STATISTICS

Association

Associations between criteria

Statistical criteria

- Qualitative criteria
- Quantitative criteria
- Temporal criteria
- Regional criteria

The population has been analyzed so far according to a single criterion. From now on, the population will be examined according to **two criteria**, i.e. in a combination table. The **purpose** of the analysis is **to determine the strength and sign of the relationship between the two criteria** examined.

Associations between criteria

- **The two criteria** (x and y) **are independent**, if belonging to x does not give extra information on belonging to y.

  (We do not deal with them.)

  (e.g. no. of storks vs no. of births; yield vs no. of movie-goers);

- There is a **stochastic association between the two criteria** (x and y), if on belonging to x we can conclude as a tendency on belonging to y $\rightarrow$ Statistics.

  (e.g. body weight vs body height; golden crown value of land vs yield);

- **The two criteria** (x and y) **are in a functional relationship**, if belonging to x clearly determines belonging to y $\rightarrow$ Mathematics

  (e.g. sugar beet harvest vs the amount of sugar produced; no. of movie-goers vs ticket sales);

Stochastic associations

Different criteria may have different effects:

x criterion: reason (explanatory variable)

y criterion: effect (resultant variable)

✓**reason-effect**

E.g. income level – meat consumption (qualification - unimployment, use of seat belts – accident severity);

? reason-effect

There are cases when the criteria mutually influence each other, namely the causal relationship is not clear, i.e. there is mutual causality (e.g. price - demand; family situation - alcohol consumption)

Conditions of causal relationship between two variables / criteria

a) the reason preceedes causation;

b) there is an empirical correlation between them;

c) this relationship is not the result of a third variable;

Can the direction of the association be interpreted? If yes, the relationship is positive or negative?

In case of nominal scale, for quality characteristics it cannot be interpreted**.**

1. e.g. gender – smoking;
2. e.g. qualification – position;

Direction: can be interpreted in case of quantitative criteria.

1. e.g. daily mean temperature – beer consumption: the hotter it is, in general, the more beer is consumed and vice versa. *Positive association*;
2. e.g. price – consumption: the more expensive the product, the less you usually buy. *Negative association;*

Real or pseudo relationship is it?

In the relationship of the variables, in the pseudo causal relationship, a third criterion plays role.

(E.g. TV sales increase vs growth in the number of divorces)

Associations between criteria

associative relationship: the associated criteria are quality related or territorial criteria (e.g. gender – smoking, gender – position, qualification – position);

mixed relationship: one criterion is regional or quality criterion and the other is quantitative criterion (e.g. qualificiation - per capita monthly income; gender - per capita monthly income; position - age);

correlation: both criteria are quantitative criteria (e.g. per capita monthly income - per capita monthly food consumption; age - per capita monthly income; learning time - marks); stochastic relationship can be examined simultaneously among multiple criteria;

rank correlation: both criteria can be measured on ordinal scale;

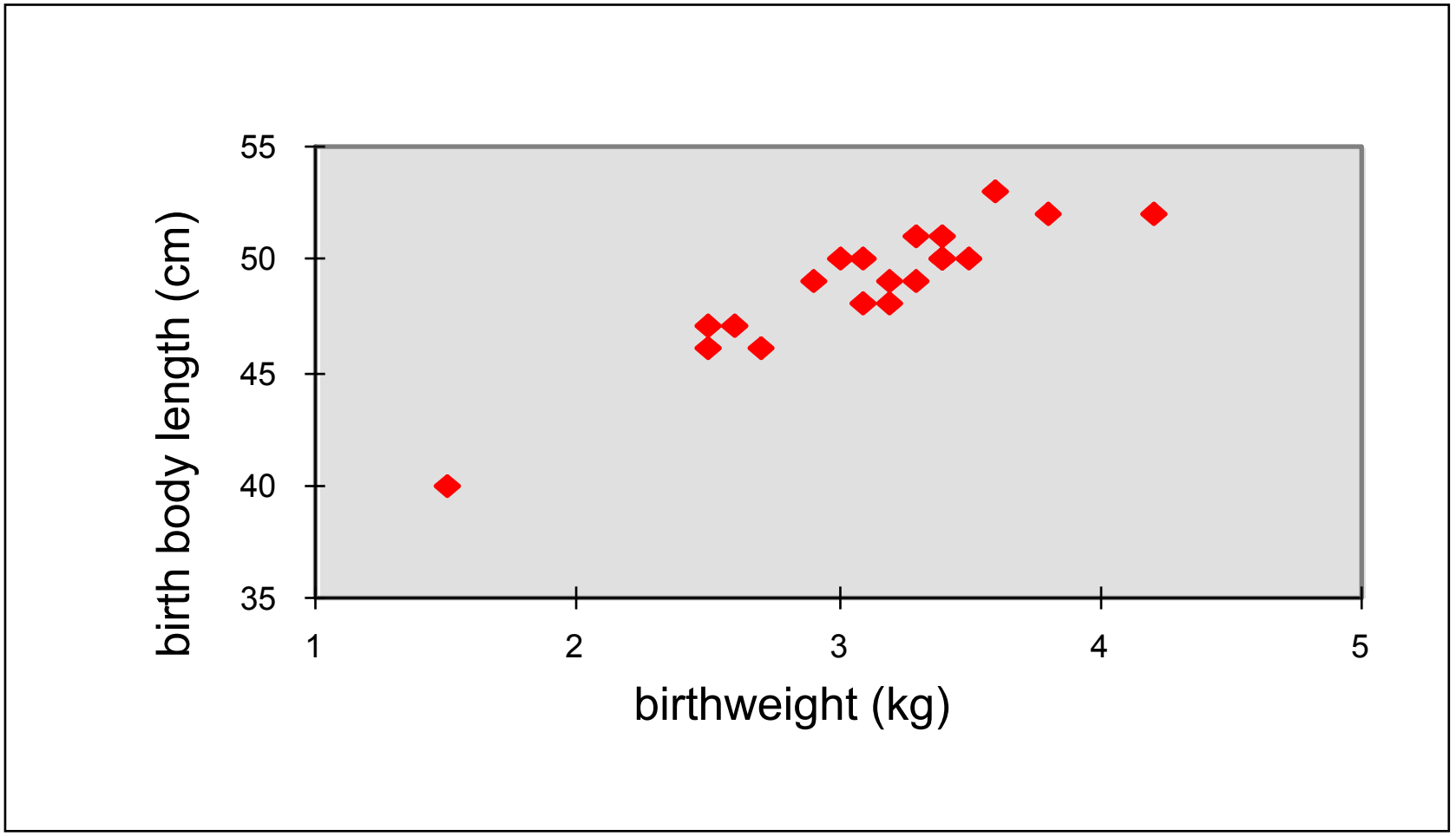
Associations between criteria

In case of **associative and mixed relationships** we can examine relationship between two criteria at a time.
We look for the answer whether:

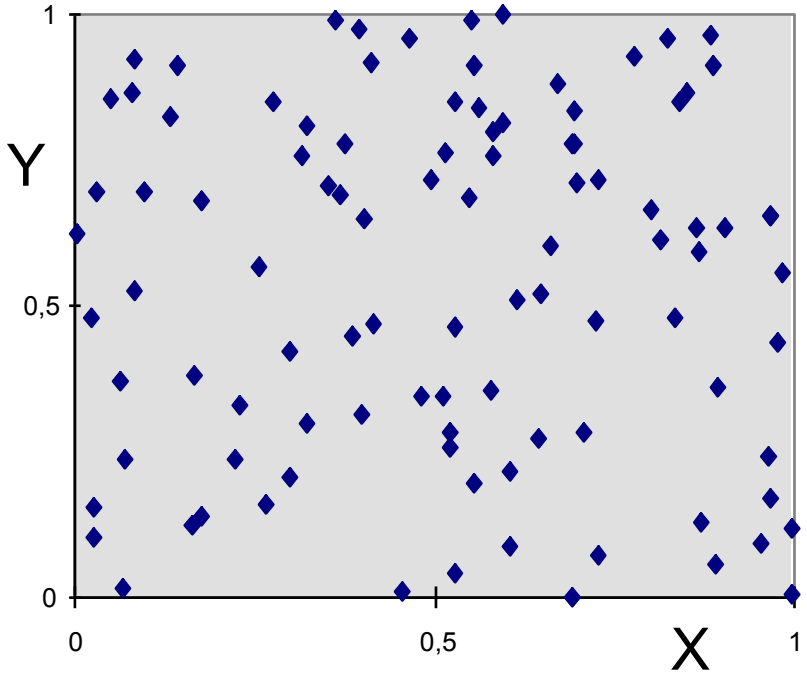
- ✓ there is a connection
- ✓ if there is a connection, how strong it is between the two criteria.

