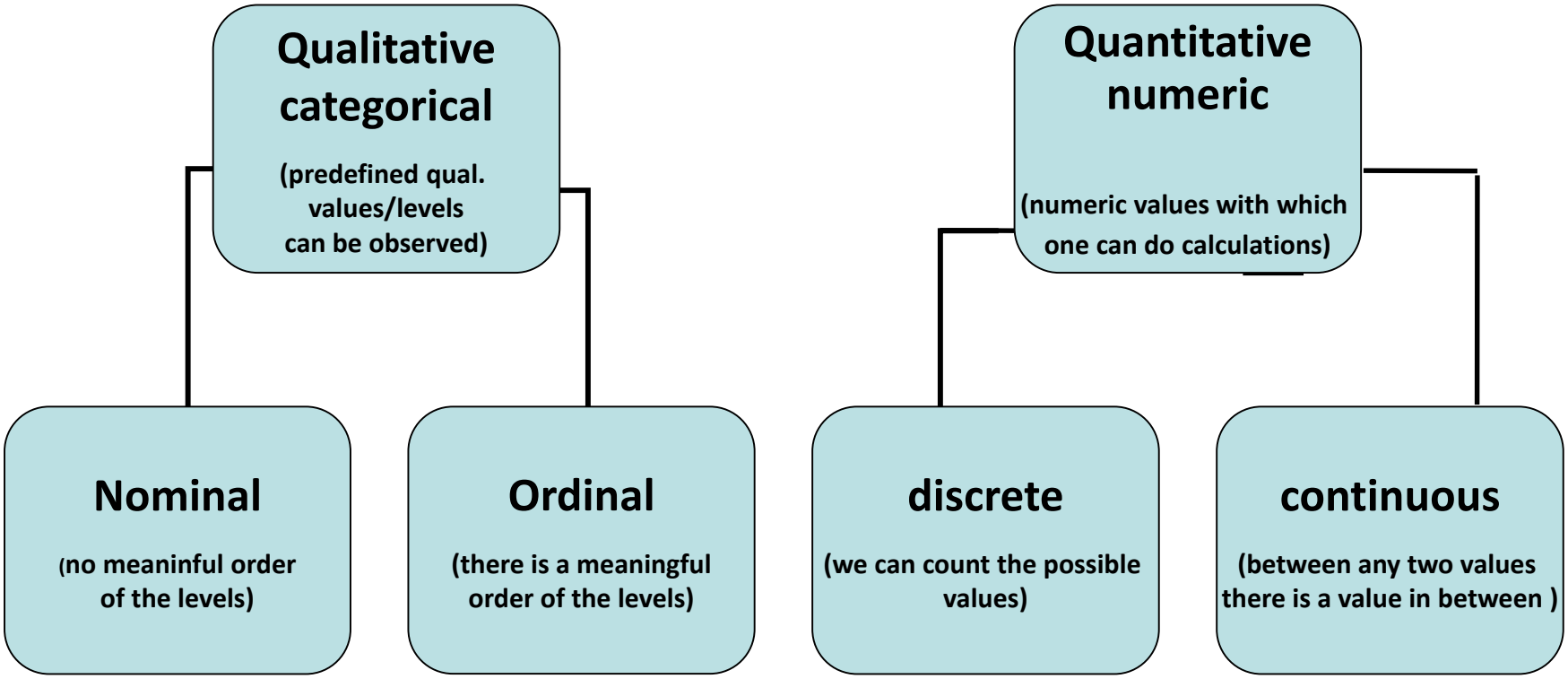
# Biostatistics Week 2

## Topics this week:

### uni-variate descriptive Analysis

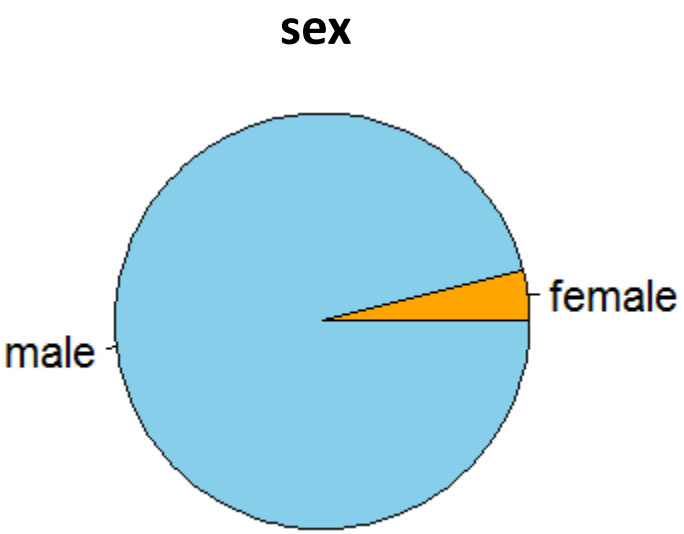
- Data types
- Measure for location
- Visualizing categorical variables: pie chart and barplots
- Visualizing continuos variables: histogram and boxplot
- Confidence intervals of location measures

# There are different types of data

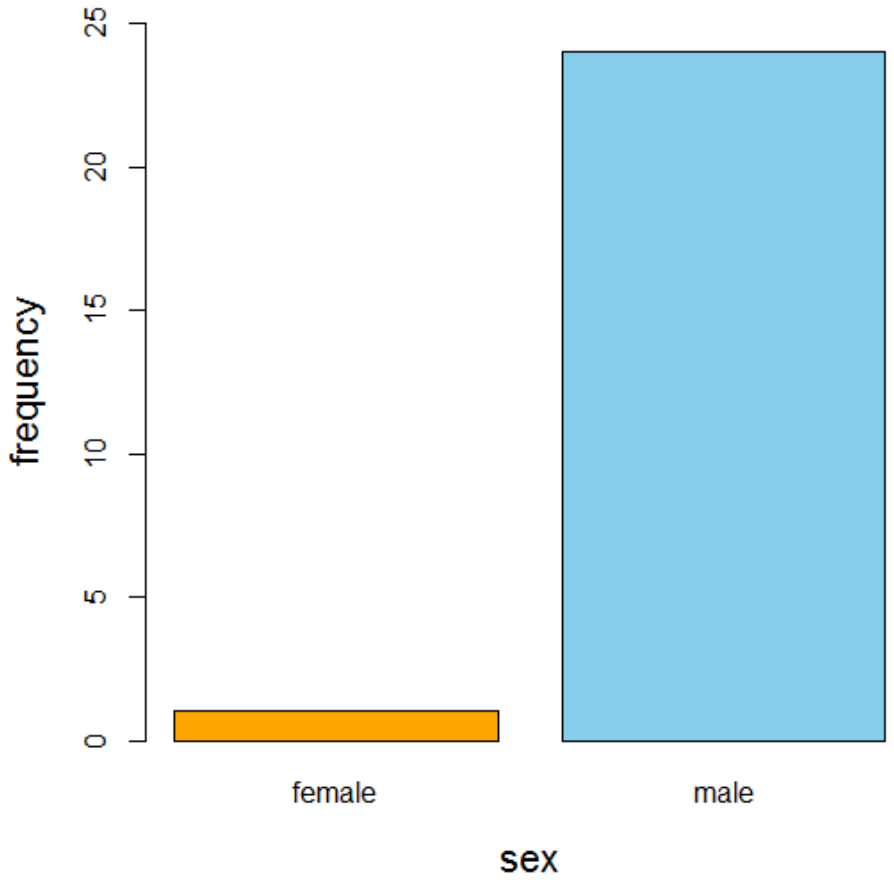


# How to summarize categorical data?

pie chart




bar chart



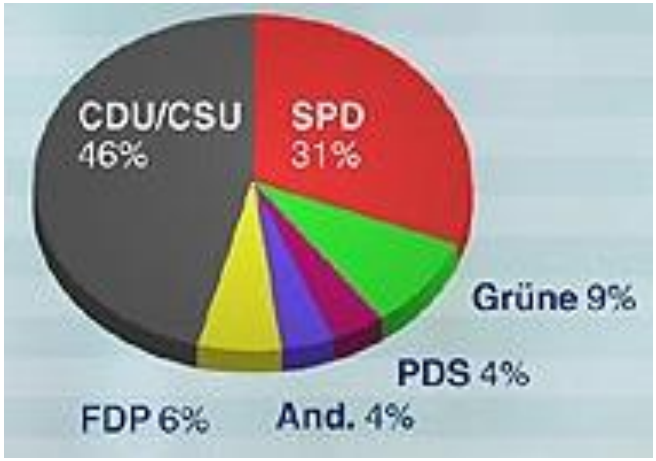

```
pie(table(dat$sex))
```

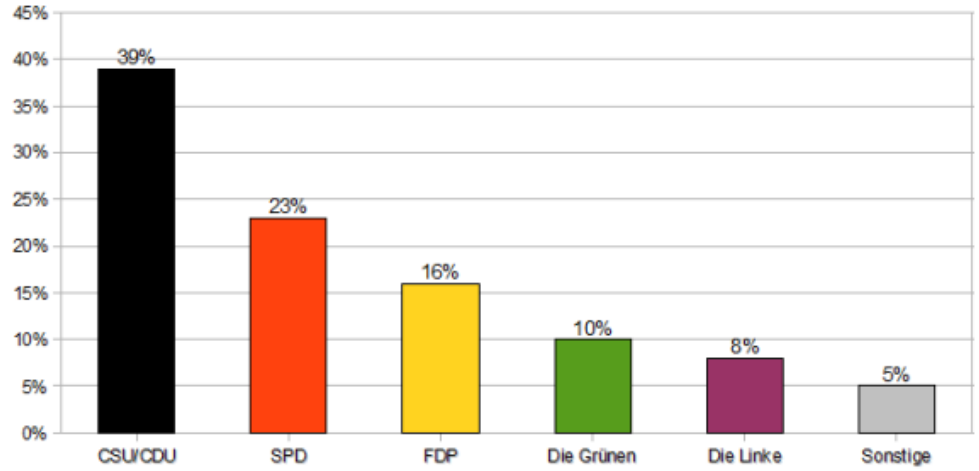
```
barplot(table(dat$sex))
```

Bar charts can transport more information than pie charts.

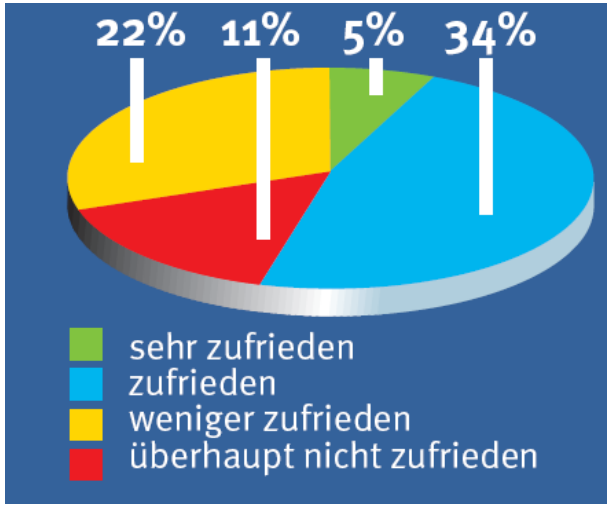
## Visualizing categorical data by Bar-Chart or Pie-Chart

## These charts are simple – is there room form manipulation?



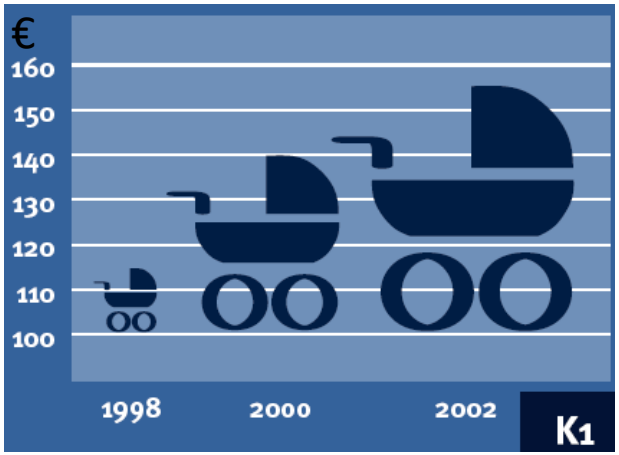
# Half of all reader are satisfied with Klinsmann - true?




