Medical statistics, informatics and telemedicine

Descriptive statistics

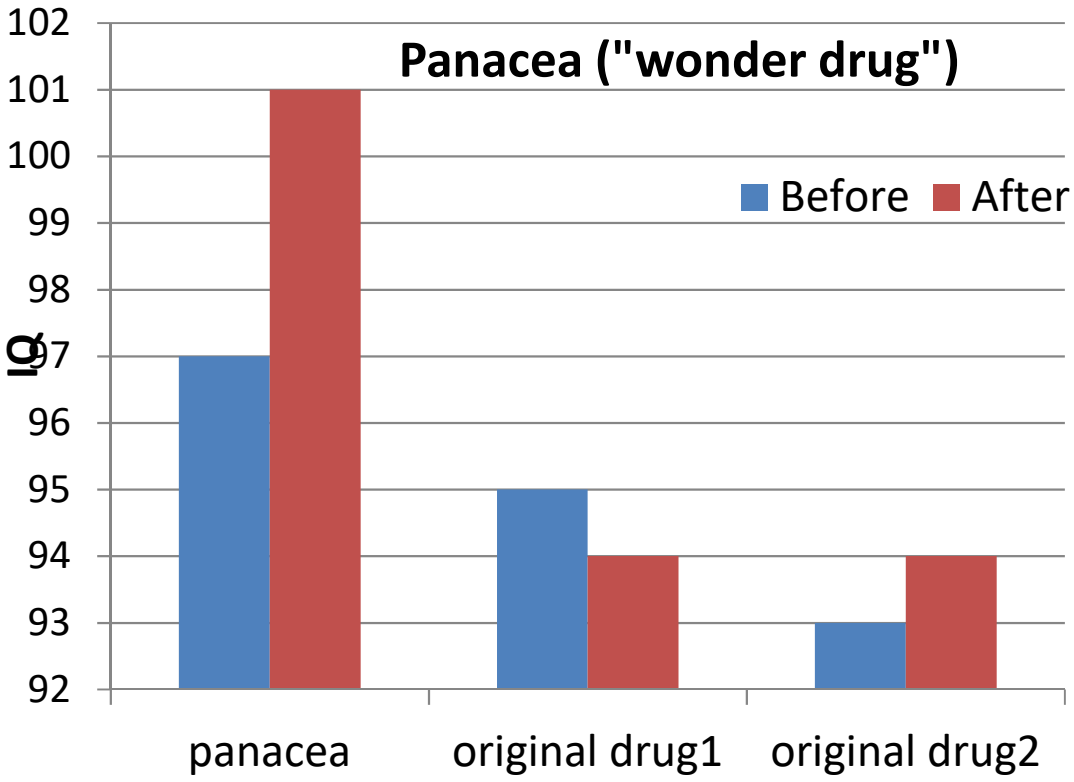
Summarizing data

by Dániel Veres

2021. September 17.

Why to learn medical statistics?

„ To decide whether we should believe in something we are reading or to see where the mistake is, that is to say, do not fall so easily into statistical „juggling”, artifacts and mistakes.”
(Reiczigel-Harnos-Solymosi: Biostatistics for non-statisticians)



Why to learn medical statistics?

- „ So that we can do our best to design, evaluate and interpret our own statistics in our work (diploma thesis…).”

- „ To make it easier for us to seek advice - to work with - a statistician if (when) we are unsure.”

VARIABILITY

To get to know and describe:

Different **way of thinking**

new **nomenclature**

little **math**

Statistics describes **random mass** phenomena.

- ***Data Collecting (Sampling)***
- ***Data Organization***

Descriptive Statistics

- ***Data Analysis***
- ***Conclusion***

Inferential Statistics
(Inductive)

Statistics describes **random mass** phenomena.

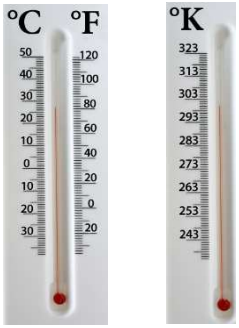
- Data Collecting (Sampling)
- Data Organization

Descriptive Statistics

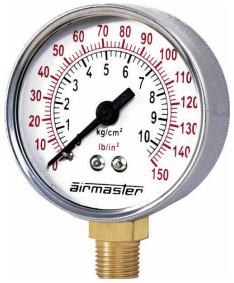
- Data Analysis
- Conclusion

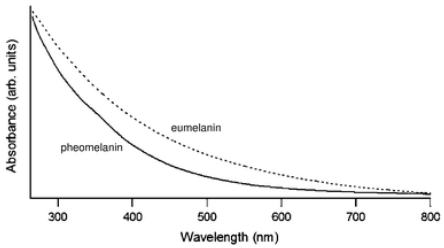
Inferential Statistics
(Inductive)

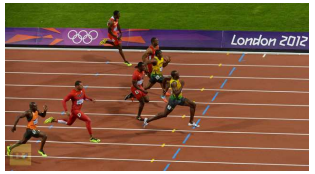
Variables


°C °F
°K


Blue
Turquoise
Amethyst
Gray
Green
Honey
Pure Hazel
TrueSapphire
*New Brilliant Blue
*New Sterling Gray
*New Gemstone Green


airmaster


Absorbance (arb. units)
eumelanin
pheomelanin
300
400
500
600
700
800
Wavelength (nm)


London 2012

Measuring scales – type of variables

qualitative (categorical)

quantitative (numerical)

nominal "scale"	ordinal scale	interval scale	ratio scale
= ≠	= ≠ < >	= ≠ < > + −	= ≠ < > + − × ÷

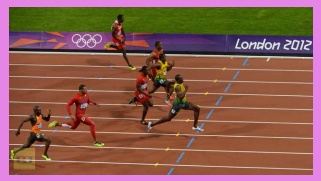
natural order or ranking

known distance

natural zero value

(QUASI-) DISCRETE

(QUASI-) CONTINUOUS



Description of nominal variables I.

