

Biostatistics Week 2

Topics this week:

➤ uni-variate descriptive Analysis

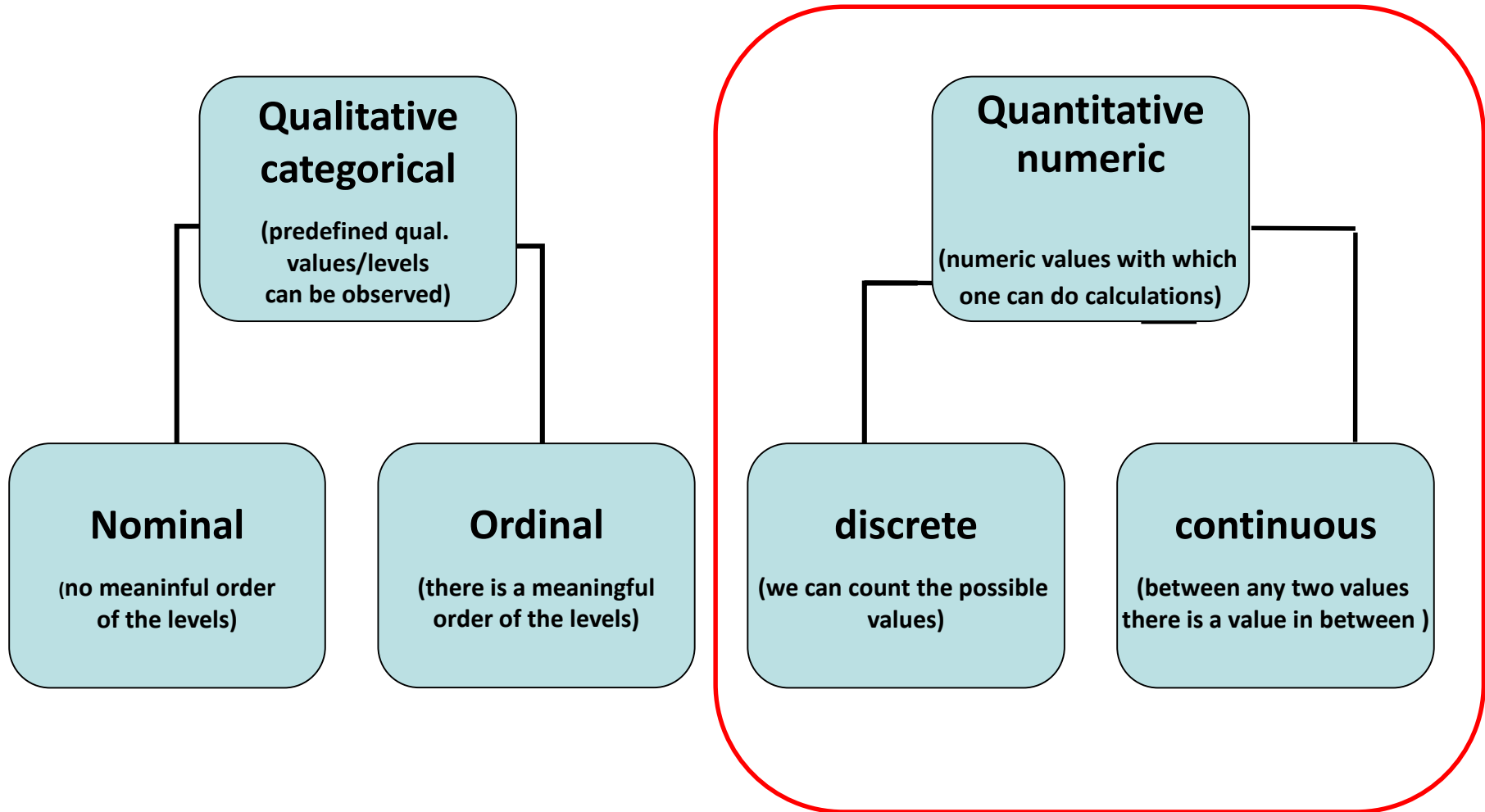
- Recap: Data types
- Measure for location: mean, median, mode, quantiles
- Visualizing categorical variables: pie chart and barplots
- Visualizing continuous variables: histogram and boxplot

➤ Bi-variate descriptive Analysis

- Continuous vs continuous: scatterplot
- Categorical vs categorical: mosaicplot
- Continuous vs categorical: grouped boxplots or stripcharts

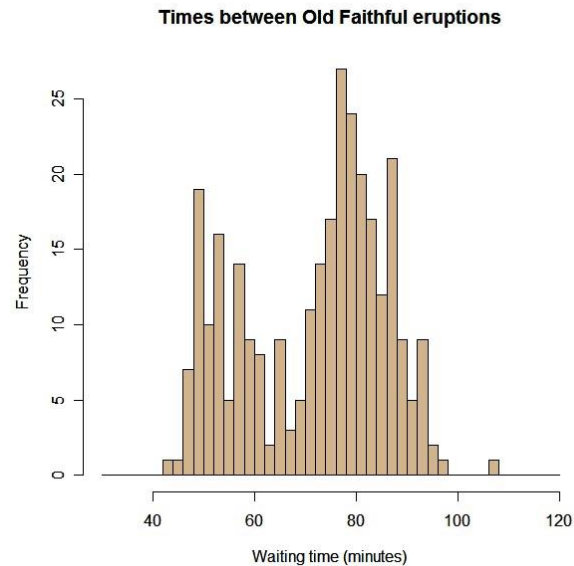
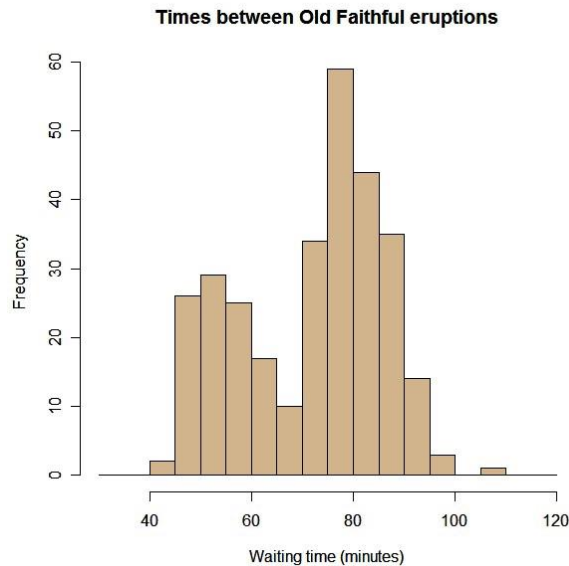
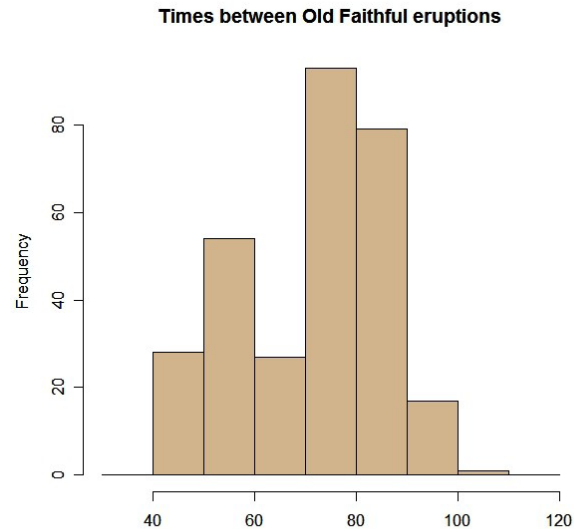
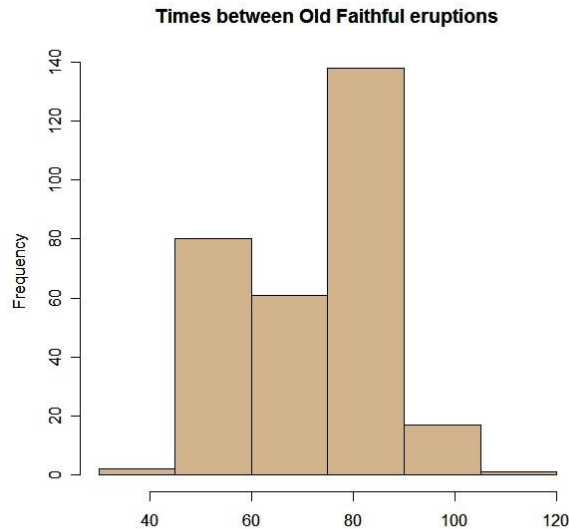
<https://bsick.github.io/Biostatistics-Fall-2018/>

There are different types of data



Visualization of quantitative continuous variables

How many classes do we need?