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

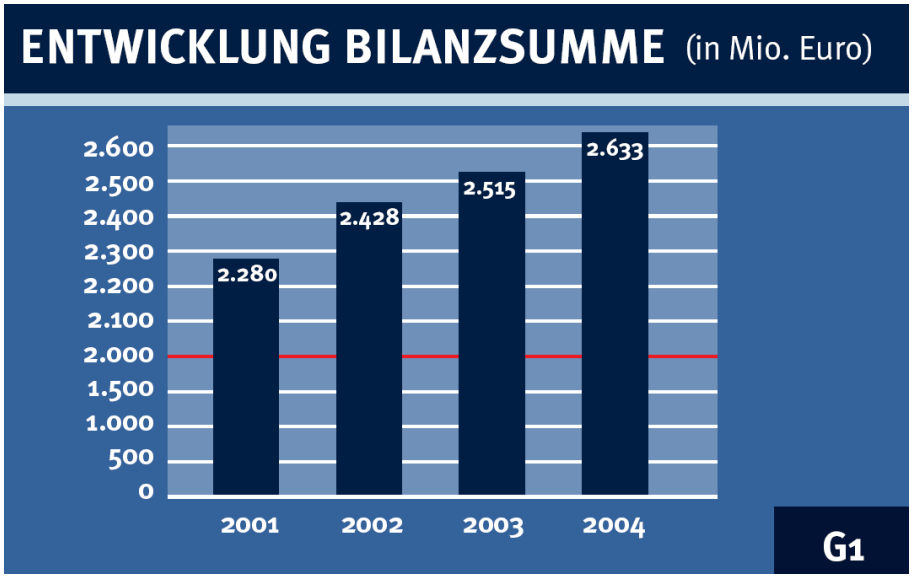
**Quelle: http://www.wdr.de.varnish.gl-systemhaus.de/tv/quarks/sendungsbeitraege/2006/1017/002_zahlen.jsp?smonat=2009-10**

# Generous increase of child allowance - true?



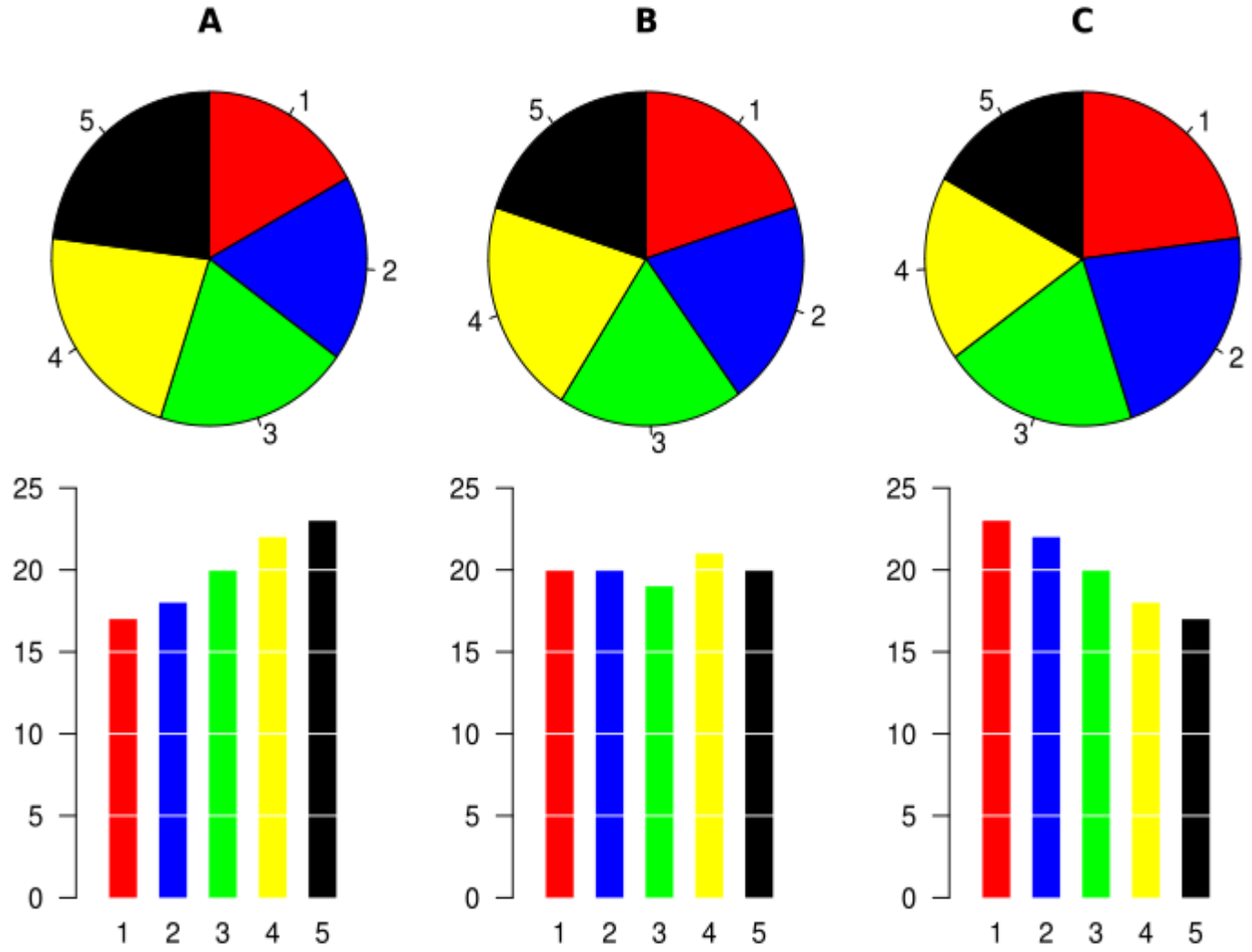

Graph from government statement in the German red-green agenda 2010.

# Good business development - true?




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

# Barplots are often to prefere over pie-charts

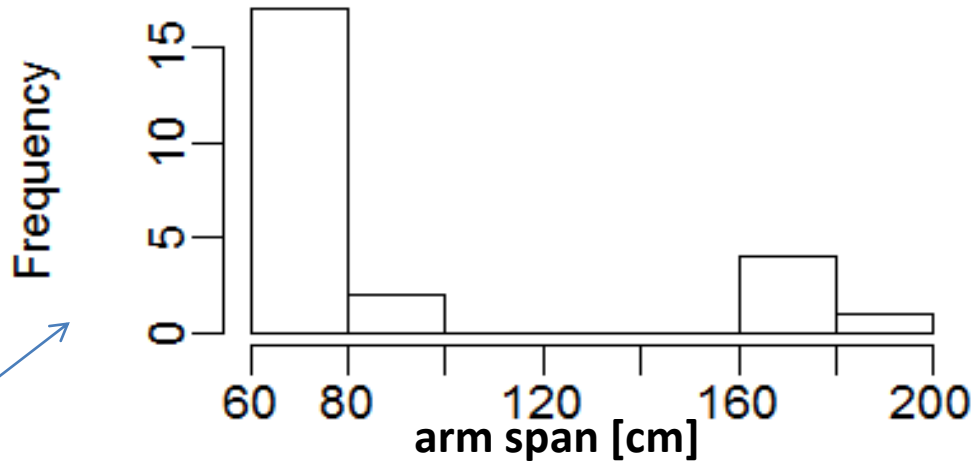



Humans are much better in comparing heights than comparing areas.

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

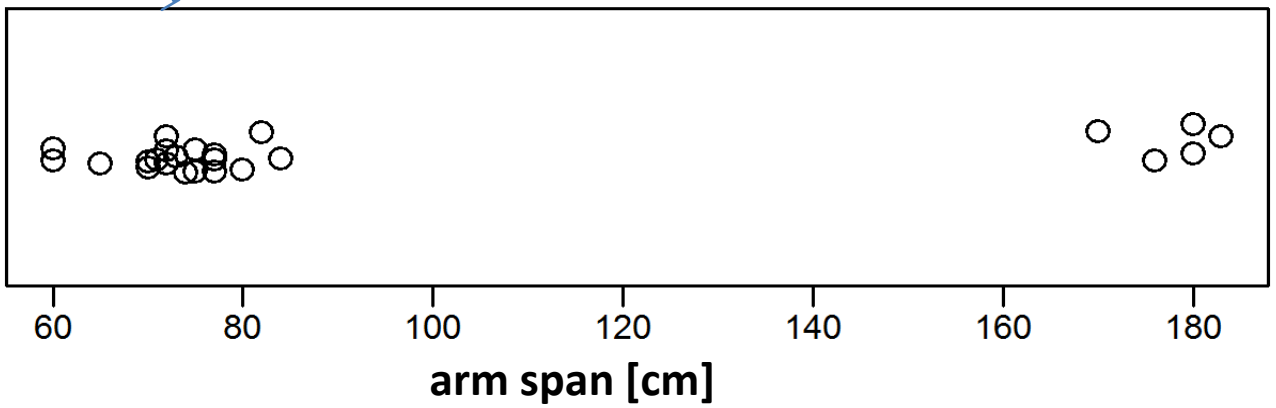
| X: arm span | frequency |
|---|---|
| [60, 80) | 17 |
| [80,100) | 2 |
| [100,120) | 0 |
| [120,140) | 0 |
| [140,160) | 0 |
| [160,180) | 4 |
| [180,200) | 1 |

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




```
hist(x)
```

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



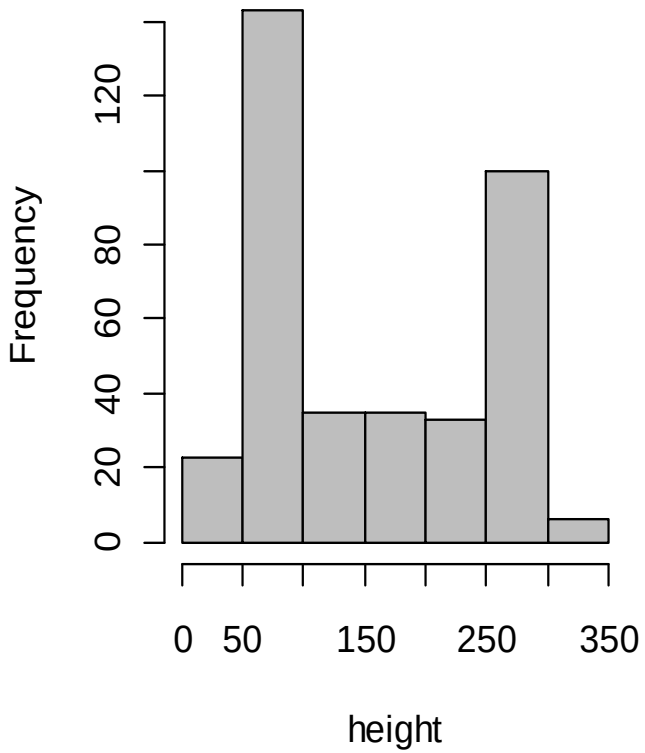
# How to visualize continuous data?

