Experimental Humanities II (HUMB002) 2016
STATISTICAL ANALYSIS

Lecture 2

Z-SCORES AND OTHER STANDARDIZED SCORES, NORMAL DISTRIBUTION

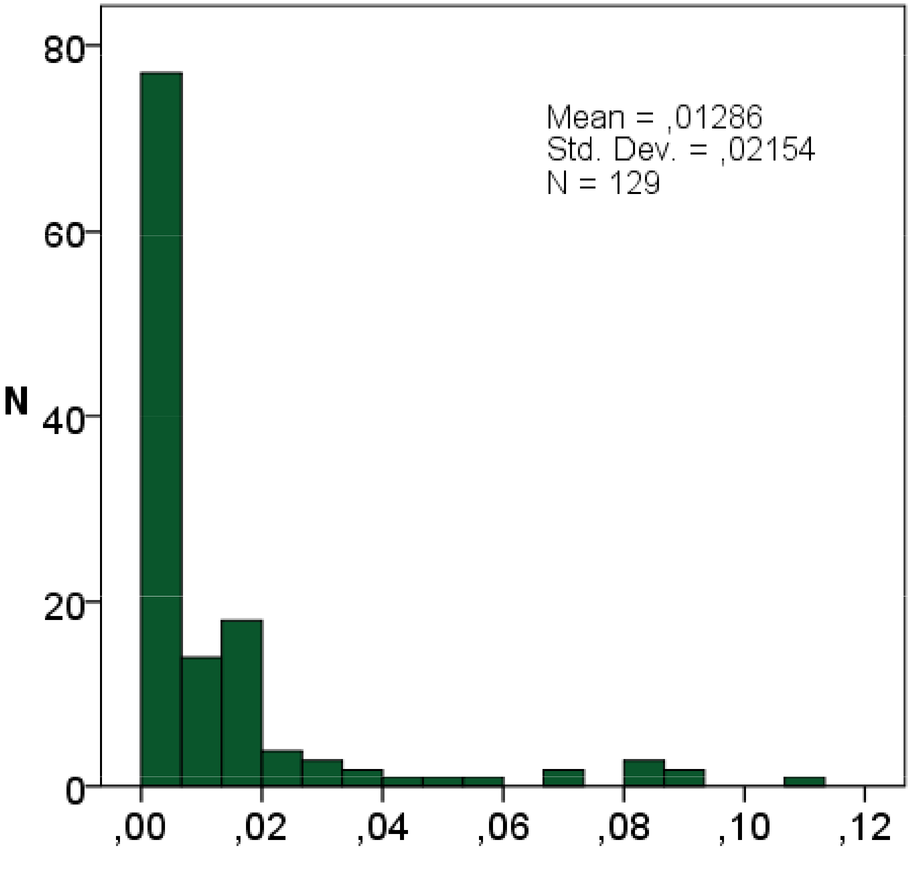
CORRELATION, SIMPLE LINEAR REGRESSION

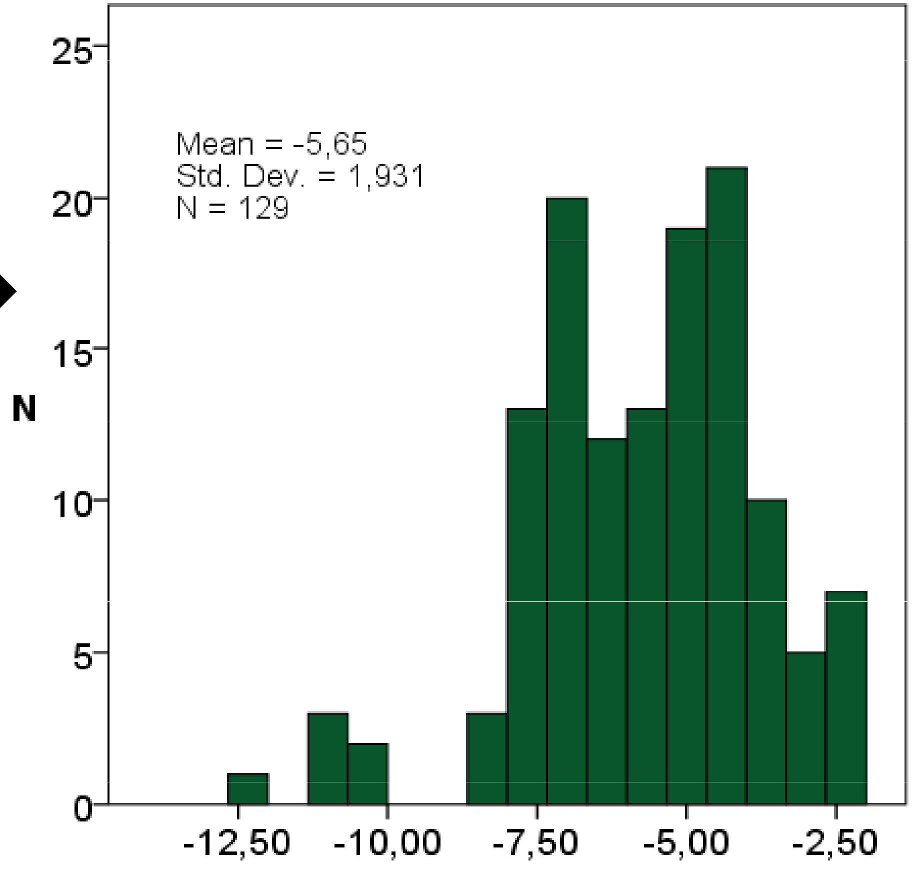
Scores transformation

- We often transform observed score for easier understanding and interpretation
- Making interpretation easier – linear transformations
  - e.g. multiplying by 10 or 100 for getting rid of the decimals
  - distribution shape remains unchanged
  - descriptive statistics will change in a predictable way
  - possibility of standardization
- Change of distribution shape – nonlinear transformations
  - log/exp functions, square(root)…
- Change of measurement level – ordinal transformation (ranking)

Nonlinear transformations

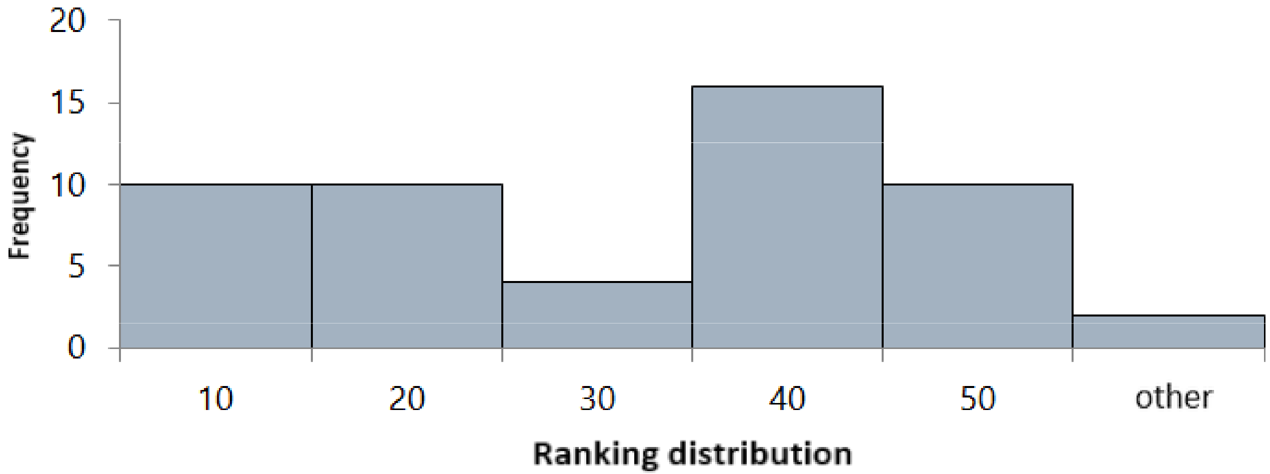
- Example: ln transformation, usually trying to make distribution „more normal“





Ordinal transformation: ranking

- Transformation from observed values on ranking:
  - Elimination of extreme values, neglecting differences magnitude between people
  - Usually ascending (the lowest value is 1)
  - The same values get average ranking (=RANK.AVG)
  - The distribution will be approximately uniform
  - **Percentiles** – standardized form of ranking transformation



Linear transformation – standardization

- The most usual transformation: **standardization (z-scores)**
  - scores transformation, so that **M=0, SD=1**
  - measurement unit becomes SD – we can easily compare scores from different scales (but differences in distribution remain!)
  - $\mathbf{z_i = (X_i - M) / SD}$
- Scores derived from z-scores:
  - T scores: M=50, SD=10; $\mathbf{T_i = 50 + 10z_i}$
  - IQ scores: M=100, SD=15
  - Stens (Standard TENs, cz: steny): M=5.5, SD=2; $Sten_i = 2z_i + 5.5$
  - Stanines (STAndard NINEs, cz: staniny): M=5, SD=2; $Stanine_i = 2z_i + 5$
- Normal distribution is always required for correct standardized scores interpretation!

Psychodiagnostic calculator

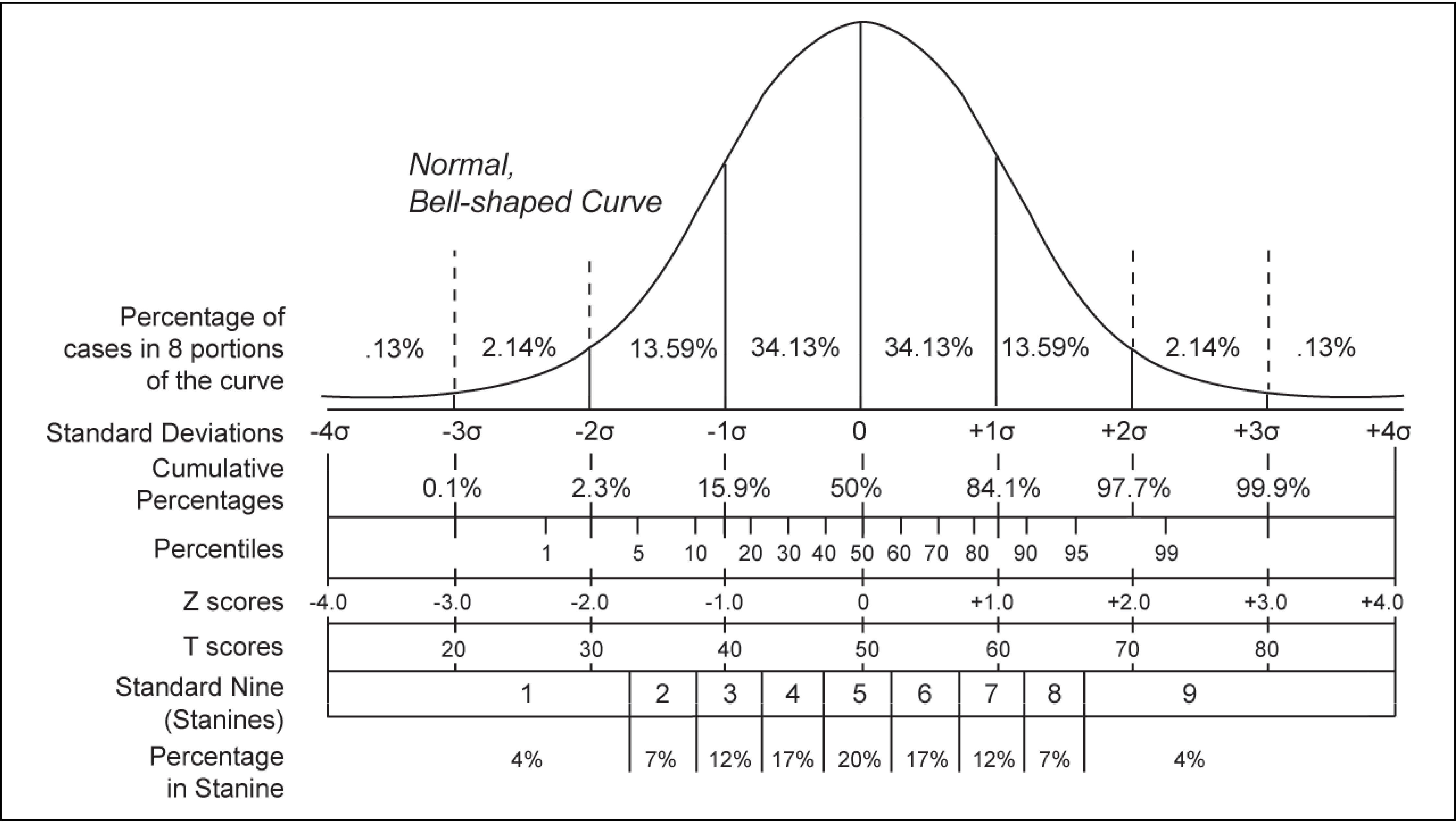
- Online tool for scores transformation
- Developed by Hynek Cígler and Martin Šmíra
  from the Department of Psychology, Faculty of Social Studies MU
- http://kalkulacka.testforum.cz/transformace-skoru
- In Czech language