Analytical

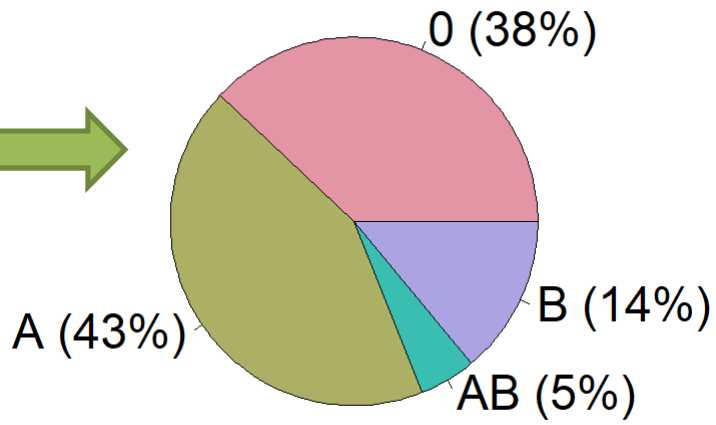
List

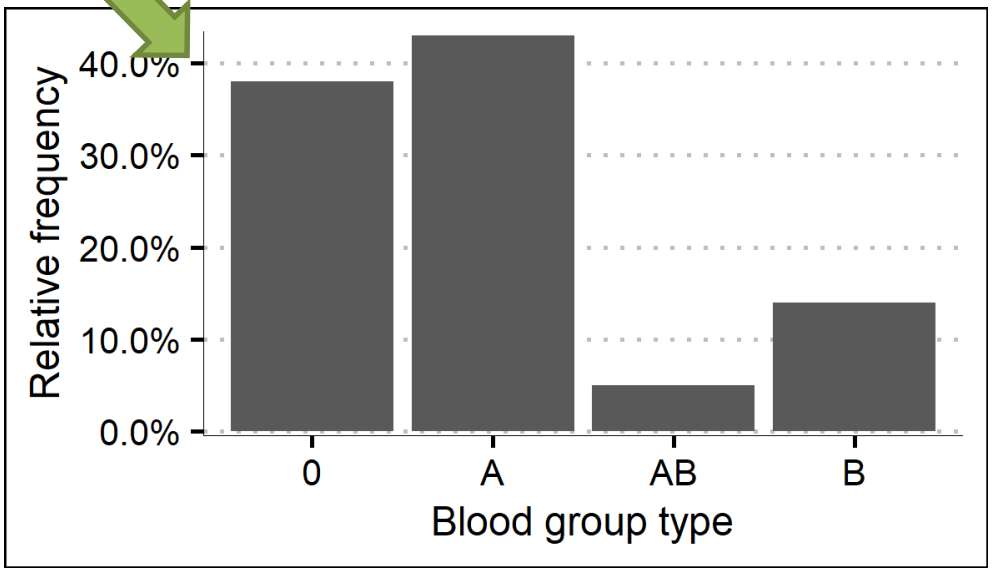
patient №	blood group (ABO)	cholesterol level (mg/dL)
1	B	148
2	AB	147
3	B	169
4	B	159
5	B	150
6	B	167
7	A	144
8	B	158
9	AB	177

Frequency table

blood group	(absolute) frequency	relative frequency
A	85	0.425
B	28	0.14
AB	10	0.05
0	77	0.385
Σ	200	1

Graphical (Plots)





Description of nominal variables II.

Analytical

Frequency table

blood group	(absolute) frequency	relative frequency
A	85	0.425
B	28	0.14
AB	10	0.05
0	77	0.385
Σ	200	1

Organization, but **loss of information**

„Typical value” (***indicator***): ~~**Mean?!**~~

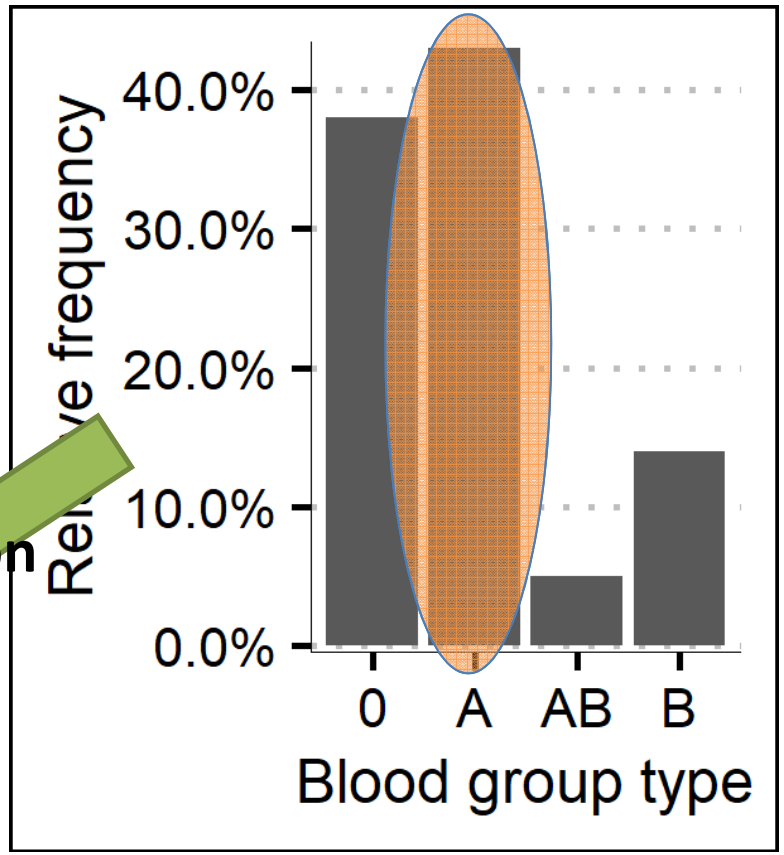
***Mode*:** most frequent element(s)

Notation: *Mod*, x_{mod}

Other parameters:

data count (*n*)***, count of categories***

Graphical



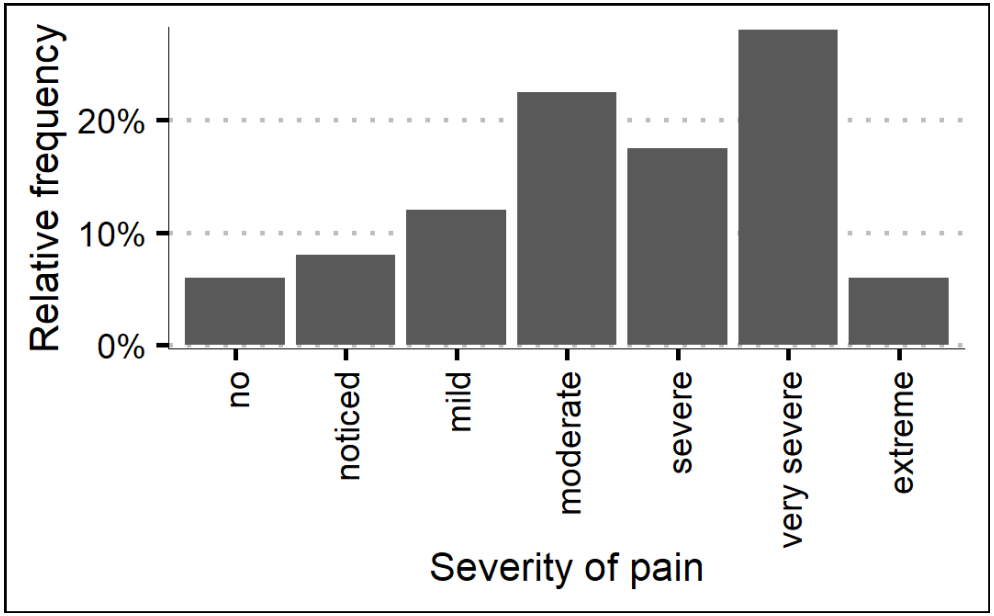
Description of ordinal variables I.

