UNIVERSITÄT BERN

Statistical methods for archaeological data analysis I: Basic methods

04 - Descriptive statistics

Martin Hinz

Institut für Archäologische Wissenschaften, Universität Bern

18.03.2025

You can download a pdf of this presentation.

Loading data for the following steps

download data

- muensingen_fib.csv

Read the Data on Muensingen Fibulae

```
muensingen <- read.csv2("muensingen_fib.csv")
head(muensingen)
```

```
##    X Grave Mno FL BH BFA FA CD BRA ED FEL  C   BW  BT FEW Coils Length
## 1  1   121 348 28 17   1 10 10   2  8   6 20  2.5 2.6 2.2     4     53
## 2  2   130 545 29 15   3  8  6   3  6  10 17 11.7 3.9 6.4     6     47
## 3  3   130 549 22 15   3  8  7   3 13   1 17  5.0 4.6 2.5    10     47
## 4  8   157  85 23 13   3  8  6   2 10   7 15  5.2 2.7 5.4    12     41
## 5 11   181 212 94 15   7 10 12   5 11  31 50  4.3 4.3  NA     6    128
## 6 12   193 611 68 18   7  9  9   7  3  50 18  9.3 6.5  NA     4    110
##   fibula_scheme
## 1             B
## 2             B
## 3             B
## 4             B
## 5             C
## 6             C
```

Descriptive 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

Central tendency [1]

mean

The classical. Suitable for metric data (interval or ratio) Sum of values/number of values, or

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

```
sum(muensingen$Length) / length(muensingen$Length)
```

```
## [1] 57.58824
```

```
mean(muensingen$Length)
```

```
## [1] 57.58824
```

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
##    2
```

Central tendency [4]

Variable is

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

after: Dolić 2004

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))
```

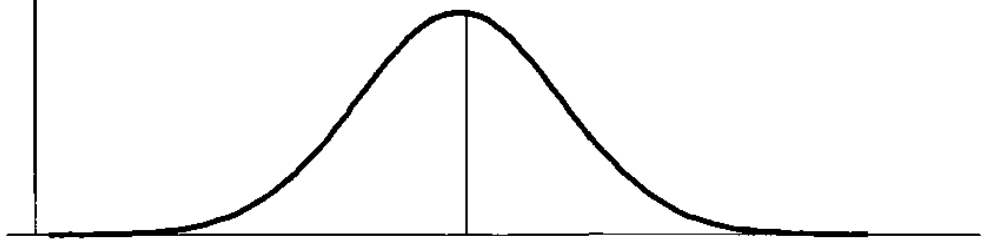
```
## 3
## 3
```

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
```

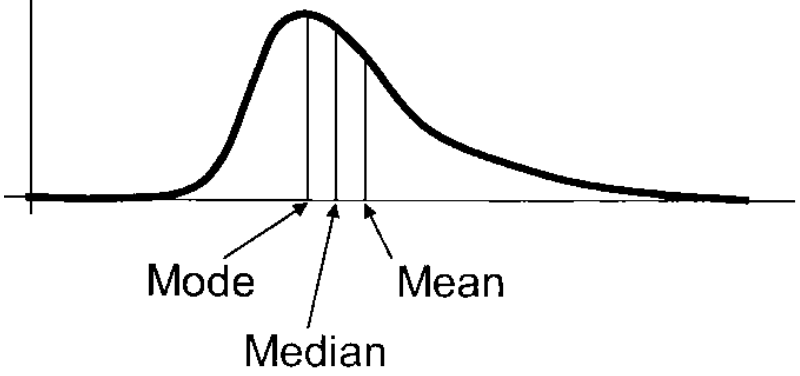
a)



Mean
Mode
Median

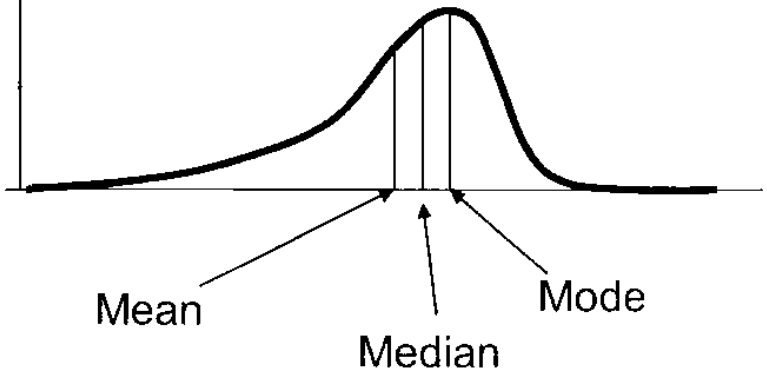
Symmetrical

b)



Positive skew

c)



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.csv

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

