Grundpraktikum Bioinfo - Week 1

Biological Data Analysis

Carl Herrmann
IPMB - Universität Heidelberg

IPMB
Institut für Pharmazie und Molekulare Biotechnologie

UNIVERSITÄT HEIDELBERG
ZUKUNFT SEIT 1386

Goal of this lecture - week 1

- Introduce the basic concepts in statistics
  - descriptive statistics - graphical representation
  - inference statistics
  - hypothesis testing; multiple testing
  - regression analysis
- Introduce notions of data analysis
  - data formats
  - data imputation
  - data cleaning
- Give you notions of data analysis in R
  - introduction to tidyverse
  - making nice plots with ggplots

morning lectures

afternoon labs

Day 1

- what kind of data? - data types
- which plot for which data? - graphical representation
- how does my data look like? - describing data QQ-plots
- are my variables related ? - correlation

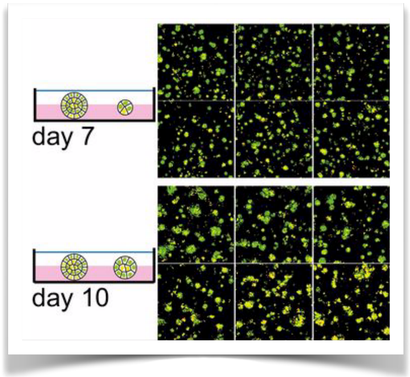
IPMB
Institut für Pharmazie und Molekulare Biotechnologie

UNIVERSITÄT HEIDELBERG
ZUKUNFT SEIT 1386

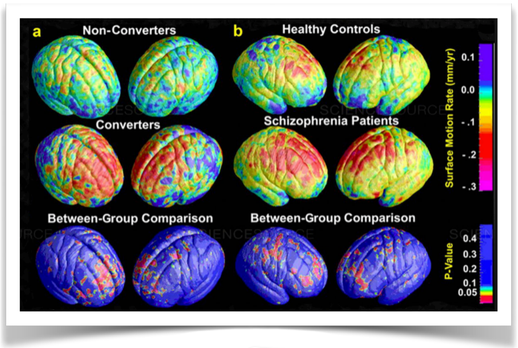
1. Basics on data types

Biological data sources

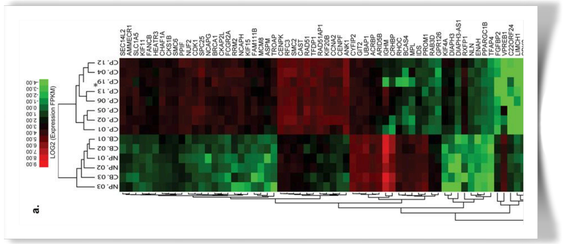
drug screens



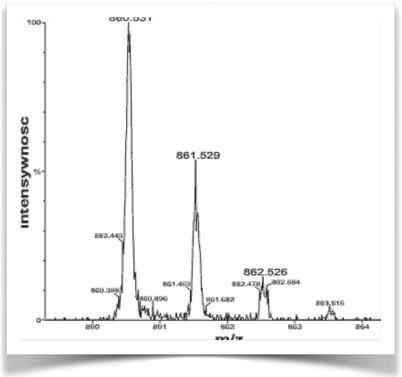
MRI imaging



Genomics



proteomics

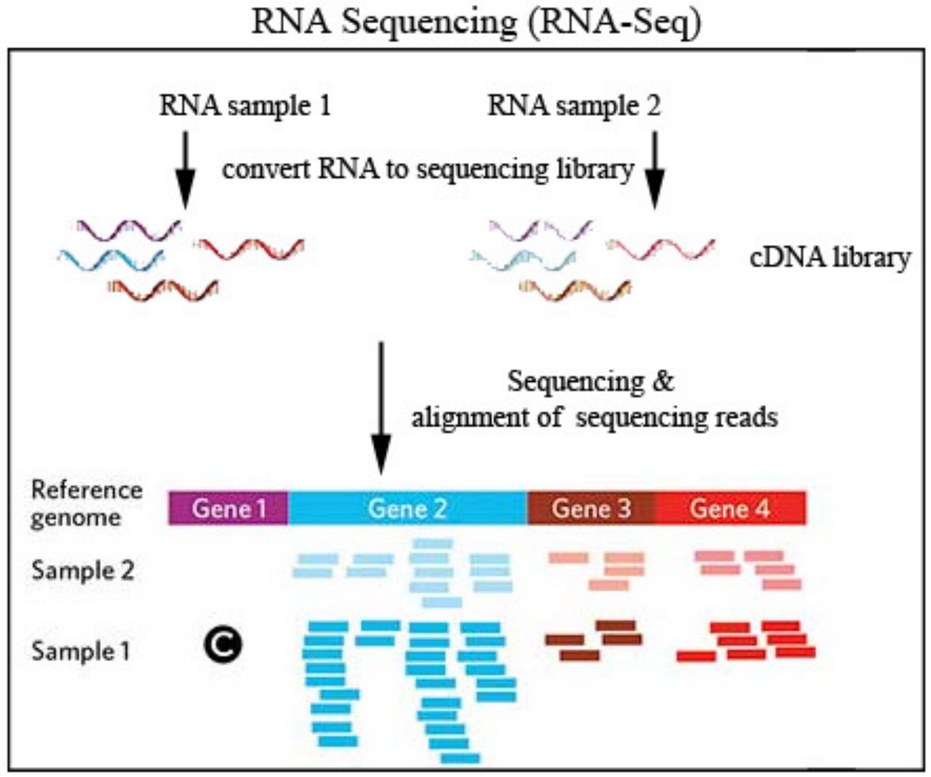


large (numerical) data matrices

Patient data

Example: gene expression

- mRNA molecules are retro-transcribed into cDNA
- cDNA libraries are sequenced into short fragments (~ 100 bp)
- Fragments are aligned to the genome
- Normalized number of fragments per gene is a measure of the gene expression
- examples of normalizations: reads per kilobase per million aligned reads (RPKM)



Example : gene expression as counts

patients (n ~ 100 - 1000)

genes (n ~30.000)

Description	GTEX-1117F-0226-SM-5GZZ7	GTEX-111CU-1826-SM-5GZYN	GTEX-111FC-0226-SM-5N9B8	GTEX-111VG-2326-SM-5N9BK	GTEX-111YS-2426-SM-5GZZQ	GTEX-1122O-2026-SM-5NQ91	GTEX-1128S-2126-SM-5H12U	GTEX-113IC-0226-SM-5HL5C
DDX11L1	3	4	1	1	0	2	1	3
WASH7P	616	395	826	364	301	419	340	451
MIR1302-11	2	1	1	0	1	0	2	3
FAM138A	1	0	1	1	0	0	2	3
OR4G4P	0	0	0	0	0	0	2	1
OR4G11P	0	2	2	0	0	1	0	0
OR4F5	0	0	0	0	2	0	0	0
RP11-34P13.7	8	3	12	12	2	4	10	9
CICP27	11	29	9	18	5	5	7	4

Tables are usually oriented such that # rows > # columns

Example : clinical data

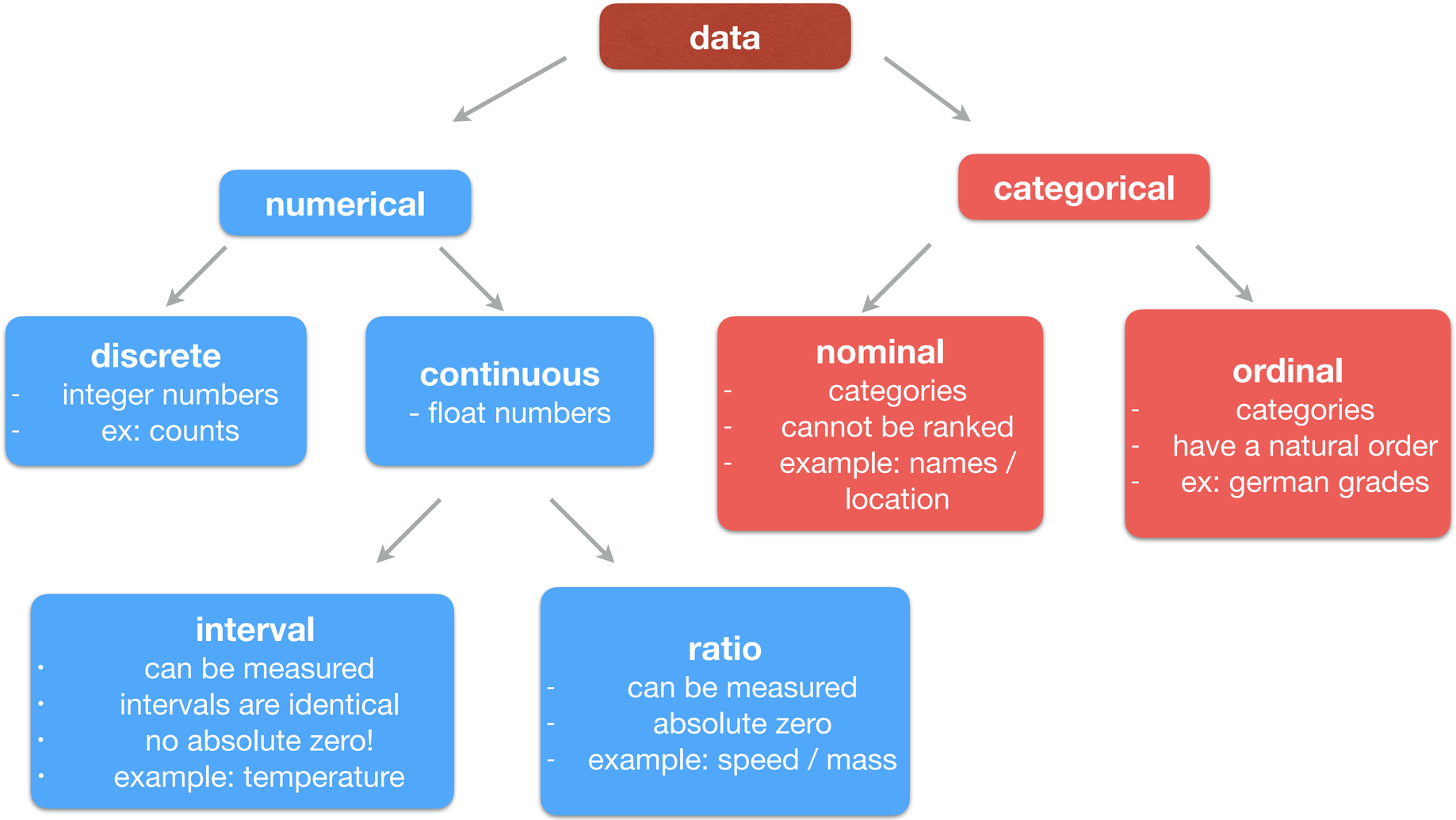
Cohort of diabetes patients with multiple clinical parameters

id	chol	stab.glu	hdl	ratio	glyhb	location	age	gender	height	weight	frame	bp.1s	bp.1d	bp.2s	bp.2d	waist	hip	time.ppn
1000	203	82	56	3.60	4.31	Buckingham	46	female	62	121	medium	118	59	NA	NA	29	38	720
1001	165	97	24	6.90	4.44	Buckingham	29	female	64	218	large	112	68	NA	NA	46	48	360
1002	228	92	37	6.20	4.64	Buckingham	58	female	61	256	large	190	92	185	92	49	57	180
1003	78	93	12	6.50	4.63	Buckingham	67	male	67	119	large	110	50	NA	NA	33	38	480
1005	249	90	28	8.90	7.72	Buckingham	64	male	68	183	medium	138	80	NA	NA	44	41	300
1008	248	94	69	3.60	4.81	Buckingham	34	male	71	190	large	132	86	NA	NA	36	42	195
1011	195	92	41	4.80	4.84	Buckingham	30	male	69	191	medium	161	112	161	112	46	49	720
1015	227	75	44	5.20	3.94	Buckingham	37	male	59	170	medium	NA	NA	NA	NA	34	39	1020
1016	177	87	49	3.60	4.84	Buckingham	45	male	69	166	large	160	80	128	86	34	40	300
1022	263	89	40	6.60	5.78	Buckingham	55	female	63	202	small	108	72	NA	NA	45	50	240
1024	242	82	54	4.50	4.77	Louisa	60	female	65	156	medium	130	90	130	90	39	45	300
1029	215	128	34	6.30	4.97	Louisa	38	female	58	195	medium	102	68	NA	NA	42	50	90
1030	238	75	36	6.60	4.47	Louisa	27	female	60	170	medium	130	80	NA	NA	35	41	720
1031	183	79	46	4.00	4.59	Louisa	40	female	59	165	medium	NA	NA	NA	NA	37	43	60
1035	191	76	30	6.40	4.67	Louisa	36	male	69	183	medium	100	66	NA	NA	36	40	225
1036	213	83	47	4.50	3.41	Louisa	33	female	65	157	medium	130	90	120	96	37	41	240
1037	255	78	38	6.70	4.33	Louisa	50	female	65	183	medium	130	100	NA	NA	37	43	180

Different data types

Variable	Explanation	Unit	Type
chol	total cholesterol		numerical
stab.glu	Stabilized Glucose		numerical
hdl	High Density Lipoprotein		numerical
ratio	Cholesterol/HDL Ratio		numerical
glyhb	Glycosolated Hemoglobin		numerical
location			**categorical**
age			numerical
gender			**categorical**
height		inches	numerical
weight		pounds	numerical
frame			**categorical**
bp.1s	systolic blood pressure		numerical
bp.1d	diastolic blood pressure		numerical
bp.2s	systolic blood pressure		numerical
bp.2d	diastolic blood pressure		numerical
waist		inches	numerical
hip		inches	numerical
time.ppn	Time since last meal	minutes	numerical

Different data types



Different data types

Variable	Explanation	Unit	Type	Subtype
chol	total cholesterol		numerical	ratio
stab.glu	Stabilized Glucose		numerical	ratio
hdl	High Density Lipoprotein		numerical	ratio
ratio	Cholesterol/HDL Ratio		numerical	ratio
glyhb	Glycosolated Hemoglobin		numerical	ratio
location			**categorical**	**nominal**
age			numerical	ratio
gender			**categorical**	**nominal**
height		inches	numerical	ratio
weight		pounds	numerical	ratio
frame			**categorical**	**ordinal (small/medium/large)**
bp.1s	systolic blood pressure		numerical	ratio
bp.1d	diastolic blood pressure		numerical	ratio
bp.2s	systolic blood pressure		numerical	ratio
bp.2d	diastolic blood pressure		numerical	ratio
waist		inches	numerical	ratio
hip		inches	numerical	ratio
time.ppn	Time since last meal	minutes	numerical	ratio

