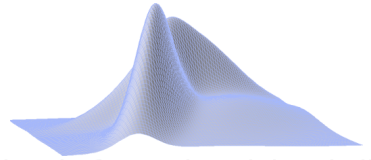
Example 3.1: Quantiles (3)

Graphical depiction of $x_{0,25}$:




```
# New plot like on the last slide
lines(x = c(0,5.7), y = c(0.25,0.25), col = "red")
lines(x = c(5.7,5.7), y = c(0,0.3), col = "red")
```

Example 3.1: Quantiles (4)



Wirtschafts- und Sozialstatistik
Universität Trier

Determination of $x_{0.2} = 4.9$ and $x_{0.25} = 5.7$ in R:

```
sort(x3_1)
```

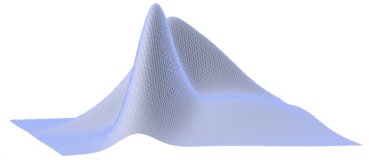
```
[1]  1.6  4.9  5.7  8.7 11.9 12.6 14.3 15.6 15.9 19.8
```

```
quantile(x = x3_1, probs = c(0.2 , 0.25), type = 1)
```

```
20% 25%
4.9 5.7
```

The determination of a quantile **with the inverse distribution function** is done in R with `quantile(x, type = 1)`.

The median



Wirtschafts- und Sozialstatistik
Universität Trier

Let a variable of at least ordinal scaling be given. Then the median

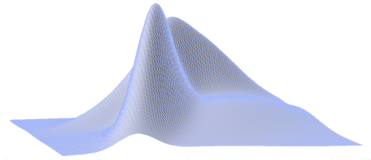
$$z := x_{0.5}$$

is the 0.5-quantile.

The median divides the smaller 50% from the larger 50% of values of a distribution.

Computation of the median:

- For uneven n: $x_{0.5} = x_{[(n+1)/2]}$
- For even n: $x_{0.5} = \frac{1}{2} \cdot (x_{[n/2]} + x_{[n/2+1]})$
  (In a strict sense, metric scaling would be needed here.)



Exemplary median computations (1)

Example 3.2:

Original values ($n = 9$):
13.1; 12.5; 8.3; 6.4; 9.1; 10.5; 10.8; 17.9; 22.3

Ordered values:
6.4; 8.3; 9.1; 10.5; 10.8; 12.5; 13.1; 17.9; 22.3

$$x_{0,5} = x_{[5]} = 10{,}8$$

Calculation in R:

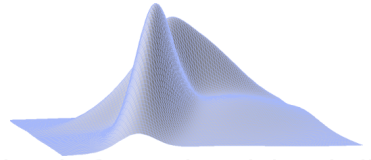
```
x3_2 <- c(13.1, 12.5, 8.3, 6.4, 9.1, 10.5,
          10.8, 17.9, 22.3)
sort(x3_2)
```

```
[1]  6.4  8.3  9.1 10.5 10.8 12.5 13.1 17.9 22.3
```

```
median(x3_2)
```

```
[1] 10.8
```

Exemplary median computations (2)



Example 3.3:

(Ordered) original values ($n = 10$):
6; 17; 22; 22; 23; 31; 34; 80; 90; 200

$$x_{0.5} = \frac{1}{2} \cdot \left(x_{[5]} + x_{[6]} \right) = 27$$

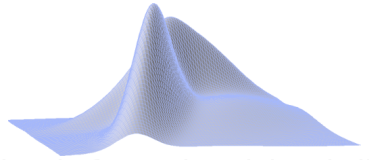
Calculation in R:

```
x3_3 <- c(6, 17, 22, 22, 23, 31, 34, 80, 90, 200)
sort(x3_3)
```

```
[1]   6  17  22  22  23  31  34  80  90 200
```

```
median(x3_3)
```

```
[1] 27
```



Determination of quantiles using grouped frequencies

Given grouped frequency distributions, the p-quantile is approximatively determined by using the empirical distribution function:

- j is the class for which $F(x^o_{j-1}) \leq p < F(x^o_j)$.
- Determination of quantile x_p by linear interpolation:

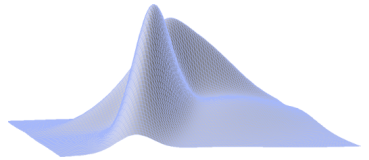
$$\frac{p - F(x^o_{j-1})}{F(x^o_j) - F(x^o_{j-1})} = \frac{x_p - x^o_{j-1}}{x^o_j - x^o_{j-1}}$$

$\Leftrightarrow$

$$x_p = x^o_{j-1} + (x^o_j - x^o_{j-1}) \cdot \frac{p - F(x^o_{j-1})}{F(x^o_j) - F(x^o_{j-1})}$$

Remark:
Values within the classes are assumed to be *uniformly* distributed.

Example 3.4: see Ex. 2.6 (1)



Remember:

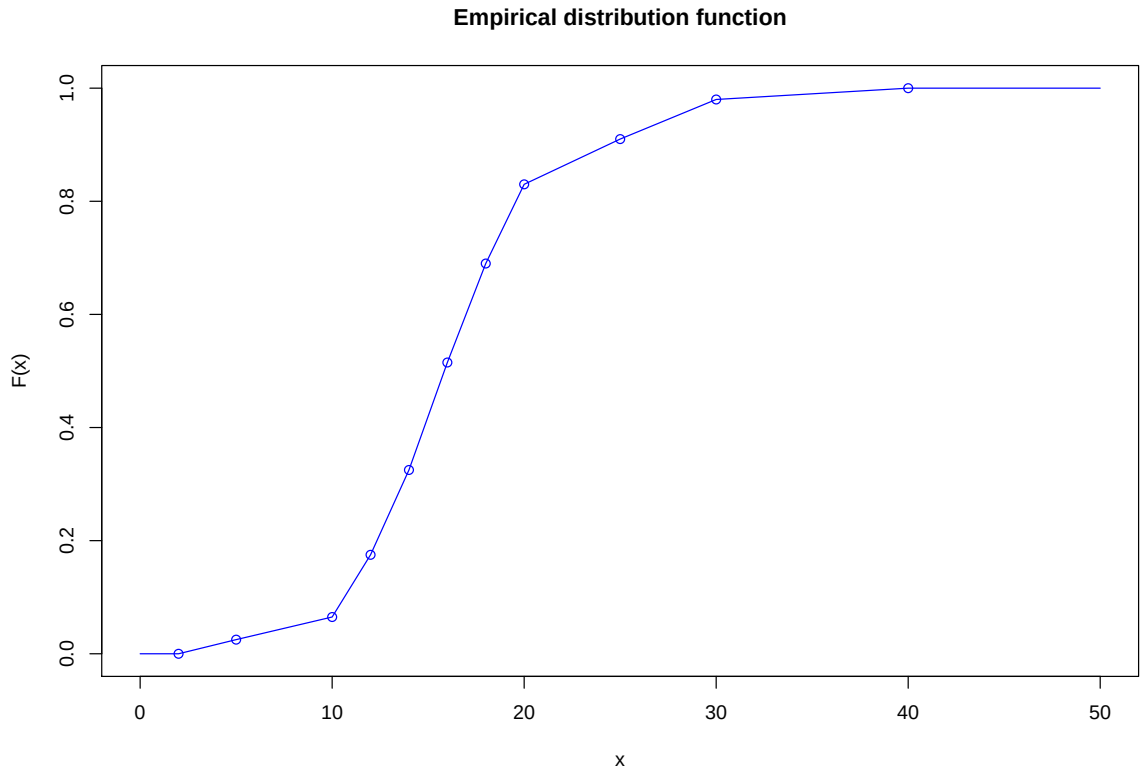
For $n = 200$ pupils, the time taken to solve a problem was recorded:

class (min.)			j	n_j	p_j
2	up to, but less than	5	1	5	0.025
5	up to, but less than	10	2	8	0.040
10	up to, but less than	12	3	22	0.110
12	up to, but less than	14	4	30	0.150
14	up to, but less than	16	5	38	0.190
16	up to, but less than	18	6	35	0.175
18	up to, but less than	20	7	28	0.140
20	up to, but less than	25	8	16	0.080
25	up to, but less than	30	9	14	0.070
30	up to, but less than	40	10	4	0.020
$\sum$				200	1.000

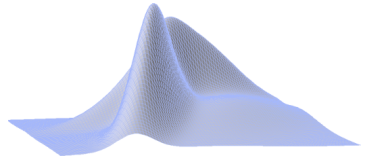
Which value will the median z have?