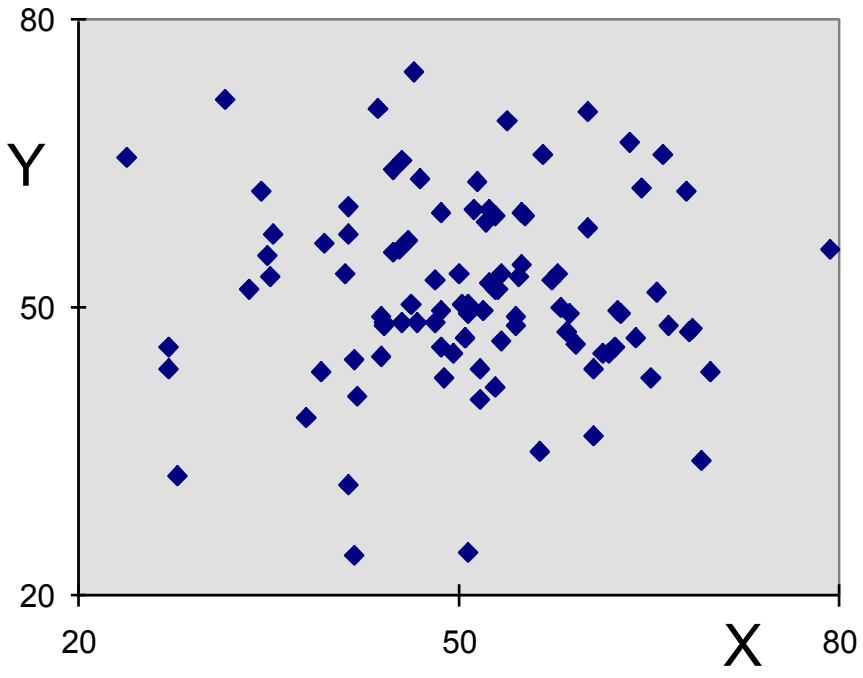
Correlation analysis (examination of relationship between quantitative criteria) provides more opportunity for analysis, because here we can also consider that one of the criteria how quantifies the other.

Whether birth body length depends on birthweight?
And vice versa?

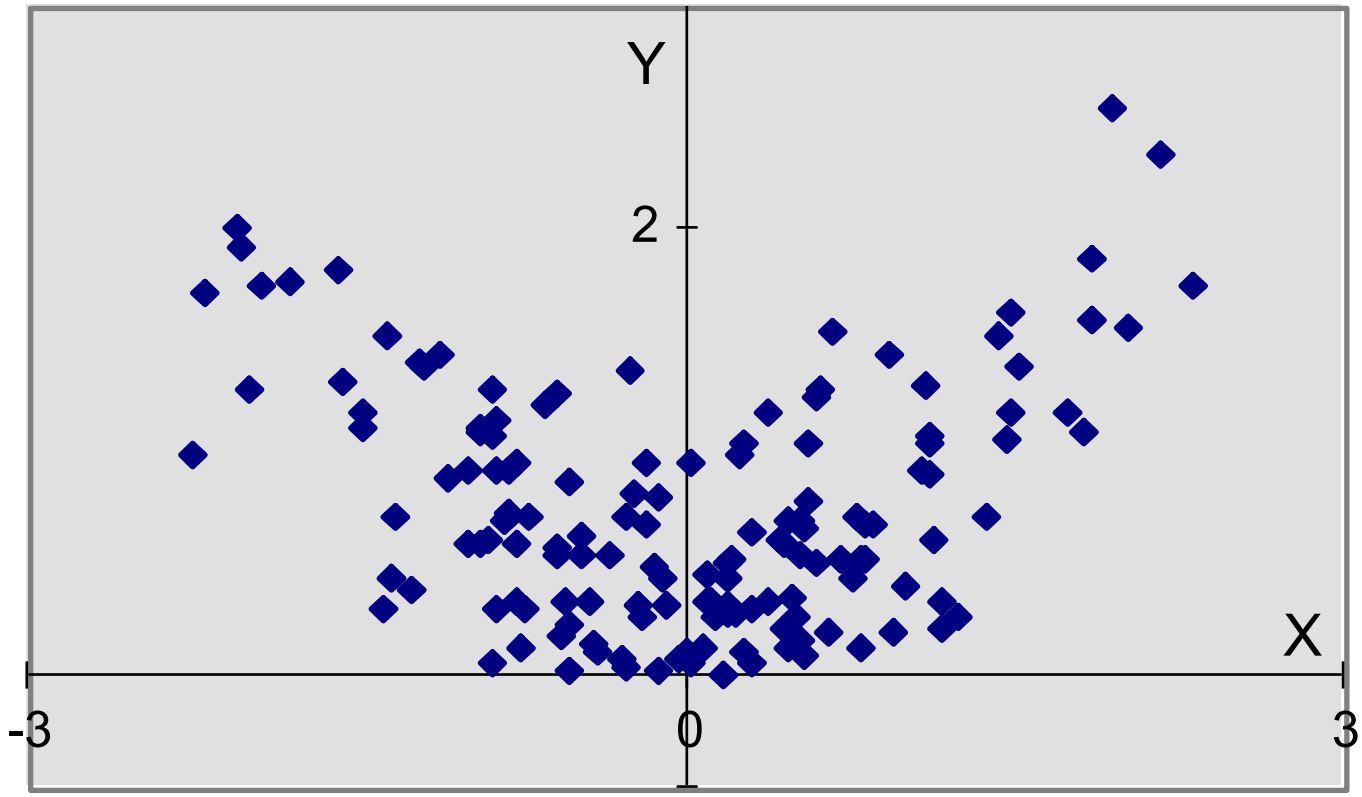


Whether Y variable depends on X variable?

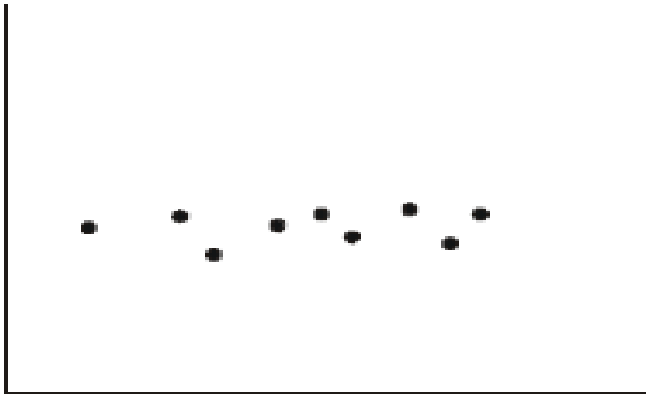




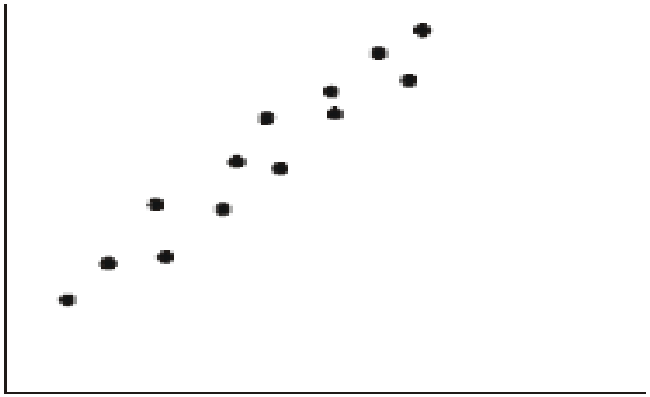
Whether Y variable depends on X variable?



