

Biostatistics Week 2

Topics this week:

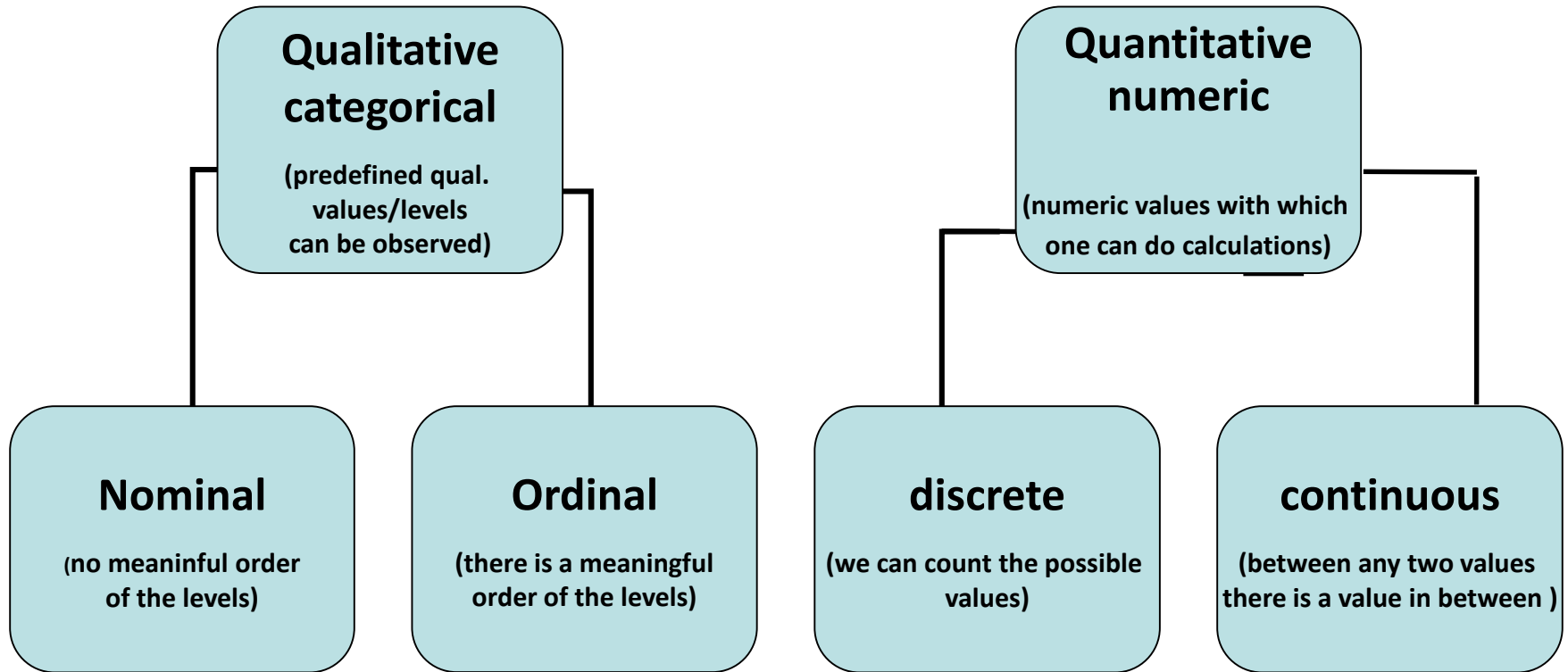
➤ uni-variate descriptive Analysis

- **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

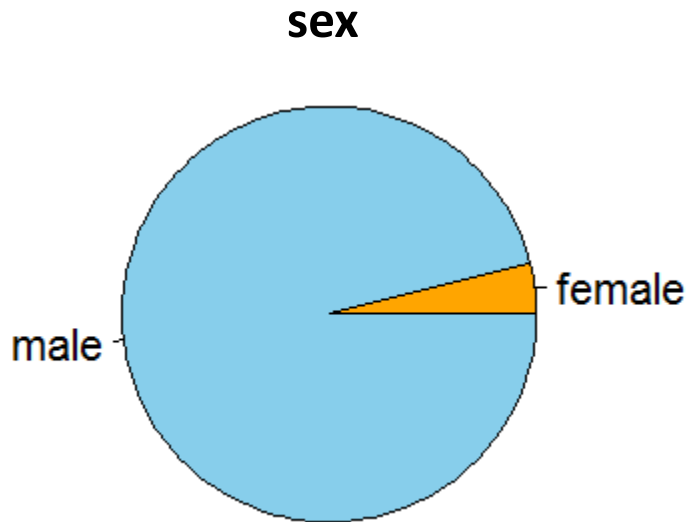
- **Continuous vs continuous: scatterplot**
- **Categorical vs categorical: mosaicplot**
- **Continuous vs categorical: grouped boxplots or stripcharts**

There are different types of data



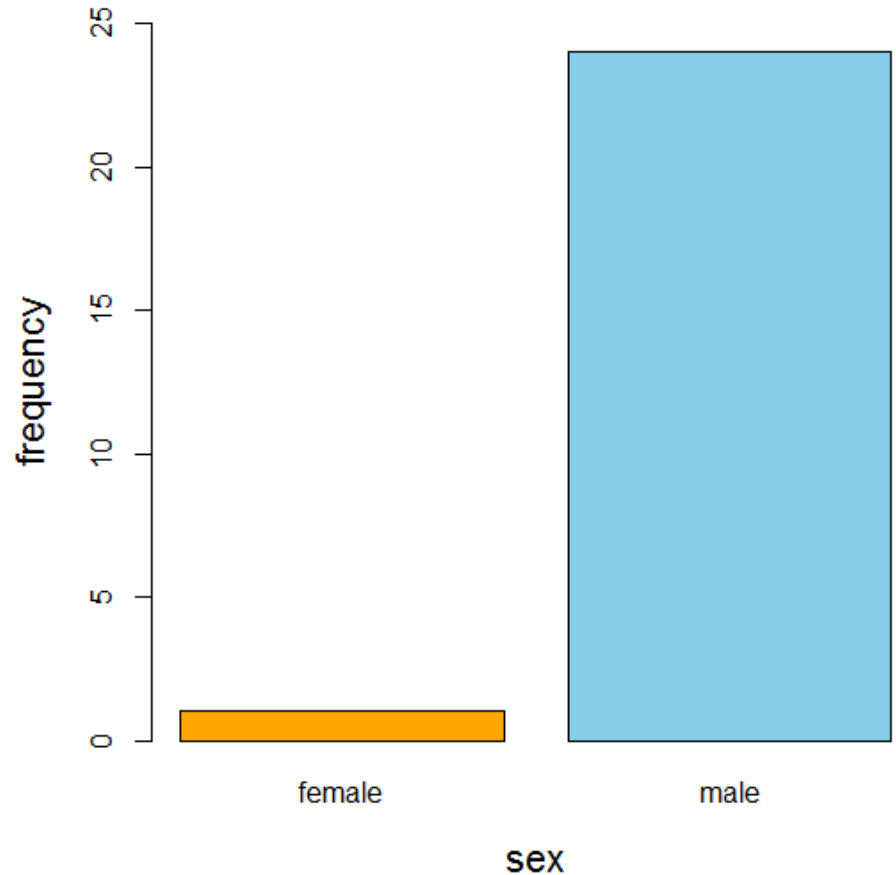
How to summarize categorical data?

pie chart



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

bar chart

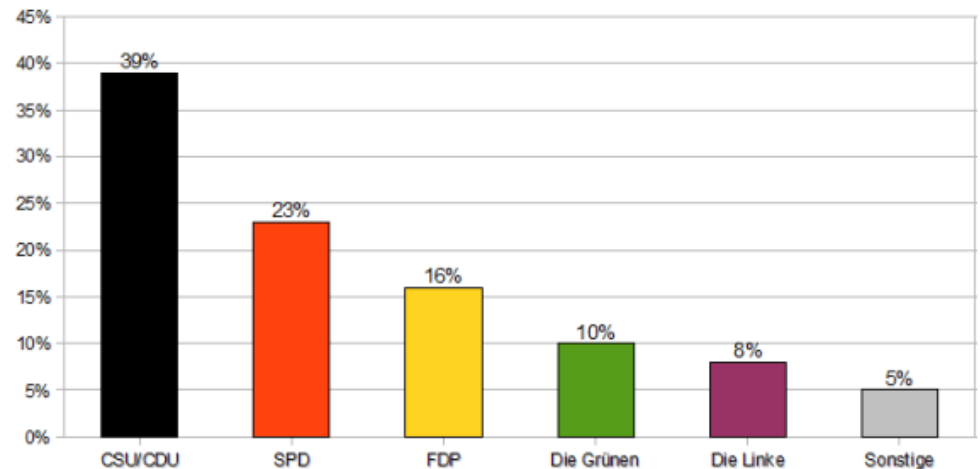
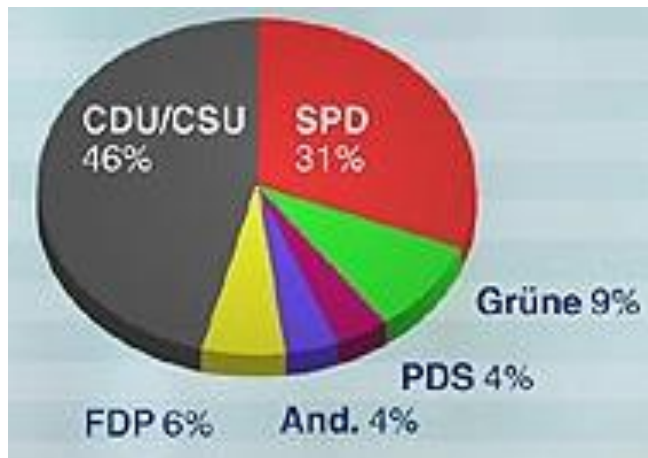


```
barplot(table(dat$sex))
```

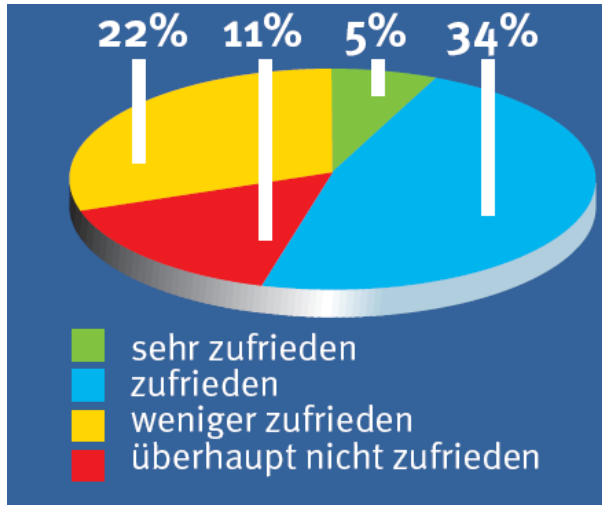
Bar charts can transport more information than pie charts.

Visualizing categorical data by Bar-Chart or Pie-Chart

These charts are simple - is there room for manipulation?

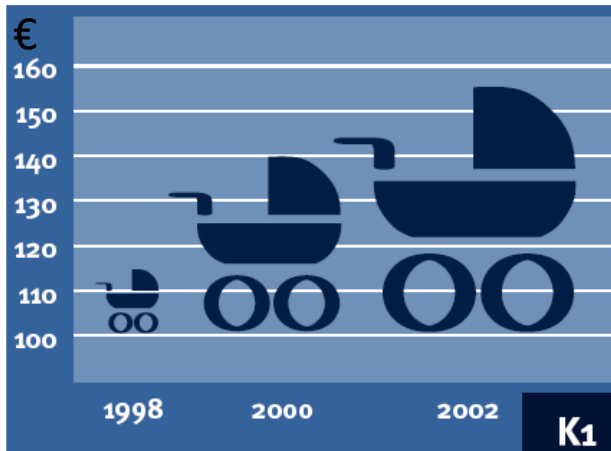


Half of all reader are satisfied with Klinsmann - true?



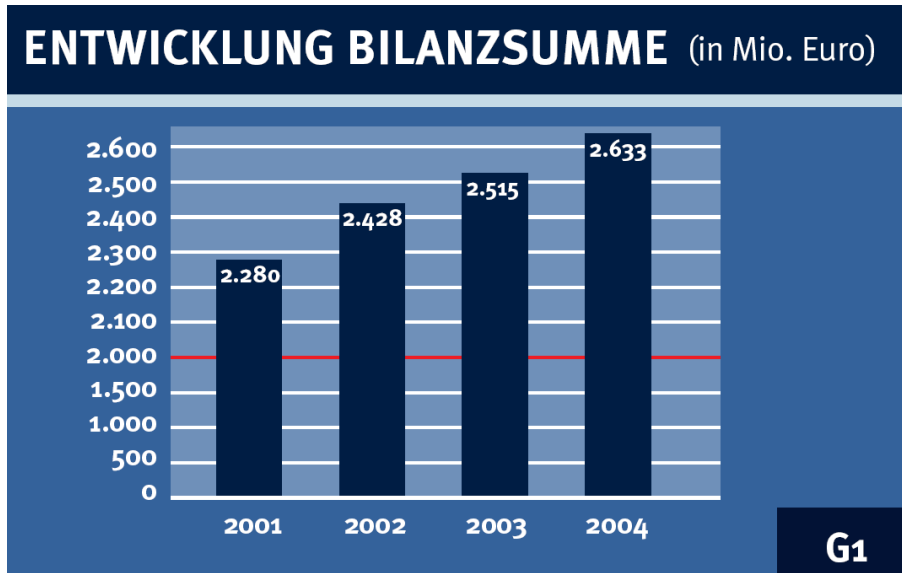
Pie-Chart from the German newspaper „Bild“ .
Reader were asked how satisfied they are with
soccer trainer Klinsmann.

