Biostatistics Week 2

Topics this week:

- **uni-variate descriptive Analysis**
  - **Recap: Data types**
  - **Measure for location: mean, median, mode, quantiles**
  - **Visualizing categorical variables: pie chart and barplots**
  - **Visualizing continuos variables: histogram and boxplot**

- **Bi-variate descriptive Analysis**
  - **Continous vs continuous: scatterplot**
  - **Categorical vs categorical: mosaicplot**
  - **Continous vs categorical: grouped boxplots or stripcharts**

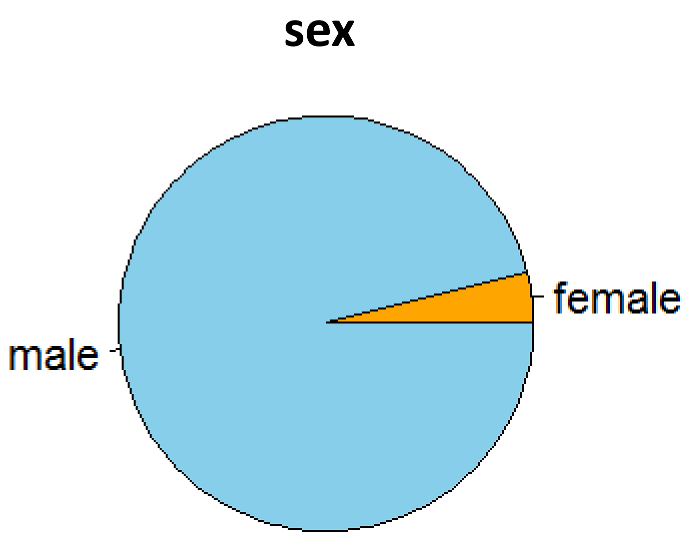
https://bsick.github.io/Biostatistics-Fall-2018/

There are different types of data

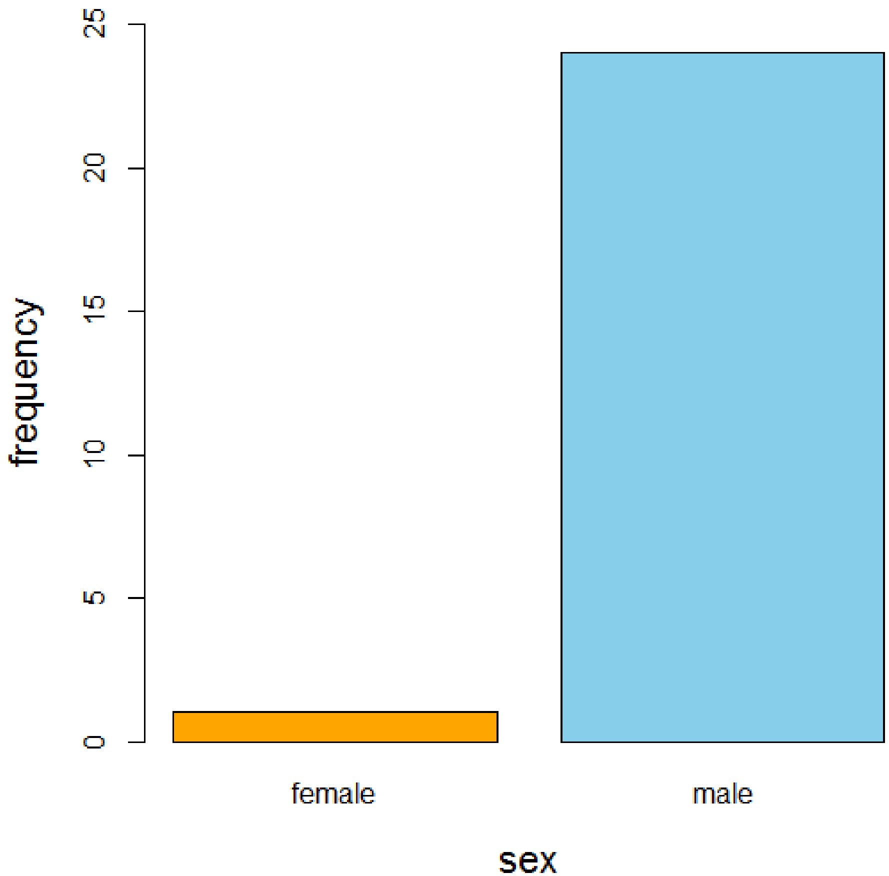
- **Qualitative categorical**
  (predefined qual. values/levels can be observed)
  - **Nominal**
    (no meaninful order of the levels)
  - **Ordinal**
    (there is a meaningful order of the levels)
- **Quantitative numeric**
  (numeric values with which one can do calculations)
  - **discrete**
    (we can count the possible values)
  - **continuous**
    (between any two values there is a value in between)

How to summarize categorical data?

pie chart



bar chart

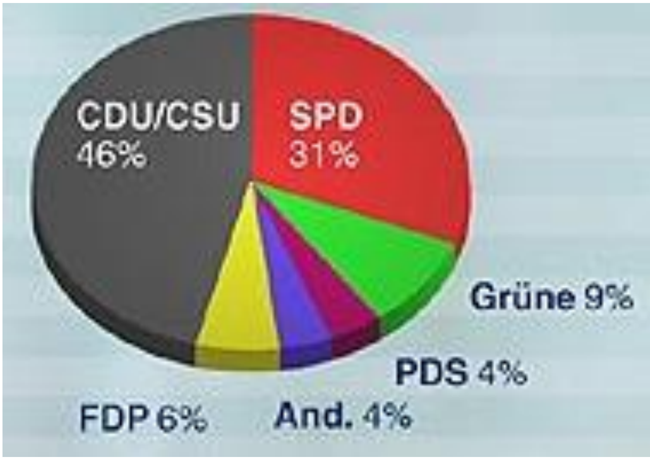


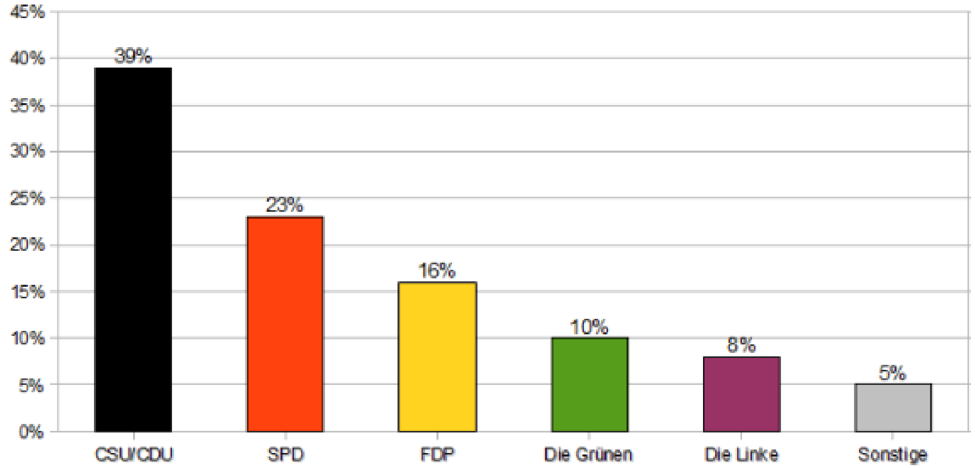
`pie(table(dat$sex))`

`barplot(table(dat$sex))`

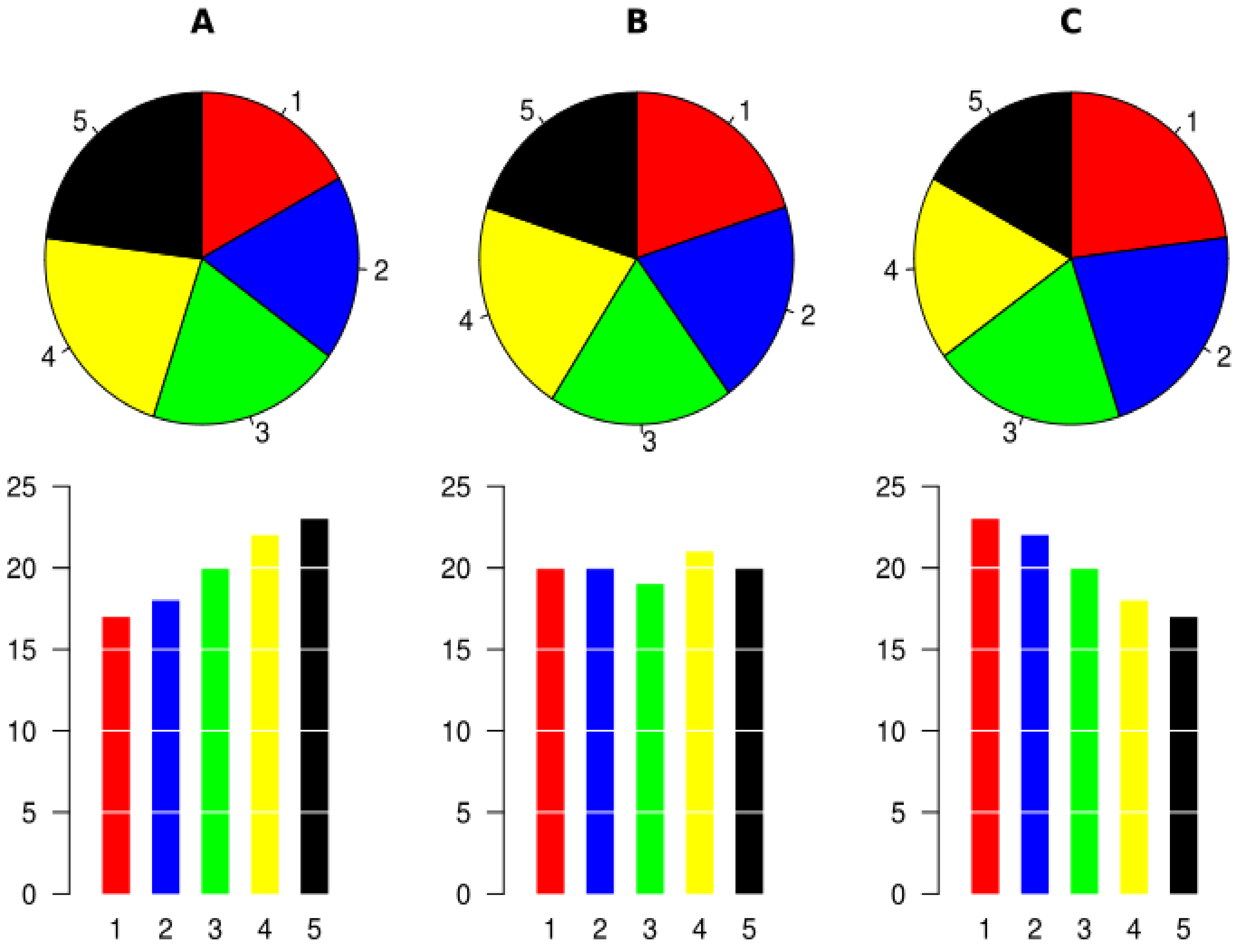
Visualizing categorical data by Bar-Chart or Pie-Chart

These charts are simple – is there room form manipulation?



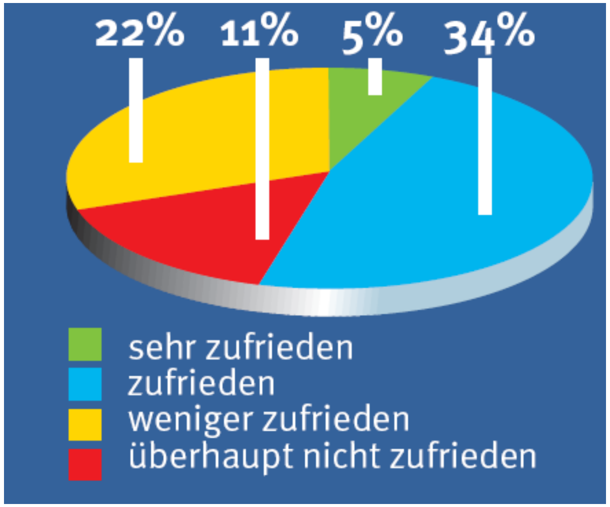


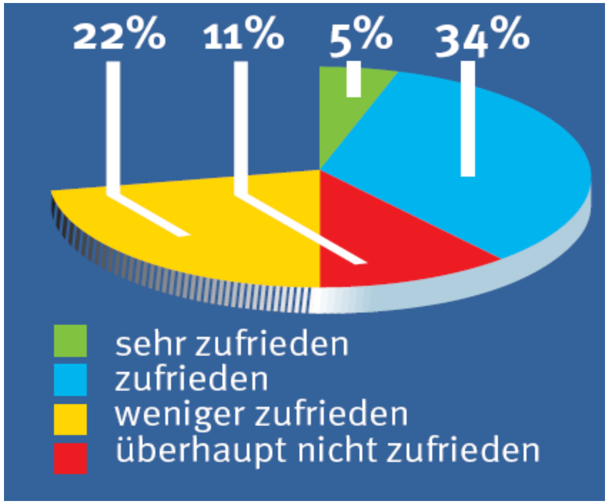
Barplots are often to prefere over pie-charts



Humans are much better in comparing heights than comparing areas.