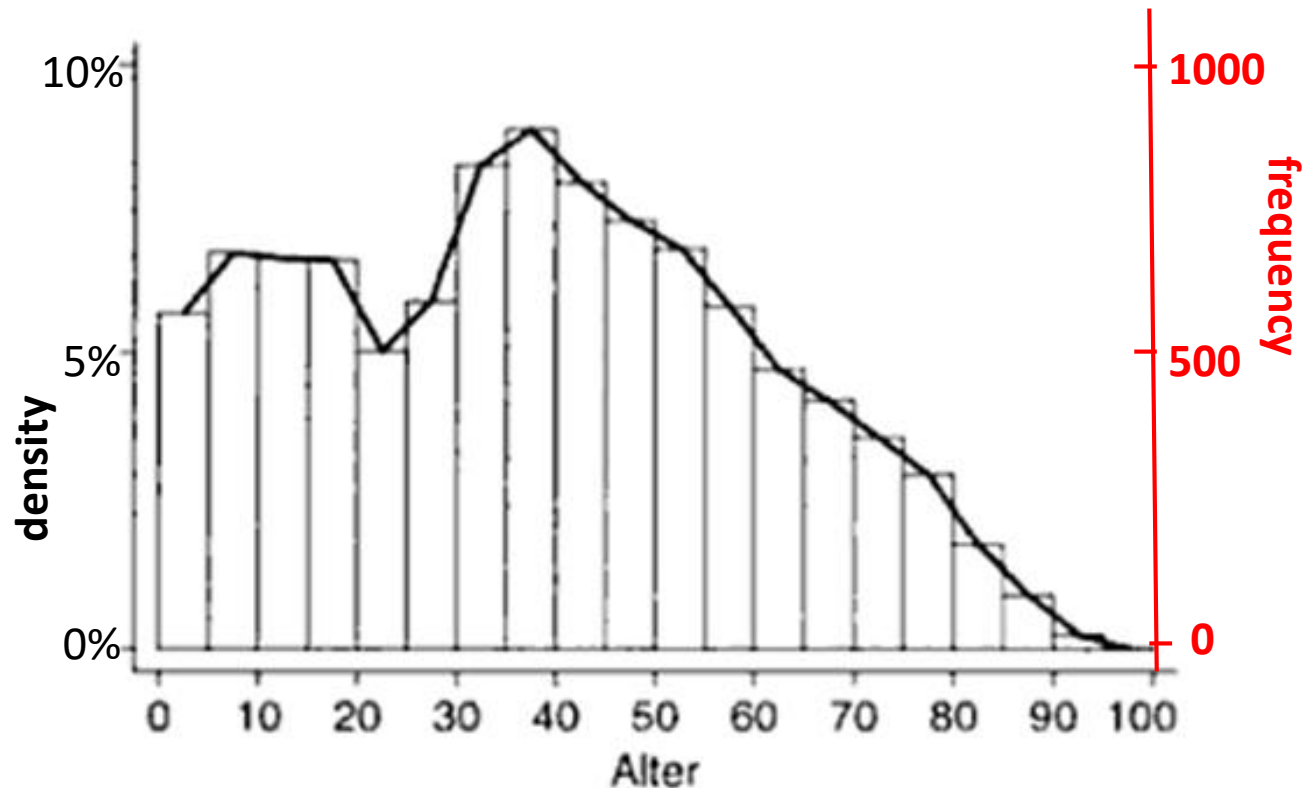
<http://www.amstat.org/publications/jse/v6n3/applets/Histogram.html>



299 eruption intervals were observed

Shape of the histogram may depend on the class choices

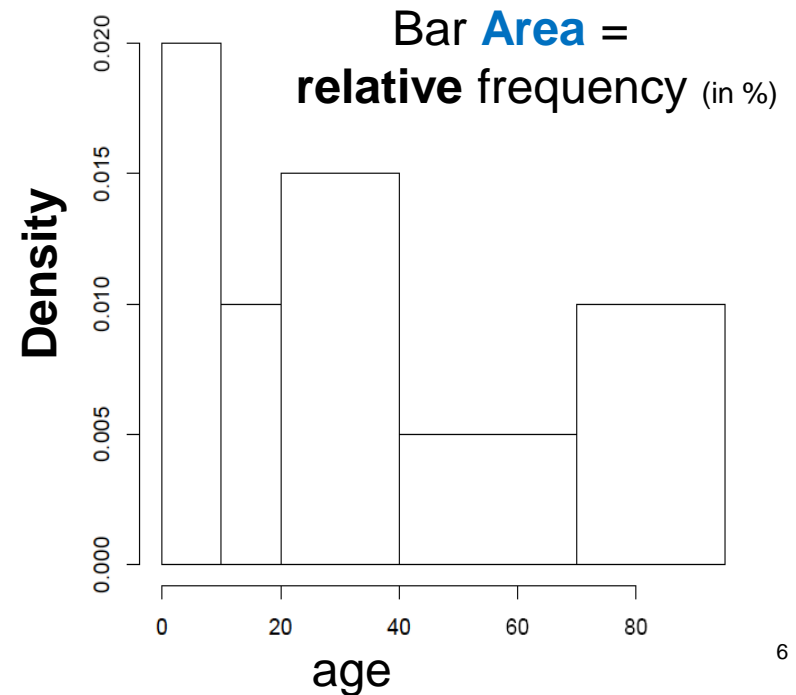
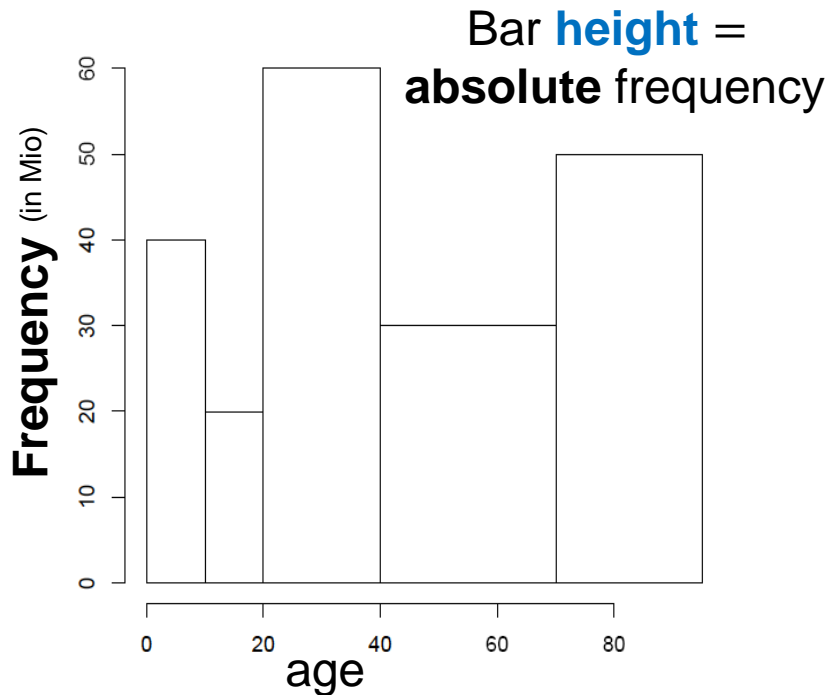
How does the distribution of age look like



Only in case of equally sized bins the **scaled** and **unscaled** histogram look the same and it is possible to label the y-axis with percentages – usually it only 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%

