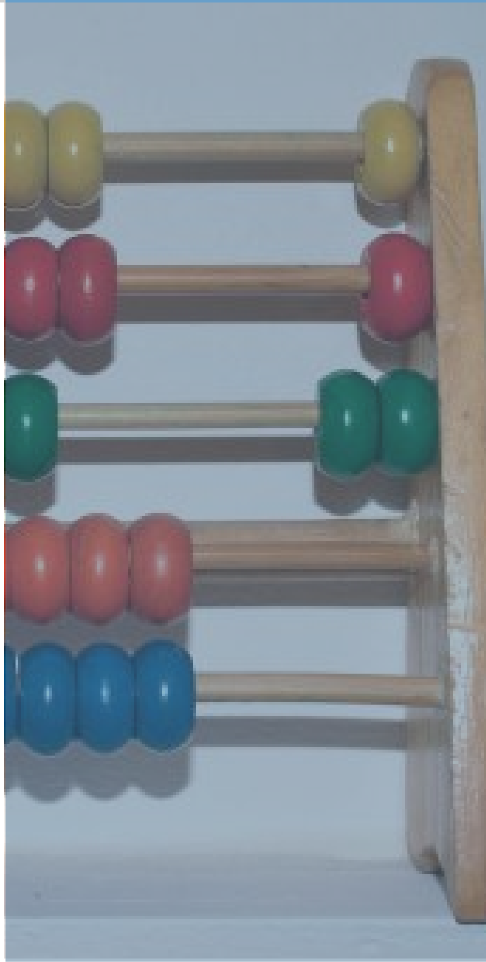
Christian-Albrechts-Universität zu Kiel

04_descriptive_statistics

Central tendency and dispersion

Loading data for the following steps

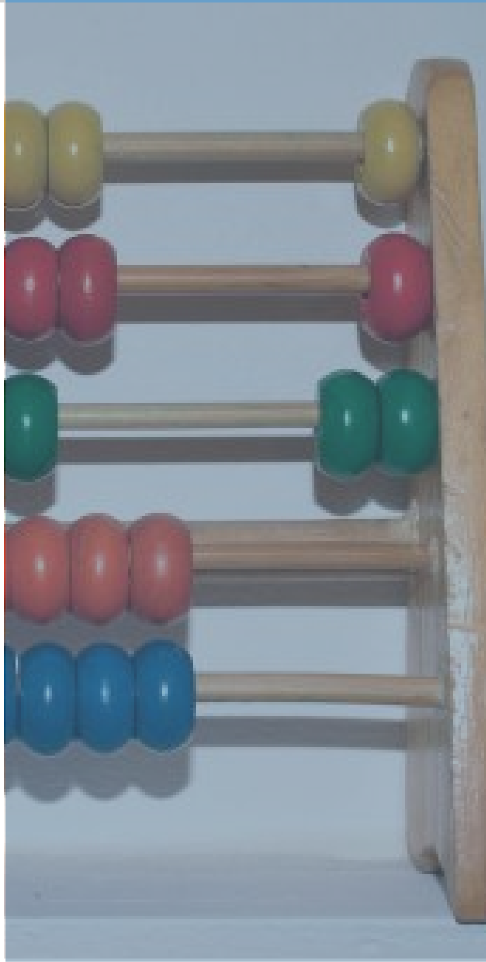


Read the data of the Kursteilnehmer:

```
> setwd("--your R-directory--")
> laender<-read.csv2("laenderdaten.csv")

> laender[1:3,]
               Name Einwohnerzahl Fläche.in.km.                                     Amtssprache        BIP
1 Königreich Dänemark       5732173     2244490.0                                         Dänisch 3.3320e+11
2         New Zealand       4445000      269652.0 Englisch, Maori, neuseeländische Gebärdensprache 1.6181e+11
3            Schweden       9644864      438575.8                                       Schwedisch 5.3820e+11
  Weltrang.nach.BIP Weltrang.CPI Einlieferer kontinent
1                32            1      breske    Europa
2                56            1      breske      <NA>
3                21            1      breske    Europa
```

Deskriptive Statistics



Summary of a amount of observed data

The distribution of the data in the sample is displayed.

Ways of display

Table – contingency table
Graphical – charts
Numeric – with specific parameters of the distribution

Descriptive statistics do (effectivly) not making statements about the population but describes the sample! (in difference to statistical inference)

Parameters of distributions

Central tendency
What is the typical individual

mean, median, mode

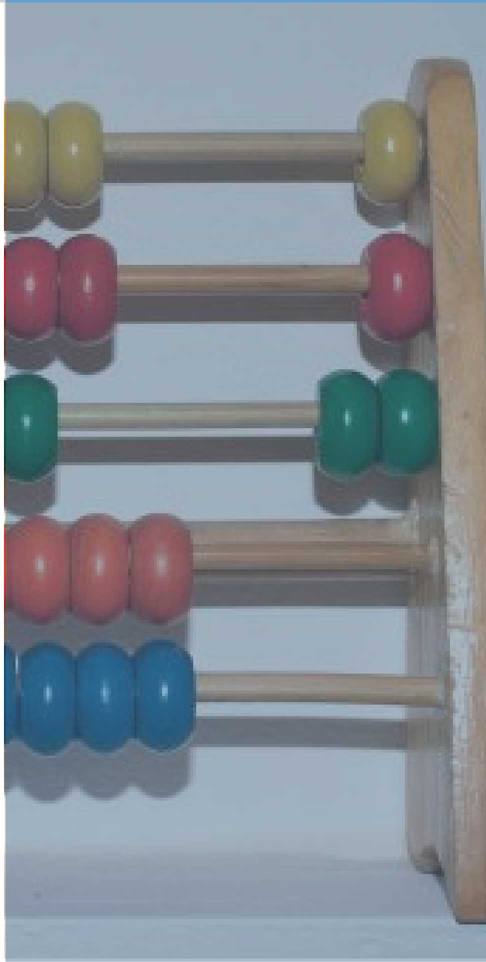
dispersion:
How much variation is there

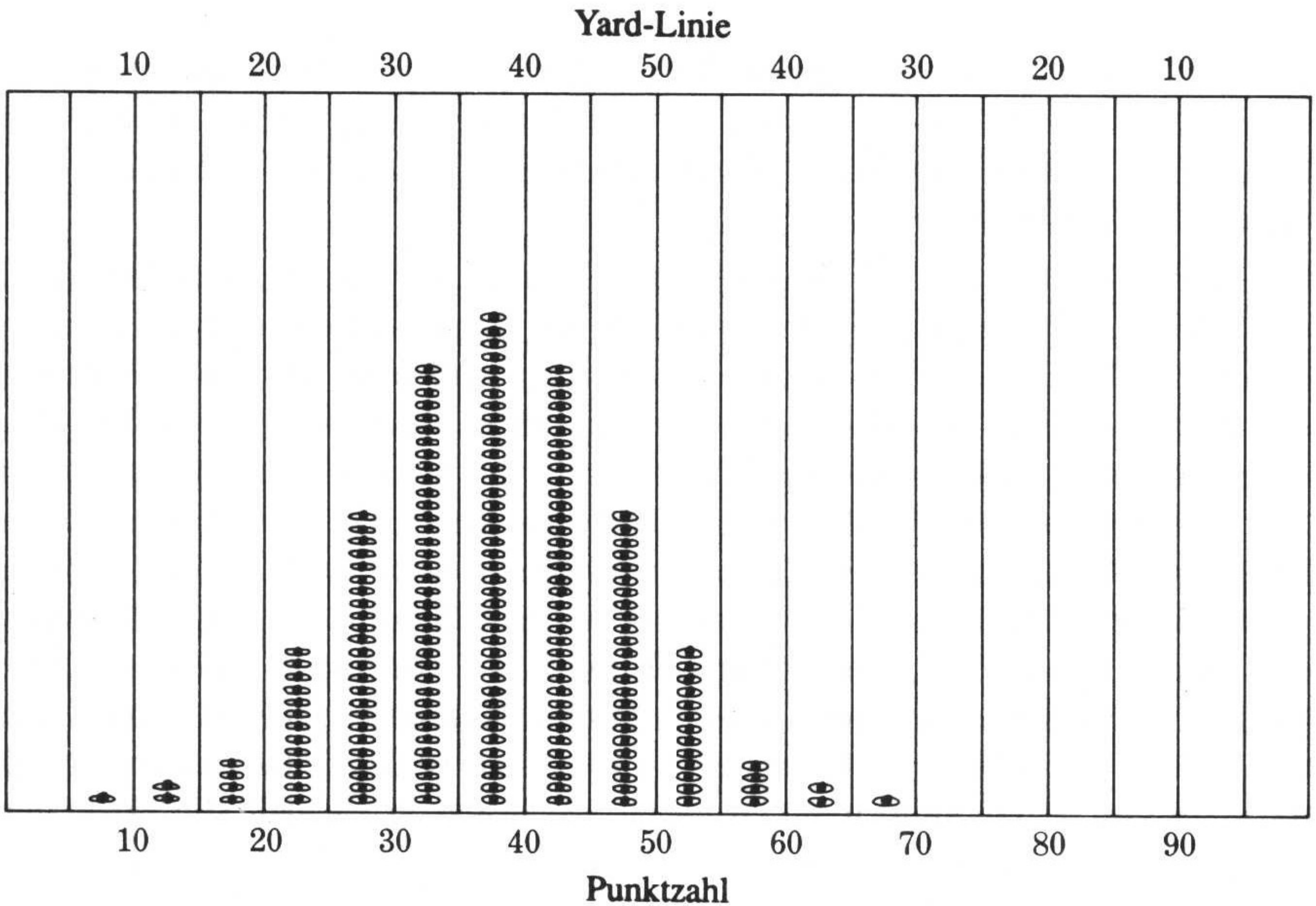
Range, variance, standard deviation, coefficient of variation

shape:
Shape of the distribution curve

symmetric/asymmetric

Skewness and curtosis





Studenten, die sich nach ihren Testergebnissen in Reihen auf einem Footballfeld aufgestellt haben – eine Häufigkeitsverteilung.