Analytical

Graphical

Frequency table

Severity of pain	Relative frequency	Cumulative relative frequency
no pain	0,06	0,06
noticed	0,08	0,14
mild	0,12	0,26
moderate	0,225	0,485
severe	0,175	0,66
very severe	0,28	0,94
extreme	0,06	1
Σ	1	



Indicator:

Mode

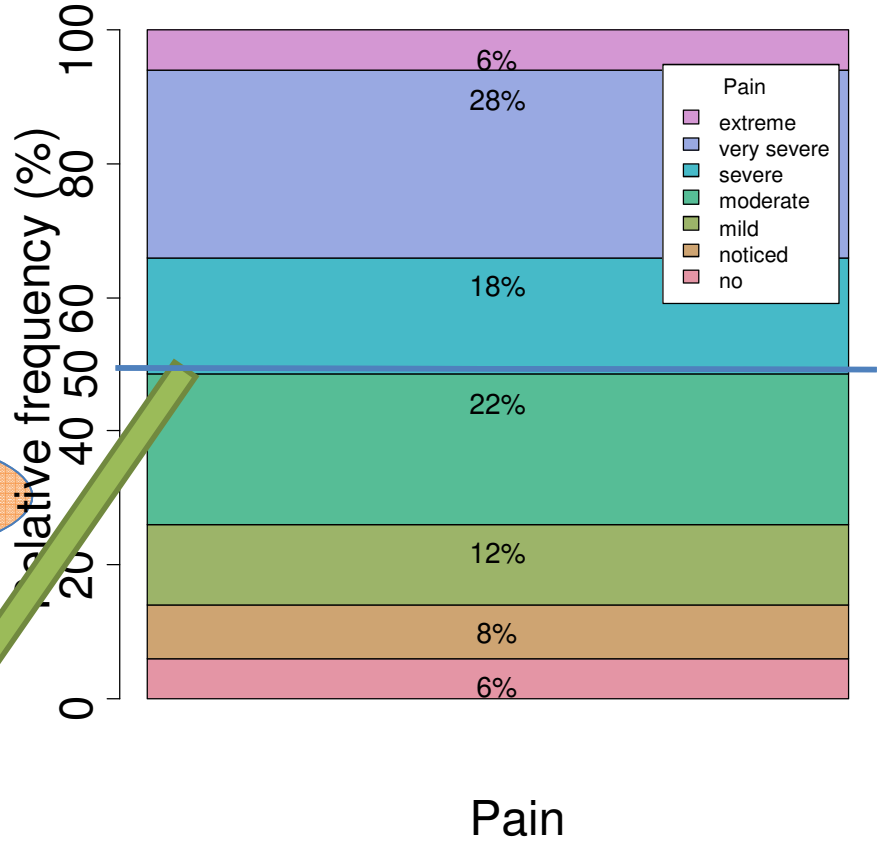
Other parameters:

data count (*n*)***, count of categories***

Description of ordinal variables II.

Frequency table

Severity of pain	Cumulative relative frequency
no pain	0,06
noticed	0,14
mild	0,26
moderate	0,485
severe	0,66
very severe	0,94
extreme	1
Σ	



New indicator:

Median: „middle" element(s)

Notation: Me, Med, x_{med}

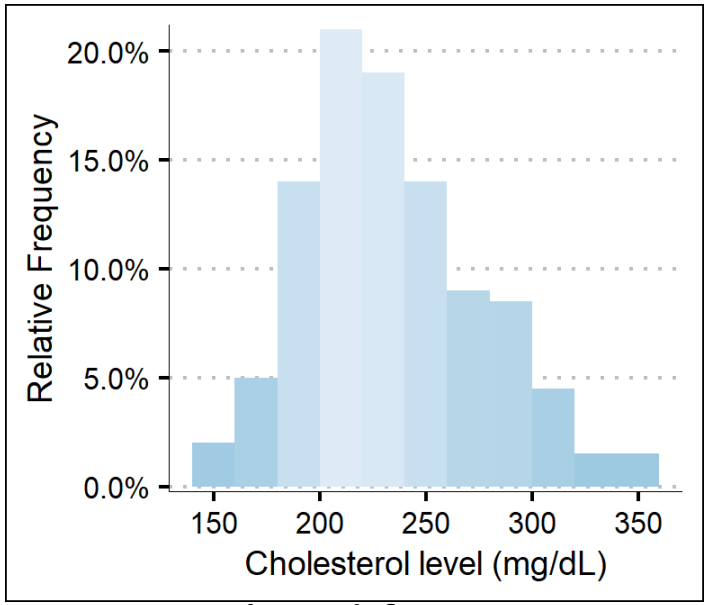
Description of quantitative variables I.

Frequency tables

frequency distributions (differential discrimination functions)

bins (classes, intervals)	(absolute) frequency (FREQUENCY)	relative frequency	(absolute) frequency density	relative frequency density
x ≤ 100	0			
100 < x ≤ 110	0	0	0	0
110 < x ≤ 120	2	0,01	0,2	0,001
120 < x ≤ 130	5	0,025	0,5	0,0025
130 < x ≤ 140	22	0,11	2,2	0,011
140 < x ≤ 150	31	0,155	3,1	0,0155
150 < x ≤ 160	48	0,24	4,8	0,024

Histogram

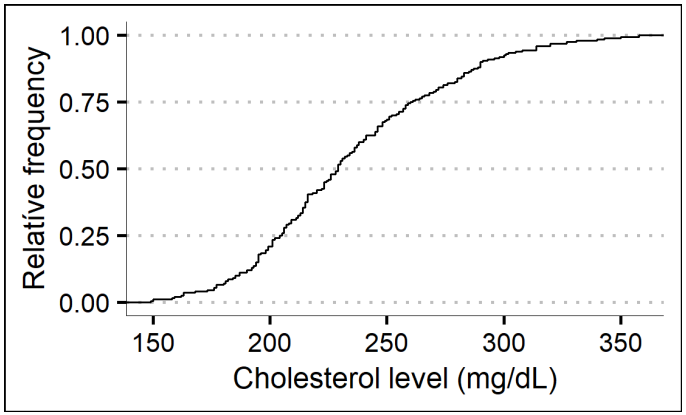


Organizing data – with **loss of information**

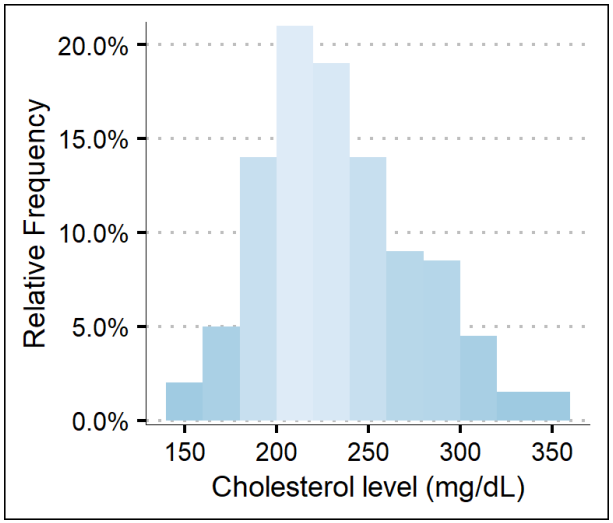
Determination of bin width:

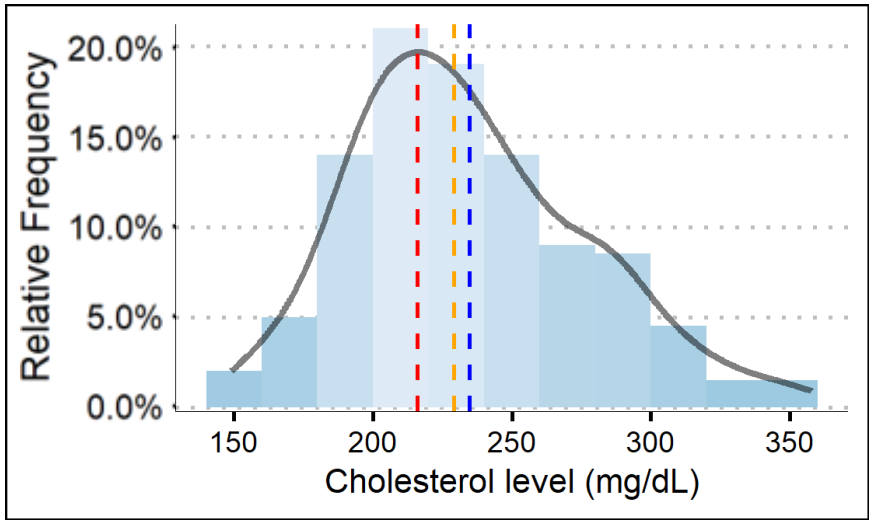
- technical and aesthetic concerns
- statistical concerns

Cumulated frequencies



Description of quantitative variables II.





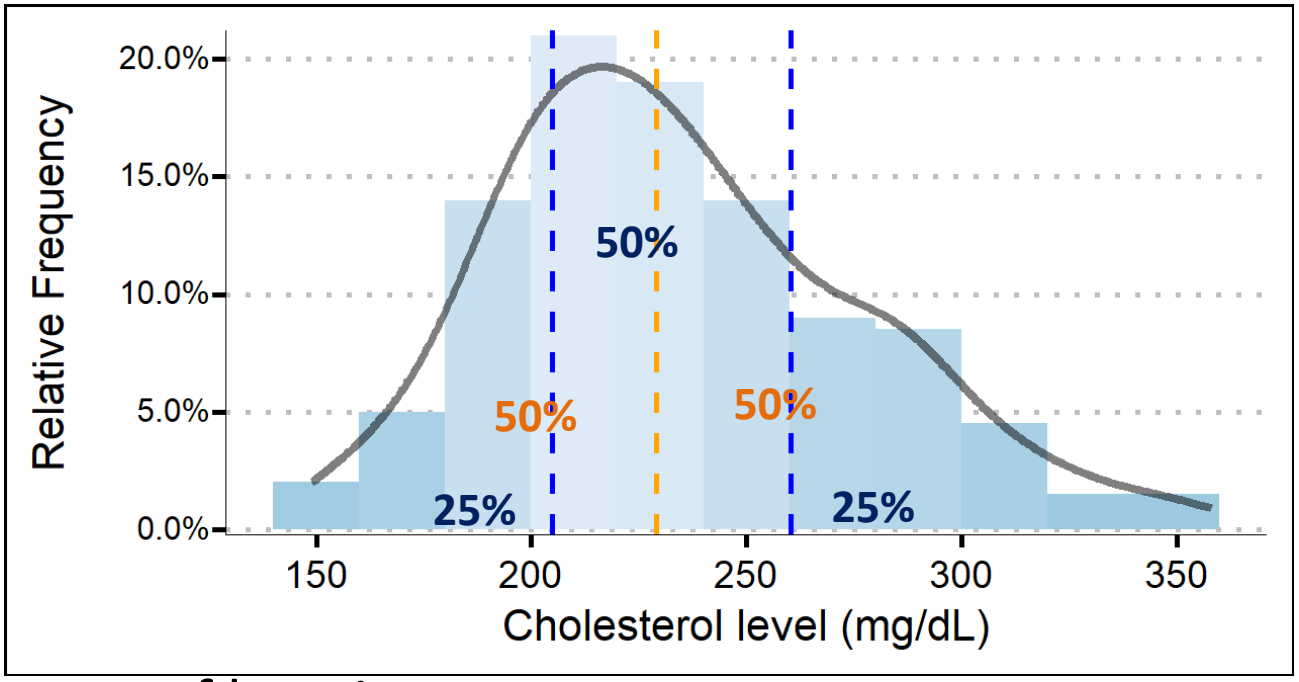
„Typical values" – ***central tendencies*** (special ***measures of location***):

- Mode: most frequent element(s) ?
- Median: „middle" element(s?)
- ***Mean*** (arithmetic mean): „gravity center" , sensitive to „outliers"?

  Notation: x_{mean},

Advantage: compact, ***could be determined from few data***

Quantiles



Other measures of location:

- Median: 50-50% (Q_2)
- ***Quartile***: lower quartile(Q_1): 25-75%; upper quartile (Q_3): 75-25%

General

p-quantile(s): is the number to which the count of data are smaller is maximum *n*p* and to which the count of data are larger is maximum *n*(1 – p)*, where p is between 0 and 1, and *n* is the count of data

Remarks

Day	Waiting time (min)		
1	1,27	median	8,48
2	3,3	lower quartile	3,59
3	3,44	mean	7,72
4	3,64		
5	6,33		
6	7,72		
7	9,23		
8	9,87		
9	10,31		
10	12,29		
11	12,3		
12	12,98		

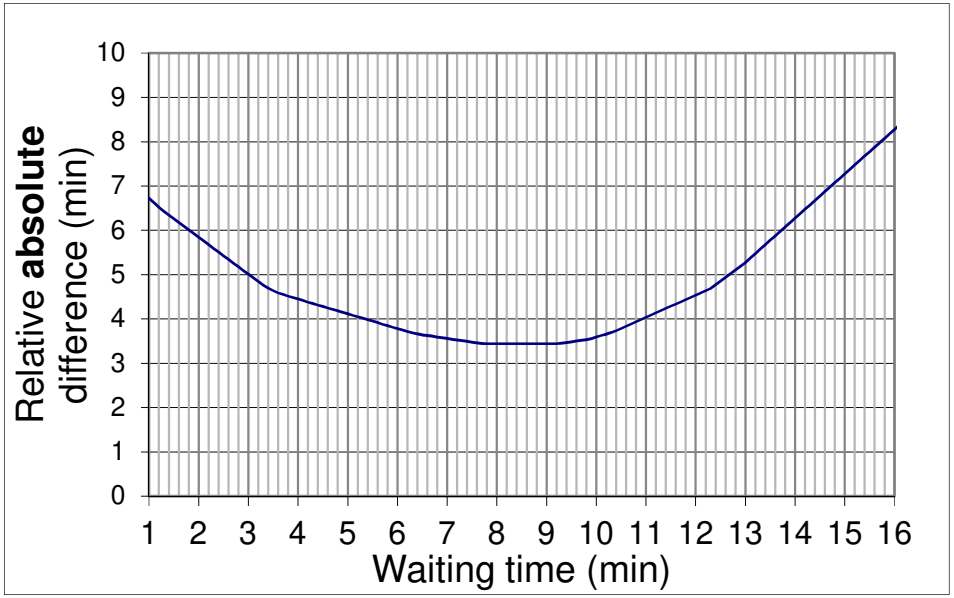
Day	Waiting time (min)		
1	1,27	median	8,48
2	3,3	lower quartile	3,59
3	3,44	mean	8,31
4	3,64		
5	6,33		
6	7,72		
7	9,23		
8	9,87		
9	10,31		
10	12,29		
11	12,3		
12	20		