Example 3.4: see Ex. 2.6 (2)

Wirtschafts- und Sozialstatistik
Universität Trier

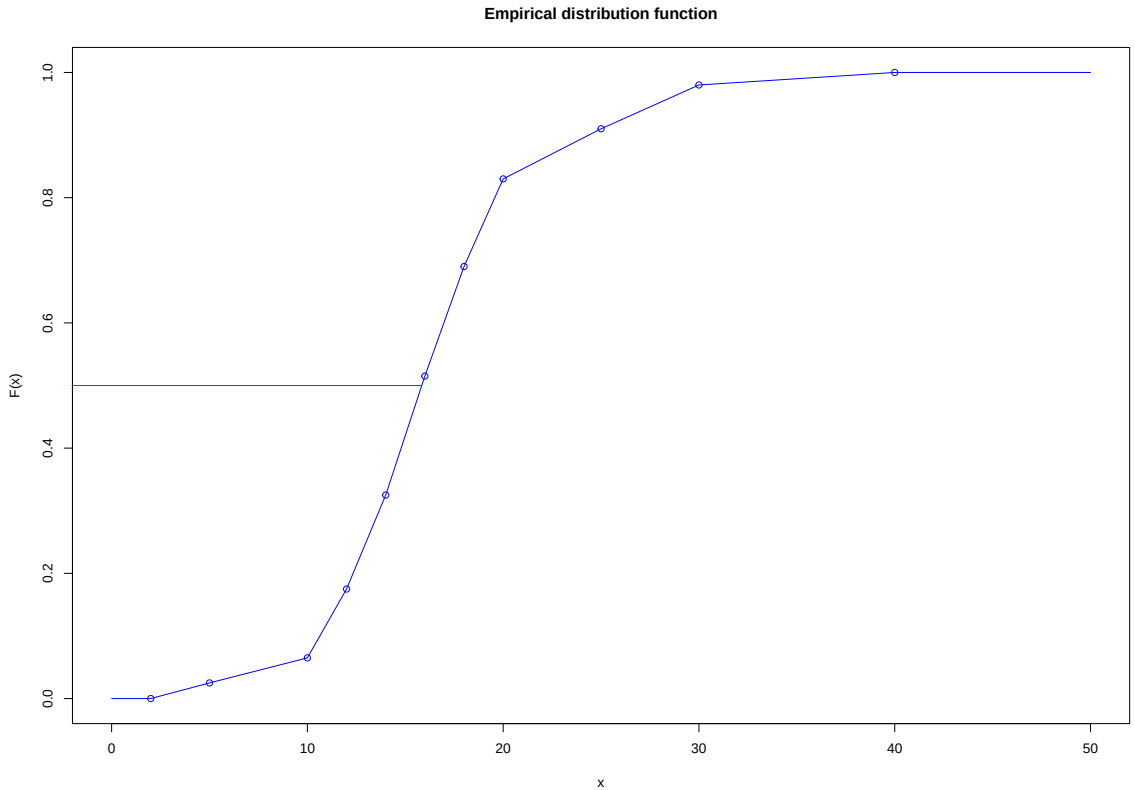



```
setwd("path"); load("Example3-4.RData") # your path
F_j <- cumsum(x = p_j); plot(x = c(0,0), type = "n",
  main = "Empirical distribution function",
  ylab = "F(x)", xlab = "x", xlim = c(0,50), ylim = c(0,1))
lines(x = c(0,x_o,50), y = c(0,0,F_j,1), col = "blue")
points(x = x_o, y = c(0,F_j), col = "blue")
```



Wirtschafts- und Sozialstatistik
Universität Trier

Example 3.4: see Ex. 2.6 (3)

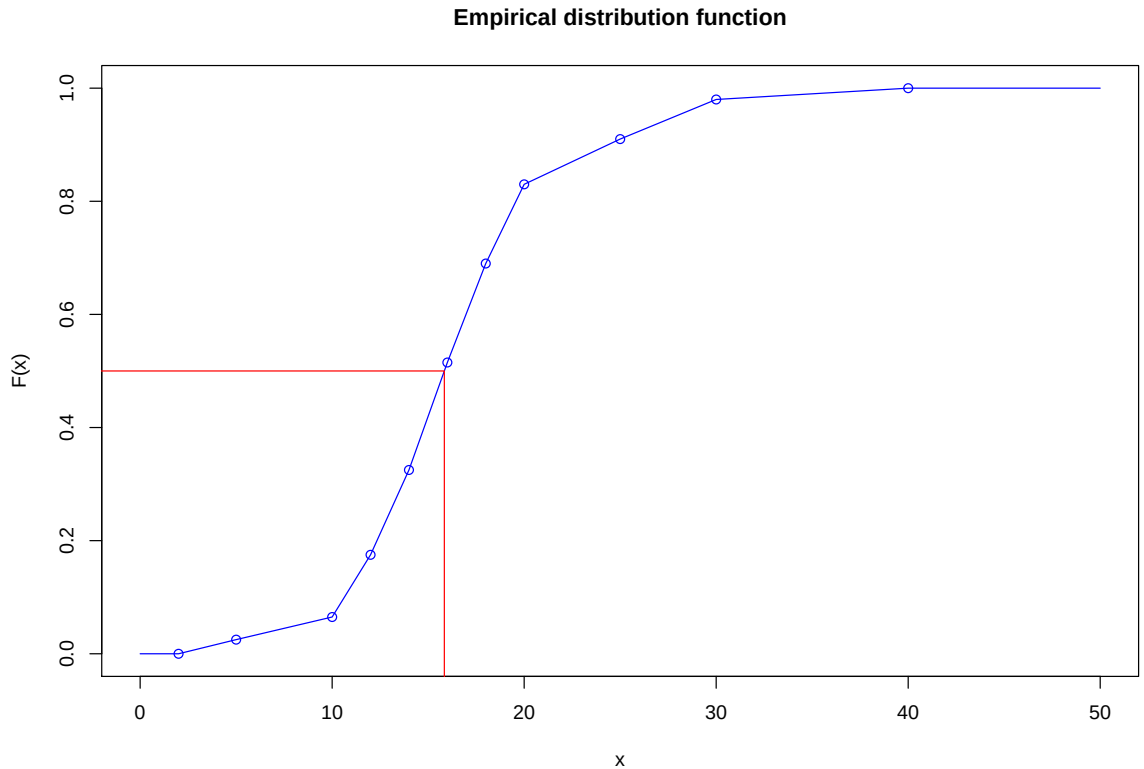


The median falls into class $j = 5$.

Graphical depiction in R:

```
lines(x = c(-5,15.8421), y = c(0.5,0.5), col = "red")
```

Example 3.4: see Ex. 2.6 (4)



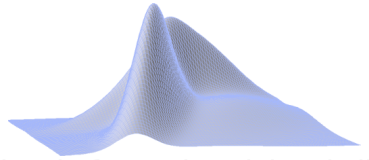
Finally, we reach

$$z = 14 + (16 - 14) \cdot \frac{0.5 - 0.325}{0.515 - 0.325} = 15.8421.$$

Graphical depiction in R:

```
lines(x = c(15.8421,15.8421), y = c(0.5,-5), col = "red")
```

Example 3.4: see Ex. 2.6 (5)



Wirtschafts- und Sozialstatistik
Universität Trier

Calculation of $z = 15.8421$ in R:

```
x_o
```

```
[1]  2  5 10 12 14 16 18 20 25 30 40
```

```
length(x_o)
```

```
[1] 11
```

```
F_j
```

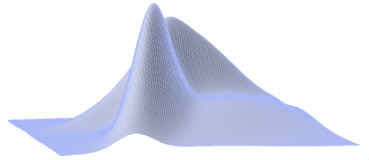
```
[1] 0.025 0.065 0.175 0.325 0.515 0.690 0.830 0.910 0.980 1.000
```

```
length(F_j)
```

```
[1] 10
```

```
j <- 5
x_median <- x_o[j] + (x_o[j+1] - x_o[j]) *
              (0.5 - F_j[j-1])/(F_j[j] - F_j[j-1])
round(x_median, 4)
```

```
[1] 15.8421
```



The arithmetic mean

Let a variable of metric scaling with n original values be given. Then

$$\overline{x} = \frac{1}{n} \cdot \sum_{i=1}^{n} x_i$$

is the arithmetic mean.