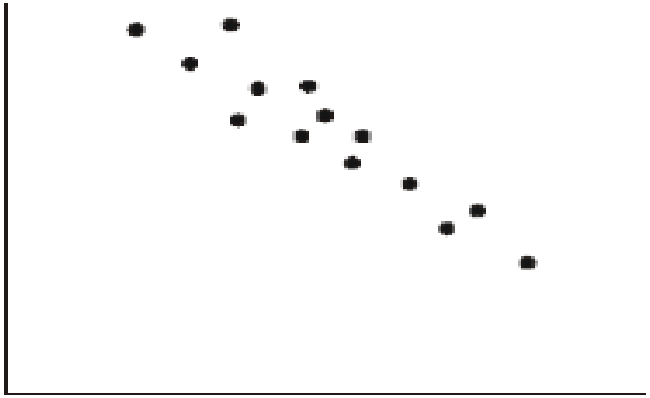
Scatter plots



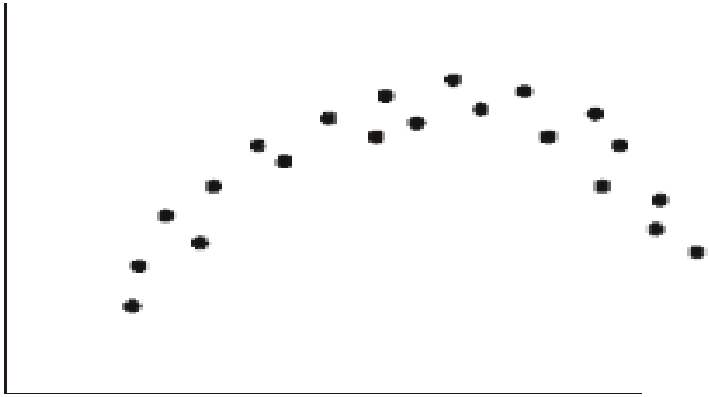
a) x and y are **independent** of each other [there is no one f(x) function for well estimating y]



b) there is a **positive** - linear - relationship between x and y



c) there is a **negative** - linear - relationship between x and y



c) there is a **nonlinear** relationship between x and y

Independence is mutual

IMPORTANT:

If Y is independent from X,

then X is also independent from Y

Contingency table

X/Y	1	2	…	j	…	t	Σ
1	f_{11}	f_{12}	…	f_{1j}	…	f_{1t}	$f_{1\cdot}$
2	f_{21}	f_{22}	…	f_{2j}	…	f_{2t}	$f_{2\cdot}$
.	.	.		.		.	.
.	.	.		.		.	.
.	.	.		.		.	.
i	f_{i1}	f_{i2}	…	f_{ij}	…	f_{it}	$f_{i\cdot}$
.	.	.		.		.	.
.	.	.		.		.	.
.	.	.		.		.	.
s	f_{s1}	f_{s2}	…	f_{sj}	…	f_{st}	$f_{s\cdot}$
Σ	$f_{\cdot 1}$	$f_{\cdot 2}$	…	$f_{\cdot j}$	…	$f_{\cdot t}$	n

Contingency table

$f_{ij} =$ joint frequencies, actual frequency in the ith row and jth column of the contingency table;

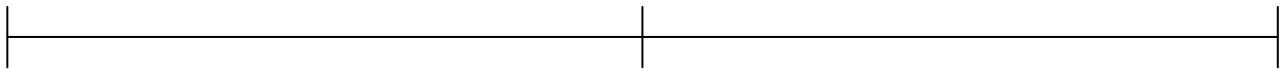
$f_{i.} =$ edge frequencies; all frequencies in the ith row; the number of elements with x variant;

$f_{.j} =$ edge frequencies; all frequencies in the jth column; the number of elements with y variant;

n = the number of elements of the population

Measuring the strength of the stochastic relationship

A general scheme to the indicators:



0

lower threshold independence

0,5

association of moderate strength

1

upper threshold

functional relationship

Requirements for indicators measuring strength of the relationship:

- they should have upper and lower limits;
- in case of total absence of relationship its value is 0;
- in case of functional relationship its value is 1;
- the value of the indicator does not depend on the number of observations;

Measuring the strength of associative relationships

1) In case of alternative criteria:

2 x 2 contingency table:

X/Y	y_1	y_2	Total
x_1	f_{11}	f_{12}	$f_{1.}$
x_2	f_{21}	f_{22}	$f_{2.}$
Total	$f_{.1}$	$f_{.2}$	n

In case of independence of the two criteria: $\frac{f_{11}}{f_{21}} = \frac{f_{12}}{f_{22}}$

Yule-coefficient (Y): $Y = \frac{f_{11}f_{22} - f_{21}f_{12}}{f_{11}f_{22} + f_{21}f_{12}}$ $\quad -1 \leq Y \leq 1$

Measuring the strength of associative relationships

Characteristics of Yule-coefficient:

$$0 \leq |Y| \leq 1$$

$Y = 0$ Total independence, i.e., total lack of relationship

$0 < |Y| < 1$ Stochastic relationship

$|Y| = 1$ Functional relationship

b) YULE-coefficient (Y)

In case of alternative criteria;

With regard to the prohibition of smoking opinion of 800 people, according to gender

Denomination	Permit		Forbid		Total	
Man (1)	440	f_{11}	60	f_{12}	500	$f_{1.}$
Woman (2)	160	f_{21}	140	f_{22}	300	$f_{2.}$
Total	600	$f_{.1}$	200	$f_{.2}$	800	n

$$\frac{f_{11}}{f_{21}} = \frac{f_{12}}{f_{22}}$$

$$f_{11} \times f_{22} - f_{21} \times f_{12} = 0$$

$$Y = \frac{f_{11} \times f_{22} - f_{21} \times f_{12}}{f_{11} \times f_{22} + f_{21} \times f_{12}}$$

$$Y = \frac{440 \times 140 - 160 \times 60}{440 \times 140 + 160 \times 60} = 0{,}73$$

The relationship is stronger than medium, namely more men than women support the authorization of smoking.

Measuring the strength of associative relationships

2) Commonly used indicator (both for alternative criteria and for criteria with more than two variants):

(where s is the no. of variants of one criterion, while t is the no. of variants of the other criterion): i.e., s = no. of raws; t = no. of columns; n = total no. of frequencies of the variants, namely: n = namely: $n = \sum f_{i.} = \sum f_{.j}$

Csuprov indicator (T):

$$T = \sqrt{\frac{\chi^2}{n \cdot \sqrt{(s-1)} \cdot \sqrt{(t-1)}}}$$ where $$\chi^2 = \sum_i \sum_j \frac{(f_{ij} - f^*_{ij})^2}{f^*_{ij}}$$

where $$f^*_{ij} = \frac{f_{i.} \cdot f_{.j}}{n}$$

$f_{i.}$ = total no. of frequency of the *i*-th row;
$f_{.j}$ = total no. of frequency of the *j*-th column;
$s \leq t$;

$f^*_{ij} =$ frequency, supposed in case of independence in the *i*-th row and *j*-th column of the contingency table

Measuring the strength of associative relationships

Characteristics of Csuprov indicator:

$\mathbf{0 \le T \le T_{max}}$

$\mathbf{s \le t}$

Cramer indicator (C) is used if:

$$0 \le C \le 1 \qquad C = \frac{T}{T_{\max}}, \text{ where } T_{\max} = \sqrt[4]{\frac{s-1}{t-1}} \quad \mathbf{s \le t}$$

$s = t = 2$ ← In case of alternative criteria, Y and T can also be applied. In this case the formula of T is as follows:

$$T = \frac{|f_{11} \cdot f_{22} - f_{12} \cdot f_{21}|}{\sqrt{f_{\cdot 1} \cdot f_{\cdot 2} \cdot f_{1\cdot} \cdot f_{2\cdot}}}$$

CRAMER indicator

Basic idea: in case of total independence, calculated frequencies f_{ij}^{*} are compared to actual frequencies f_{ij} ;

The bigger the difference, the stronger the relationship.