Normal distribution (Gaussian, bell curve)

- The distribution of nature phenomena influenced by many independent factors: many variables
- The distribution of random errors
- Advantages: we can assume how many of which values there are in the target population
- Many statistical procedures work with normal distribution (-> many statistical procedures require normal distribution)

Characteristics of normal distribution

- Symmetrical, unimodal
- Mean = median = mode
- Skewness = 0
- Kurtosis = 3
- Standardized normal distribution: scores transformed to z-scores (M=0, SD=1)


Normal,
Bell-shaped Curve
Percentage of cases in 8 portions of the curve
.13%
2.14%
13.59%
34.13%
34.13%
13.59%
2.14%
.13%
Standard Deviations
-4σ
-3σ
-2σ
-1σ
0
+1σ
+2σ
+3σ
+4σ
Cumulative Percentages
0.1%
2.3%
15.9%
50%
84.1%
97.7%
99.9%
Percentiles
1
5
10
20
30
40
50
60
70
80
90
95
99
Z scores
-4.0
-3.0
-2.0
-1.0
0
+1.0
+2.0
+3.0
+4.0
T scores
20
30
40
50
60
70
80
Standard Nine (Stanines)
1
2
3
4
5
6
7
8
9
Percentage in Stanine
4%
7%
12%
17%
20%
17%
12%
7%
4%

Counting quantiles in Excel

- NORM.S.DIST(z;1) – returns corresponding percentile for given z-score (=how many people have the same or lower z-score)
  - Percentage of people between two given z-scores: NORM.S.DIST(higher z;1) minus NORM.S.DIST(lower z;1)

- NORM.S.INV(p) – returns corresponding z-score for given percentile

Example

- John got score 7 from math test and score 13 in English tests. The math tests scores have M=5 and SD=2.2, the English test has M=9 and SD=3.6.
  In which of the tests was John better?
- Math
  - z-score: (7-5)/2.2 = 0.91
  - T-score: 50 + 10*0.91 = 59.1
  - percentile: NORM.S.DIST(0.91;1) = 0.82 = 82nd percentile
  - John was in the math test same or worse than 82% of other students
- English
  - z-score: (13-9)/3.6 = 1.11
  - T-score: 50 + 10*1.11 = 61.1
  - percentile: NORM.S.DIST(1.11;1) = 0.87 = 87th percentile
  - John was in the English test same or worse than 87% of other students

ASSOCIATIONS BETWEEN VARIABLES

CORRELATION

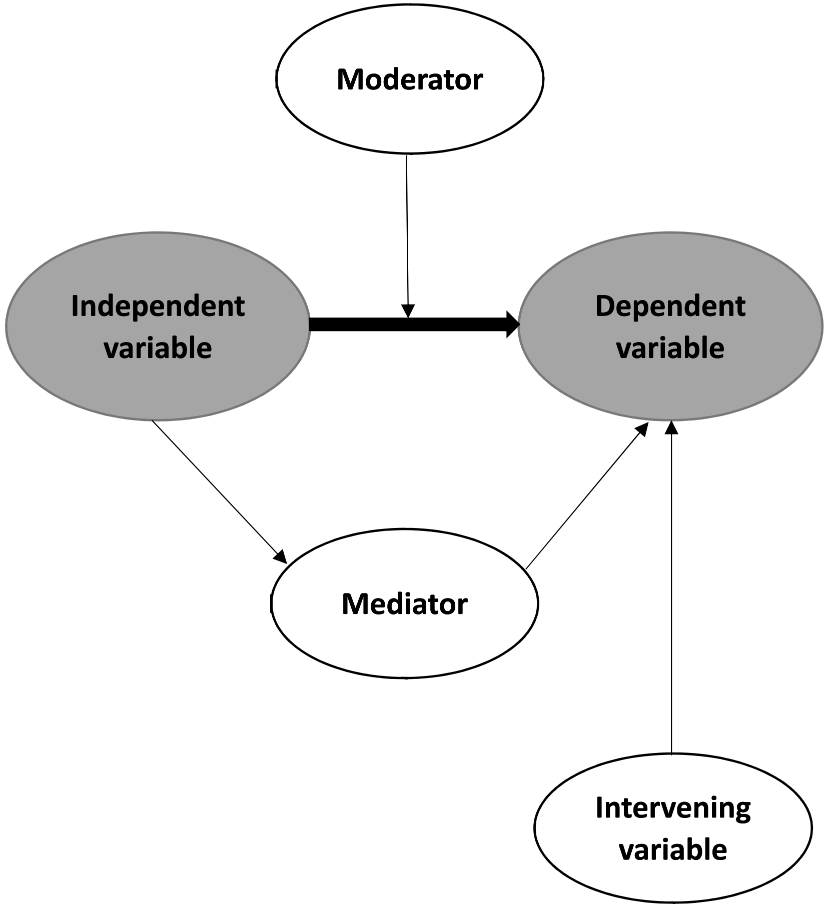
Associations between variables

- Statistical mapping of association between variables depends on measurement level: categorical vs. metric variables

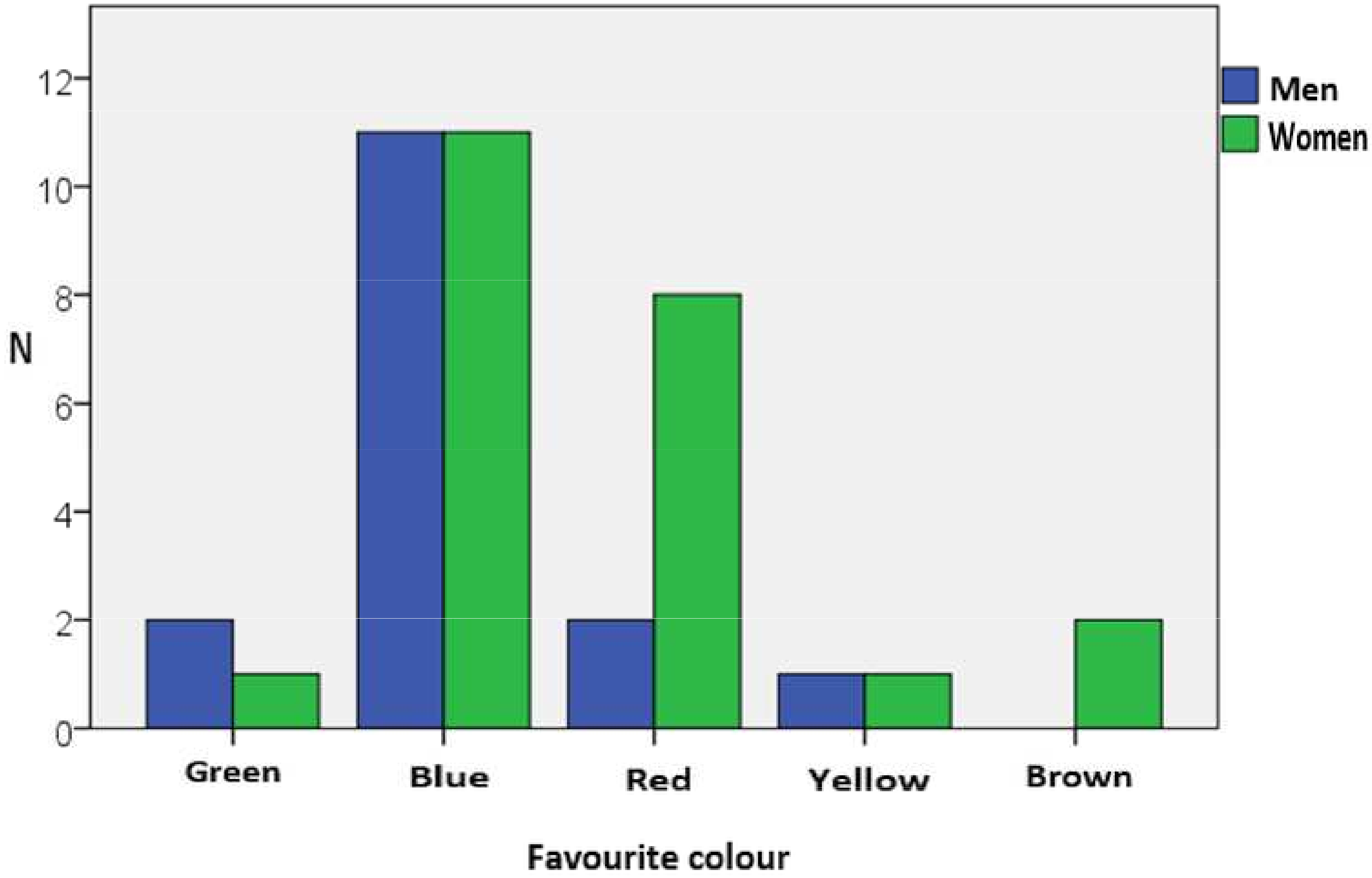
	categorical	**metric**
categorical	contingency table compound bar chart *chi square*	
metric	compound unidimensional graphs *difference in descriptive statistics*	scatterplot *correlation*

Variables classification according to their function in the association

- We are usually interested in causal relationships
  - Statistics itself cannot detect or test causality
  - Causality can be determined by research design and theoretical assumptions
- Variables classification
  - Dependent, independent, intervening
  - Exogenous, endogenous, mediators, moderators
  - Usually we can’t identify all intervening variables…