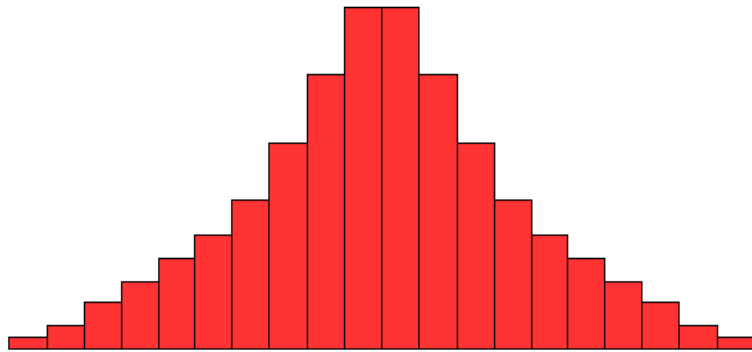


Rules for histograms

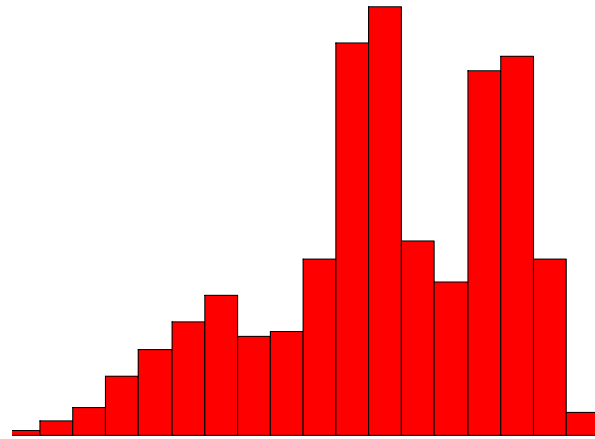
- Avoid classes with different width! (shape will change)
- How many classes: \sqrt{n} classes for n observations.
- The shape can depend on the number classes and the class limits.

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.

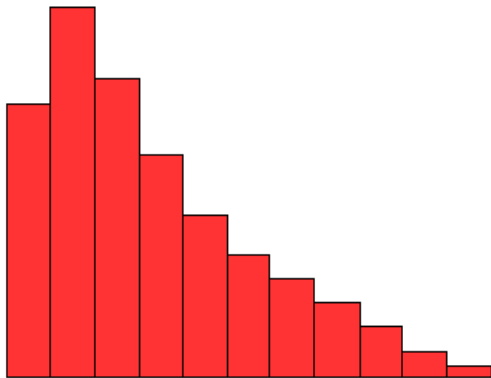
Shapes of distributions



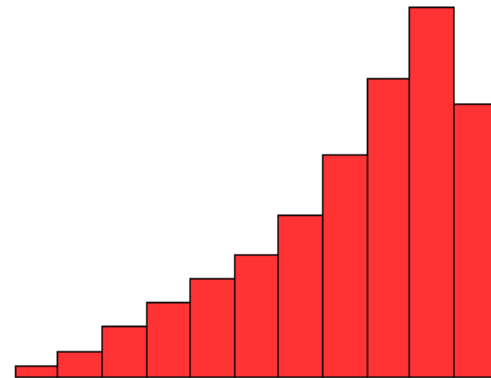
Symmetric, uni-modale



Multi-modale, slightly left-skewed



Right-skewed, uni modale



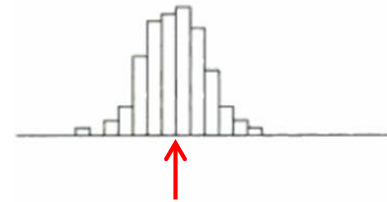
Left-skewed, uni-modale

Measures for the location and variation

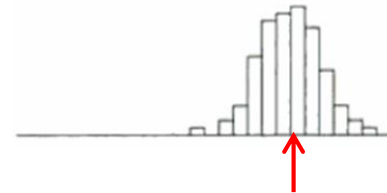
Data can be summarized by summary statistics. Most important key figure describe the center and the width of a distribution..

Measures for the location

Where is the center?

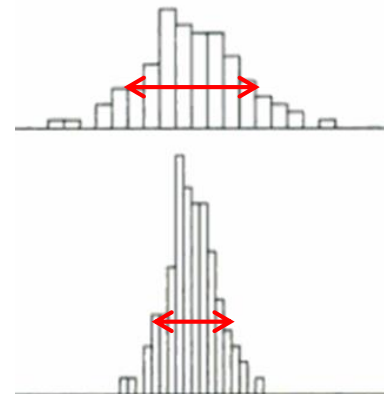


What is a typical value?



Measures for the variation

A number which quantifies the width of the distribution.



Is the mean salary a «typical salary»?

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



Schweizer Arbeitsplätze



[Grafik vergrößern](#)

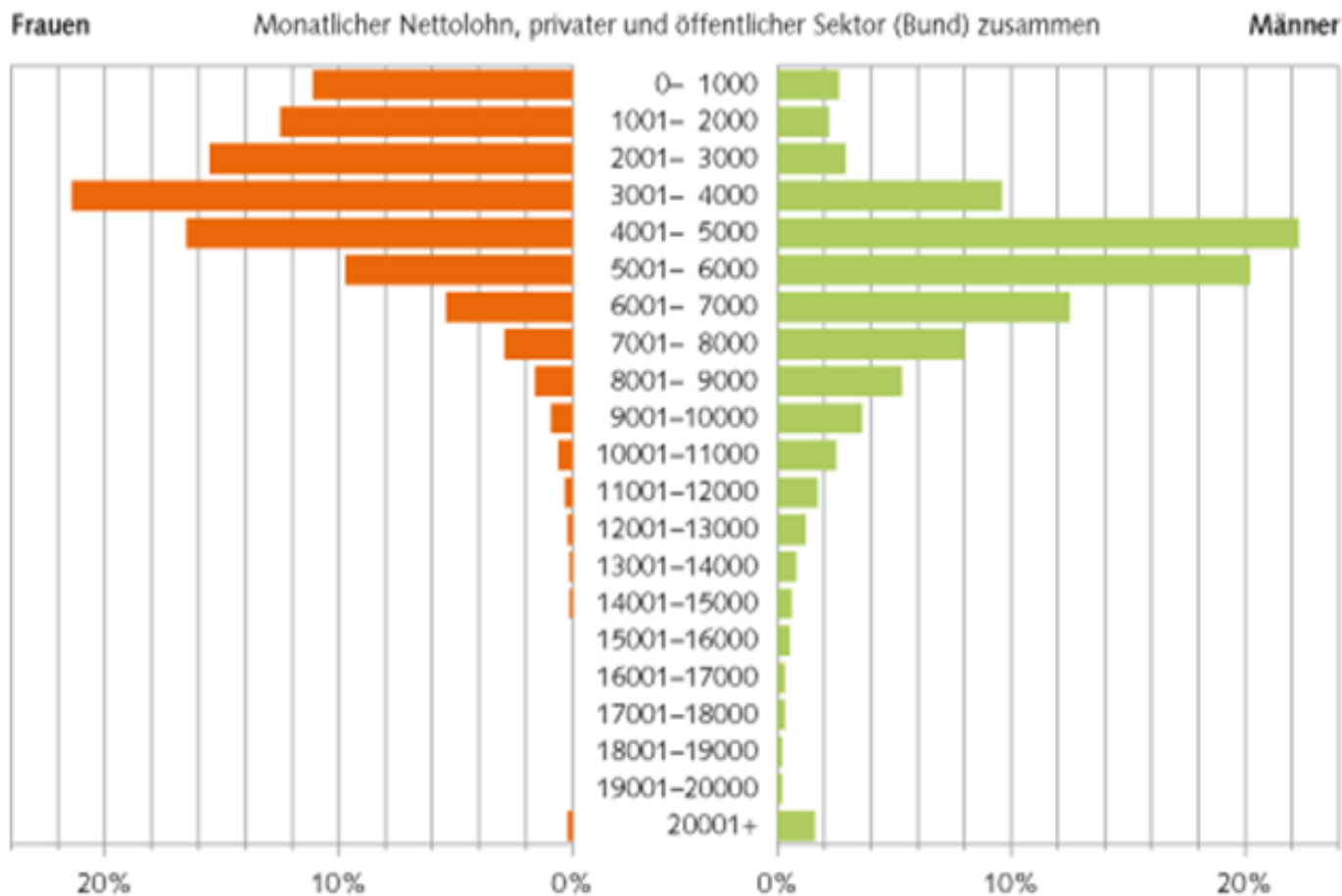
Novartis beschäftigt weltweit zurzeit rund 99.800 Mitarbeitende. Davon arbeiten rund 12.000 in der Schweiz – verteilt auf die acht Standorte in Basel BS/BL, Stein AG, Embrach ZH, Cham ZG, Bern BE, St-Aubin FR, Nyon VD und Locarno TI. Eine kürzlich veröffentlichte Studie hat ergeben, dass für jeden direkten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