```
##  [1] 12.0 19.5 18.6 12.9 13.2  9.9 19.5  8.4 21.0 18.9  7.5 18.9  8.1  9.0  7.8
## [16]  9.9 10.2  8.1 12.0  9.0 26.1 20.4
```

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

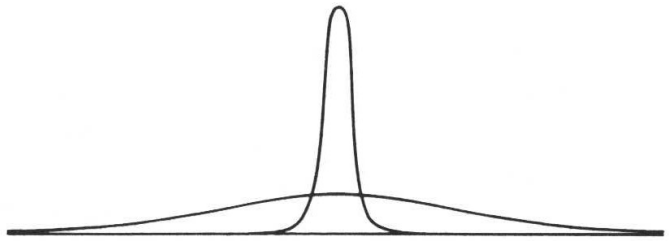
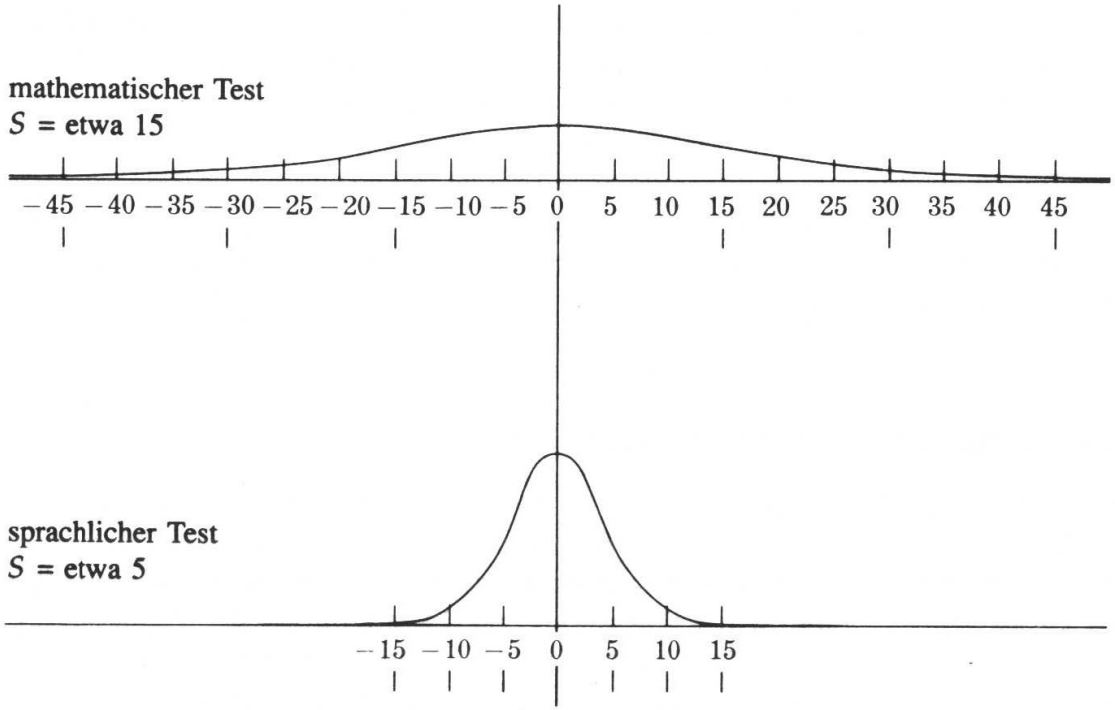


Abb. 4.1 Zwei Verteilungen mit denselben *N*s, aber unterschiedlicher Streuung.



source: Phillips 1997

Dispersion [1]

Range

Simply the range of the values of a data vector.