Half of all reader are satisfied with Klinsmann - true?





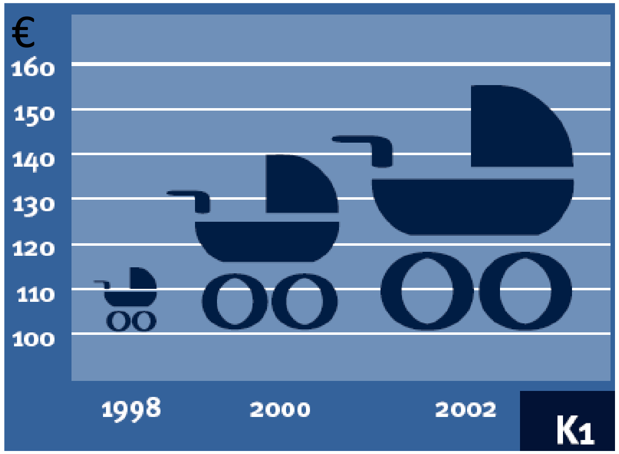
Pie-Chart from the German newspaper „Bild“ . Reader were asked how satisfied they are with soccer trainer Klinsmann.

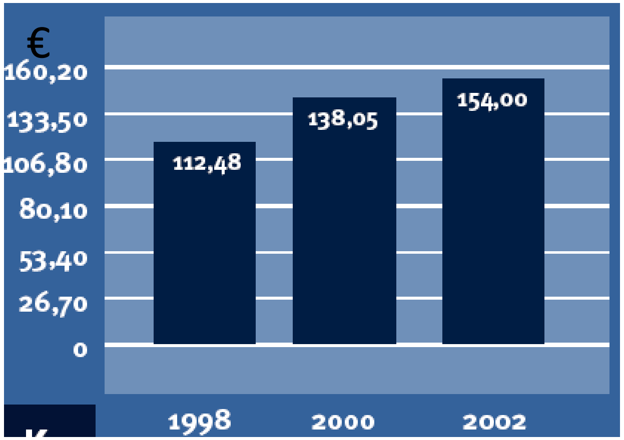
Manipulation:
pie segment does not correspond to percentage.

trick:
Percentages do not add up to 100%

Quelle: http://www.wdr.de.varnish.gl-systemhaus.de/tv/quarks/sendungsbeitraege/2006/1017/002_zahlen.jsp?smonat=2009-10

Generous increase of child allowance - true?



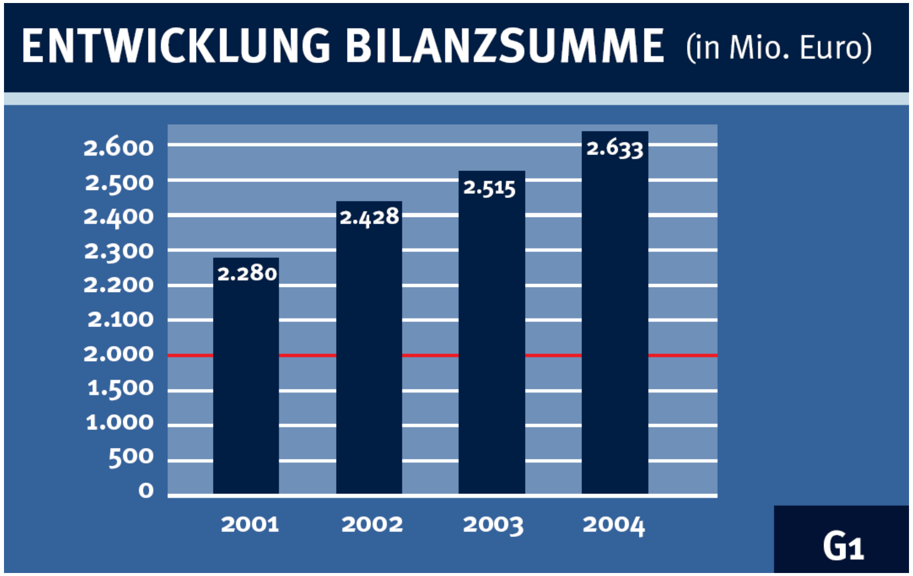


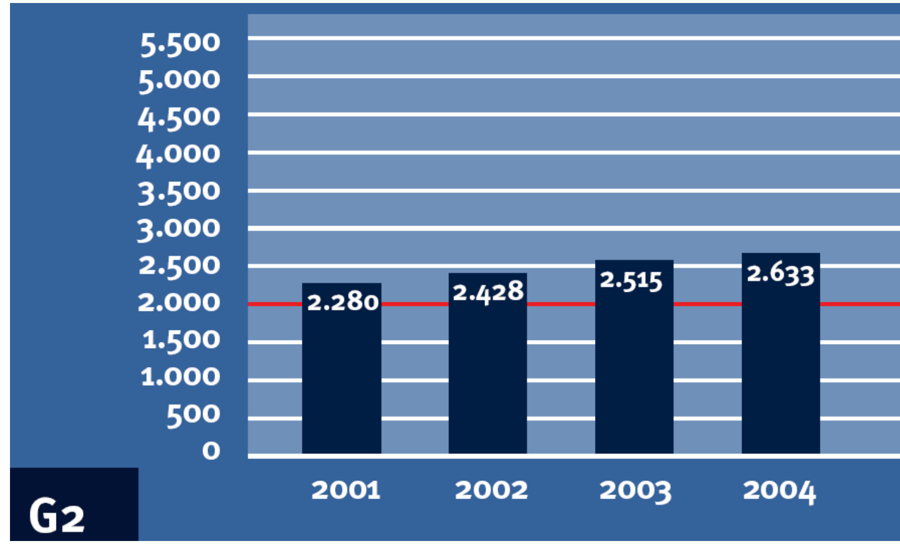
Graph from government statement in the German red-green agenda 2010.

Manipulation: area of the buggies are not proportional to the child allowance.

trick: scale starts not at 0 and the shape of the buggies are almost circles where the area increases proportional to the square of the height.

Good business development - true?





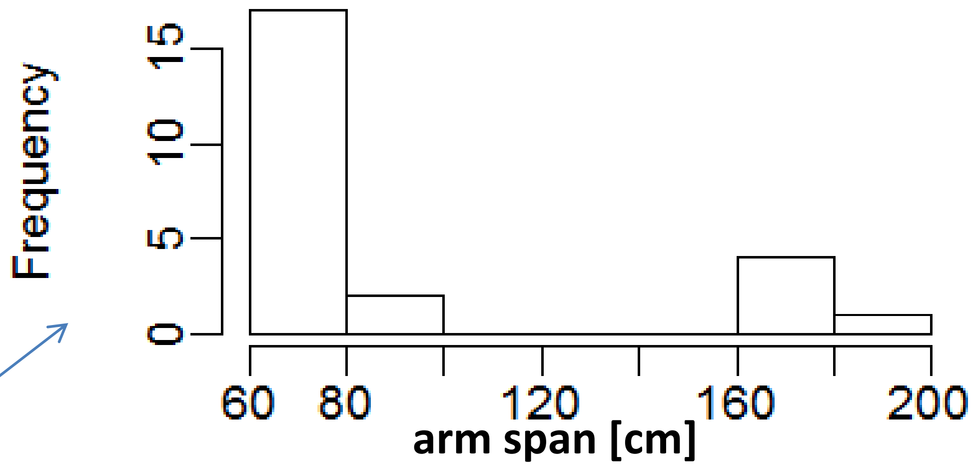
Bar Chart in the business report of a german bank
(psd-Bank Rhein-Ruhr 2004)

Manipulation: bar heights are not proportional to actual numbers.
trick: scale is changing above 2000.

How to summarize continuous data – e.g. arm span?

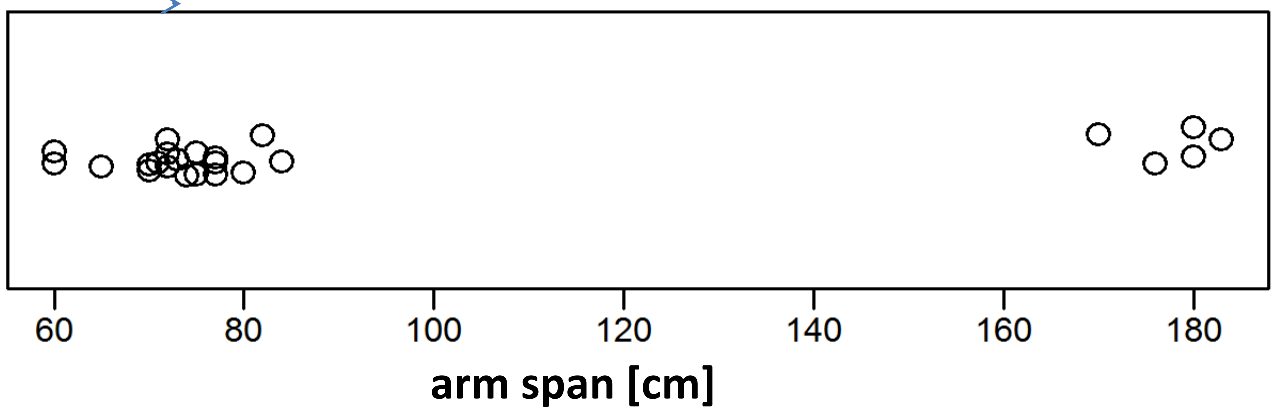
X: arm span	frequency
[60, 80)	17
[80,100)	2
[100,120)	0
[120,140)	0
[140,160)	0
[160,180)	4
[180,200)	1

- define non-overlapping classes/bins
- count number of observation per class
- draw histogram (no gaps between bars)



`hist(x)`

`stripchart(x,method="jitter")`



How to visualize continuous data?

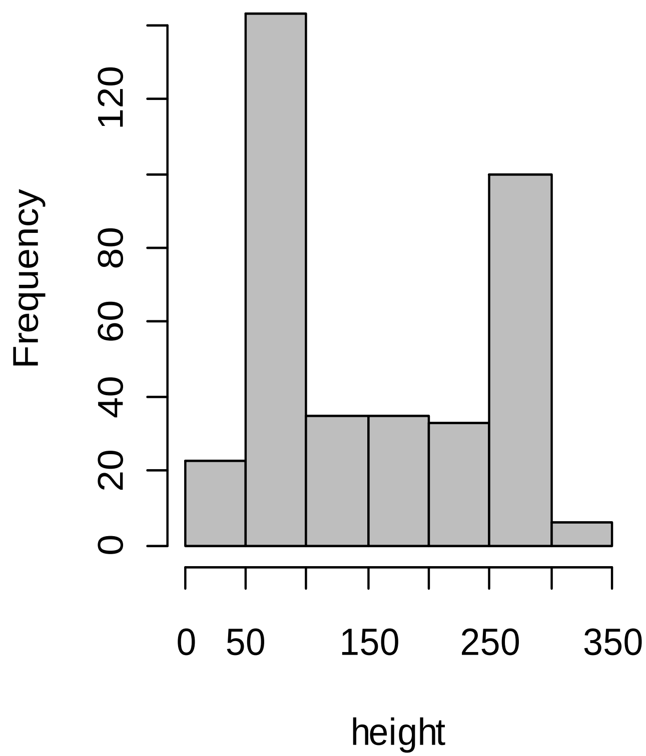
The height (cm) of 376 plants were measured.