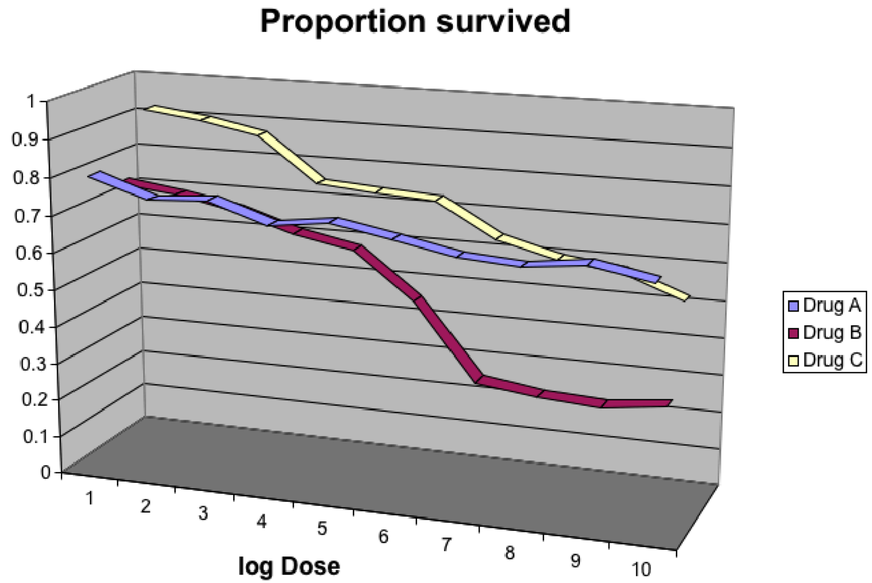
IPMB
Institut für Pharmazie und Molekulare Biotechnologie

UNIVERSITÄT HEIDELBERG
ZUKUNFT SEIT 1386

2. Visualizing data with plots

Graphical representation

- Appropriate graphical representation depends on the type of data
  - categorical
  - counts
  - continuous data
- Aim of good data graphics: **display data accurately and clearly**
  - (Karl Broman https://www.biostat.wisc.edu/~kbroman/topten_worstgraphs/)
  - https://www.polymersearch.com/blog/10-good-and-bad-examples-of-data-visualization
- **Bad practice:**
  - as little information as possible
  - make things obscure through inappropriate graphics
  - pseudo 3D
  - poor scales



Example of bad plot

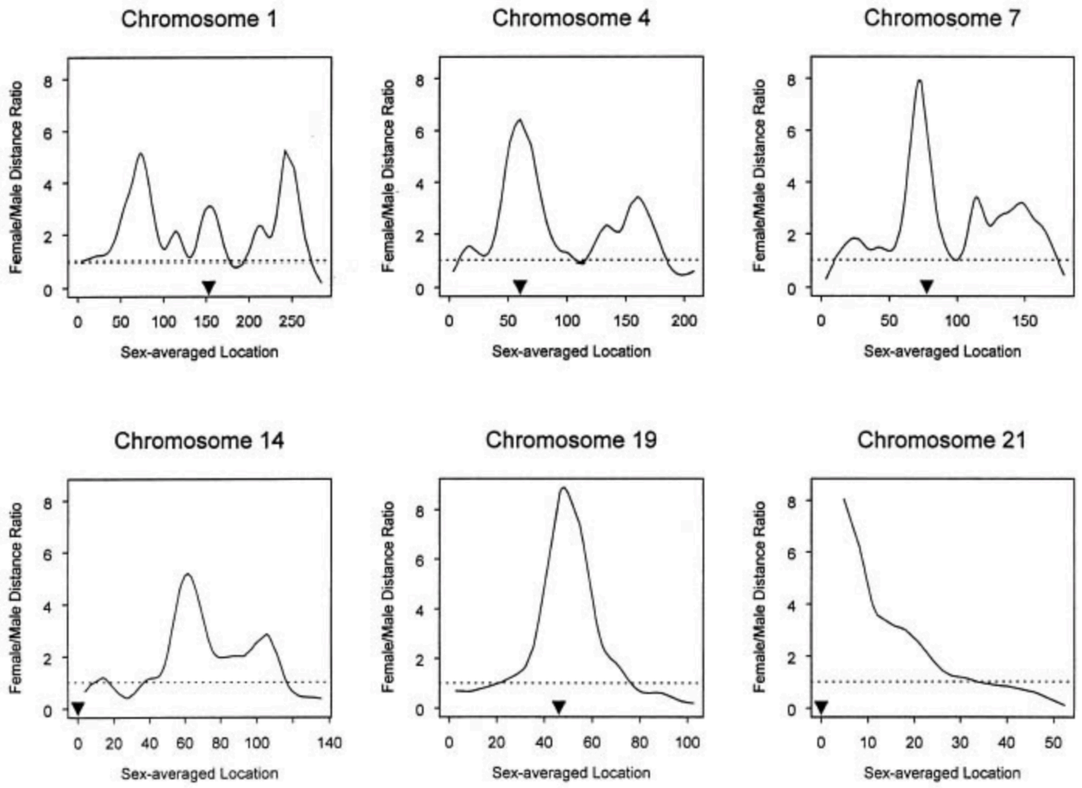


Figure 1 Plots of the female:male genetic-distance ratio against sex-averaged genetic location (in cM) along six selected chromosomes. Approximate locations of the centromeres are indicated by the triangles. The dashed lines correspond to equal female and male distances.

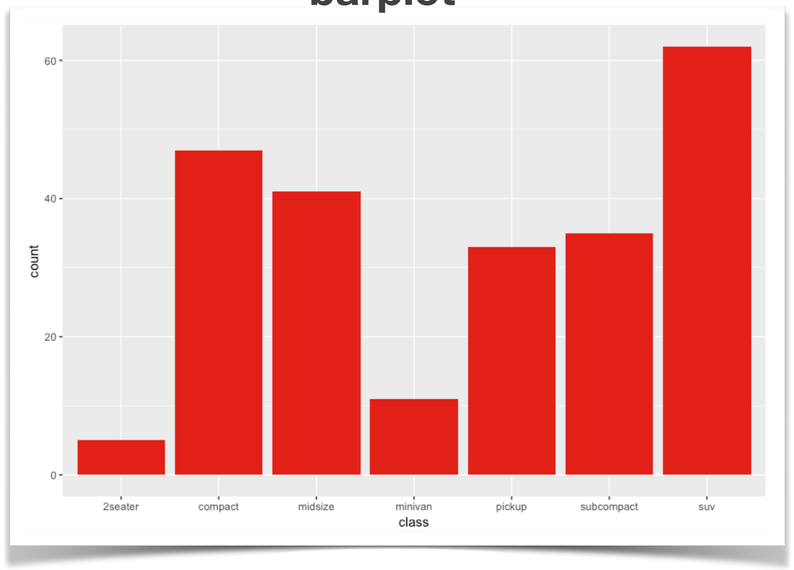
What's wrong with this plot?

https://www.biostat.wisc.edu/~kbroman/topten_worstgraphs/)

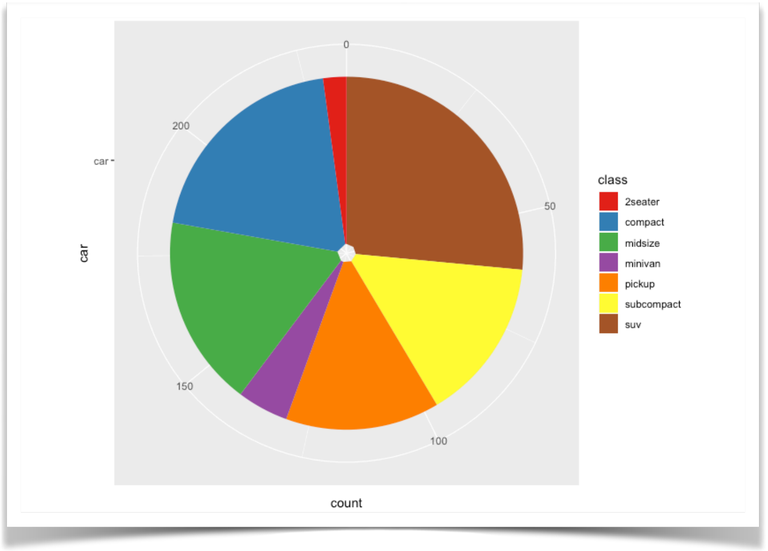
Categorical Data
Barplots

- How many instances in each category?
- Only meaningfull measure: **MODE** (= category with highest counts)
- Possible plots: **barplots**; **piecharts**

barplot



piechart

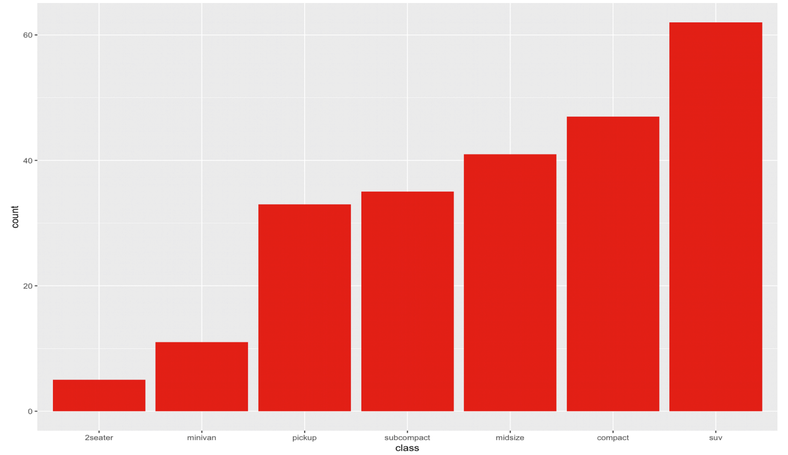


Avoid piechart : areas are more difficult to judge than length!

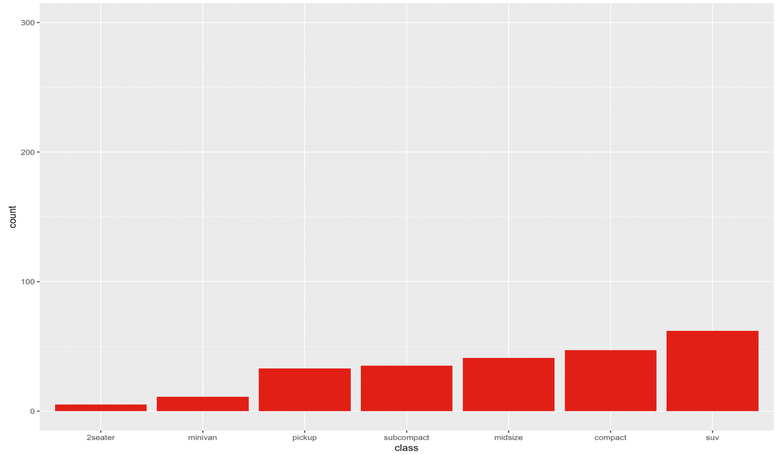
Categorical Data

Barplots

- Consider ordering the data by counts (no natural order of categories for nominal data)



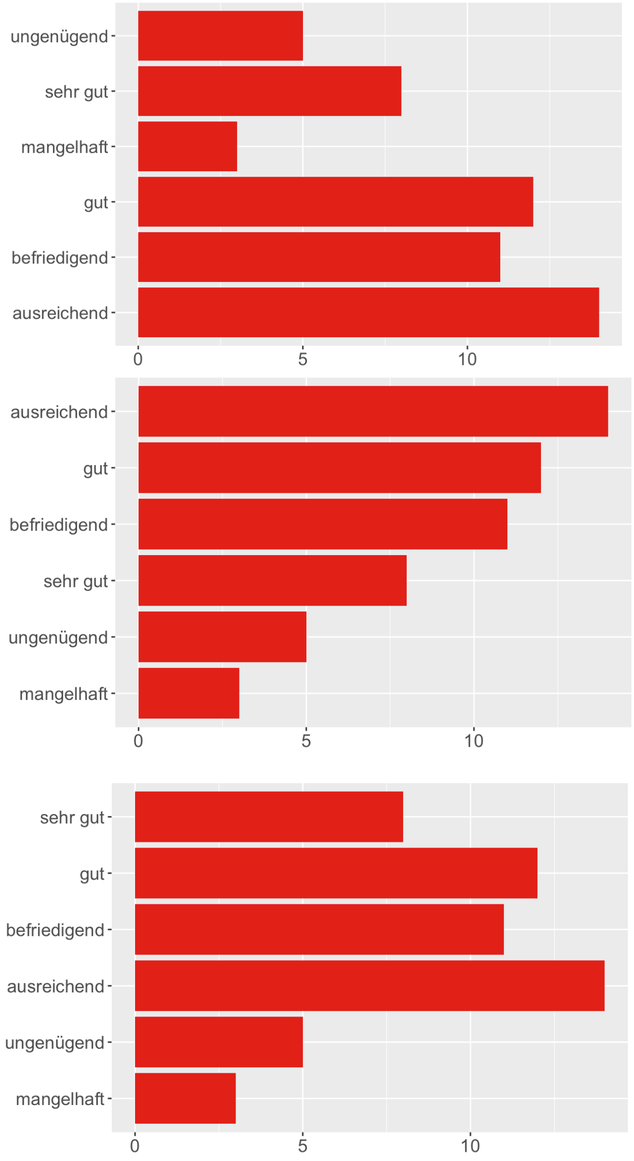
- Beware of selecting the proper scales for plotting!