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

Distribution of salaries in Switzerland

Häufigkeitsverteilung der Arbeitnehmenden nach Lohnhöhenklassen 2008



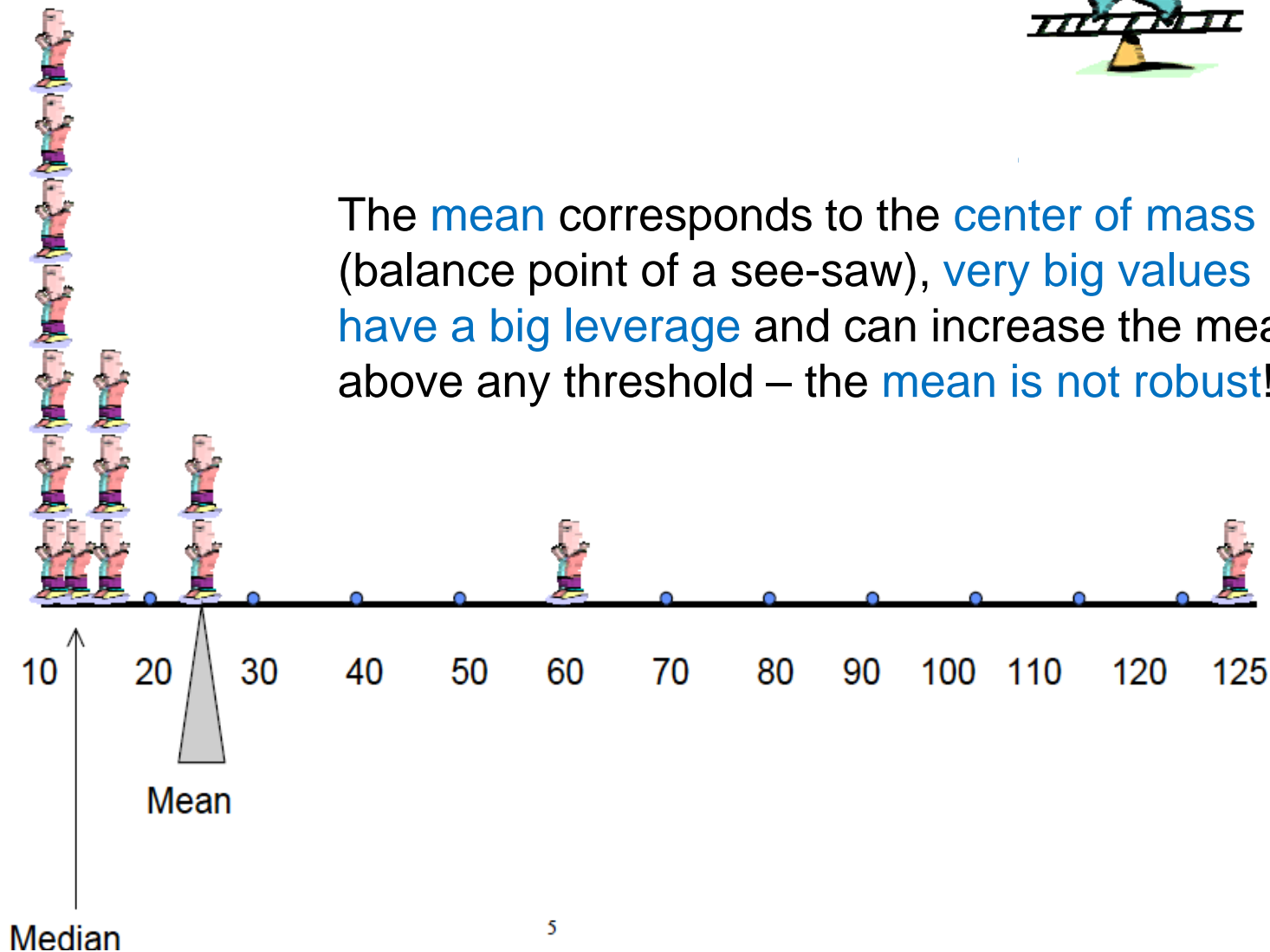
Quelle: Schweizerische Lohnstrukturerhebung

© BFS

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 \Rightarrow median=3
 - In case of an odd-numbered number of observation, take mean of the two center values:
 - 1,2,3,1000 \Rightarrow median=2.5
- Median
 - Create a ordered sample:

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

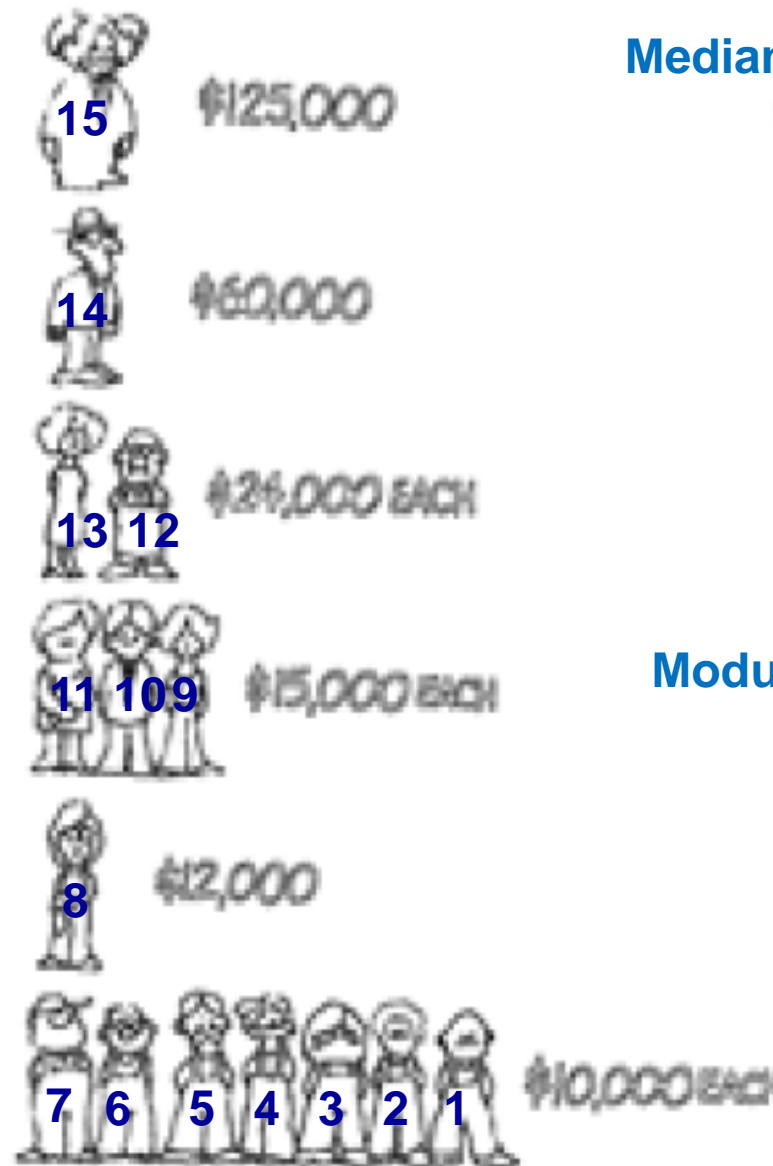
$$\tilde{x} = \begin{cases} x_{[n+1/2]} & , \text{ for odd } n \\ \frac{1}{2} \cdot \left(x_{[n/2]} + x_{[n/2+1]} \right) & , \text{ for even } n \end{cases}$$

What is „the mean“ income in this company?