Quelle: Phillips 1997

Central tendency [1]

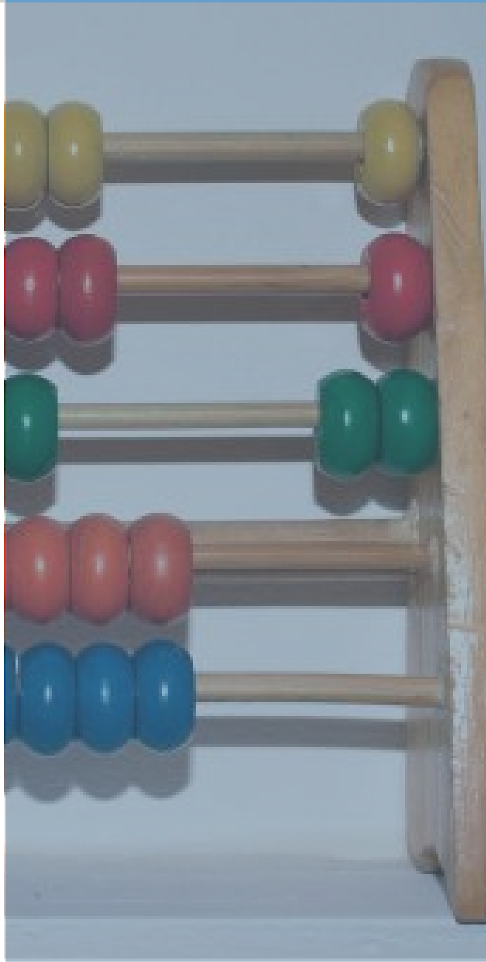
mean
The classic. Suitable for metric data (interval or ratio)

Sum of values/number of values, or

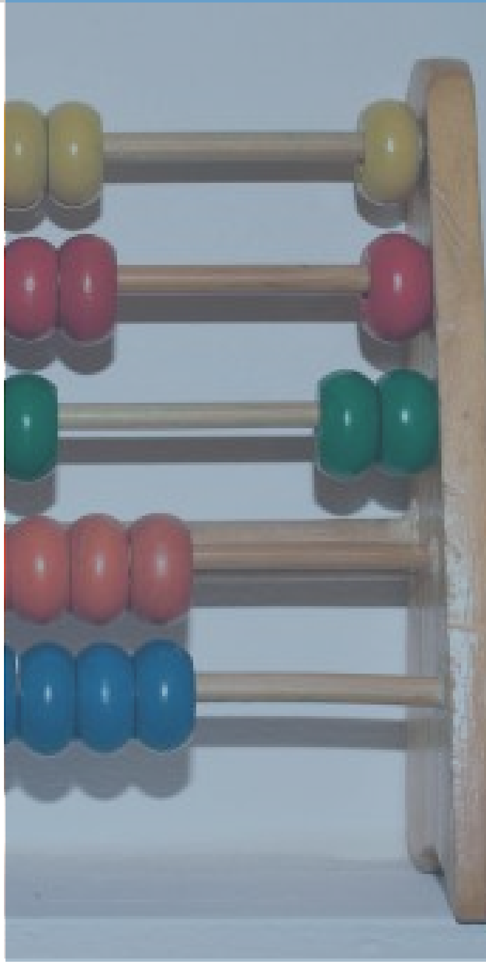
$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

R:

```
> sum(laender$Fläche)/length(laender$Fläche)
[1] 943844
> mean(laender$Fläche)
[1] 943844
```



Central tendency [2]



Median

Suitable for metric and ordinal variables.

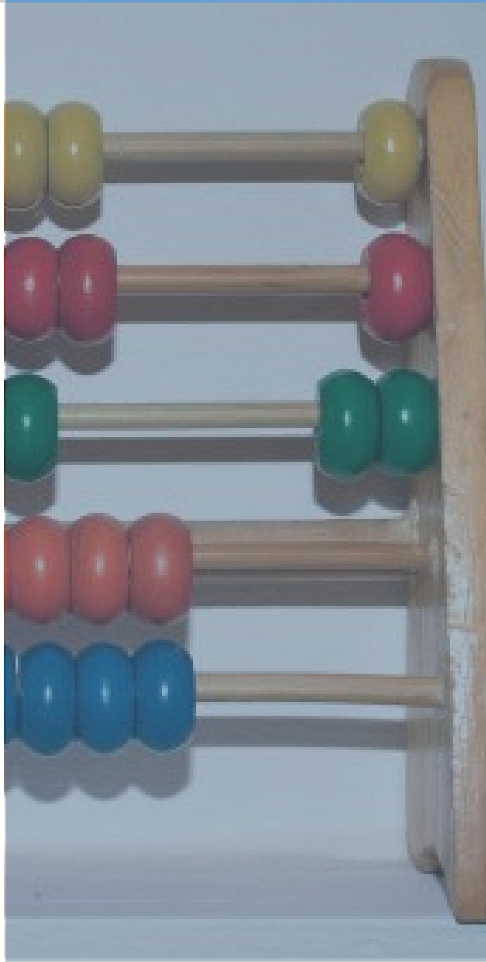
Uneven number: the central value of a sorted vector.

```
1 2 3 4 5 6 7
      |
R:
> median(c(1,2,3,4,5,6,7))
[1] 4
```

Even number: the mean of the two central values of a sorted vector.

```
1 2 3 4 5 6 7 8
       |
R:
> median(c(1,2,3,4,5,6,7,8))
[1] 4.5
```

Central tendency [3]



Mode

The most frequent value of a vector. Suitable for metric, ordinal and nominal variables.

goat sheep goat cattle cattle goat pig goat

Modus: goat

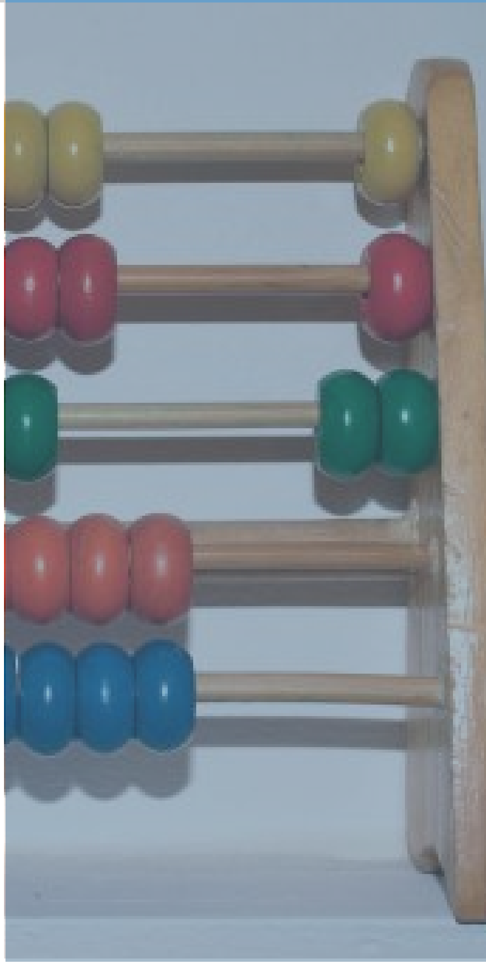
In R:

```
> which.max(table(c("goat", "sheep", "goat", "cattle",
"cattle", "goat", "pig", "goat")))
 goat
    4
```

Central tendency [4]

Variable is		
nominal	ordinal	intervall+
mode	mode	mode
-	median	median
-	-	mean

after: Dolić 2004



Christian-Albrechts-Universität zu Kiel

Central tendency [5]