χ^2 value to be calculated:

$$\chi^2 = \sum_{i=1}^{s} \sum_{j=1}^{t} \frac{\left(f_{ij} - f_{ij}^*\right)^2}{f_{ij}^*}$$

Cramer coefficient:

$$C = \sqrt{\frac{\chi^2}{n(s-1)}} \qquad (s \le t)$$

$$0 \le C \le 1$$

C=0, if the two criteria are independent;

C=1, if the relationship of the two criteria is functional relationship;

Example:
(Distribution of white collar employees according to the occupational main group and gender, 2000, III-th quarter, thousand people)

Occupational main group	Man (1)	Woman (2)	Total
I. (1)	159	83.2	242.2
II. (2)	185.2	254.4	439.6
III. (3)	179.4	325.1	504.5
IV. (4)	17.6	235.4	253.0
Total	541.2	898.1	1439.3

Calculation of χ^2

Criterion	f_{ij}	f_{ij}^*	$\left(f_{ij} - f_{ij}^*\right)^2 / f_{ij}^*$
1.1.	159	91,1	50,6
2.1.	185,2	165,3	2,4
3.1.	179,4	189,7	0,6
4.1.	17,6	95,1	63,2
1.2.	83,2	151,1	30,5
2.2.	254,4	274,3	1,4
3.2.	325,1	314,8	0,3
4.2.	235,4	157,9	38,0
Total:	1439,3	1439,3	187,0

$$\chi^2 = 187$$

$$C = \sqrt{\frac{187}{1439,3(2-1)}} = 0,36$$

There is a weaker than medium relationship between the work area of white collar workers and gender.

Mixed relationship

Mixed relationship: one criterion is areal or qualitative criterion (namely not quantitative criterion), while the other is quantitative criterion (e.g. qualification – per capita gross monthly income);

When analyzing a mixed relationship, we examine whether to what extent the dispersion of the quantitative criterion is influenced by the classification according to the qualitative or the regional criterion.

Heterogeneous populations

They consist of complex and qualitatively different parts. Average of the heterogeneous population is a weighted average of the sub-population averages.

Indications:

$\bar{x}_j$: average of the j-th group

n_j : element no. of the j-th group

$j = 1, \ldots, m$: no. of groups

$$\bar{x} = \frac{\sum_{j=1}^{m} n_j \bar{x}_j}{\sum_{j=1}^{m} n_j} = \sum_{j=1}^{m} w_j \bar{x}_j$$

$\frac{n_j}{n} = w_j$: weight ratio

$\bar{x}$: average of the whole population

Heterogeneous populations

Indications:

n = element no. of the population

m = no. of groups

n_j = element no. of the *j*-th population

$\overline{x}_j$ = average of the *j*-th group

$\overline{x}$ = average of the whole population (main average)

x_{ij} = *i*-th element no. of the *j*-th population

Groups	Element no.	Mean of groups	*St. dev. of groups
C_1	n_1	$\bar{x}_1$	σ_1
C_2	n_2	$\bar{x}_2$	σ_2
…			
C_k	n_k	$\bar{x}_k$	σ_k
…			
C_m	n_m	$\bar{x}_m$	σ_m
Total	n	$\bar{x}$	σ

*Standard deviation

Example:
Properties to be sold at a rural real estate agency in a big city

Sale price (million HUF)	No. of flats made of prefabricated concrete slabs (PCS) (piece)	No. of non-PCS flats (piece)	Total no. of flats (piece)
6,1 – 8,0	8	2	10
8,1 – 10,0	15	5	20
10,1 – 15,0	34	12	46
15,1 – 20,0	24	14	38
20,1 – 25,0	7	19	26
25,1 – 30,0	2	8	10
Total	90	60	150
	$\bar{x}_p$	$\bar{x}_{np}$	$\bar{x}$
	σ_P	σ_{NP}	σ

$$\bar{x}_P = \frac{8 \cdot 7{,}0 + 15 \cdot 9{,}0 + 34 \cdot 12{,}5 + 24 \cdot 17{,}5 + 7 \cdot 22{,}5 + 2 \cdot 27{,}5}{90} = \frac{1\,248{,}5}{90} = 13{,}872$$

$$\bar{x}_{NP} = \frac{2 \cdot 7{,}0 + 5 \cdot 9{,}0 + 12 \cdot 12{,}5 + 14 \cdot 17{,}5 + 19 \cdot 22{,}5 + 8 \cdot 27{,}5}{60} = \frac{1\,101{,}5}{60} = 18{,}358$$

$$\bar{x} = \frac{10 \cdot 7{,}0 + 20 \cdot 9{,}0 + 46 \cdot 12{,}5 + 38 \cdot 17{,}5 + 26 \cdot 22{,}5 + 10 \cdot 27{,}5}{150} = \frac{2\,350}{150} = 15{,}67$$

$$\sigma_P = \sqrt{\frac{8 \cdot (7{,}0 - 13{,}872)^2 + 15 \cdot (9{,}0 - 13{,}872)^2 + \ldots + 2 \cdot (27{,}5 - 13{,}872)^2}{90}} = \sqrt{\frac{2\,006{,}3}{90}} = 4{,}72$$

$$\sigma_{NP} = \sqrt{\frac{2 \cdot (7{,}0 - 18{,}358)^2 + 5 \cdot (9{,}0 - 18{,}358)^2 + \ldots + 8 \cdot (27{,}5 - 18{,}358)^2}{60}} = \sqrt{\frac{2\,112{,}55}{60}} = 5{,}93$$

$$\sigma = \sqrt{\frac{10 \cdot (7{,}0 - 15{,}67)^2 + 20 \cdot (9{,}0 - 15{,}67)^2 + \ldots + 10 \cdot (27{,}5 - 15{,}67)^2}{150}} = \sqrt{\frac{4\,843{,}335}{150}} = 5{,}68$$

For easier viewing, calculation of the sold housing is tabulated:

Sold houses			
Type of flat	**No. of properties**	**Avarage sales price of housing groups**	***St. dev. of sales price of housing groups**
PCS	**90**	**13872**	**4.72**
Non-PCS	**60**	**18358**	**5.93**
Total	**150**	**15670**	**5.68**
	n	$\bar{x}$	σ

*Standard deviation

Indications

$x_{ij} - \bar{x}$ = full difference (d_{ij})

$(x_{ij} - \bar{x}_j)$ = internal difference (B_{ij})

$(\bar{x}_j - \bar{x})$ = external difference (K_{ij})

σ^2 = full variance

σ_B^2 = internal variance

σ_K^2 = external variance

Calculation of variance

$$\sigma^2 = \frac{\sum_{i=1}^{n_j}\sum_{j=1}^{m}(x_{ij}-\bar{x})^2}{n} = \frac{S}{n}$$

S: Total sum of squares

$$\sigma_B^2 = \frac{\sum\sum(x_{ij}-\bar{x}_j)^2}{n} = \frac{\sum n_j\sigma_j^2}{n} = \frac{S_B}{n}$$

S_B: Internal sum of squares

$$\sigma_K^2 = \frac{\sum n_j(\bar{x}_j-\bar{x})^2}{n} = \frac{S_K}{n}$$

S_K: External sum of squares

Relationships

$$d_{ij} = B_{ij} + K_{ij}$$

$$x_{ij} - \bar{x} = (x_{ij} - \bar{x}_j) + (\bar{x}_j - \bar{x})$$

full difference internal difference external difference

$$\sigma^2 = \sigma_B^2 + \sigma_K^2$$

full variance internal variance external variance

$$S = S_B + S_K$$

Total sum of squares Internal sum of squares External sum of squares

Example:

In a college, bachelor training occurs in 4 professions. The time spent on the students' daily learning is as follows.

Profession	Time spent on daily learning (hr)		Students (%)
	Mean x_j	*St. deviation σ_j	
Human resources	1,5	1,2	24
Management	2,25	0,8	26
International management	1,75	1,5	20
Finance & accounting	2,75	1,3	30

*Standard deviation

Calculate $\sigma_B, \sigma_K, \sigma$ and interpret them!