Mean:

Median:

Modus:



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 \leq the median-value.

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

- **Mean:**

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

Quartiles und Quantiles or Percentiles

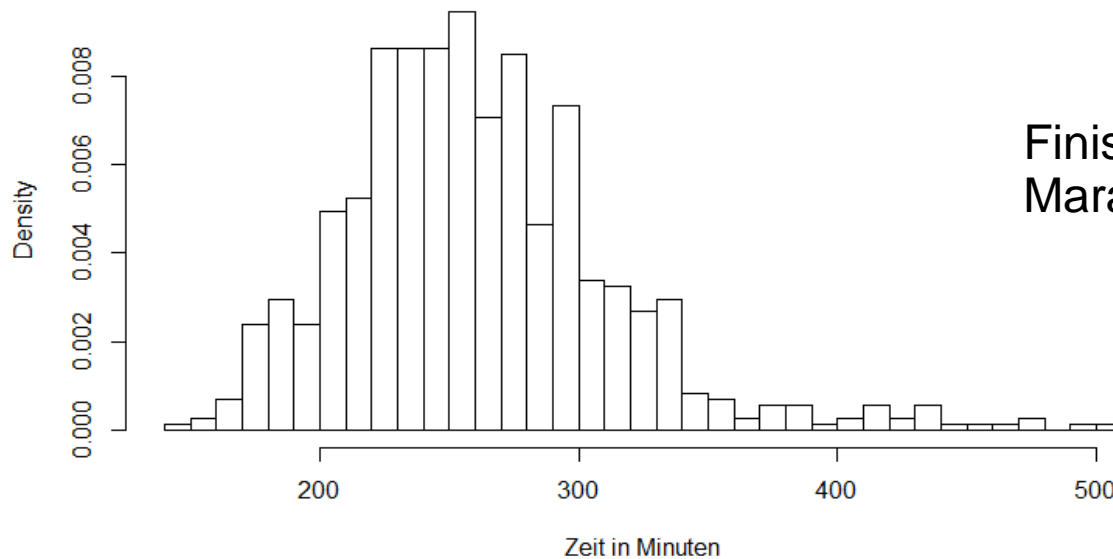
The first Quartile $Q1$ or $^{25\%}q$ splits the ordered data in a ratio 25:75.

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

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

In analogy an $\alpha\%$ -Percentile or Quantile $^{\alpha\%}q$ splits the ordered data in a ratio $\alpha : (1 - \alpha)$ – meaning $^{\alpha\%}q$ is the value in a sample for which $\alpha\%$ 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 you need at least to be among the 10% best runners?

10%-Quantil=201.96

Quartile, Perzentile/Quantile

Formula*: $Q_\alpha = x_{[\lceil h \rceil]} + (h - \lceil h \rceil)(x_{[\lceil h \rceil + 1]} - x_{[\lceil h \rceil]})$
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 $Q_{0.75}$?