Generous increase of child allowance - true?



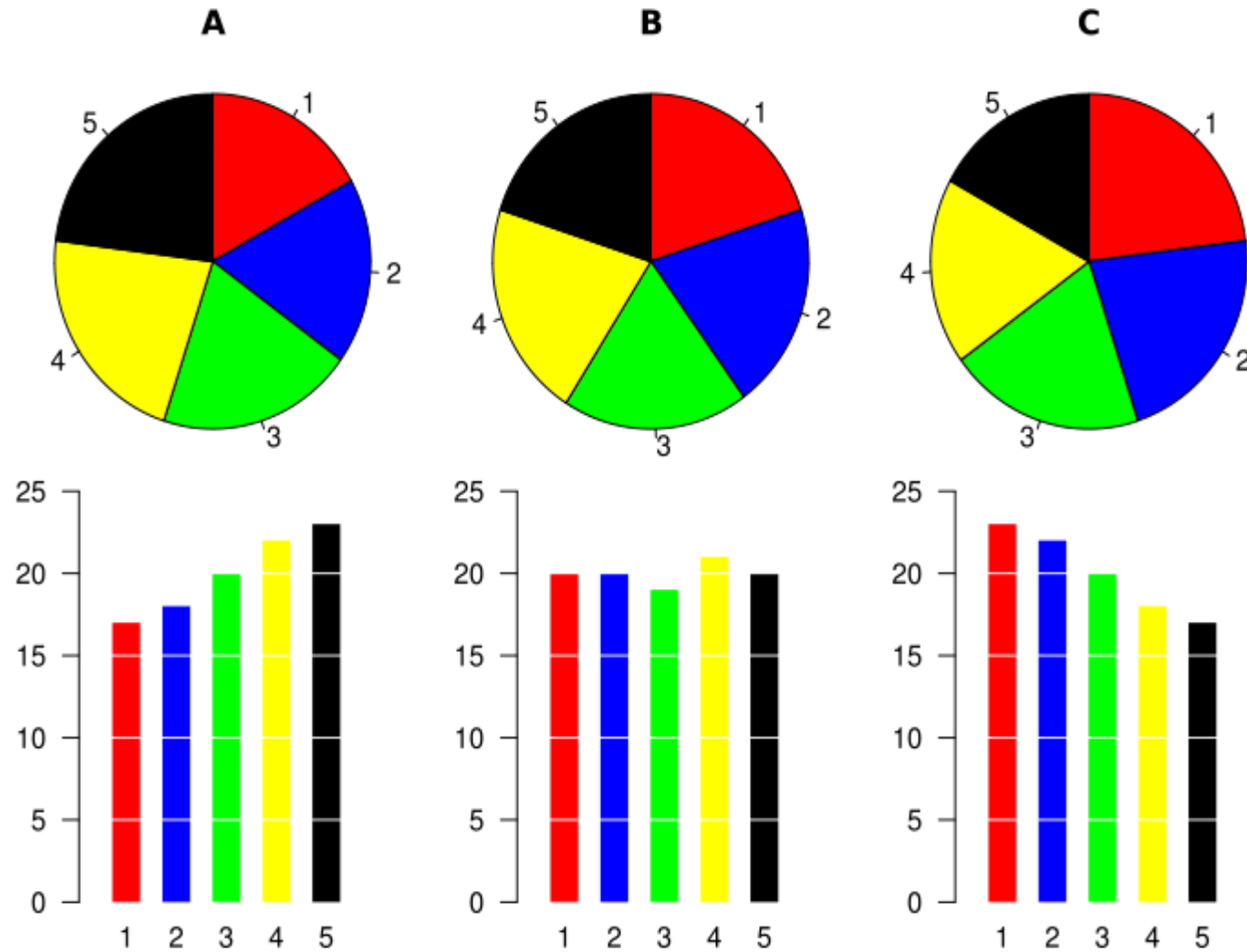
Graph from government statement in the German red-green agenda 2010.

Good business development - true?



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

Barplots are often to prefer over pie-charts

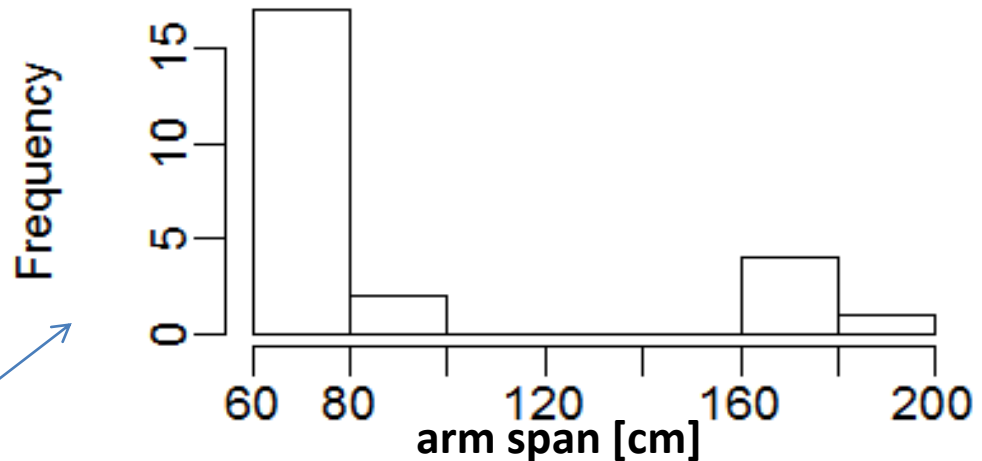


Humans are much better in comparing heights than comparing areas.

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

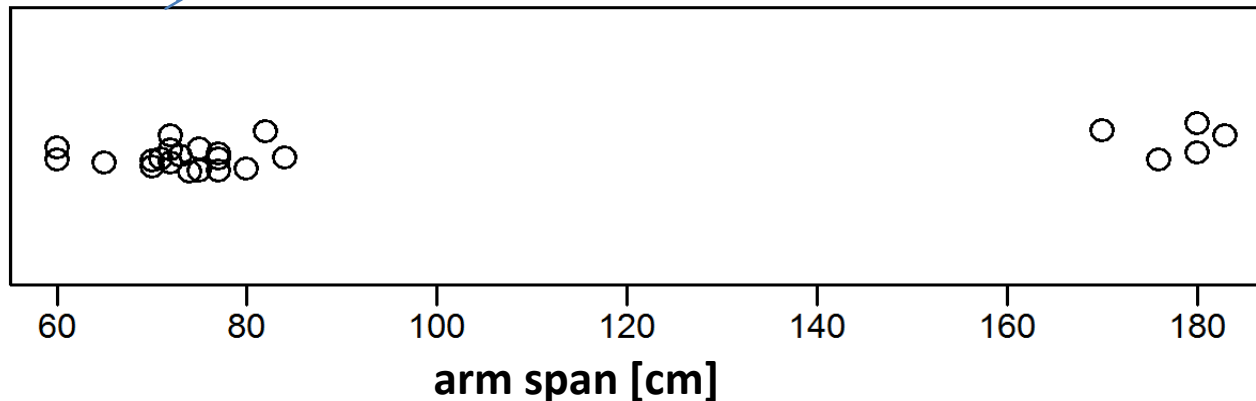
X: arm span	frequency y
[60, 80)	17
[80,100)	2
[100,120)	0
[120,140)	0
[140,160)	0
[160,180)	4
[180,200)	1

- 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)`

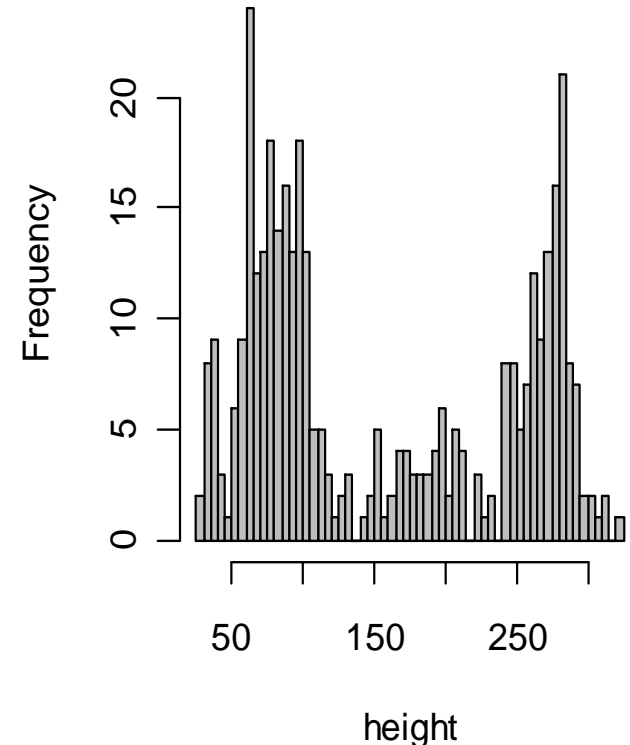
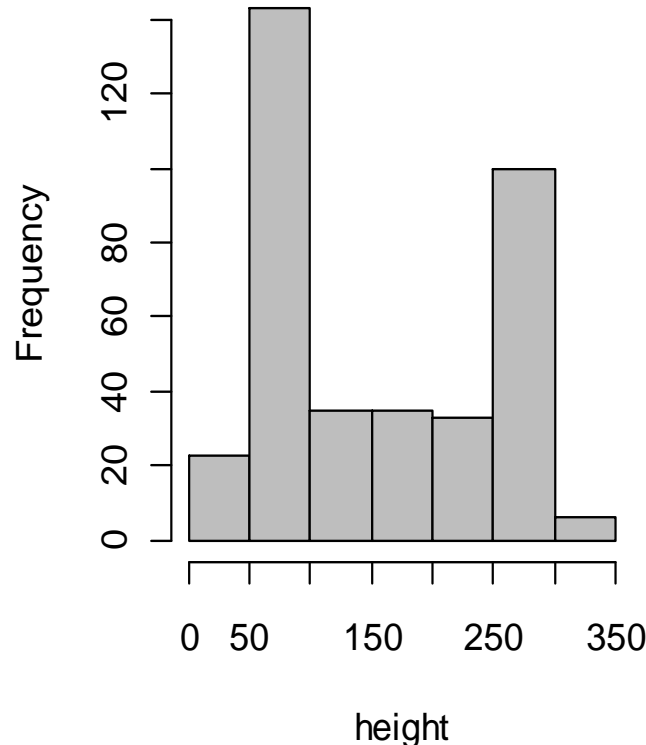
G
height
57.9
62.1
55.8
61.5
68
52.8
70.5
60.4
75.2
77.1
70.4
70.1
27.6
35
⋮

`# hist, few classes, big bin width`

`hist(dat$height, nclass=7)`

`# hist, many classes, small bin width`

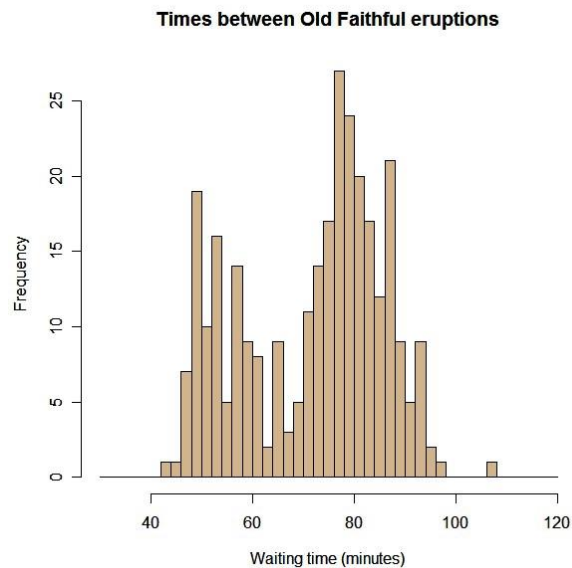
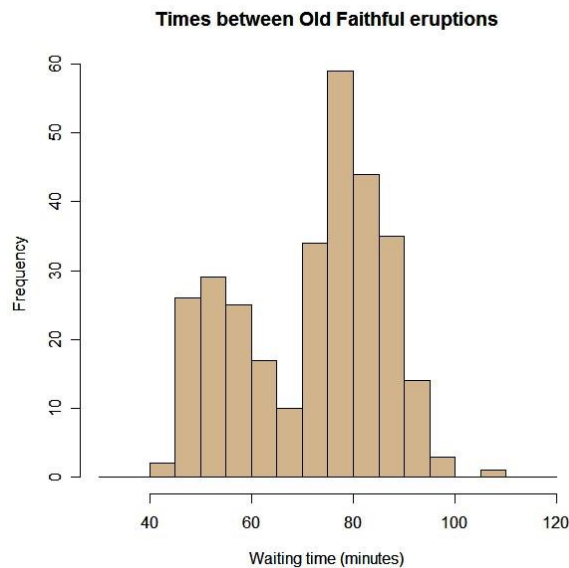
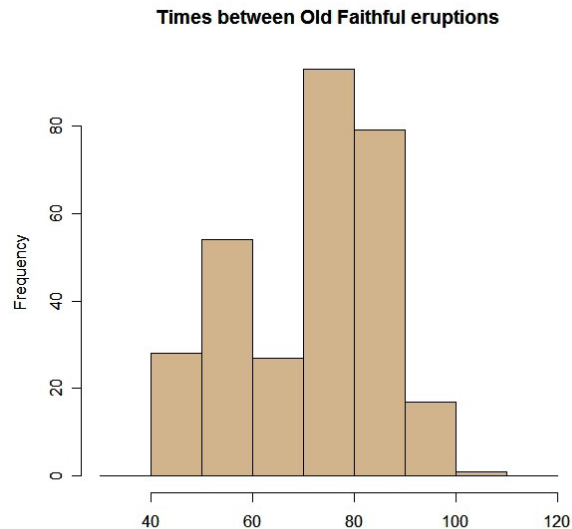
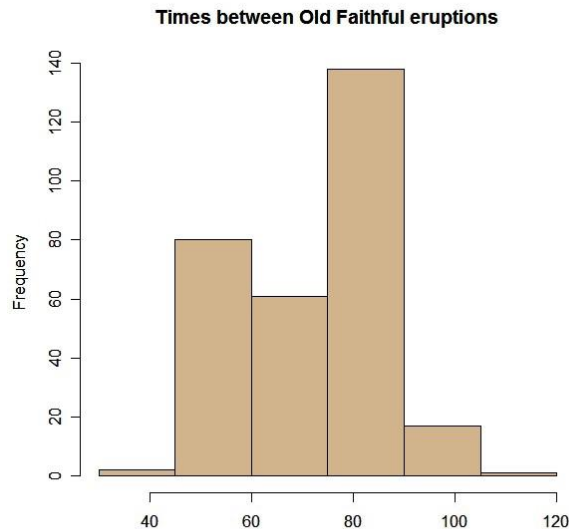
`hist(dat$height, nclass=100)`



Are there subgroups? If yes – how many?