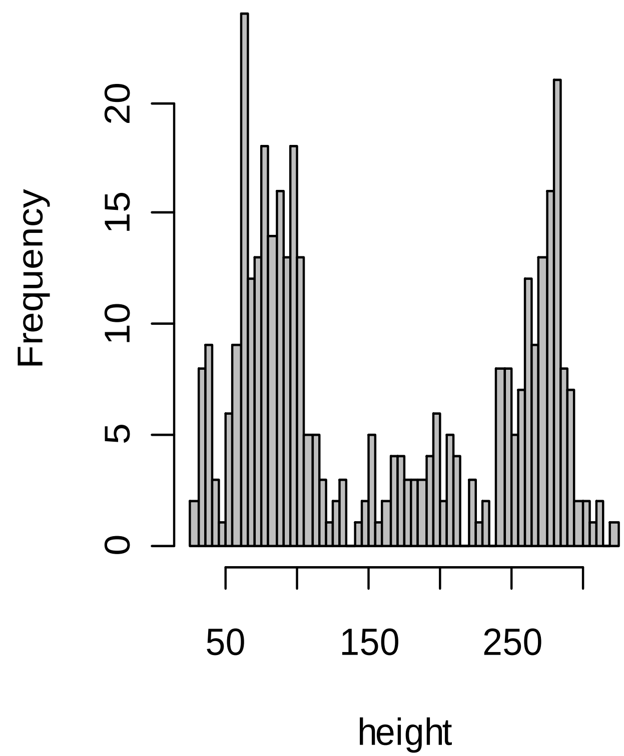
```
head(dat$height)
```

G
height
57.9
62.1
55.8
61.5
68
52.8
70.5
60.4
75.2
77.1
70.4
70.1
27.6
35

⋮

```
# hist, few classes, big bin width
hist(dat$height, nclass=7)
# hist, many classes, small bin width
hist(dat$height, nclass=100)
```

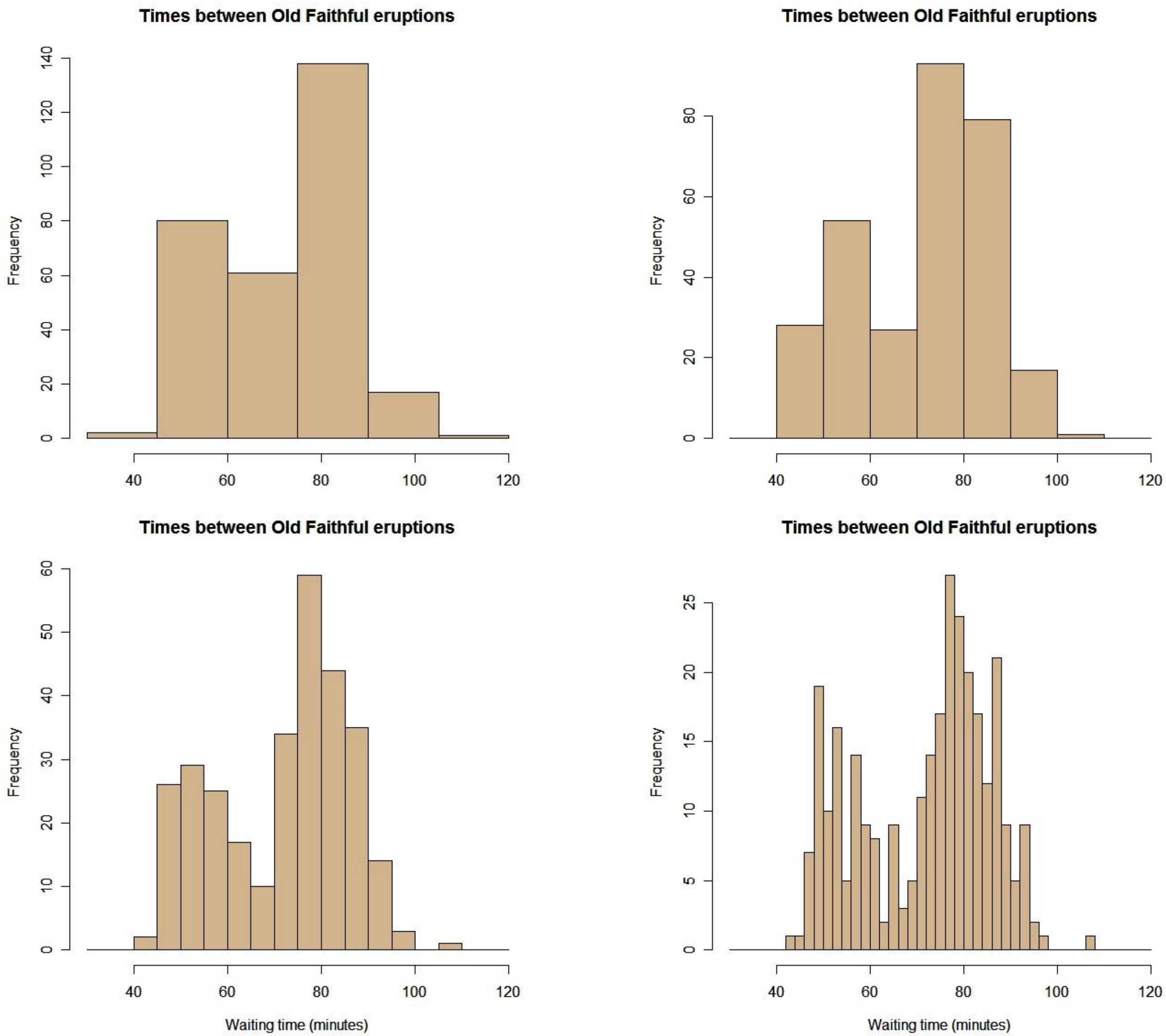




Are there subgroups? If yes – how many?

How reliable is the height of a bar?

How many classes do we need?

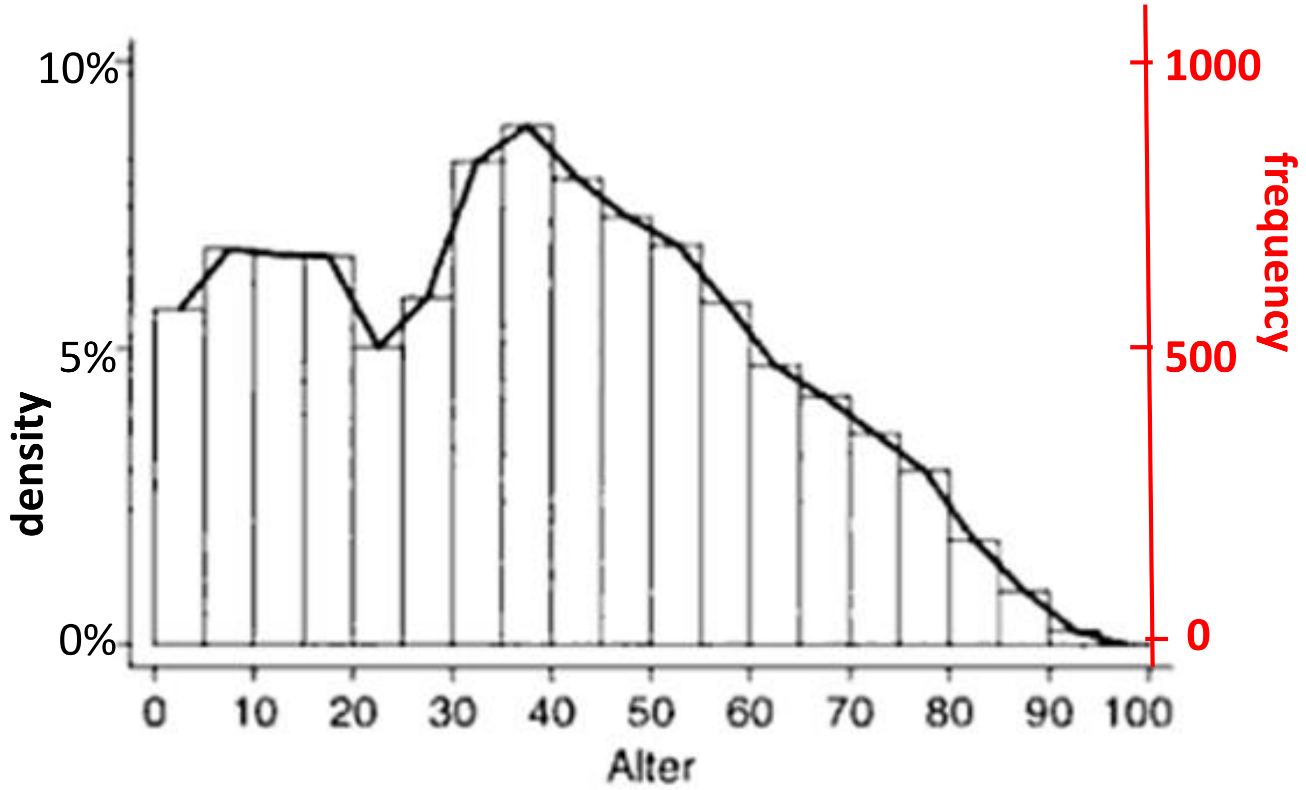


http://www.amstat.org/publications/jse/v6n3/applets/Histogram.html

299 eruption intervals were observed

Shape of the histogram may depend on the class choices

How does the distribution of age look like

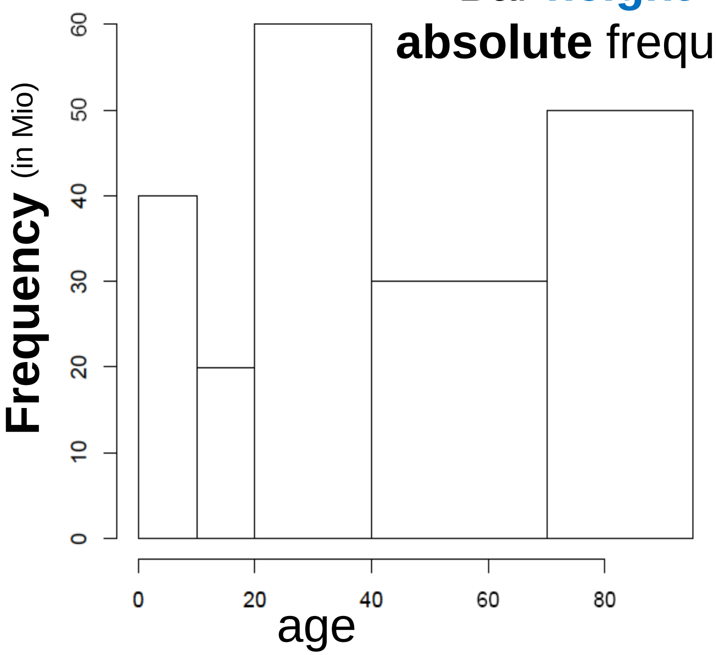


Only <u>in case of equally sized bins</u> the scaled and unscaled histogram look the same and it is possible to label the y-axis with percentages – usually it ony shows the density!

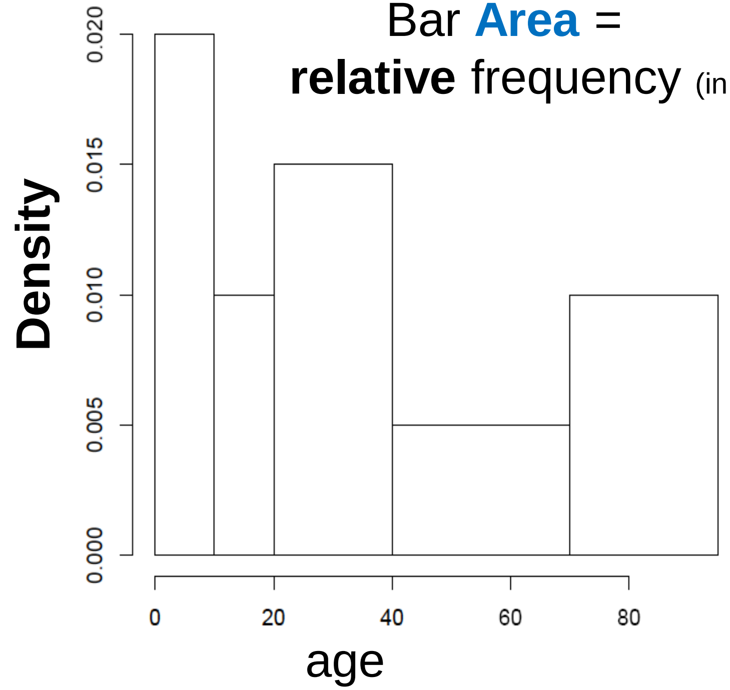
Unscaled and scaled histograms

Age	Total Population (Millions)	Percentage of Population
$[0, 10)$	40	20%
$[10, 20)$	20	10%
$[20, 40)$	60	30%
$[40, 70)$	30	15%
$[70, 95)$	50	25%

Bar **height** = **absolute** frequency



Bar **Area** = **relative** frequency (in %)

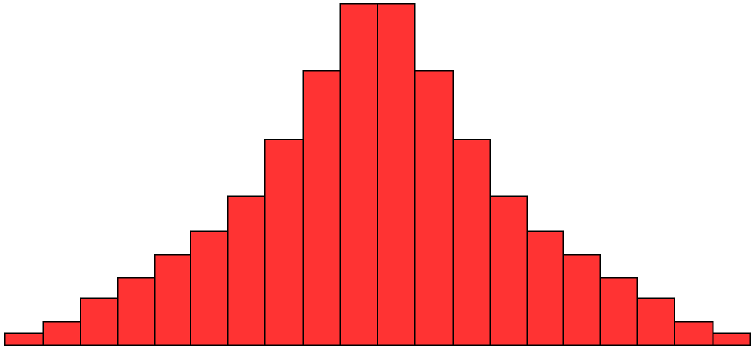


Rules for histograms

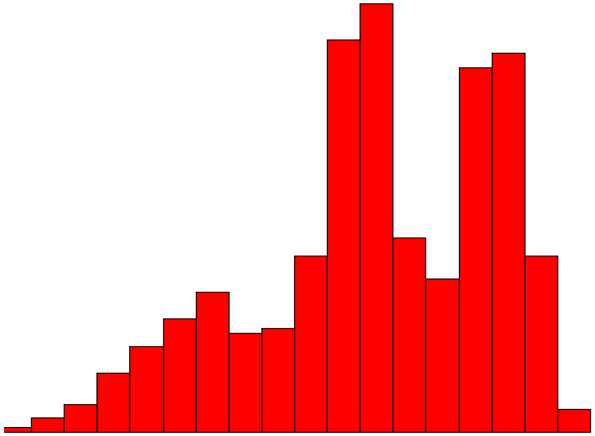
- Avoid classes with different width! (shape will change)
- How many classes: $\sqrt{n}$ classes for n observations.
- The shape can depend on the number classes and the class limits.