* 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(x, probs=...)
```



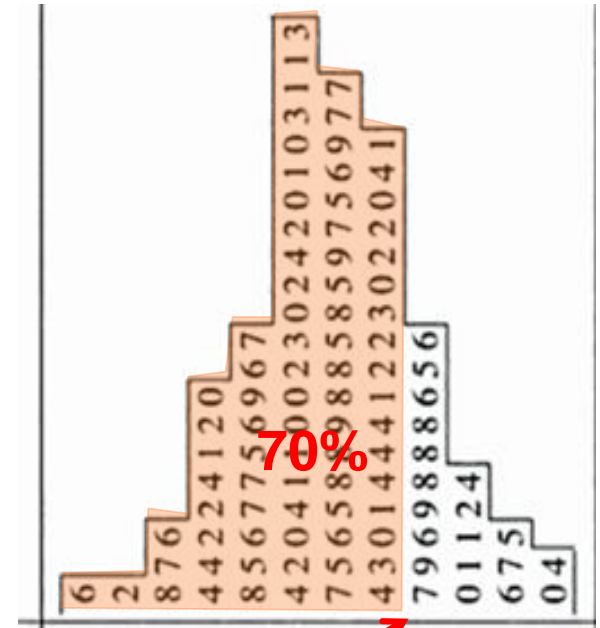
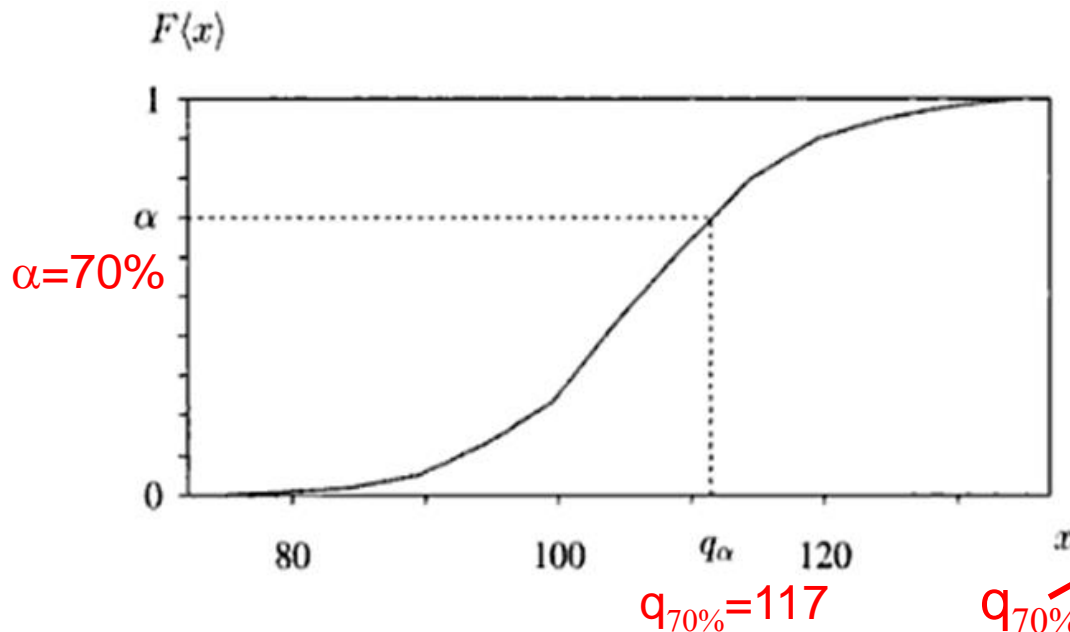
In R there are 9 options for interpolation - jfyi

Quantile, Verteilungsfunktion und Histogramm

The area in the scaled histogram 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 \rightarrow :

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



Bemerkung: das Histogramm der abs. Häufigkeiten und das Histogramm der Dichte (rel. Häufigkeiten), haben bei gleicher Klassenbreite die gleiche Form \Rightarrow Flächenverlauf hat auch gleiche Form.

How to determine the α -Quantil of a sample

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

$${}^{\alpha}q_x = \begin{cases} x_{[\overline{\alpha \cdot n}]} , & \text{if } \alpha \cdot n \notin \mathbb{Z} , \overline{\alpha \cdot n} : \text{ceil to next bigger integer} \\ \frac{1}{2} \cdot (x_{[\alpha \cdot n]} + x_{[\alpha \cdot n + 1]}) , & \text{falls } \alpha \cdot n \in \mathbb{Z} \end{cases}$$

Example:

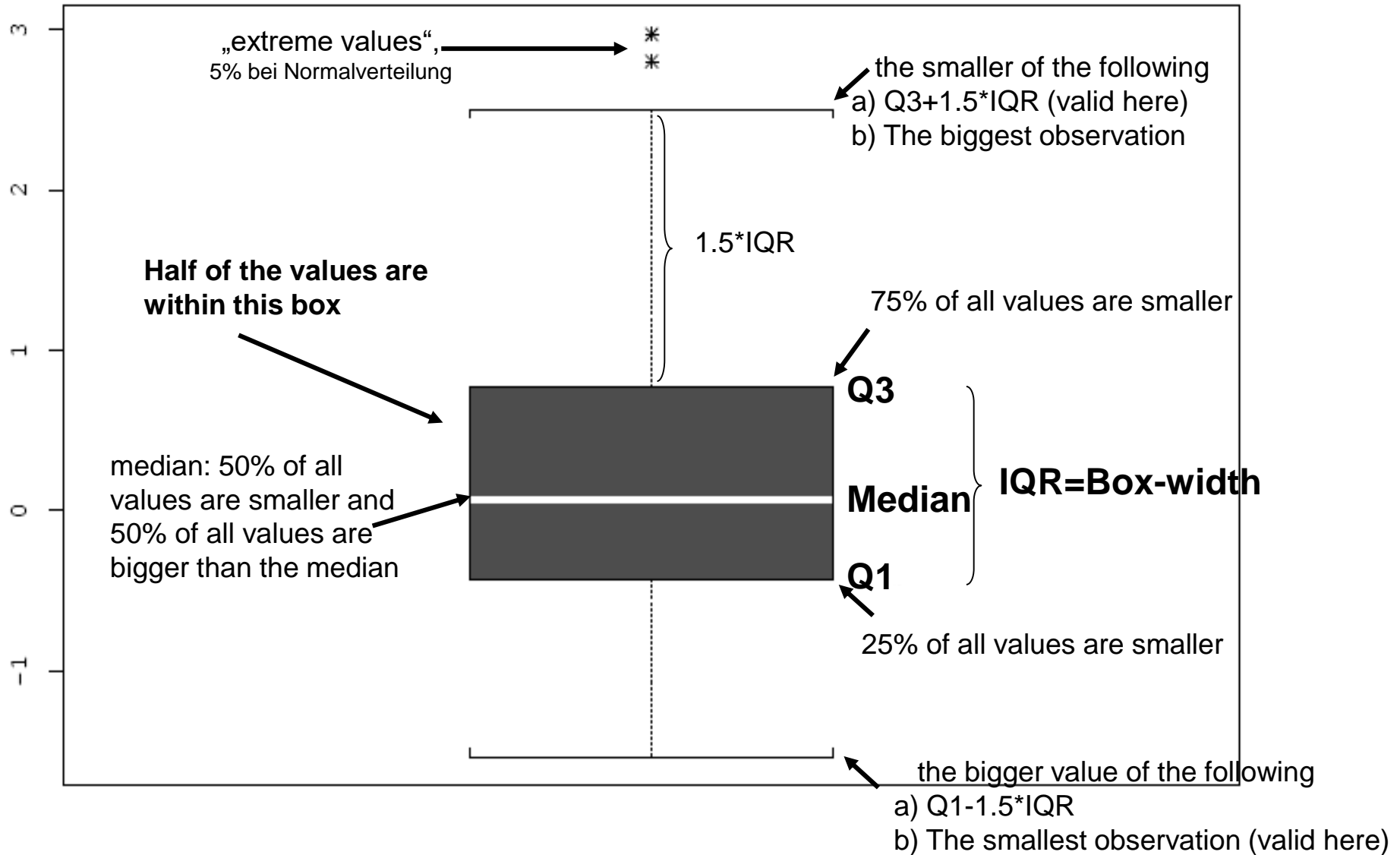
Sample: 3, 4, 7, 5, 0.5, 6 : $n=6$; ordered sample: 0.5, 3, 4, 5, 6, 7

To determine the Median= $0.5q$ we determine the ordinal numbers

$[\alpha \cdot n] = [0.5 \cdot 6] = [3] \rightarrow$ Median are the average of the third- and forth-smallest value:

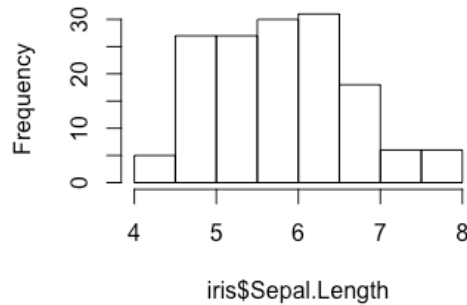
Median= $0.5 \cdot (x_{[3]} + x_{[4]}) = 0.5 \cdot (4 + 5) = 4.5$.

Definition of the Boxplot to visualize continuous data



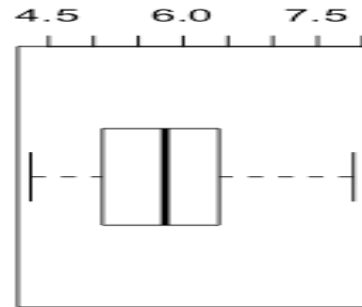
Univariate Plotting: Categorical Data

- Histogram

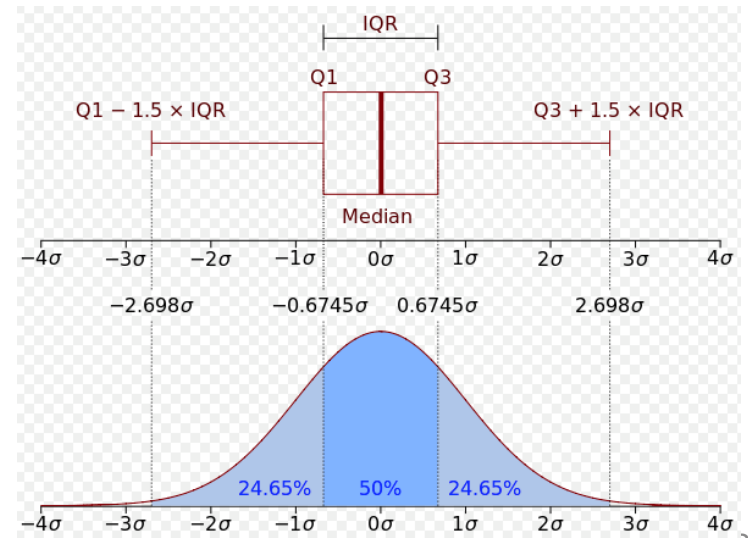


```
hist(iris$Sepal.Length)
```

- Box Plot

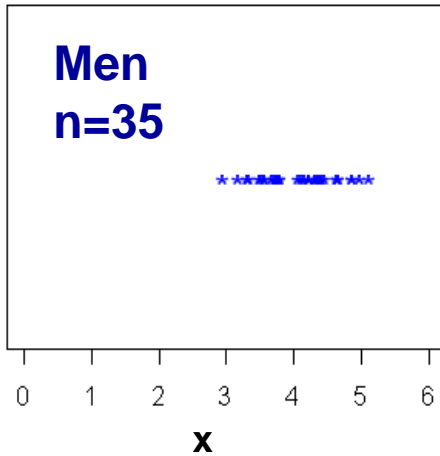