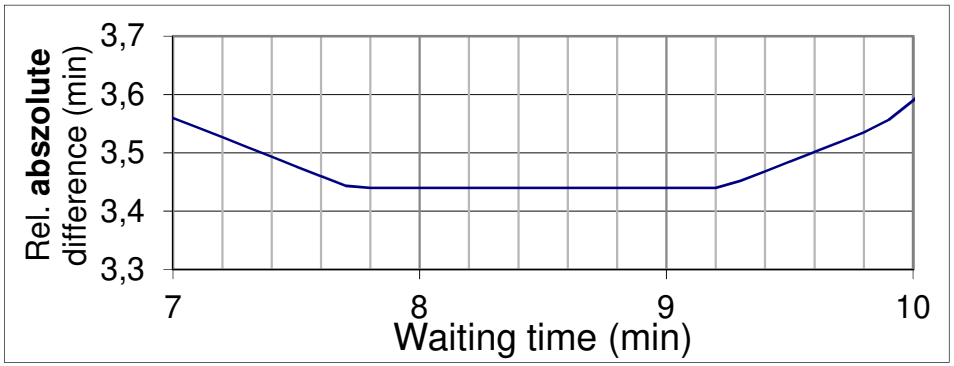
Median, quantiles could differ in theory and practice.
Mean is sensitive to the outliers, but quantiles not (...).
Mode?

Distances

$$\frac{1}{n}\sum\left|x_i - x^*\right|$$

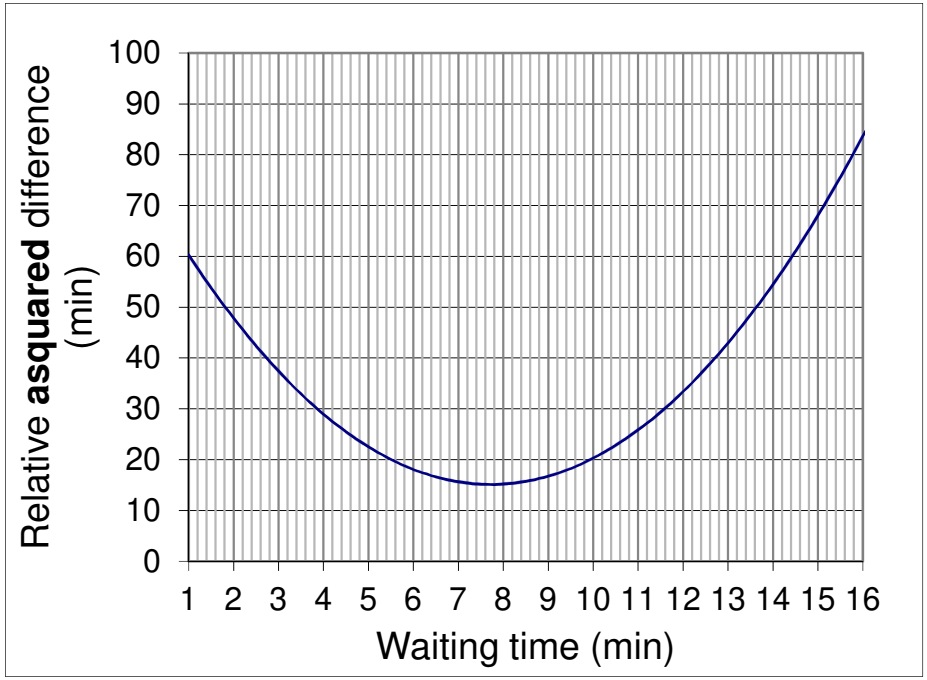
Minimal, if: $x^* = Median$



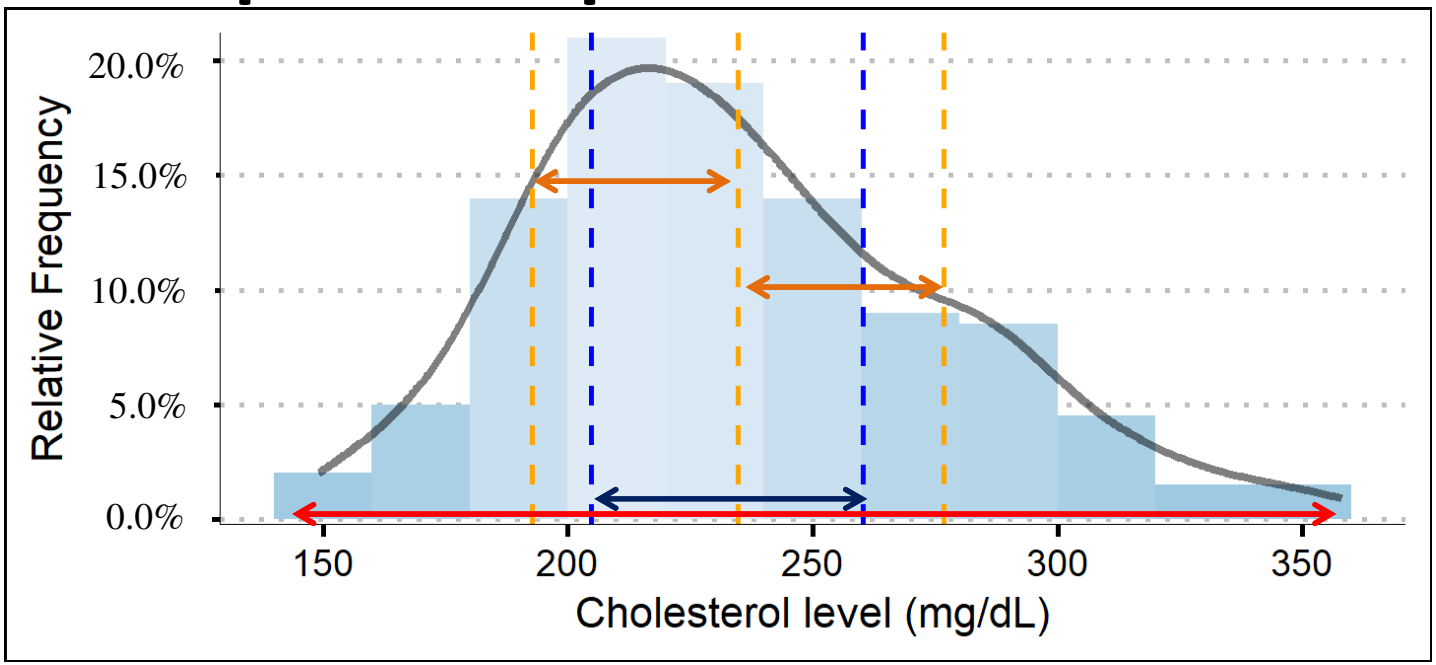


$$\frac{1}{n}\sum\left(x_i - x^*\right)^2$$

Minimal, if: $x^* = Mean$



Description of quantitative variables III.