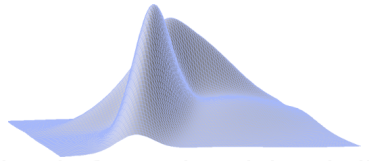
Example 3.5: see Ex. 3.2

$$\overline{x} = \frac{1}{9} \cdot (13.1 + 12.5 + \cdots + 22.3) = \frac{1}{9} \cdot 110.9 \approx 12.32$$

Calculation in `R`:

```
mean(x3_2)
```

```
[1] 12.32222
```



The arithmetic mean for grouped data

The arithmetic mean can also be computed as a weighted mean of the means in the different classes:

$$\overline{x} = \frac{1}{n} \cdot \sum_{j=1}^{m} n_j \cdot \overline{x}_j = \sum_{j=1}^{m} p_j \cdot \overline{x}_j \quad .$$

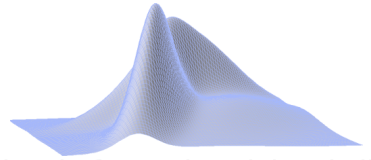
It holds that

$$\overline{x} = \frac{1}{n} \cdot \sum_{j=1}^{m} n_j \cdot \underbrace{\left(\frac{1}{n_j} \sum_{\nu=1}^{n_j} x_{\nu j} \right)}_{\overline{x}_j} = \frac{1}{n} \cdot \underbrace{\sum_{j=1}^{m} \sum_{\nu=1}^{n_j} x_{\nu j}}_{\text{Overall sum of values}} \quad .$$

If no information on class means ($\overline{x}_j$) is available, the arithmetic mean may still be determined by using the approximations $\overline{x}_j \approx x'_j$ and

$$\overline{x}' = \sum_{j=1}^{m} p_j \cdot x'_j \quad .$$

Example 3.6: see Ex. 2.5 (1)



Wirtschafts- und Sozialstatistik
Universität Trier

For the following classification we get:

class	from	up to, but less than	n_j	$\overline{x}_j$	x'_j
1	0	1500	5	1000.00	750
2	1500	3000	7	2000.00	2250
3	3000	4500	3	3266.67	3750
4	4500	6000	5	5040.00	5250

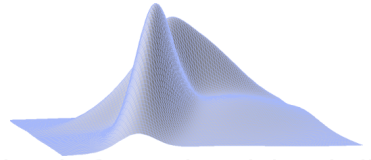
The arithmetic mean of the originalvalues is $\overline{x} = 2700$.

Calculation in R:

```
x2_5 <- c(3500,3200,2100,500,1800,2100,5600,4500,1400,1200,
          1500,2200,3100,1500,2800,1100,5200,4500,5400,800)
mean(x2_5)
```

```
[1] 2700
```

Example 3.6: see Ex. 2.5 (2)



The same value results when using the grouped data

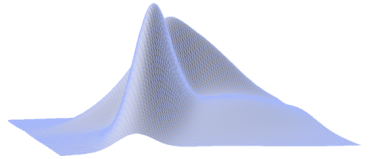
$$\overline{x} = \frac{5}{20} \cdot 1000 + \frac{7}{20} \cdot 2000 + \frac{3}{20} \cdot 3266.67 + \frac{5}{20} \cdot 5040 = 2700 \quad .$$

Calculation in R:

```
x_o <- c(0,1500,3000,4500,6000)
x2_5_kl <- cut(x2_5, x_o, right = FALSE)
n_j <- table(x2_5_kl)

x_mean_j <- tapply(X = x2_5, INDEX = x2_5_kl, FUN = mean)

x_mean_kl <- sum(n_j/sum(n_j) * x_mean_j)

x_mean_kl
```

```
[1] 2700
```

Example 3.6: see Ex. 2.5 (3)



Wirtschafts- und Sozialstatistik
Universität Trier

If the arithmetic means of the classes were not available we would get

$$\overline{x}' = \frac{5}{20} \cdot 750 + \frac{7}{20} \cdot 2250 + \frac{3}{20} \cdot 3750 + \frac{5}{20} \cdot 5250 = 2850 \quad .$$

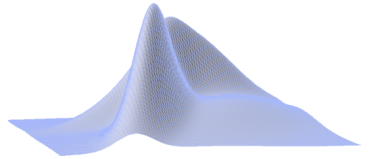
Calculation in R:

```
x_mean_approx_j <- x_o[-5]+(x_o[-1]-x_o[-5])/2
x_mean_aprrox_kl <- sum(n_j/sum(n_j) * x_mean_approx_j)
x_mean_aprrox_kl
```

```
[1] 2850
```

What happens if we approximate and have *over .. up to* classes?

The geometric mean



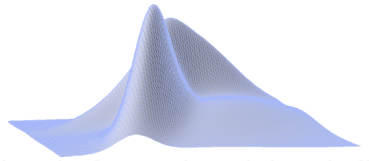
Wirtschafts- und Sozialstatistik
Universität Trier

Let a variable with strictly positive values and exhibiting a ratio scale be given ($x_i > 0; i = 1, \ldots, n$). Then

$$g = \sqrt[n]{\prod_{i=1}^{n} x_i} = \left(\prod_{i=1}^{n} x_i \right)^{\frac{1}{n}}$$

is the geometric mean.

Example 3.7: GDP growth



The growth of EU-25 GDP (in %) for the years from 1996 through 2011 is given in the New Cronos database as follows:

1.8; 2.8; 3.0; 3.0; 3.9; 2.1; 1.3; 1.4; 2.5; 2.1; 3.3; 3.2; 0.3; -4.3; 2.1; 1,5.

```
x3_7 <- c(1.8,2.8,3.0,3.0,3.9,2.1,1.3,1.4,2.5,2.1,3.3,3.2,
          0.3,-4.3,2.1,1.5)
```

By transitioning to growth factors, the mean growth is calculated as

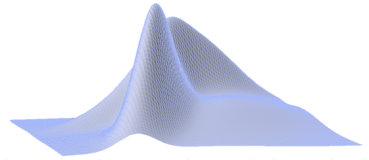
$$g = \sqrt[16]{1.018 \cdot 1.028 \cdot \dots \cdot 1.015} = \sqrt[16]{1.342574} = 1.018582 \quad .$$

```
wa <- x3_7/100 + 1
x_geom <- prod(wa)^(1/length(wa))
x_geom
```

```
[1] 1.018582
```

The use of the arithmetic mean would have yielded $\overline{x} = 1.01875$.
After 16 years an overall growth of 34.2574% would have to be reported instead of 34.6114%.

Example 3.8: German population (1)



Wirtschafts- und Sozialstatistik
Universität Trier

Population figures for the FRG (in thousands) and the years from 1995 through 2005 are available in the New Cronos database as well:

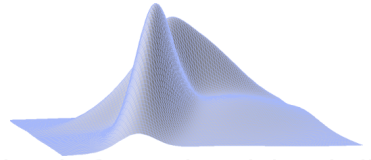
81,538.6; 81,817.5; 82,012.2; 82,057.4; 82,037.0; 82,163.5; 82,259.5; 82,440.3; 82,536.7; 82.531.7; 82,500.8.

Official DESTATIS figures for 2006 – 2011: 82,314.9; 82,217.8; 82,002.4; 81,802.3; 81,751.6; 81,843.7.