```
range(muensingen$Length)
```

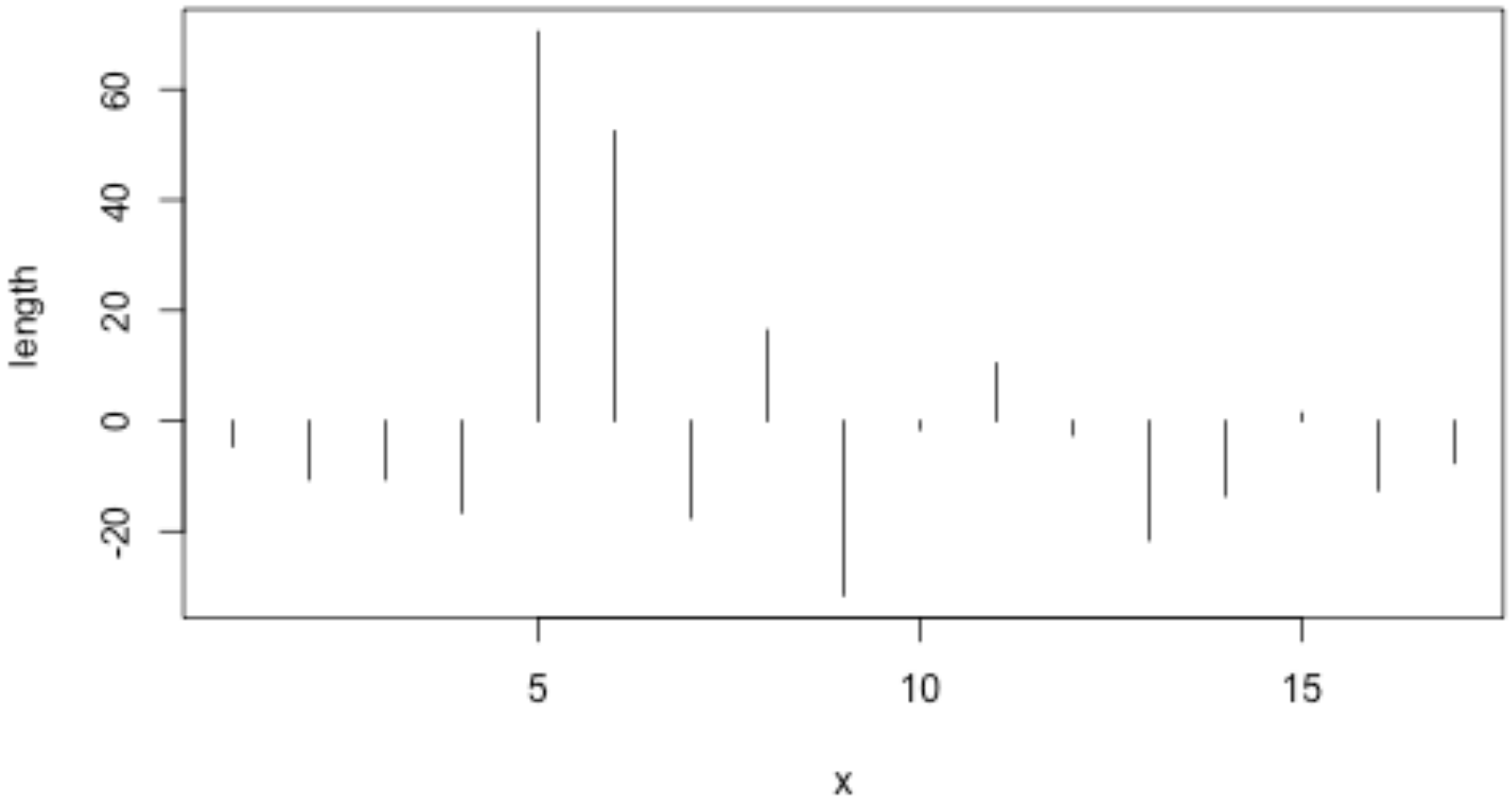
```
## [1]  26 128
```

```
range(tassen$x)
```

```
## [1]  7.5 26.1
```

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

How far deviates the individual values from the mean in the mean?



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)
```

```
## [1] 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

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))
```

```
## [1] 5.577756
```

```
sd(tassen$x)
```

```
## [1] 5.577756
```

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 length and total length equal?

```
sd(muensingen$Length)/mean(muensingen$Length)
```

```
## [1] 0.4508988
```

```
sd(muensingen$FL)/mean(muensingen$FL)
```

```
## [1] 0.7732486
```