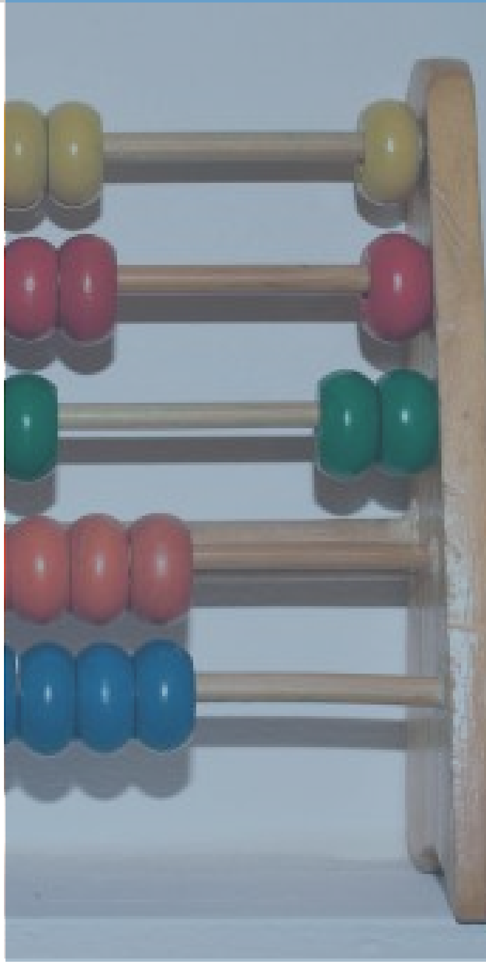
Comparison of central values:

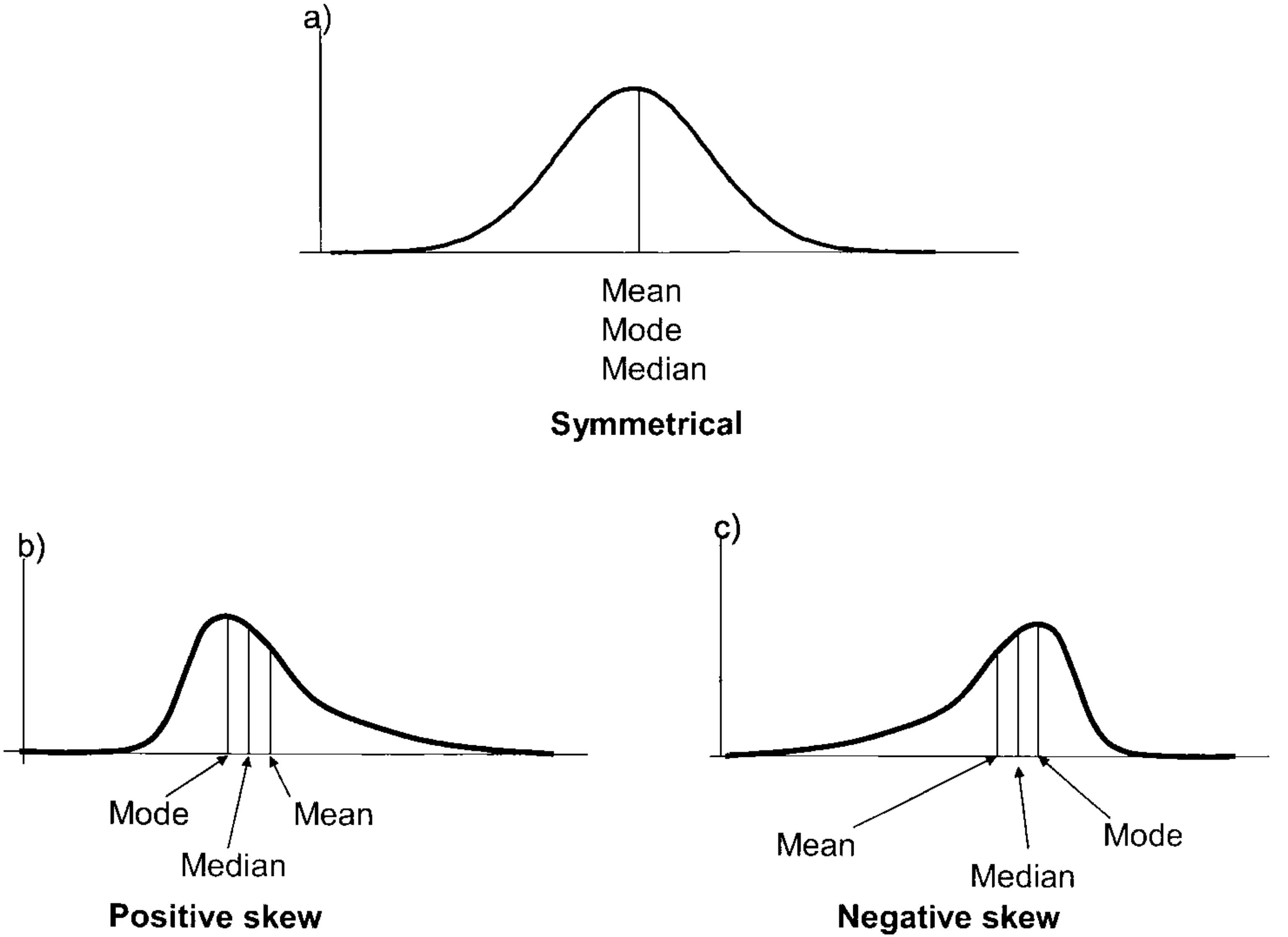
Strongly affected by outliers: the mean is very sensitive for outliers, the median less, the mode hardly

```
> test<-c(1,2,2,3,3,3,4,4,5,5,6,7,8,8,8,9,120)
> mean(test)
[1] 11.64706
> median(test)
[1] 5
> which.max(table(test))
3
3
```

The mode is of little value for describing metric or ordinal data, only when a more or less symmetric distribution is present

```
> which.max(table(c(1,2,2,3,3,3,4,4,4,4,5,5,5,6,6,7)))
4
4
```




a)
Mean
Mode
Median
Symmetrical
b)
Mode
Mean
Median
Positive skew
c)
Mean
Mode
Median
Negative skew

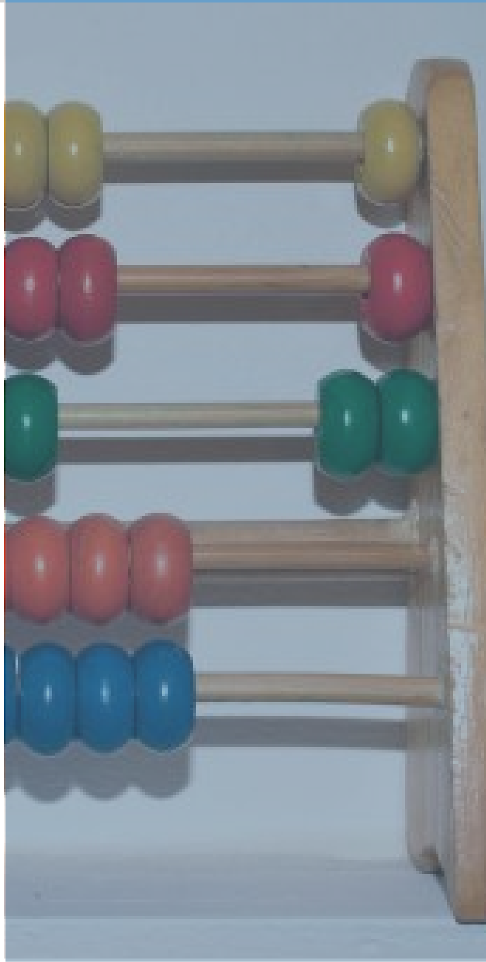
Central tendency exercise

Describe the central tendency

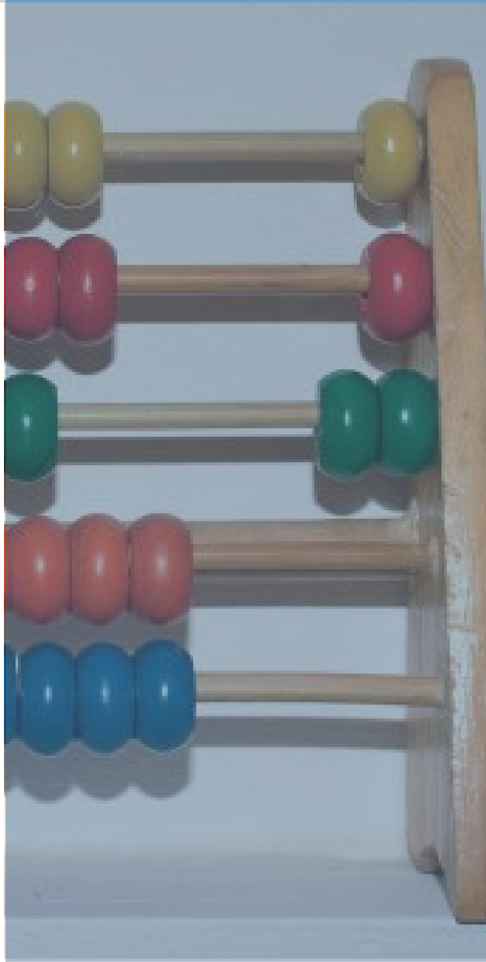
Analyse the measurements of the width of cups (in cm) from the burial ground Walternienburg (Müller 2001, 534; selection):

```
> tassen<-read.csv2("tassen.csv",row.names=1)
> tassen$x
```

Identify the mode, median and mean and determine if the distribution is symmetric, positive or negative skewed.