Categorical Data

Barplots

- Random order
- Order by increasing / decreasing counts
- Natural order of ordinal data



Mode

Numerical variables

- Numerical data are instances of underlying **random variable**

- Random variables *X* have
  - **Density** distributions *p(X)*
  - **Expectation values** *E(X)*
  - **Variances** *Var(X)*

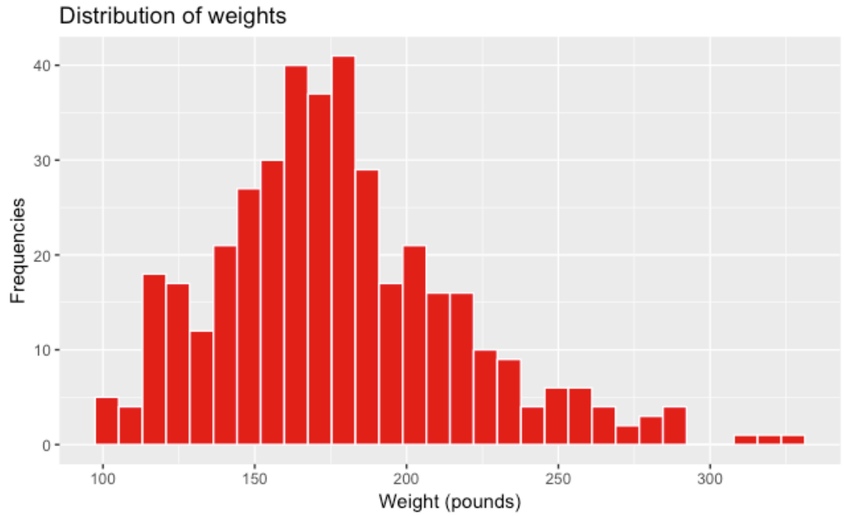
random variable X

1, 5, 4, 2, 6, ...

instances x_i

Numerical data
Histograms

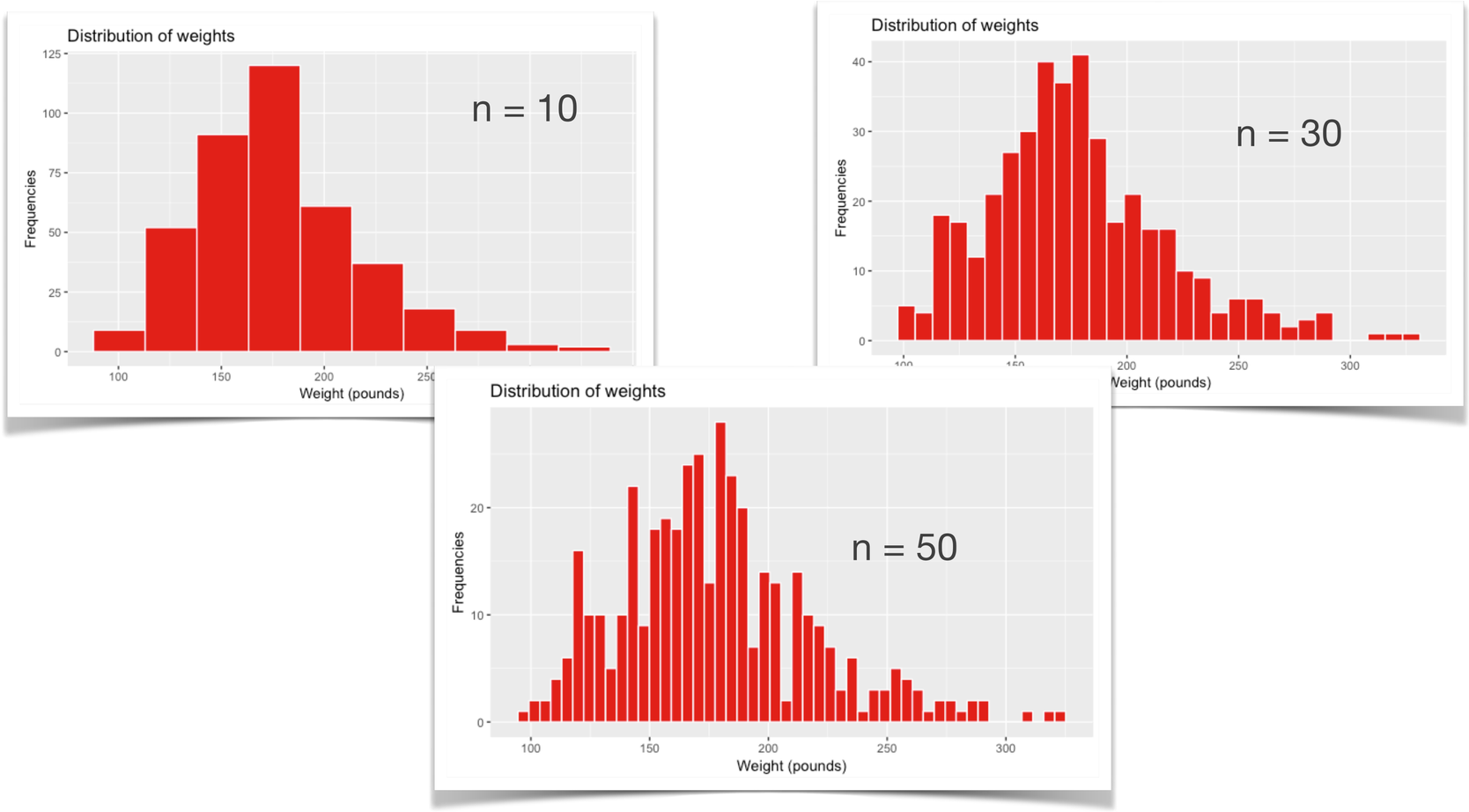
- Categorical data → counts (**barplot**)
- Numerical data → counts within intervals (**histogram**)
  - define **discrete intervals** for numerical variable → **ordinal variable** e.g. [0,10), [10,20), [20,30), …
  - count occurences within intervals and plot
- histograms represent the **distribution of the variable**



Numerical data

Histograms

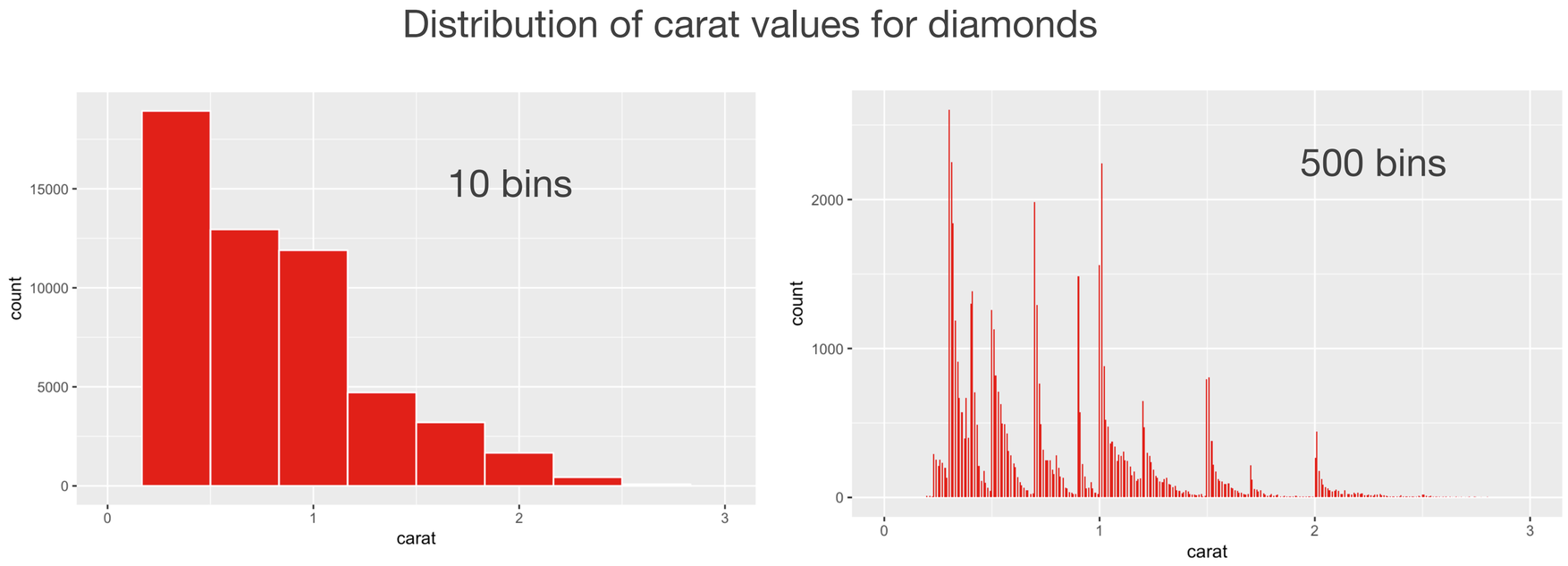
- Right choice of interval depends on the data type
- **Number of bins has a strong impact on appearance of plot!**



Numerical data

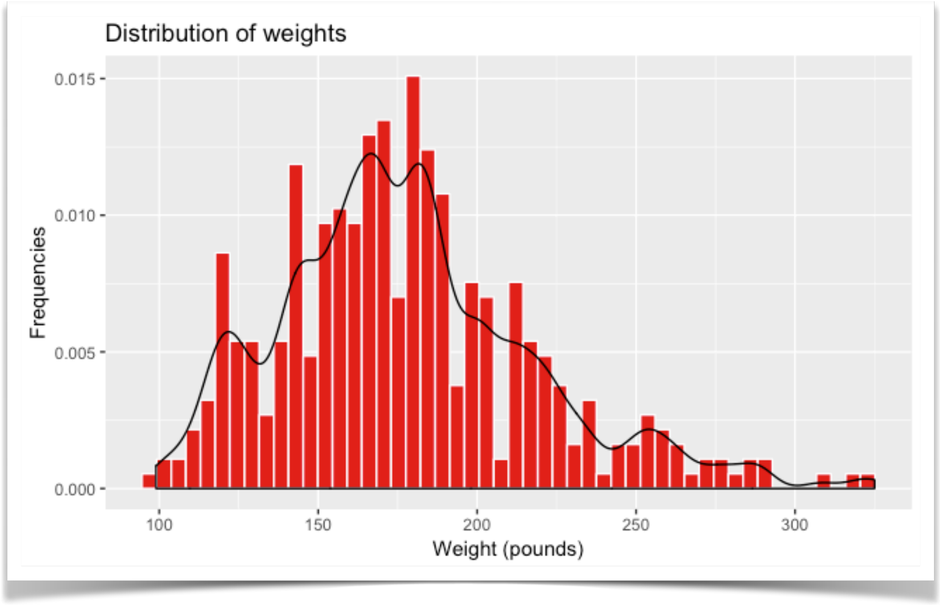
Histograms

- pattern becomes visible at high resolution
- peaks around integer values (why?)

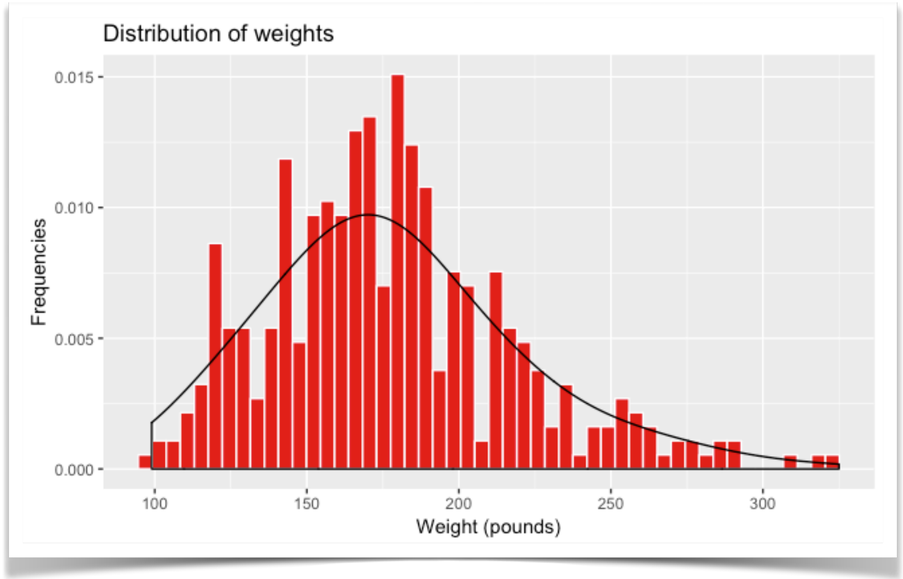


Numerical data
Histograms

- Frequency distributions (= histograms) can be shown using a smoothed **density curve**
- Smoothing depends on the bandwidth (~size of the interval over which to smooth)



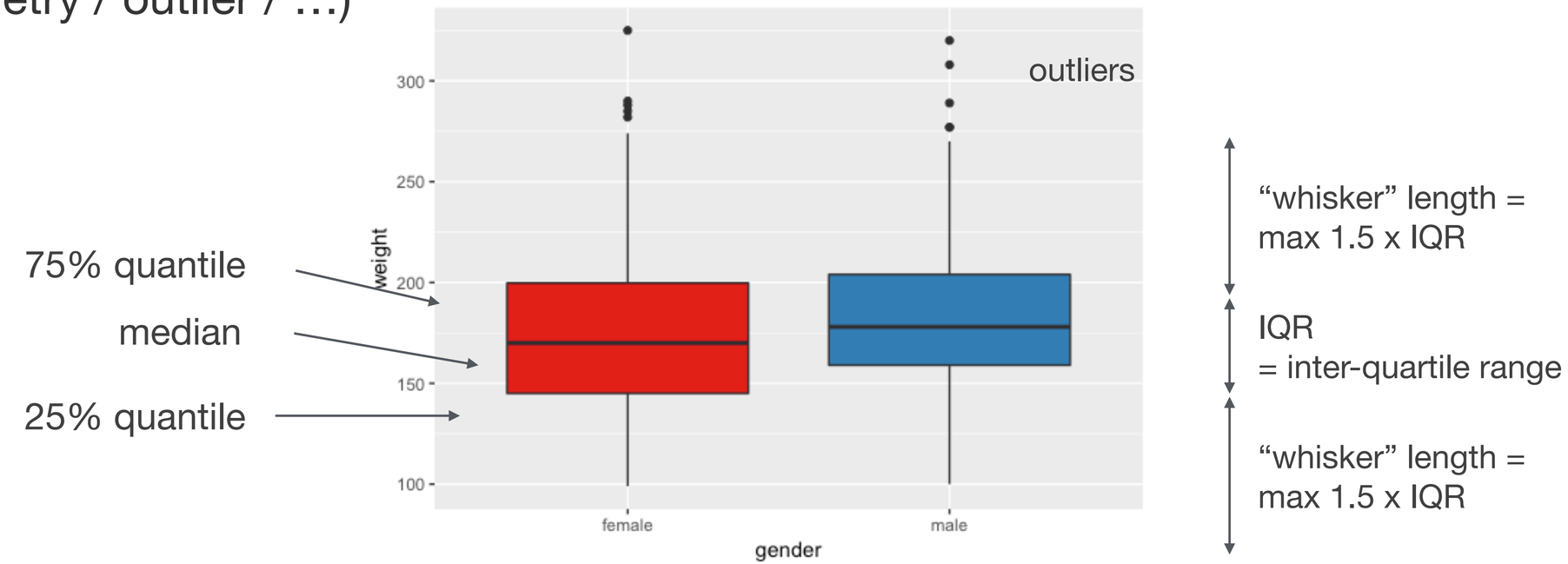
bandwith = 5



bandwith = 20

Numerical values
boxplots

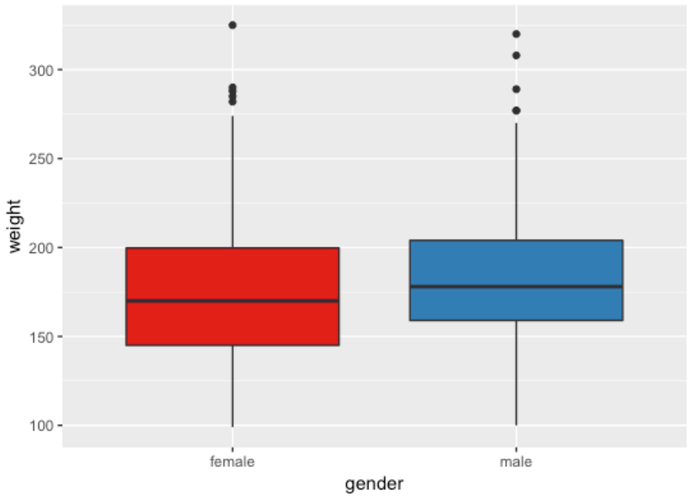
- Boxplot give an indication on the shape of the distribution (median / symmetry / outlier / …)