Find the mean of the yearly population growth rate for the FRG in the time span from 1995 until 2005.

Data input in R:

```
x3_8 <- c(81538.6, 81817.5, 82012.2, 82057.4, 82037,82163.5,
          82259.5, 82440.3, 82536.7, 82531.7, 82500.8)
```



Example 3.8: German population (2)

First, the 10 growth factors x_t/x_{t-1} have to be determined.

```
wa <- x3_8[-1] / x3_8[-length(x3_8)]
```

The corresponding geometric mean is

$$g = \sqrt[10]{\frac{81,817.5}{81,538.6} \cdot \frac{82,012.2}{81,817.5} \cdot \ldots \cdot \frac{82,500.8}{82,531.7}} = \sqrt[10]{1.011801} = 1.001174 \quad .$$

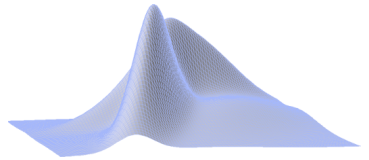
```
x_geom <- prod(wa)^(1/length(wa))
x_geom
```

```
[1] 1.001174
```

It follows that $81{,}538{,}600 \cdot g^{10} = 82{,}500{,}800$.
This matches the actual population in 2005.

Using the arithmetic mean, we would get
$81,538,600 \cdot 1.001175^{10} = 82,501,380$ after 10 years, thus 580 more than actually existing.

The harmonic mean



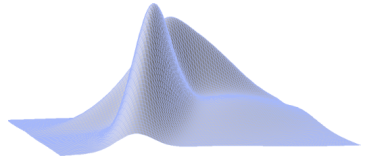
Wirtschafts- und Sozialstatistik
Universität Trier

Let a variable with strictly positive values and exhibiting a ratio scale be given $(x_i > 0; i = 1, \ldots, n)$. Then

$$h = \frac{1}{\frac{1}{n} \cdot \sum_{i=1}^{n} \frac{1}{x_i}} = \frac{n}{\sum_{i=1}^{n} \frac{1}{x_i}}$$

is the harmonic mean.

Example 3.9: New tires (1)



It takes a master mechanic 8 minutes to mount a new set of tires to a car. His apprentice completes this task in 12 minutes. How long does it take the two of them *together* to mount 100 sets of tires?

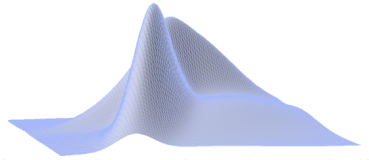
$$h = \frac{1}{\frac{1}{2} \cdot \left(\frac{1}{8} + \frac{1}{12} \right)} = \frac{48}{5} \quad \left[\frac{\text{min}}{\text{set}} \right]$$

Calculation in R:

```
x3_9 <- c(8,12)
n <- length(x3_9)

x_harm <- n / sum(1 / x3_9)
x_harm
```

```
[1] 9.6
```

Example 3.9: New tires (2)



Wirtschafts- und Sozialstatistik
Universität Trier

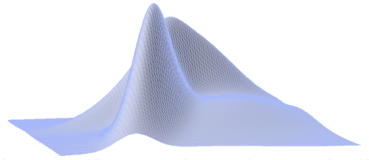
Hence, it takes *each* mechanic on average 48 minutes to mount 5 sets of tires. To mount 100 sets, both of them together need

$$100 \text{ [sets]} \cdot \frac{24}{5} \left[\frac{\text{min}}{\text{set}}\right] = 480 \text{ [min]} \quad ,$$

filling 8 hours.

Using the arithmetic mean, the mounting of one set would have taken 10 minutes per mechanic, resulting in 8 hours and 20 minutes.

Comparison of arithmetic and harmonic mean



Wirtschafts- und Sozialstatistik
Universität Trier

Find the respective average speeds for the following two cases (v_1 and v_2):

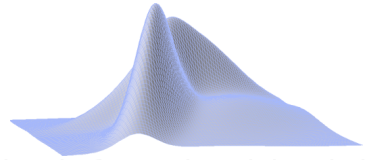
- A car drives at 100 km/h for one hour and at 120 km/h for three hours.
  We use the weighted arithmetic mean:

$$v_1 = \frac{1}{4} \cdot 100 + \frac{3}{4} \cdot 120 = 115 \qquad [km/h]$$

- A car drives at 100 km/h for 115 km and at 120 km/h for 345 km.
  We use the weighted harmonic mean:

$$v_2 = \frac{1}{\frac{115}{460} \cdot \frac{1}{100} + \frac{345}{460} \cdot \frac{1}{120}} = 114.2857 \qquad [km/h]$$

The second car will take 90 seconds longer.



Influence of changes in the unit of measurement

The observations x_i $(i = 1, \ldots, n)$ are transformed linearly:

$$y_i = a \cdot x_i + b \quad , \quad a \neq 0, b \in \mathbb{R}\,.$$

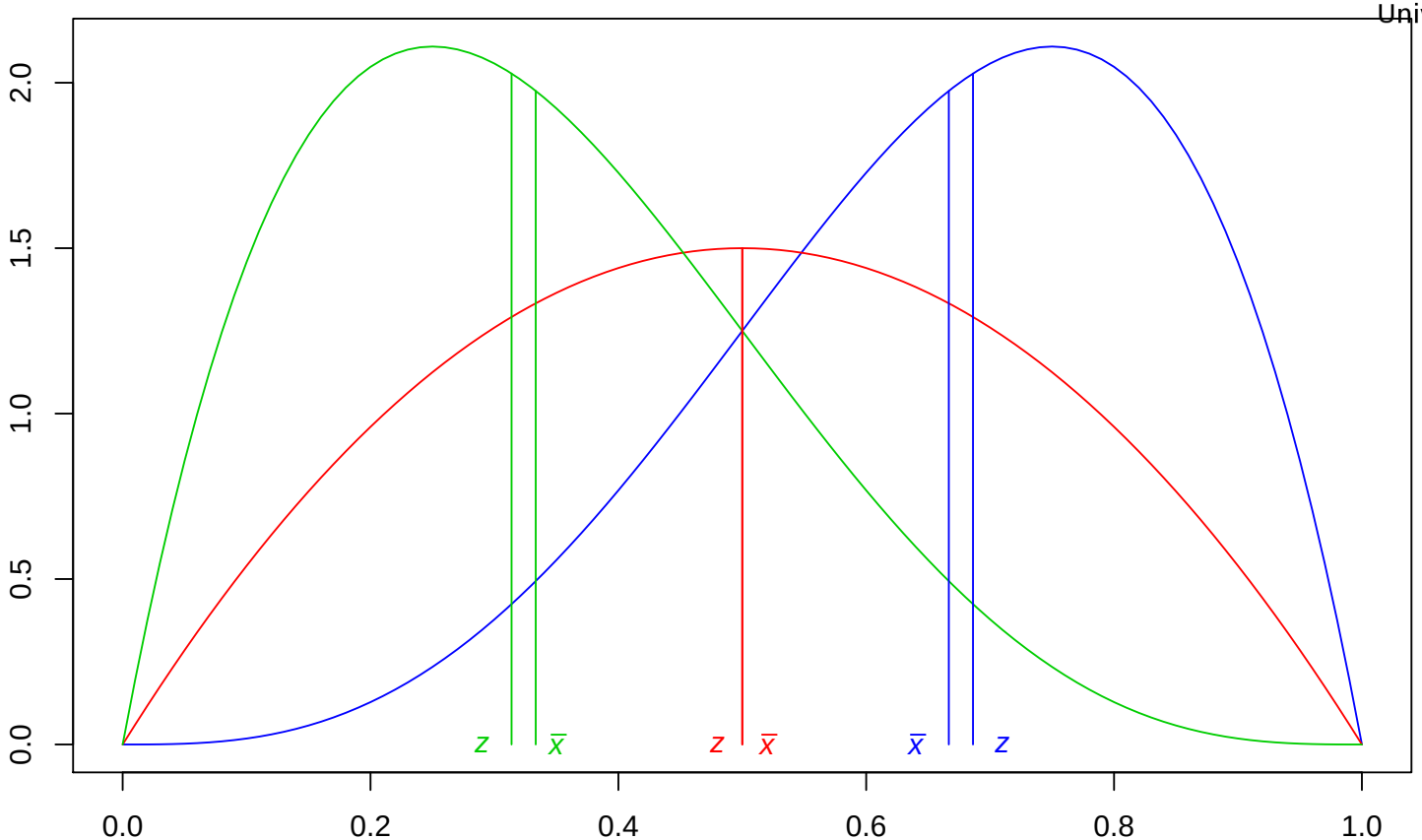
It follows that:

$$y_{\mathrm{M}} = a \cdot x_{\mathrm{M}} + b$$
$$z_y = a \cdot z_x + b$$
$$\overline{y} = a \cdot \overline{x} + b$$

g and h only satisfy this transformation for $b = 0$. Therefore, the axiom of translation does not hold. For $b = 0$ it follows that:

$$g_y = a \cdot g_x$$
$$h_y = a \cdot h_x$$

Comparison of distributions (1)