Central tendency exercise



Describe the central tendency

Analyse the measurements of the width of cups (in cm) from the burial ground Walternienburg (Müller 2001, 534; selection):

```
> tassen<-read.csv2("tassen.csv",row.names=1)
> tassen$x
```

Identify the mode, median and mean and determine if the distribution is symmetric, positive or negative skewed.

```
> mean(tassen$x)
[1] 13.67727
> median(tassen$x)
[1] 12
> which.max(table(tassen$x))
8.1
  3
```

The median is bigger than the mean: positiv skewed.

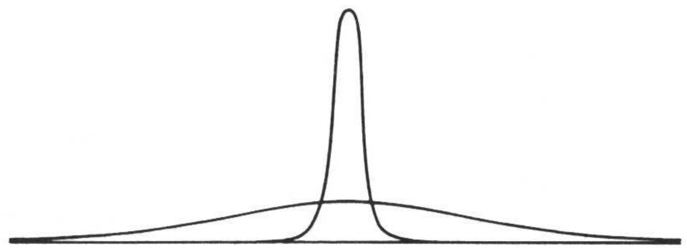
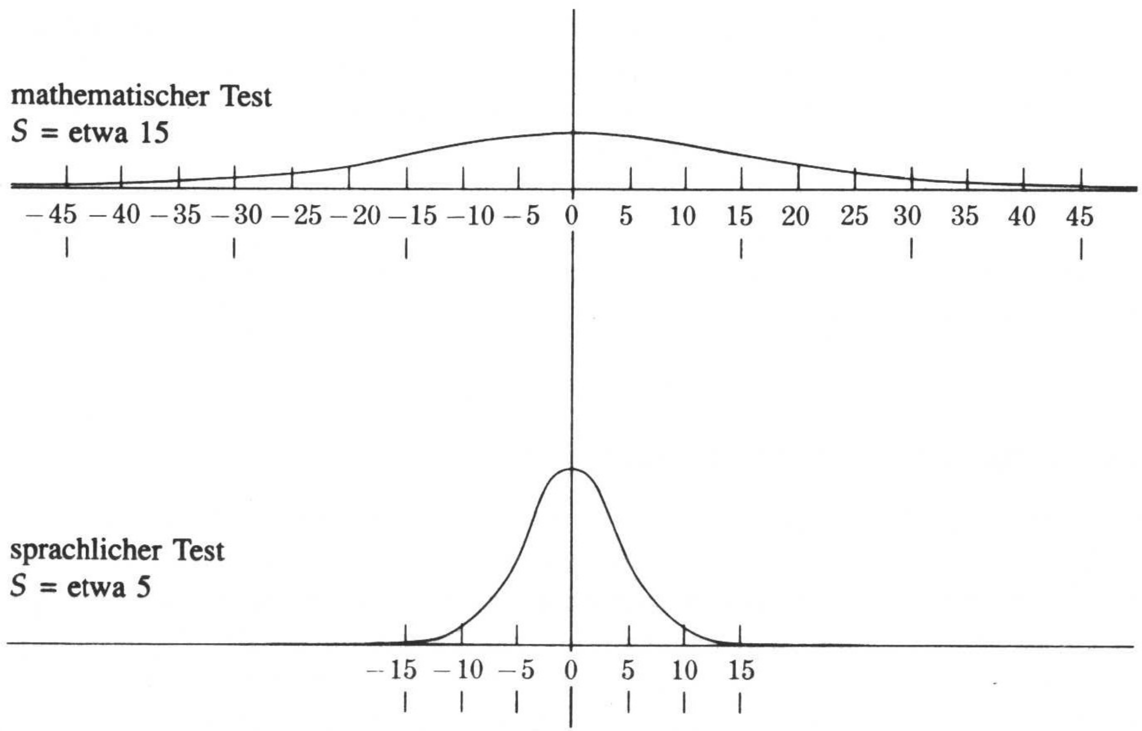
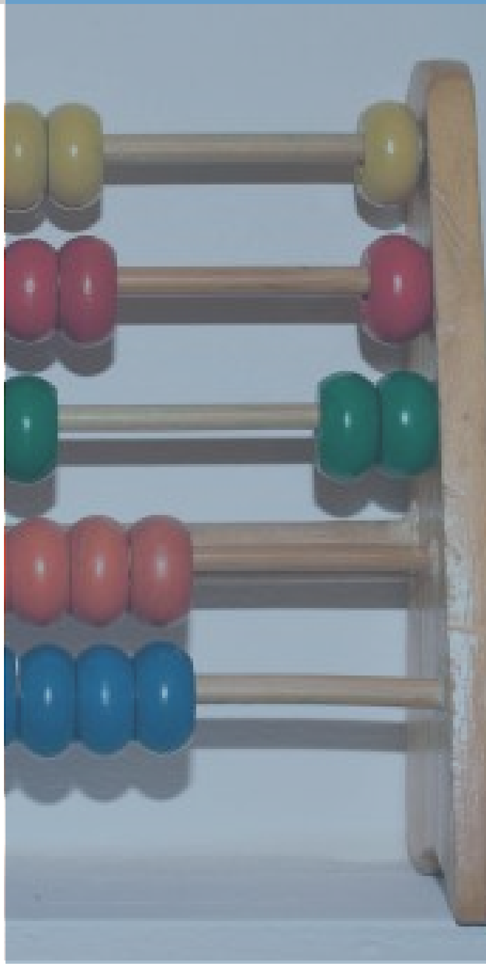


Abb. 4.1 Zwei Verteilungen mit denselben *N*s, aber unterschiedlicher Streuung.



Quelle: Phillips 1997

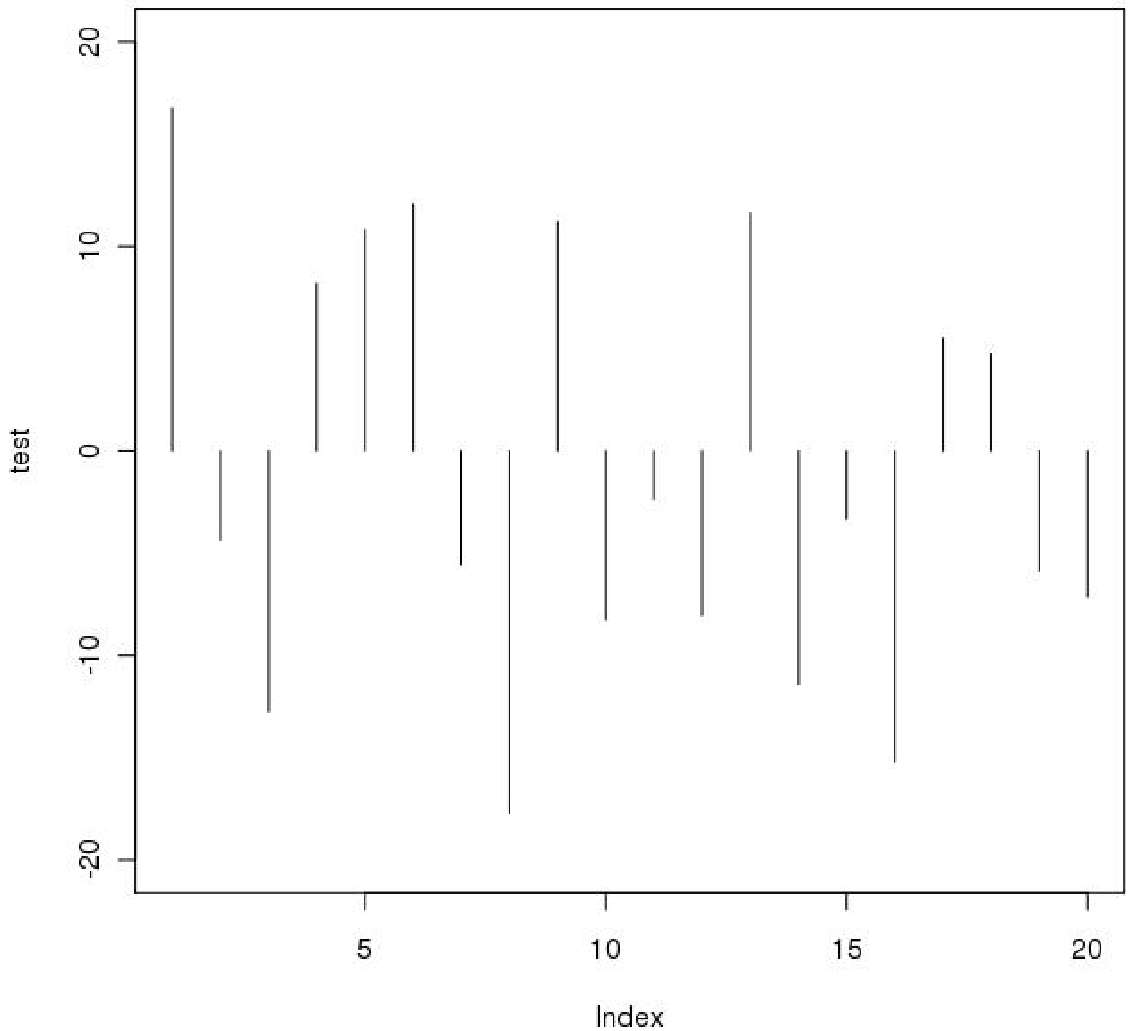
Dispersion [1]