How reliable is the height of a bar?

How many classes do we need?



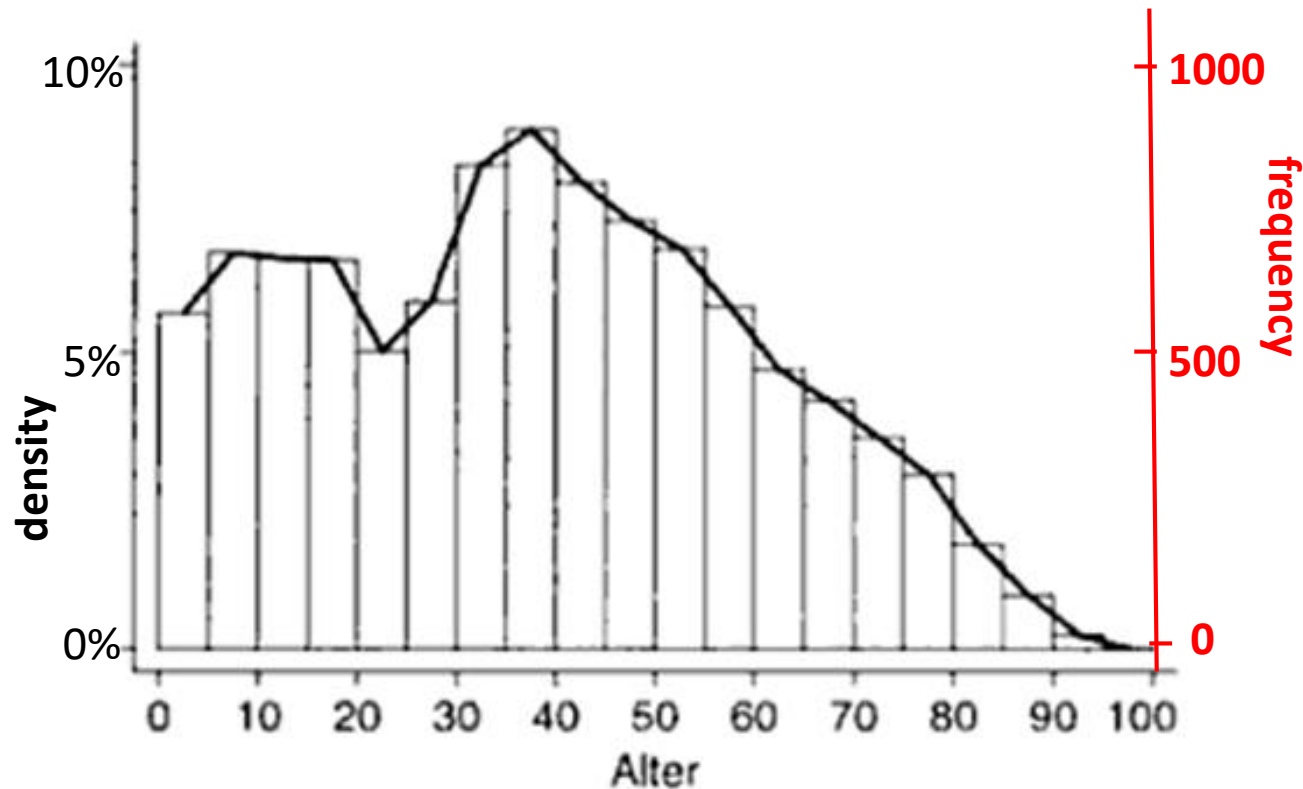
<http://www.amstat.org/publications/jse/v6n3/applets/Histogram.html>



299 eruption intervals were observed

Shape of the histogram may depend on the class choices

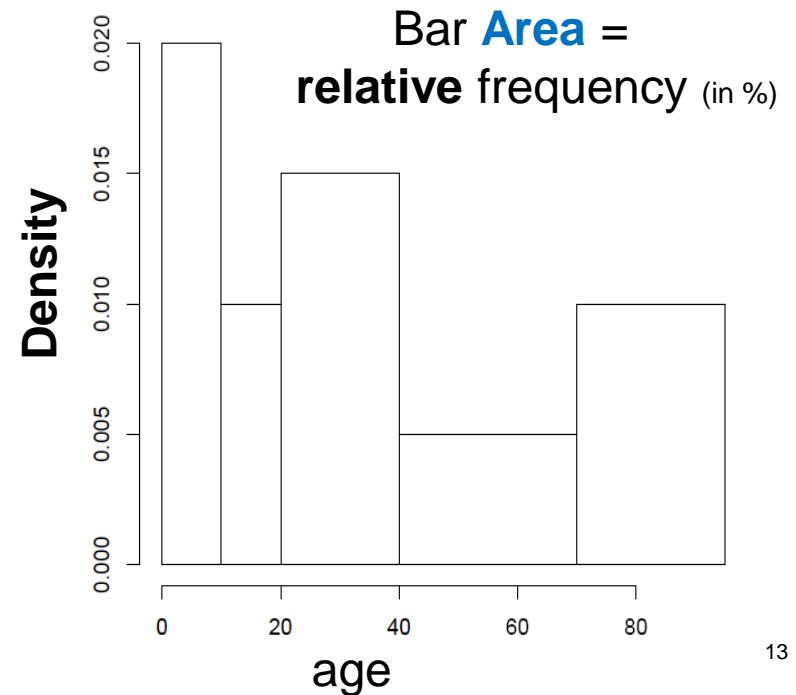
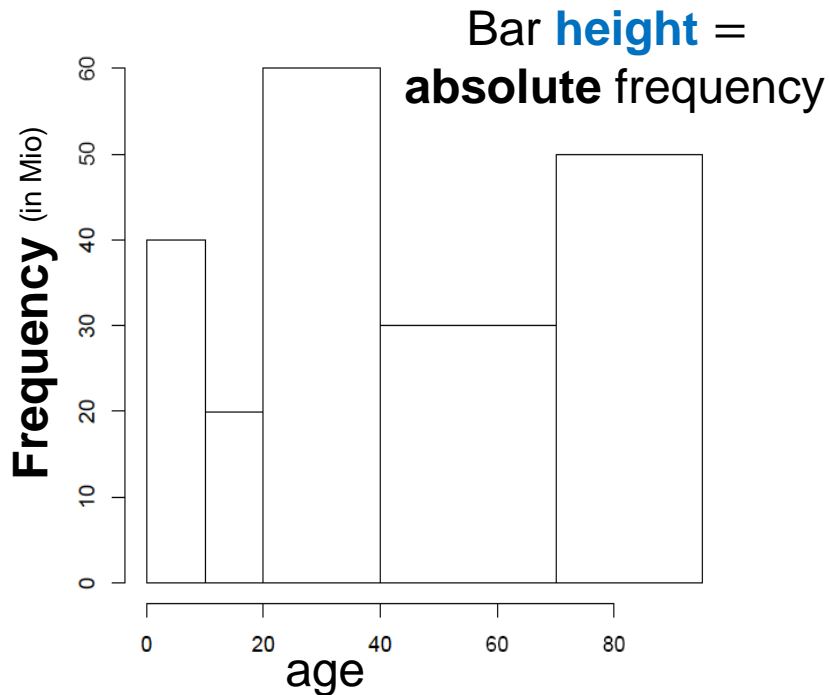
How does the distribution of age look like



Only 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%

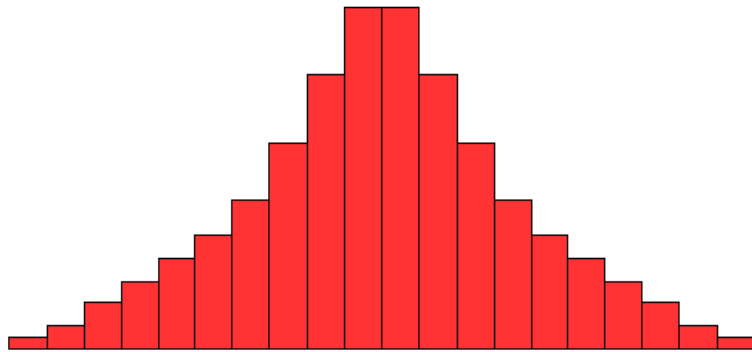


Rules for histograms

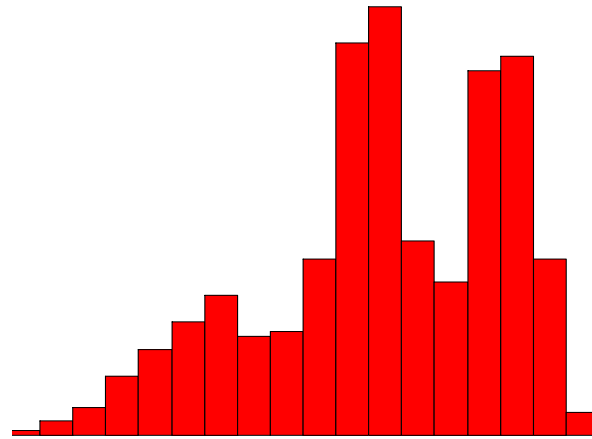
- Avoid classes with different width! (shape will change)
- How many classes: \sqrt{n} classes for n observations.
- The shape can depend on the number classes and the class limits.

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.

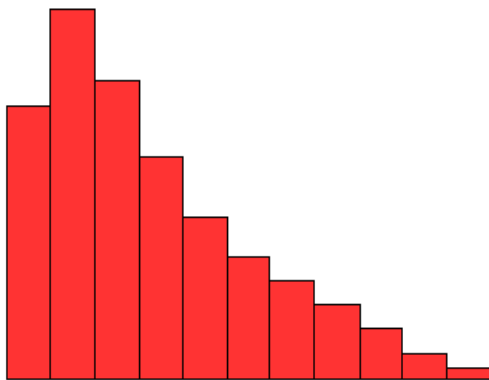
Shapes of distributions



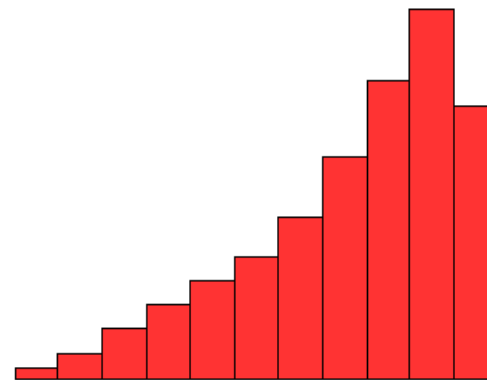
Symmetric, uni-modale



Multi-modale, slightly left-skewed



Right-skewed, uni modale



Left-skewed, uni-modale

Measures for the location and variation

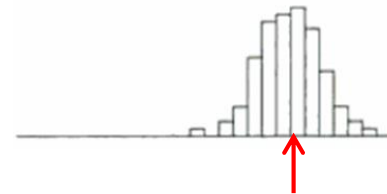
Data can be summarized by summary statistics. Most important key figure describe the center and the width of a distribution..

Measures for the location

Where is the center?

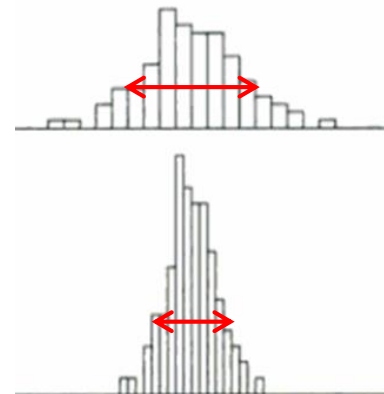


What is a typical value?