Attention: in a scaled histogram the ***area*** of the bar indicates the relative frequency, whereas in a unscaled histogram the ***height*** of the bar indicates the absolute frequency -> in case of unequal bin-widths the shape of the unscaled and scaled histograms can differ substantially.

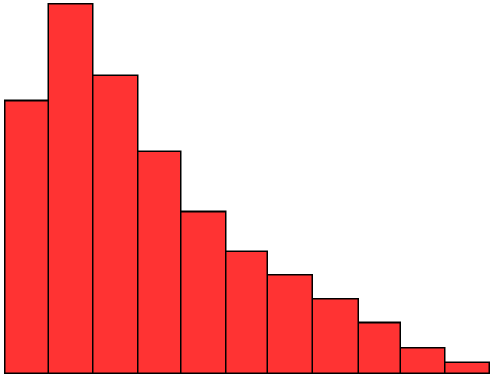
Shapes of distributions



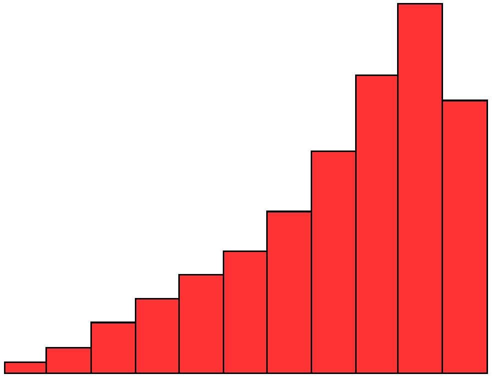
Symmetric, uni-modale



Multi-modale, slightly left-skewed



Right-skewed, uni modale

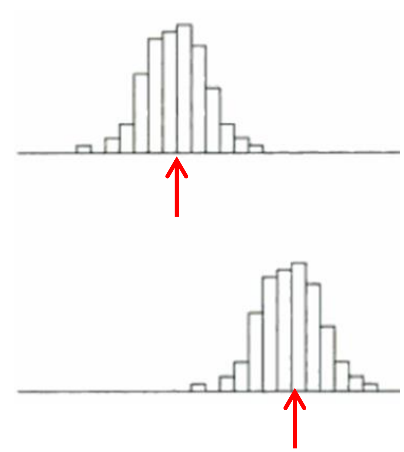


Left-skewed, uni-modale

Measures for the location and variation

Data can be summarized by summary statistics. Most important key figure describe the center and the width of a distribution..

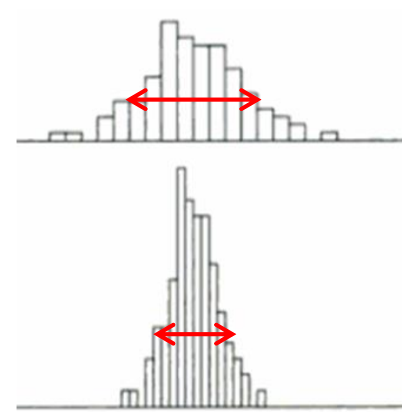
Measures for the location



Where is the center?

What is a typical value?

Measures for the variation



A number which quantifies the width of the distribution.

Bild: W. Stahel

Is the mean salary a «typical salary»?

The mean salary for Novartis employees was in 2009 around 220’000 CHF.

Willkommen bei **Novartis Schweiz**

Home | Über uns | F & E | Investoren | Medien | Soziale Verantwortung

Über uns — print | text | A+

Schweizer Arbeitsplätze

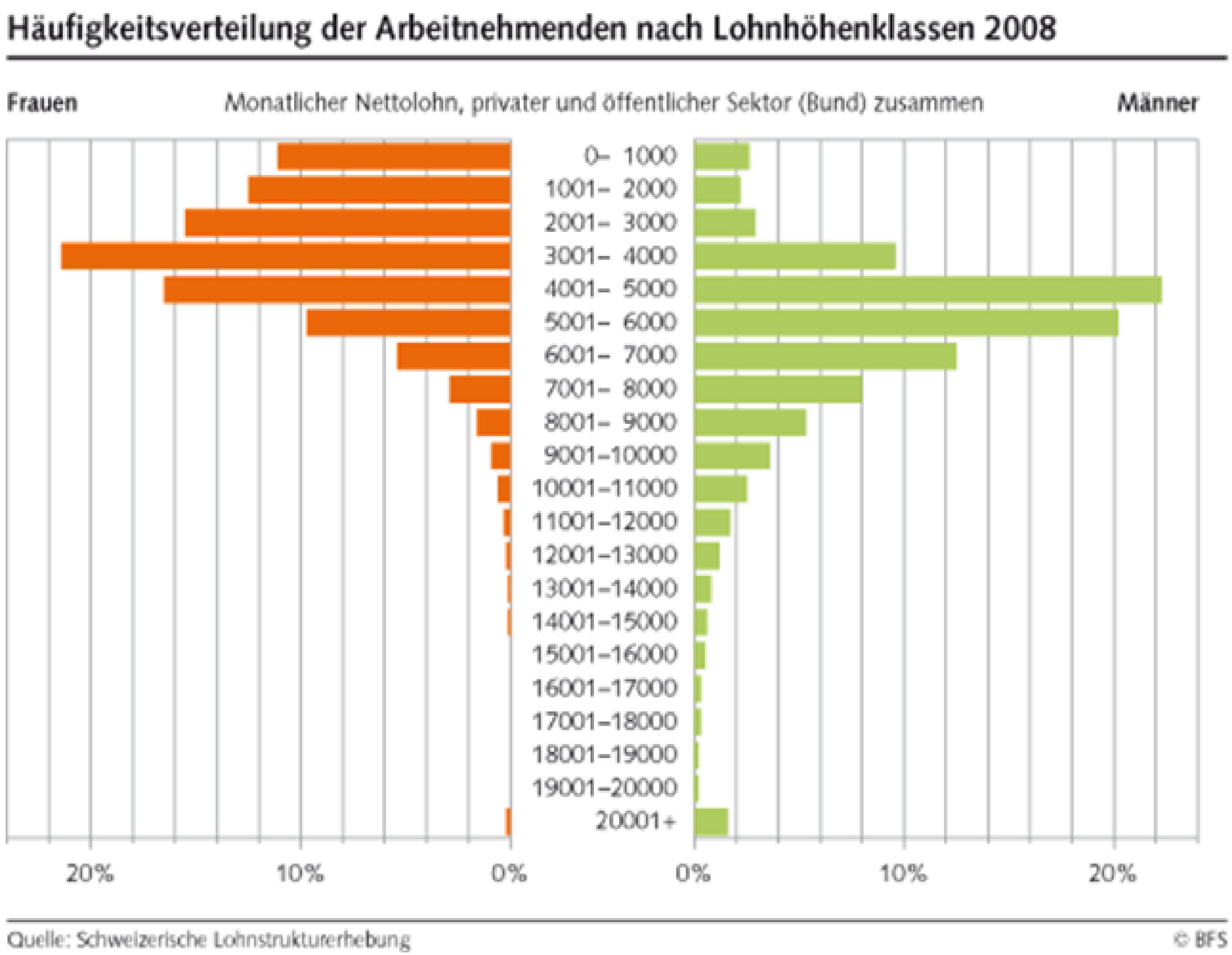
Grafik vergrössern

Novartis beschäftigt weltweit zurzeit rund 99.800 Mitarbeitende. Davon arbeiten rund 12.000 in der Schweiz – verteilt auf die acht Standorte in Basel BS/BL, Stein AG, Embrach ZH, Cham ZG, Bern BE, St-Aubin FR, Nyon VD und Locarno TI. Eine kürzlich veröffentlichte Studie hat ergeben, dass für jeden direkten Arbeitsplatz bei Novartis in der Schweiz indirekt 2,5 weitere Arbeitsplätze geschaffen werden.

Die Gesamtsumme der Lohn- und Sozialleistungen für Mitarbeitende von Novartis in der Schweiz betrug im Jahr 2009 rund 2,6 Milliarden Franken.

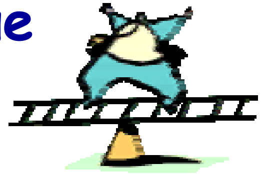
$$\text{mean.salary} \approx \frac{2.6Mrd.}{12000} \approx 220000 \ CHF$$

Distribution of salaries in Switzerland

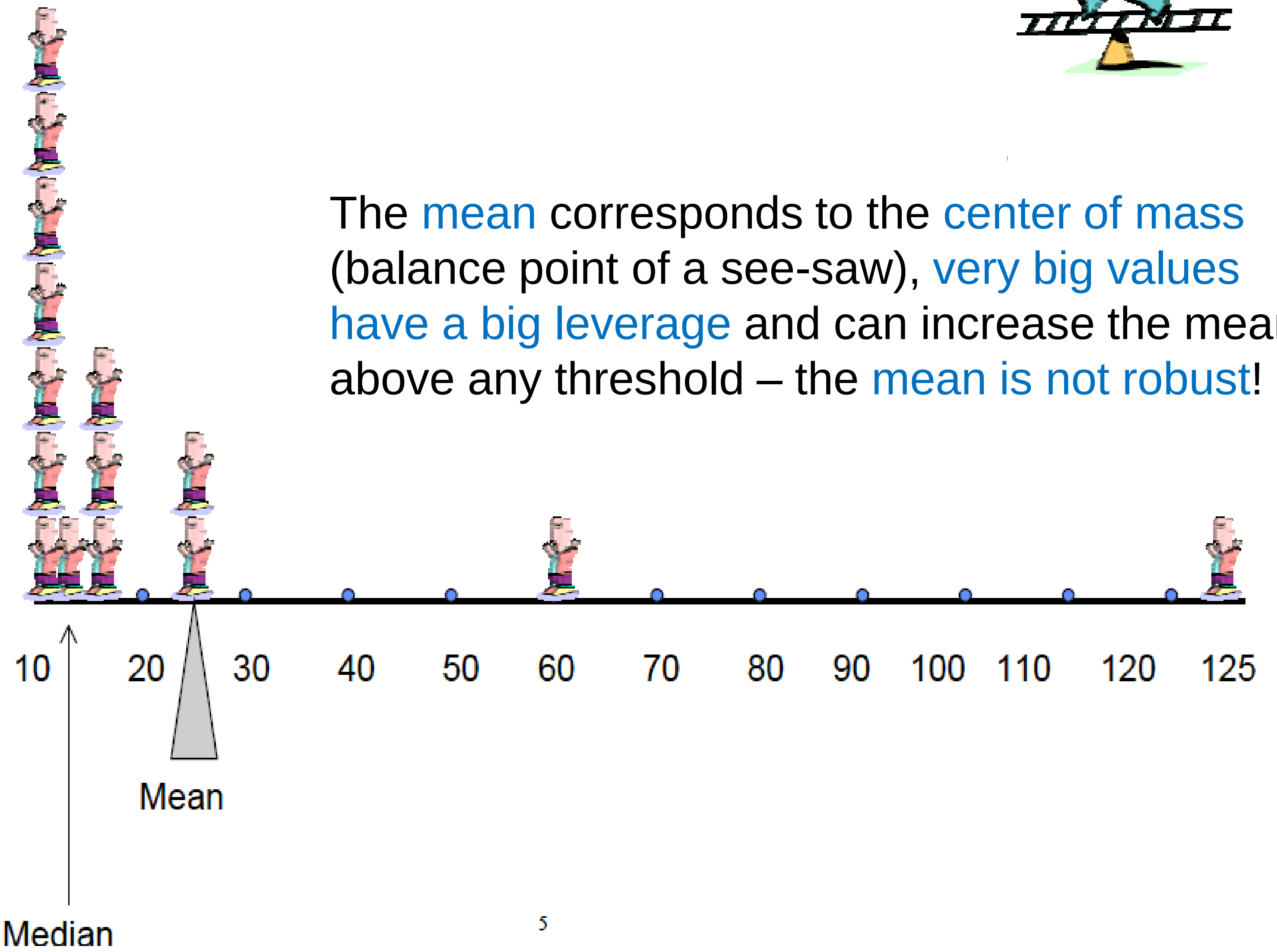


18

For right-skewed distributions the mean is not a typical value



The mean corresponds to the center of mass (balance point of a see-saw), very big values have a big leverage and can increase the mean above any threshold – the mean is not robust!