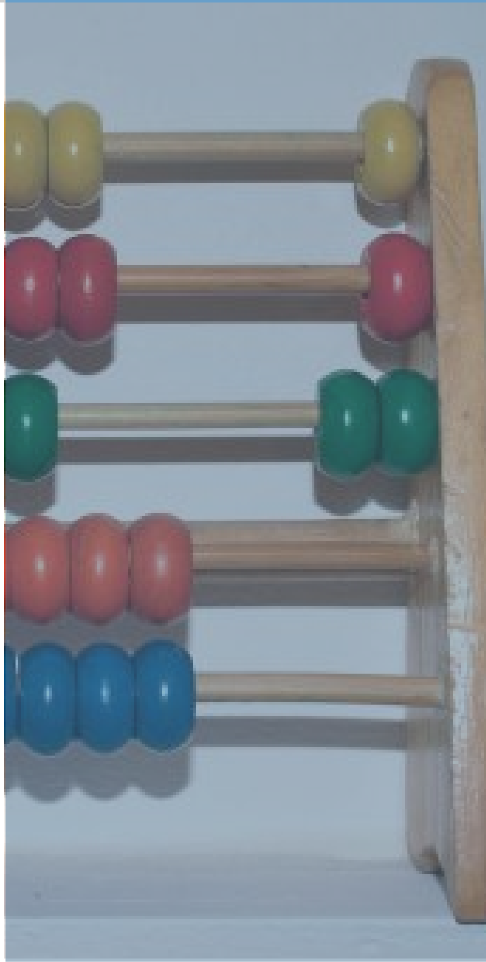
Range
Simply the range of the values of a data vector.

```
> range(laender$Fläche)
[1]   14954 9826675
> range(tassen$x)
[1]  7.5 26.1
```

Because the measurement is related to the extreme values it is very sensitive for outliers.


20
10
0
-10
-20
test
5
10
15
20
Index

Dispersion [2]



(empirical) variance

Measure for the variability of the data, more insensitive against outliers
Equals to the sum of the squared distances from the mean divided by the number of observations

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

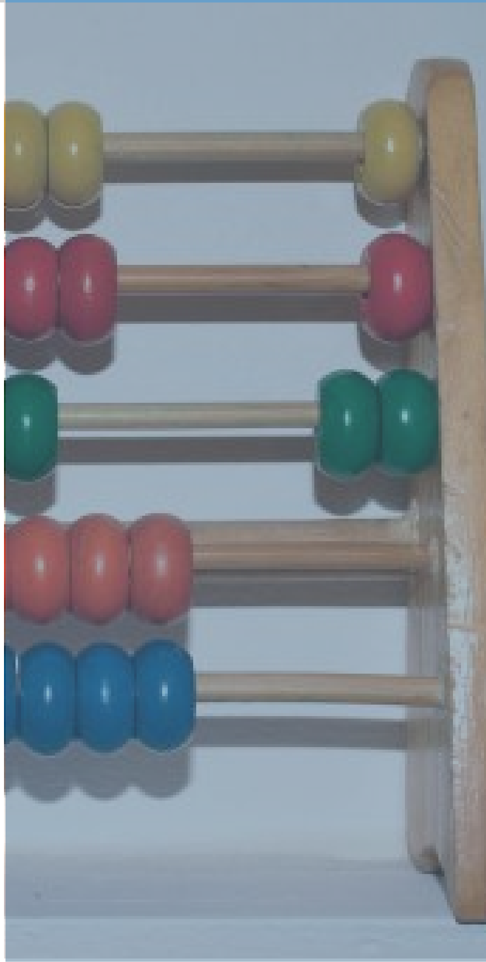
In R:

```
> sum((tassen$x-mean(tassen$x))^2)/(length(tassen$x)-
1)
[1] 31.11136
> var(tassen)
         x
x 31.11136
```

Attention: there is another variance σ² (with n instead of n-1) which is only suitable for analysis of the population (which is not known most of the times), not for samples

Christian-Albrechts-Universität zu Kiel

Dispersion [3]



(empirical) standard deviation

Variance has through the squaring squared units (mm → mm²)

For a parameter with the original units: square root → standard deviation

$$s=\sqrt{\frac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{n-1}}$$

```
> sqrt(sum((tassen$x-mean(tassen$x))^2)/(length(tassen$x)-1))

> sd(tassen$x)
```

Equals the mean distance from the mean

Attention: there is another standard deviation σ (with n instead of n-1) which is only suitable for analysis of the population (which is not known most of the times), not for samples

Dispersion [4]

coefficient of variation

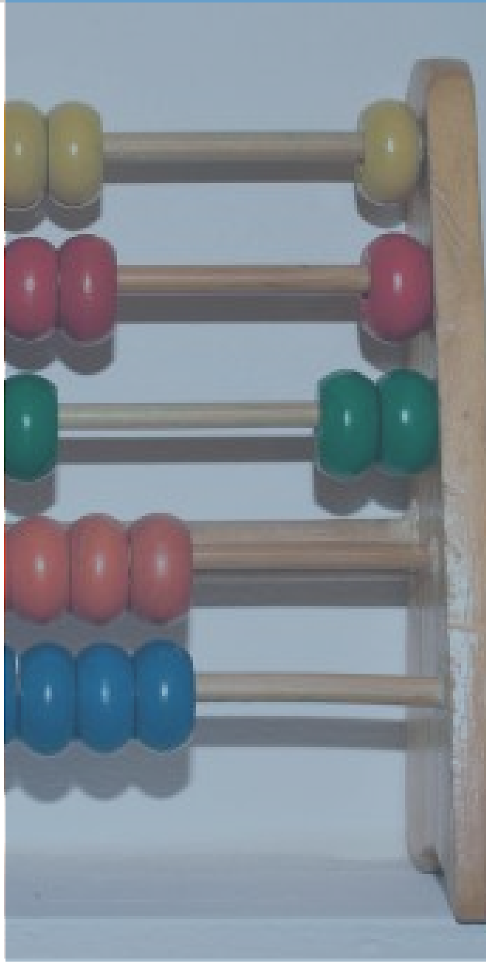
Standard deviation has the unit of the original data (e.g. mm).

To compare two distributions with different units:
coefficient of variation = standard deviation/mean

Example: Vary foot size and body height equal?

```
> sd(laender$Fläche)/mean(laender$Fläche)
[1] 2.576648
> sd(laender$Einwohnerzahl)/mean(laender$Einwohnerzahl)
[1] 2.479968
```

Foot size vary more than body height



Dispersion [5]

Quantile

Oh, we've done that one...
The 1., 2., 3. and 4. quarter of the data (sorted and counted) resp. there boundaries

```
> quantile(tassen$x)
  0%  25%  50%  75% 100%
 7.5  9.0 12.0 18.9 26.1
```

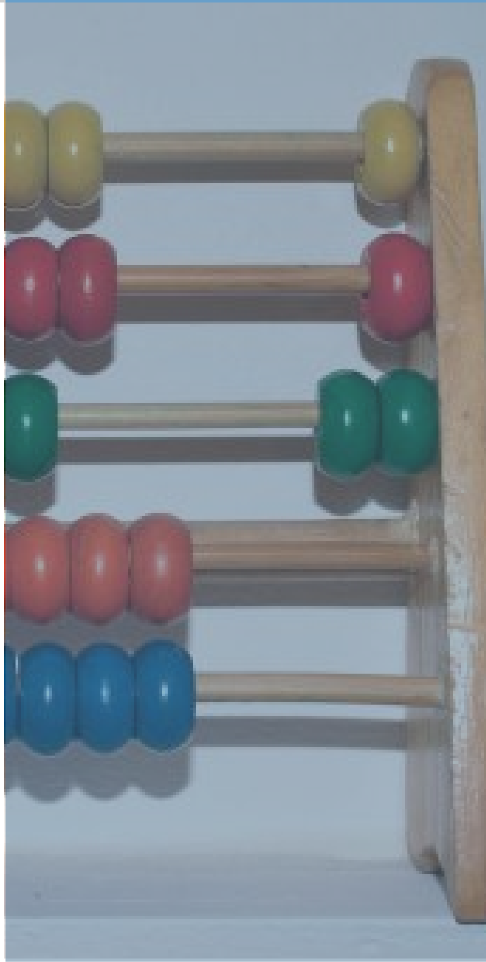
new: percentile (the same for percents)

```
> quantile(tassen$x, probs=seq(0,1,0.1))
   0%   10%   20%   30%   40%   50%   60%   70%   80%   90%  100%
 7.50  8.10  8.52  9.27 10.02 12.00 13.08 18.81 19.38 20.31 26.10
```