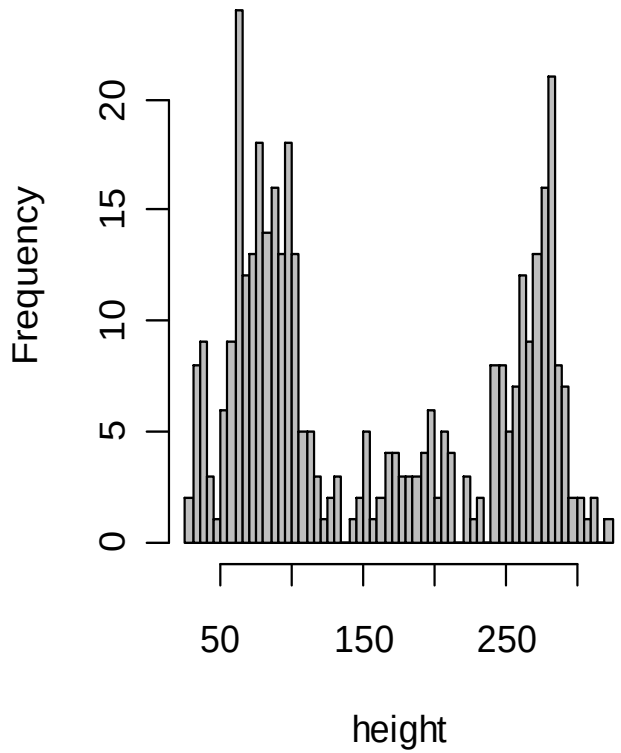
The height (cm) of 376 plants were measured.

```
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)
```

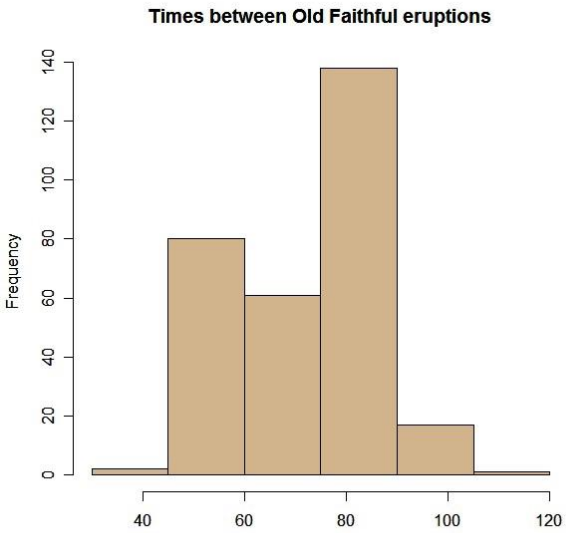


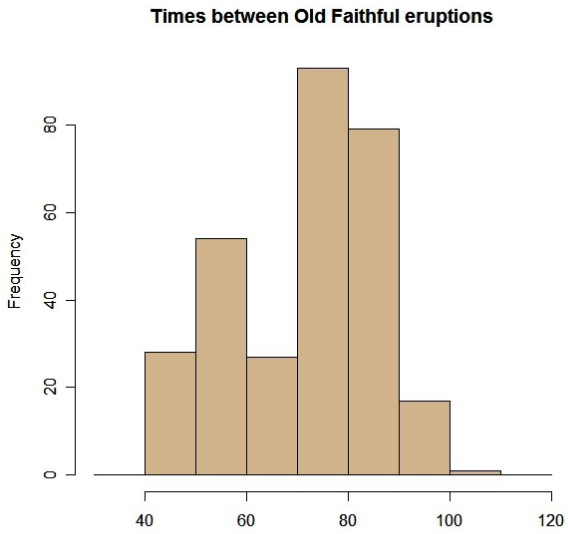



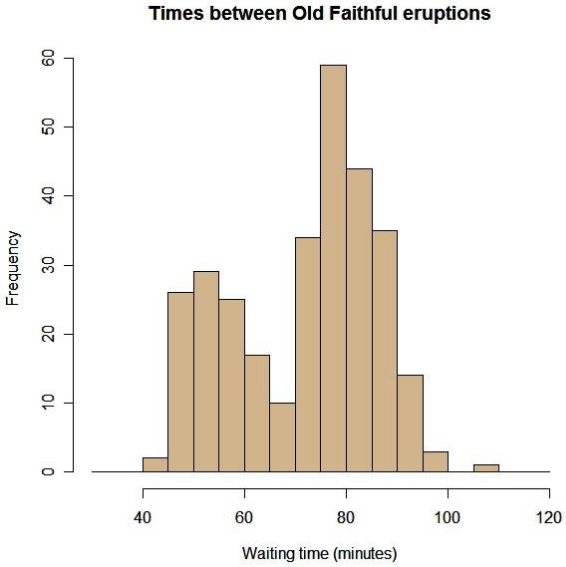

Are there subgroups? If yes – how many?

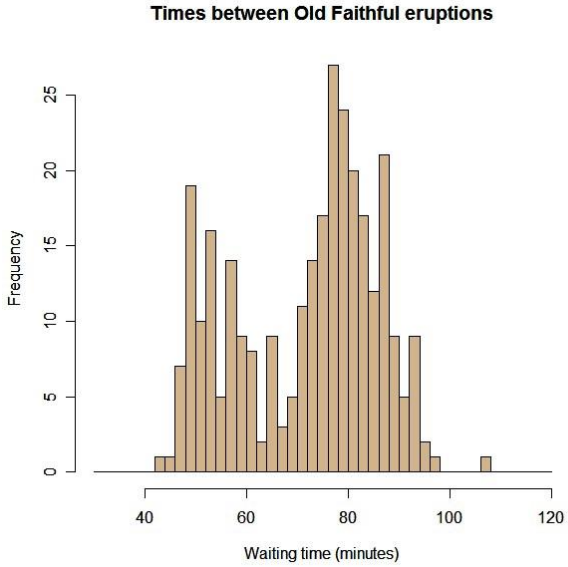
How reliable is the height of a bar?

# How many classes do we need?



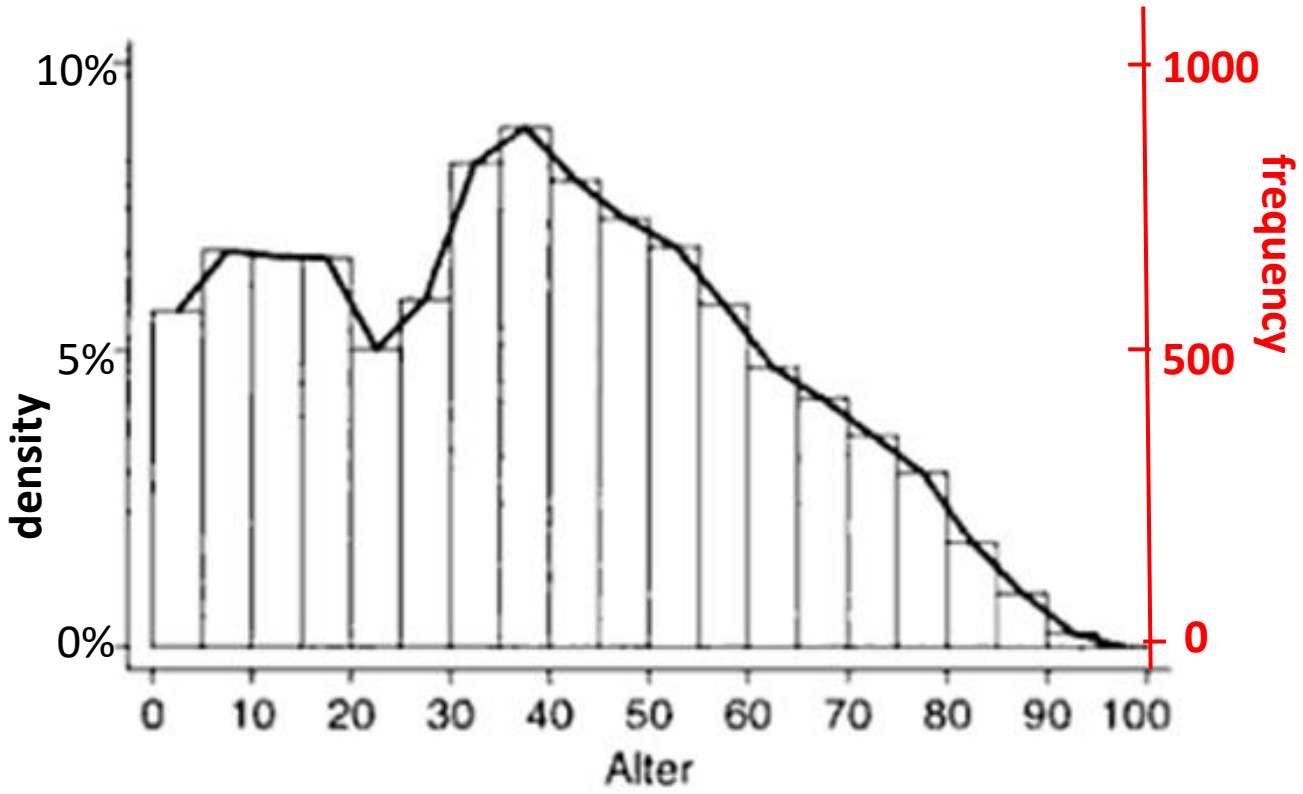




http://www.amstat.org/publications/jse/v6n3/applets/Histogram.html

299 eruption intervals were observed

Shape of the histogram may depend on the class choices

# How does the distribution of age look like

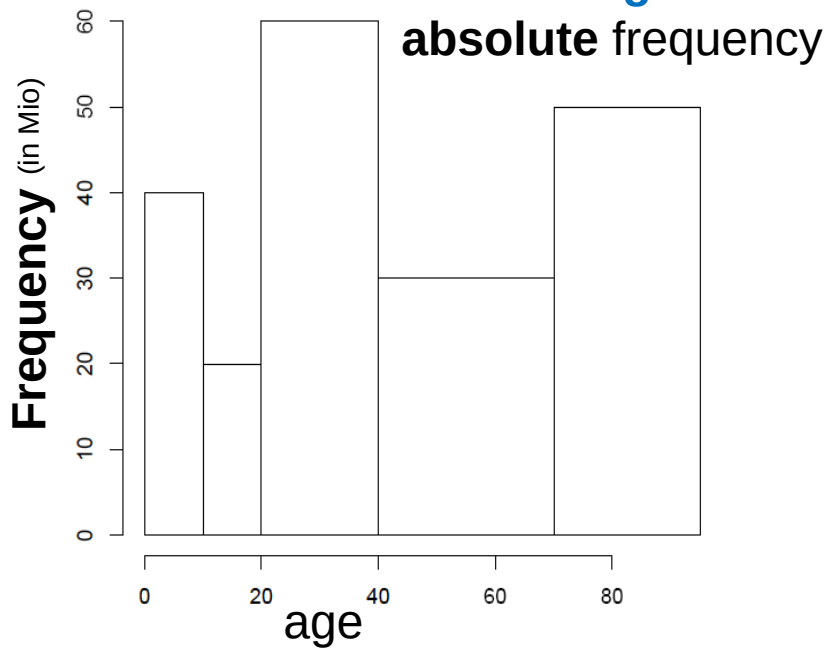



**Only <u>in case of equally sized bins</u> the scaled and unscaled histogram look the same and it is possible to label the y-axis with percentages – usually it ony shows the density!**

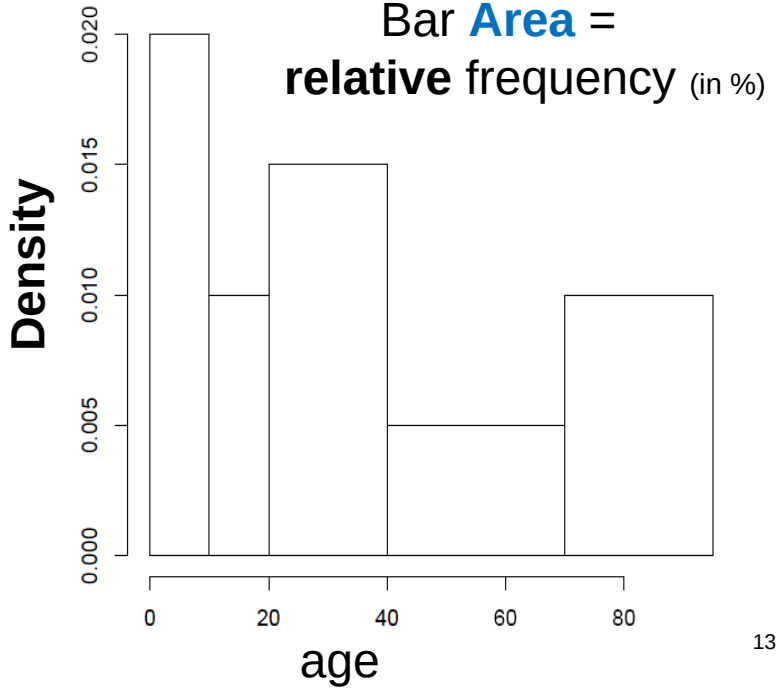
# Unscaled and scaled histograms

| 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

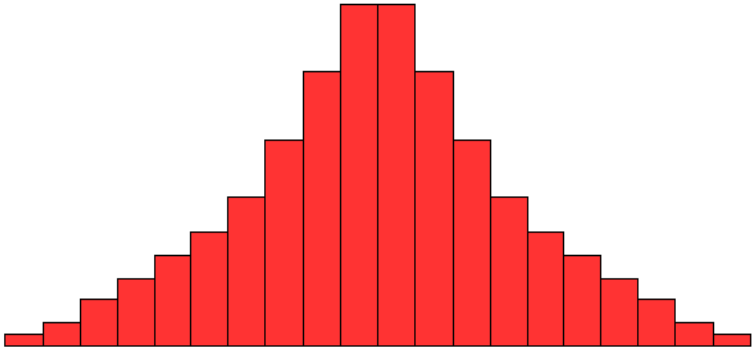
Zurich University of Applied Sciences

zh aw

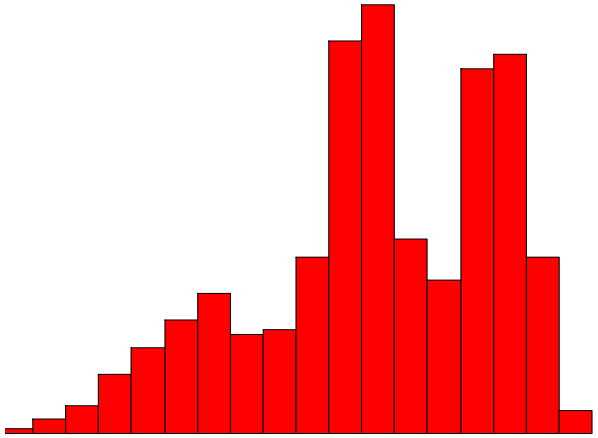
- Avoid classes with different width! (shape will change)
- How many classes: $\sqrt{n}$ classes for n observations.
- The shape can depend on the number classes and the class limits.