Compound bar chart



Contingency table

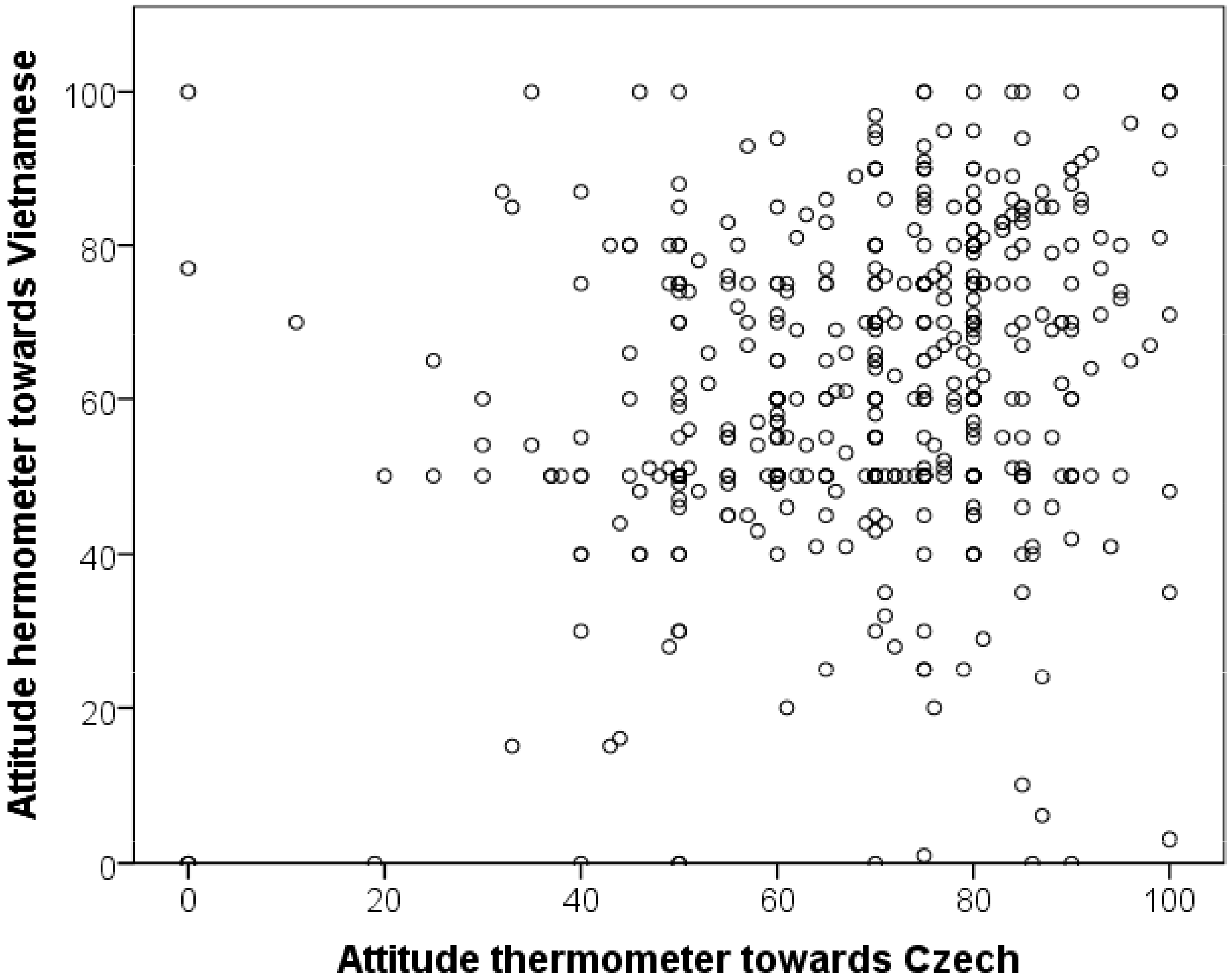
		Math grades					Total
		1	2	3	4	5	
Czech language grades	1	82	40	8	1	0	131
	2	71	200	73	17	0	361
	3	4	75	109	25	0	213
	4	1	7	23	24	1	56
	5	0	0	2	1	2	5
Total		158	322	215	68	3	766

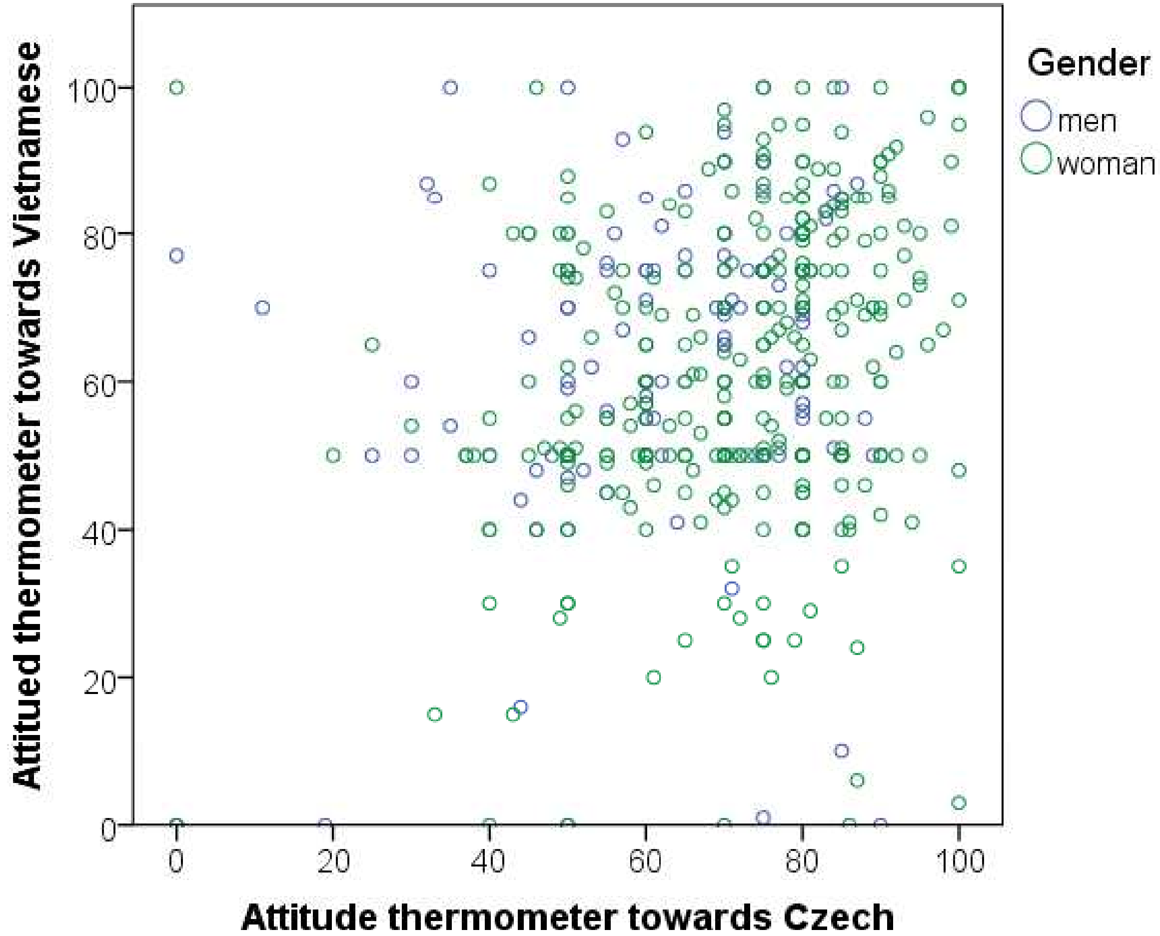
- For any variables, but most suitable for discrete variables with not many values
- The cells can contain both absolute or relative frequencies (row, column and total frequencies)
- The last row and column contain so called row/column marginal frequencies
- Graphical representation of contingency table is 3D bar chart or 3D histogram
- High frequencies in diagonals indicate linear relationship between variables

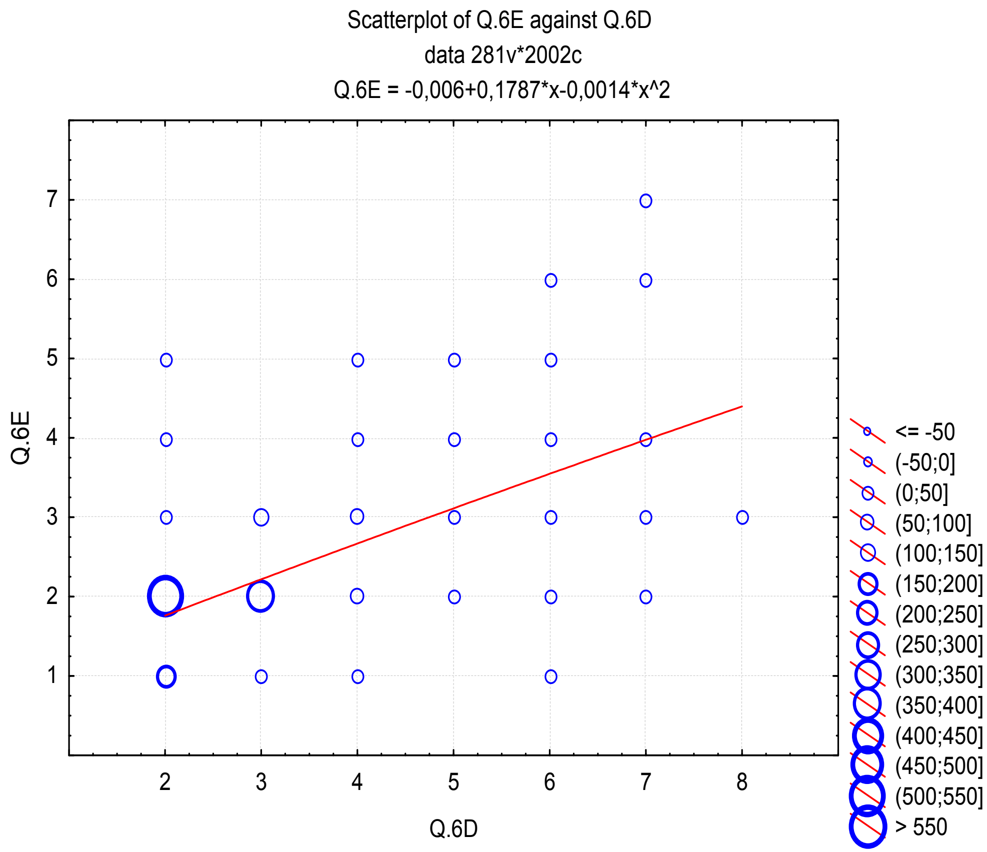
			How often do you personally encounter people of Vietnamese origin?					Total
			Not at all	Rarely	Sometimes	Often	Very often	
Gender	Men	N	2	34	46	43	6	131
		% (gender)	1,5%	26,0%	35,1%	32,8%	4,6%	100,0%
		% (contact)	15,4%	37,8%	21,6%	28,3%	16,7%	26,0%
	Women	N	11	56	167	109	30	373
		% (gender)	2,9%	15,0%	44,8%	29,2%	8,0%	100,0%
		% (contact)	84,6%	62,2%	78,4%	71,7%	83,3%	74,0%
Total		N	13	90	213	152	36	504
		% (gender)	2,6%	17,9%	42,3%	30,2%	7,1%	100,0%
		% (contact)	100,0%	100,0%	100,0%	100,0%	100,0%	100,0%

Scatterplot

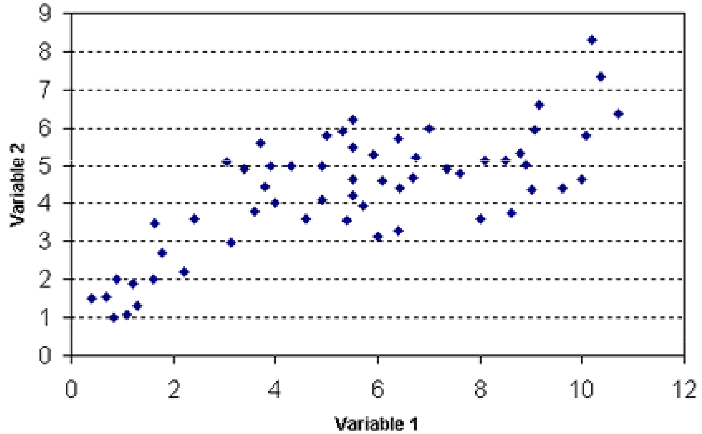
- Substitutes contingency table for continuous variable
- Each axis represents one variable
- Each point represents one subject (unit)
- Frequency of the same values can be represented e.g. by the dot size