Dispersion measure inner quartile range

```
> IQR(tassen$x)
[1] 9.9
```

More insensitive against outliers than the standard deviation, but information is lost

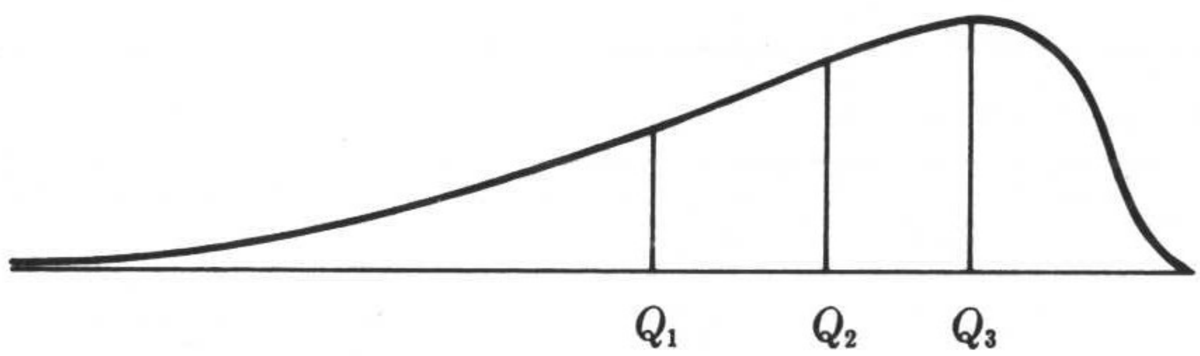


Dispersion [5]

Quantile

Oh, we've done that one...

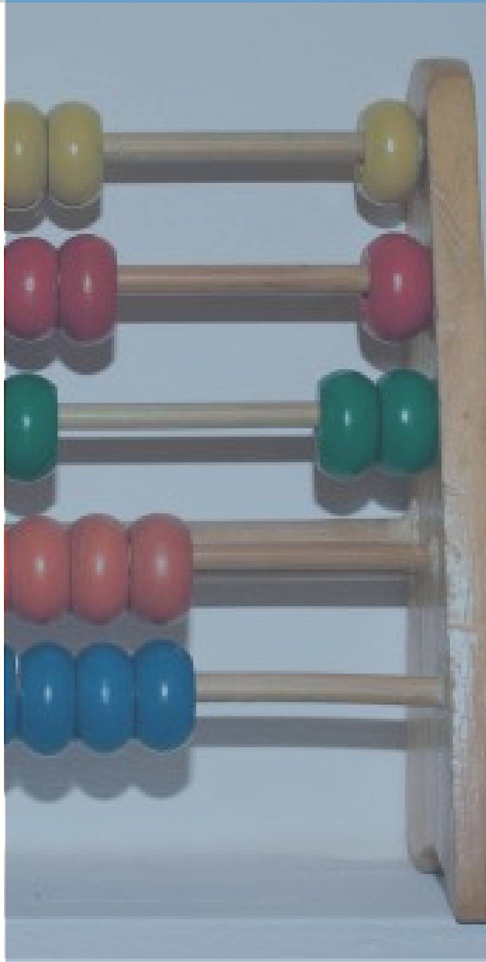
The 1., 2., 3. and 4. quarter of the data (sorted and counted) resp. there boundaries



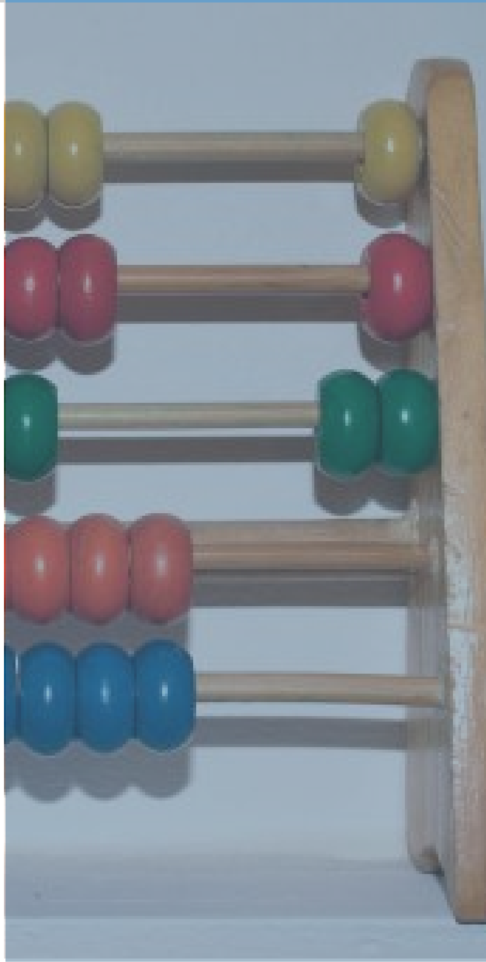
Linksschiefe Verteilung mit einer in Viertel geteilten Fläche.

Quelle: Phillips 1997

More insensitive against outliers than the standard deviation, but information is lost



Dispersion [5]



Quantile

Oh, we've done that one...
The 1., 2., 3. and 4. quarter of the data (sorted and counted) resp. there boundaries

```
> quantile(tassen$x)
  0%  25%  50%  75% 100%
 7.5  9.0 12.0 18.9 26.1
```

new: percentile (the same for percents)

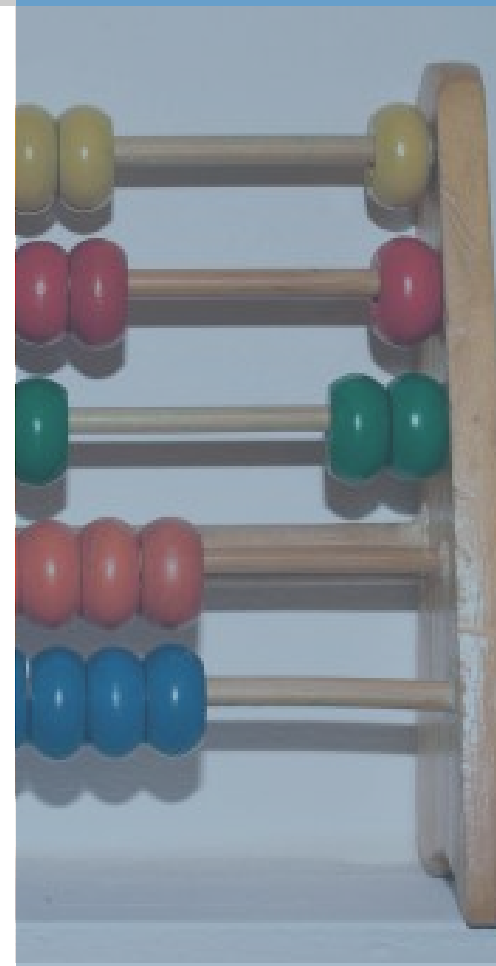
```
> quantile(tassen$x, probs=seq(0,1,0.1))
   0%   10%   20%   30%   40%   50%   60%   70%   80%   90%  100%
 7.50  8.10  8.52  9.27 10.02 12.00 13.08 18.81 19.38 20.31 26.10
```

Dispersion measure inner quartile range

```
> IQR(tassen$x)
[1] 9.9
```

More insensitive against outliers than the standard deviation, but information is lost

Dispersion exercise



Determine the dispersion of the data

Analyse the sizes of areas visible from different megalithic graves of the Altmark (Demnick 2009):

```
> altmark<-read.csv2("altmark_denis.csv",row.names=1)
> altmark$sichtflaeche
```

Find out in which region the visible area is more equal (less disperse).

Streuung Aufgabe

Determine the dispersion of the data

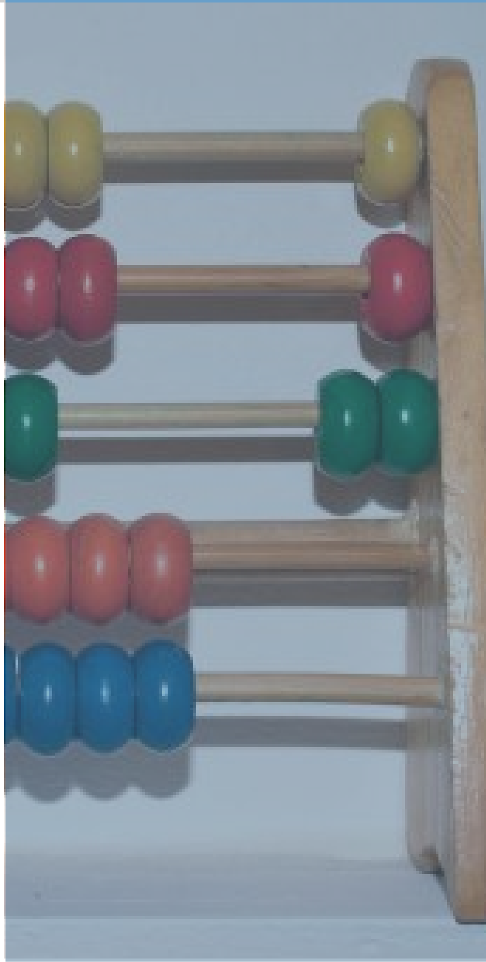
Analyse the sizes of areas visible from different megalithic graves of the Altmark (Demnick 2009):

```
> altmark<-read.csv2("altmark_denis.csv",row.names=1)
> altmark$sichtflaeche
```

Find out in which region the visible area is more equal (less disperse).

```
> sd(altmark[altmark$region=="Mitte",1])
[1] 60.56687
> sd(altmark[altmark$region=="Ost",1])
[1] 51.46048
> sd(altmark[altmark$region=="West",1])
[1] 28.73535
```

The standard deviation is the smallest for the region West, therefore are the visible areas more similar.