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.

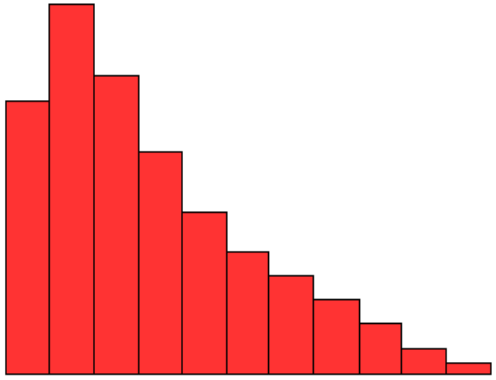
# Shapes of distributions



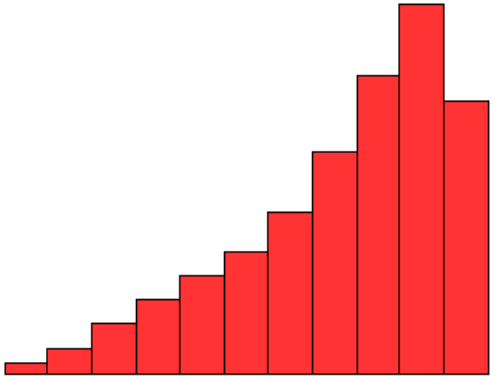
Symmetric, uni-modale



Multi-modale, slightly left-skewed



Right-skewed, uni modale



Left-skewed, uni-modale

# Measures for the location and variation

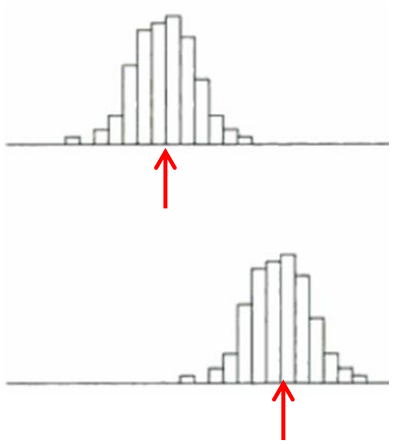
**Data can be summarized by summary statistics. Most important key figure describe the center and the width of a distribution..**

---

## 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.**

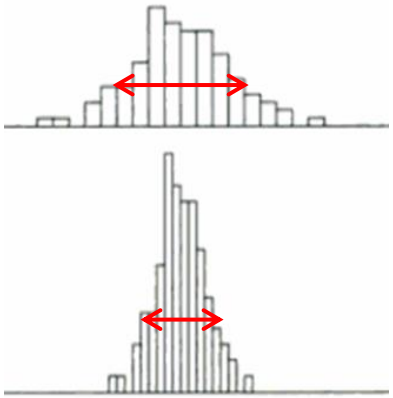




Bild: W. Stahel

# Is the mean salary a «typical salary»?

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


**Schweizer Arbeitsplätze**

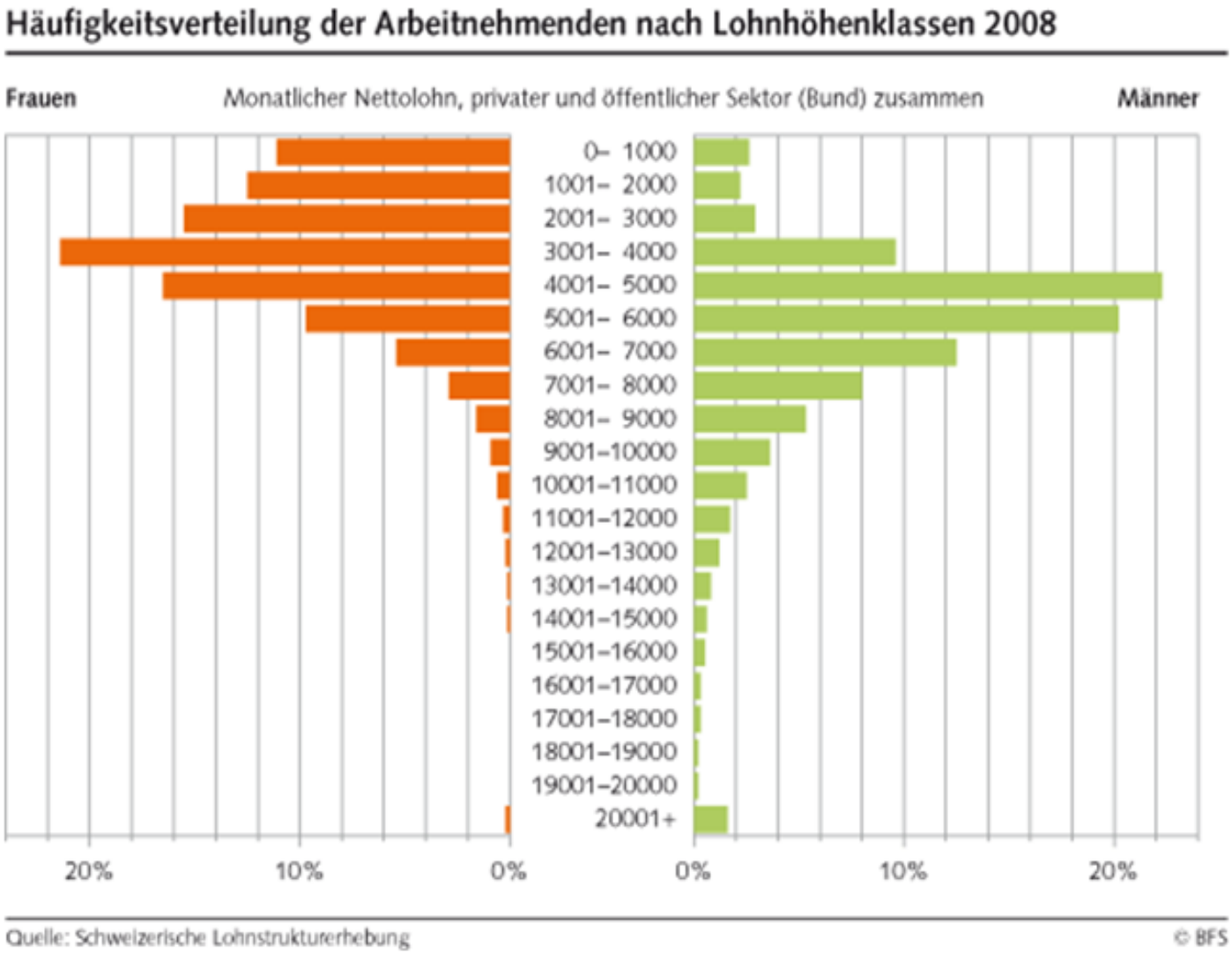
Grafik vergrössern

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.