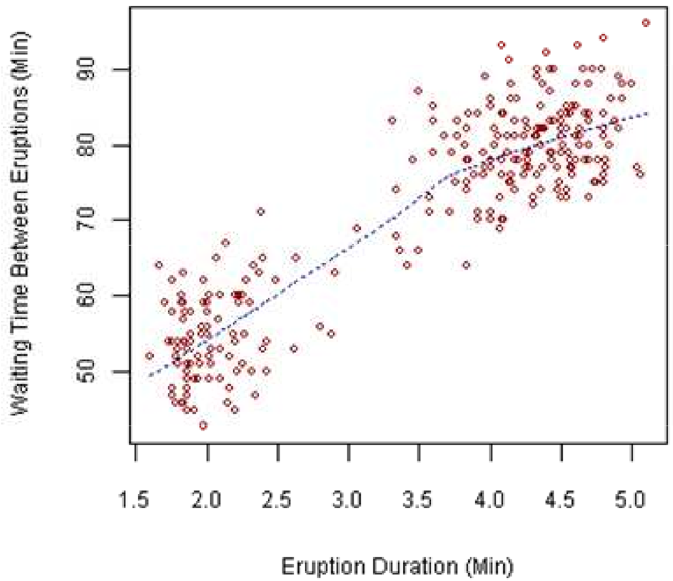


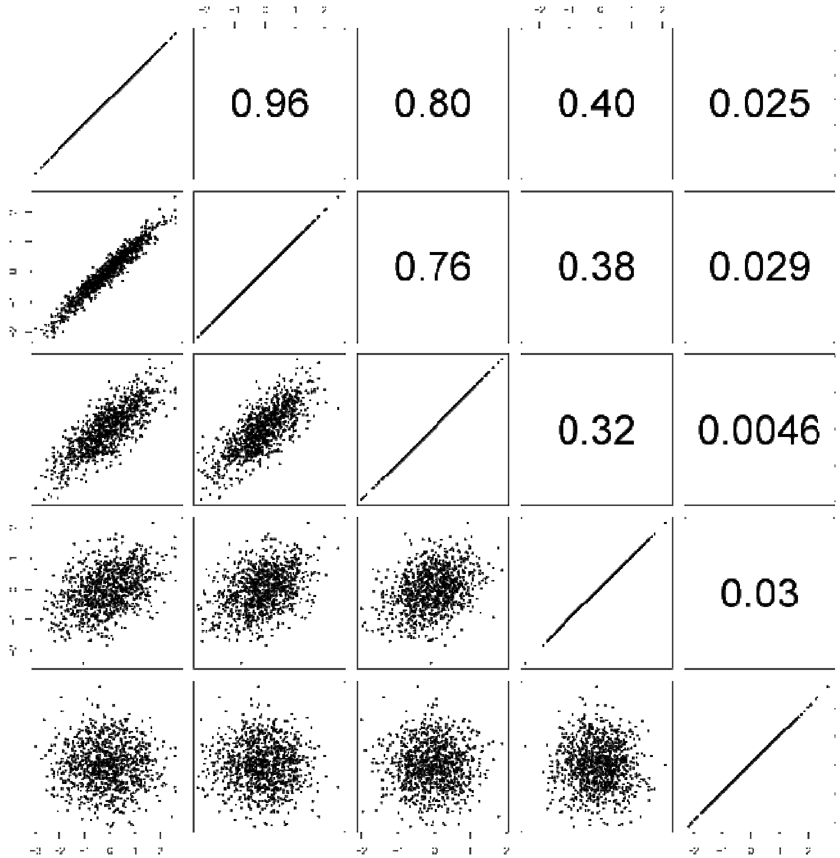

Attitued thermometer towards Vietnamese
Attitude thermometer towards Czech
Gender
men
woman
0
20
40
60
80
100

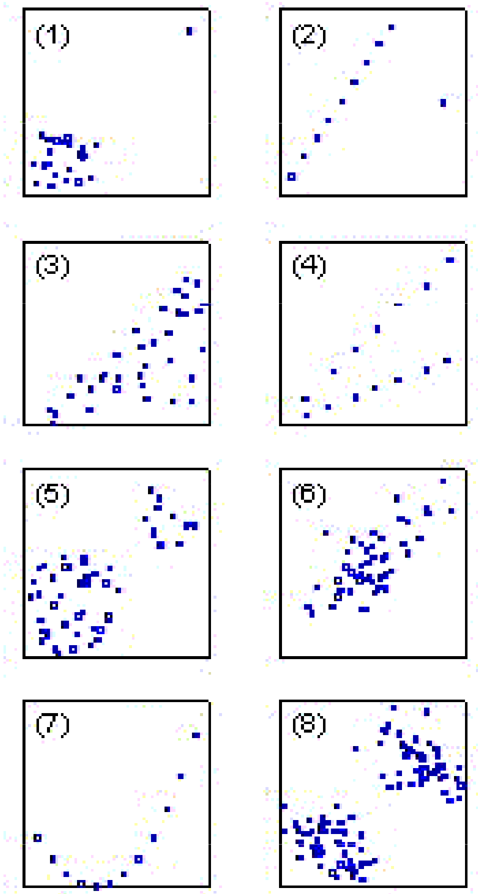

Scatterplot of Q.6E against Q.6D
data 281v*2002c
Q.6E = -0,006+0,1787*x-0,0014*x^2
Q.6E
Q.6D
<= -50
(-50;0]
(0;50]
(50;100]
(100;150]
(150;200]
(200;250]
(250;300]
(300;350]
(350;400]
(400;450]
(450;500]
(500;550]
> 550