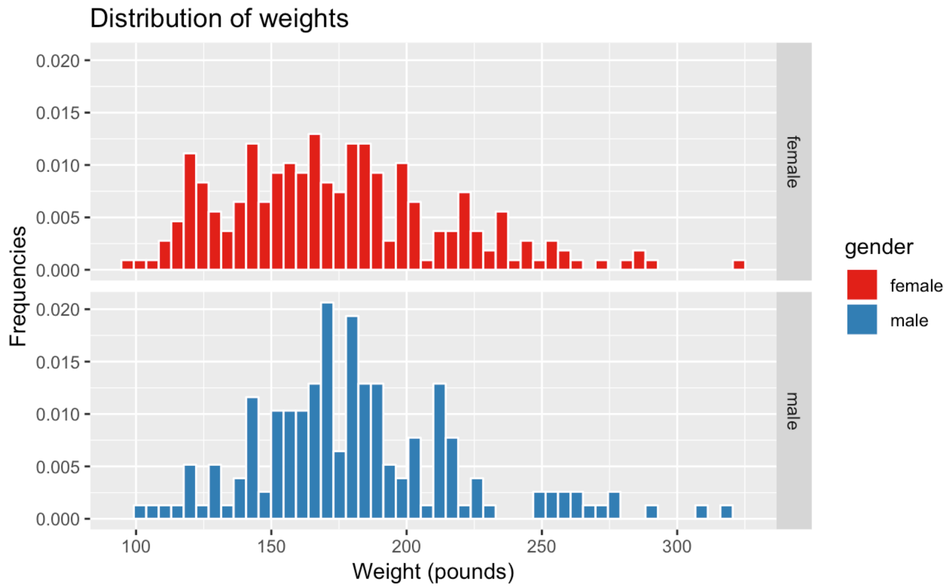
- Upper Whisker extend to the last point that is not larger than Q75 + 1.5*IQR
- Lower Whisker extends to the last point that is not smaller than Q25-1.5*IQR
- Whisker does not go not beyond maximun or minimum value ! (Hence both whisker can have different length < 1.5 x IQR)

Numerical values
boxplots

Boxplot

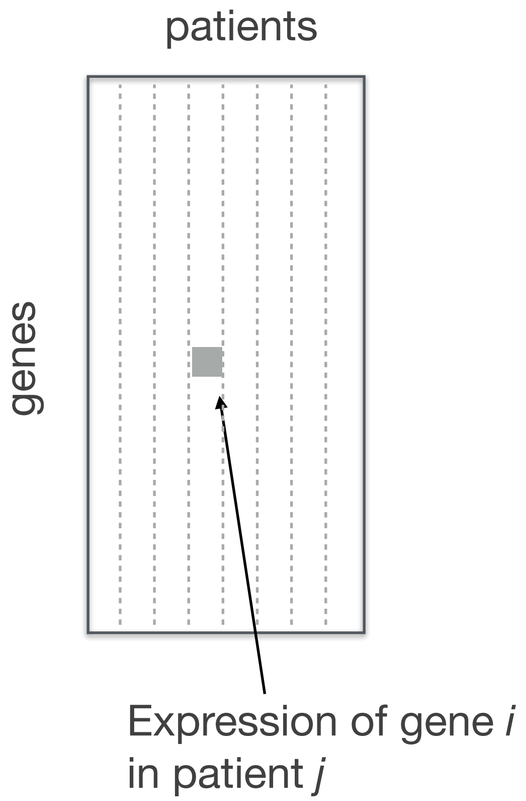


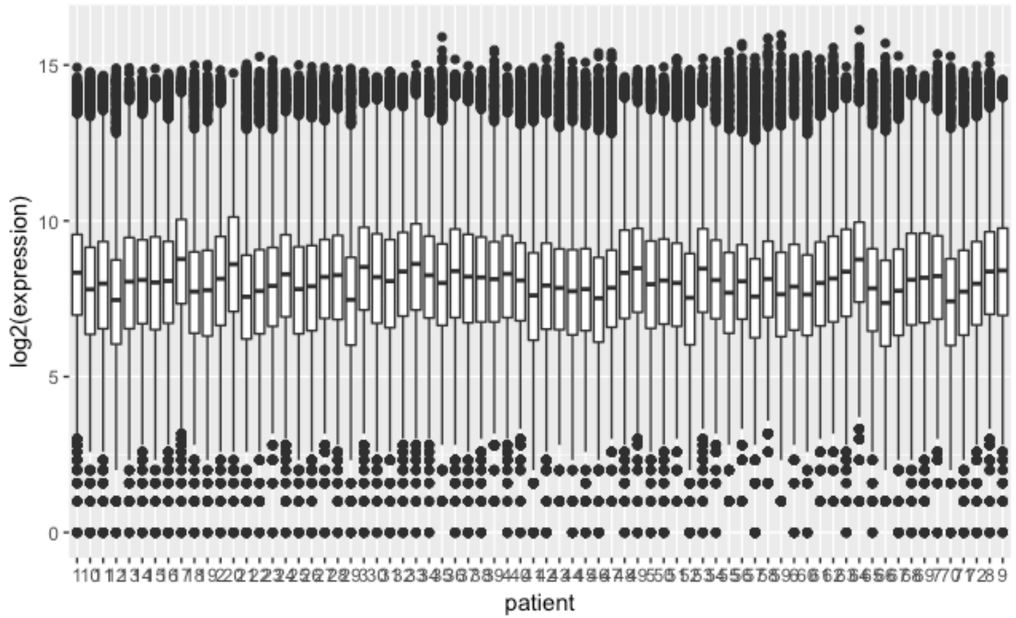
Histogram



Boxplots summarize the properties of the distribution
Usefull to compare many distributions side-by-side

Numerical values
boxplots



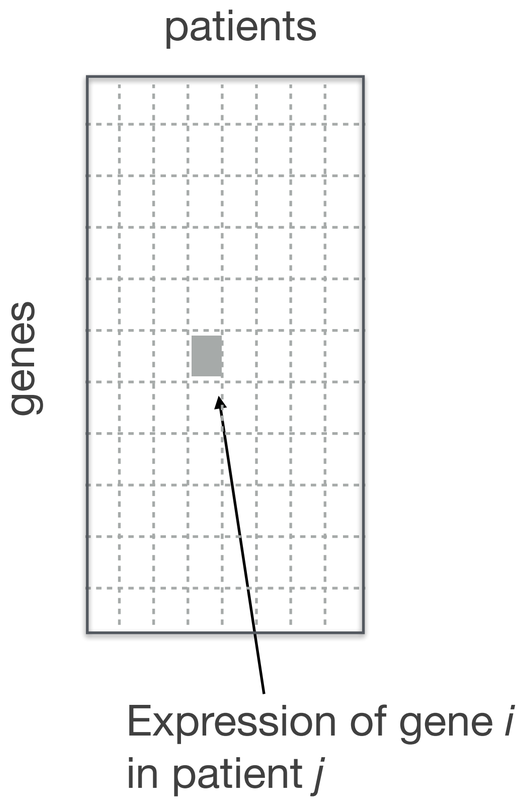


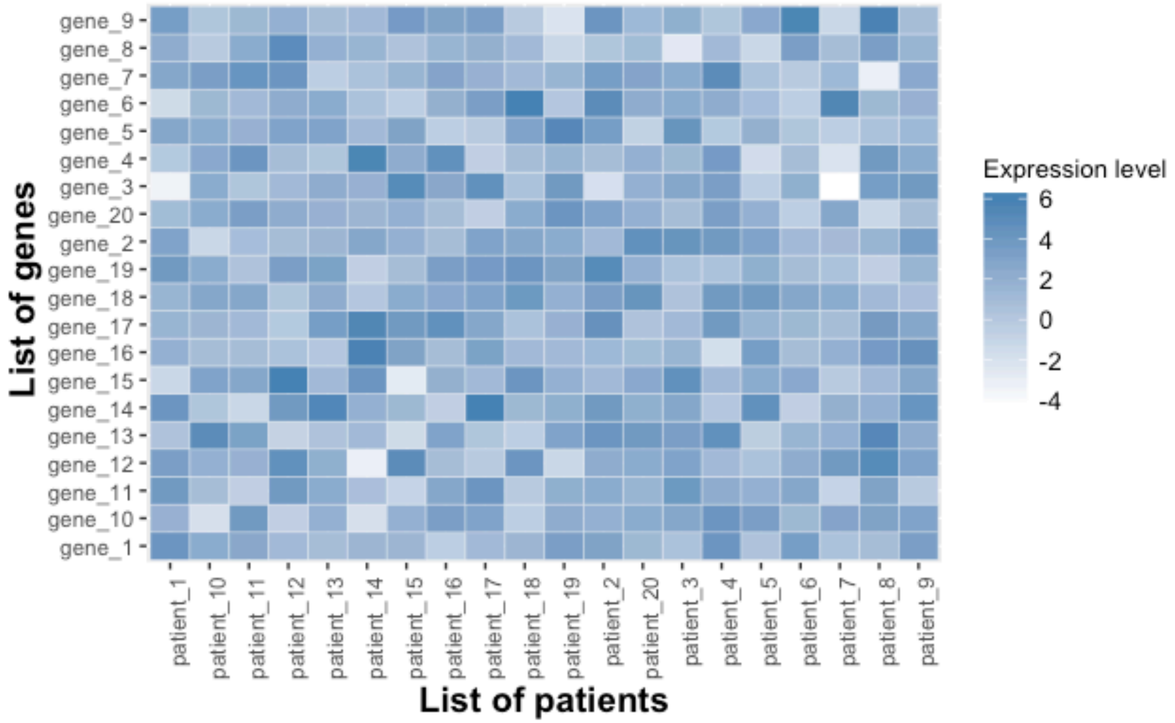
Question: do some patients have a different median gene expression?

→ *values for individual genes are lost in this type of plot!*

Numerical Data :
heatmaps

Heatmaps display numerical values in a data matrix using a color scheme

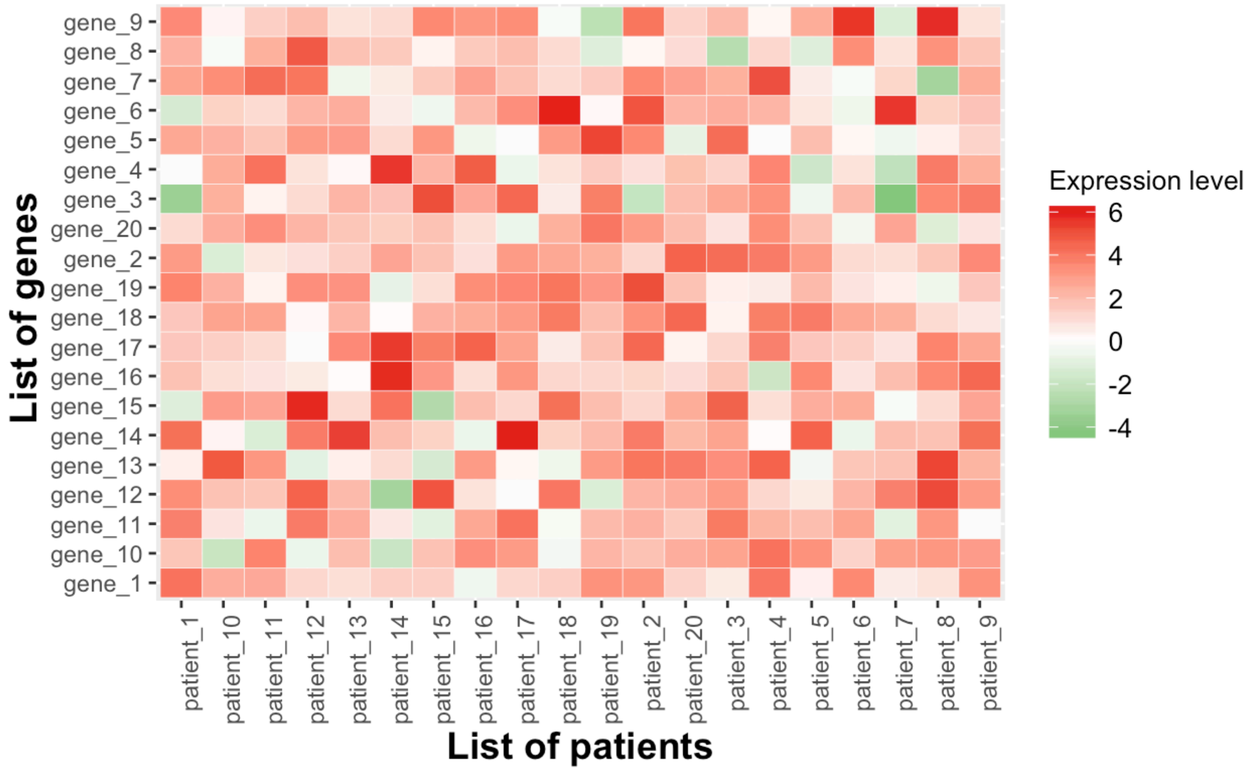




Question: do some genes in some patients have a different gene expression?

Numerical Data :
heatmaps

use symmetrical color scales for symmetrical ranges !

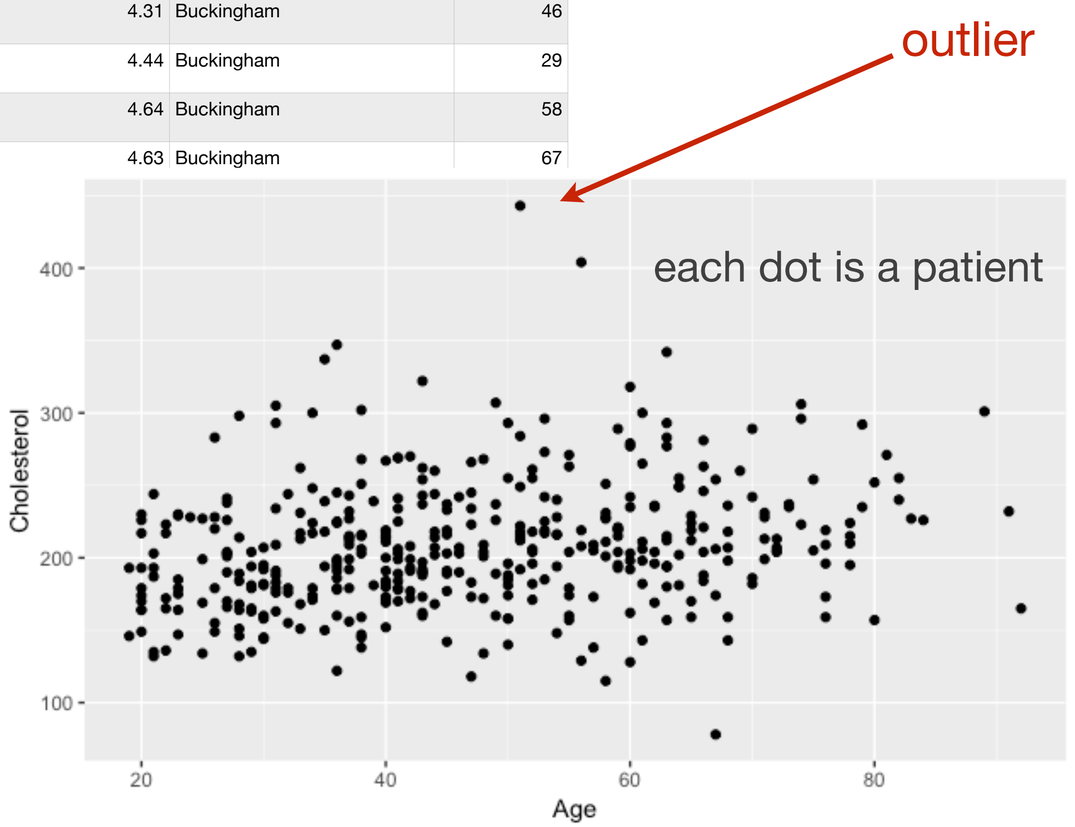


Numerical data

comparing variable with scatter plots

any relation between age and cholesterol?