Solution

$$\bar{x} = 0{,}24 \cdot 1{,}5 + 0{,}26 \cdot 2{,}25 + 0{,}2 \cdot 1{,}75 + 0{,}3 \cdot 2{,}75 = 2{,}12$$

$$\sigma_K^2 = 0{,}24 \cdot (1{,}5 - 2{,}12)^2 + \ldots + 0{,}3 \cdot (2{,}75 - 2{,}12)^2 = 0{,}2431$$

$$\sigma_k = 0{,}49$$

$$\sigma_B^2 = 0{,}24 \cdot 1{,}2^2 + 0{,}26 \cdot 0{,}8^2 + 0{,}2 \cdot 1{,}5^2 + 0{,}3 \cdot 1{,}3^2 = 1{,}469$$

$$\sigma_B = 1{,}212$$

$$\sigma^2 = \sigma_B^2 + \sigma_K^2$$

$$\sigma^2 = 1{,}469 + 0{,}2431 = 1{,}7121 \rightarrow \sigma = 1{,}308$$

Indicators of mixed relationships

Variance-ratio: show that to what extent (in percentage) the classification of a quality or areal criterion affects dispersion of a quantitative criterion.

$$H^2 = \frac{\sigma_K^2}{\sigma^2} = 1 - \frac{\sigma_B^2}{\sigma^2} = \frac{S_K}{S} = 1 - \frac{S_B}{S}$$

Quotient of standard deviations (square root of variance-ratio): shows that how strong is the relationship between the non-quantitative (grouping) and quantitative criteria.

$$H = \sqrt{H^2} = \sqrt{\frac{\sigma_K^2}{\sigma^2}} = \frac{\sigma_K}{\sigma} = \sqrt{1 - \frac{\sigma_B^2}{\sigma^2}} = \sqrt{\frac{S_K}{S}} = \sqrt{1 - \frac{S_B}{S}}$$

Interpretation of the indicators of the mixed relationships

$$0 < H < 1$$
$$0 < H^2 < 1$$

Stochastic relationships

$$H = H^2 = 0$$

Total independence, total lack of relationship

$$H = H^2 = 1$$

Functional, deterministic relationship

CORRELATION AND REGRESSION

Basic concepts

- The relationship between quantitative criteria is called correlation;
- **Correlation analysis:** measurement of the strength of relationship between quantitative criteria;
- **Regression analysis:** deals with quantification of the impact of the quantitative criteria on each other, as well as with the direction and extent of these impacts;

If the correlation is based on a one-way causal relationship:

- the criterion as cause is called ***explanatory variable*** *(X)*;
- the criterion as effect is called ***resultant variable*** *(Y)*;

Indicators of the strength of the relationship

Covariance

It shows the direction of the relationship between X and Y quantitative variables.

It is based on the differences from the means of X and Y variables, namely on $x - \bar{x}$ and $y - \bar{y}$

$$d_x = x - \bar{x} \qquad d_y = y - \bar{y}$$

$$C = \frac{\Sigma d_x d_y}{n-1} = \frac{\Sigma xy}{n-1} - \bar{x} \cdot \bar{y}$$

$$C = r \cdot s_x \cdot s_y$$

s_x: standard deviation of x; s_y: standard deviation of y; r: correlation coefficient between x and y;

Characteristics of covariance

- The sign of the covariance shows the direction of the relationship.
- There is no upper limit of the absolute value of the covariance.
- The absolute value of the covariance is maximum, if there is a linear relationship between x and y, namely:

$$\left|C_{\max}\right| = \sigma_x \cdot \sigma_y$$

- The covariance is symmetrical in the two variables, i.e. X and Y are interchangeable in the formula.
- If the criteria are independent, then C = 0. (However, this is not true conversely: if C = 0, then the relationship is uncorrelated, but not necessarily independent. Independence means more severe conditions than uncorrelation.)

- Since C is unit-dependent, so it is advisable to divide by the maximum value, and you get a signed indicator: namely the linear correlation coefficient:

$$r_{xy} = \frac{C_{xy}}{\sigma_x \cdot \sigma_y}$$ it can also be calculated: $$r_{xy} = \frac{\sum d_x d_y}{\sqrt{\sum d_x^2 \cdot \sum d_y^2}}$$

- This indicator measures the strength of the linear relationship, so if it shows a value close to 0, you may have a linear relationship, but weak; but it is also possible that the relationship is strong, but it is not linear.

- In a case of one variable, the covariance regarding itself is variance. The variance is therefore a special case of covariance:

$$C_{xx} = \sigma_x^2$$

Earnings and monthly saving of a company's employees

Worker	Wage (x) (HUF / worker)	Monthly savings (y) (HUF/month)	d_x	d_y	$d_x d_y$	d_x^2	d_y^2
1	120000	13000	-13000	-3010	39130000	169000000	9060100
2	90000	10000	-43000	-6010	258430000	1849000000	36120100
3	220000	35000	87000	18990	1652130000	7569000000	360620100
4	150000	18000	17000	1990	33830000	289000000	3960100
5	100000	12000	-33000	-4010	132330000	1089000000	16080100
6	115000	12500	-18000	-3510	63180000	324000000	12320100
7	160000	20000	27000	3990	107730000	729000000	15920100
8	130000	13800	-3000	-2210	6630000	9000000	4884100
9	145000	14000	12000	-2010	-24120000	144000000	4040100
10	100000	11800	-33000	-4210	138930000	1089000000	17724100
Total	**1330000**	**160100**	0	0	2408200000	13200000000	480729000

$$\bar{y} = \frac{13000 + 10000 + \cdots + 14000 + 11800}{10} = 16010$$

$$\bar{x} = \frac{120000 + 90000 + \cdots + 145000 + 100000}{10} = 133000$$

Covariance

$$C = \frac{\Sigma d_x d_y}{n-1} = \frac{\Sigma xy}{n-1} - \bar{x} \cdot \bar{y} = \frac{2408200000}{9} = 267577777{,}8$$

Interpretation: the relationship between the earnings of the employees and the monthly amount saved is positive.

Correlation coefficient (r)

- Correlation coefficient (r) is the most important measure of the strength of linear correlations.
- The lack of relationship (uncorrelated) is indicated by the value r = 0.
- The sign of *r* shows the direction of the correlation. Functional linear relationship – depending on the direction – fits either r = +1, or r = -1, respectively.
- Between extreme positions, the absolute value of the coefficient informs on the strength of the relationship.

Correlation coefficient (r)

$$\mathbf{r} = \frac{\mathbf{C}}{s_{\mathbf{x}} \cdot s_{\mathbf{y}}} = \frac{\sum \mathbf{d_x d_y}}{\sqrt{\sum \mathbf{d_x}^2 \sum \mathbf{d_y}^2}} = \frac{\sum \mathbf{xy} - \mathbf{n} \cdot \bar{\mathbf{x}} \cdot \bar{\mathbf{y}}}{\sqrt{(\sum \mathbf{x}^2 - \mathbf{n} \cdot \bar{\mathbf{x}}^2)(\sum \mathbf{y}^2 - \mathbf{n} \cdot \bar{\mathbf{y}}^2)}}$$

$$\Sigma d_x \cdot d_y = \Sigma xy - n \bar{x} \bar{y}$$

$$\Sigma d_x^2 = \Sigma x^2 - n \bar{x}^2$$

$$\Sigma d_y^2 = \Sigma y^2 - n \bar{y}^2$$

Correlation coefficient (r)

Worker	Wage (HUF / worker)	Monthly savings (HUF/month)	$\sum d_x$	$\sum d_y$	$\sum d_x d_y$	$\sum d_x^2$	$\sum d_y^2$
Összesen	1330000	160100	0	0	2408200000	13260000000	480729000

$$r = \frac{C}{s_x \cdot s_y} = \frac{\sum d_x d_y}{\sqrt{\sum d_x^2 \sum d_y^2}} = \frac{2408200000}{\sqrt{13260000000 \cdot 480729000}} = 0{,}954$$

Interpretation: the relationship between the earnings of the employees and monthly amount saved is positive and strong.

Coefficient of determination (r^2)

- The coefficient of determination shows that to what extent (in percentage) the explanatory variables affects the dispersion of the dependent variable.
- Indication: r^2
- The coefficient of determination characterizes:
  - ✓ the fit of the regression function,
  - ✓ the explanatory power of the model.