The Median

- Median (50% of all observations are smaller 50% are larger)
  - Order observation: take value in the center
  - 1,2,3,4,1000 => median=3
  - In case of an odd-numbered number of observation, take mean of the two center values:
  - 1,2,3,1000 => median=2.5
- Median
  - Create a ordered sample:

$$x_1, x_2, \dots, x_n \qquad x_{[1]}, x_{[2]}, \dots, x_{[n]}$$

$$\tilde{x} = \begin{cases} x_{\left[\frac{n+1}{2}\right]} & \text{, for odd n} \\ \frac{1}{2} \cdot \left(x_{\left[\frac{n}{2}\right]} + x_{\left[\frac{n}{2}+1\right]} \right) & \text{, for even n} \end{cases}$$

What is „the mean" income in this company?

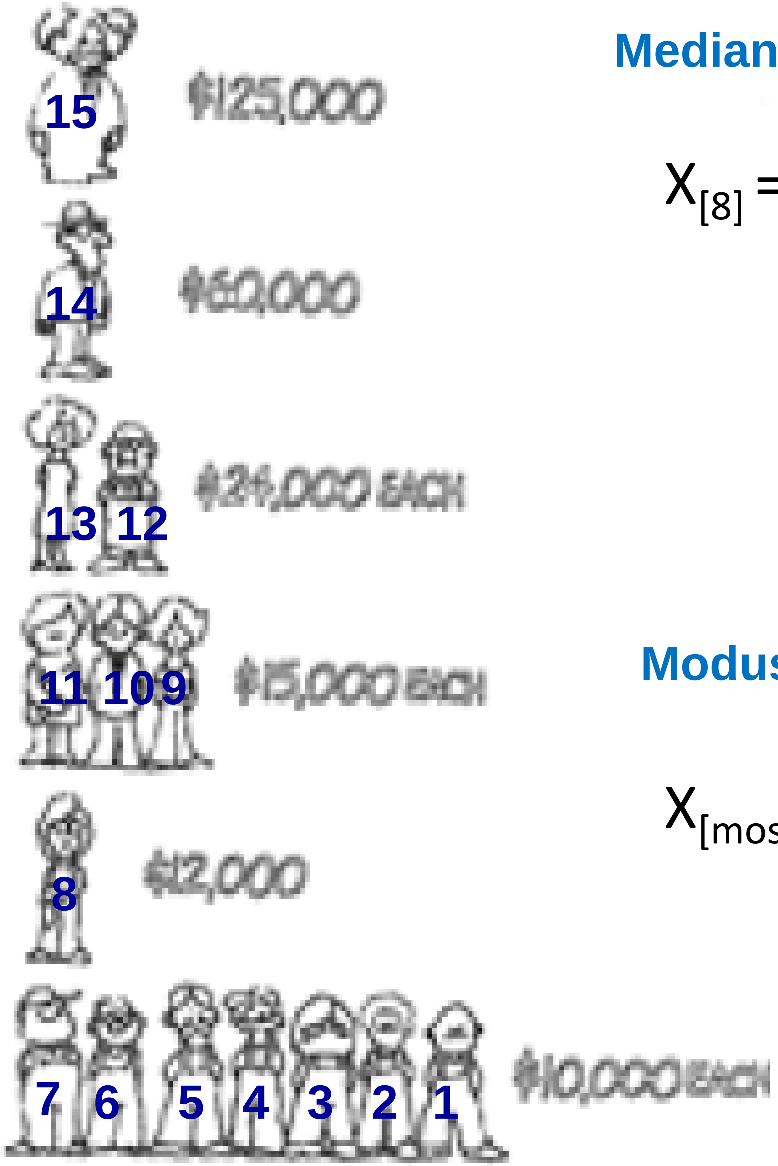
Mean:

Sum:

125,000
+ 60,000
+ 2 x 24,000
+ 3 x 15,000
+ 12,000
+ 7 x 10,000

= 360,000

$$\frac{\$360{,}000}{15} = \boxed{\$24{,}000}$$



Median:

$X_{[8]}$ = 12'000 \$

Modus:

$X_{[\text{most frequent}]}$ = 10'000 \$

The most important measures for the location

- **Mode:** The most frequent value

- **Median:** Value „in the center“ of an ordered sample, i.e. 50% of all values in the sample are ≤ the median-value.

$$median = \begin{cases} x_{[(n+1)/2]} \quad , & falls \;\; n \;\; ungerade \\ \tfrac{1}{2}\,(x_{[n/2]} + x_{[(n+2)/2]}) \;, & falls \;\; n \;\; gerade \end{cases}$$

- **Mean:**

$$\bar{x} = \frac{1}{n} \cdot (x_1 + x_2 + \ldots + x_n) = \frac{1}{n} \sum_{i=1}^{n} x_i$$

Quartiles und Quantiles or Percentiles

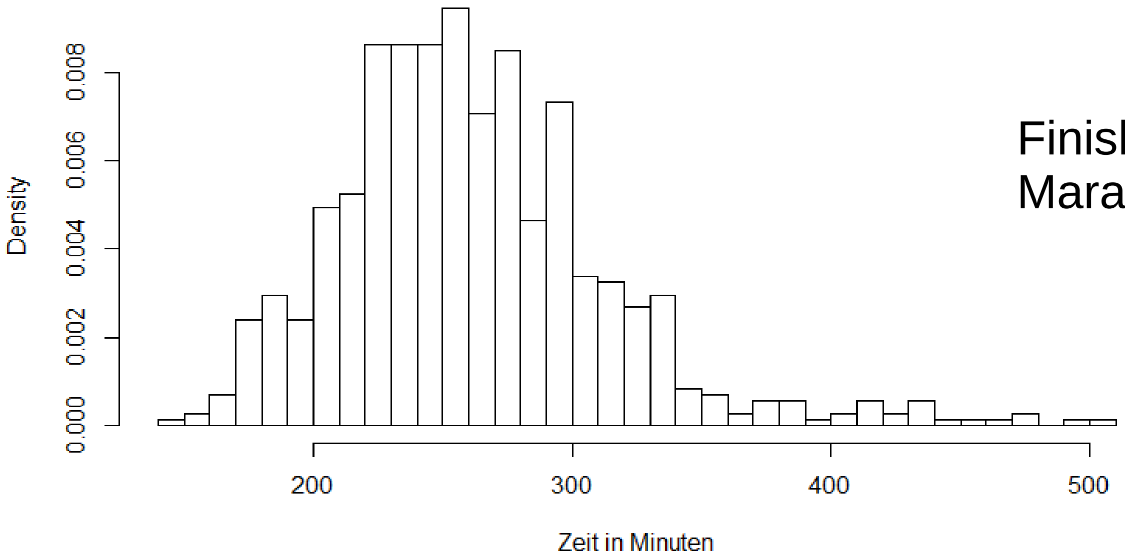
The first Quartile Q1 or $^{25\%}q$ splits the ordered data in a ratio 25:75.

Q2 or $^{50\%}q$ is the median of the data – it splits the ordered data in a ratio 50:50

The third Quartile Q3 or $^{55\%}q$ splits the ordered data in a ratio 75:25.

In analogy an α%-Percentile or Quantile $^{\alpha\%}q$ splits the ordered data in a ratio α :(1- α) – meaning $^{\alpha\%}q$ is the value in a sample for which α% of all values are smaller than this value.

Motivation Quantiles



Finisher times in New York Marathon 2002

Frage:

- Which time is needed to be in the center?

  Median = 256.02 minutes

- Which time do yout need at least to be among the 10% best runners?

  10%-Quantil=201.96

Quartile, Perzentile/Quantile

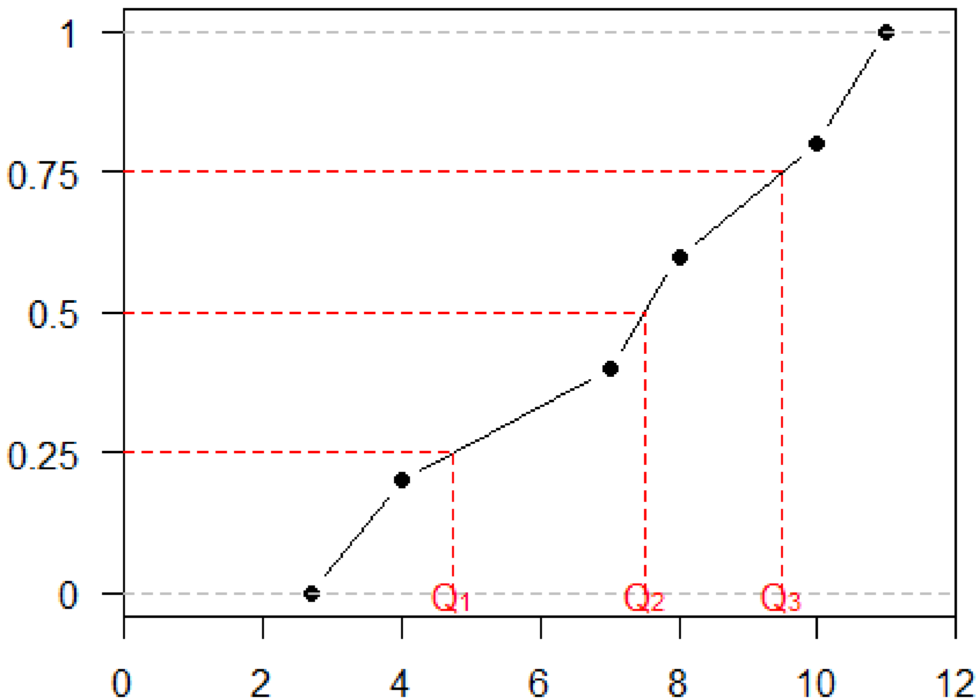
example: 2.7, 4, 7, 8, 10, 11

6 observations: 1. observation divides data 0:100

2. observation divides data 20:80

…

→ lineare Interpolation between two next neighbours



Q_1: ¼ on the way between 2 and 3 observation:

→ $Q_1 = 0.75 * x_{[2]} + 0.25\, x_{[3]} = 0.75 * 4 + 0.25 * 7 = 4.75$

Q_2: in the center between observation 3 and 4

→ $Q_{0.5} = 0.5 * x_{[3]} + 0.5\, x_{[4]} = 0.5 * 7 + 0.5 * 8 = 7.5$

Q_3: ¾ on the way between observation 4 and 5

→ $Q_{0.75} = ¼ * x_{[4]} + ¾ * x_{[5]} = ¼ * 8 + ¾ * 10 = 9.5$

Quartile, Perzentile/Quantile

Formla*: $Q_\alpha = x_{[\lfloor h \rfloor]} + (h - \lfloor h \rfloor)(x_{[\lfloor h \rfloor + 1]} - x_{[\lfloor h \rfloor]})$

$$\text{mit } h = (N-1) \cdot \alpha + 1$$

ceiling: $\lceil x \rceil$ (Beispiel $\lceil 7.2 \rceil = 8$)

floor: $\lfloor x \rfloor$ (Beispiel: $\lfloor 5.7 \rfloor = 5$)

example: 2.7, 4, 7, 8, 10, 11

What is $Q_{0.75}$?

$\alpha = 0.75$

$h=(6-1)\bullet 0.75+1=4.75$

$$Q_{0.75} = x_{[4]} + (4.75 - 4) \cdot (x_{[4+1]} - x_{[4]}) = 8 + 0.75 \cdot (10 - 8) = 9.5$$

* We do that with R!