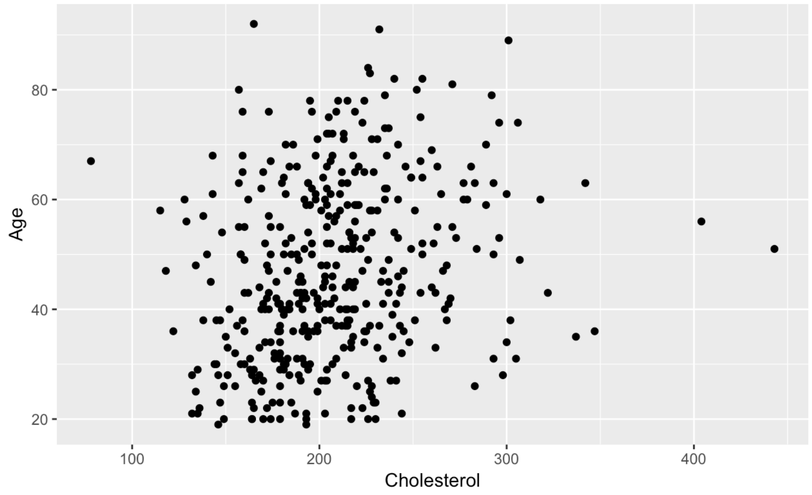
id	chol	stab.glu	hdl	ratio	glyhb	location	age
1000	203	82	56	3.60	4.31	Buckingham	46
1001	165	97	24	6.90	4.44	Buckingham	29
1002	228	92	37	6.20	4.64	Buckingham	58
1003	78	93	12	6.50	4.63	Buckingham	67
1005	249	90	28	8.90			
1008	248	94	69	3.60			
1011	195	92	41	4.80			
1015	227	75	44	5.20			
1016	177	87	49	3.60			
1022	263	89	40	6.60			
1024	242	82	54	4.50			
1029	215	128	34	6.30			
1030	238	75	36	6.60			
1031	183	79	46	4.00			
1035	191	76	30	6.40			
1036	213	83	47	4.50			
1037	255	78	38	6.70			



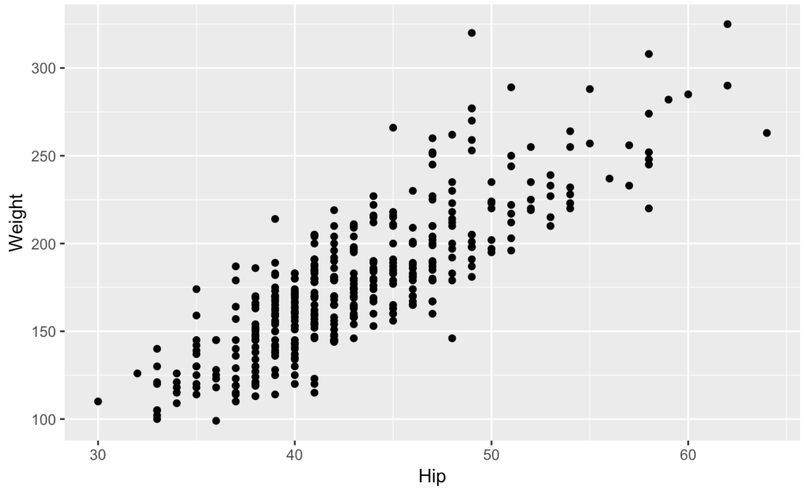
Numerical data

comparing variable with scatter plots

- we will later **quantify** this relationship in terms of **covariance / correlation** and determine how **significant** this relationship is!



Weak relationship
between Cholesterol and age

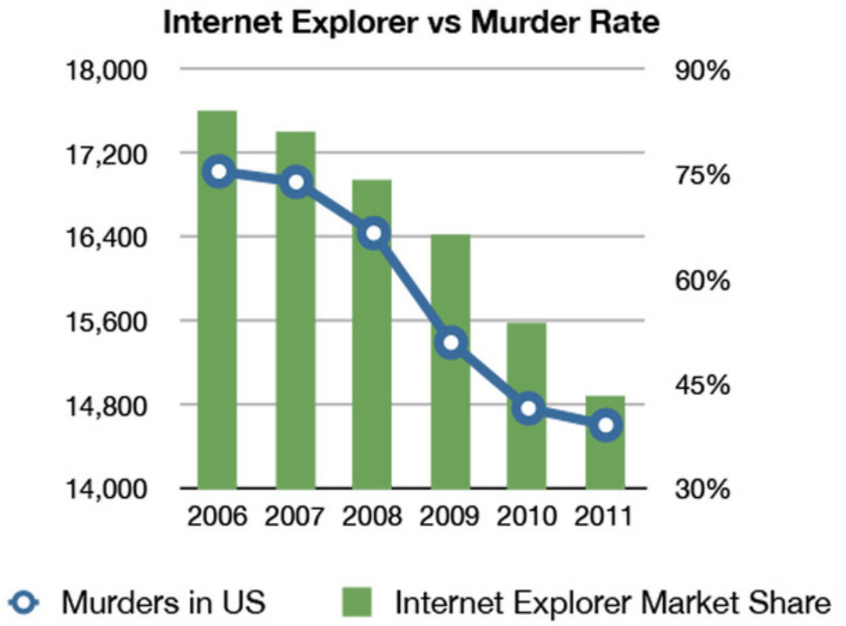


Strong relationship
between Hip and Weight

Numerical data

comparing variable with scatter plots

- Do not over-interpret scatter plots!
- Existence of relation between variables does not mean that there is a causal relationship between them!
- Correlation is NOT causality!!

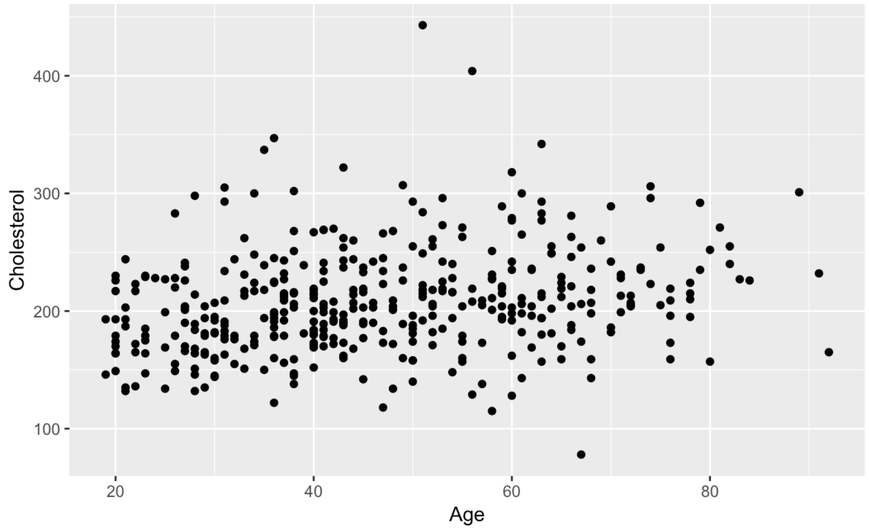


http://www.tylervigen.com/spurious-correlations

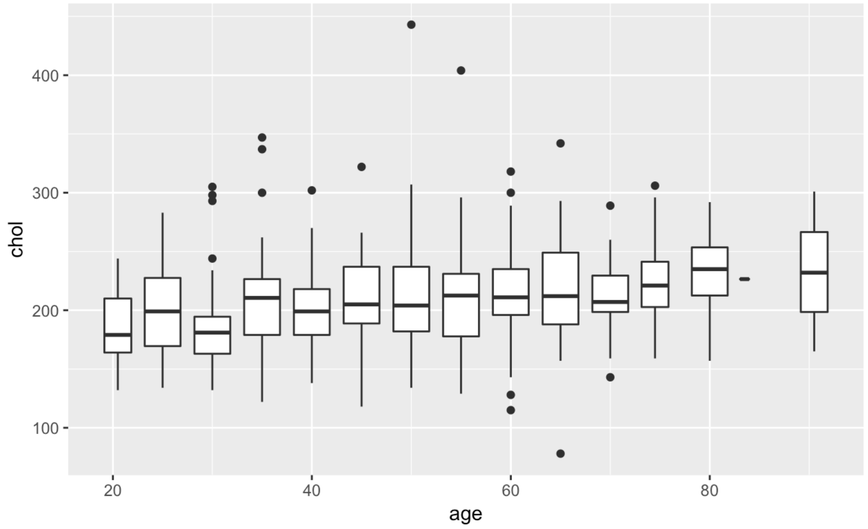
Numerical data

comparing variable with scatter plots

- A **continuous numerical** variable can always be transformed into an **ordinal categorical** variable through binning



Age = continuous variable



Age = ordinal variable (bins)

Numerical data

comparing variable with scatter plots

- Additional categorical / numerical variables can be added using color, shape, size of dot ,…

Summary on visualization

Single variable plot
plot counts

	type of plot
continuous variable	histogram
categorical variable	barplot

Two variable plot
plot relationship

	continuous variable	categorial data
continuous variable	scatter plot	boxplot
categorical variable		heatmap

3. Describing data

Descriptive statistics

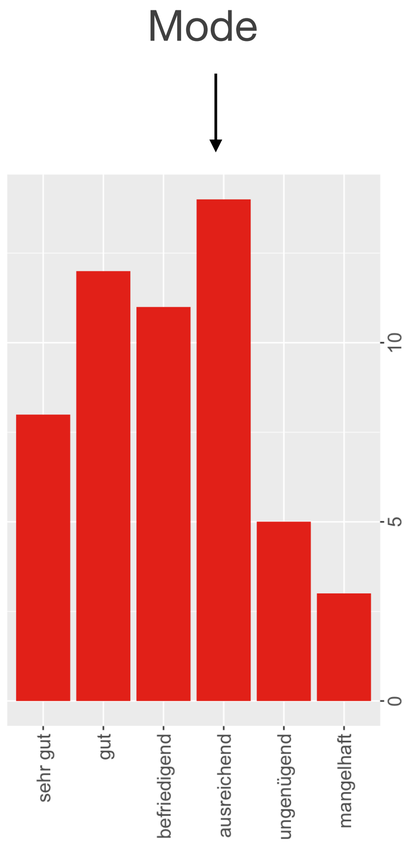
- Datasets represent samples taken from a larger population
- Example
  - diabetes dataset: sample representing a group of diabetes patients
  - gene expression dataset: samples (e.g. 10 mice before and after treatment) represent a group of mice extracted from a larger population
  - opinion polls: opinion of 1000 people extracted from a more general population

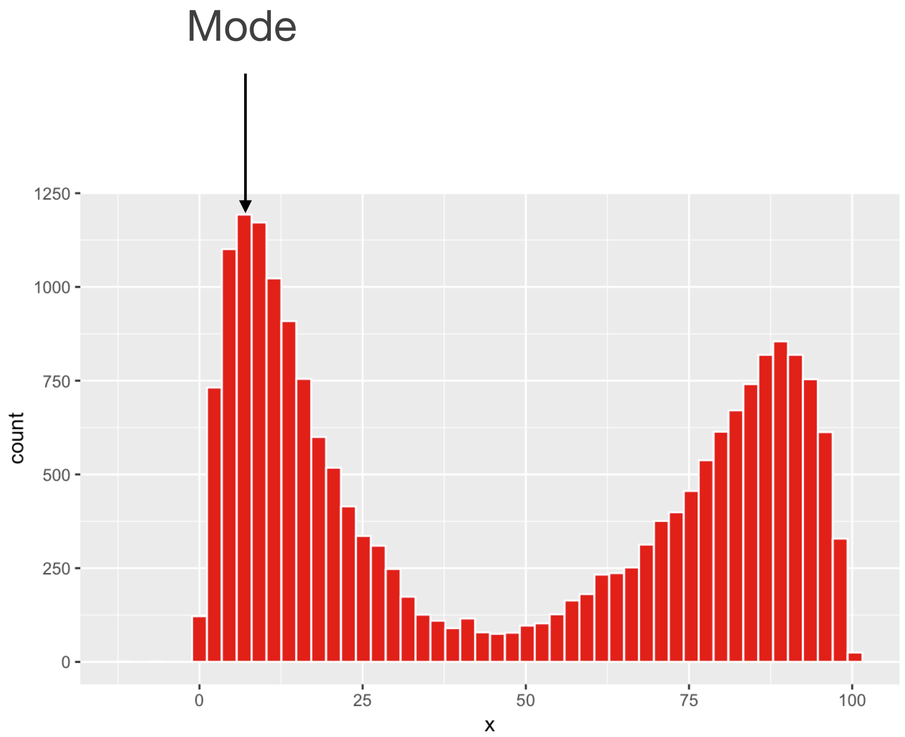
$$\{x_1, x_2, \ldots, x_N\}$$

Descriptive statistics provides a description of the sample dataset, without making any extrapolation to the more general population!