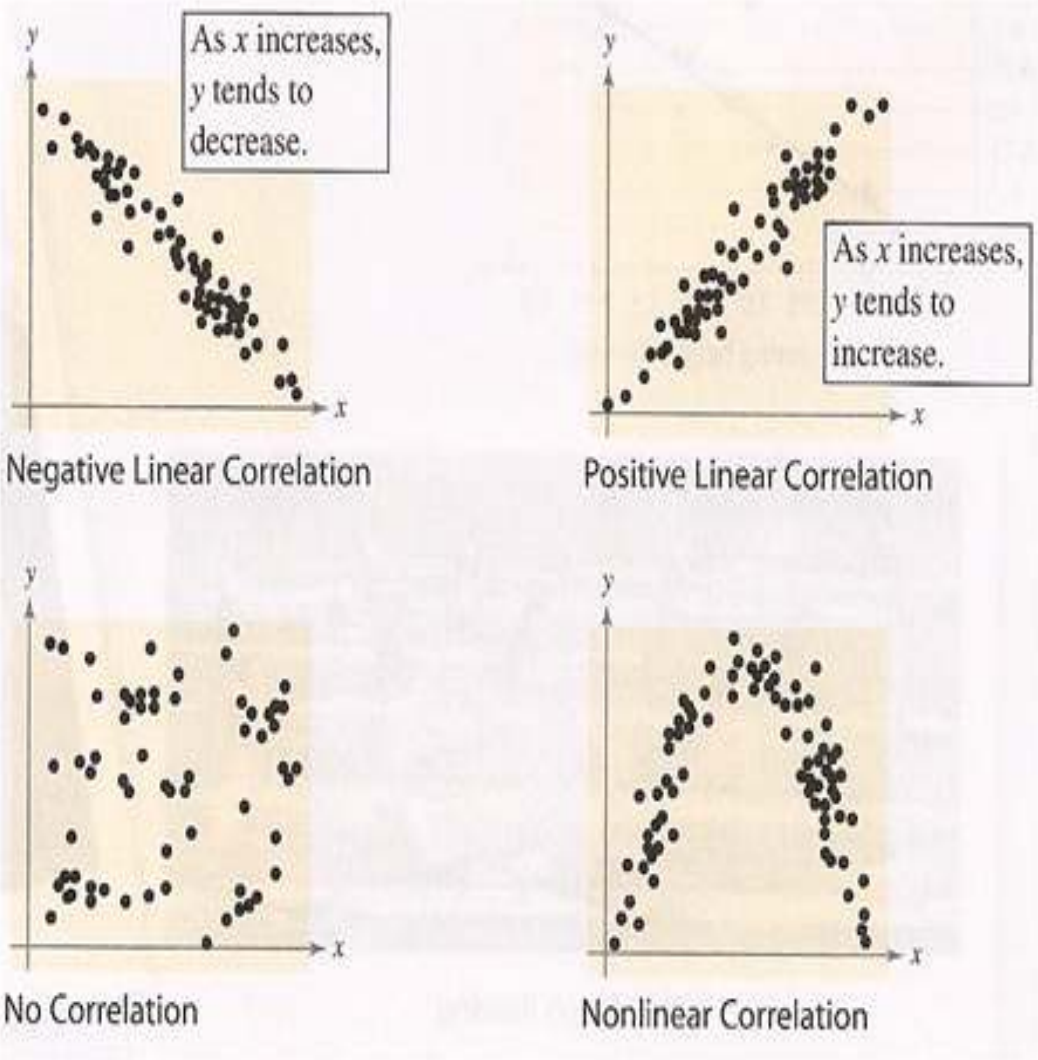
Coefficient of determination (r^2)

$$r^2 = 0,954^2 = 0,9098 = 90,98\%$$

Interpretation: the earnings of workers influence in 90.98% the dispersion of the monthly amount saved.

Correlation

- Representation of pairs of (x, y) points belonging together;
- If there is an imaginary line, along which the point pairs occur → linear correlation;
- Depending on the direction of the relationship: positive or negative correlation;
- If there is no such a line → variables are uncorrelated (but not necessarily independent!)



The most important characteristics of the linear correlation coefficient

1. **If there is no linear correlation, then the correlation coefficient is 0, while in case of functional relationship, the value of the correlation coefficient is +1.00 or -1.00.**
2. **The value of the correlation coefficient is independent from the units [eg. the correlation between the body height and body weight is independent from the units of the variables (kg, pounds, cm, inch)].**
3. **The correlation coefficient is symmetric** (correlation of x with y = correlation of y with x), that is r(x, y) = r(y, x);
4. **The linear correlation coefficient measures the linear relationship,** not the relationship in general;

The most important characteristics of the linear correlation coefficient

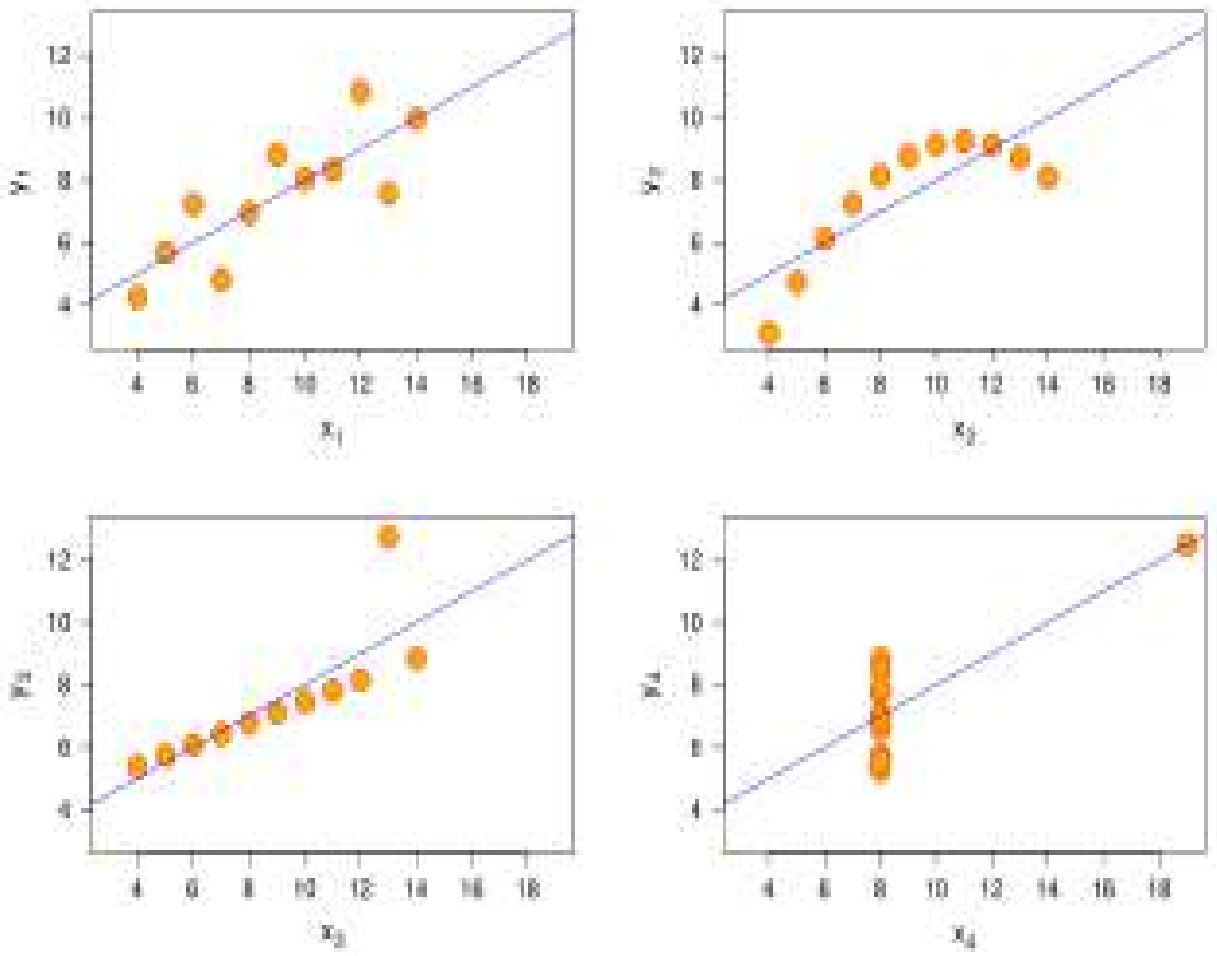
5. **The value of the correlation coefficient is strongly influenced by outliers. The outlier can be**
   - **a result of an irrgeular, distorted distrubiton ⇒ transformation;**
   - **measurement error ⇒ repeat of the measurement / exclusion of the value;**
6. **The correlation does not necessarily imply a causal relationship, since the**
   - **the variable x may influence the variable y;**
   - **the variable y may influence the variable x;**
   - **a third factor may affect both x and y**
     - ✓ **to one direction (positive correlation), or**
     - ✓ **to different directions (negative correlation);**