Different forms of associations





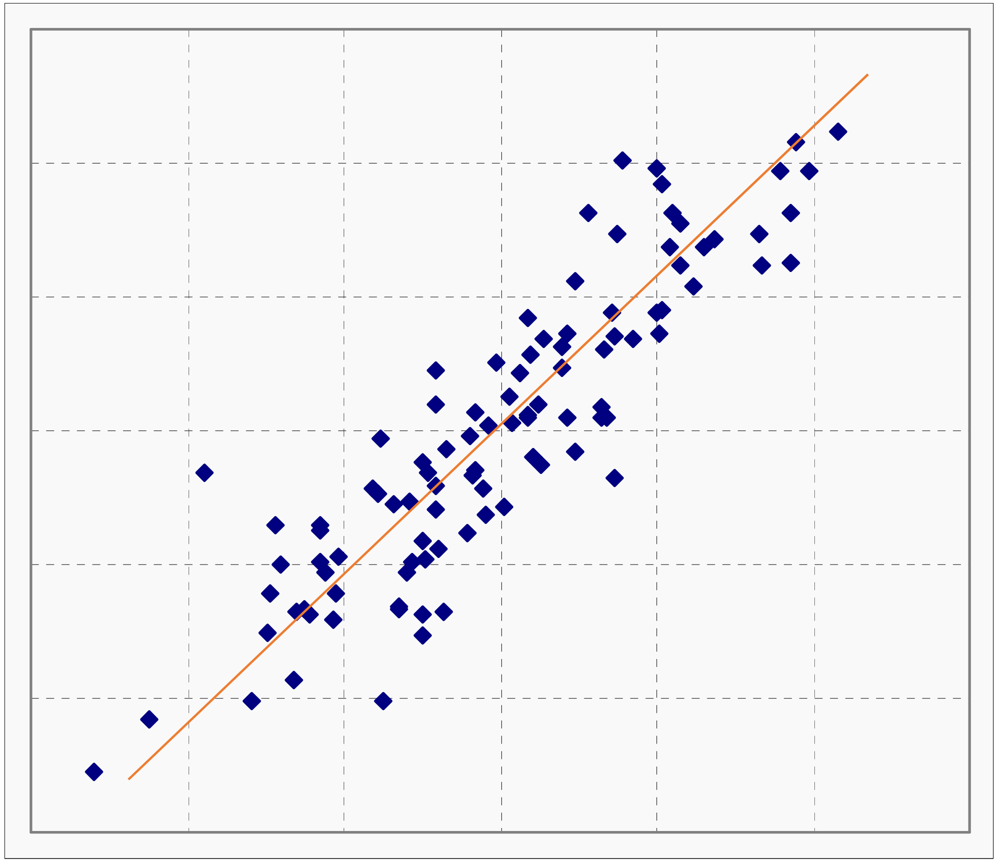




LINEAR RELATIONSHIPS

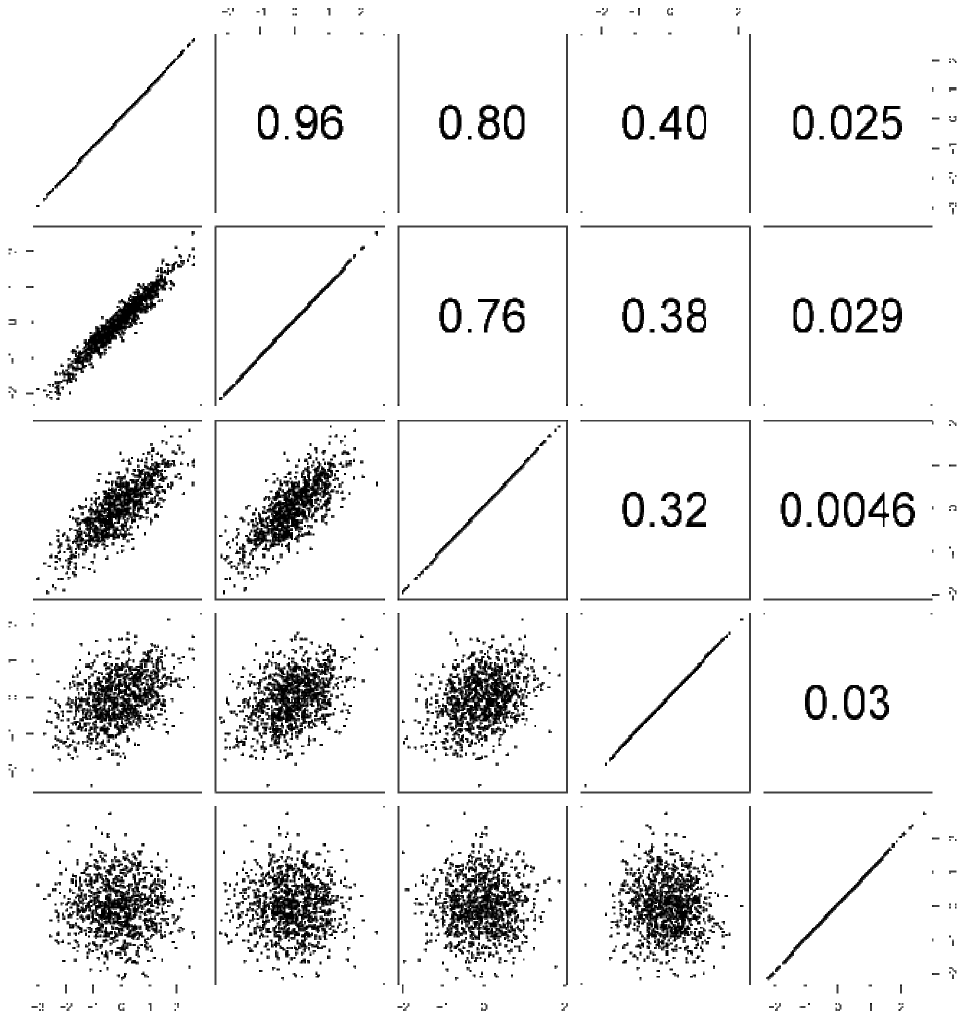
Linear association

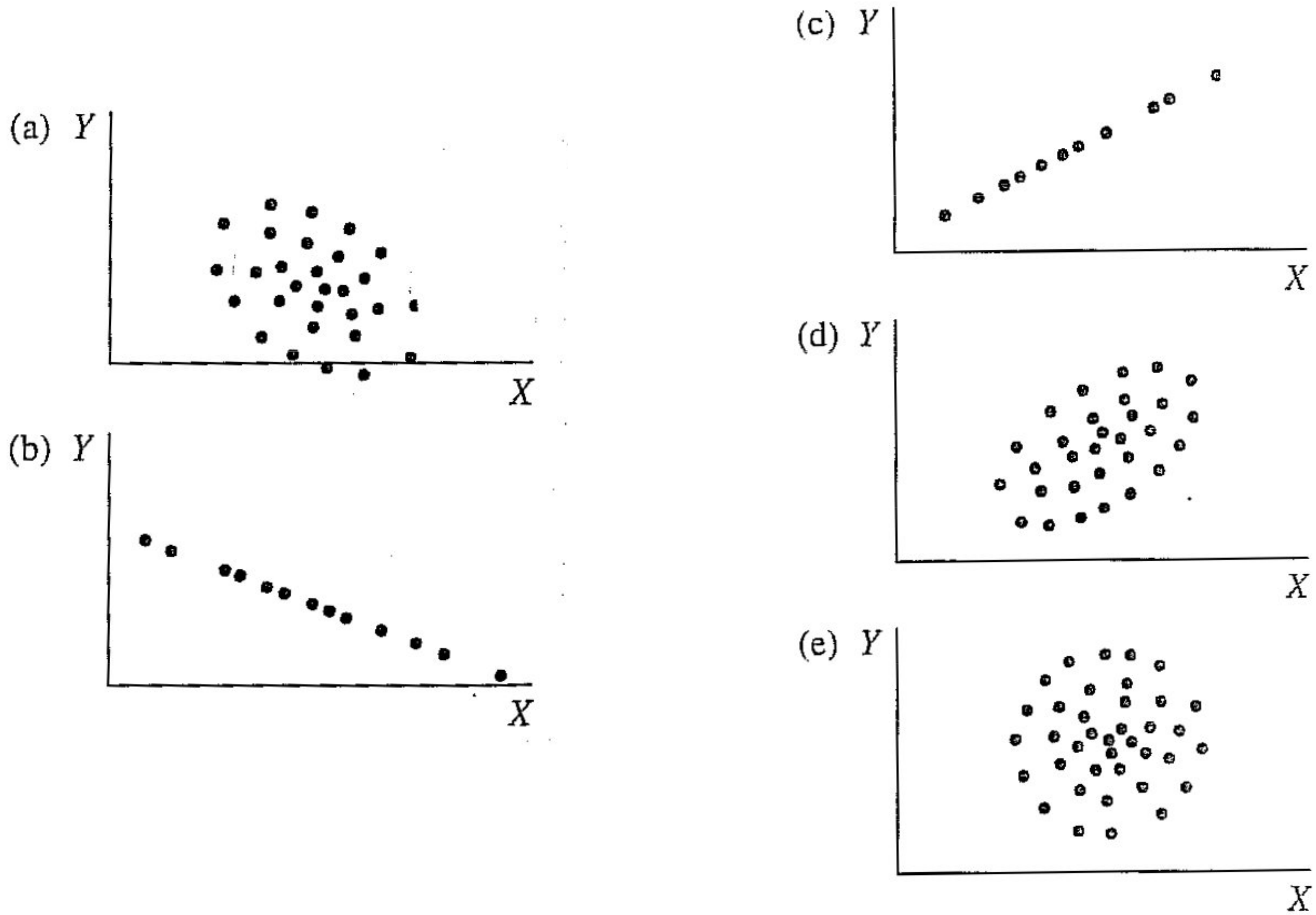
- Monotonous relationship which can be described in words:
  the higher X, the higher/lower Y
- By „correlation“ we usually mean linear association
- In the scatterplot, the „ideal“ line can be placed
- The linear function (line) Y = a + bx indicates the association slope
- Correlation describes the strength of the linear association
  (cz: těsnost vztahu)



Strength of association

- The stronger the association, the closer are the points to the line
- Strength of association isn't related to the line's slope
- Strength of association can be described by correlation coefficient from -1 to 1
- -1 means maximum negative association, 0 mean no association, 1 means maximum positive association
- + values: the higher X, the higher Y
- - values: the higher X, the lower Y




(a)
Y
X
(b)
Y
X
(c)
Y
X
(d)
Y
X
(e)
Y
X

Covariance (shared variance)

- Covariance expresses the extent of the shared variance
- It is a numerical expression of association strength

$$c_{xy} = \frac{1}{n-1}\sum_{i=1}^{n} x_i y_i$$

Remember the formula for variance calculation: $\Sigma x^2 / (n-1)$. This formula is: $\Sigma\, x*y / (n-1)$.
So, instead of x*x, we have here x*y, that's why it's called co-variance.

This sum is the higher the more pairs of xy there are, where both x and y value is above-average or below-average. The sum is the lower the more pairs of xy there are, where one of the values is above-average and the other below-average.

- x and y are deviation scores (deviations from the average)
- covariance is not very practical, similarly as variance, because it is in „squared unit“

Correlation (=standardized shared variance)

- For better interpretation, we standardize the covariance – same as with z-scores, we divide the deviation score by standard deviation

$$r_{xy} = \frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{X_i - m_x}{s_x}\right)\left(\frac{Y_i - m_y}{s_y}\right) = \frac{c_{xy}}{s_x s_y}$$

- We already know the circled part: that's z-score transformation, so to make it easier:

$$r_{xy} = \frac{\sum_{i=1}^{n} z_{X_i} z_{Y_i}}{n-1}$$

= Pearson's product-moment correlation
(cz: součinový, momentový koeficient korelace)

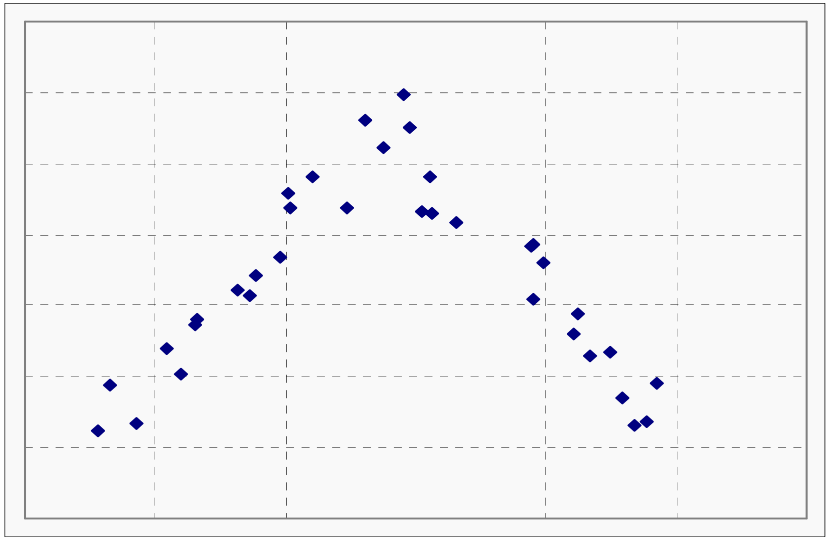
Characteristics of Pearson's correlation coefficient

- A deviation statistics:
  - interval or higher measurement level required
  - great impact of extreme values
  - suitable for normally distributed variables description (approximately normal distribution of both variables required)
  - expresses only the association strength, not causality!!!

- Takes values between -1 a +1
  - 0 = no association
  - +1(-1) = perfect positive (negative) association = variables identity
    = direct (indirect relationship)

Characteristics of Pearson's correlation coefficient

- r^2 = coefficient of determination (R^2, D) = proportion of shared variance
- Consequence: r 0,3 – r 0,1 ≠ r 0,7 – r 0,5 (0,09-0,01=0,08; 0,49-0,25=0,24)
- r=0 doesn't mean there isn't any relationship between variables, it only means there is no linear association between them



Computing correlation

1. Check assumptions: interval or higher measurement level, normal distribution of both variables, extreme values, assumption of linear relationship **(plot <u>scatterplot</u> and histograms)**
2. Compute z-scores for all observed scores – you will need M and SD for both groups: $\mathbf{z_i = (X_i – M) / SD}$

   Excel: =AVERAGEA(data), =PRŮMĚR(data), =STDEVA(data), =SMODCH.VÝBĚR.S(data), =STANDARDIZE(X;M;SD)
3. Compute correlation:

$$r_{xy} = \frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{X_i - m_x}{s_x}\right)\left(\frac{Y_i - m_y}{s_y}\right) = \frac{c_{xy}}{s_x s_y}$$

$\mathbf{r_{xy}}$ = [(-1,52*1,67) + (-1,01*1,01) + (0,00*0,02) + (1,52*-0,97) + (0,51*-0,97) + (0,01*-0,31) + (0,25*-0,47)] / (7-1) = **-0,94**

Excel: =COVARIANCE.P(var1, var2), =COVARIANCE.S(var1, var2), =CORREL(var1, var2)

Weeks of Exercise	*Resting Heart Rate*
2	82
4	78
8	72
14	66
10	66
9	70
9	69

Week of exercise	Resting heart rate
M=8	M=71,86
SD=3,96	SD=6,07

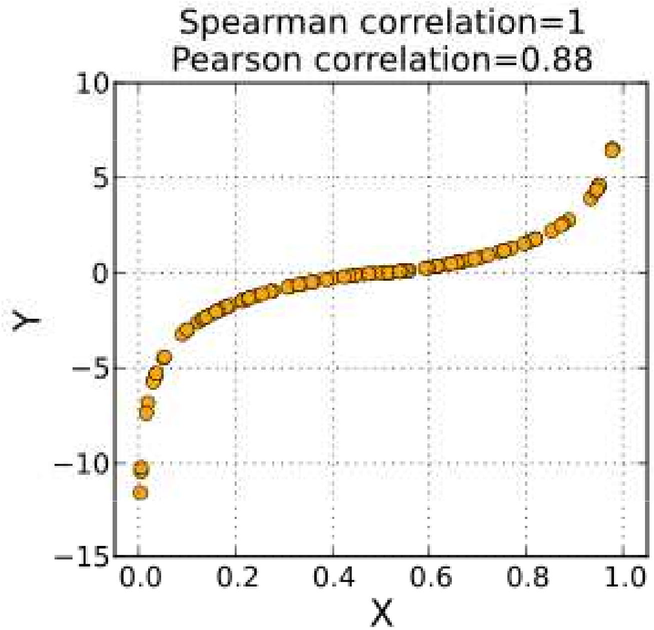
Week of exercise z scores	Resting heart rate z scores
-1,52	1,67
-1,01	1,01
0,00	0,02
1,52	-0,97
0,51	-0,97
0,01	-0,31
0,25	-0,47

Characteristics of Pearson's correlation coefficient

- When correlation doesn't make sense?
  - Q1: How many hours daily do you watch TV?
  - Q2: How many hours daily do you watch TV news?
  - …why?
- Correlation of variables with the same cause:
  - Priests' salaries and vodka prices correlate
  - Children's IQ and their height correlate as well…
  - …why?
  - Age and number of birthdays…
  - Covariance of variables with the same cause is the basis for other analysis methods: scale reliability analysis and factor analysis

Rank (ordinal) correlation coefficients

- Suitable not only for ordinal data, but also for interval data with deviations from normal distribution
- Capture also nonlinear monotonic relationships
- To what extent are ranking of the two correlated variables the same
- Spearman rho coefficient - ρ, r_s
  - Based on differences magnitude in rankings
  - Ordinal equivalent for Pearson's correlation
  - r^2 can be interpreted
  - Usually used as a more resistant variant of Pearson's r
  - Calculated in the same way as Pearson's, but on rankings
- Kendall tau coefficient – t (+ „b“ and „c“ variants)
  - Based on number of values „out of order“
  - No effect of outliers
  - b and c variants deal with more values of the same ranking