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

# Distribution of salaries in Switzerland

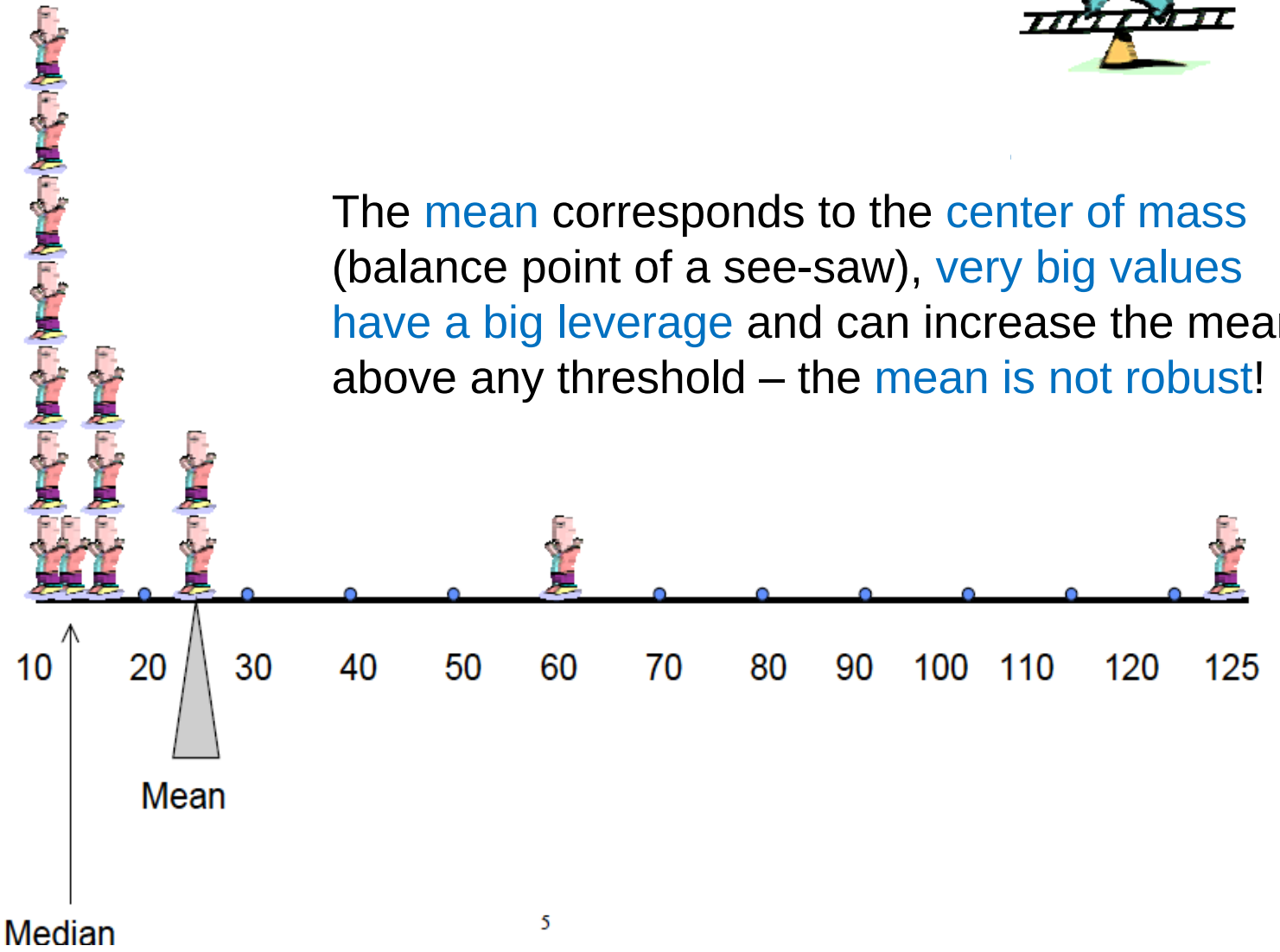





18

# 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:

$$x_1, x_2, \ldots, x_n \qquad x_{[1]} \le x_{[2]} \le \cdots \le x_{[n]}$$

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

# What is „the mean“ income in this company?

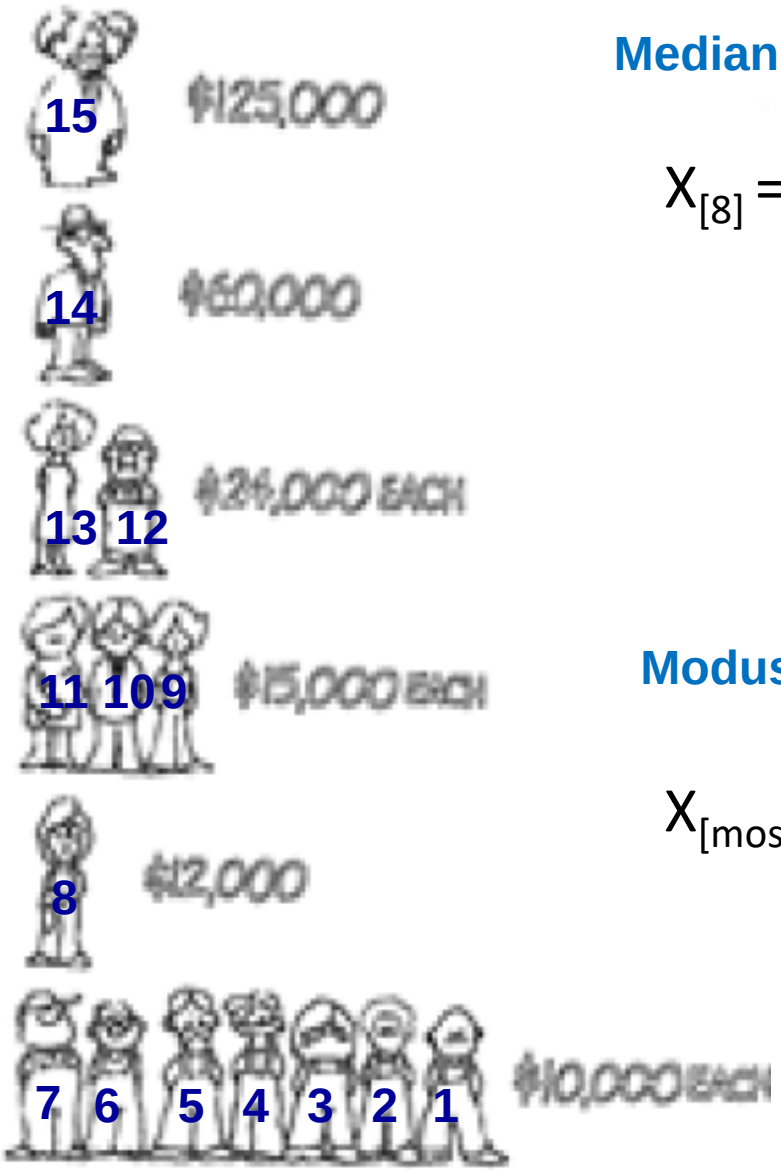
**Mean:**

**Sum:**

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

= 360,000

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




**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.

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

• **Mean:**

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

# 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 $\alpha$%-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.

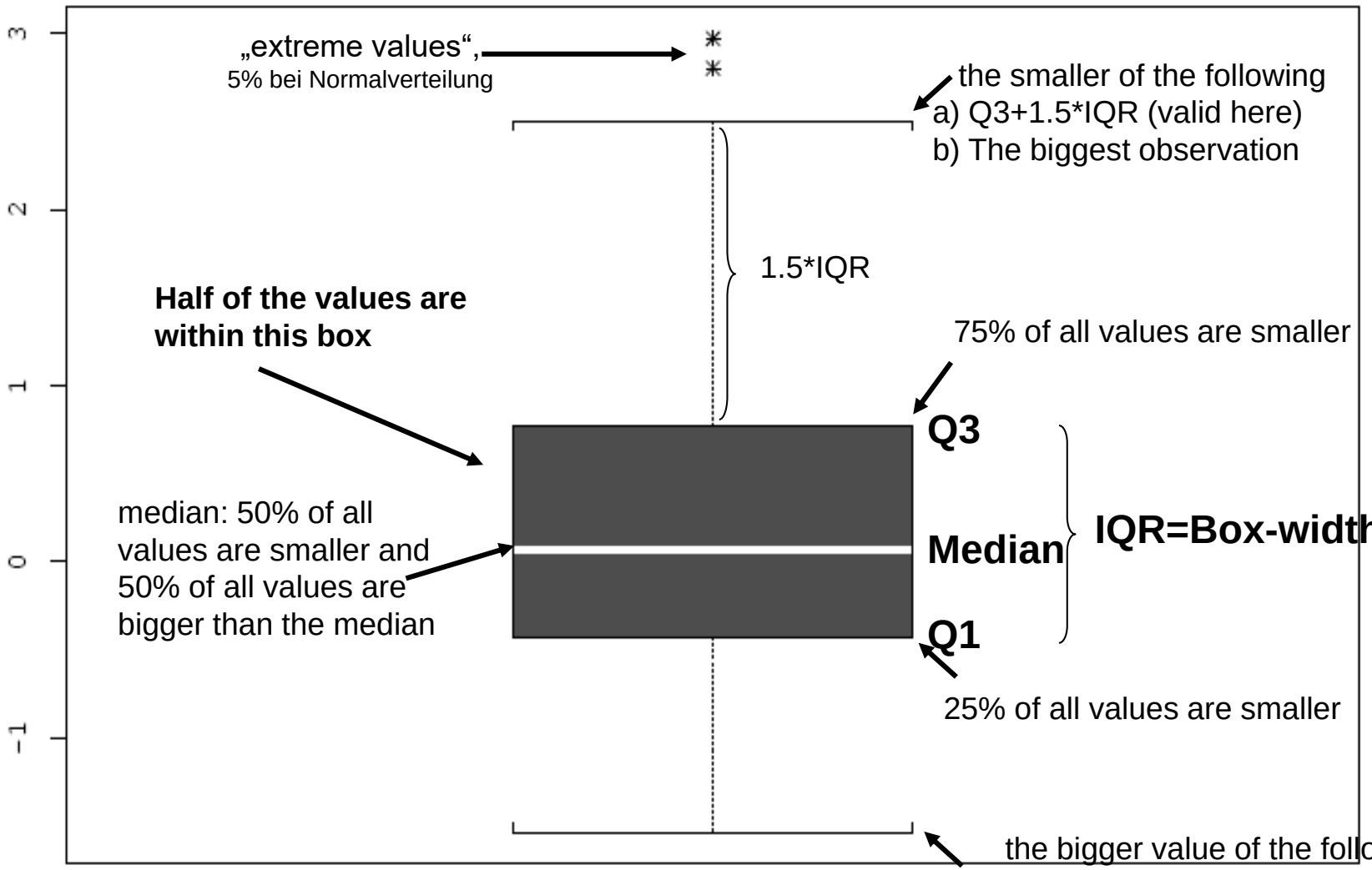
# How to determine the $\alpha$-Quantil of a sample

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.

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

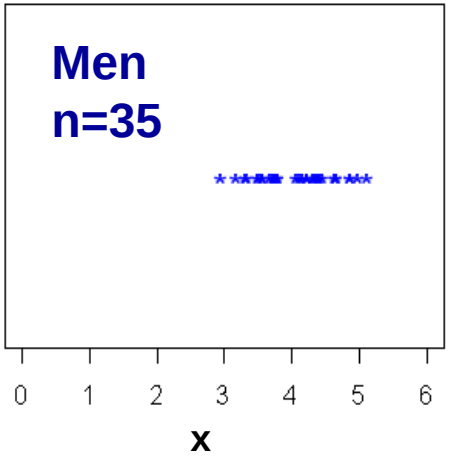
Example:
Sample: 3, 4, 7, 5, 0.5, 6 : n=6 ; ordered sample: 0.5, 3, 4, 5, 6, 7
To determine the Median=$^{0.5}$q we determine the ordinal numbers
[$\alpha$*n]=[0.5*6] =[3] -> Median are the average of the third- and forth-smallest value:
Median=0.5*($x_{[3]}$+$x_{[4]}$)=0.5*(4+5)=4.5 .

# Definition of the Boxplot to visualize continuous data

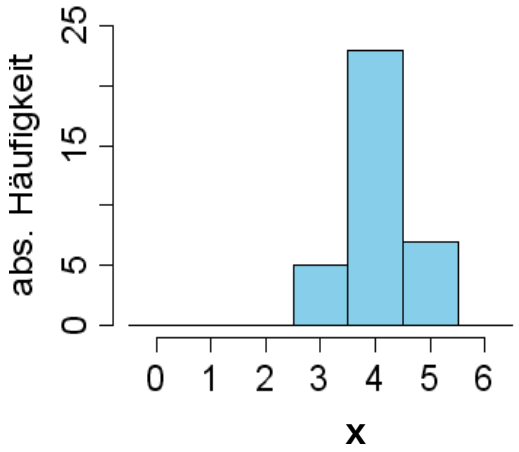


# How to best visualize continuous data

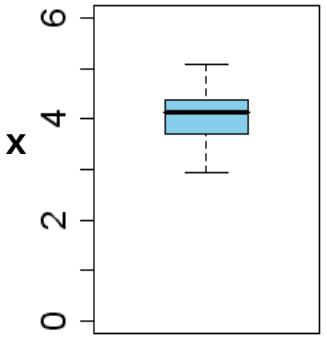
X: reduction of the BMI after take-in of a new drug