Quantile in R

```
vals = c(2.7, 4, 7, 8, 10, 11)
quantile(vals)
0%   25%   50%   75%  100%
2.70  4.75  7.50  9.50 11.00
```

```
quantile(vals, type=…)
```

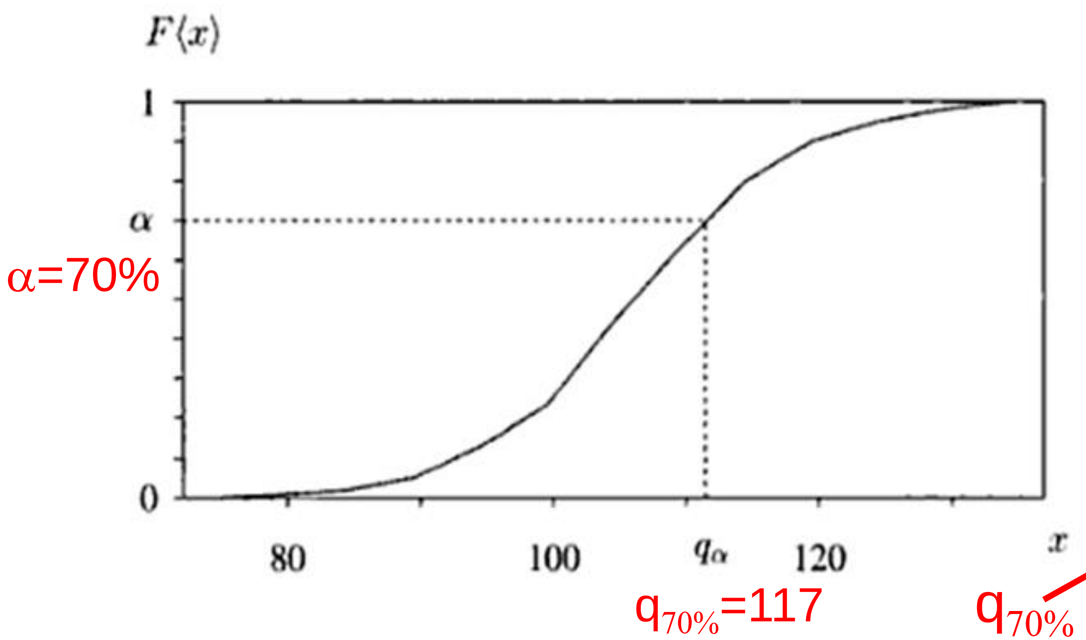
In R there are 9 options for interpolation - jfyi

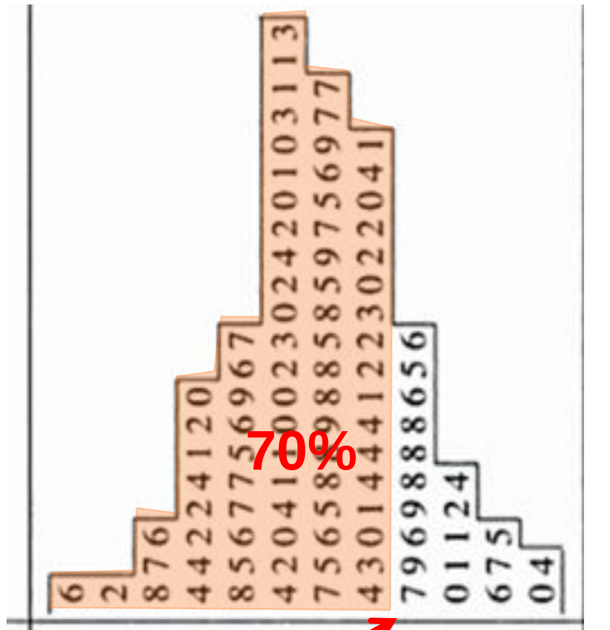
Quantile, Verteilungsfunktion und Histogramm

The area in the scaled historgram left from q_α is $=\alpha$.

The cumulative distribution function F is the integral of the density f and the inverse function of the quantile function → :

$$F(q_\alpha) = \alpha$$

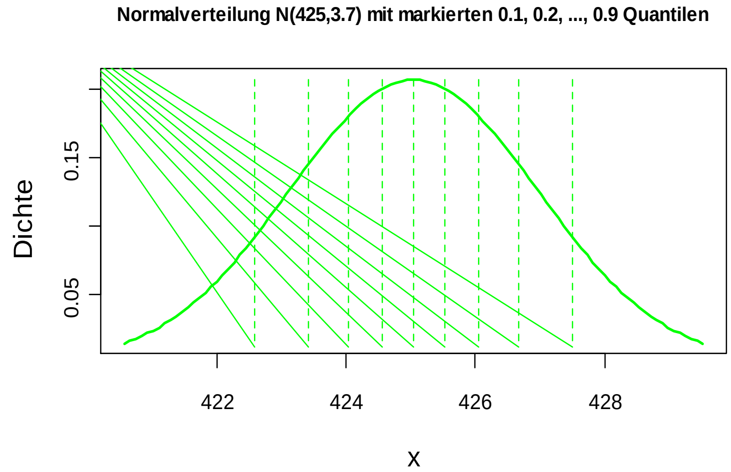


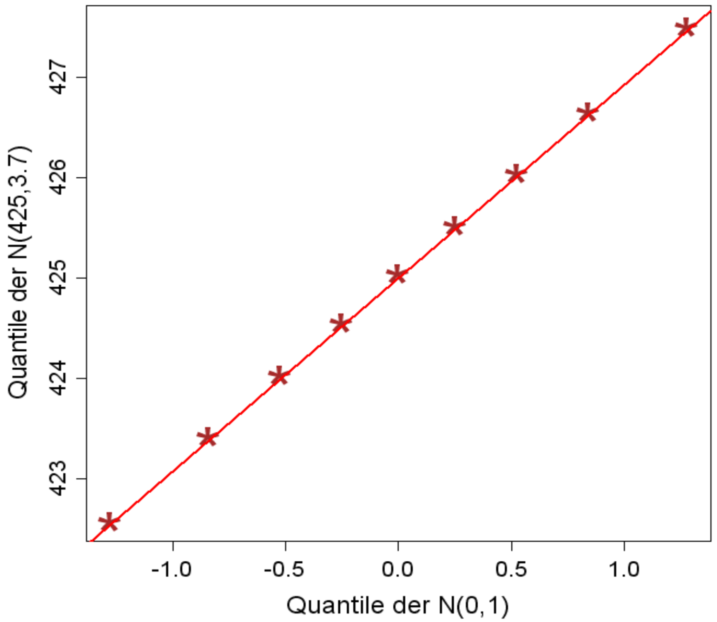


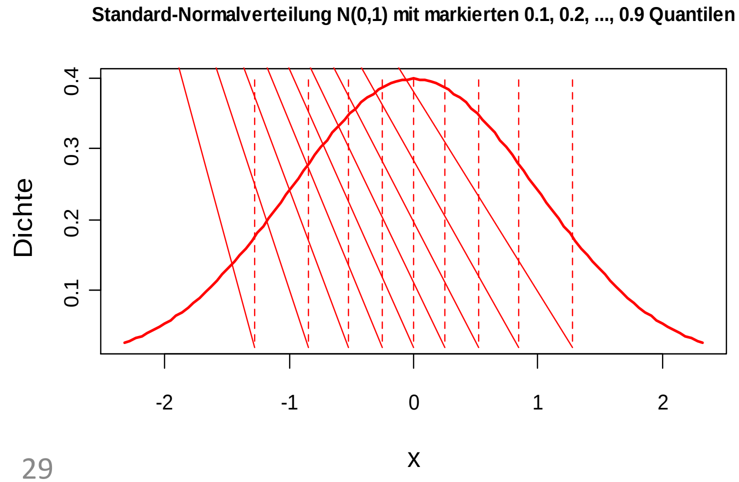
<u>Bemerkung</u>: das Histogramm der abs. Häufigkeiten und das Histogramm der Dichte (rel. Häufigkeiten), haben bei gleicher Klassenbreite die gleiche Form => Flächenverlauf hat auch gleiche Form.

Quantil-Quantil Plots

---: Position of the 10%-, 20%, …, 90%-Quantils





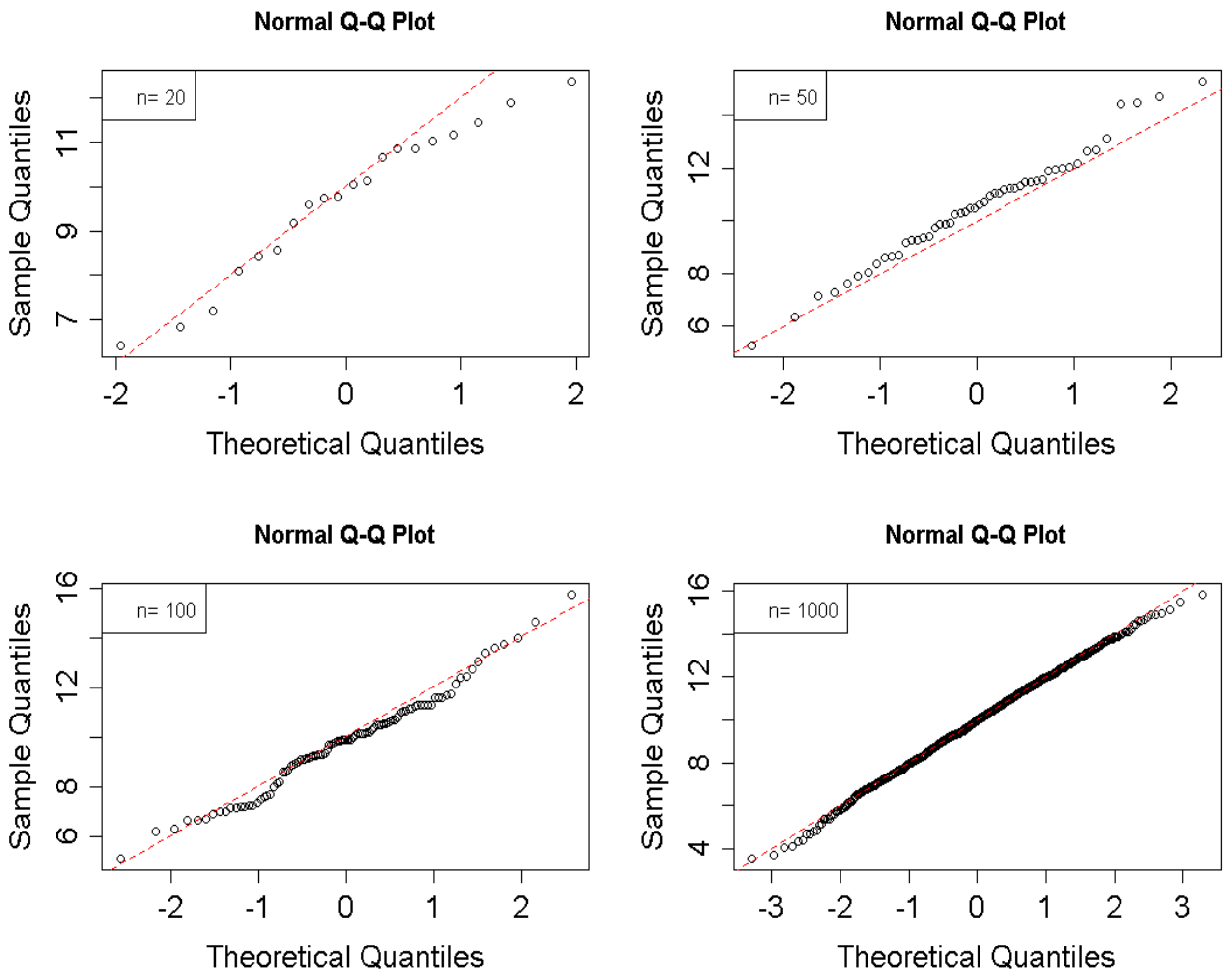


Normal-Distributions have all the same bell-shape – regardless of the parameters.

We can use a Normal QQ-Plot, where the empirical quantilse are plotet versus the quantilse of a N(0,1)-distribution, if we want to check if our sample is normal distributed.

29

Draw data from Normal-Distribution and generate the Normal-Quantil-Quantil-Plots



Because of sampling variation we do not get all points on a straight line, however the bigger the sample is the less important is sample variation.