Measures of spread:

- ***Range***: the difference between the maximum and the minimum
- ***Variance (s^2):*** the average of the squared distance from the mean
- ***Standard deviation (s, sd, SD)***: the square root of the variance

  the width of the curve
- ***Interquartile range (IQR)*** : the difference between the upper and the lower quartile

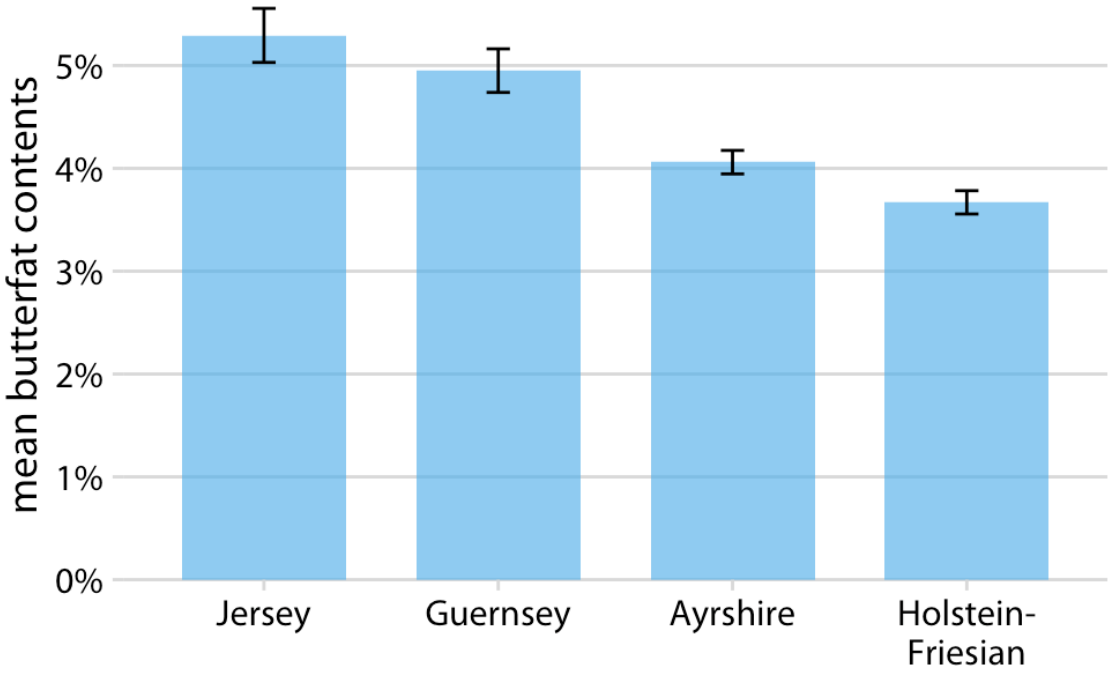
  not sensitive to the „outliers”

Do NOT

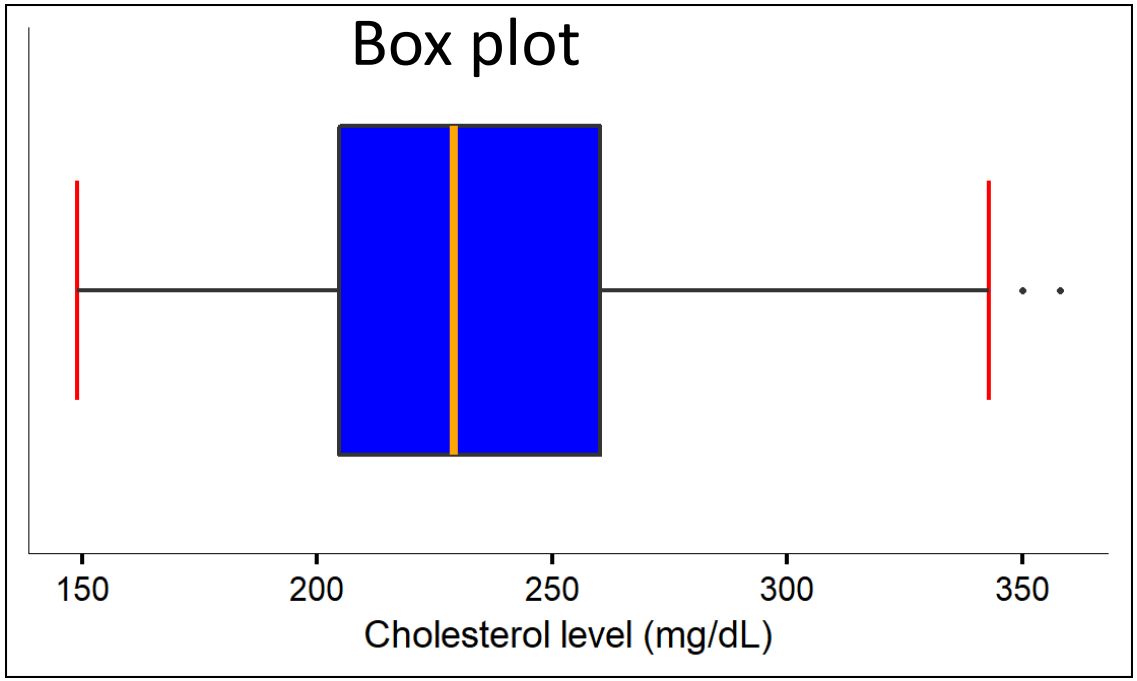
Bar chart for numerical characteristics:

actually the area under the curve is highlighted in the figure,

- but we look the "top" not even the area under the curve ...
- It must start at 0 (should)
- the standard deviation is a y-range, the value (mean) is an area
- and if the standard deviation is not symmetric? and if it is higher than average?
- if the average is negative?
- ...



Description of quantitative variables IV.