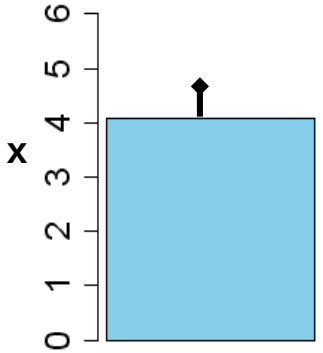

**Stripchart**




**Histogramm**

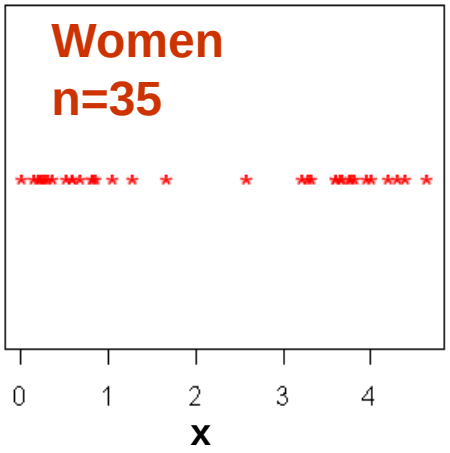


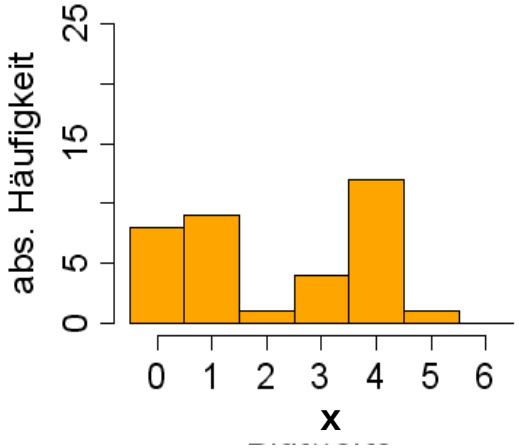

**Boxplot**

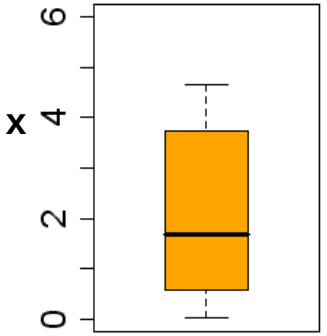


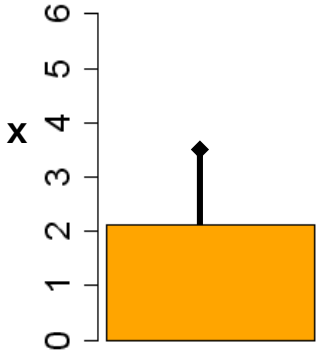

**Mean-plot**

**(not recommended)**



# How to best visualize continuous data

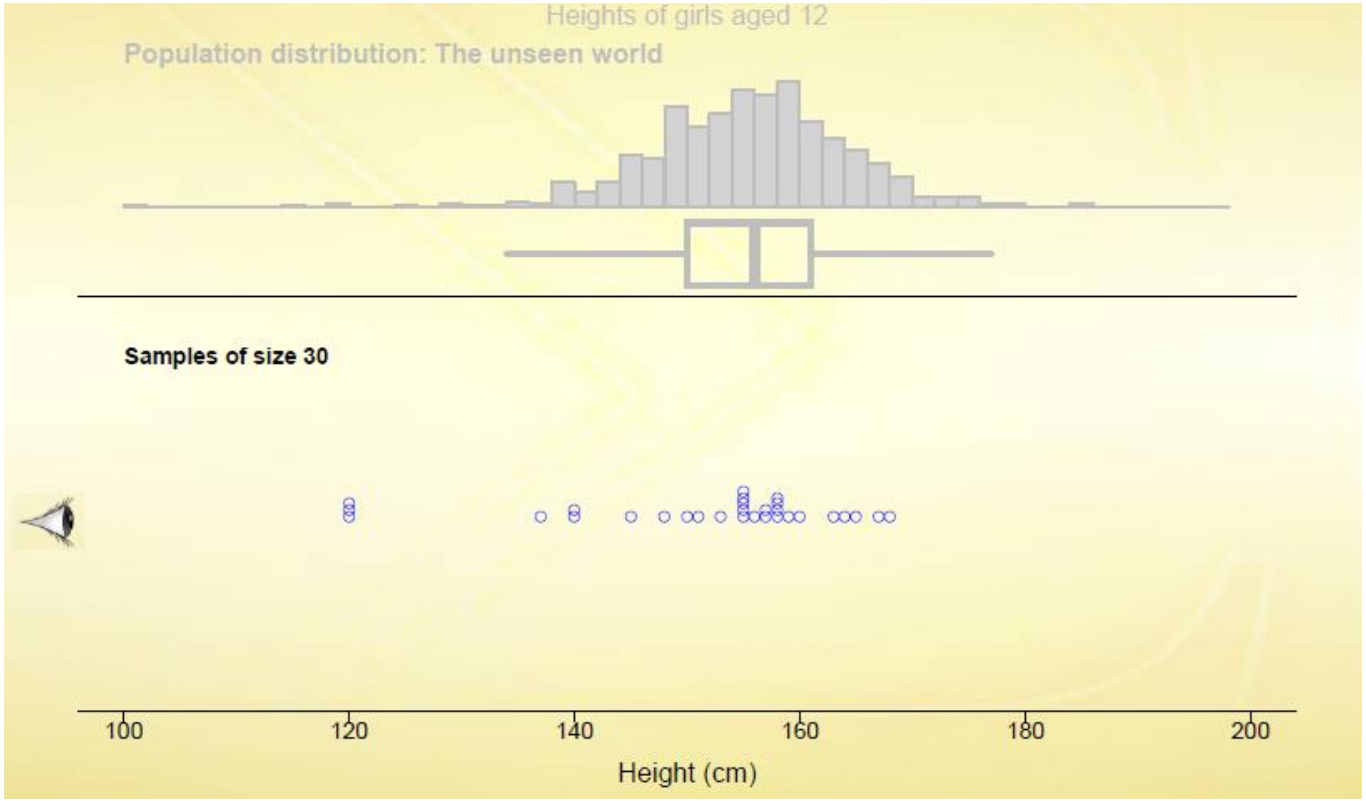
Stripchart:
for n<20 it is a good plot, since it shows each data point and gives an impression about the distribution.

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. .

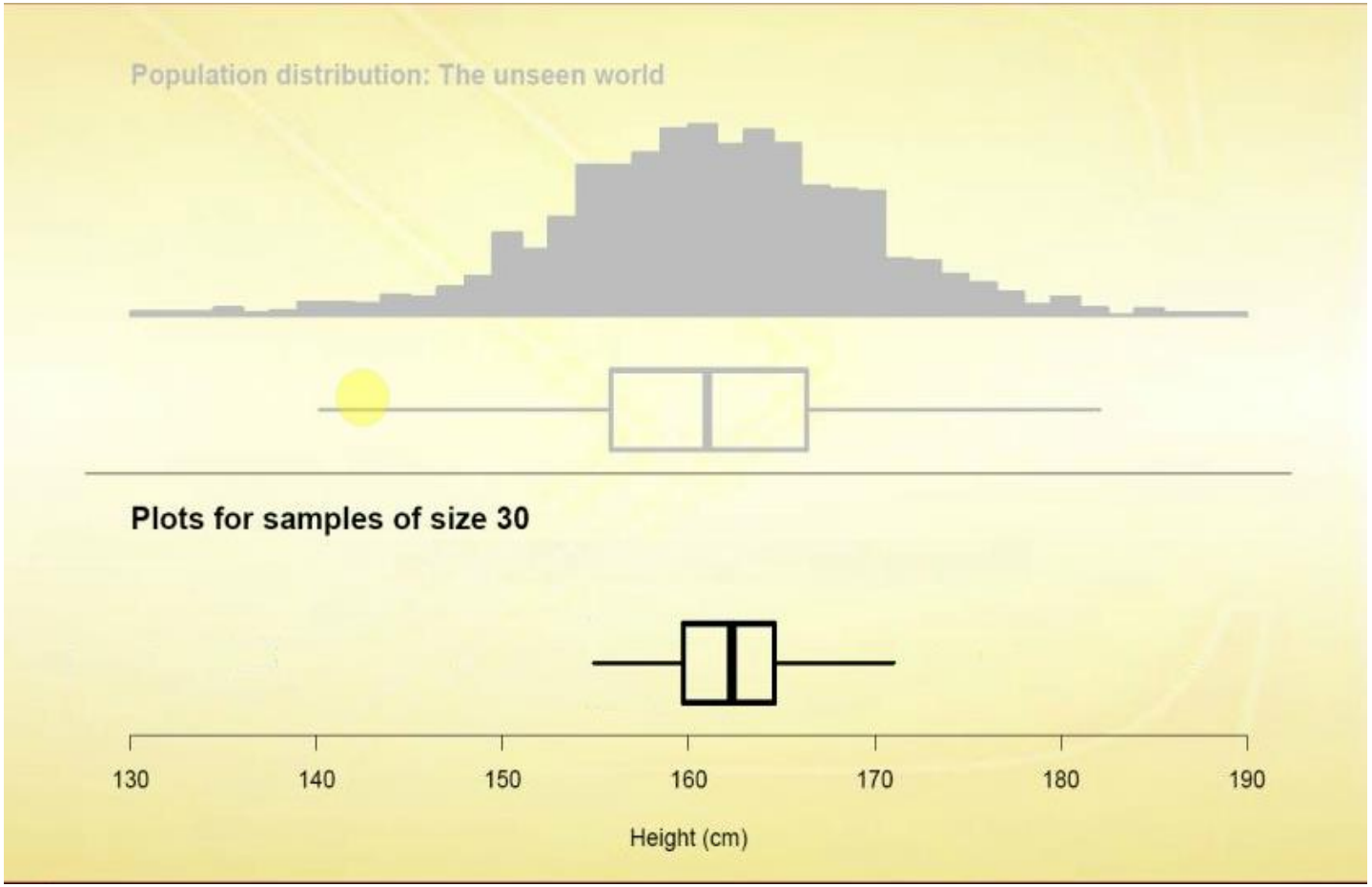
# Where is the center of the population?




n=30

CtsVar_1samp_Dots30.pdf

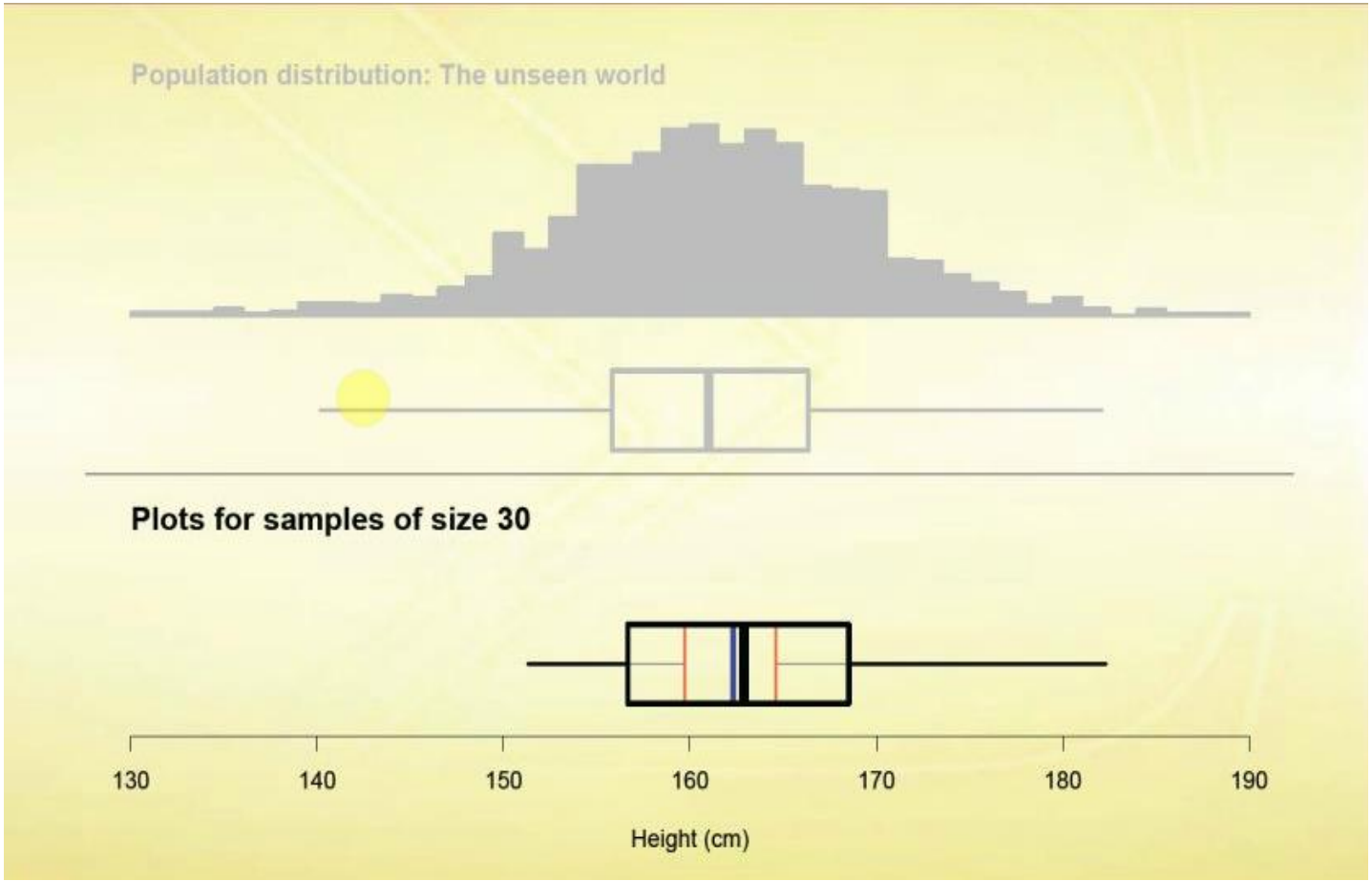
# Where is the center of the population?