Outliers, linearity

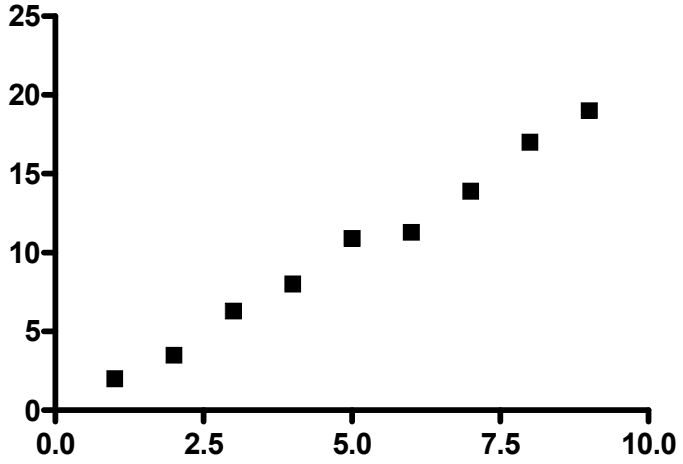
- Regression equation:

  y=0.5x + 3
- r = 0.816

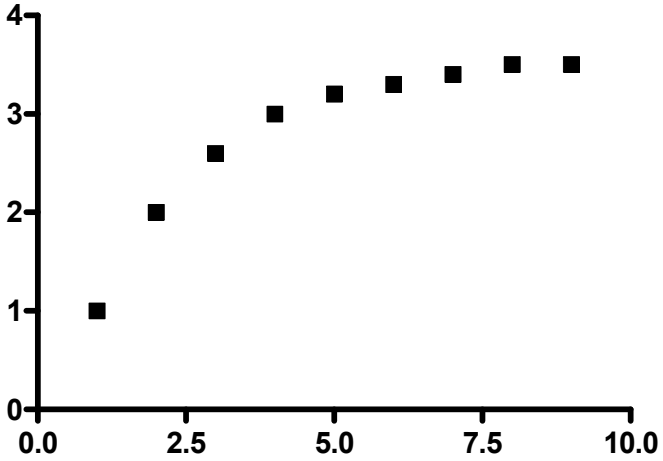
- 2. non-linear relationship!
- Without outlier
  - 3. r=1
  - 4. r=0



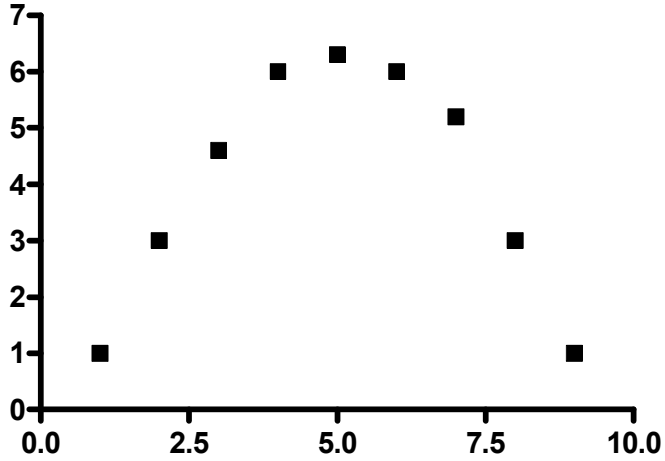
a


25
20
15
10
5
0
0.0
2.5
5.0
7.5
10.0

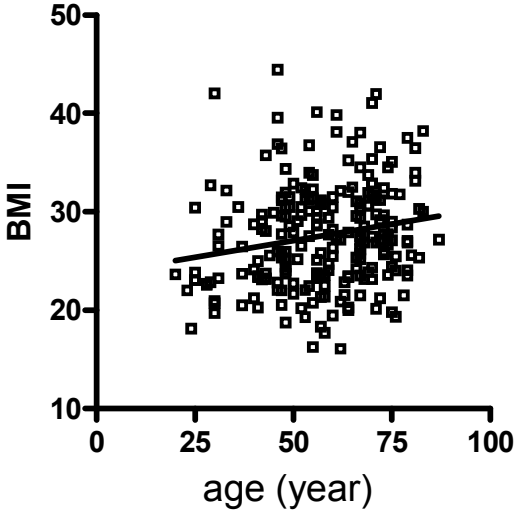
b


4
3
2
1
0
0.0
2.5
5.0
7.5
10.0

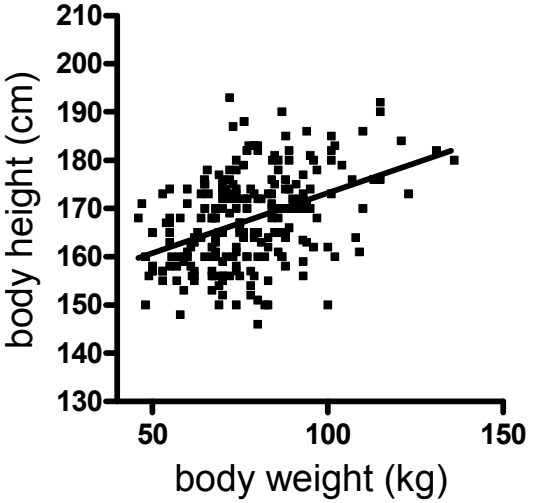
c


7
6
5
4
3
2
1
0
0.0
2.5
5.0
7.5
10.0

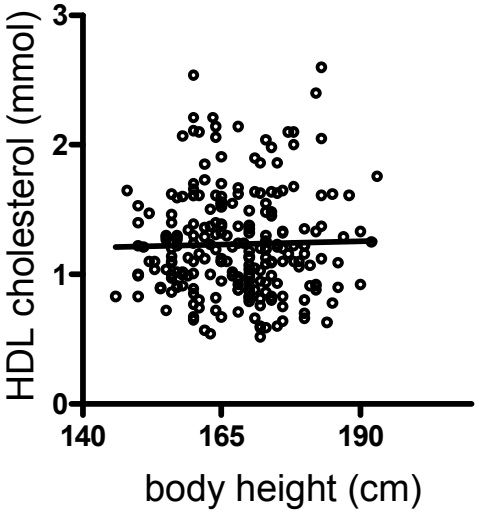
weak positive relationship



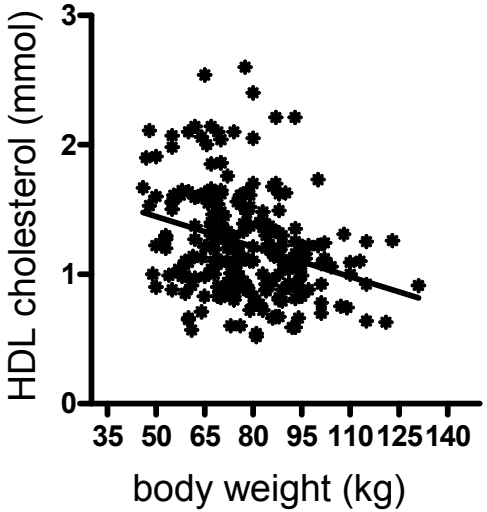
strong positive relationship



no relationship



strong negative relationship



What to do?