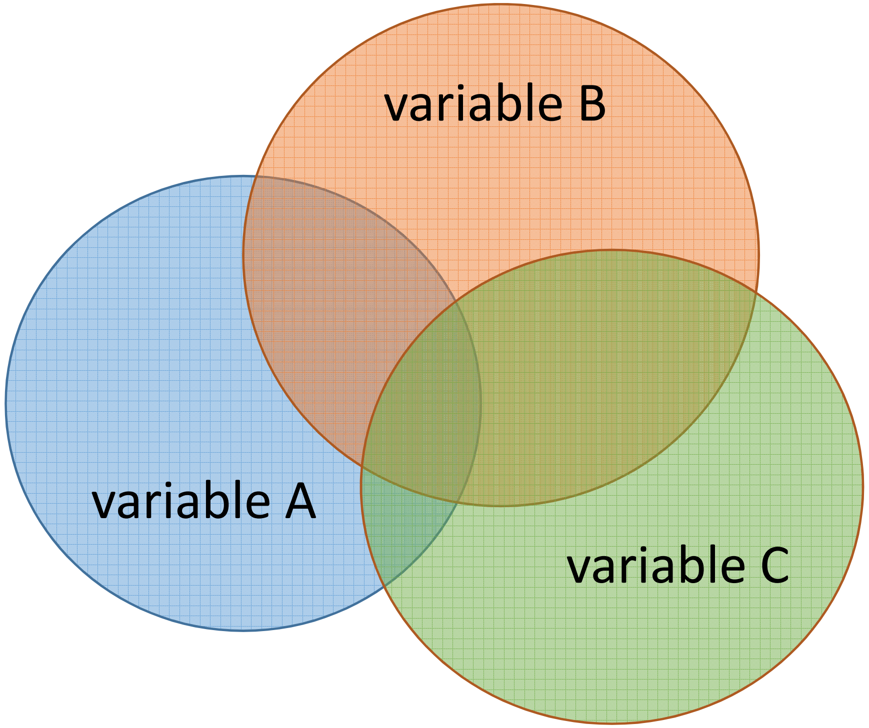


Kendall rank correlation: example

Math grade	Head circumference	Math rank	Head c. rank	Math rank	Head c. rank	K+, D-
3	48	3	3	1	5	----
2	43	2	2	2	2	++-
1	50	1	5	3	3	+-
4	49	4	4	4	4	-
5	40	5	1	5	1	

τ **= (K-D) / [*N* (*N* -1)/2]** = (3-7)/(5.4/2) = -4/10 = -0,4

Partial correlation: portioning variance



STATISTICAL PREDICTION
LINEAR REGRESSION

Statistical prediction

- **Statistical prediction** is qualified estimating of the most probable variable value from data we already know by **modelling the relationship** between the variable and its **correlates**
- From one (or more) variable (predictor, independent variable) we are trying to predict another variable (predicted variable, dependent variable)
- E.g. How well can intelligence test at 10 years predict grades in the end of high school?
- We make a model: we collect data from both variables, that is results of an intelligence test and high school grades from the same people (we already need them to be finishing their high school).
- If the model works successfully, we can use the intelligence test scores in 10-year-old children to predict their future grades…

Statistical prediction

- **Example 1:** Imagine that all students have exactly the same grades from math and physics. The variables would be identical:
  - **What would be the value of correlation between the variables?**

    $r = 1$
  - **What would be the value of coefficient of determination?**

    r^2 (R^2, D) = $1^2 = 1$
  - **What would be the proportion of shared variance? What does it mean?**

    $R^2 * 100 = 1 * 100 = 100\%$

    It means that we can predict 100% of math grades values correctly from physics values (or the other way).
  - **What of the information above would change if all students had exactly opposite math grades than physics grades?**

    $r = -1$, $R^2 = -1^2 = 1$, $R^{2\,*}\,100 = 100\%$
- Of course, usually we can predict with much less precision, but we're trying to predict with the highest precision. For that, we need correlates with high correlation with the predicted variable.

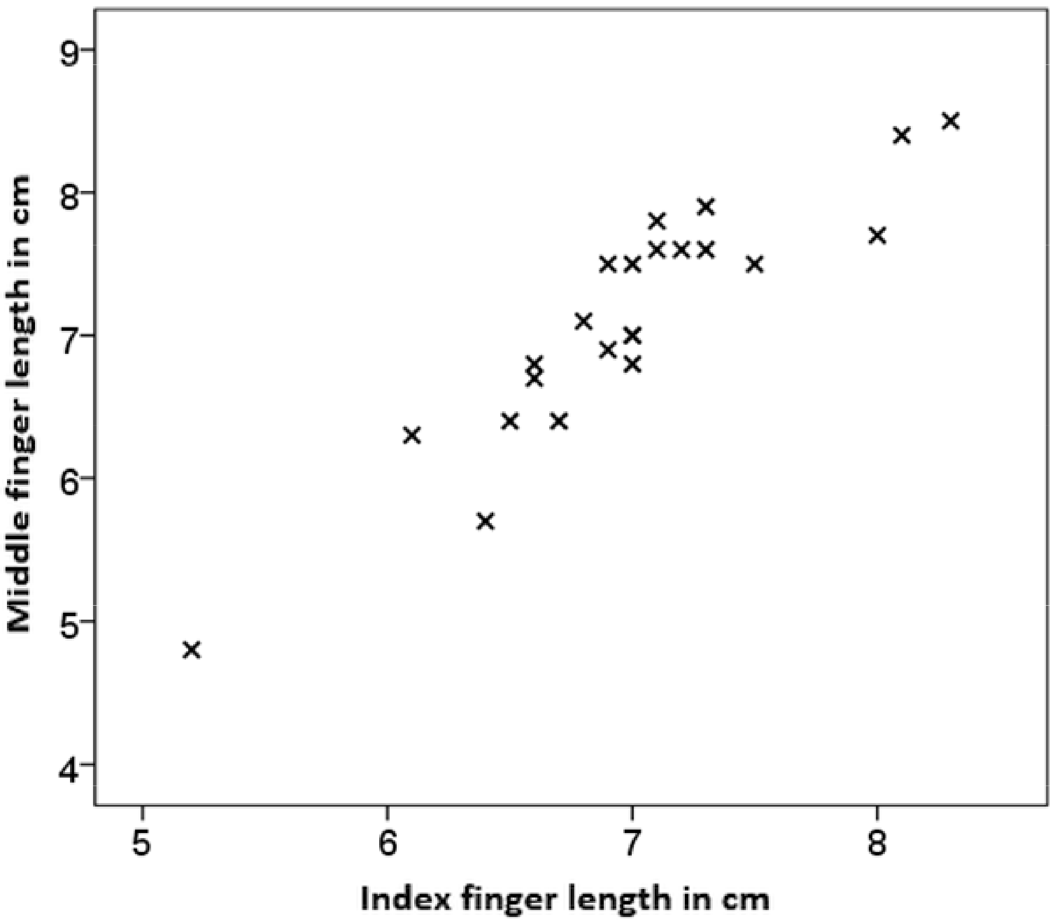
Math	Physics
1	1
2	2
3	3
4	4
5	5

Math	Physics
1	5
2	4
3	3
4	2
5	1

Statistical prediction

- Example 2: What score in an intelligence test could we predict for a random respondent, if we know that the test has approximately normal distribution with M=100 and SD=15?
- What information could make our prediction more precise?
  - Height?
  - Education?
  - Score from a memory test?
  - …

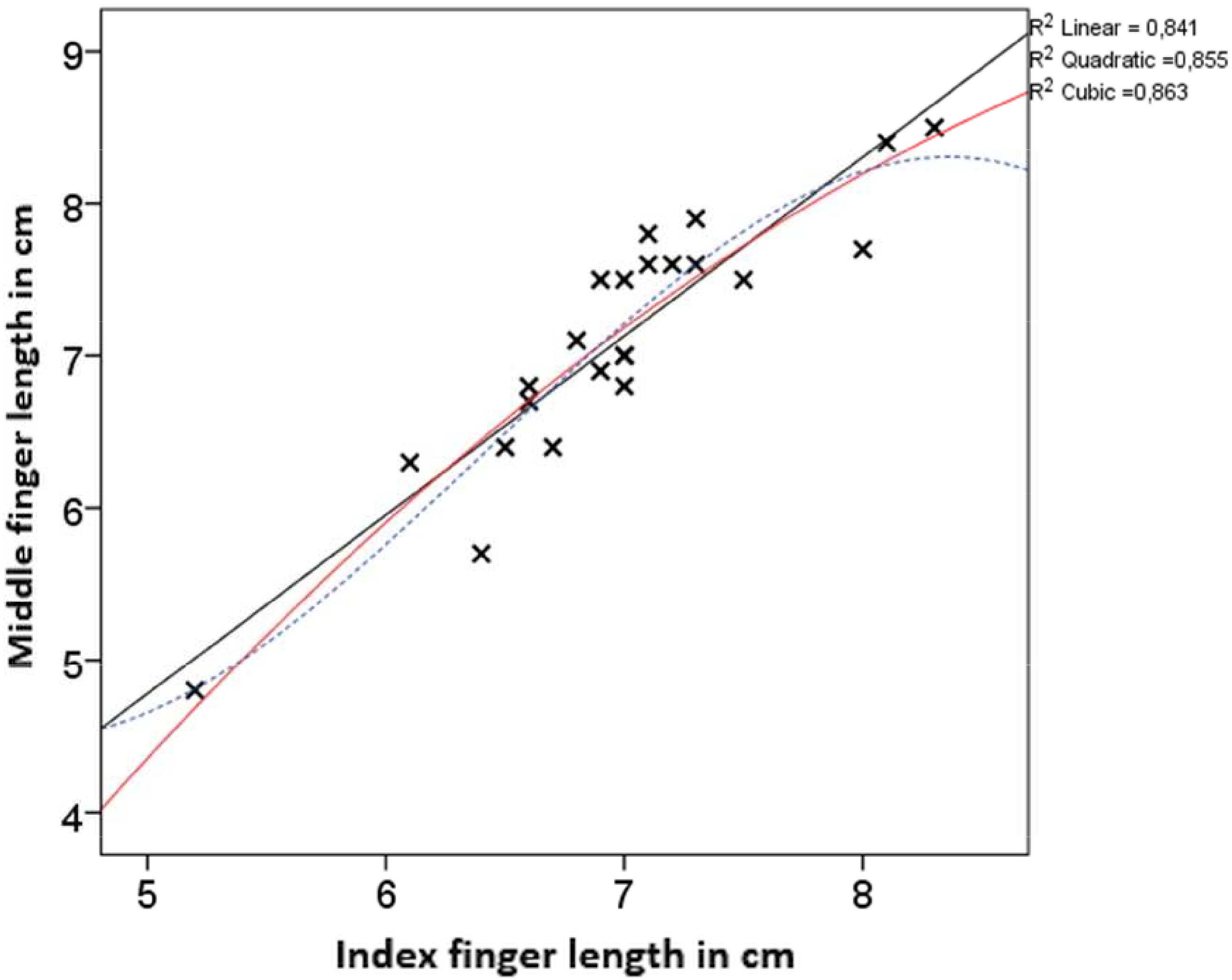
Prediction of middle finger length from index finger length