Middle point: *median*

Box: *interquartile range (IQR)*

Whisker: minimum and maximum without outliers

outliers: *>1.5*IQR difference from quartiles*

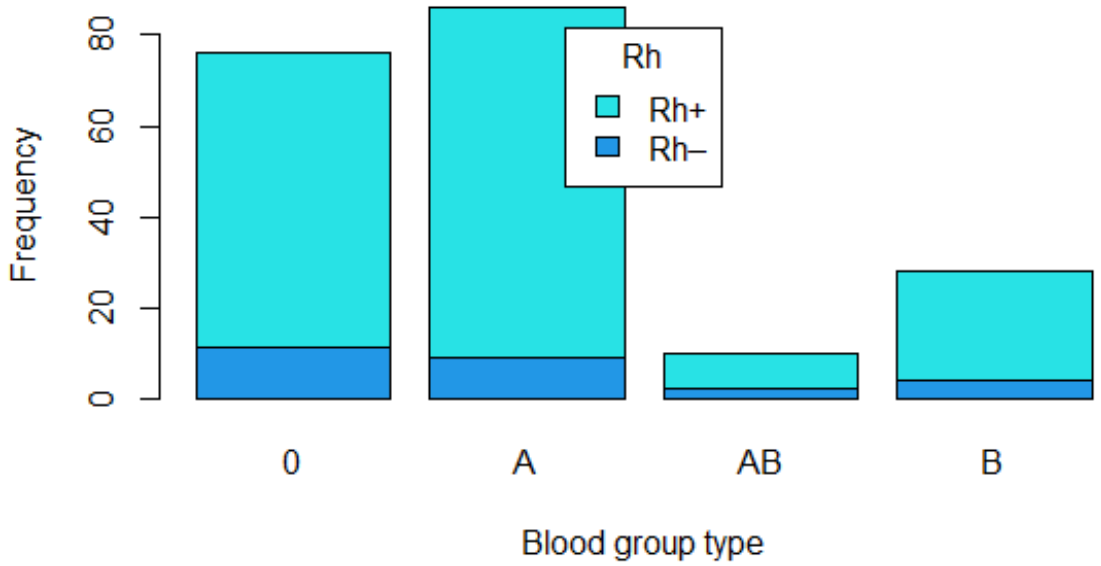
out of whiskers: ***outliers***

Qualitative bivariate description

Contingency table

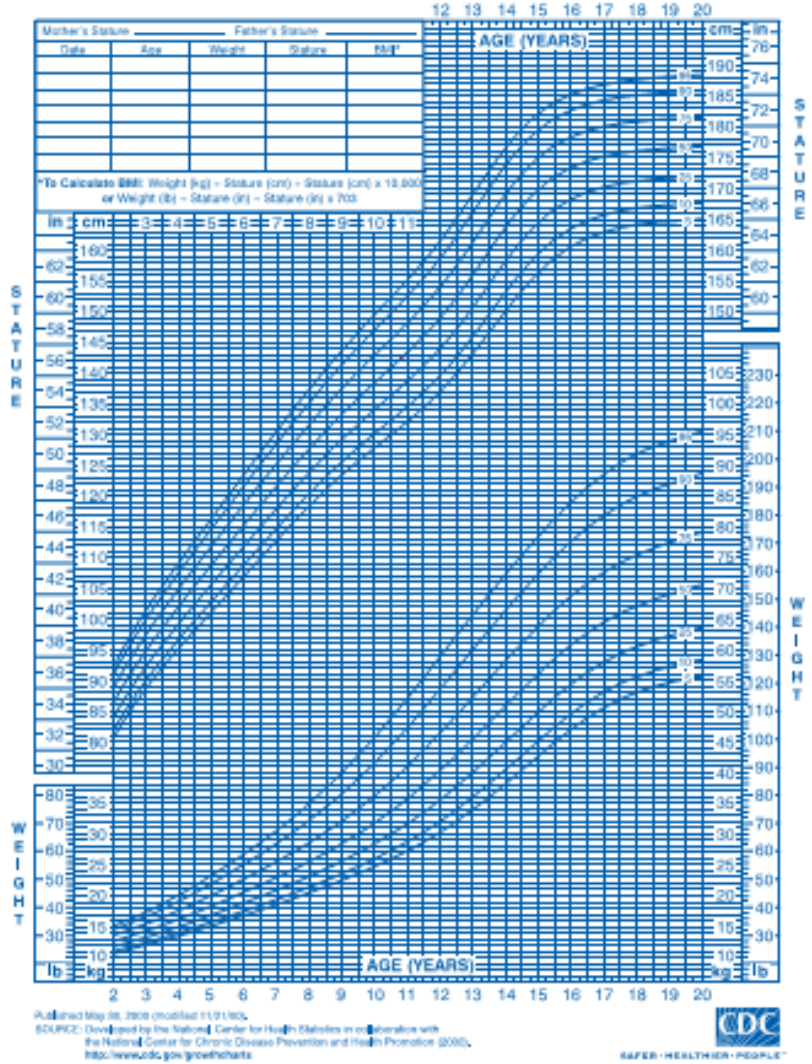
	A	B	AB	0	Σ
Rh+	77	24	8	65	154
Rh–	9	24	2	11	46
Σ	86	48	10	76	200

Stacked bar chart

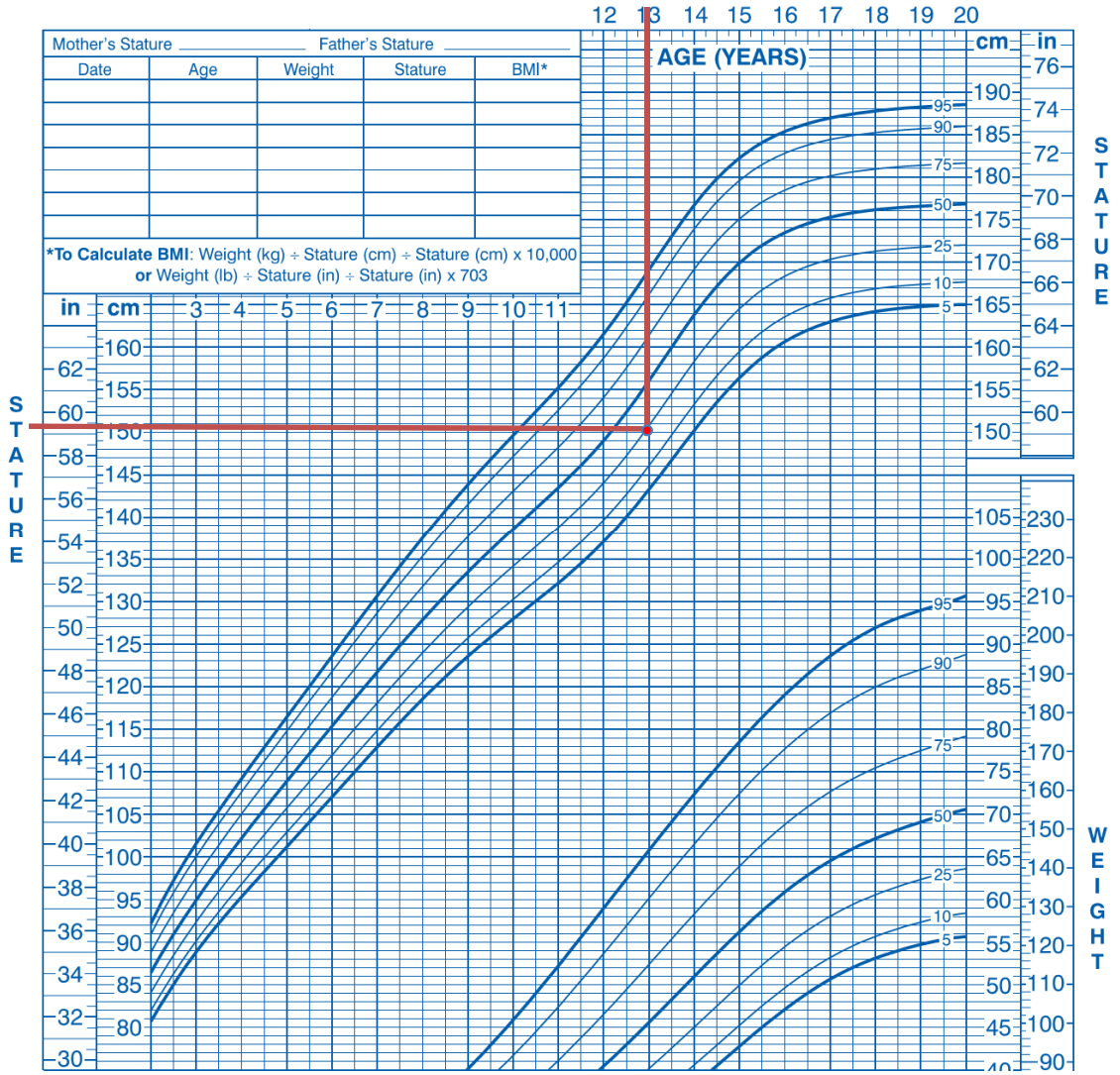


Quantitative bivariate description

Percentile curves



Percentile curves



Colors

Which one?