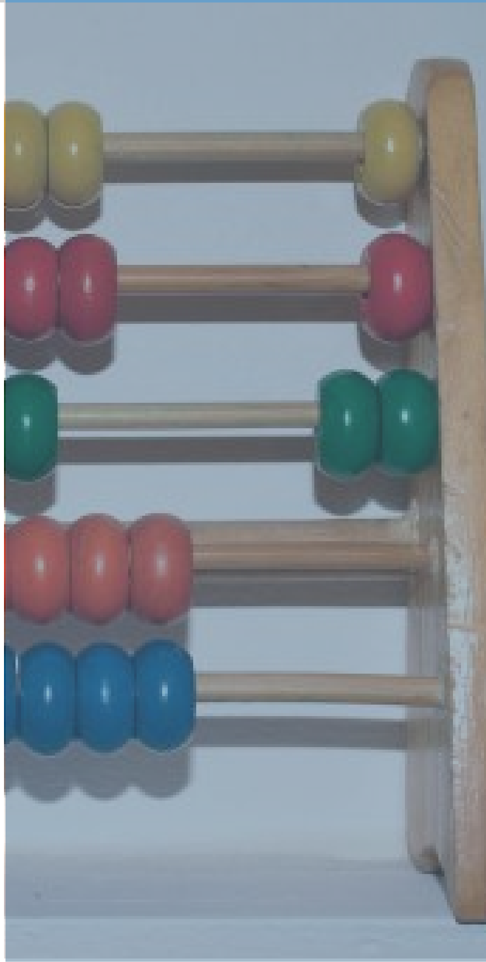
Shape of the distribution [1]

Important Parameters

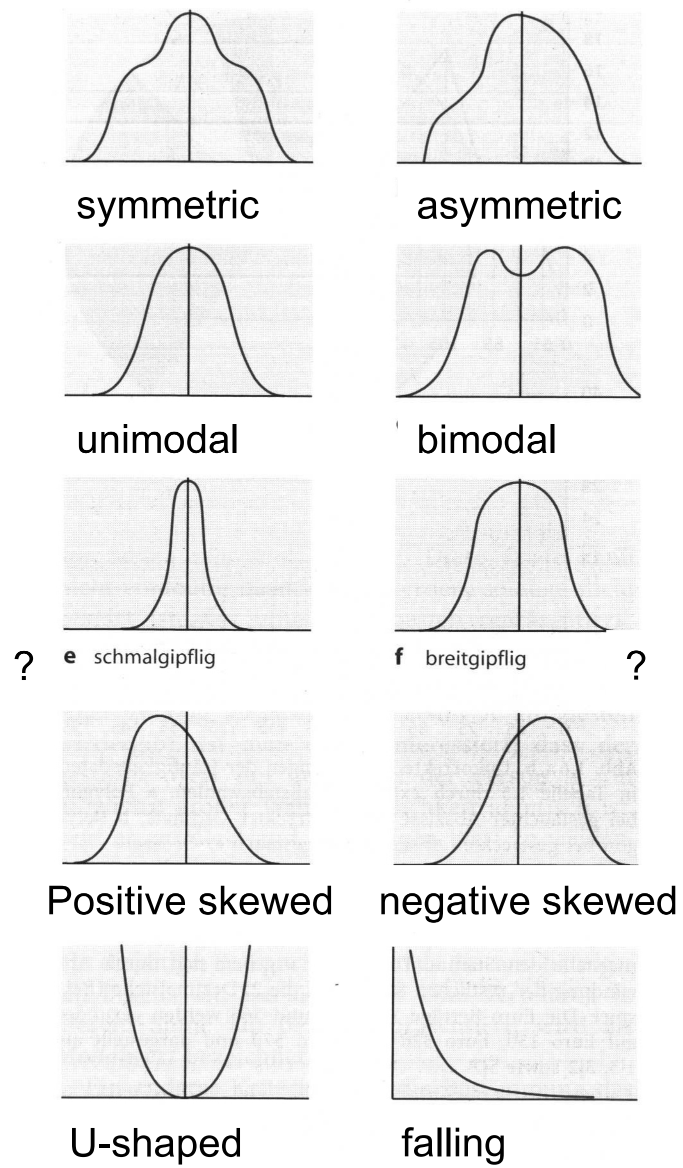
Number of peaks of the distribution: unimodal, bimodal, multimodal

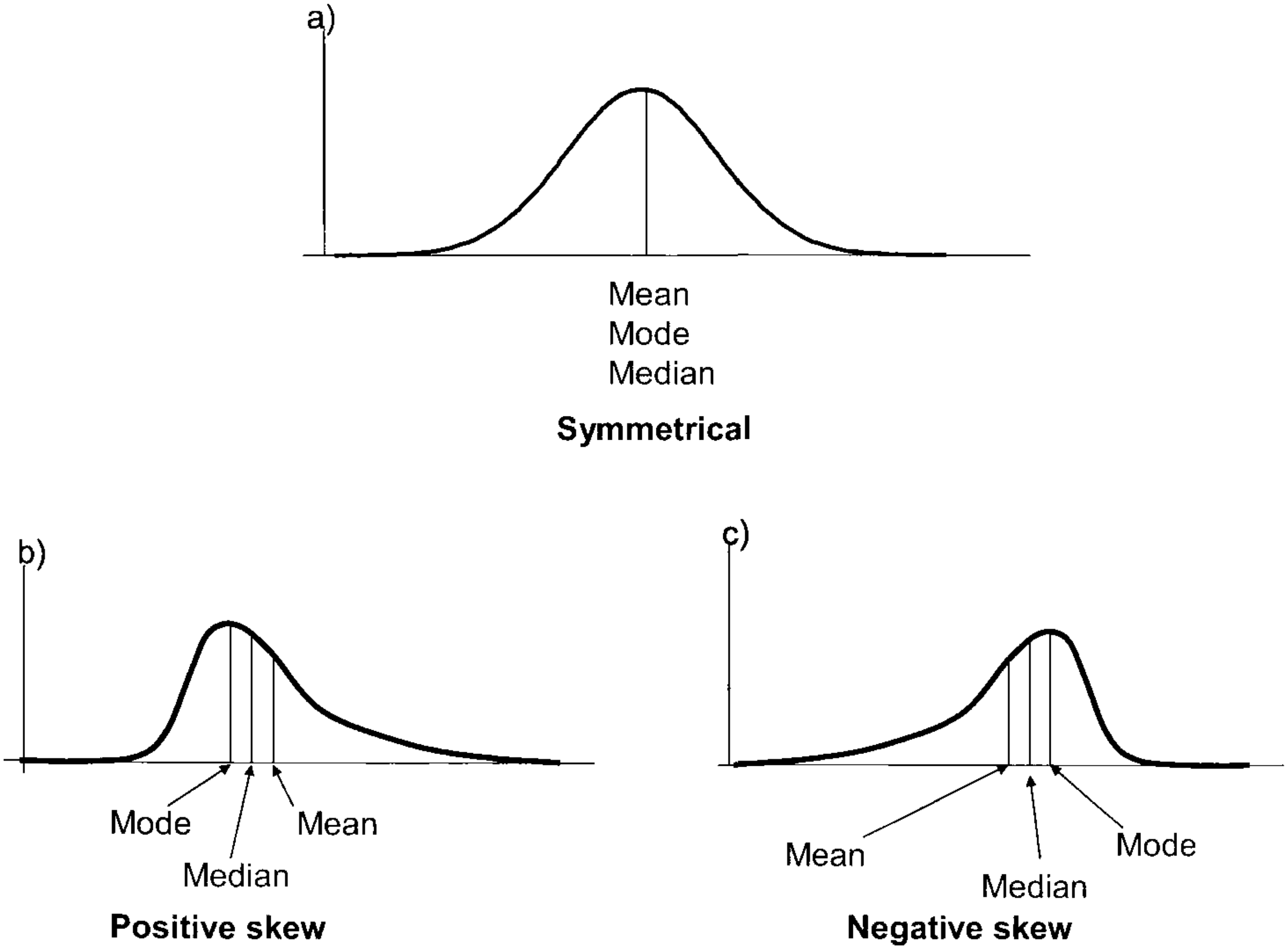
Skewness of the distribution: positive, negative