Symmetric distribution $\overline{x} \approx z_x \approx x_M$

Right-skewed (positively skewed) distribution $\overline{x} > z_x > x_M$

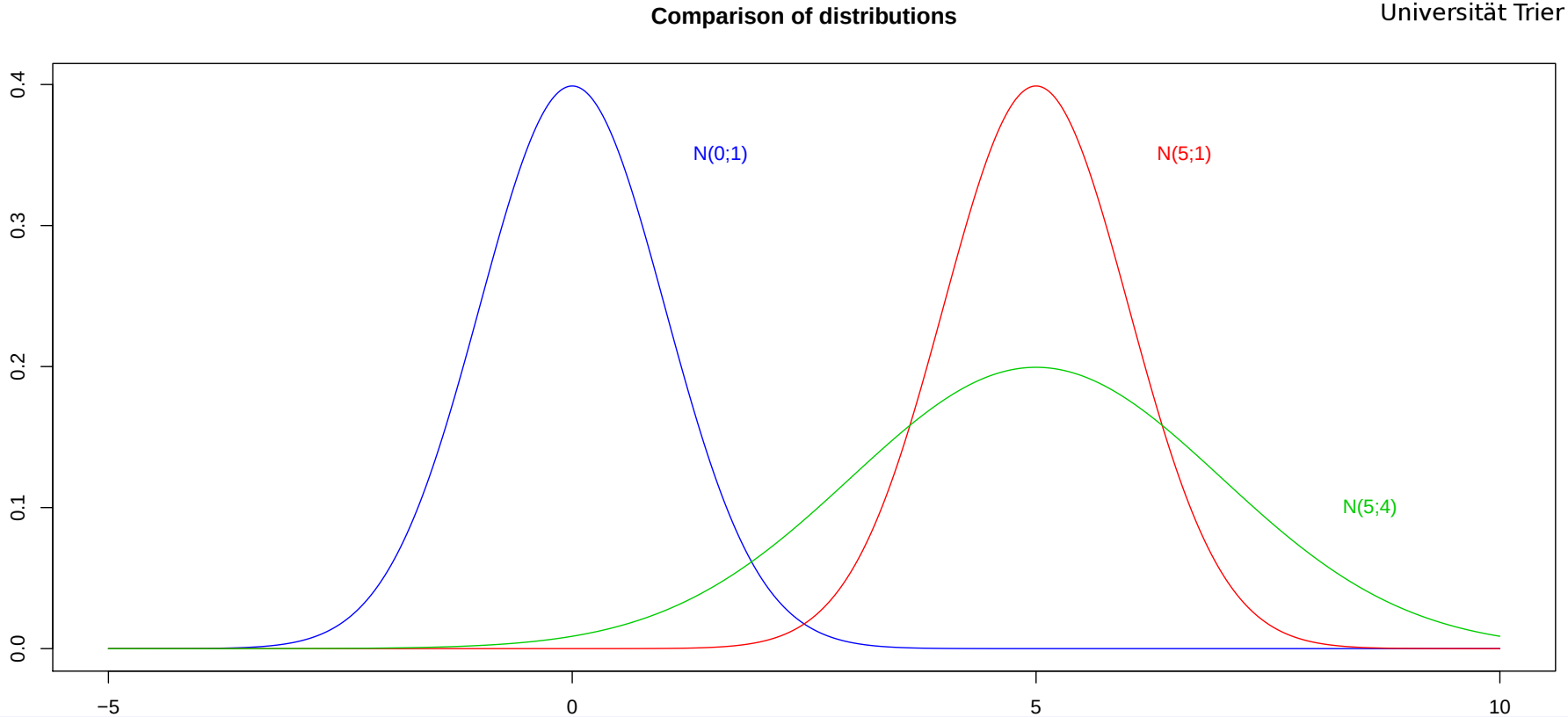
Left-skewed (negatively skewed) distribution $\overline{x} < z_x < x_M$

Always:

$\overline{x} \geq g \geq h$

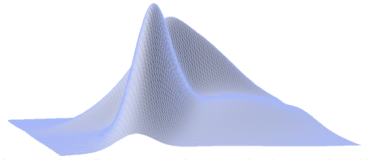
Comparison of distributions (2)

Wirtschafts- und Sozialstatistik
Universität Trier




```
y <- seq(from = -5, to = 10, by = 0.01)
plot(y, dnorm(y, mean = 0, sd = 1), type = "l", xlab = "",
     ylab = "", main = "Comparison of distributions", col=4)
lines(y, dnorm(y, mean = 5, sd = 1), col = 2)
lines(y, dnorm(y, mean = 5, sd = 2), col = 3)
text(x = 1.6, y = 0.35, label = "N(0;1)", col = 4)
text(6.6, 0.35, "N(5;1)",col=2);text(8.6,0.1,"N(5;4)",col=3)
```

Properties of measures of variation



Wirtschafts- und Sozialstatistik
Universität Trier

Degenerate distribution

If all observations are equal, there is no variation in the data and the measure of variation should take on the value zero.

Positive value

When there are at least two unique observations, the measure of variation should take on a positive value.

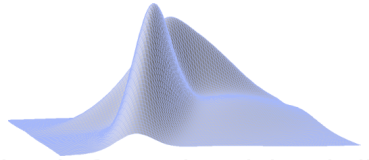
Translation invariance

If each observation is shifted by a common constant, the measure of variation should remain unchanged.

Axiom of homogeneity

If the frequencies of all m different values are (multiplicatively) changed by a common value in such a way that the relative frequencies stay constant, the value of the measure of variation should not change (homogeneity of degree zero). The empirical distribution function will not be changed!

Range and interquartile range



Wirtschafts- und Sozialstatistik
Universität Trier

Range

The range w is defined by

$$w = \max_{i=1,\ldots,n} x_i - \min_{i=1,\ldots,n} x_i = x_{[n]} - x_{[1]} \quad .$$

$x_{[1]}$ ($x_{[n]}$) is the first (n-th) element of the ordered list of values.

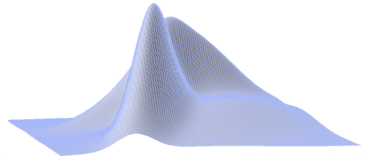
Interquartile range

The interquartile range IQR is

$$\text{IQR} = x_{0.75} - x_{0.25}$$

and therefore equal to the difference of the third and first quartile.

Mean linear deviation



Wirtschafts- und Sozialstatistik
Universität Trier

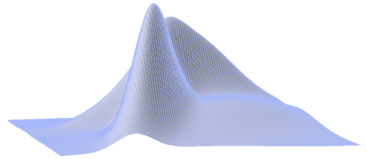
The mean linear deviation l is given by

$$l = \frac{1}{n} \sum_{i=1}^{n} |x_i - z| \quad .$$

The median minimises the mean linear deviation as a *measure of distance:*

$$\arg\min_{t \in \mathbb{R}} l(t) = \arg\min_{t \in \mathbb{R}} \frac{1}{n} \sum_{i=1}^{n} |x_i - t| = z \, .$$

Variance



Wirtschafts- und Sozialstatistik
Universität Trier

The variance s^{*2} is defined as

$$s^{*2} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2 \quad .$$

The arithmetic mean $\overline{x}$ minimises the variance as a *measure of distance:*

$$\arg\min_{t\in\mathbb{R}} \frac{1}{n}\sum_{i=1}^{n}(x_i - t)^2 = \overline{x}$$

Proof using derivative. . .