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](#)

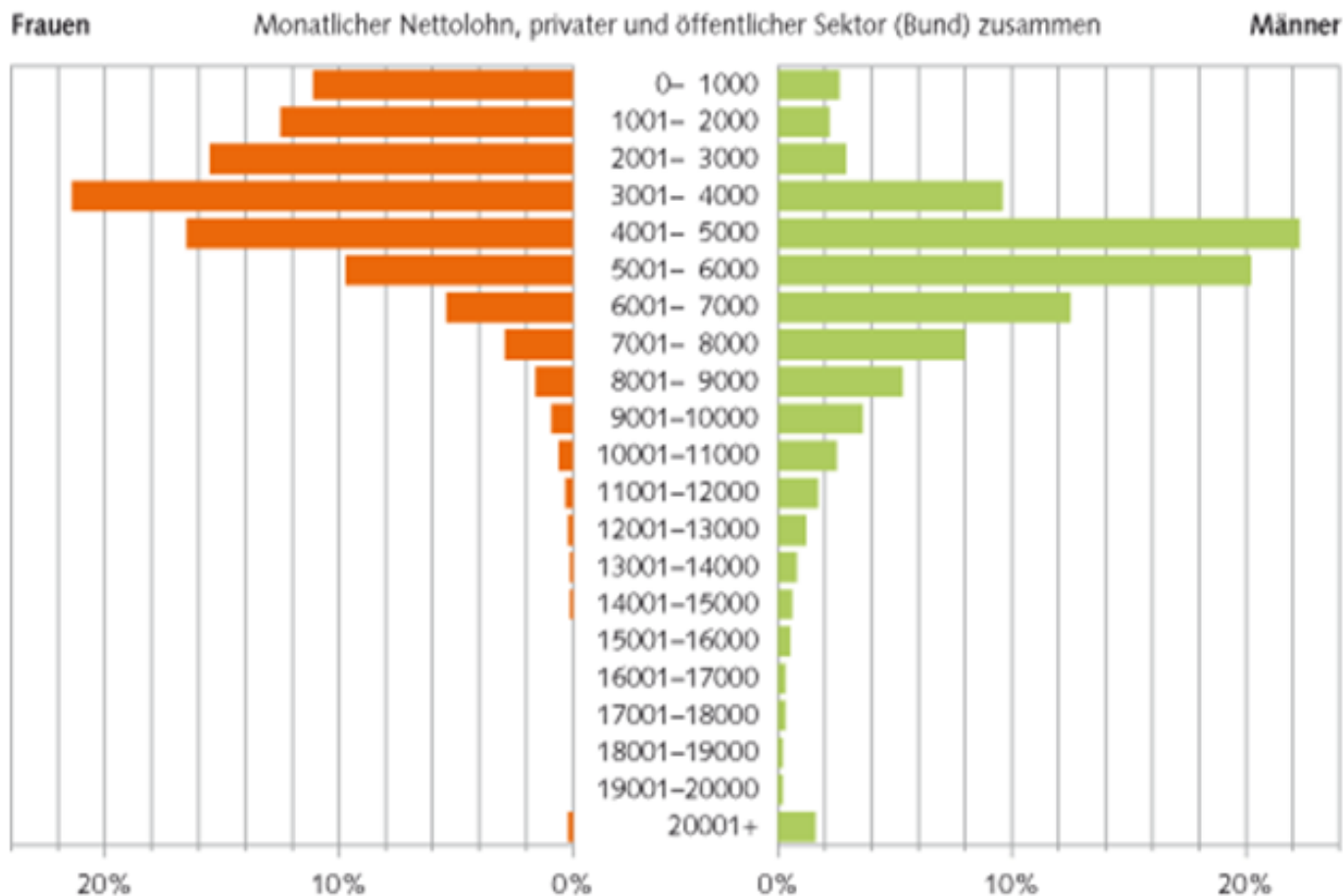
Novartis beschäftigt weltweit zurzeit rund 99.800 Mitarbeitende. Davon arbeiten rund 12.000 in der Schweiz – verteilt auf die acht Standorte in Basel BS/BL, Stein AG, Embrach ZH, Cham ZG, Bern BE, St-Aubin FR, Nyon VD und Locarno TI. Eine kürzlich veröffentlichte Studie hat ergeben, dass für jeden direkten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



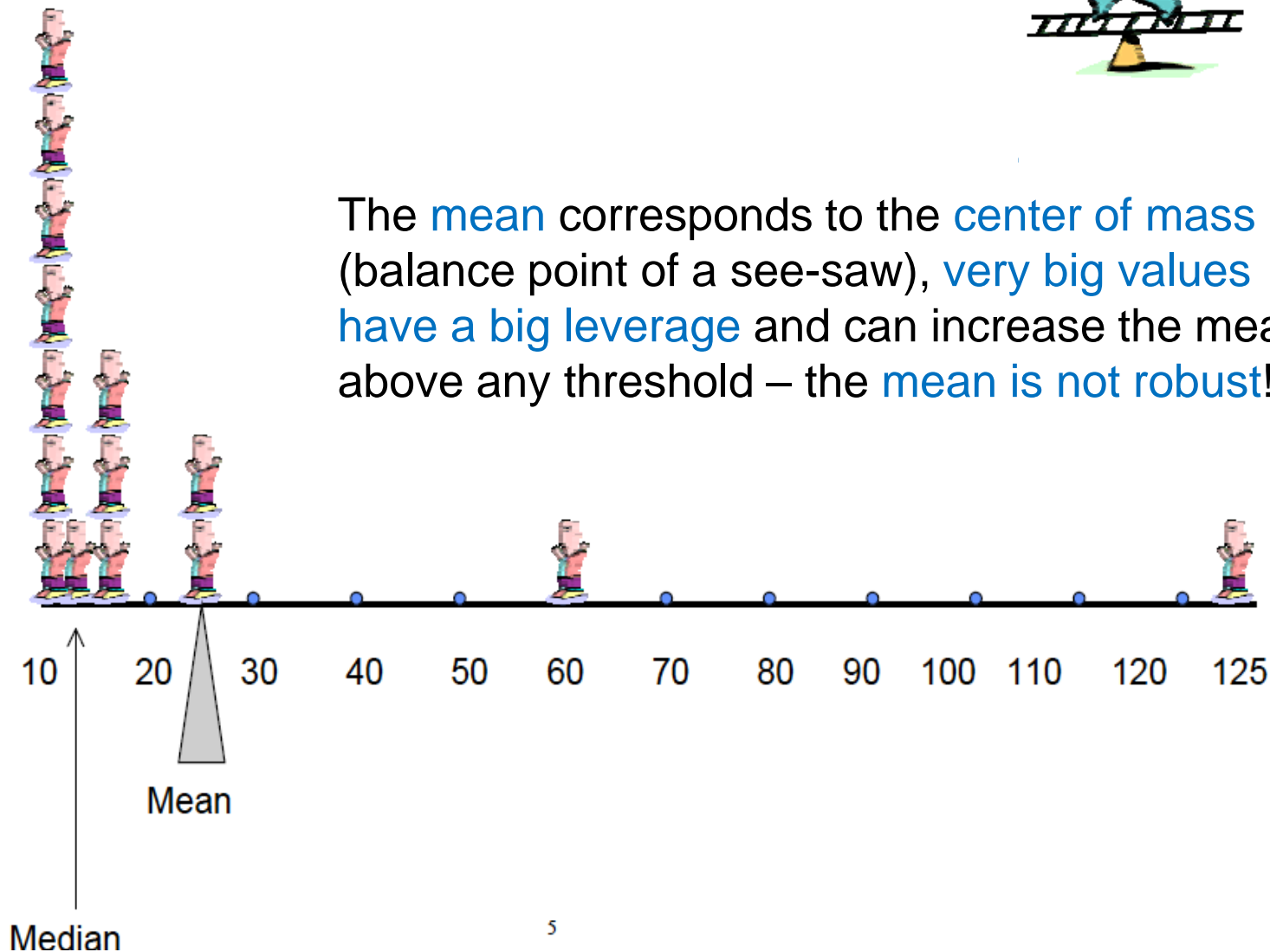
Quelle: Schweizerische Lohnstrukturerhebung

© 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:

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

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

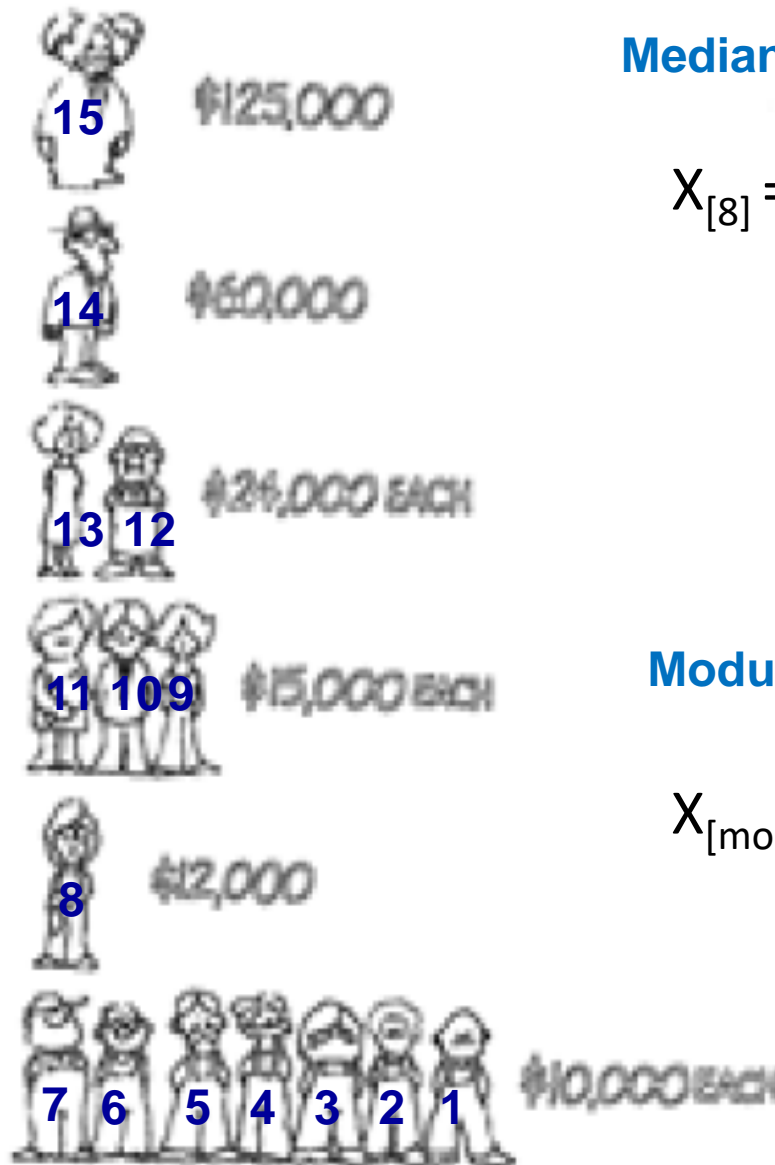
What is „the mean“ income in this company?

Mean:

Sum:

$$\begin{aligned}
 &125,000 \\
 &+ 60,000 \\
 &+ 2 \times 24,000 \\
 &+ 3 \times 15,000 \\
 &+ 12,000 \\
 &+ 7 \times 10,000 \\
 \hline
 &= 360,000
 \end{aligned}$$

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



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 \leq the median-value.

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

- **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.

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:

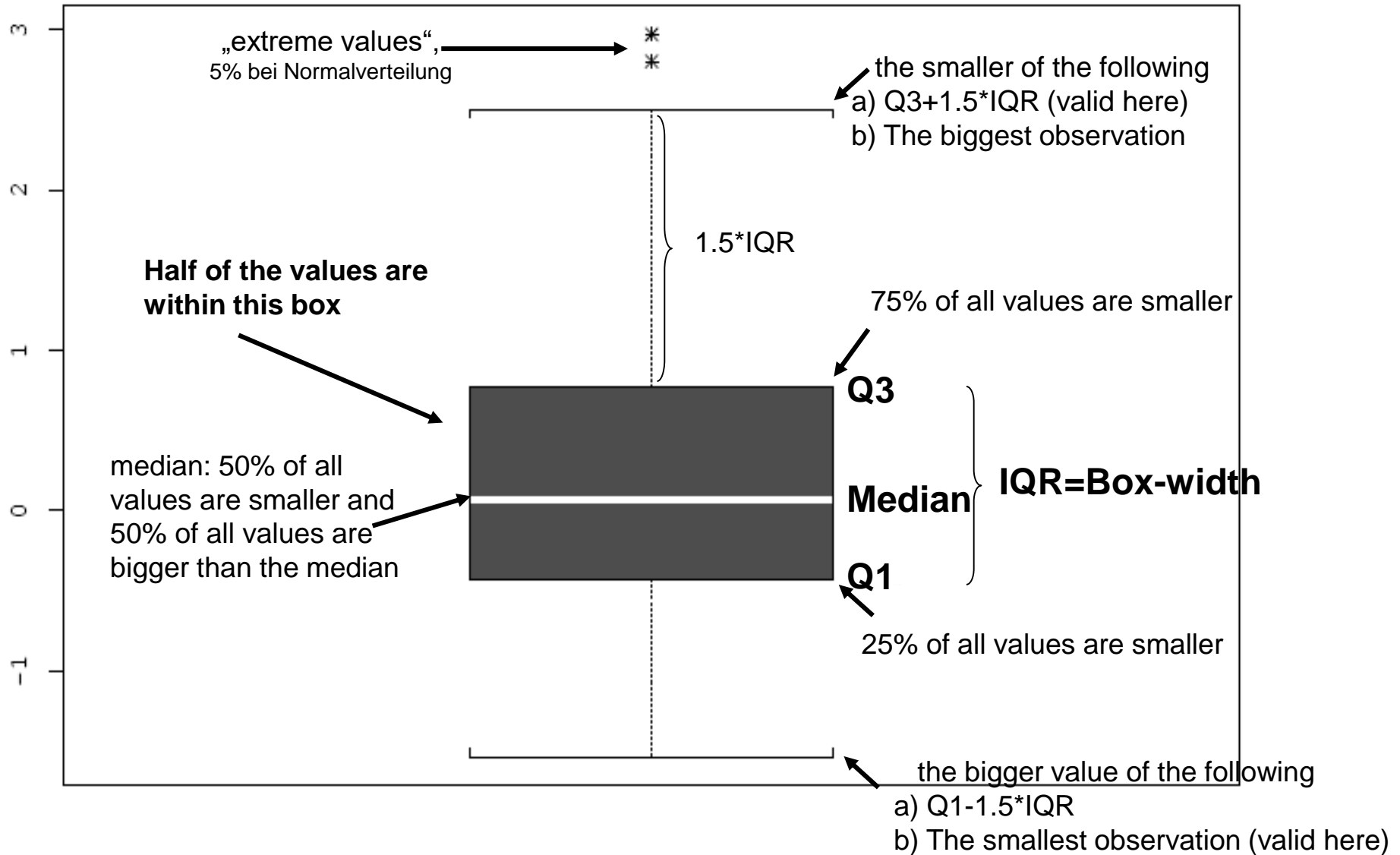
Sample: 3, 4, 7, 5, 0.5, 6 : $n=6$; ordered sample: 0.5, 3, 4, 5, 6, 7