Curtosis (curvature) of the distribution: flat, medium, steep

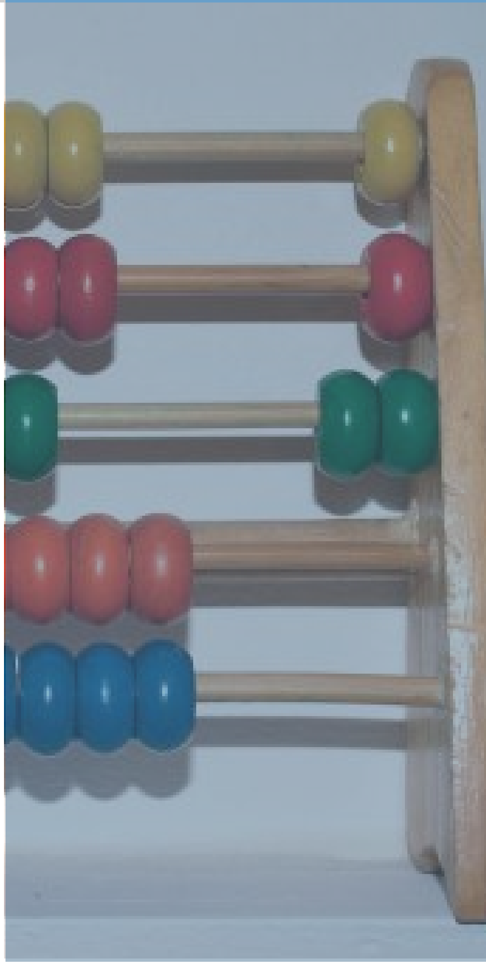


Shape of distributions (after Bortz 2006)




a)
Mean
Mode
Median
Symmetrical
b)
Mode
Mean
Median
Positive skew
c)
Mean
Mode
Median
Negative skew

Shape of the distribution [2]



Skewness

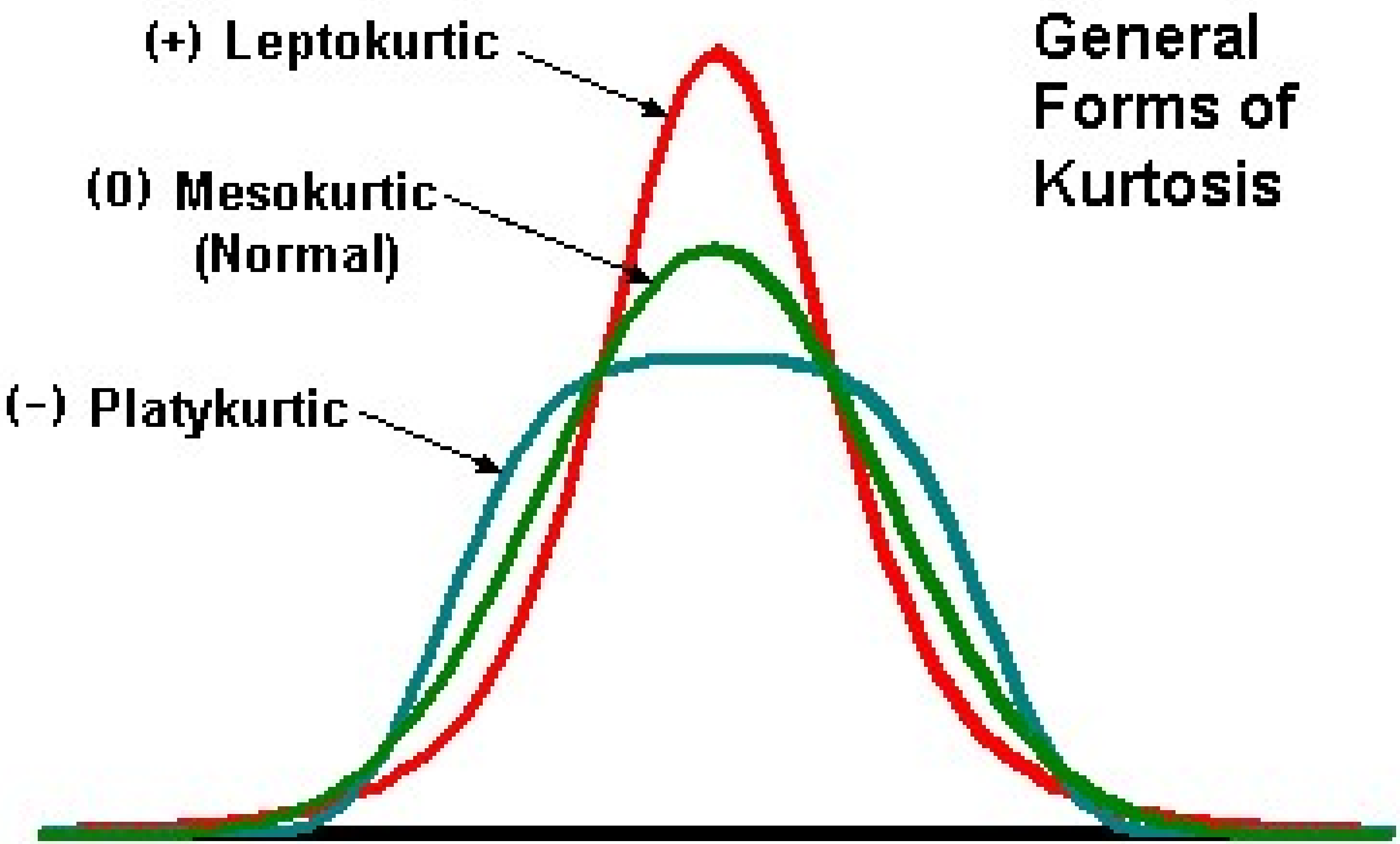
Mean right or left of the median
Read from the chart ;-)

calculate: $$\hat{S} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^3}{n * s^3}$$

Positive value indicates positive skew, negative resp.

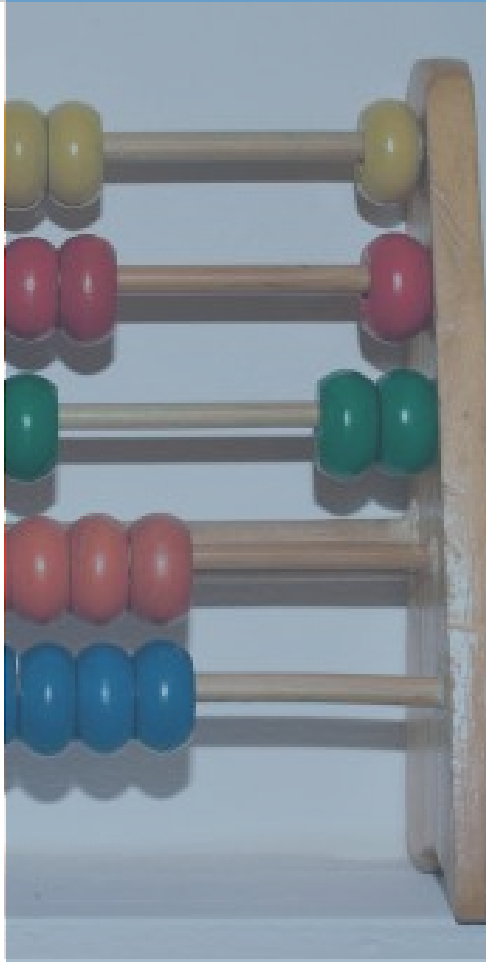
In R:

```
schiefe <- function (x) {
+ m3 <- sum((x-mean(x))^3) #Zähler
+ skew <- m3 / ((sd(x)^3)*length(x)) #Nenner
+ skew}
> test<-c(1,1,1,1,1,1,1,1,1,1,2,3,4,5)
> schiefe(test)
[1] 1.406826
> test<-c(3,3,3,3,3,3,3,3,3,3,3,3,2,1)
> schiefe(test)
[1] -2.231232
```



Quelle: http://grants.hhp.coe.uh.edu/doconnor/PEP6305/Topic%20002%20Organizing%20Data2.3.htm

Shape of the distribution [3]



Curtosis

The curvature of the distribution
Read from the chart ;-)

calculate:

$$K=\frac{\sum_{i=1}^{n}(x_i-\bar{x})^4}{n*s^4}-3$$

Positive if steeper, negative if flatter curve than the normal distribution

In R:

```
> kurtosis <- function (x) {
+ m3 <- sum((x-mean(x))^4)
+ skew <- m3 / ((sd(x)^4)*length(x))-3
+ skew}
> test<-c(1,2,3,4,4,5,6,7)
> kurtosis(test)
[1] -1.46875
> test<-c(1,2,3,4,4,4,4,4,4,4,4,4,4,4,4,4,4,4,4,5,6,7)
> kurtosis(test)
[1] 2.011364
```