Wirtschafts- und Sozialstatistik
Universität Trier

Variance and *displacement law* - *Der Verschiebungssatz*

We have

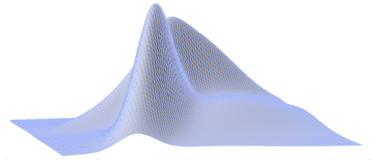
$$s^{*2} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2 = \frac{1}{n}\sum_{i=1}^{n}x_i^2 - \overline{x}^2 \quad .$$

Furthermore, we use the *inferential variance*:

$$\begin{aligned} s^2 &= \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2 = \frac{n}{n-1}\cdot s^{*2} \\ &= \frac{1}{n-1}\sum_{i=1}^{n}x_i^2 - \frac{n}{n-1}\cdot\overline{x}^2 \quad . \end{aligned}$$

The variance s^{*2} is often called the *empirical variance*.

Standard deviation and coefficient of variation



Wirtschafts- und Sozialstatistik
Universität Trier

The standard deviation is the square root of the variance:

$$s^* = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2} \qquad \text{or} \qquad s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2} \quad .$$

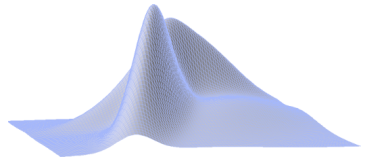
The standard deviation has the same unit of measurement as the arithmetic mean.

As a unit-independent measure of variation, we may use the coefficient of variation

$$v = \frac{s^*}{\overline{x}} \cdot 100\% \quad .$$

It requires a ratio scale.

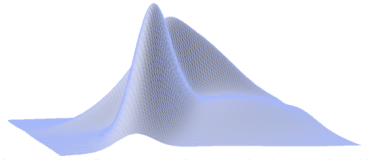
Example 3.10: Screws (1)



Quality control yielded the following actual (and squared actual) lengths of two sets of 10 screws which were supposed to be 10 mm and 80 mm long:

i	$x_{1,i}$	$x_{1,i}^2$	$x_{2,i}$	$x_{2,i}^2$
1	10.40	108.16	81.00	6561.00
2	9.90	98.01	80.30	6448.09
3	10.30	106.09	80.00	6400.00
4	9.80	96.04	79.90	6384.01
5	9.90	98.01	79.80	6368.04
6	10.30	106.09	80.60	6496.36
7	10.40	108.16	80.30	6448.09
8	10.10	102.01	80.20	6432.04
9	10.20	104.04	80.30	6448.09
10	9.70	94.09	80.60	6496.36
$\sum$	101.00	1020.70	803.00	64482.08

Example 3.10: Screws (2)



Wirtschafts- und Sozialstatistik
Universität Trier

Data input in R:

```
x_1 <- c(10.4, 9.9 ,10.3, 9.8, 9.9, 10.3, 10.4, 10.1, 10.2,
         9.7)
x_2 <- c(81, 80.3, 80, 79.9, 79.8, 80.6, 80.3, 80.2, 80.3,
         80.6)
```

First, we calculate:

$$\overline{x}_1 = \frac{1}{10} \cdot (10.4 + \ldots + 9.7) = 10.1$$

$$\overline{x}_2 = \frac{1}{10} \cdot (81.0 + \ldots + 80.6) = 80.3$$

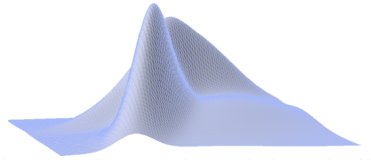
```
mean(x_1)
```

```
[1] 10.1
```

```
mean(x_2)
```

```
[1] 80.3
```

Example 3.10: Screws (3)



We calculate further:

$$s_1^{*2} = \frac{1}{10} \cdot (10.4^2 + \ldots + 9.7^2) - 10.1^2 = 0.06$$

$$s_2^{*2} = \frac{1}{10} \cdot (81.0^2 + \ldots + 80.6^2) - 80.3^2 = 0.118$$

Calculation in R:

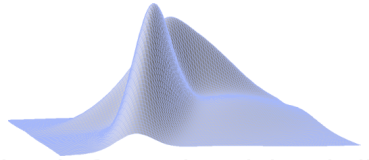
```
x_1_var <- (length(x_1) - 1) / length(x_1) * var(x_1)
x_2_var <- (length(x_2) - 1) / length(x_2) * var(x_2)
```

```
x_1_var
```

```
[1] 0.06
```

```
x_2_var
```

```
[1] 0.118
```



Example 3.10: Screws (4)

From this, we get $s_1^* = 0.244949$ and $s_2^* = 0.3435113$.

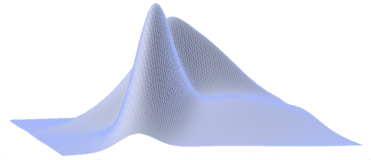
```
sqrt(x_1_var)
```
```
[1] 0.244949
```
```
sqrt(x_2_var)
```
```
[1] 0.3435113
```

The relative dispersion are $v_1 = 2.425237\%$ and $v_2 = 0.4277849\%$

Calculation in R:

```
x_1_var_koeff <- sqrt(x_1_var) / mean(x_1) * 100
x_2_var_koeff <- sqrt(x_2_var) / mean(x_2) * 100
```

```
x_1_var_koeff
```
```
[1] 2.425237
```
```
x_2_var_koeff
```
```
[1] 0.4277849
```



Variance decomposition

If a population is divided into m subpopulations, the variance can be decomposed as follows:

$$s^{*2} = s_b^{*2} + s_w^{*2}$$

(b: between; w: within) with

$$s_b^{*2} = \sum_{j=1}^{m} p_j \cdot (\overline{x}_j - \overline{x})^2$$

$$s_w^{*2} = \sum_{j=1}^{m} p_j \cdot s_j^{*2}$$

The two parts are the *external and internal variance, where*

$$s_j^{*2} = \frac{1}{n_j} \sum_{\nu=1}^{n_j} (x_{\nu j} - \overline{x}_j)^2 \quad , \quad \overline{x} = \sum_{j=1}^{m} p_j \cdot \overline{x}_j$$

Variance and grouped data

Wirtschafts- und Sozialstatistik
Universität Trier

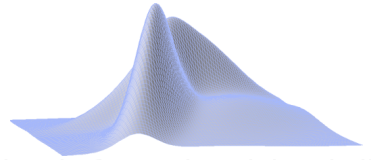
- If class means and class variances are known: Variance decomposition
- If class means and class variances are unknown:

$$s^{*2} = \sum_{j=1}^{m} p_j \cdot \underbrace{(\overline{x}_j - \overline{x})}_{x'_j - \overline{x}'}{}^2 + \underbrace{\sum_{j=1}^{m} p_j \cdot s_j^{*2}}_{\approx 0} \quad \text{with} \quad \overline{x}' = \sum_{j=1}^{m} p_j \cdot x'_j.$$

Therefore, we have:

$$s'^{*2} = \sum_{j=1}^{m} p_j \cdot (x'_j - \overline{x}')^2 = \sum_{j=1}^{m} p_j \cdot x_j'^2 - \overline{x}'^2 .$$

Example 3.11: Income classes (1)



Let the following $n = 10$ income values be given:
3500; 3200; 2100; 500; 1800; 2100; 5600; 8500; 1400; 1200 Furthermore,

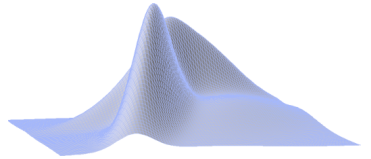
let the class boundaries be:
$x_0^o = 0$; $x_1^o = 2500$; $x_2^o = 5000$; $x_3^o = 7500$; $x_4^o = 10000$.