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:

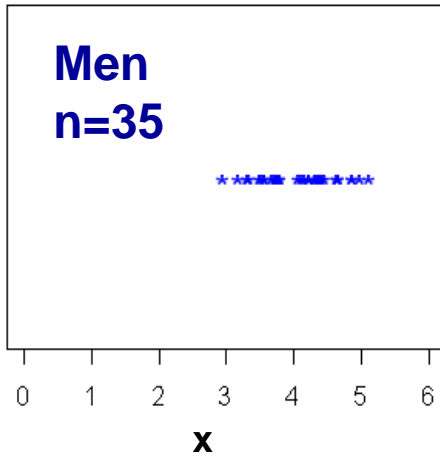
Median= $0.5 \cdot (x_{[3]} + x_{[4]}) = 0.5 \cdot (4 + 5) = 4.5$.

Definition of the Boxplot to visualize continuous data

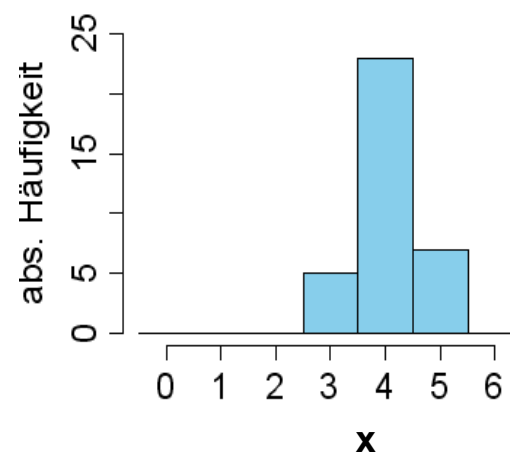


How to best visualize continuous data

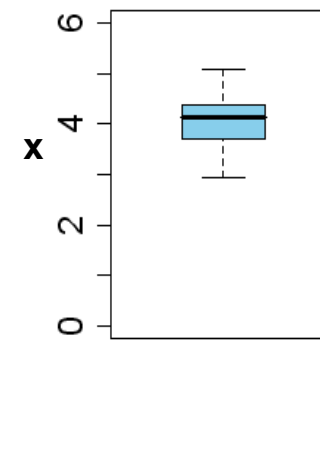
X: reduction of the BMI after take-in of a new drug



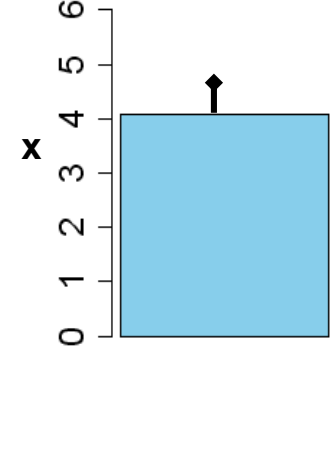
Stripchart



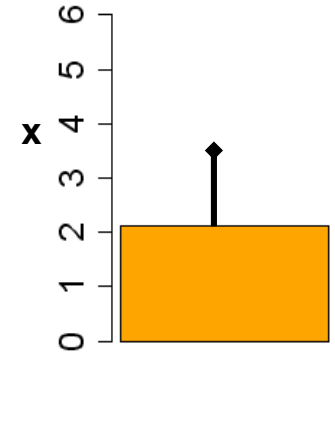
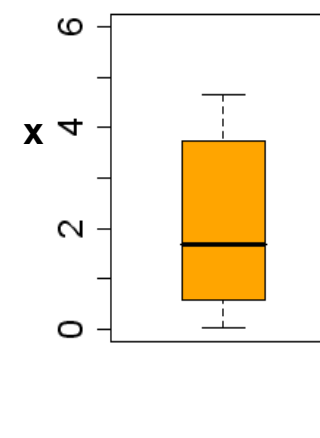
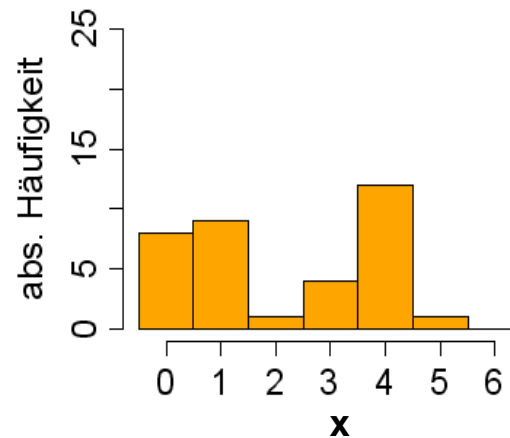
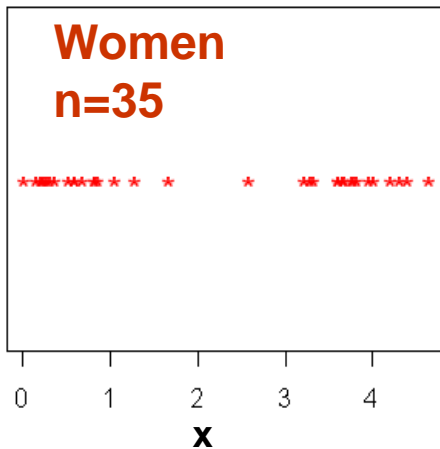
Histogramm



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

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

quantitativ vs. quantitativ:

- Scatterplots, Quantil-Quantil-Plot

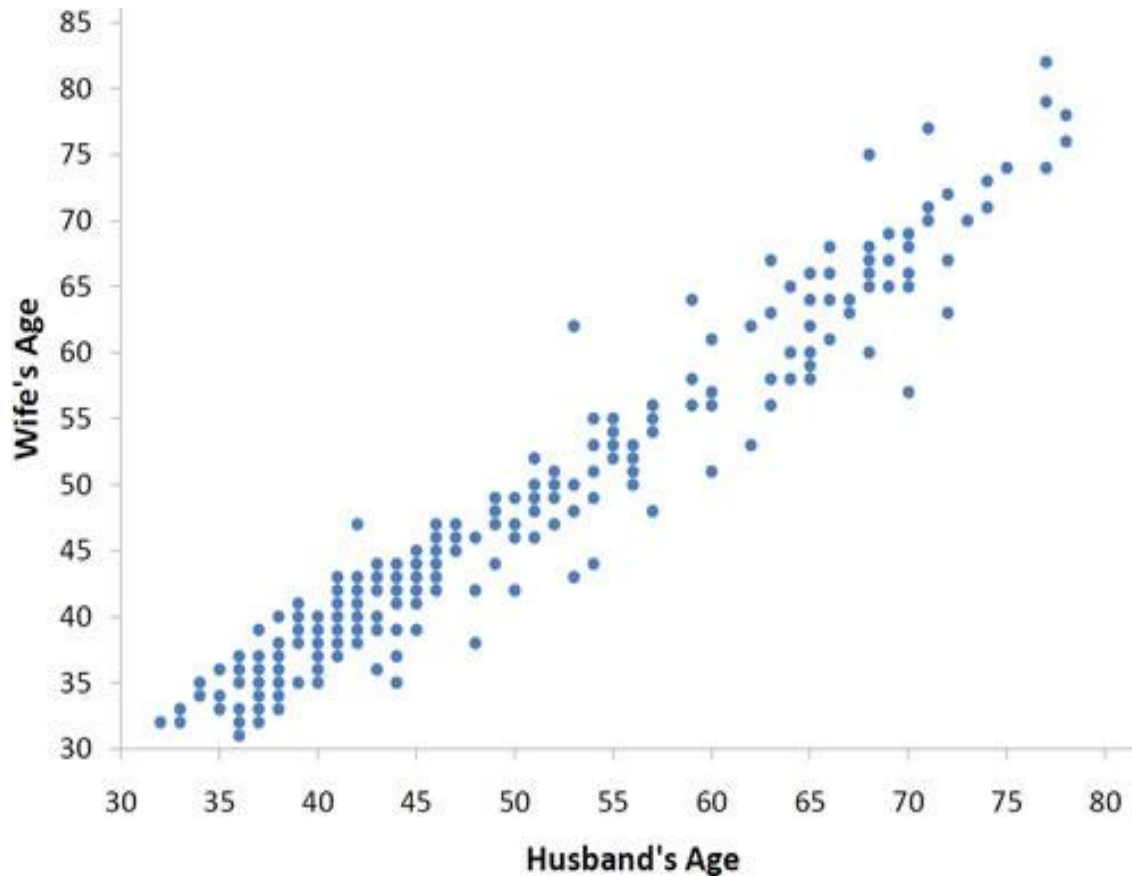
categorical vs. categorical

- mosaicplot

quantitativ vs. categorical:

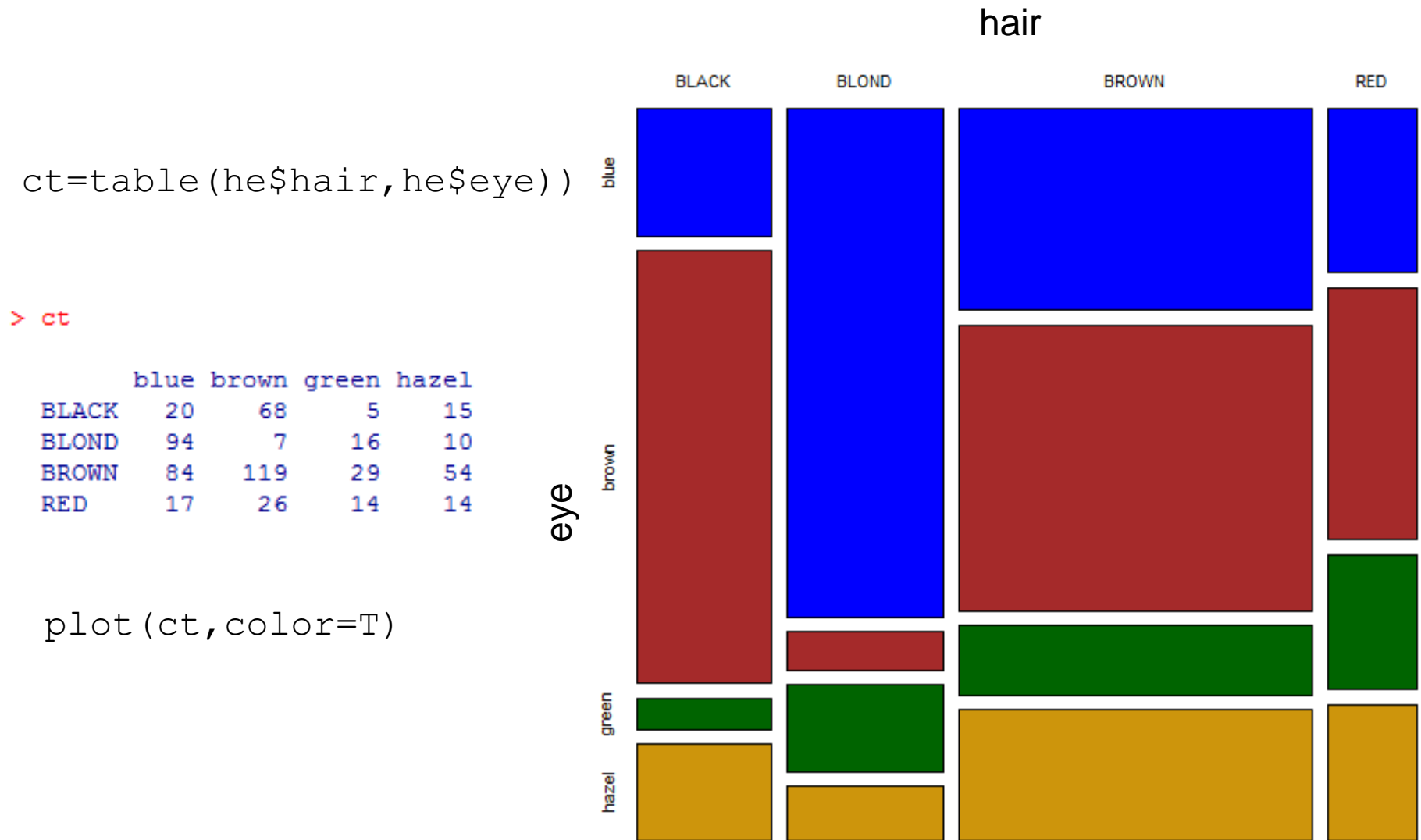
- Grouped boxplots or grouped stripcharts (mean-plots)