Computing the average

- Categorical data: **mode** = category with the highest count





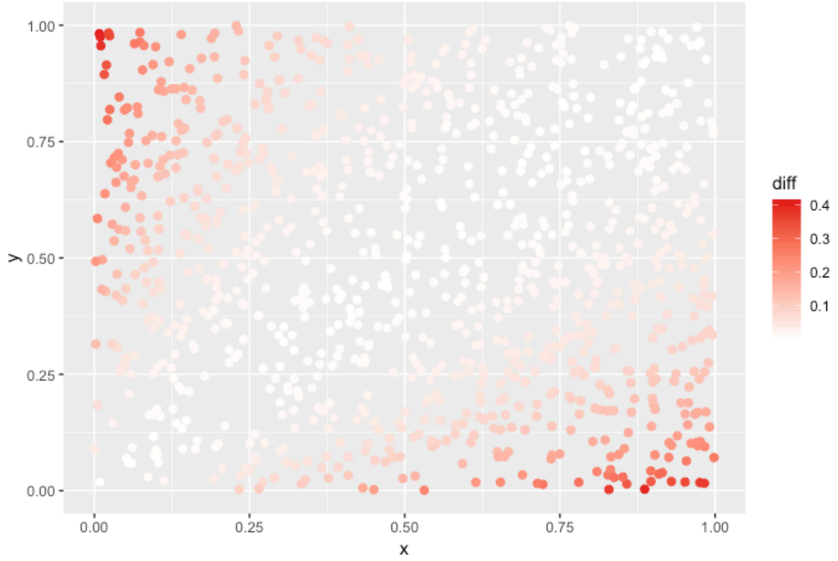
Such a distribution is
called a ***"bimodal distribution"***

Computing the average

- Numerical data: **mean** of the data

$$\text{arithmetic mean} = \frac{1}{n}\sum_{i=1}^{n} x_i \qquad \text{geometric mean} = \left(\prod_{i=1}^{n} x_i\right)^{1/n}$$

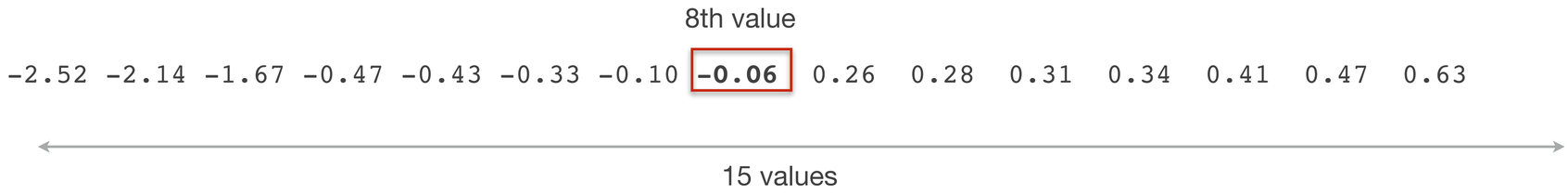
- arithmetic mean is always greater/equal the geometric mean



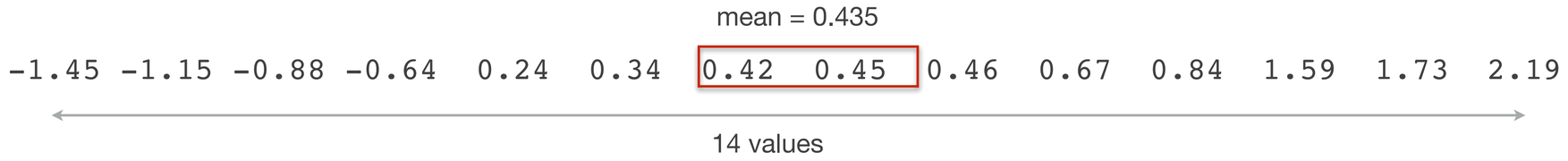
difference arithm. - geom.
for 2 values (x,y)

Computing the average

- Numerical data: **median**
- median = numerical value which splits the dataset into two equal parts
  - 50% of the values are larger, 50% of the values are smaller that the median value
- **Odd** number of values : sort and take the middle one

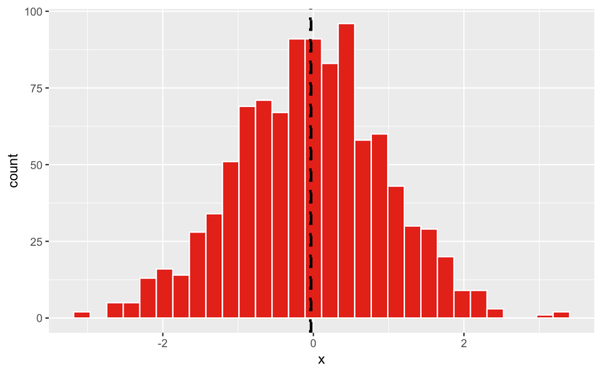


- **Even** number of values: sort and mean of the middle 2 values

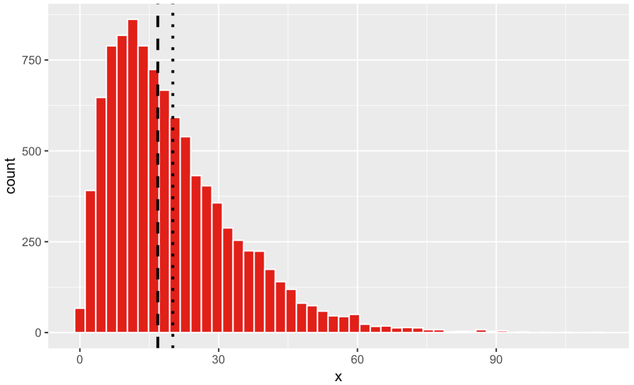


Mean / Median

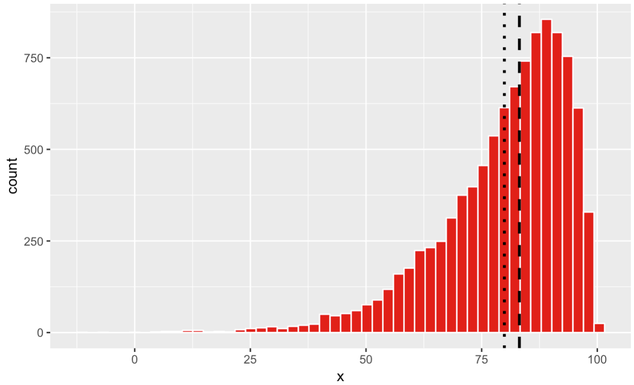
Which line is the median, which is the mean ?



symmetrical distribution
median ~ mean



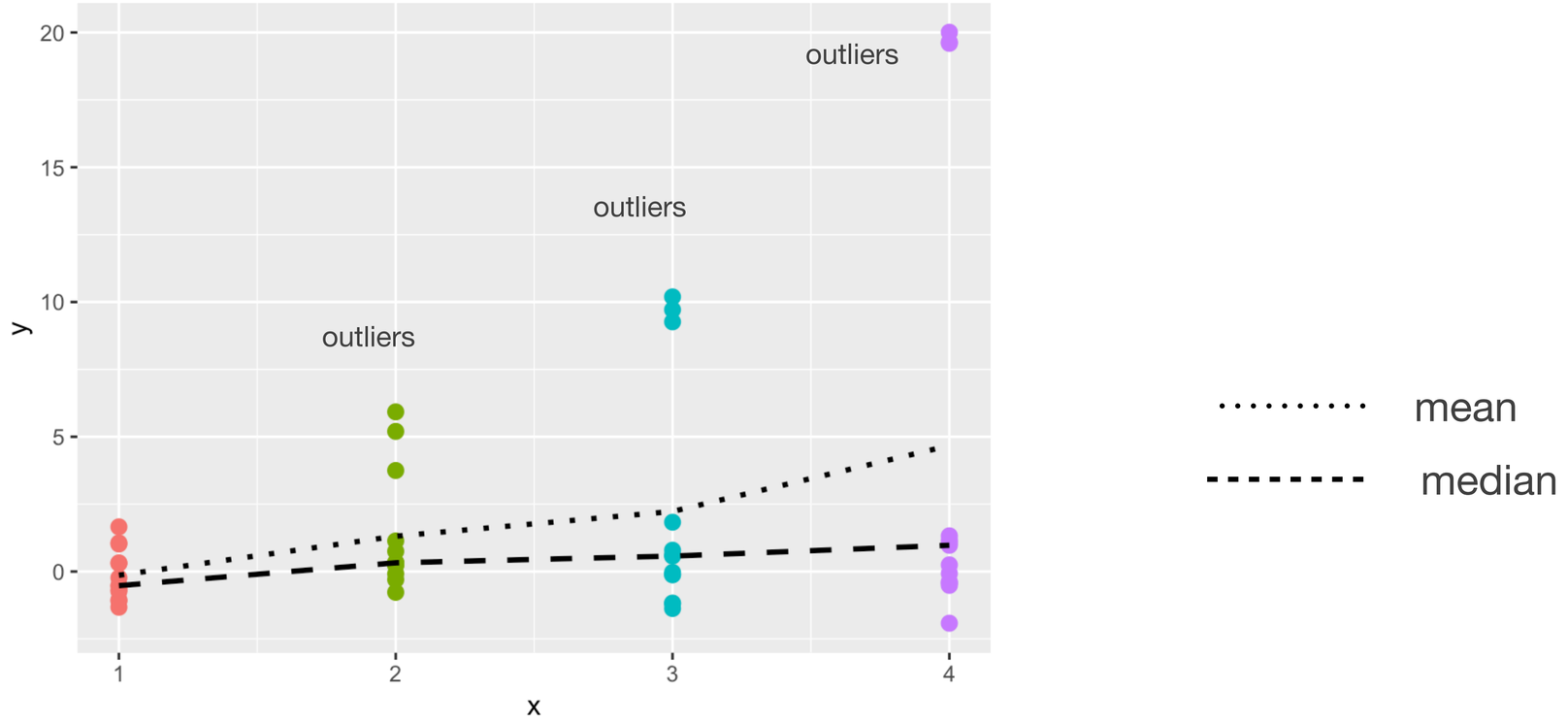
right-skewed distribution
mean > median



left-skewed distribution
mean < median

Mean vs. median

- 4 groups, with 3 outliers with increasing effect
- *median is more robust to outliers than mean*



Measuring data spread

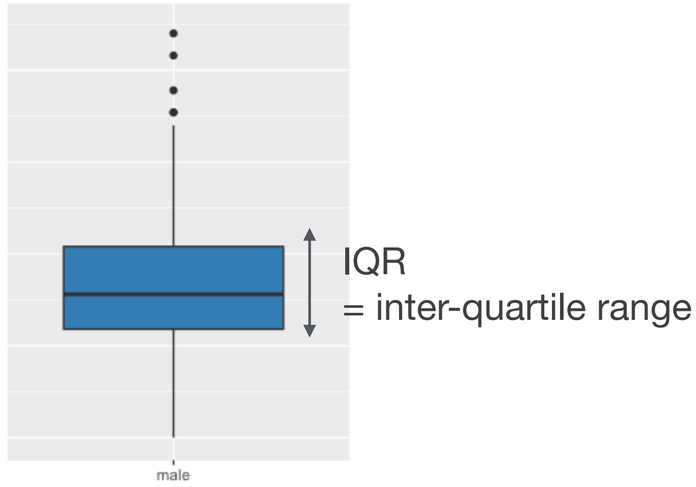
- How much spread is there in the dataset?
- **Sample variance** :

$$Var(x) = \frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})^2 \qquad \bar{x} = \text{mean}$$

- **Standard deviation** :

$$s_x = \sqrt{Var(x)}$$

- **Interquartile Range (IQR)** : difference between the 75% and 25% percentile

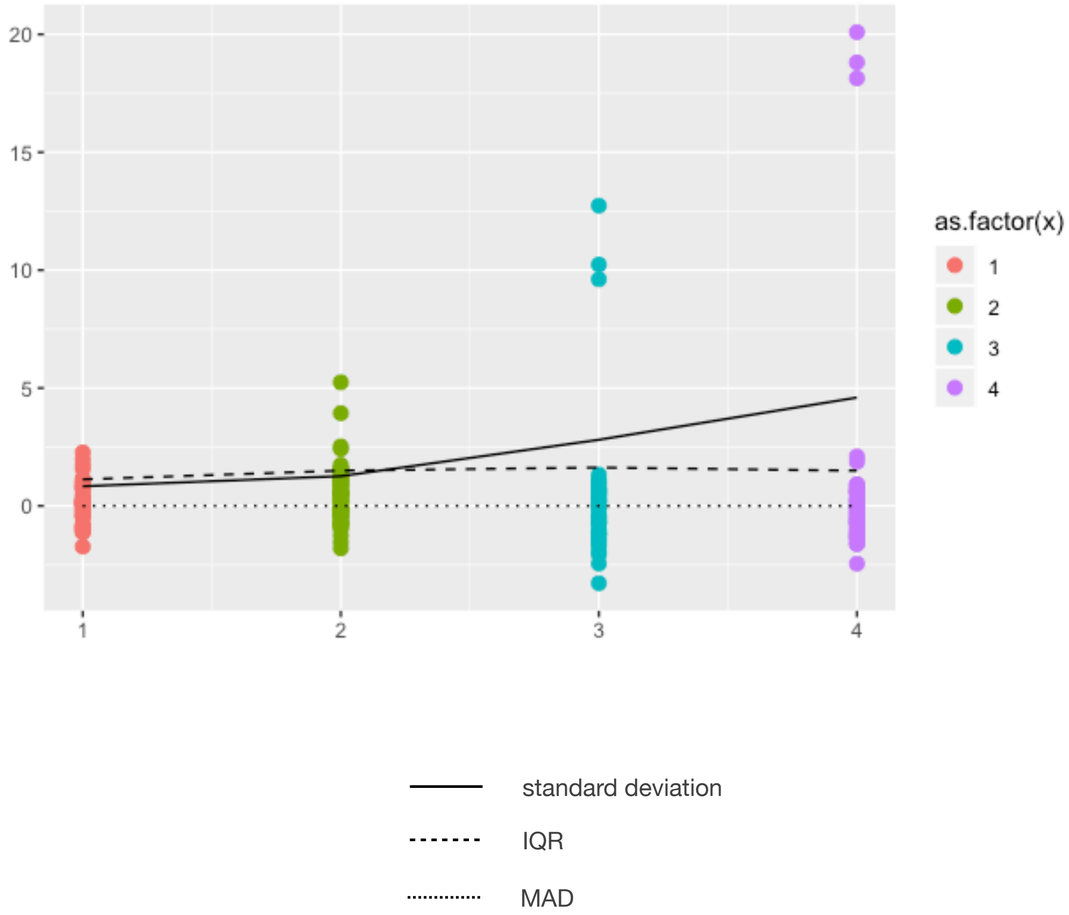


Measuring data spread

- **Median absolute deviation (MAD)**
  median absolute deviation of the data points from the median…

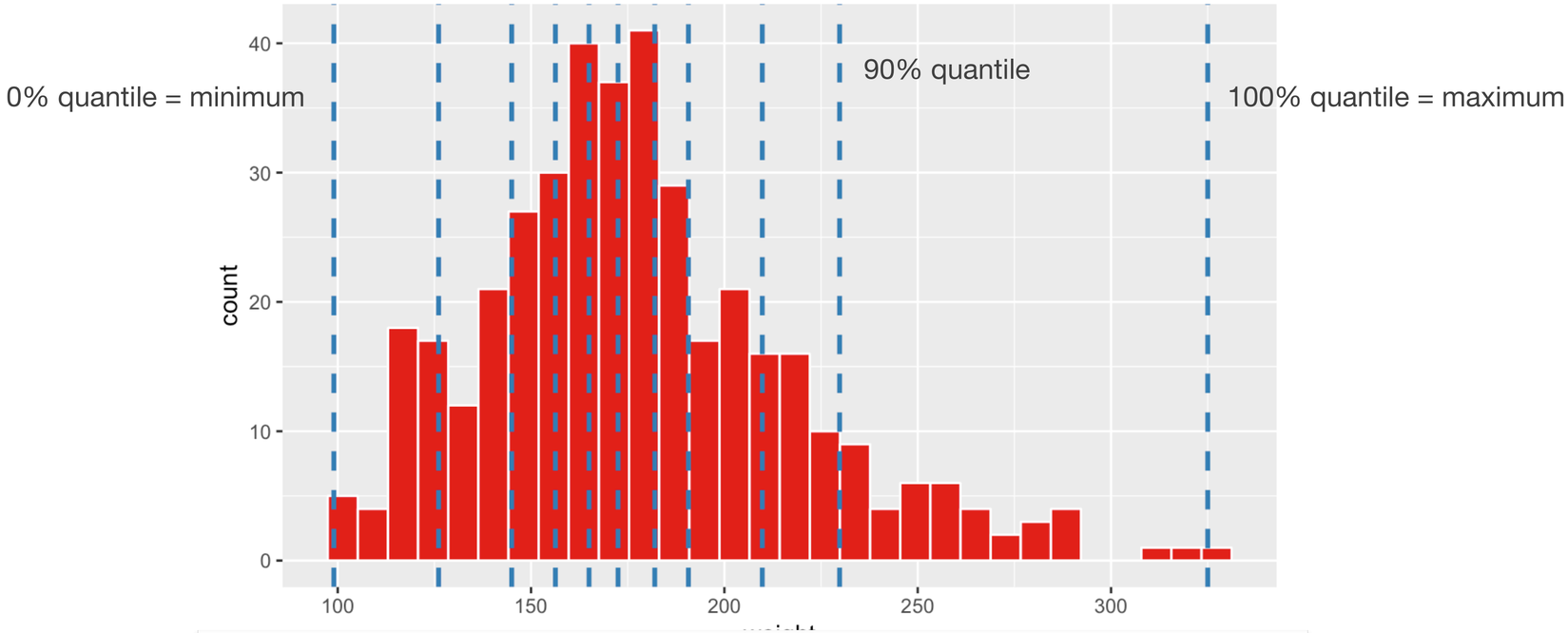
- How to compute?
  1. compute the median *med(x)* of the dataset
  2. compute the absolute deviation of all points from the median
     $d_i = |x_i - med(x)|$
  3. compute the median of the d_i

- IQR and MAD are less sensitive to outliers than standard deviation !