```
boxplot(iris$Sepal.Length)
```

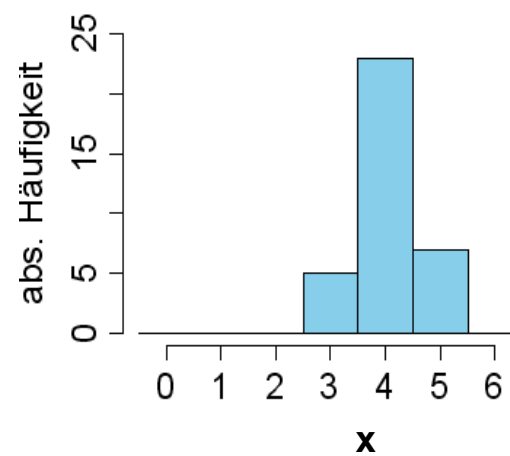


How to best visualize continuous data

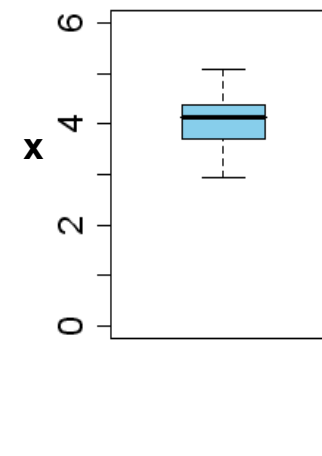
X: reduction of the BMI after take-in of a new drug



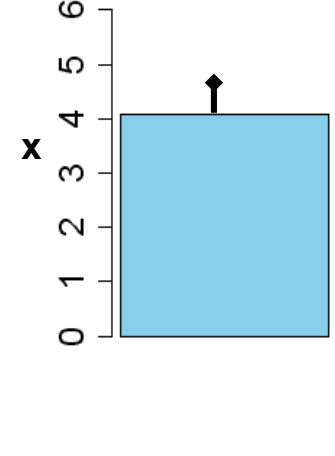
Stripchart



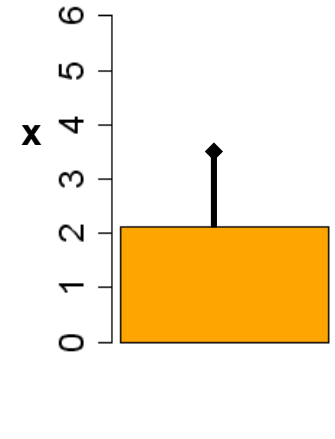
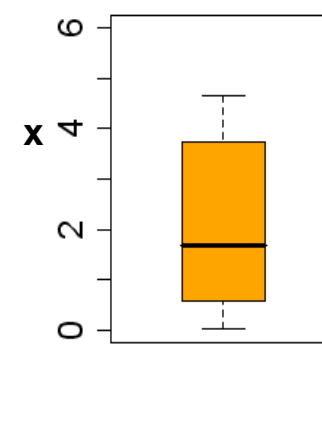
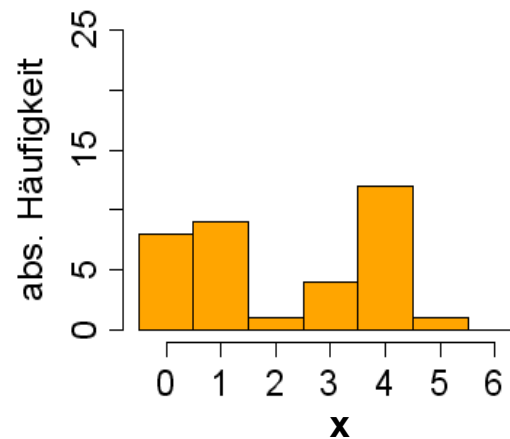
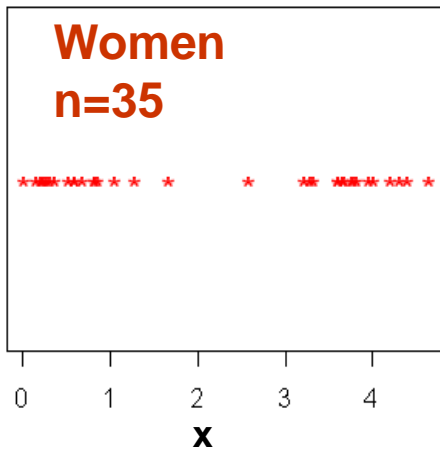
Histogramm



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:

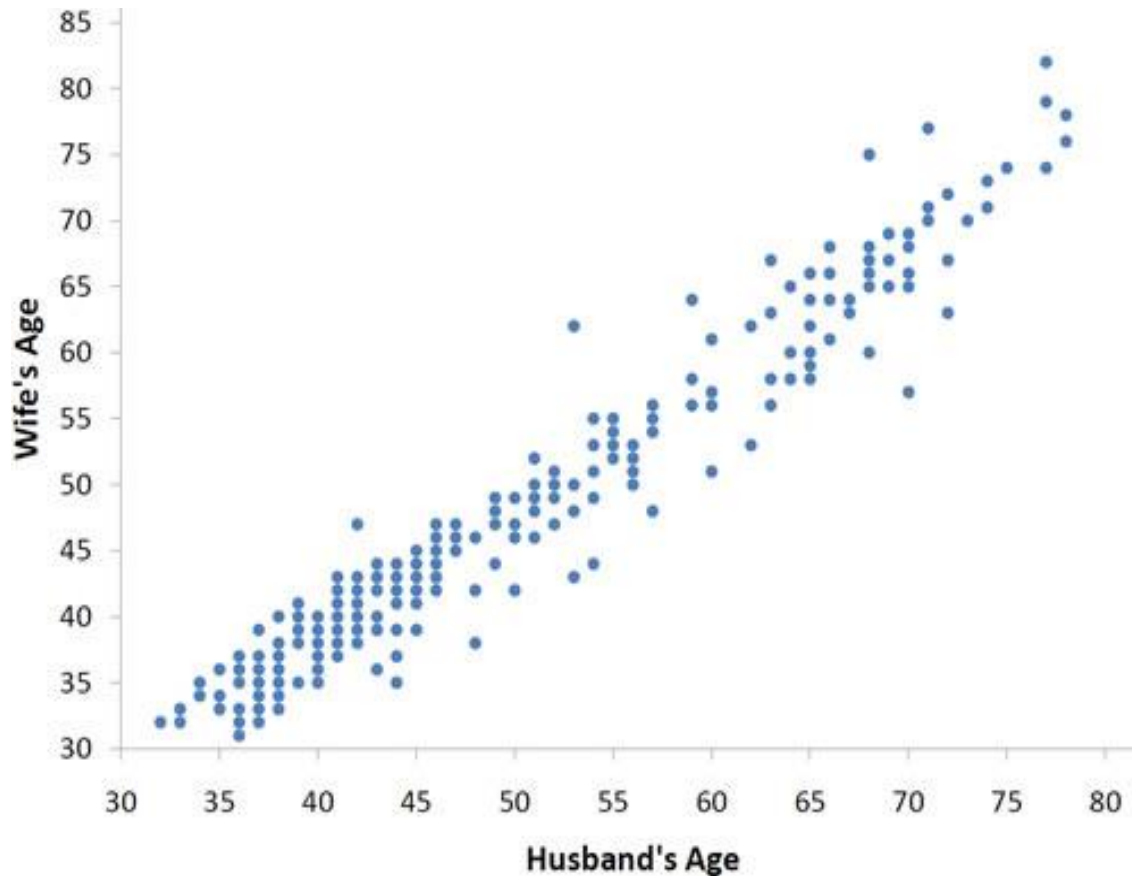
Very good for $n > 10$ and uni-modale distributions – especially good for comparing distributions across different groups.

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 misleading. .

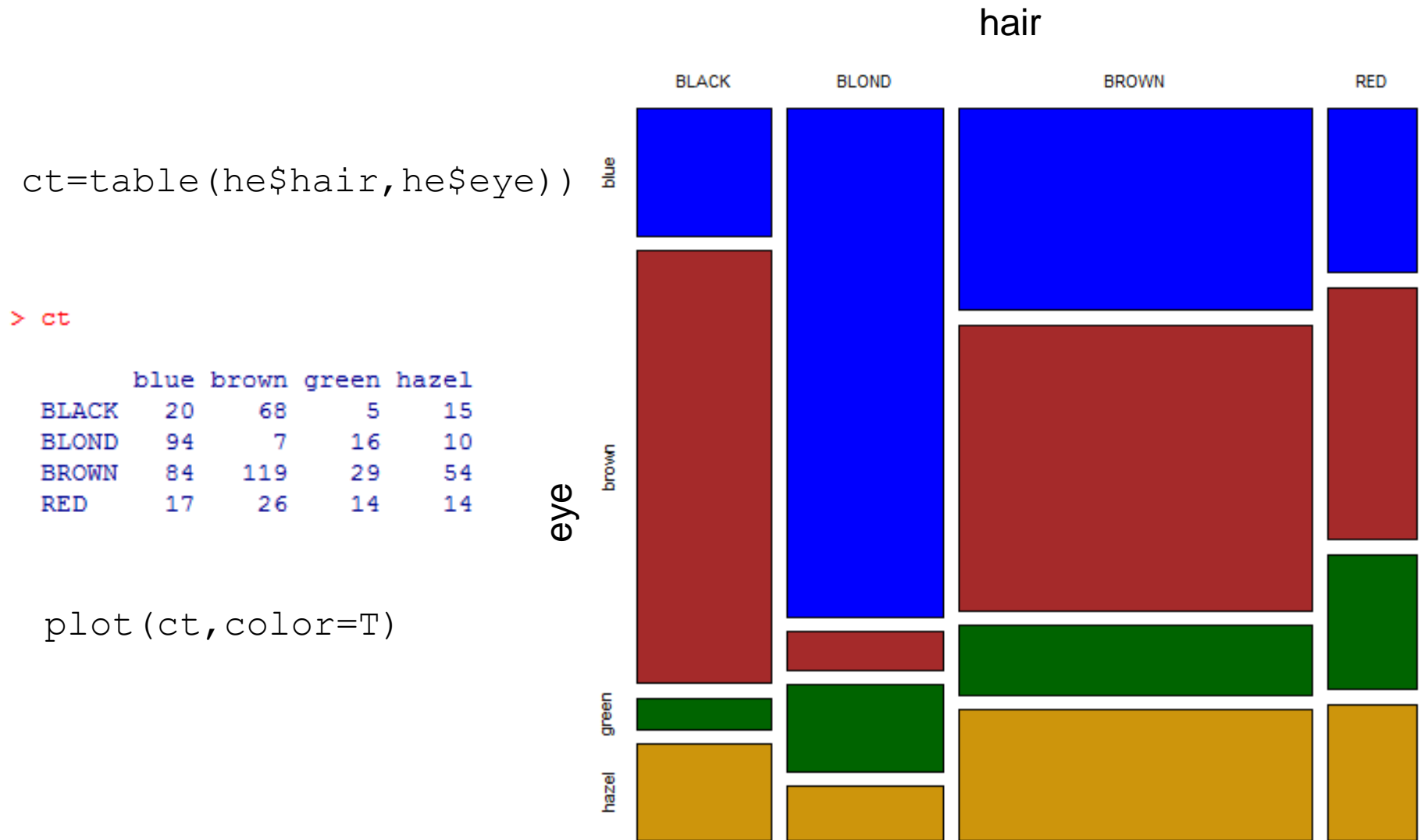
Bi-variate Visualization

Scatterplot for two quantitative variables

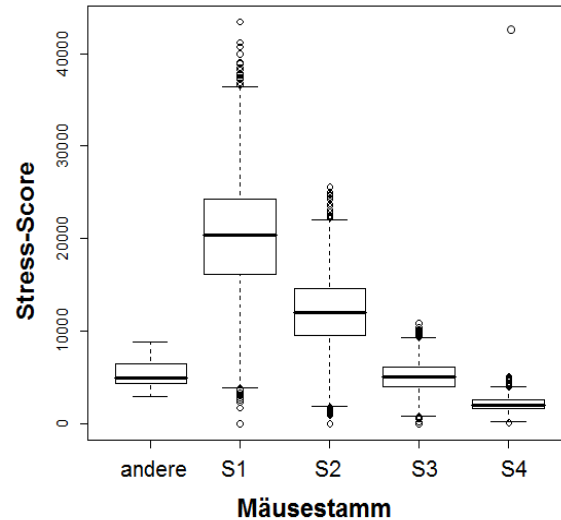
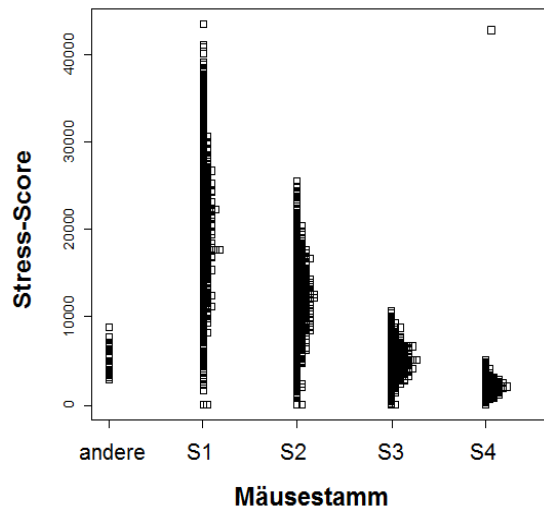