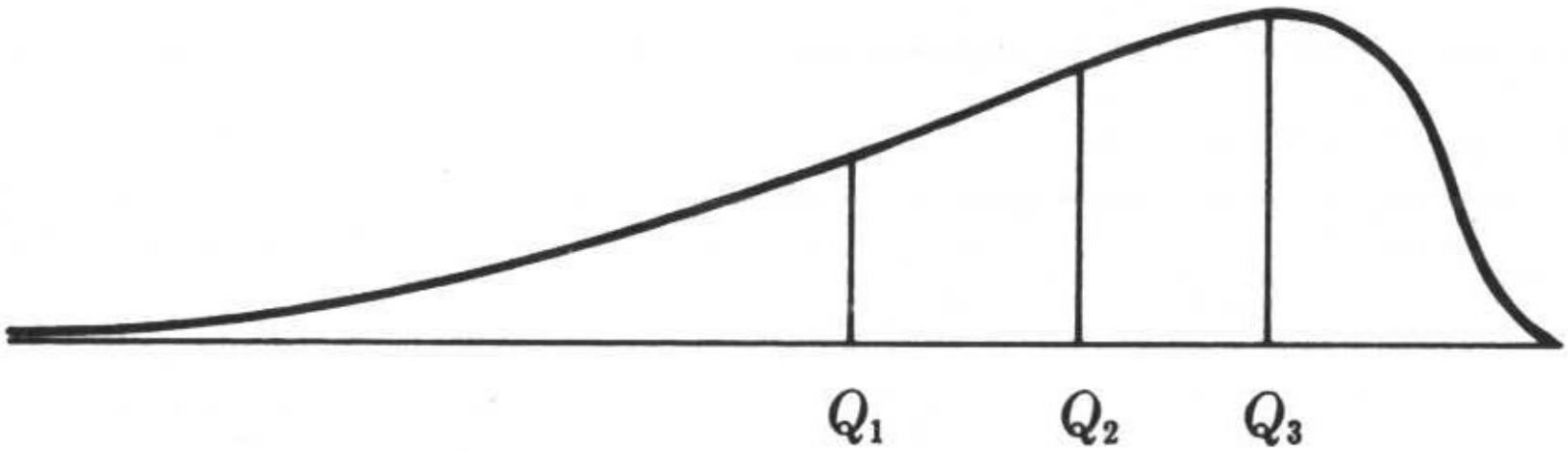
Foot length vary more than total length

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.

Phillips 1997

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_denis2.csv

```
altmark<-read.csv2("altmark_denis2.csv",row.names=1)
head(altmark)
```

```
##          sichtflaeche region
## La01             2.72  Mitte
## Lg1             26.78  Mitte
## Li02            26.96  Mitte
## Sa01            27.05  Mitte
## Li06            32.93  Mitte
## K\xf601         34.76  Mitte
```

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

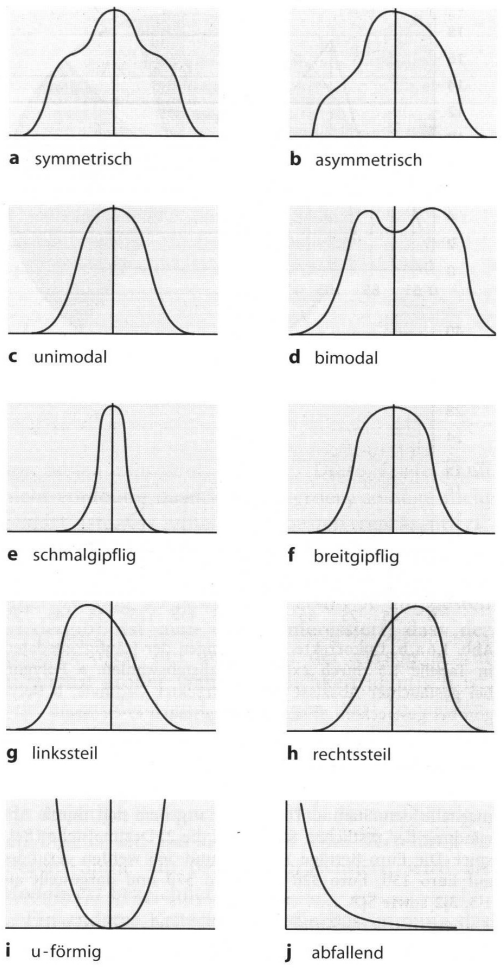
Shape of the distribution [1]

Important Parameters

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

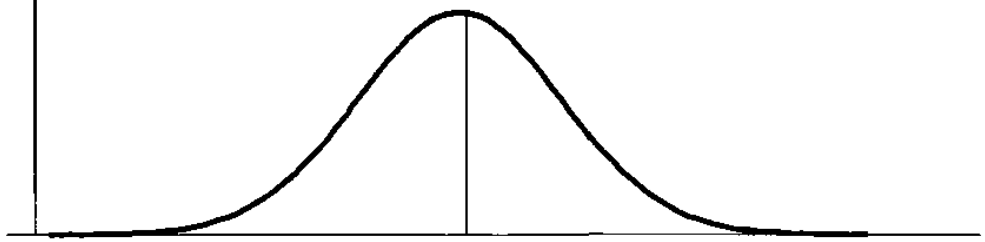
Skewness of the distribution: positive, negative

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



Shape of distributions (after Bortz 2006)