Data input in R:

```
x3_11 <- c(3500,3200,2100,500,1800,2100,5600,8500,1400,1200)
x_o <- c(0,2500,5000,7500,10000)
x3_11_kl <- cut(x3_11, x_o, right = FALSE)
```

j	p_j	$\overline{x}_j$	s_j^{*2}	$p_j(\overline{x}_j - \overline{x})^2$	$p_j s_j^{*2}$	x'_j	$p_j x'_j$	$p_j(x'_j - \overline{x}')^2$
1	0.6	1516.67	318056	1302427	190833	1250	750	1837500
2	0.2	3350,00	22500	25920	4500	3750	750	112500
3	0,1	5600.00	0	681210	0	6250	625	1056250
4	0.1	8500.00	0	3036010	0	8750	875	3306250
$\sum$				5045567	195333		3000	6312500

Example 3.11: Income classes (2)



j	p_j	$\overline{x}_j$	s_j^{*2}	$p_j(\overline{x}_j - \overline{x})^2$	$p_j s_j^{*2}$	x'_j	$p_j x'_j$	$p_j(x'_j - \overline{x}')^2$
1	0.6	1516.67	318056	1302427	190833	1250	750	1837500
2	0.2	3350,00	22500	25920	4500	3750	750	112500
3	0,1	5600.00	0	681210	0	6250	625	1056250
4	0.1	8500.00	0	3036010	0	8750	875	3306250
$\sum$				5045567	195333		3000	6312500

Calculation of p_j and n_j in R:

```
p_j <- prop.table(x = table(x3_11_kl))
n_j <- p_j * length(x3_11)
```

Calculation of $\overline{x}_j$ in R:

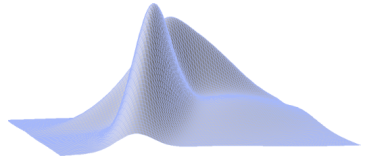
```
x_mean_j <- tapply(X = x3_11, INDEX = x3_11_kl, FUN = mean)
```

Calculation of s_j^{*2} in R:

```
x_var_j <- (n_j - 1)/n_j *
             tapply(X = x3_11, INDEX = x3_11_kl, FUN = var)
x_var_j
```

```
[0,2.5e+03) [2.5e+03,5e+03) [5e+03,7.5e+03) [7.5e+03,1e+04)
   318055.6         22500.0
```

Example 3.11: Income classes (3)



j	p_j	$\overline{x}_j$	s_j^{*2}	$p_j(\overline{x}_j - \overline{x})^2$	$p_j s_j^{*2}$	x'_j	$p_j x'_j$	$p_j(x'_j - \overline{x}')^2$
1	0.6	1516.67	318056	1302427	190833	1250	750	1837500
2	0.2	3350,00	22500	25920	4500	3750	750	112500
3	0,1	5600.00	0	681210	0	6250	625	1056250
4	0.1	8500.00	0	3036010	0	8750	875	3306250
$\sum$				5045567	195333		3000	6312500

$$s^{*2} = s_b^{*2} + s_w^{*2} = 5045567 + 195333 = 5240900$$

Calculation of s^{*2} in R:

```
x_var_j[is.na(x_var_j)] <- 0

x_var_between <- sum(p_j * (x_mean_j - mean(x3_11))^2)
x_var_within <- sum(p_j * x_var_j)

x_var_kl <- x_var_between + x_var_within
```

```
x_var_between
```

```
[1] 5045567
```

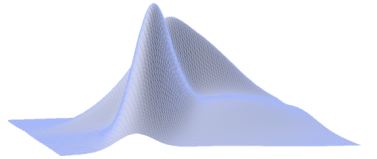
```
x_var_within
```

```
[1] 195333.3
```

```
x_var_kl
```

```
[1] 5240900
```

Example 3.11: Income classes (4)



j	p_j	$\overline{x}_j$	s_j^{*2}	$p_j(\overline{x}_j - \overline{x})^2$	$p_j s_j^{*2}$	x'_j	$p_j x'_j$	$p_j(x'_j - \overline{x}')^2$
1	0.6	1516.67	318056	1302427	190833	1250	750	1837500
2	0.2	3350,00	22500	25920	4500	3750	750	112500
3	0,1	5600.00	0	681210	0	6250	625	1056250
4	0.1	8500.00	0	3036010	0	8750	875	3306250
$\sum$				5045567	195333		3000	6312500

$$s'^{*2} = 6312500$$

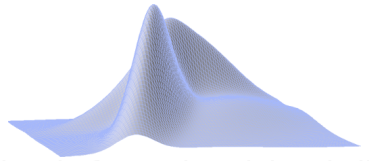
Calculation of x'_j and $\overline{x}'$ in R:

```
x_mean_approx_j <- x_o[-5] + (x_o[-1] - x_o[-5]) / 2
x_mean_approx_kl <- sum(p_j * x_mean_approx_j)
```

Calculation of s'^{*2} in R:

```
x_var_approx_kl <- sum(p_j * (x_mean_approx_j -
                              x_mean_approx_kl)^2)
x_var_approx_kl
```

```
[1] 6312500
```



Example 3.12: see Ex. 3.11

An alternative grouping according to

- $x_0^o = 0;\ x_1^o = 2000;\ x_2^o = 4000;\ x_3^o = 10000$ would lead to

$$s'^{*2} = 4,800,000\,.$$

- $x_0^o = 0;\ x_1^o = 2000;\ x_2^o = 5000;\ x_3^o = 10000$ would lead to

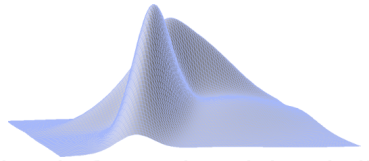
$$s'^{*2} = 5,660,000\,.$$

In both cases we have

$$s_b^{*2} = 4,570,900 \qquad \text{and} \qquad s_w^{*2} = 670,000\,.$$

Calculation in R with use uf the R code from Ex. 3.11.
For this, the vector **x_o** from Ex. 3.11 must be changed.
E.g. for the first case above of alternative grouping of the data:

```
x_o <- c(0,2000,4000,10000)
```



Wirtschafts- und Sozialstatistik
Universität Trier

Influence of changes in the unit of measurement

If the n original values x_i are linearly transformed ($y_i = a_0 + a_1 \cdot x_i$), we have

$$\overline{y} = a_0 + a_1 \cdot \overline{x} \quad , \quad s_y^{*2} = a_1^2 \cdot s_x^{*2} \quad , \quad s_y^* = |a_1| \cdot s_x^*$$

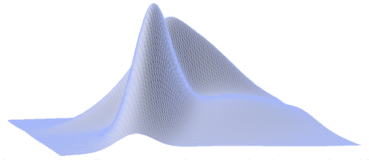
and

$$v_y = \frac{|a_1| \cdot s_x^*}{a_0 + a_1 \cdot \overline{x}} \quad .$$

For the **standard transformation**

$$y_i = \frac{x_i - \overline{x}}{s_x^*}$$

we specifically have $\overline{y} = 0$ and $s_y^{*2} = 1$.



Entropy

The *spread* of values of variables on a nominal or an ordinal scale can be measured by entropy, which is:

$$E = -\sum_{j=1}^{m} p_j \cdot \ln p_j = \ln n - \frac{1}{n} \sum_{j=1}^{m} n_j \cdot \ln n_j \quad .$$

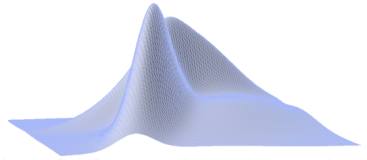
Entropy reaches its maximum when frequencies are identical.
As relative entropy

$$E_r = \frac{E}{\ln m}$$

is used. Then we have $0 \leq E_r \leq 1$.

Notice that measures of variation are usually based on metric scales!