Linear regression

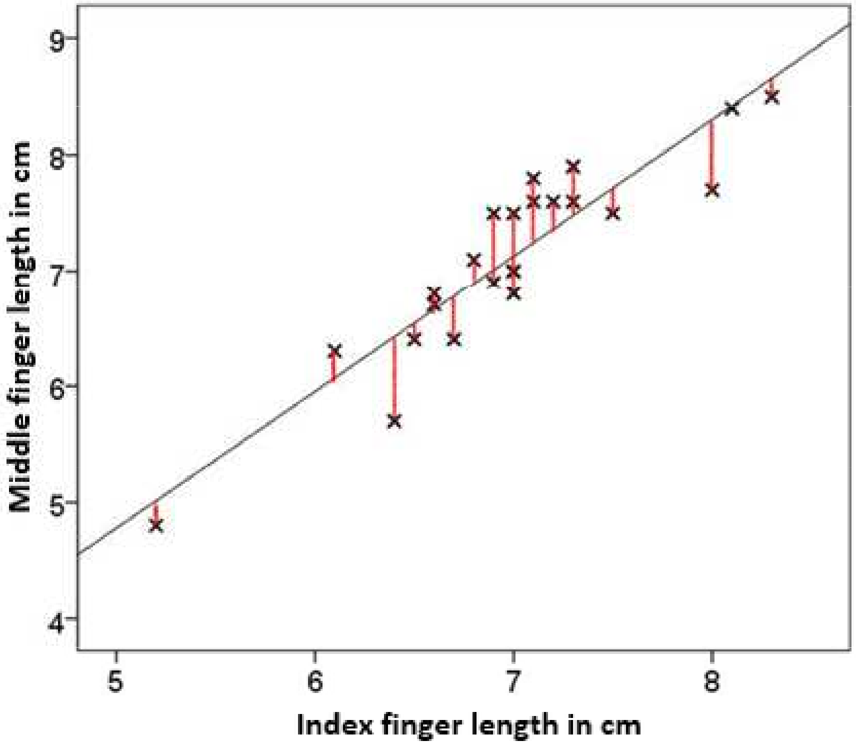
- For prediction, we need a function (how to compute variable Y from know variable X)
- For linear regression – prediction based on linear relationship, it is linear equation: **Y = a + bX** (a straight line, regression line)
- We are modelling the linear function: we're making estimations of variable Y values by computing the linear equation using variable X values
- Variable Y estimated = **Y'**
- Regression of Y on X: **Y' = Y + e = f(x) + e, where e = Y' – Y**
  - e is **residual value**, Y is dependent variable, X is independent variable (predictor)
  - e represents all other variance sources except X

Linear regression



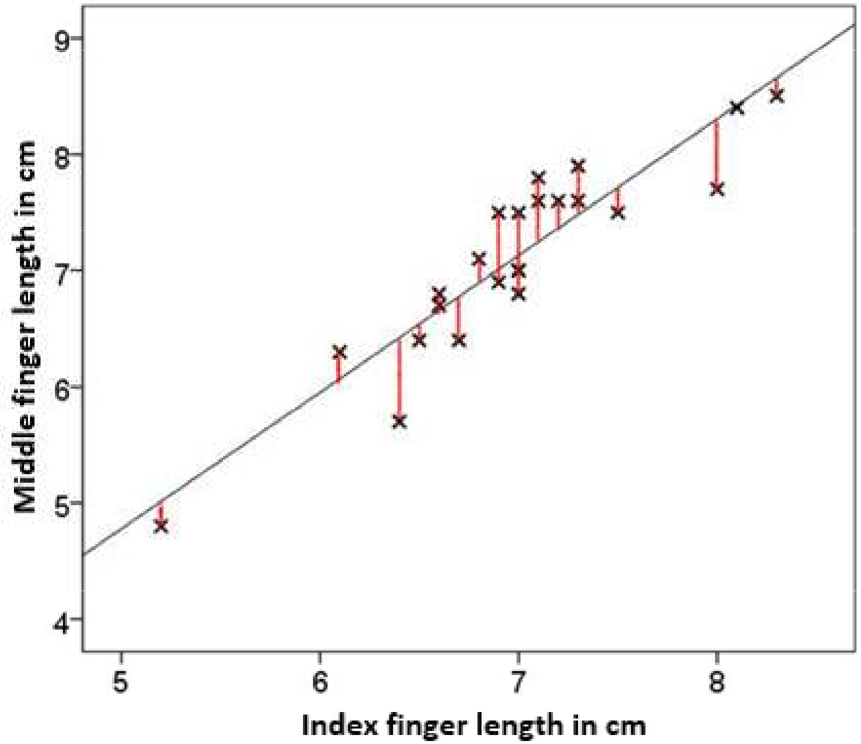
Linear regression

- If Pearson correlation well describes relationship between two variables, we can express the relationship by linear function:
- Y' = a + bX; *Y = Y' + e = a + bX + e*
- a = intercept (cz: průsečík), b = slope (směrnice)
- How can we find the best regression line?
  - Estimate by **least squares estimation** – we are trying to minimize the sum of residual squares
- $b = r_{xy}(SD_y/SD_x)$
- $a = M_y - bM_x$
- If the values of X and Y are in z-scores, then $b = r_{xy}$
- a, b – correlation coefficients



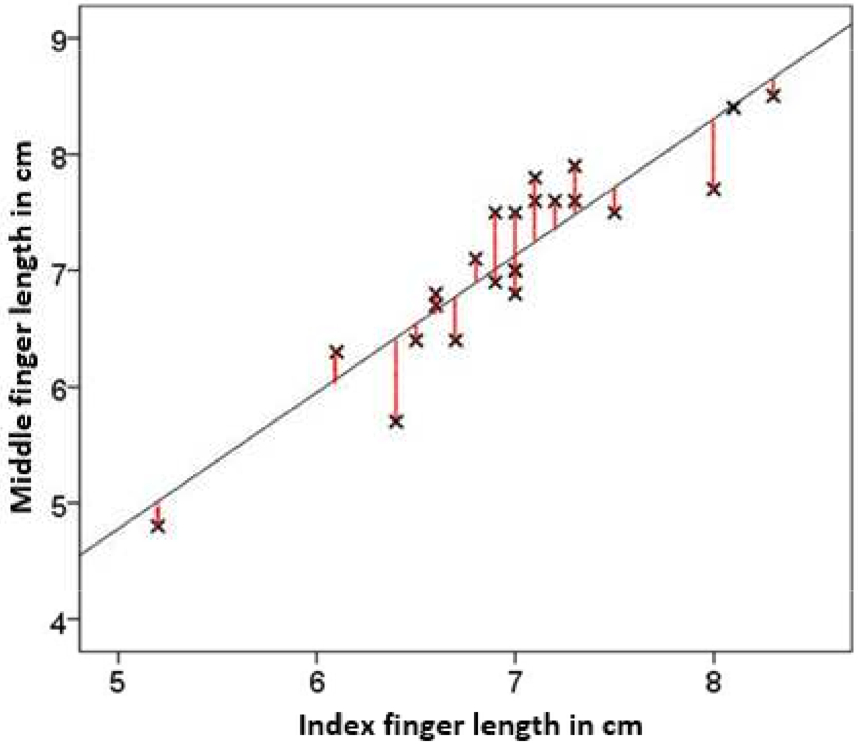
Linear regression

- $Y' = a + bX$
- $b = r_{xy}(SD_y/SD_x)$
- $a = M_y - bM_x$
- If the values of X and Y are in z-scores, then $b = r_{xy}$
- The line goes through values M_x and M_y
- The sum of residuals is zero, the sum of squared residuals is the least possible



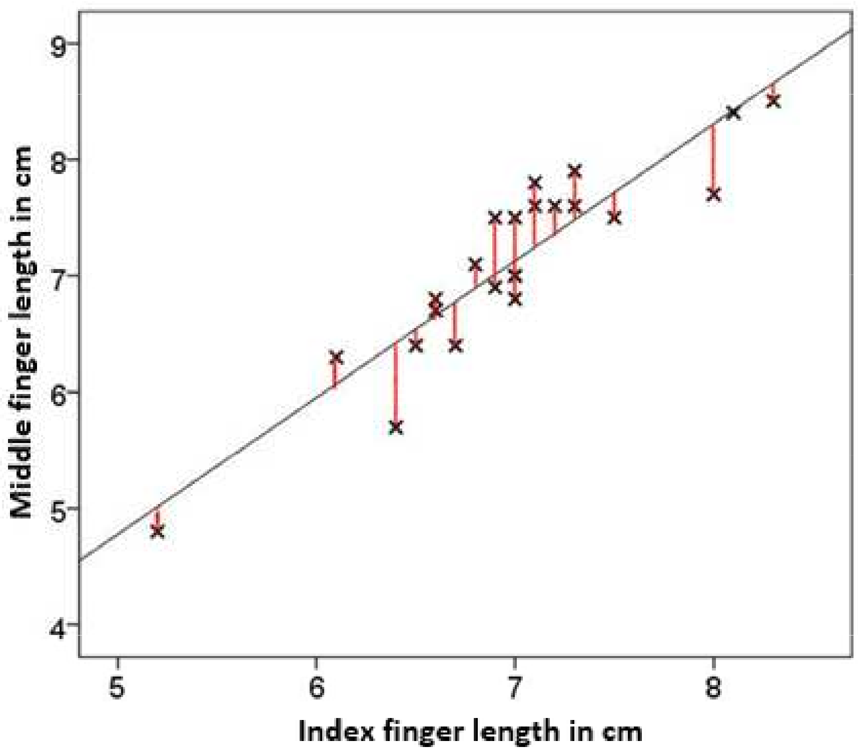
Prediction of middle finger length from index finger length

- $M_m = 7{,}109$; $SD_m = 0{,}843$ Y
- $M_i = 6{,}983$; $Sd_i = 0{,}658$ X - predictor
- $r_{mi} = 0{,}917$
- $b = r_{xy}(SD_y/SD_x) = 0{,}917(0{,}843/0{,}658) = 1{,}175$
- $a = M_y - bM_x = 7{,}109 - 1{,}175 * 6{,}983 = -1{,}096$
- $Y' = 1{,}175 * X - 1{,}096$



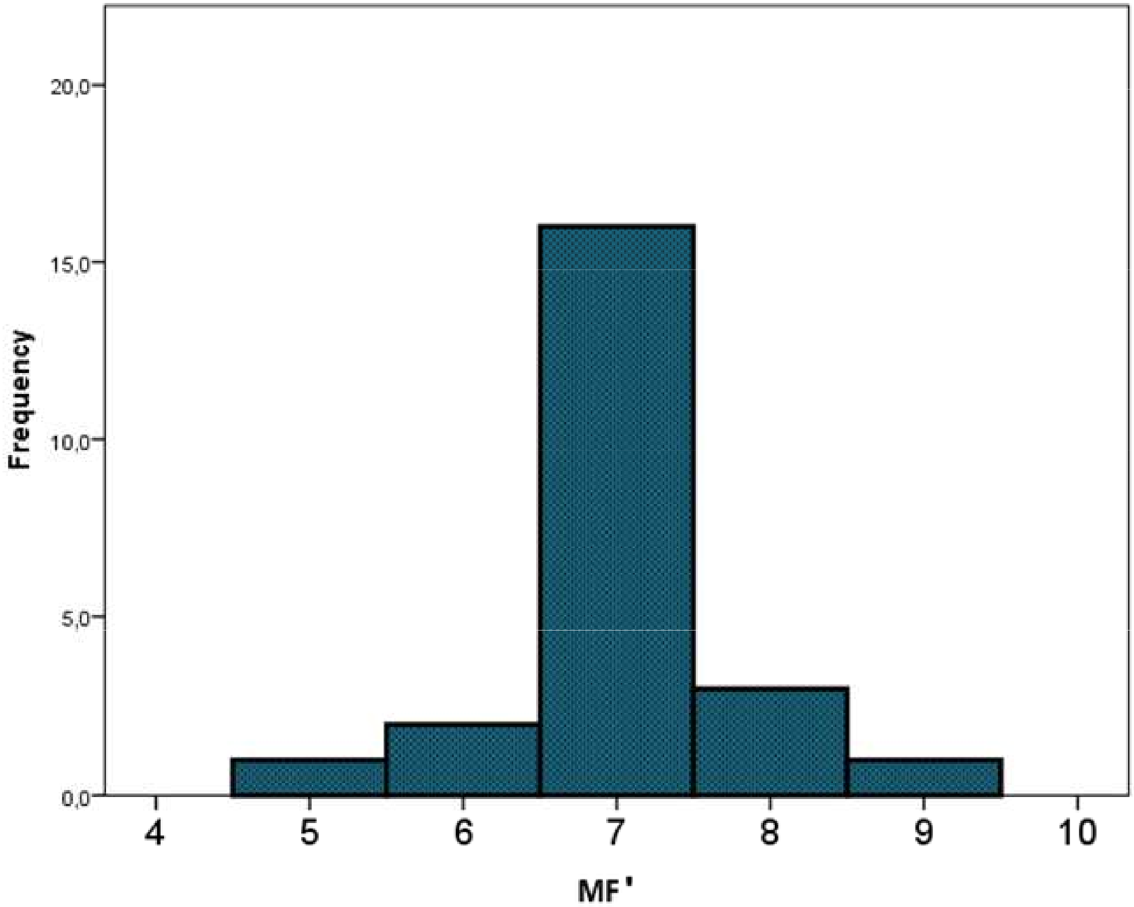
Predicted values

IF	MF	MF'
6,5	6,4	6,5413
7	7	7,1291
7,5	7,5	7,7169
5,2	4,8	5,0130
6,6	6,7	6,6589
6,6	6,8	6,6589
7	7	7,1291
6,8	?	