Scatterplot for two quantitative variables

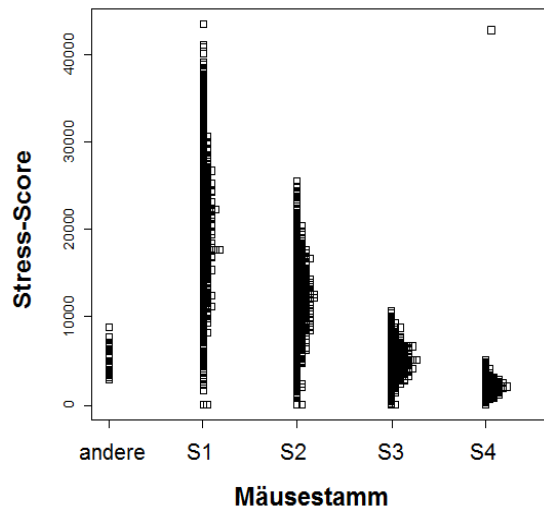


```
plot(dat$m.age, dat$w.age, xlab="Husband's Age", ylab="Wifes's Age" )  
plot( w.age ~ m.age, data=dat)
```

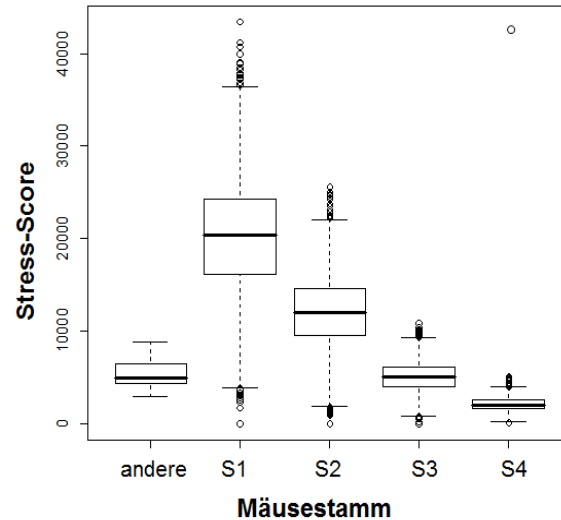
Bi-variate visualization of 2 categorical variables



Bi-variate visualization: continuous vs categorical variable



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

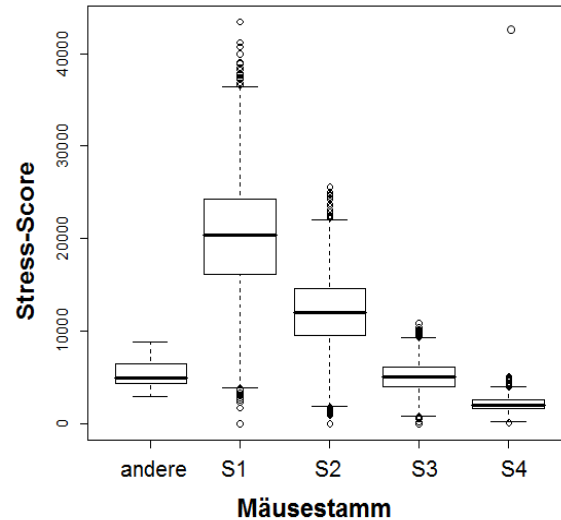
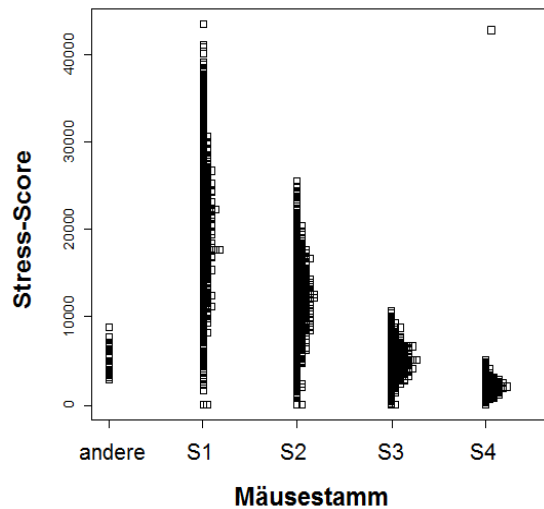


```
boxplot( stress ~ stamm, data=mouse)
```

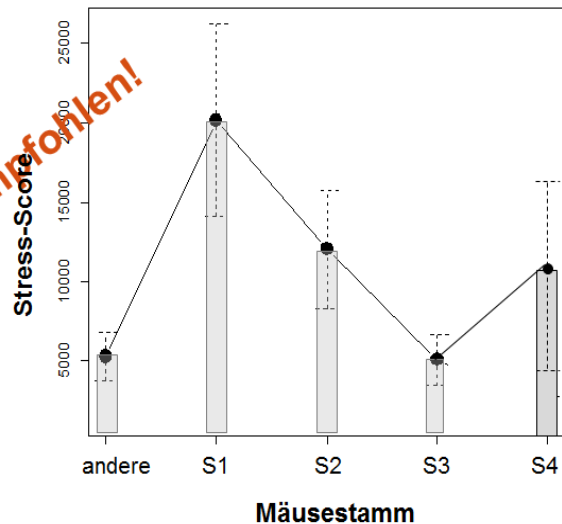
→ Compare distribution of continuous variable in different groups (categories).

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

Bi-variate visualization of 1 categorical and 1 continuous variable



Plot of Means



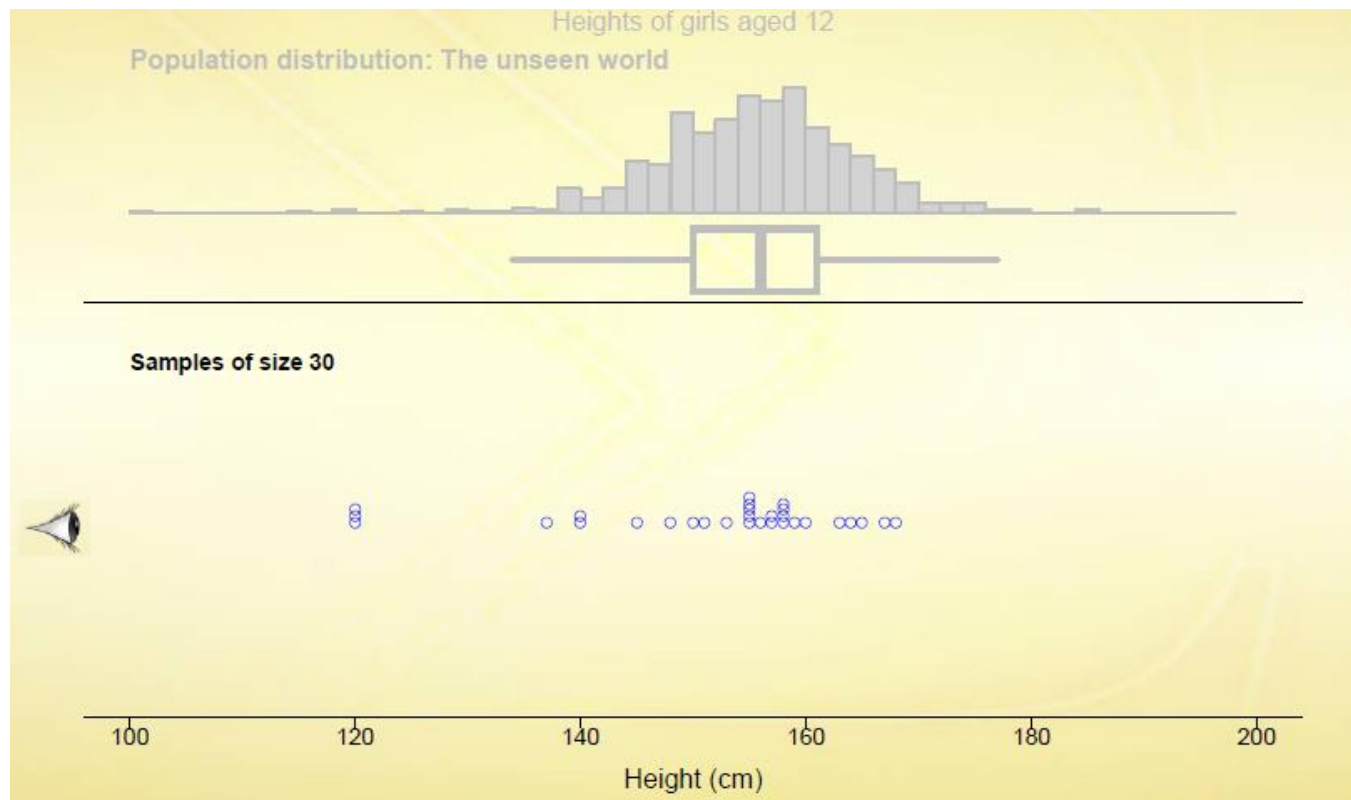
nicht empfohlen!

- => 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!

End week 2?

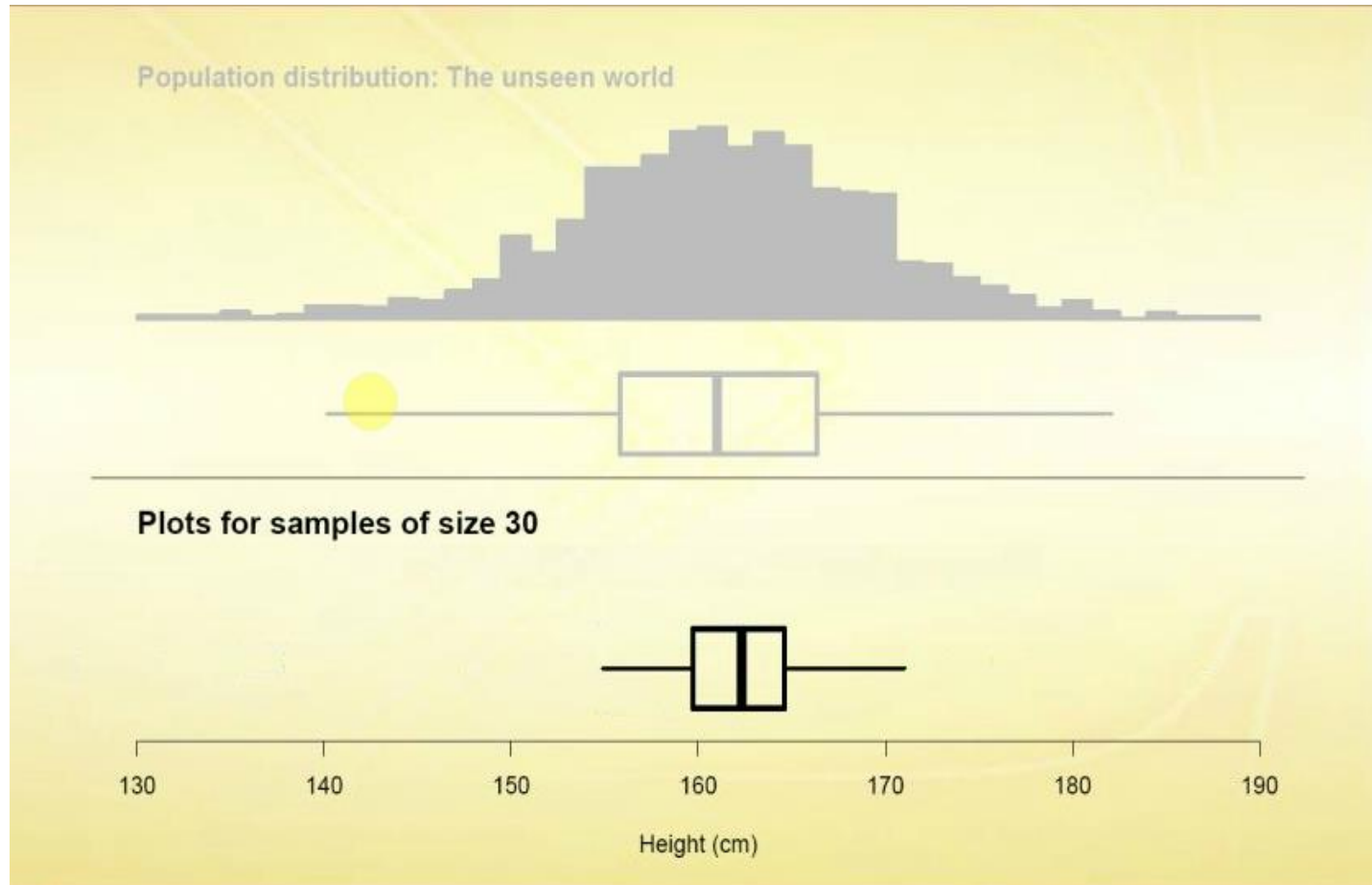
Where is the center of the population?