Skewness and kurtosis



Wirtschafts- und Sozialstatistik
Universität Trier

- The skewness of a distribution is measured by

$$m_3 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^3 \qquad \text{or} \qquad \frac{m_3}{s^{*3}} \quad .$$

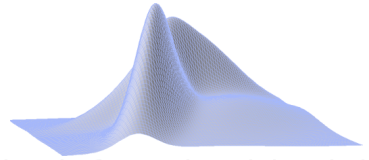
- The kurtosis of a distribution is measured by

$$m_4 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^4 \qquad \text{or} \qquad \frac{m_4}{s^{*4}} \quad .$$

The normal distribution exhibits a kurtosis of 3.

(see *Comparison of distributions (1)* and *(2)*)(slide 33)

Stem-and-leaf plot



As a stem, the first digit is used $(0, \ldots, 5)$. Then, the leafs (next digit) will be attached on the stem in increasing order. The number of stems can be variated corresponding to the size of the data.

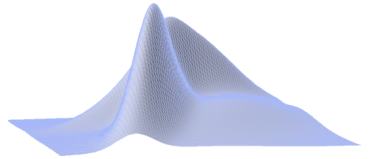
Example 3.13: see Ex. 2.5 resp. 3.6

```
sort(x2_5)
```

```
[1]  500  800 1100 1200 1400 1500 1500 1800 2100 2100 2200 2800 3100 3200
[15] 3500 4500 4500 5200 5400 5600
```

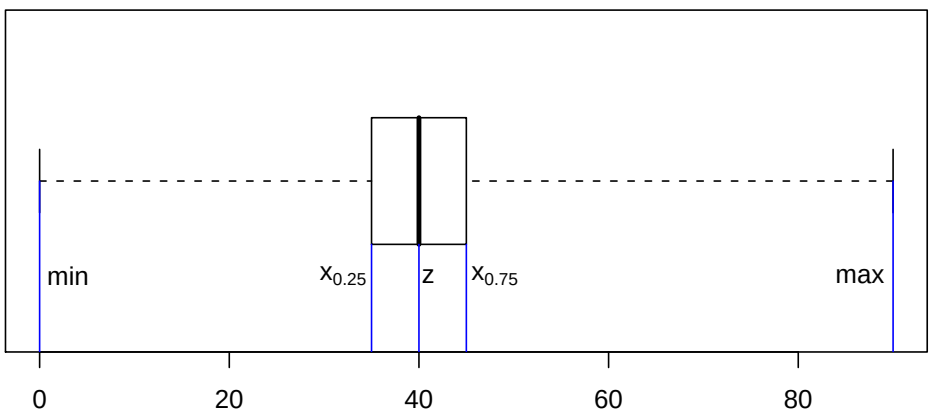
```
stem(x = x2_5, scale = 2)
```

```
  The decimal point is 3 digit(s) to the right of the |
  0 | 58
  1 | 124558
  2 | 1128
  3 | 125
  4 | 55
  5 | 246
```



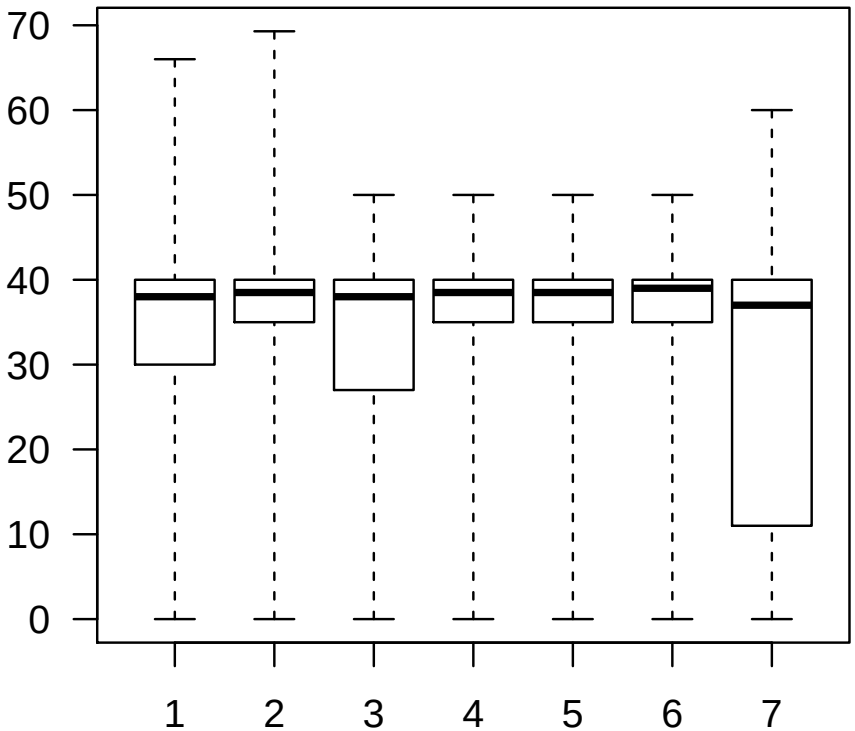
The boxplot

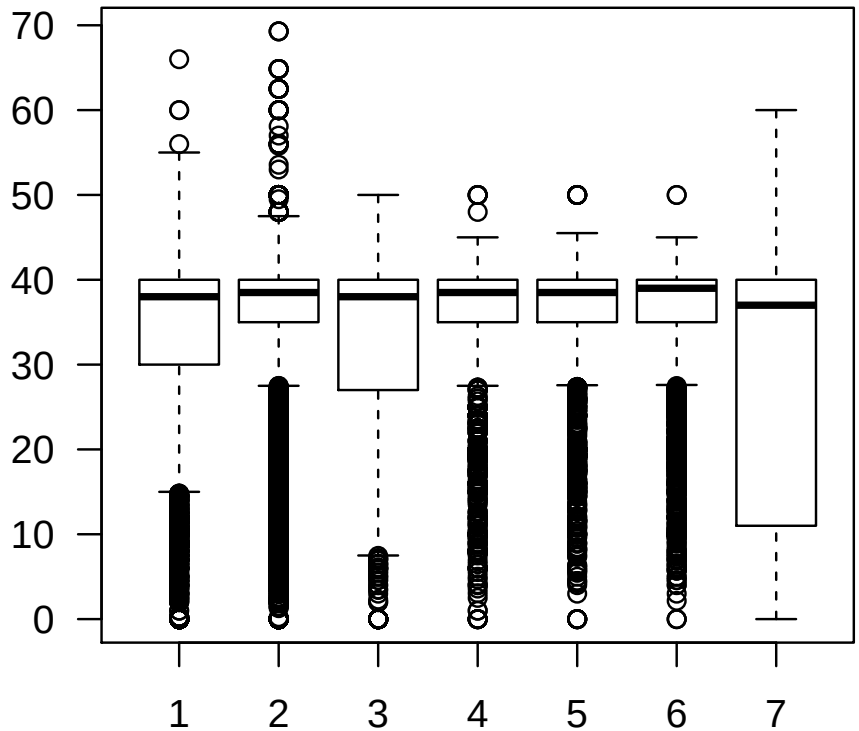
A boxplot (or box-whisker-plot) describes a distribution by means of five chosen parameters: minimum, $x_{0.25}$, z, $x_{0.75}$ and maximum. The box consists of $x_{0.25}$, z and $x_{0.75}$, the length of the box being equal to the interquartile range $\text{IQR} := x_{0.75} - x_{0.25}$. The *whiskers* reach from the ends of the box to the minimum and maximum, respectively.
On a modified boxplot the whiskers may be bounded by the values $x_{0.25} - 1.5 \cdot \text{IQR}$ and $x_{0.75} + 1.5 \cdot \text{IQR}$, respectively. Each observation outside of those boundaries is plotted as an individual value.



Working hours by qualification

Wirtschafts- und Sozialstatistik
Universität Trier





Box-Plots in R:

```
boxplot(Hours ~ Qualification, data = AZ, range = 0)
boxplot(Hours ~ Qualification, data = AZ, range = 1.5)
```