Visualize boxplot with memory

Quelle: http://www.stat.auckland.ac.nz/~wild/09.USCOTSTalk.html

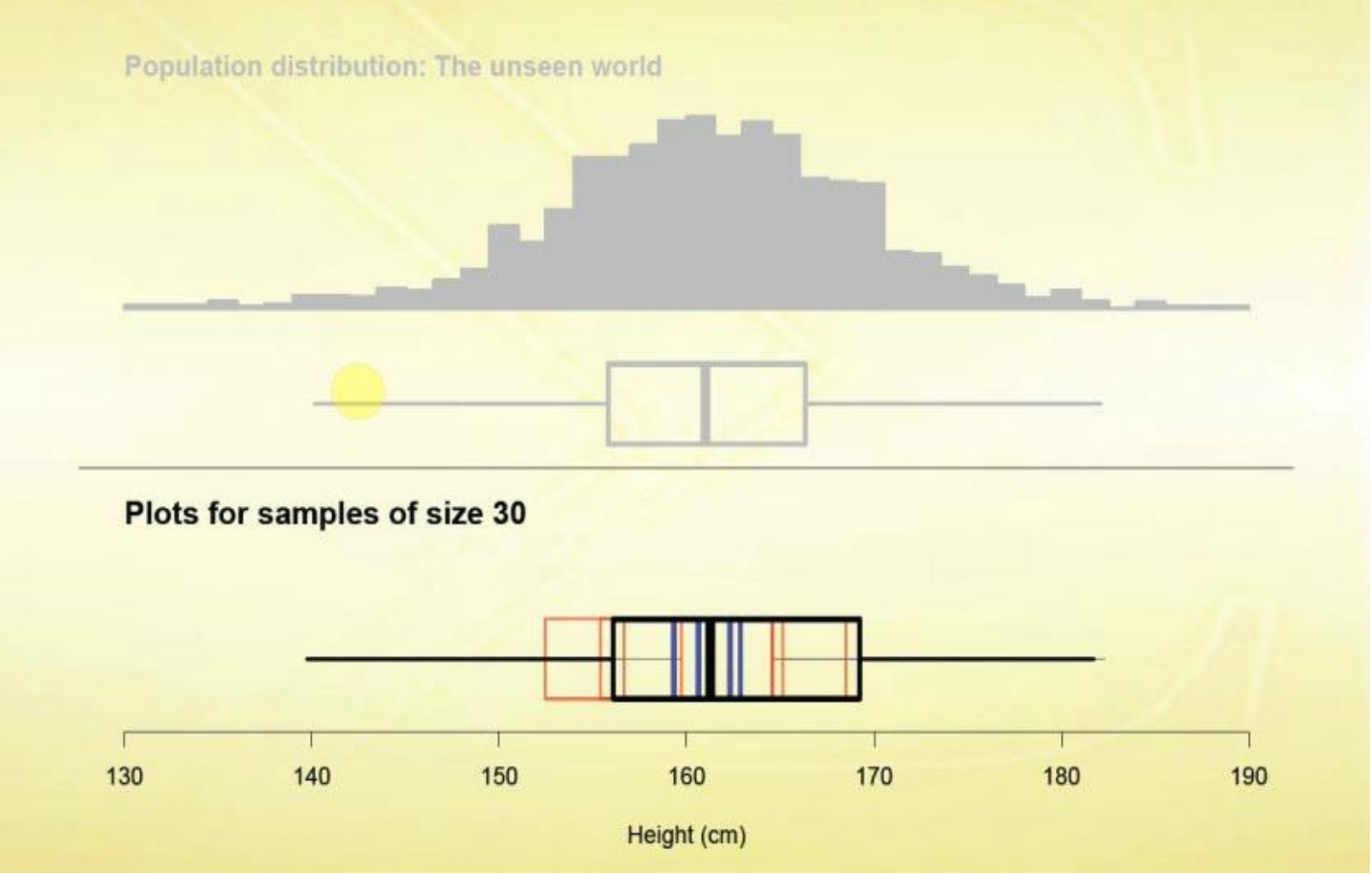
# Where is the center of the population?




Visualize boxplot with memory

Quelle: http://www.stat.auckland.ac.nz/~wild/09.USCOTSTalk.html

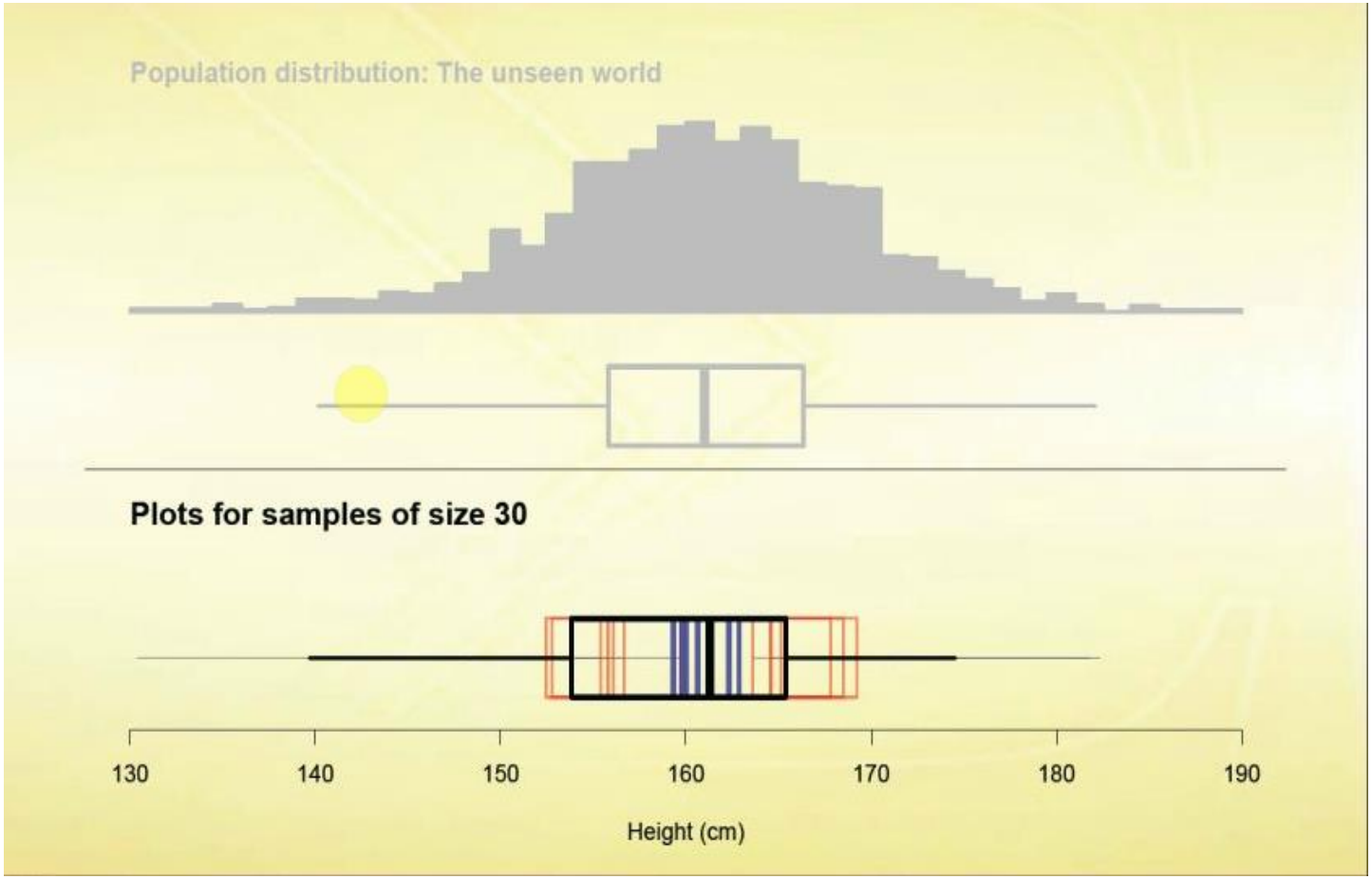
# Where is the center of the population?




Visualize boxplot with memory

Quelle: http://www.stat.auckland.ac.nz/~wild/09.USCOTSTalk.html

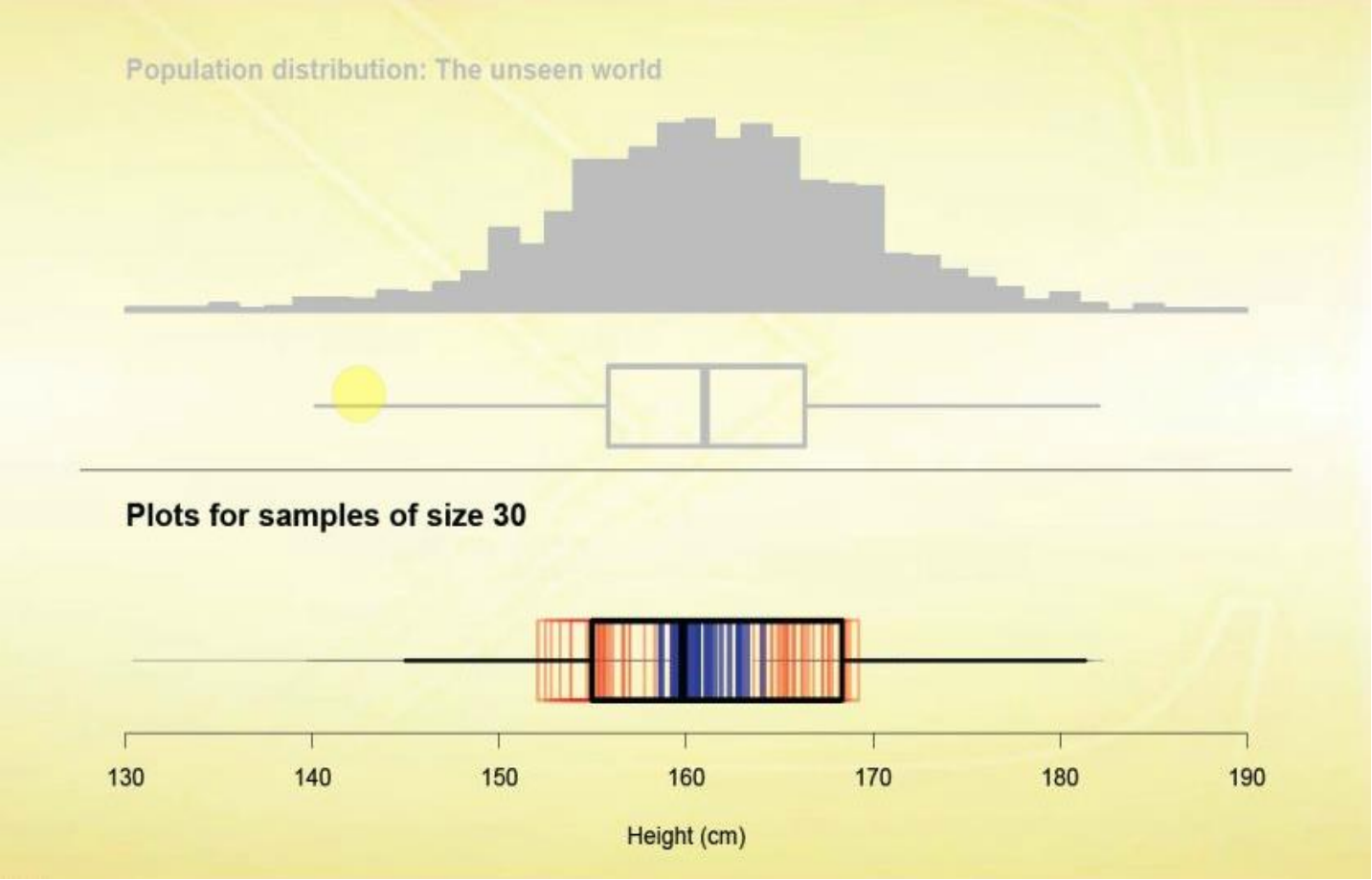
# Where is the center of the population?