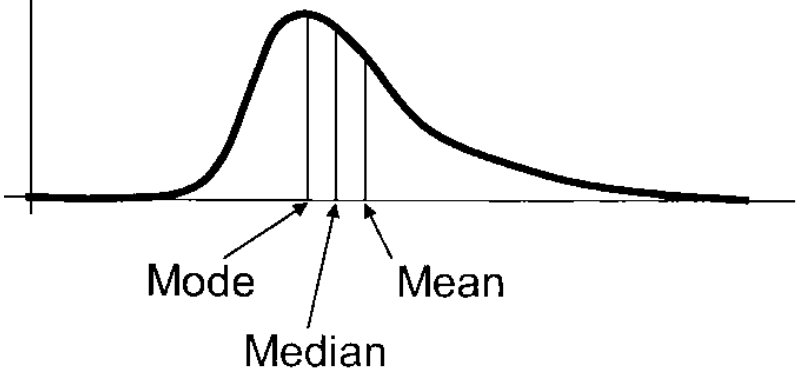
a)



Mean
Mode
Median

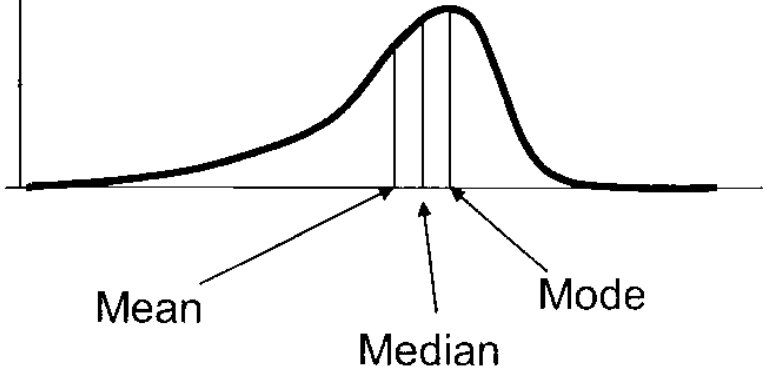
Symmetrical

b)



Positive skew

c)



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.

Shape of the distribution [2]

Skewness

There is no function in R currently available to calculate this. So we build our own:

```
skewness <- function(x) {
  m3 <- sum((x-mean(x))^3) #numerator
  skew <- m3 / ((sd(x)^3)*length(x)) #denominator
  skew
  }
```

Test:

```
test<-c(1,1,1,1,1,1,1,1,1,1,2,3,4,5)
skewness(test)
```

```
## [1] 1.406826
```

```
test<-c(3,3,3,3,3,3,3,3,3,3,3,3,2,1)
skewness(test)
```

```
## [1] -2.231232
```

Shape of the distribution [3]

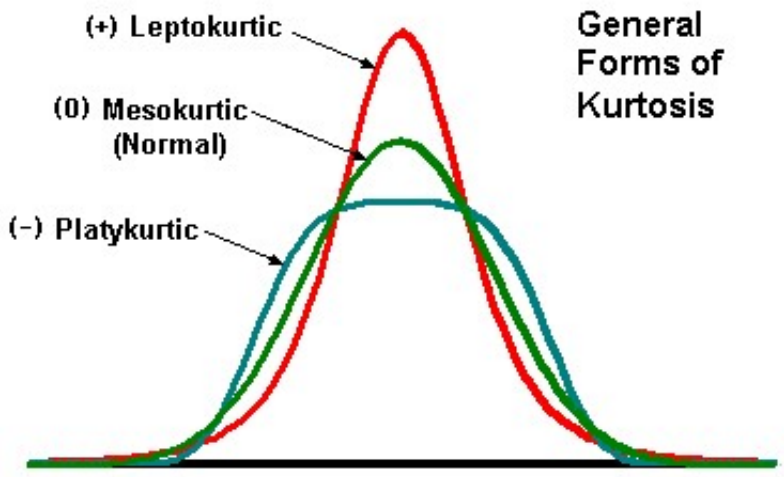
Kurtosis

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

calculate:

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

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



Shape of the distribution [3]

Kurtosis

We write a function for that, too:

```
kurtosis <- function (x) {
  m3 <- sum((x-mean(x))^4)
  skew <- m3 / ((sd(x)^4)*length(x))-3
  skew
  }
```

Test:

```
test<-c(1,2,3,4,4,5,6,7)
kurtosis(test)
```

```
## [1] -1.46875
```

```
test<-c(1,2,3,4,4,4,4,4,4,4,4,4,4,4,4,4,4,4,5,6,7)
kurtosis(test)
```

```
## [1] 2.011364
```