$n=30$

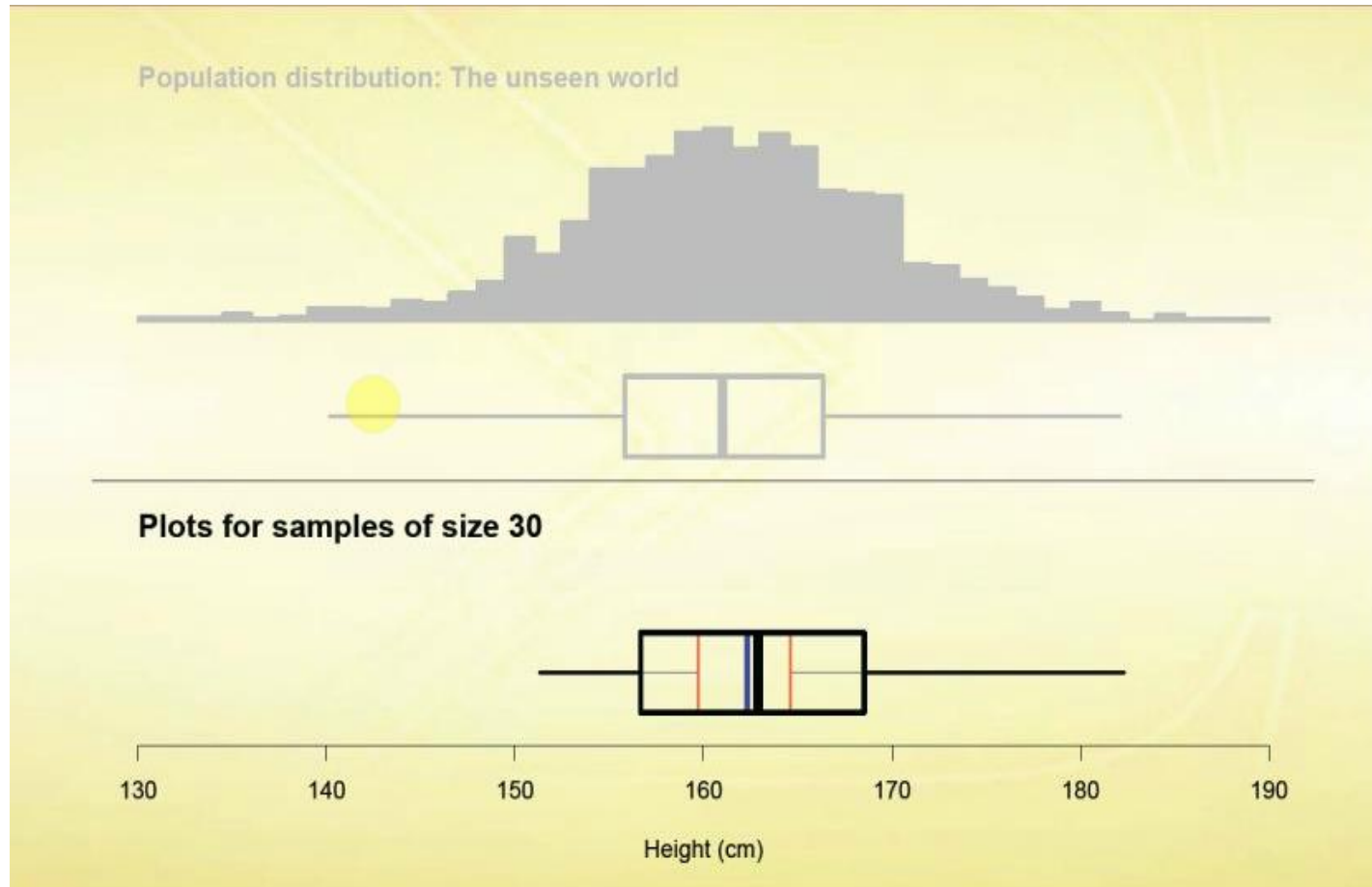
[CtsVar_1samp_Dots30.pdf](#)

Where is the center of the population?



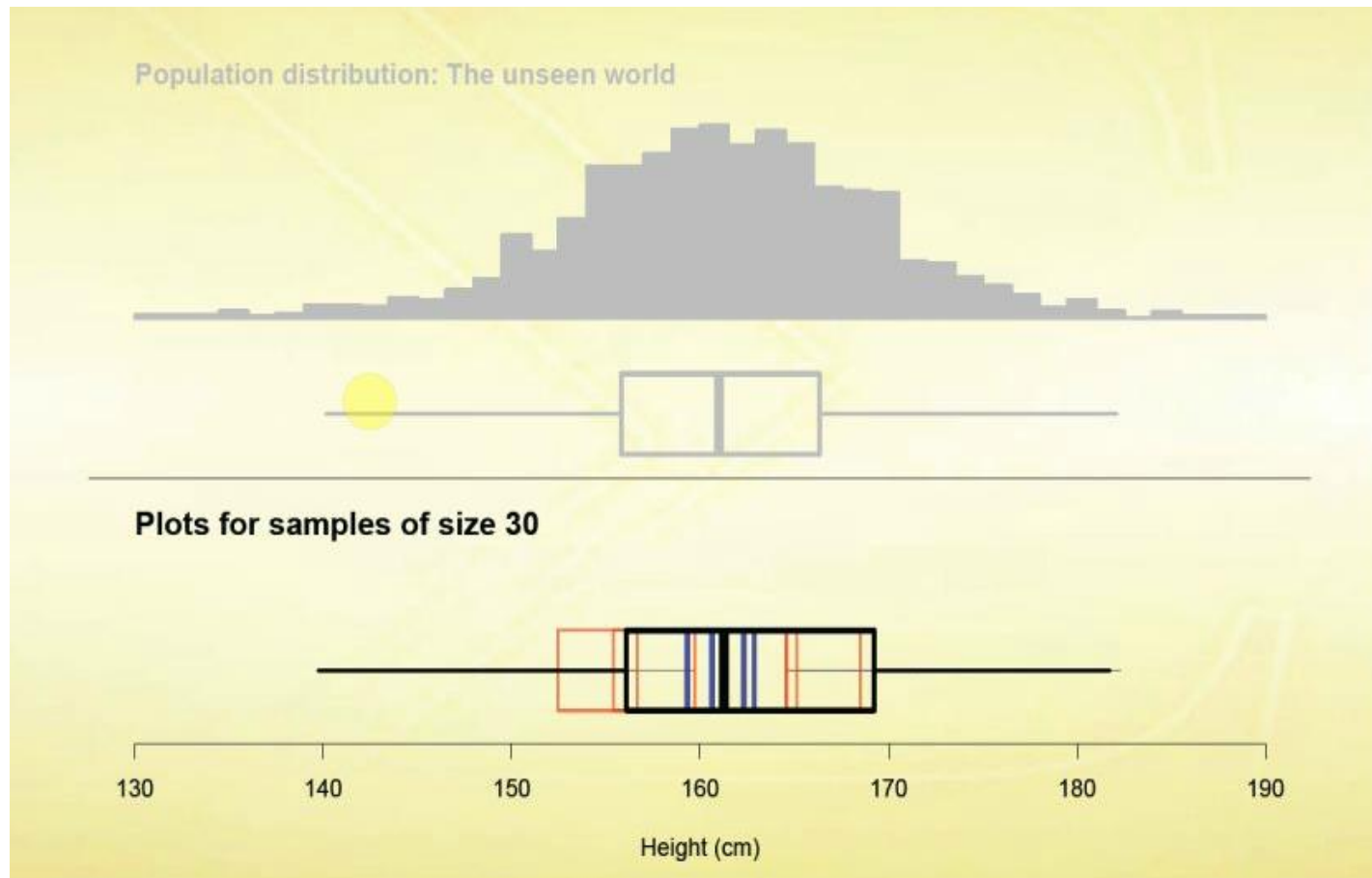
Visualize boxplot with memory

Where is the center of the population?



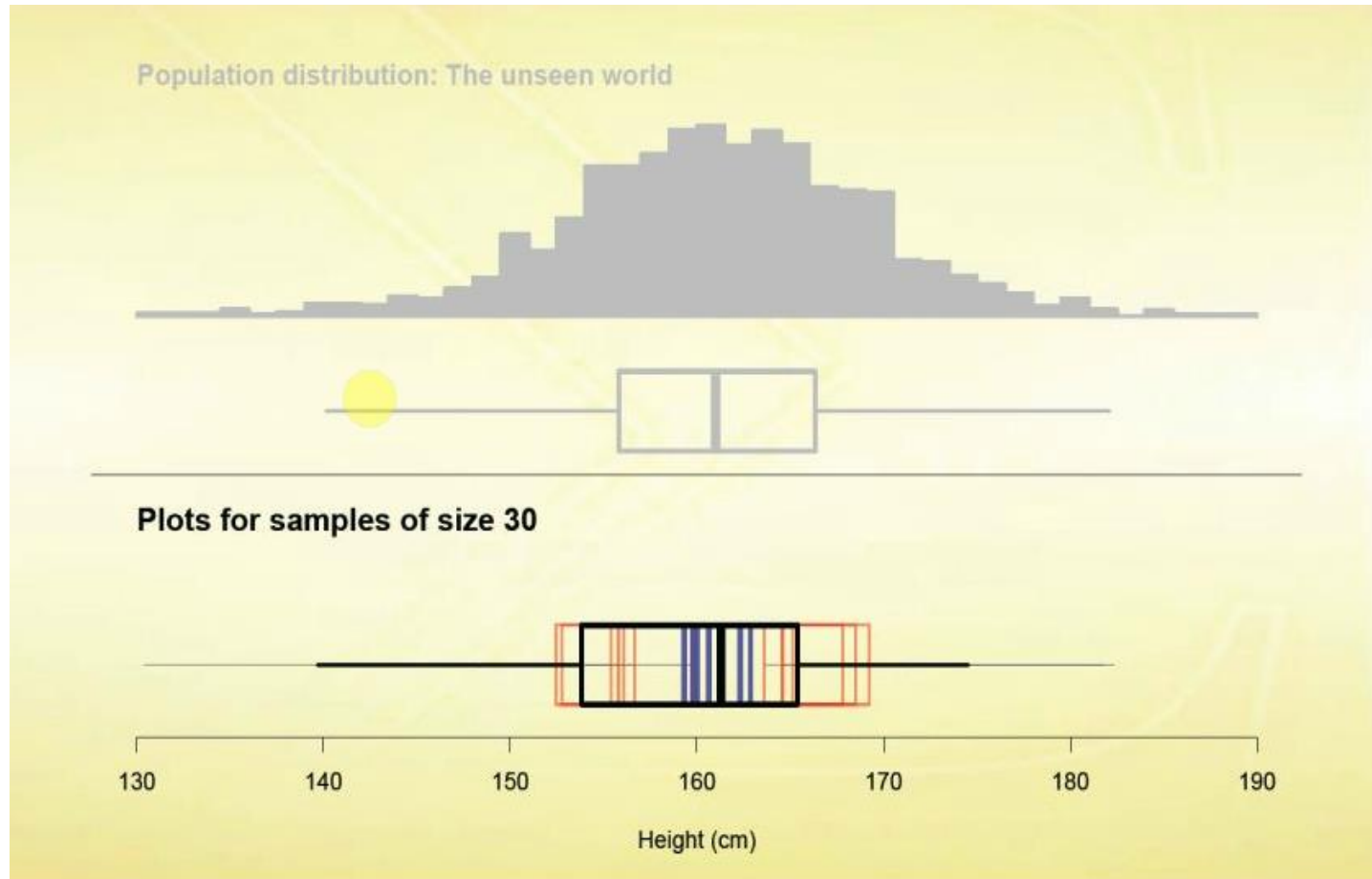
Visualize boxplot with memory

Where is the center of the population?



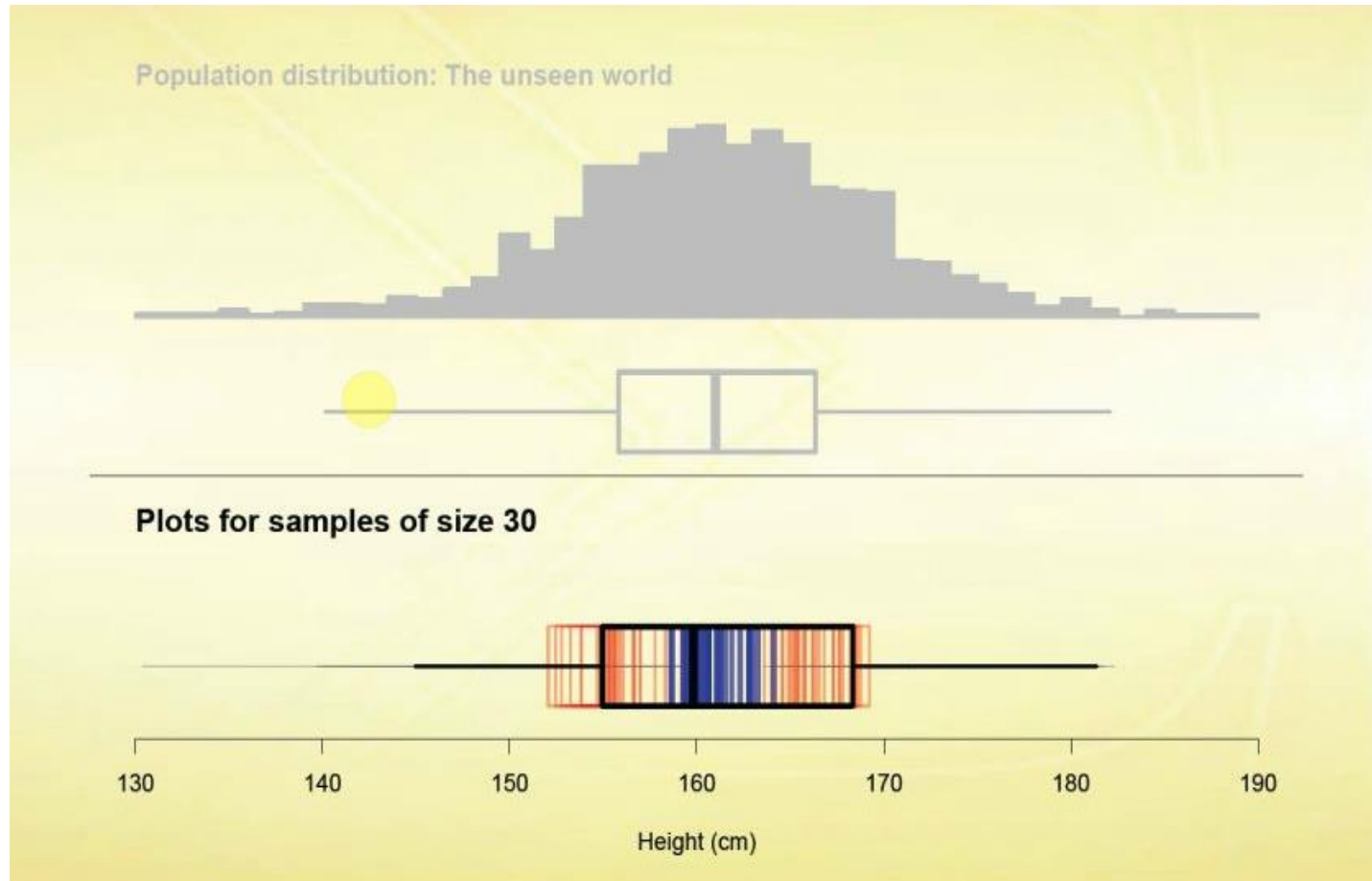
Visualize boxplot with memory

Where is the center of the population?