```
plot(dat$m.age, dat$w.age, xlab="Husband's Age", ylab="Wifes's Age" )  
plot( w.age ~ m.age, data=dat)
```

Bi-variate visualization of 2 categorical variables



Bi-variate visualization: continuous vs categorical variable



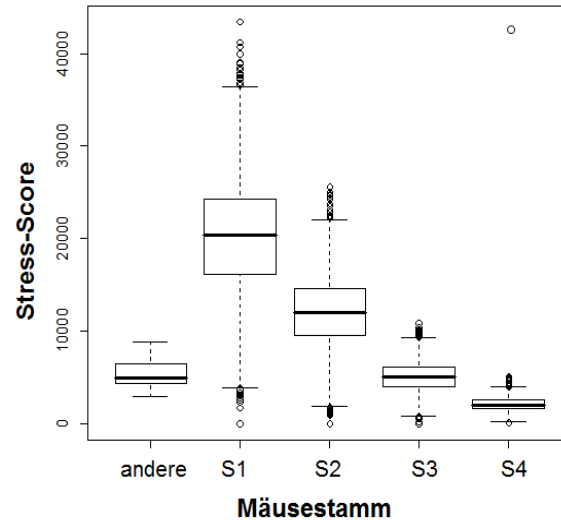
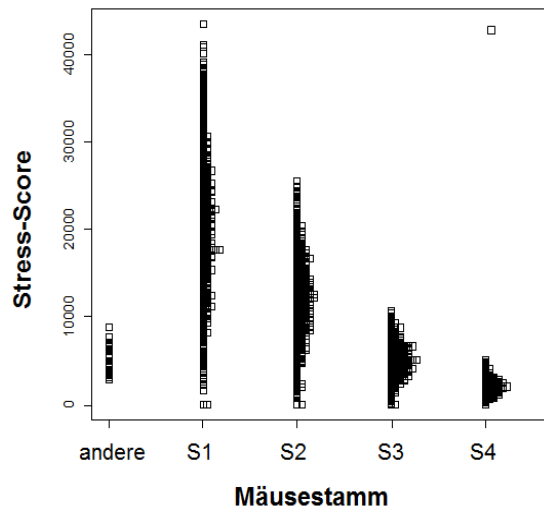
```
stripchart( stress ~ stamm,  
           method="jitter", data=mouse)
```

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

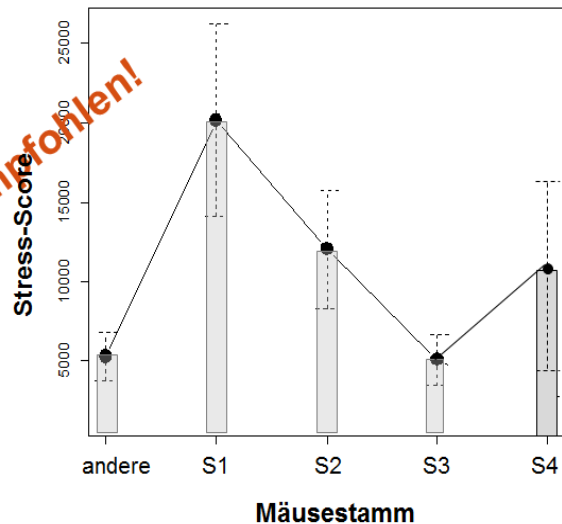
→ Compare distribution of continuous variable in different groups (categories).

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

Bi-variate visualization of 1 categorical and 1 continuous variable



Plot of Means



=> *Boxplots*

=> *Stripcharts*

=> *Histogram*

=> *Mean plots*

very good, if uni-modal

good, if «stacked»

good for revealing the shape

but takes a lot of space

not good, not robust,

Can be miss-leading!

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)