Visualize boxplot with memory

Quelle: http://www.stat.auckland.ac.nz/~wild/09.USCOTSTalk.html

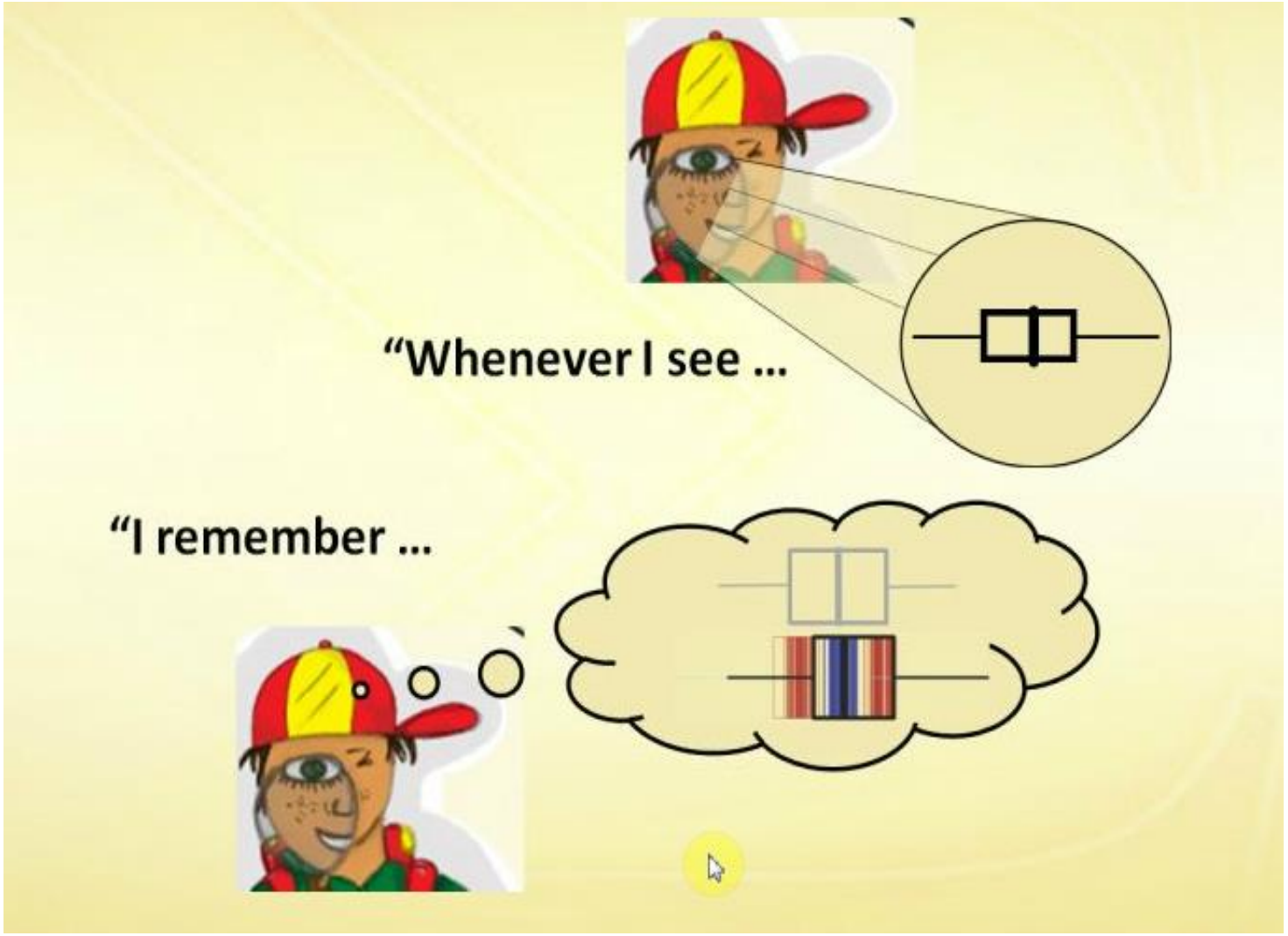
# Where is the center of the population?




Vizualize boxplot with memory

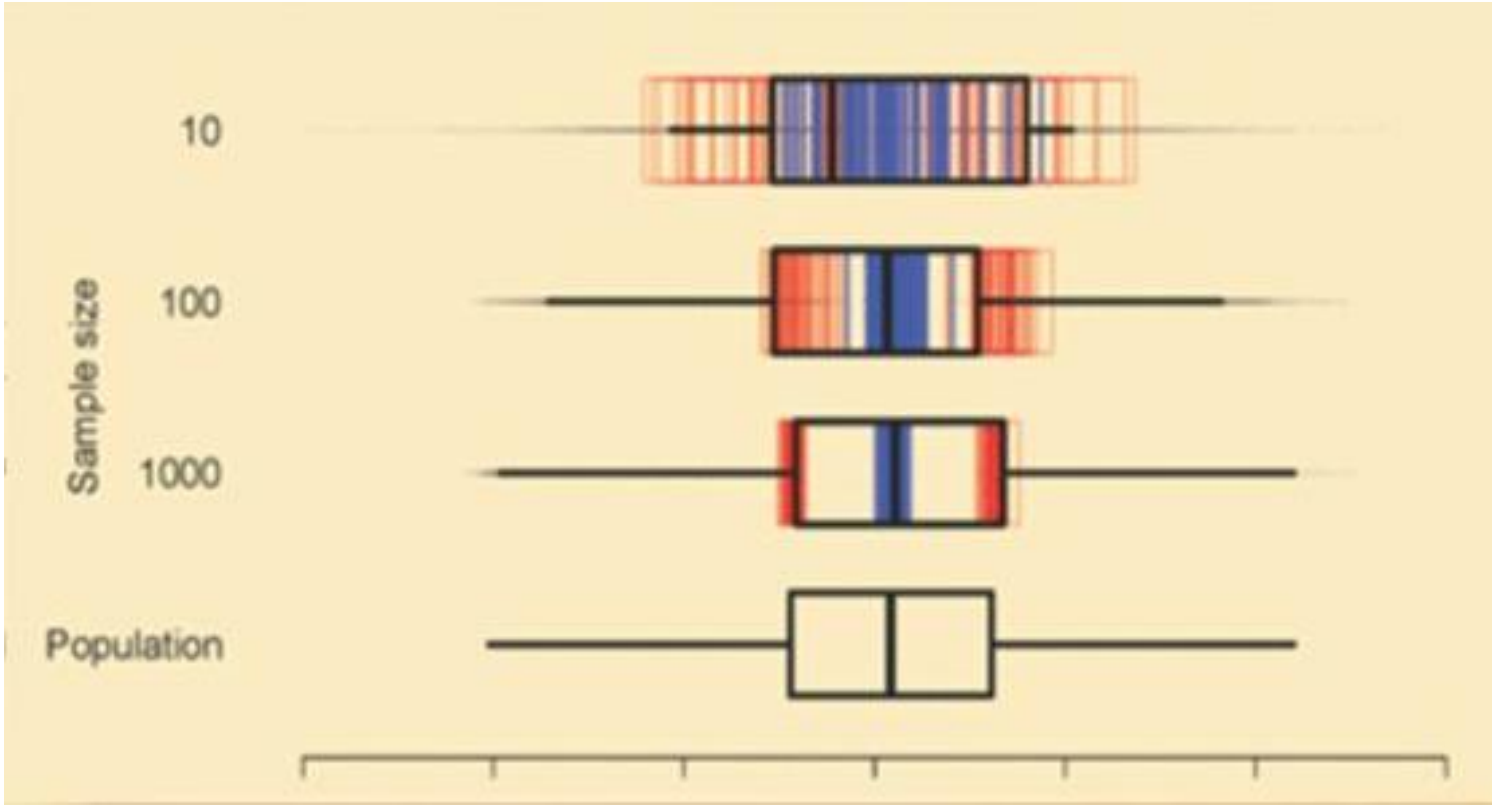
Quelle: http://www.stat.auckland.ac.nz/~wild/09.USCOTSTalk.html

# Where is the center of the population?




Quelle: http://www.stat.auckland.ac.nz/~wild/09.USCOTSTalk.html

# Where is the center of the population?
# We get more certain with increasing sample size

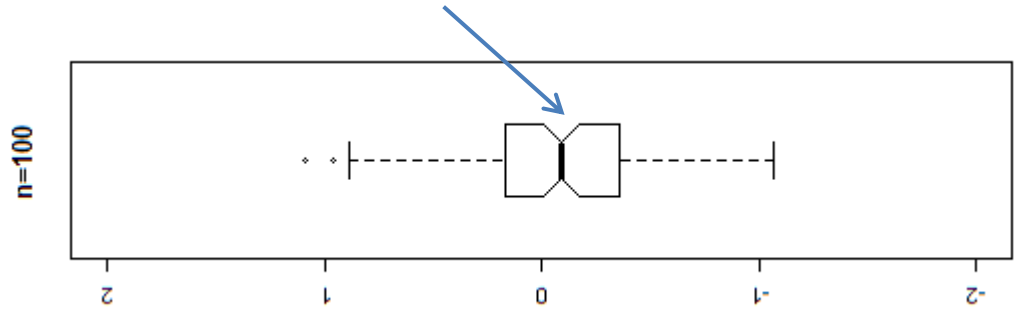



Quelle: http://www.stat.auckland.ac.nz/~wild/09.USCOTSTalk.html

# How sure can I be about the true parameter value?

**Goal:**
We would like to determine from our sample/observations an interval, which covers the true parameter value with a probability of 95%.

```
boxplot(x,notch=TRUE)
```

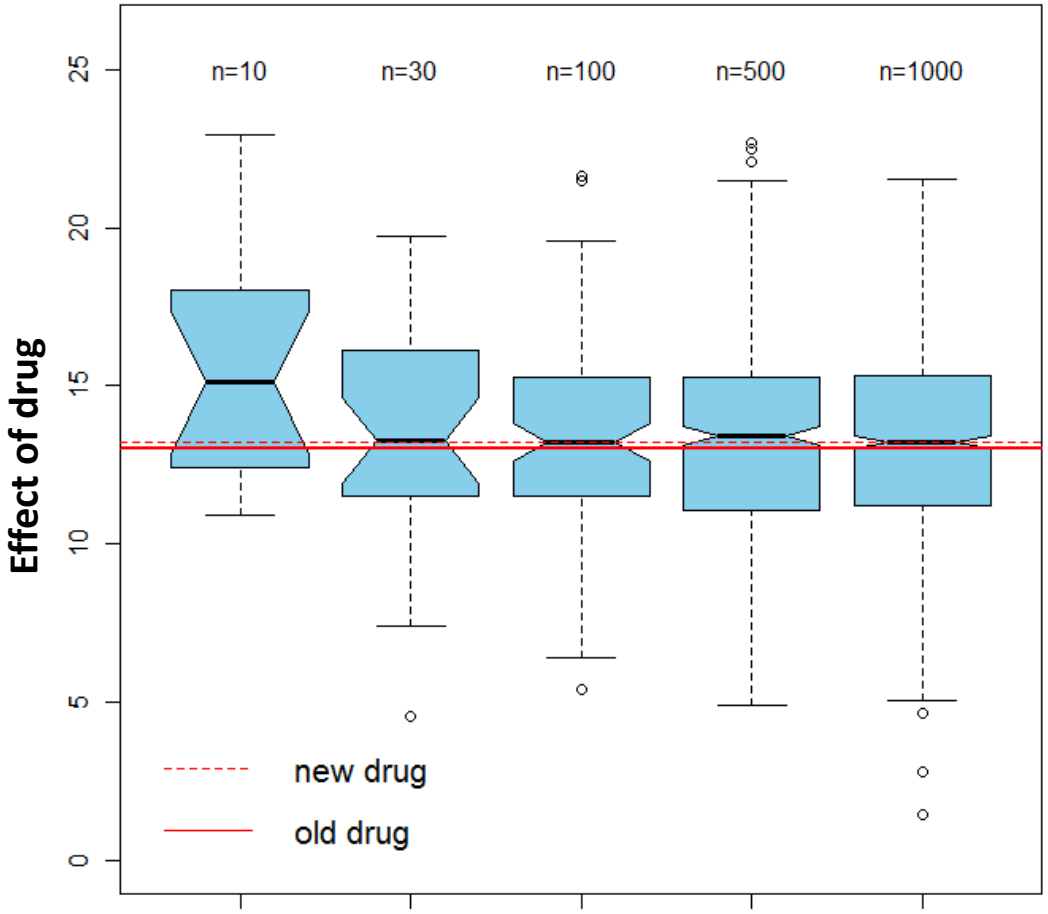



+/-1.58 IQR/sqrt(n)

The notch covers the population median «quite certain»

# Significance does not imply relevance
# Everything gets significant if the sample size is large enough

Sample with different sample sizes drawn from a Normal distribution with expected value of 13.1




$H_0$: $\mu_0 = \mu_{old\text{-}drug} = 13$
$H_A$: $\mu > 13$

Assume true median of the new drug is13.1 which would be no relevant improvement compared to old drug value 13

With increasing sample size the 95% confidence interval for the true median gets smaller, while $\alpha$ stays the same and the power increases to find a significant difference to the old mean of 13.

- To ensure relevance of an significant test one should formulate a relevant $H_A$.
- Non-significance could be caused by a too small sample.