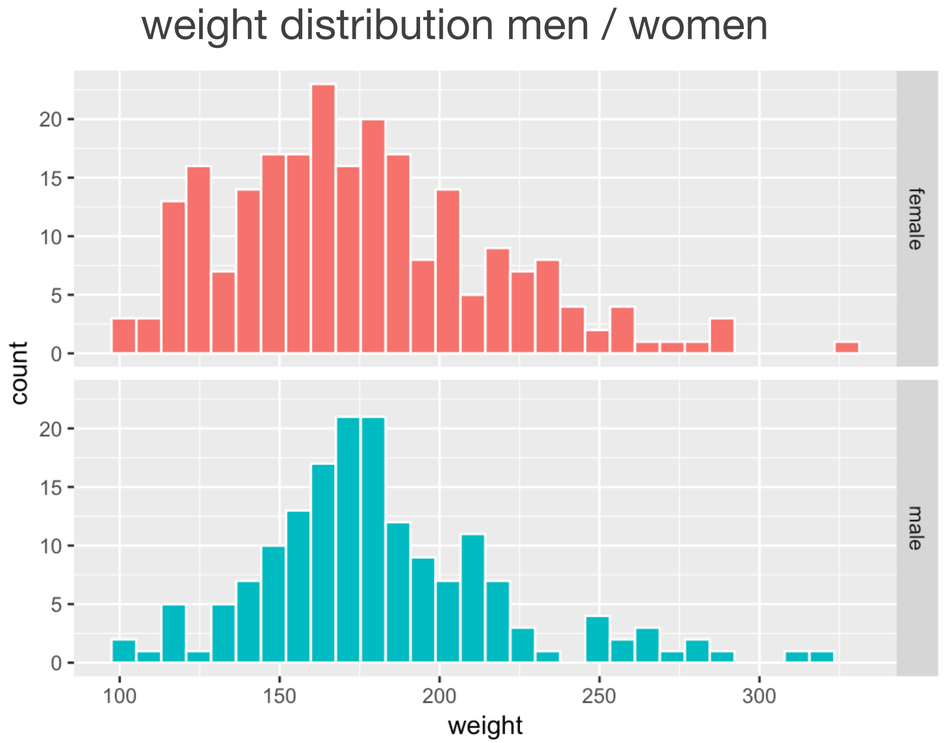
Quantiles

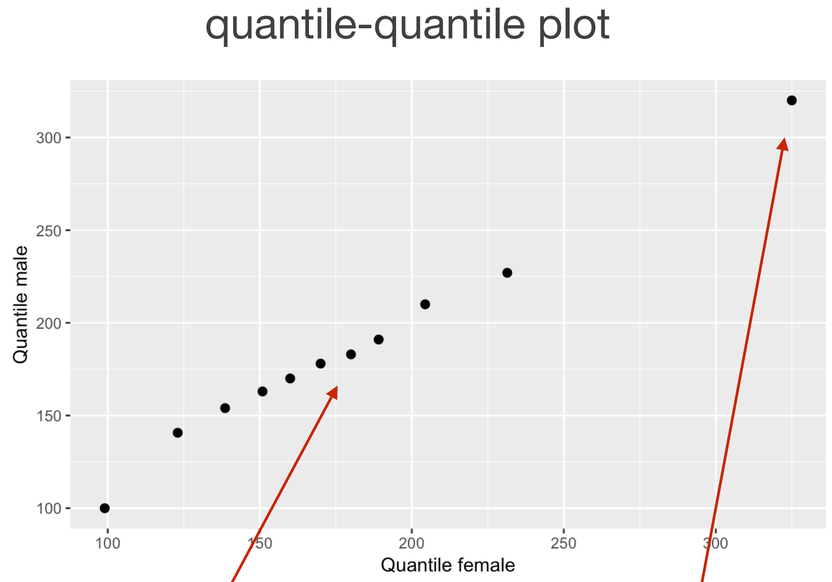
- **p-quantile** = value such that p% of the values in the data set are smaller
- distribution of quantile value given an indication about the **shape** of the distribution



0%	10%	20%	30%	40%	50%	60%	70%	80%	90%	100%
99.0	126.1	145.0	156.3	165.0	172.5	182.0	190.7	209.8	229.8	325.0

Quantile-quantile plots

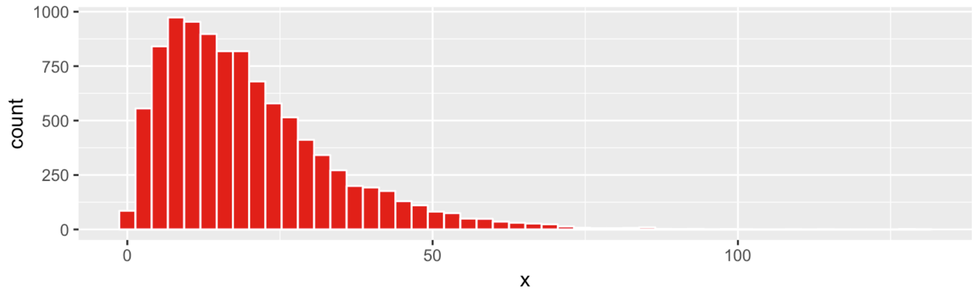


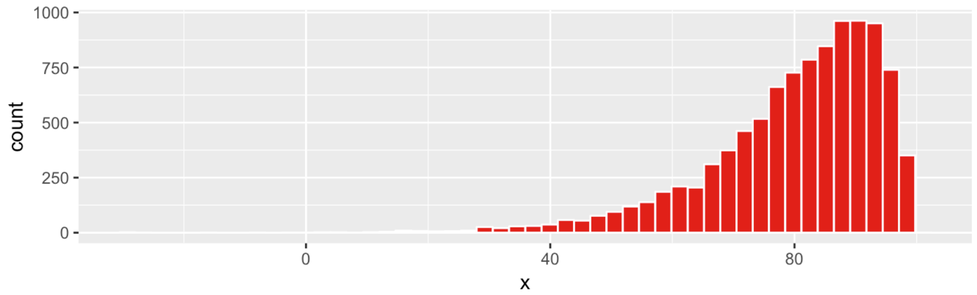


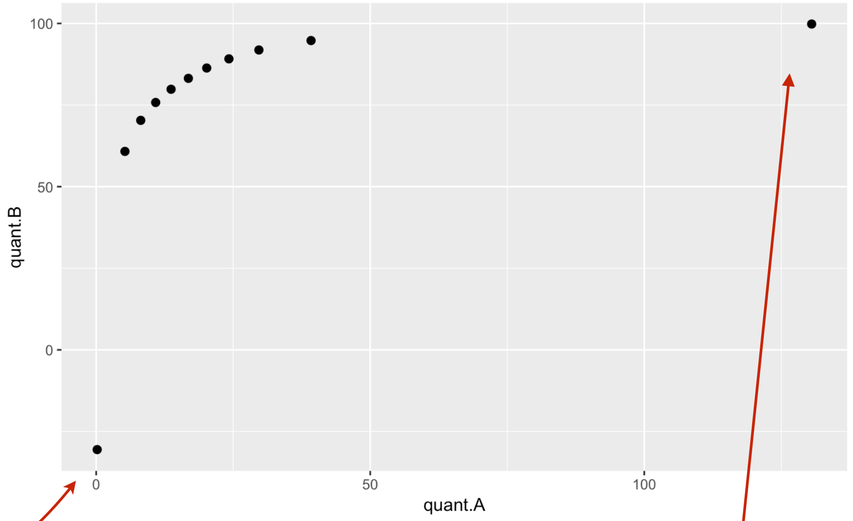
	0%	10%	20%	30%	40%	50%	60%	70%	80%	90%	100%
quant.female	99	123.0	138.6	150.9	160	170	180	189.1	204.4	231.4	325
quant.male	100	140.7	154.0	163.0	170	178	183	191.0	210.0	227.0	320

if QQ-plot is a straight line, then both distributions have a similar shape (up to translations / dilatation)

Quantile-quantile plots



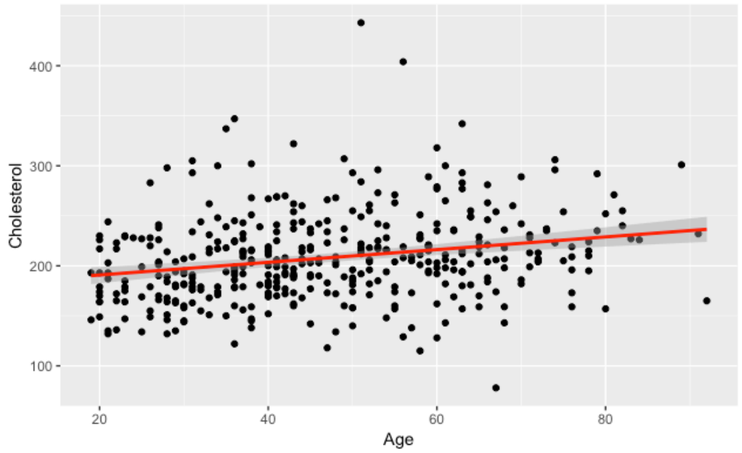


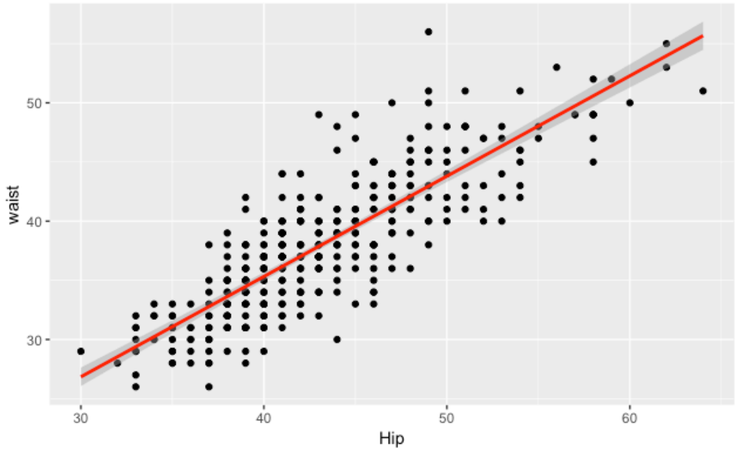


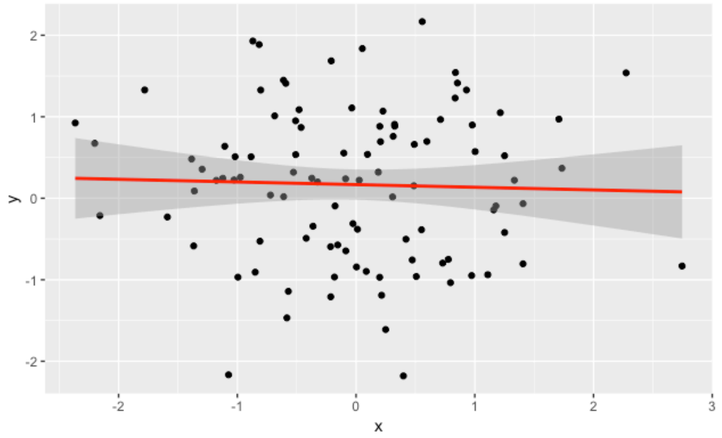
	0%	10%	20%	30%	40%	50%	60%	70%	80%	90%	100%
quant.A	0.18	5.24	8.13	10.85	13.66	16.82	20.16	24.22	29.69	39.20	130.55
quant.B	-30.55	60.80	70.31	75.78	79.84	83.18	86.34	89.15	91.87	94.76	99.82

Relation between numerical variables

- How easy is it to draw a line through a scatter plot?







Relation between numerical variables

- Variance:

$$Var(x) = (s_x)^2 = \frac{1}{N-1}\sum_{i=1}^{N}(x_i - \bar{x})^2$$

dimension: [x]²

- Covariance :

$$Cov(x,y) = \frac{1}{N-1}\sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y})$$

dimension: [x][y]