How to determine the α-Quantil of a sample

First order your sample $x_1, x_2, \ldots, x_n$, to get a ordered sample $x_{[1]}, x_{[2]}, \ldots, x_{[n]}$. In a ordered sample $x_{[1]}$ is the smallest value in the sample and $x_{[n]}$ is the biggest value in the sample.

$$
{}^{\alpha}q_x = \begin{cases} x_{\left[\lceil \alpha\cdot n \rceil\right]} \ , \ if\ \alpha\cdot n \notin \mathbb{Z} \ , \ \lceil \alpha\cdot n \rceil : ceil\ to\ next\ bigger\ \mathrm{int}eger \\ \frac{1}{2}\cdot\left(x_{[\alpha\cdot n]} + x_{[\alpha\cdot n+1]}\right) \ , \ falls\ \alpha\cdot n \in \mathbb{Z} \end{cases}
$$

Example:

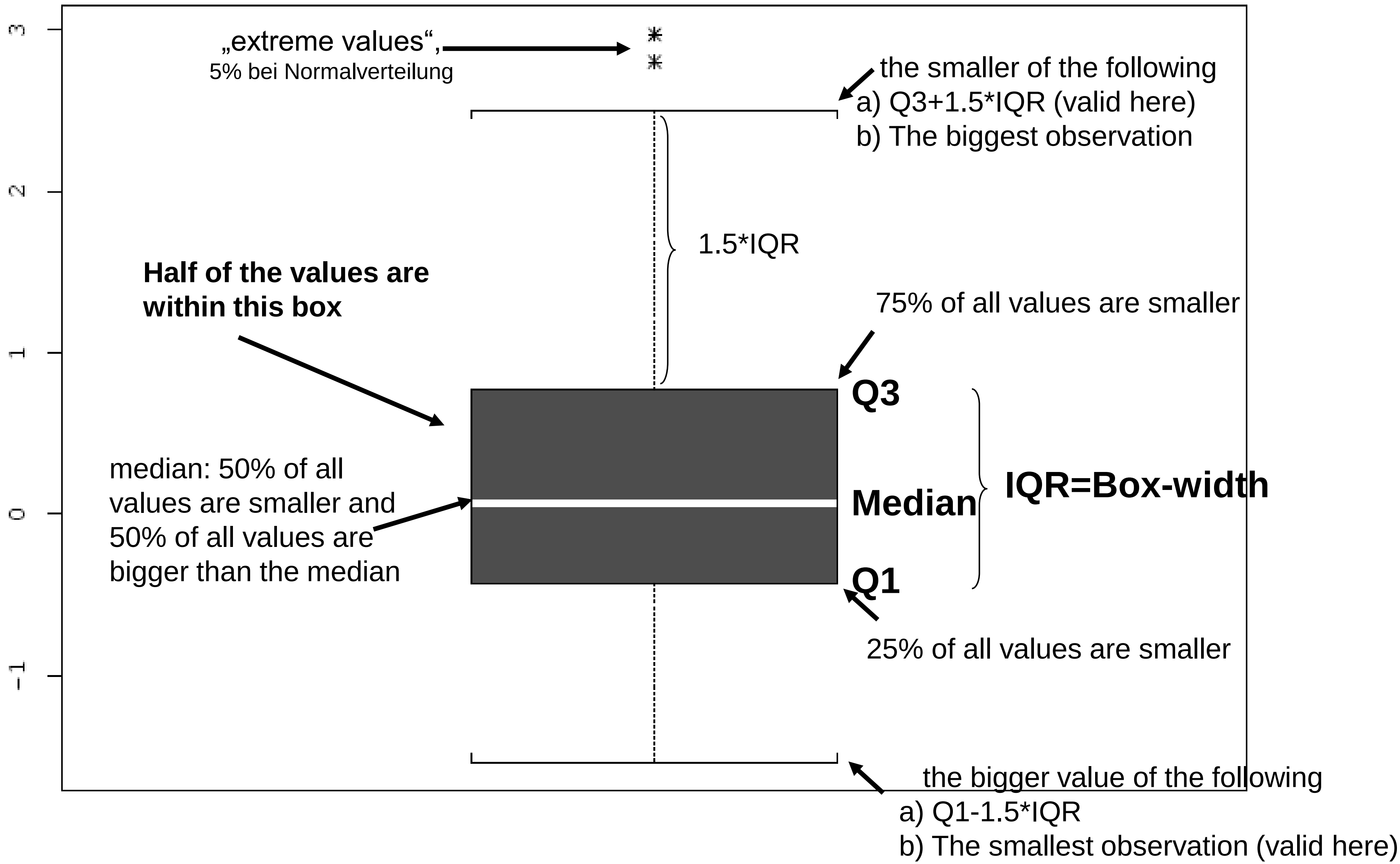
Sample: 3, 4, 7, 5, 0.5, 6 : n=6 ; ordered sample: 0.5, 3, 4, 5, 6, 7

To determine the Median=${}^{0.5}q$ we determine the ordinal numbers

[α*n]=[0.5*6] =[3] -> Median are the average of the third- and forth-smallest value:

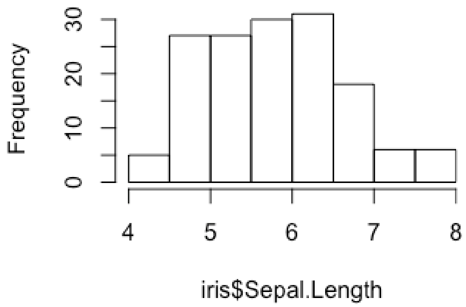
Median=0.5*($x_{[3]}$+$x_{[4]}$)=0.5*(4+5)=4.5 .

Definition of the Boxplot to visualize continuous data



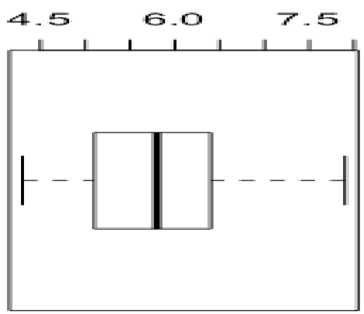
Univariate Plotting: Categorical Data

- Histogram

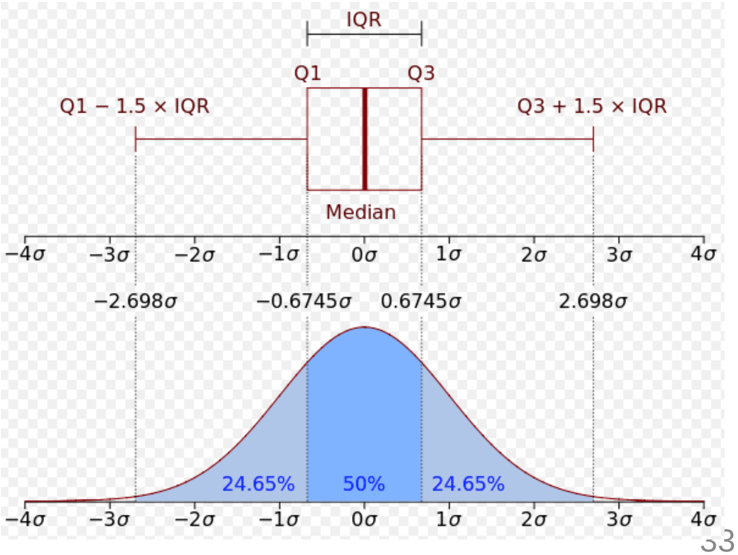


hist(iris$Sepal.Length)

- Box Plot

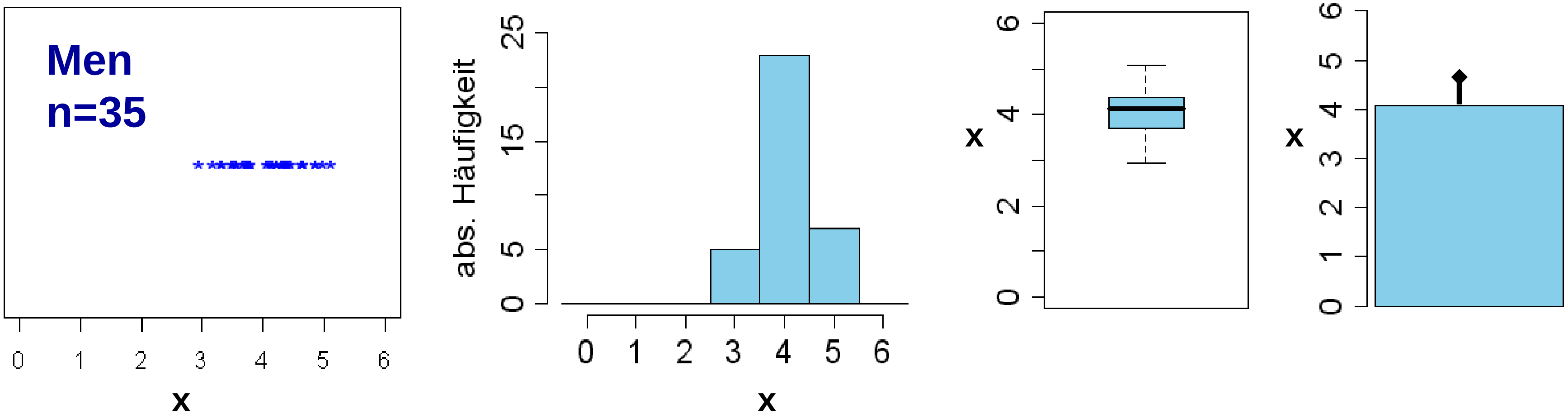


boxplot(iris$Sepal.Length)



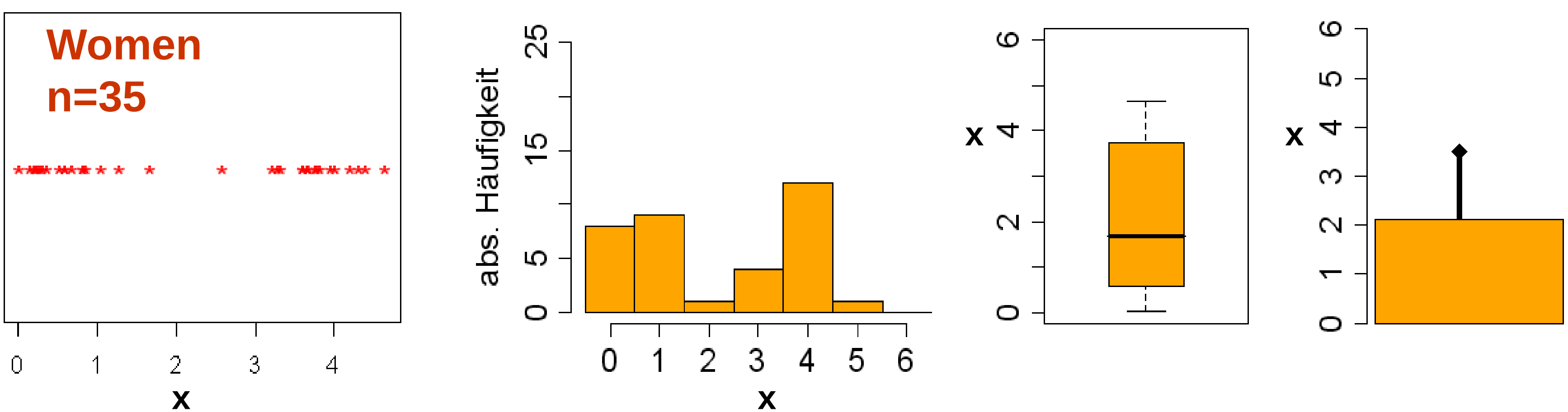
How to best visualize continuous data

X: reduction of the BMI after take-in of a new drug



Stripchart **Histogramm** **Boxplot** **Mean-plot**

(not recommended)



How to best visualize continuous data

Stripchart:
for n<20 it is a good plot, since it shows each data point and gives an impression about the distribution.

Histogram:
Good for n>20, it reveals the shape of the distributions
(classes should be well set.)

Boxplot:
Very good for n>10 and uni-modale distributions – especially good for comparing distributions across different groups.

Mean-Plots:
Does carry only very little information. Only o.k. if distribution is symmetric, uni-modal and outlier-free – in all other cases this representation is missleading. .