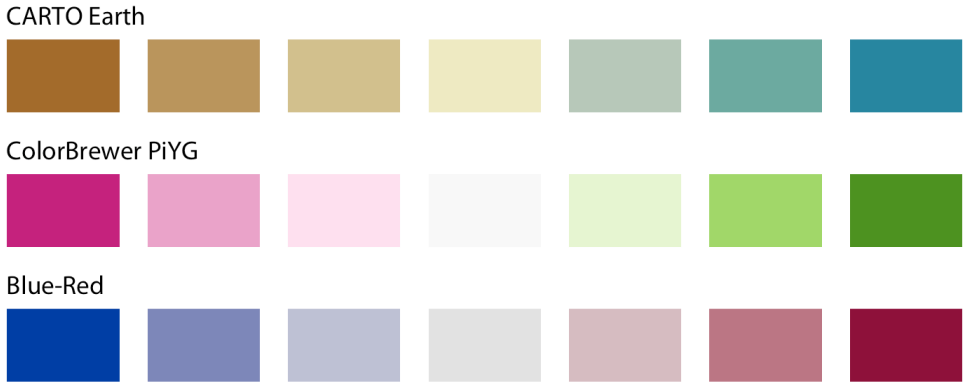
- **if no order**: qualitative color scale
  well distinguishable
  do not show order, „uniform” :
  eg. same „brightness”, „saturation”

Colors

Which one?

- **There is an order:** sequential color scale
  show order (smaller-larger)
  show the magnitude of order

If symmetric

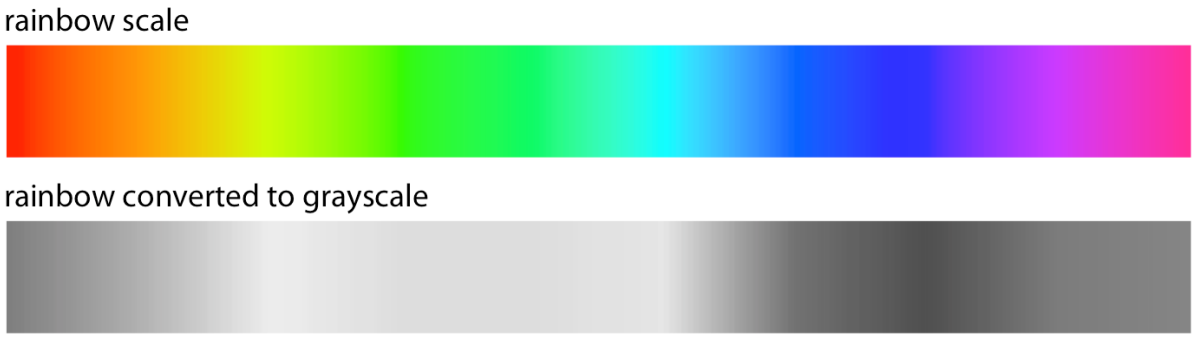


Colors

Which one?

- **There is an order:** sequential color scale
  show order (smaller-larger)
  show the magnitude of order

NOT like that: rainbow (too „fast” or „slow” change, same at the end)



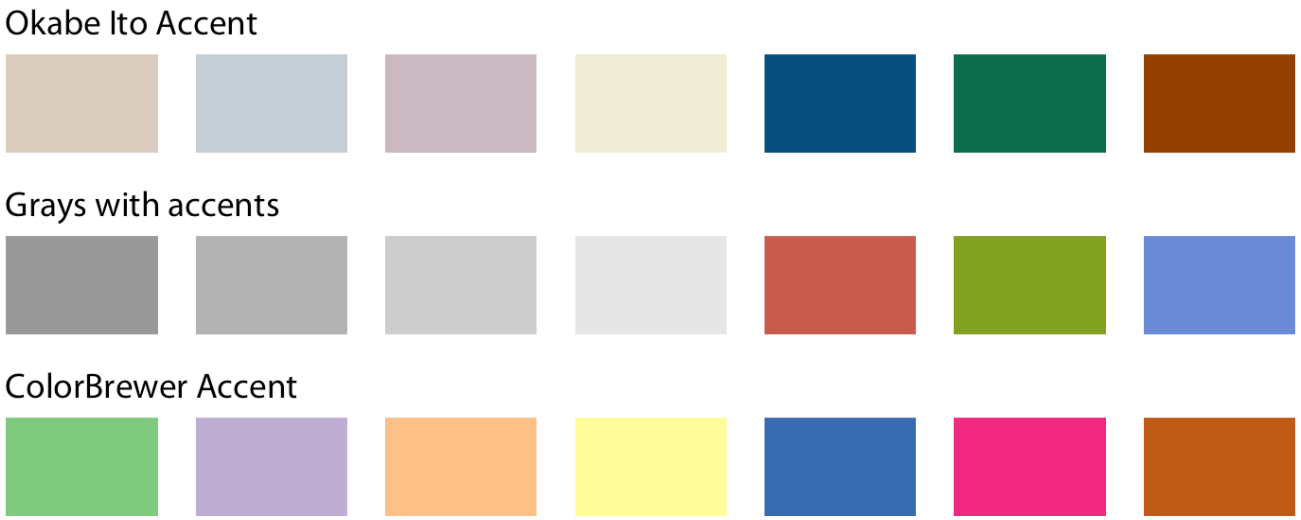
Colors

Which one?

- **If one have to be highlighted**:

  let it be different: darker or more saturated

  can have more base colors and some highlights



Colors

Which one?

Think about eg. me – COLORBLIND (old projector: red-)

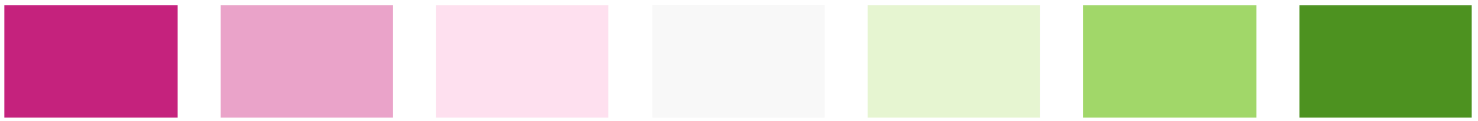
(red-green [...] ~ 8% males, 0.5% females)

Good choice, eg:

No order

Symmetric order

ColorBrewer PiYG



Softwares for testing! (it „symulates” colorblinds)

eg. ImageJ: VischeckJ

Photoshop: CUD

Feedback