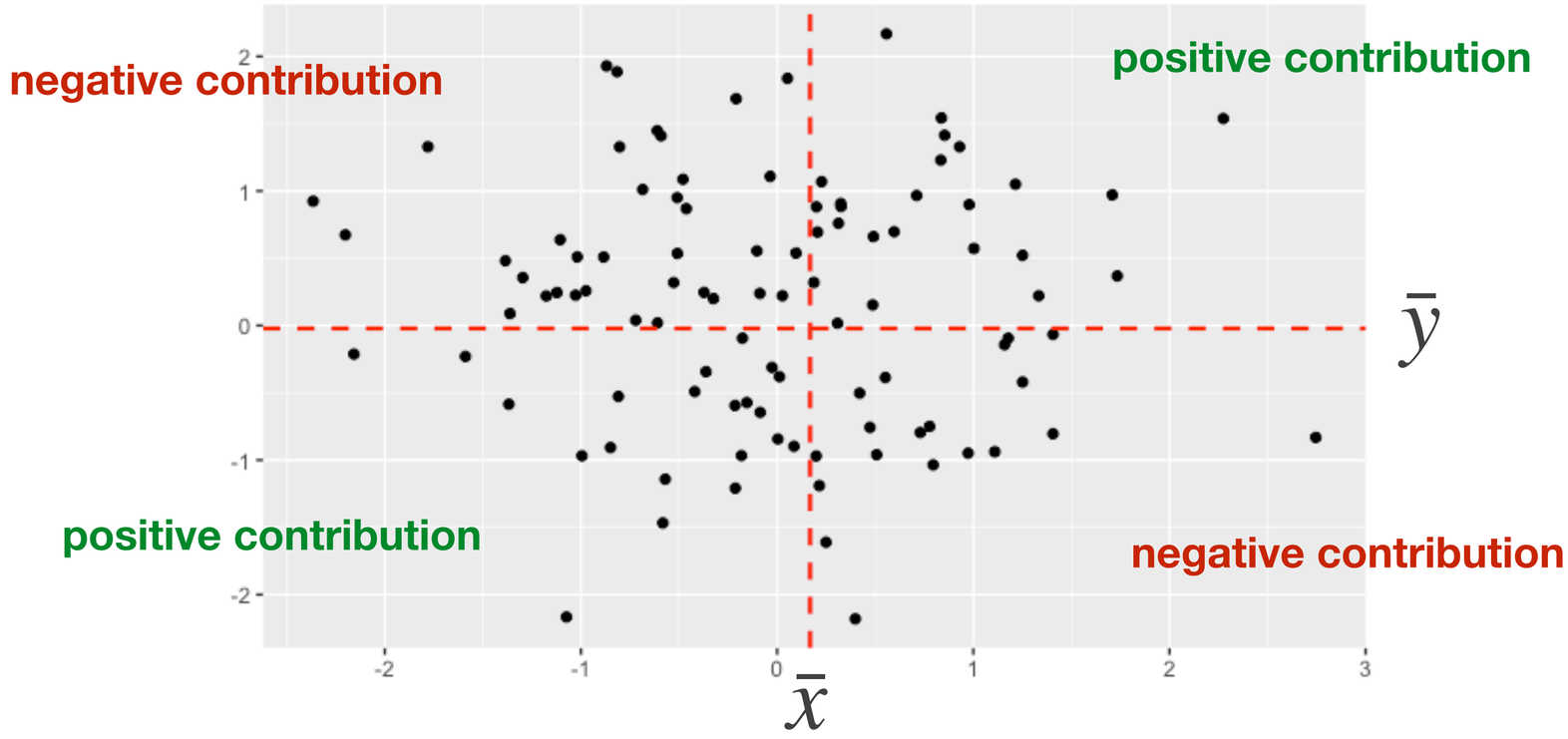
- Pearson Correlation :

$$Corr(x,y) = r = \frac{1}{N-1}\sum_{i=1}^{N}\frac{(x_i - \bar{x})}{s_x}\frac{(y_i - \bar{y})}{s_y}$$

dimension: none

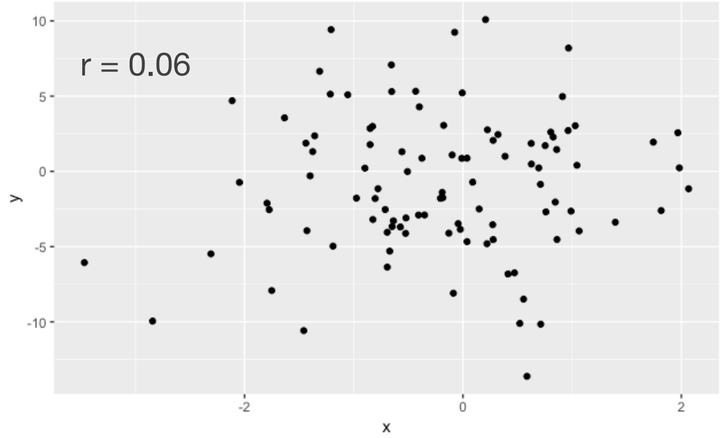
- Properties:
  - correlation is scale invariant, covariance is not!
  - cor(x,x) = 1
  - -1 =< cor(x,y) =< +1

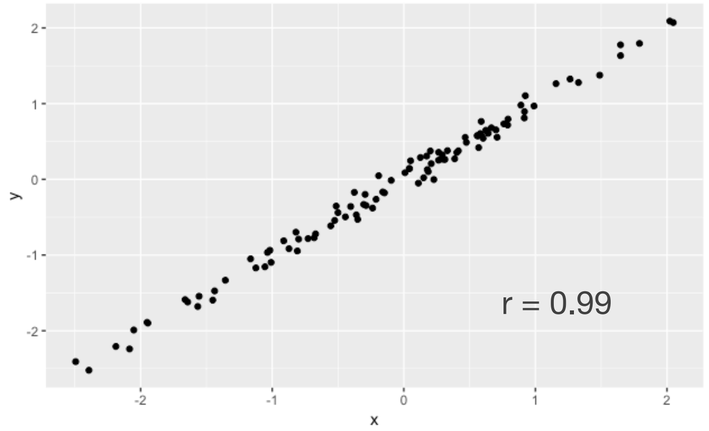
Relation between numerical variables

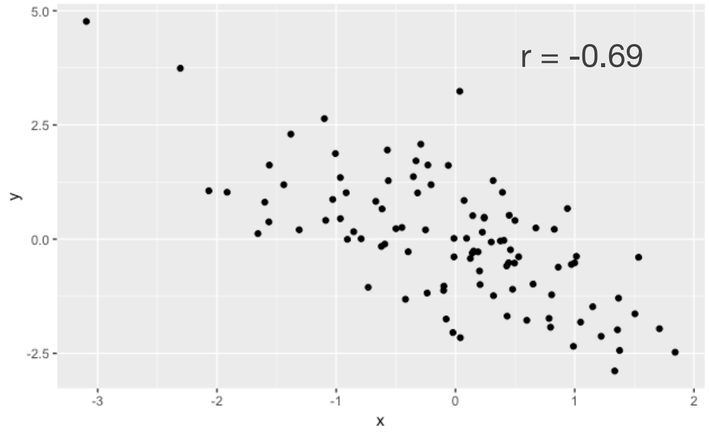


$$Corr(x,y) = \frac{1}{N-1}\sum_{i=1}^{N}\frac{(x_i-\bar{x})}{s_x}\frac{(y_i-\bar{y})}{s_y}$$

Relation between numerical variables

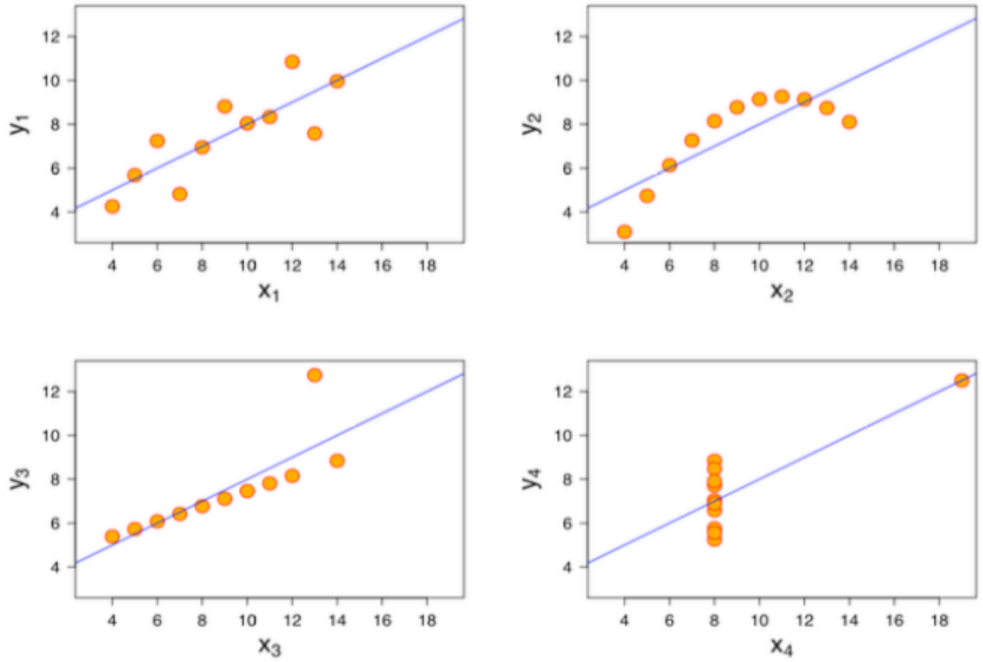






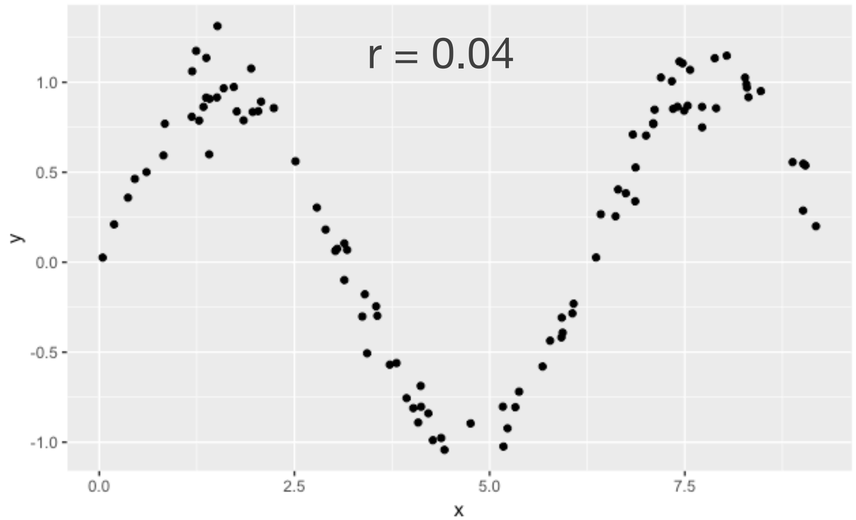
Relation between numerical variables

- Anscombe quartet is a dataset to highlight possible caveats of correlation measures
- Correlation is equal in all 4 cases (r = 0.82)



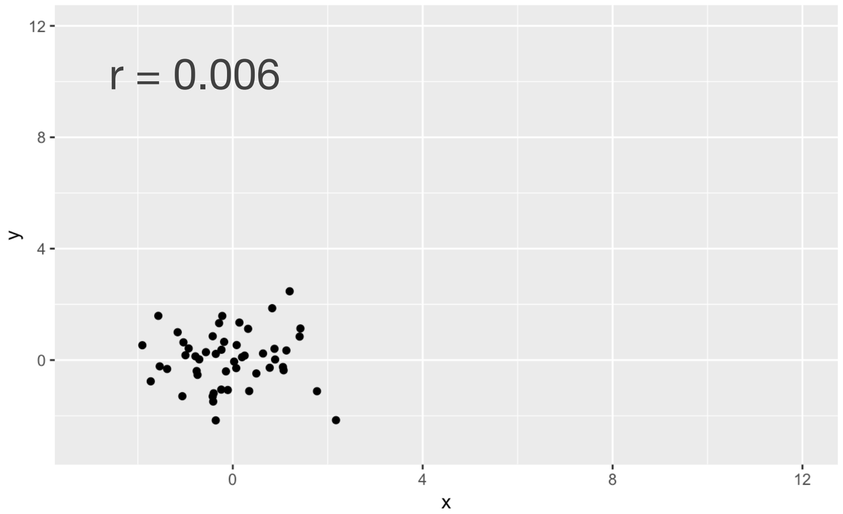
Relation between numerical variables

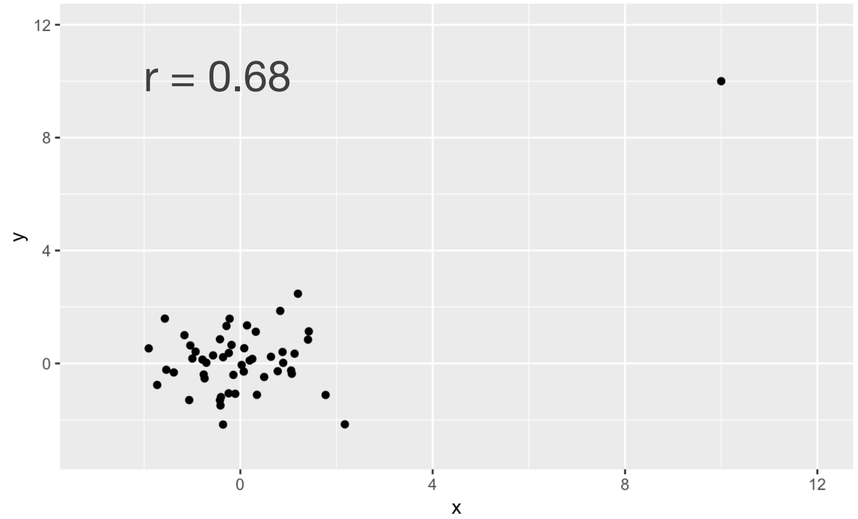
- Beware that correlation estimates a **linear relationship**!
- Weak correlation **does NOT mean** that there is no relation between both variables!



Correlation and outliers

- Pearson correlation is strongly influenced by outliers!
- we need to find robust measure which is insensitive to outliers.





Correlation and outliers

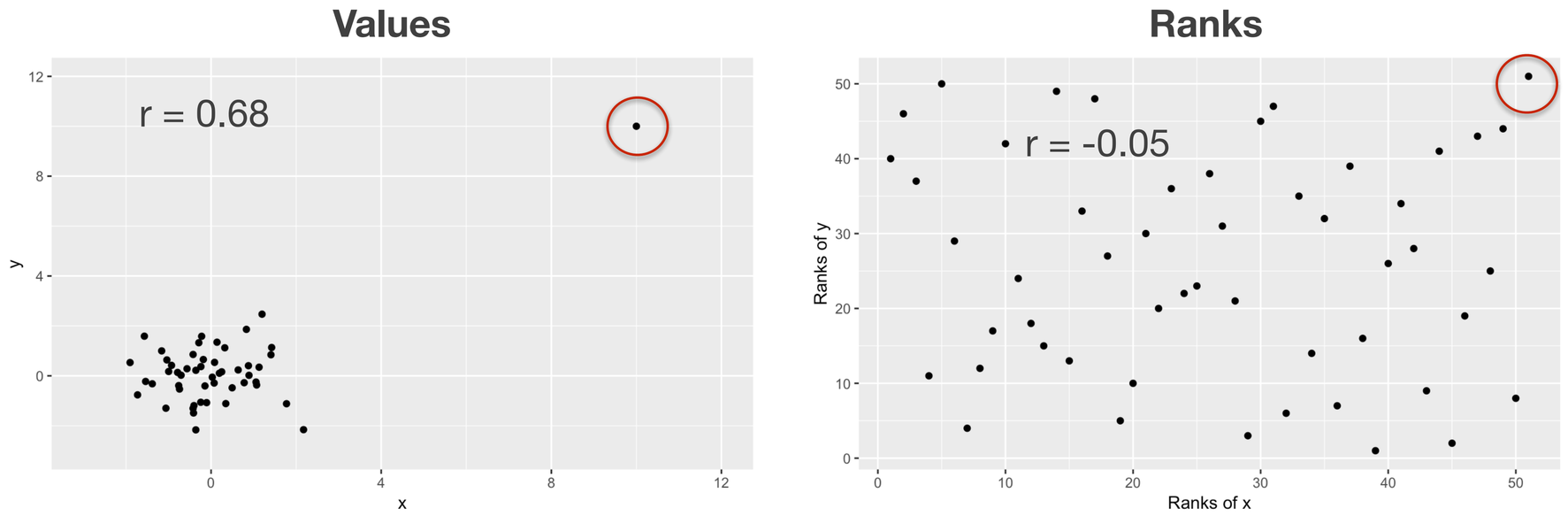
- Transform values to ranks

  x = {2.3 , -1.7 , 1.1 , -0.6 , 12.3} → R(x) = {4 , 1 , 3 , 2 , 5}

  y = {-2.7 , 0.8 , 3.1 , -1.7 , 14.3} → R(y) = {1 , 3 , 4 , 2 , 5}

- Compute the correlation **on the ranks :** cor(R(x) , R(y))

  = **Spearman correlation**



Correlating multiple variables