Bi-variate Visualization

How to visualize bivariate data, with 2 features per observation?

quantitativ vs. quantitativ:

- Scatterplots, Quantil-Quantil-Plot

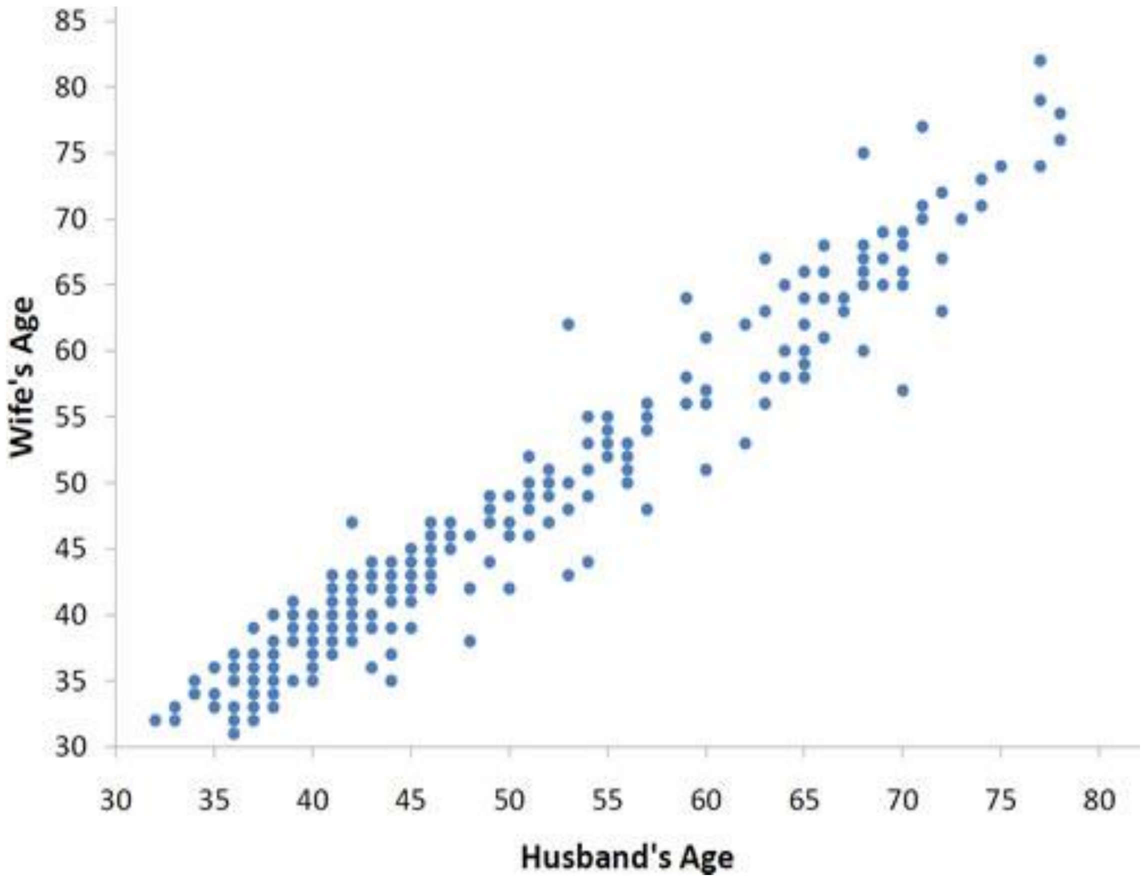
categorical vs. categorical

- mosaicplot

quantitativ vs. categorical:

- Grouped boxplots or grouped stripcharts (mean-plots)

Scatterplot for two quantitative variables



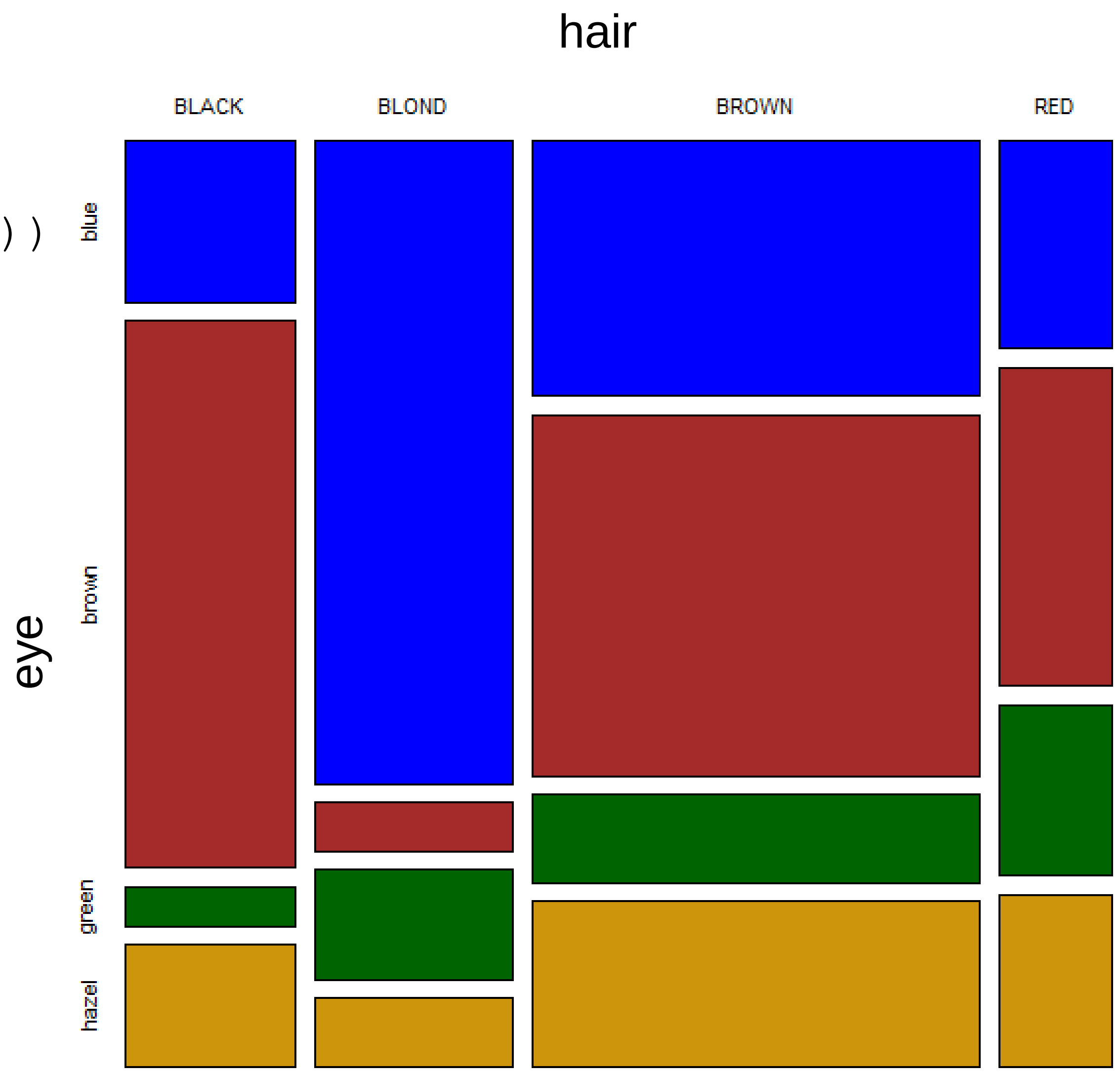

```
plot(dat$m.age, dat$w.age, xlab="Husband's Age", ylab= "Wifes's Age" )
plot( w.age ~ m.age, data=dat)
```

Bi-variate visualization of 2 categorical variables

```
ct=table(he$hair,he$eye))
```

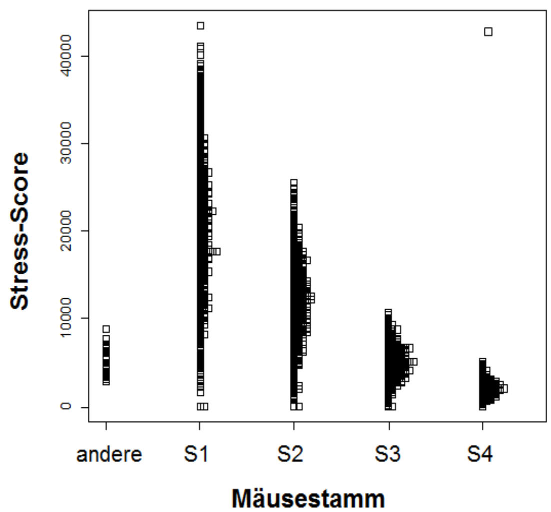
```
> ct

       blue brown green hazel
  BLACK   20    68     5    15
  BLOND   94     7    16    10
  BROWN   84   119    29    54
  RED     17    26    14    14
```

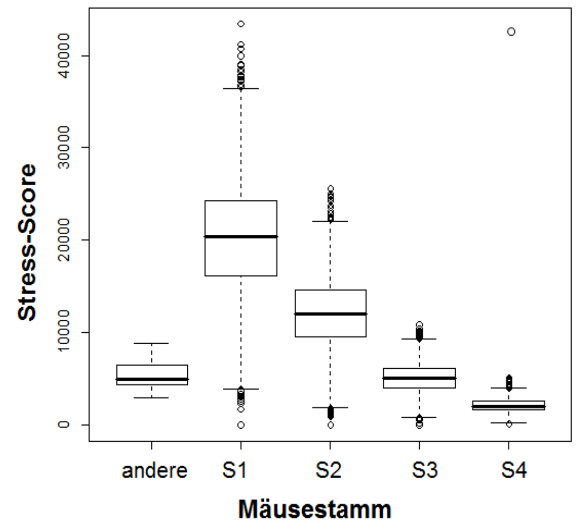
```
plot(ct,color=T)
```



Are hair-color and eye-color independent?

Bi-variate visualization: continuous vs categorical variable





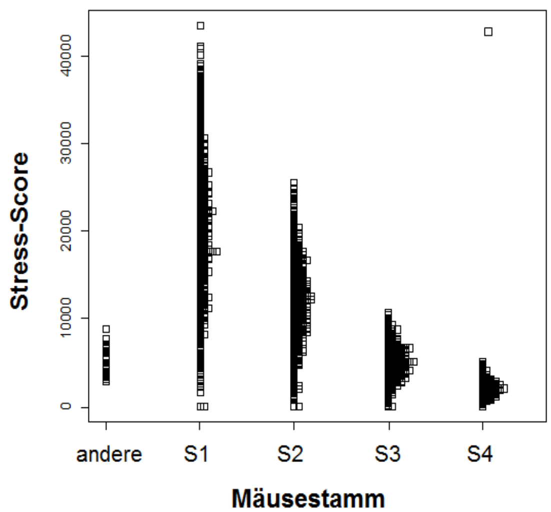

```
stripchart( stress ~ stamm,
         method="jtter",data=mouse)
```

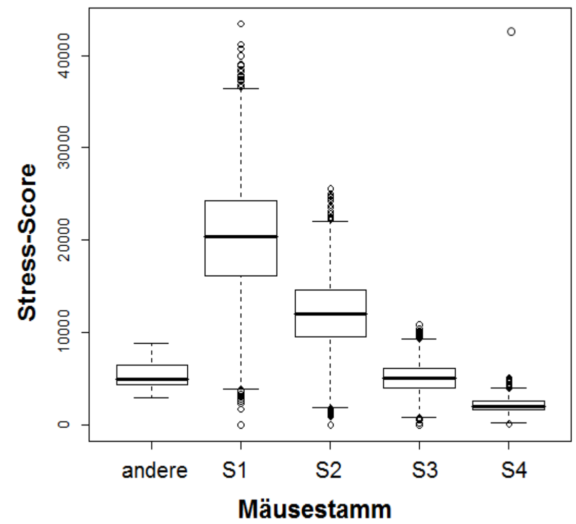
```
boxplot( stress ~ stamm, data=mouse)
```

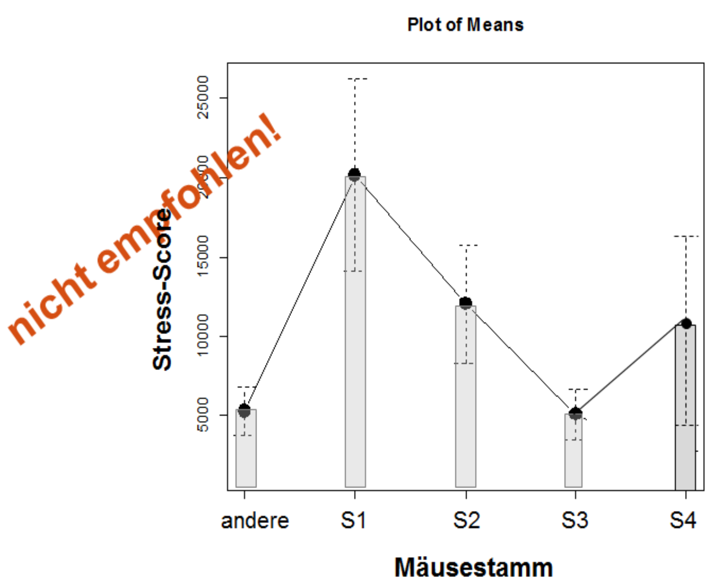
→ Compare distribution of continuous variable in different groups (categories).

Hübsche Demo: https://stekhoven.shinyapps.io/barplotNonsense

Bi-variate visualization of 1 categorical and 1 continuous variable







=> *Boxplots*	*very good, if uni-modal*
=> *Stripcharts*	*good, if «stacked»*
=> *Histogram*	*good for revealing the shape*
	but takes a lot of space
=> *Mean plots*	*not good, not robust,*
	Can be miss-leading!

40