$$Y' = 1{,}175 * X - 1{,}096 = 1{,}175 * 6{,}8 - 1{,}096 = 6{,}894$$

Distribution of predicted values



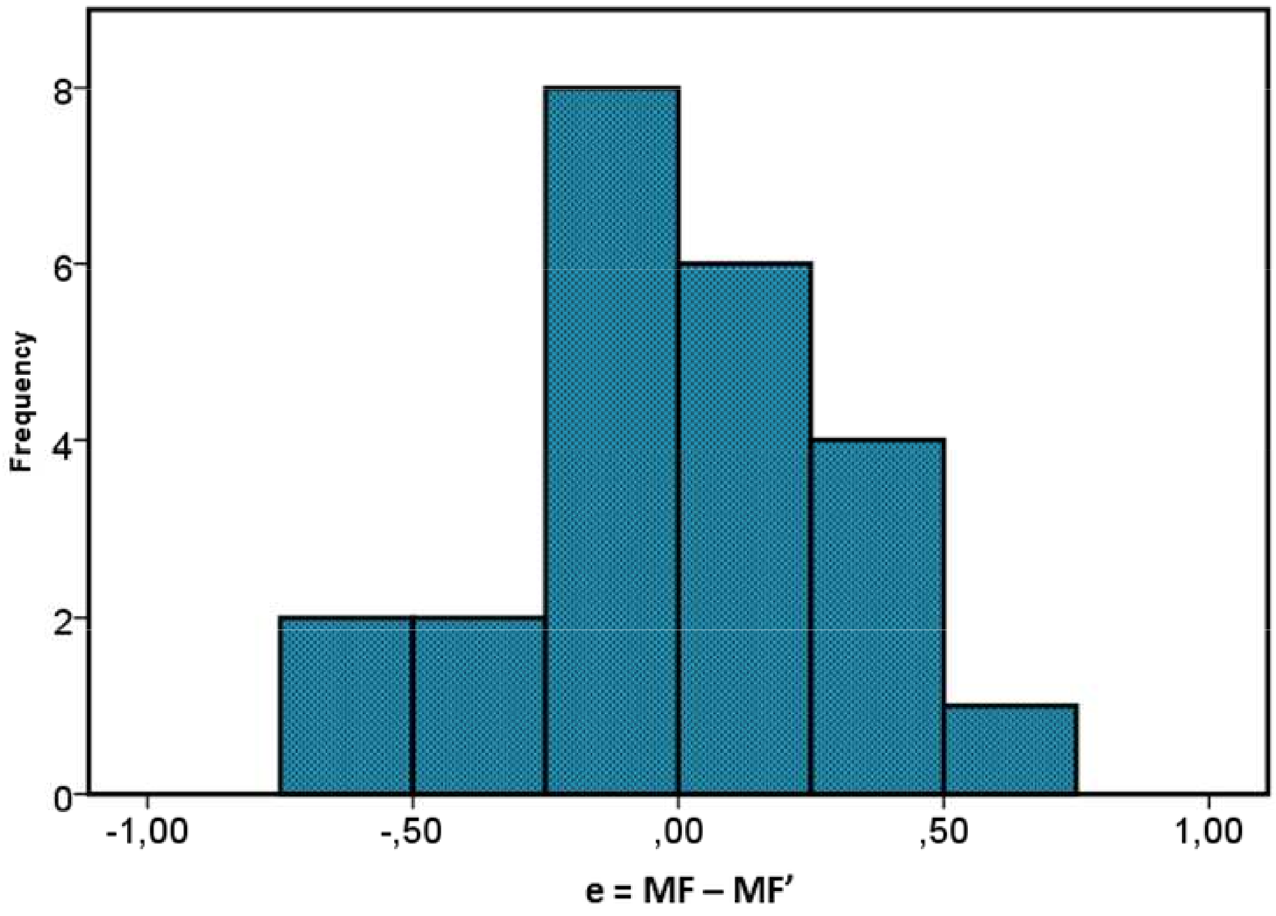
- $M_{MF'} = 7{,}109 = M_{MF}$
- $SD_{MF'} = 0{,}773$

Linear regression: model fit

- How well, precise are the predicted values?
- Precision = the least residuals
- How large are the residuals?

IF	MF	MF'	*e* = (MF - MF')
6,5	6,4	6,5413	-0,1413
7	7	7,1291	-0,1291
7,5	7,5	7,7169	-0,2169
5,2	4,8	5,0130	-0,2130
6,6	6,7	6,6589	0,0411
6,6	6,8	6,6589	0,1411
7	7	7,1291	-0,1291

Distribution of residuals



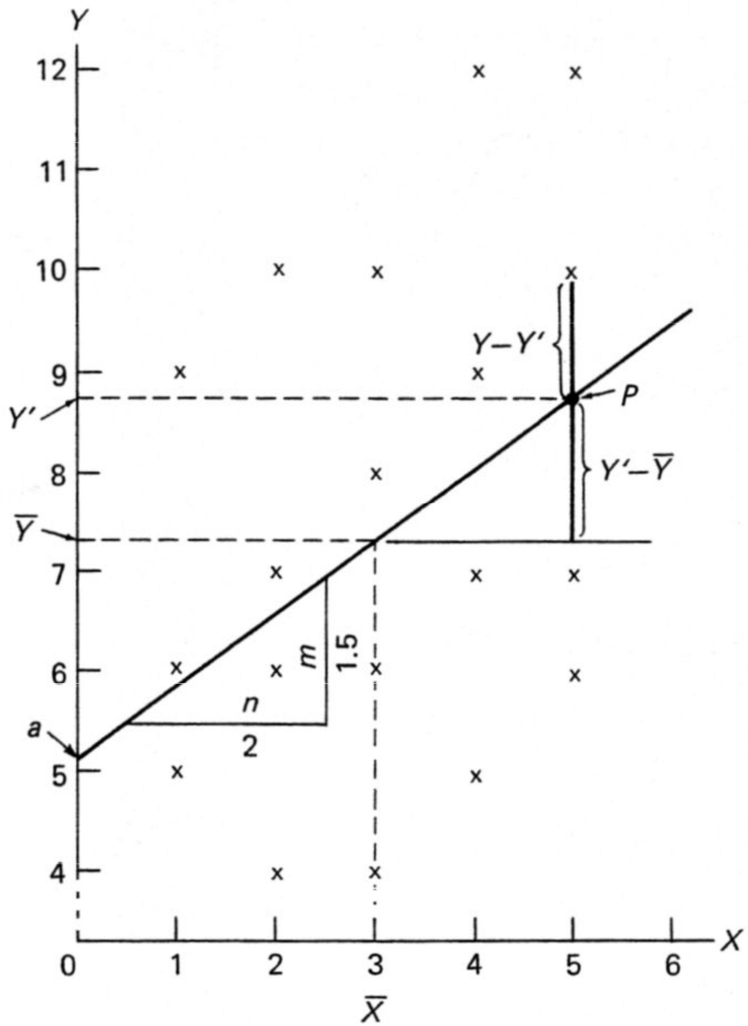
- $M_e = 0$
- $SD_e = 0,337$

Linear regression: model fit

$$s_{reg}^2 = \frac{\sum (m_y - Y')^2}{n-1} \qquad s_{res}^2 = \frac{\sum (Y - Y')^2}{n-1}$$

$$s_y^2 = \frac{\sum (Y - m_y)^2}{n-1}$$

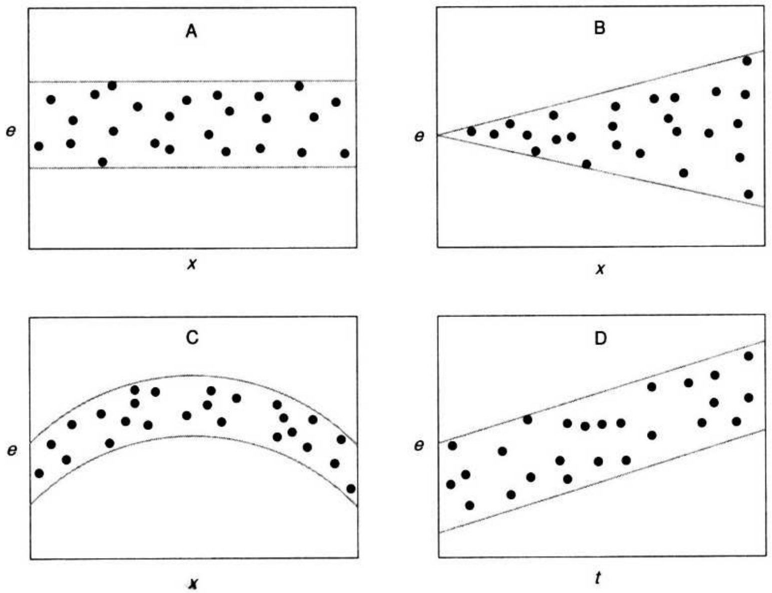
- $s_y^2 = s_{reg}^2 + s_{res}^2$
- $R^2 = s_{reg}^2 / s_y^2$ … $s_{res}^2 = s_y^2(1-R^2)$
- Coefficient of determination: R^2
    - Exaplained variance proportion
    - Measure of model fit with the data (regression success)
- **For simple linear regression it applies: $R^2 = r^2$**



Linear regression: assumptions

- Assumptions are the same as for Pearson correlation:
  - The basic assumption: the relationship really is linear
  - The residuals have normal distribution with M=0 and SD = S_{res}
  - It means the 95% of estimation residuals lie approx. between $-2s_{res}$ a $+2s_{res}$
  - **homoscedascity** (cz: homoskedascita): residuals independency = the residual variance won't change with increasing X

  - The model validity depends on data from which was the model extrapolated
  - Watch out for extreme values (as with all deviation statistics)



Other regression types

- Simple linear regression: one independent and one dependent variable
- Multiple linear regression: more independent variables (predictors)
    - $Y = a + b_1 X_1 + b_2 X_2 + \ldots + b_m X_m$
    - complicated by relationships between the predictors
- Logistic regression:
    - Dependent variable is dichotomous (nominal)
    - Prediction of dependent variable values probability
- If the relationship isn‘t linear:
    - We can try to transform the variables, so that the relationship becomes linear
    - We can divide the sample into subgroups in which the relationship is linear