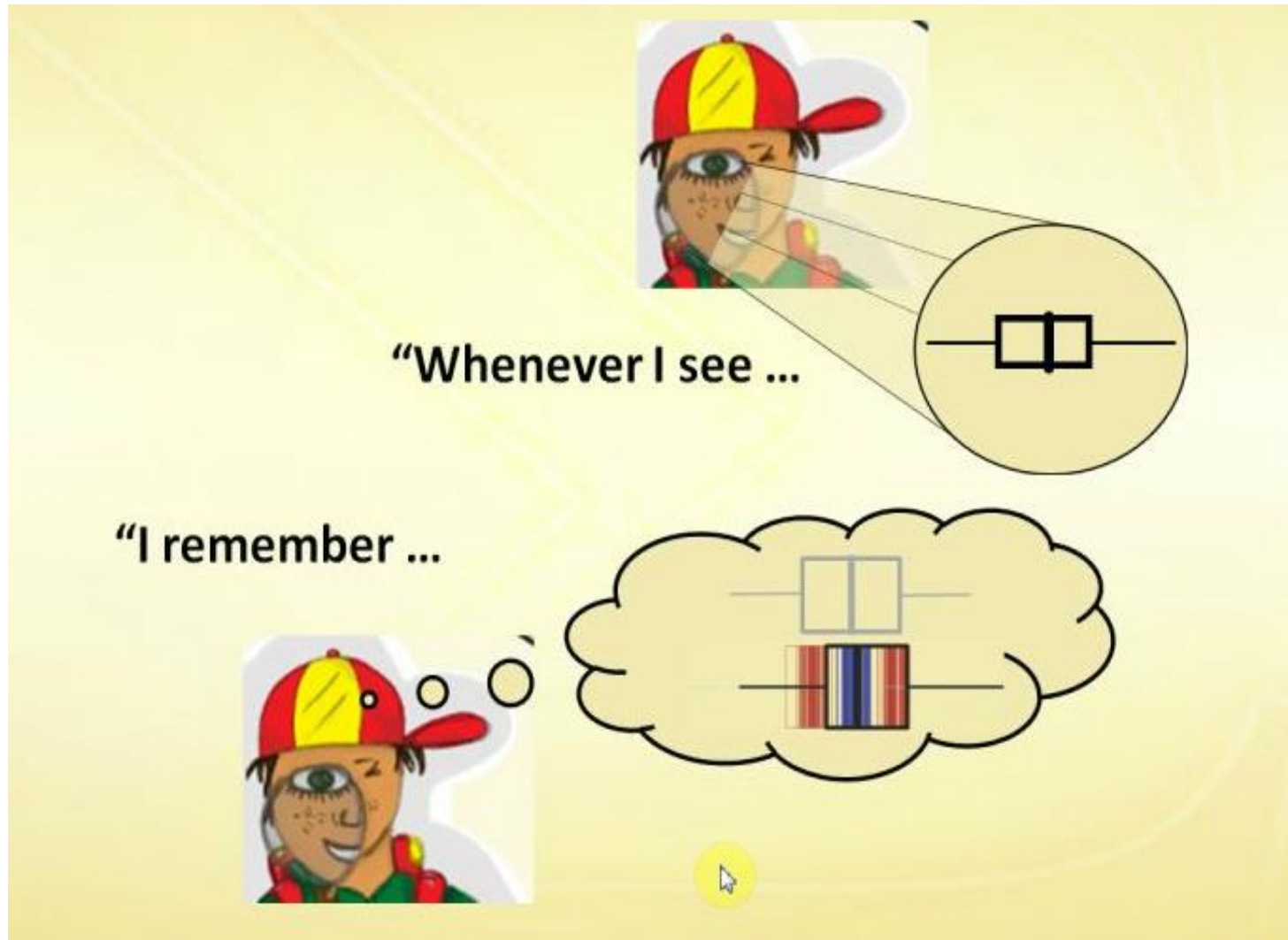
Visualize boxplot with memory

Where is the center of the population?

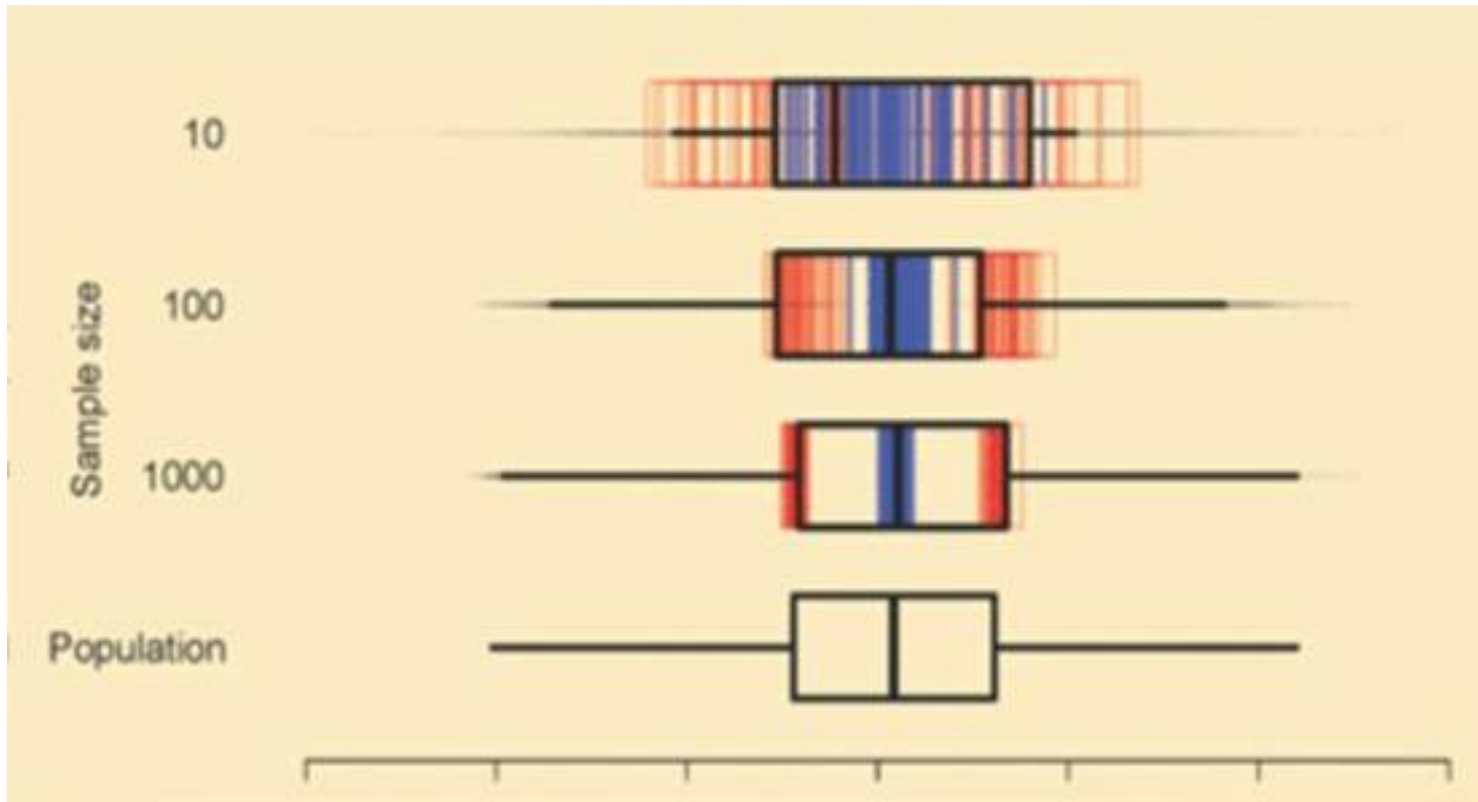


Vizualize boxplot with memory

Where is the center of the population?



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

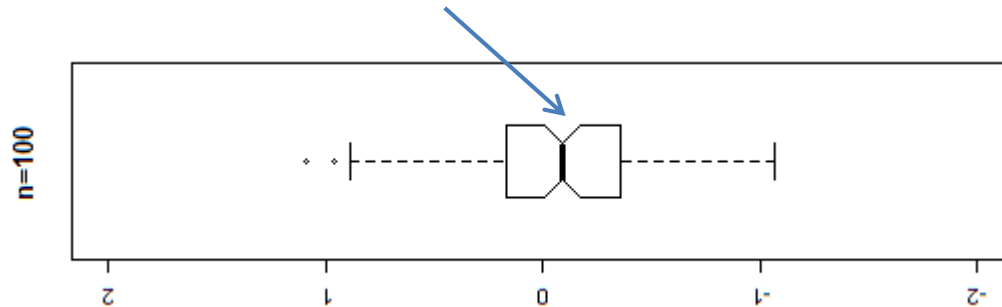


How sure can I be about the true parameter value?

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)
```



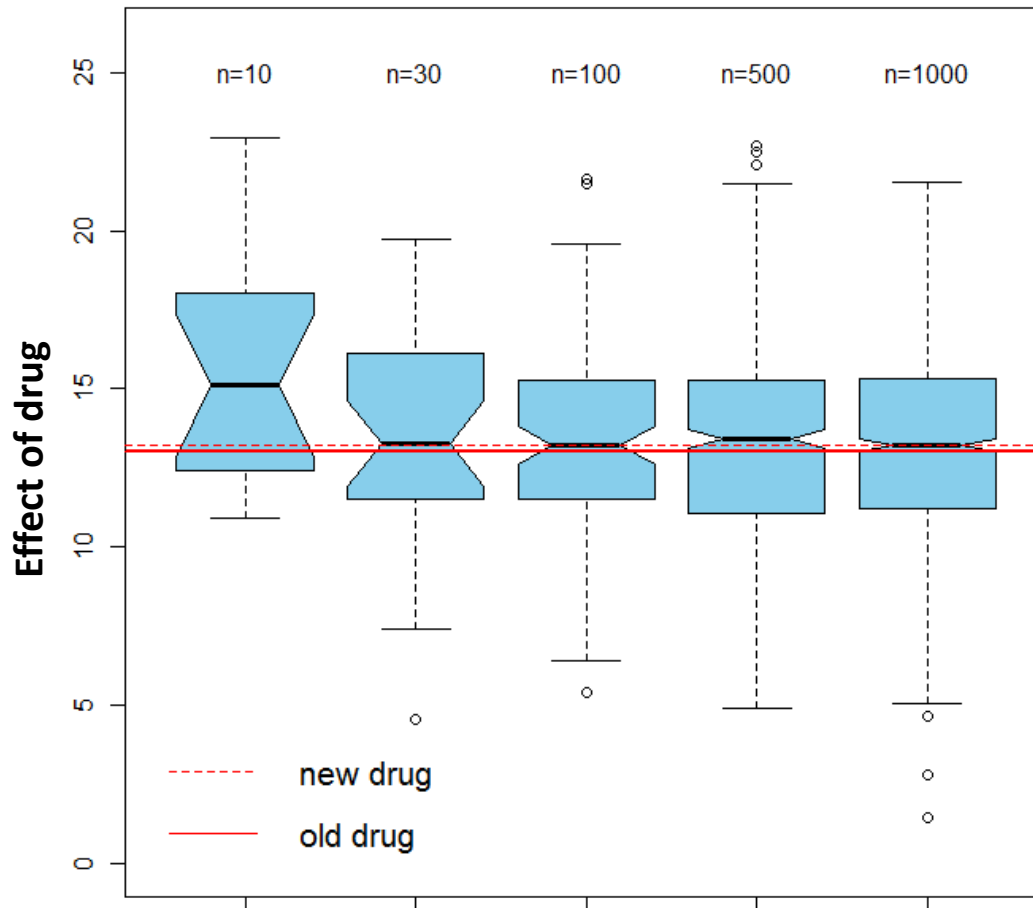
$\pm 1.58 \text{ IQR} / \sqrt{n}$

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



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

$$H_A: \mu > 13$$

Assume true median of the new drug is 13.1 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 α 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.