Outliers

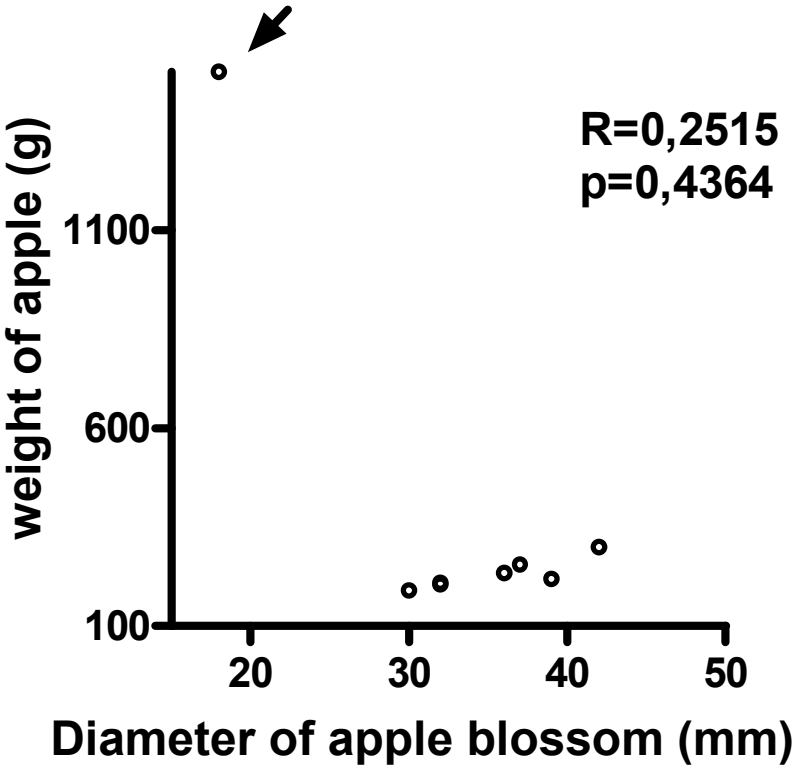
- It is subjective, what the outlier is (usually outside of 2 standard deviations);
- Check the database
  - Really fair value? Batting? Measurement error?
- If real data – unique evaluation:
  - it is inadvisable automatically to exclude;
  - if it really distorts the overall picture, it is possible;
- Check if outliers have significance. Are they surely outliers, or just they do not fit our theory?

Non-linearity

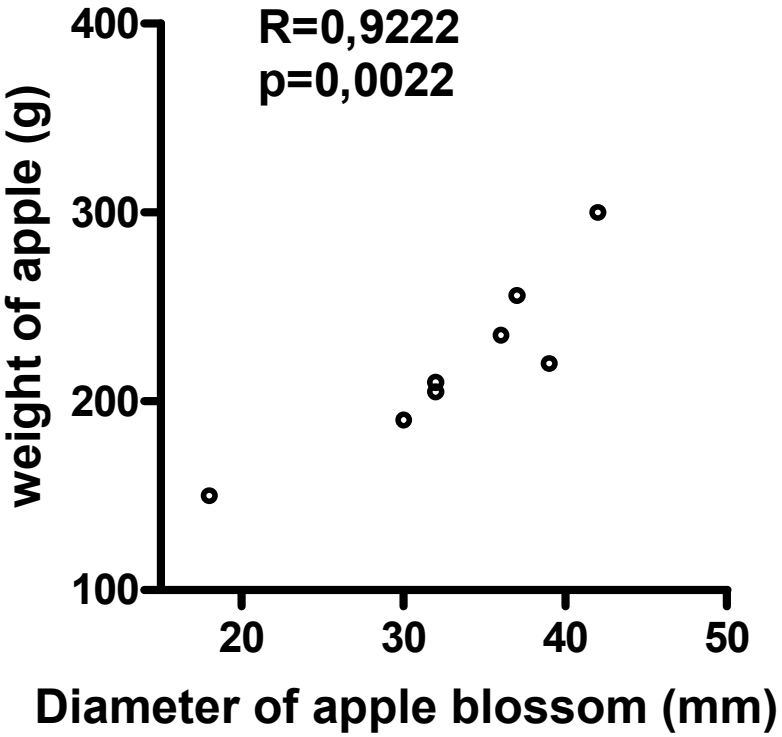
- If not monotonous, no sense of correlation. If monotonous …
- Linear transformation can be made? [e.g. logarithmic transformation – at first the figure scaling can be tested (Axis/Scaling)];
- Performing a non-parametric test (Spearman rank correlation test);
  - ✓ less sensitive;
- A function should be looked for that matches it, and describes it correctly;
- Along one variable the sample is shared into 4-5 groups of equal width. ANOVA is performed so that this variable is the grouping variable.

THE IMPACT OF AN OUTLIER TO THE STRENGTH AND SIGNIFICANCE OF THE CORRELATION COEFFICIENT

An outlier from the eight value



After eliminating the outlier



Evaluation of the strength of the correlation (relationship between the two variables). A simplificated solution.

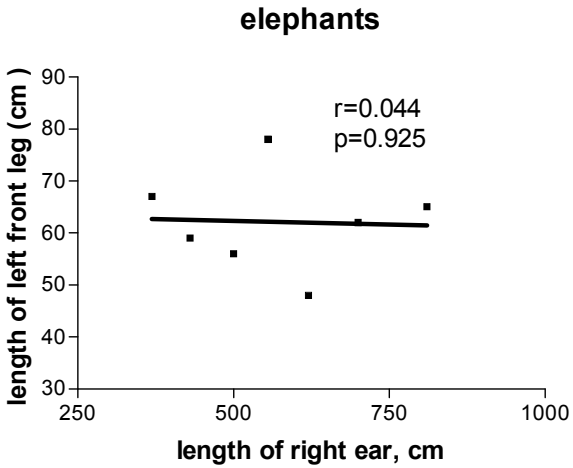
Correlation coefficient	Strength of the relationship
0.00-0.25	No, or very weak
0.25-0.50	Weak
0.50-0.75	Moderately strong or strong
0.75-1.00	Very strong

ATTENTION! A correlation coefficient of higher than |0.95 | is suspected. It suggests that one of the values follows from the other, or determined by the other.

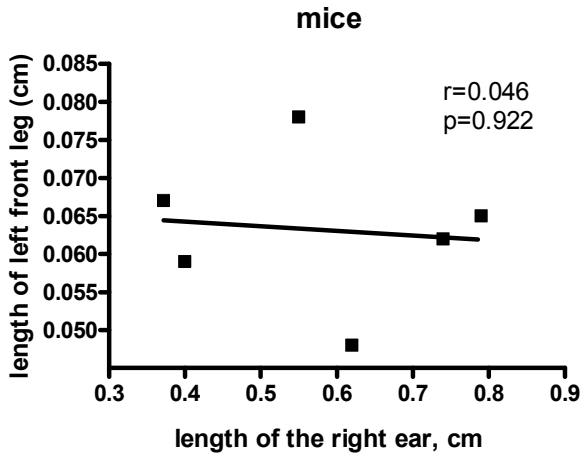
Conditions of calculating linear (Pearson) correlation coefficient

- **Versions of the criteria variants are selected randomly of a larger population**;
- **Observations** should be **independent of each other**;
- **A sample should never be selected from different populations,** because it will show a false-significant correlation, nevertheless there is no relationship between the two variables either in the one, or in the other sample;
- **Both x and y samples should be selected from a population of normal distribution;**
  - If this is not the case, non-parametric procedure (Spearman correlation coefficient) should be performed and calculated.

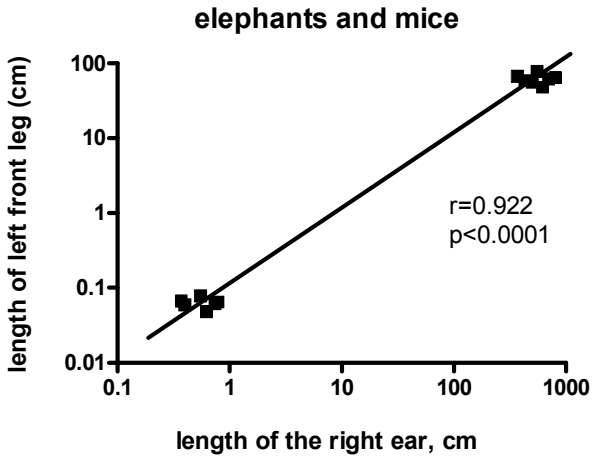
HOW IS IT PROHIBITED CALCULATING LINEAR CORRELATION COEFFICIENT?



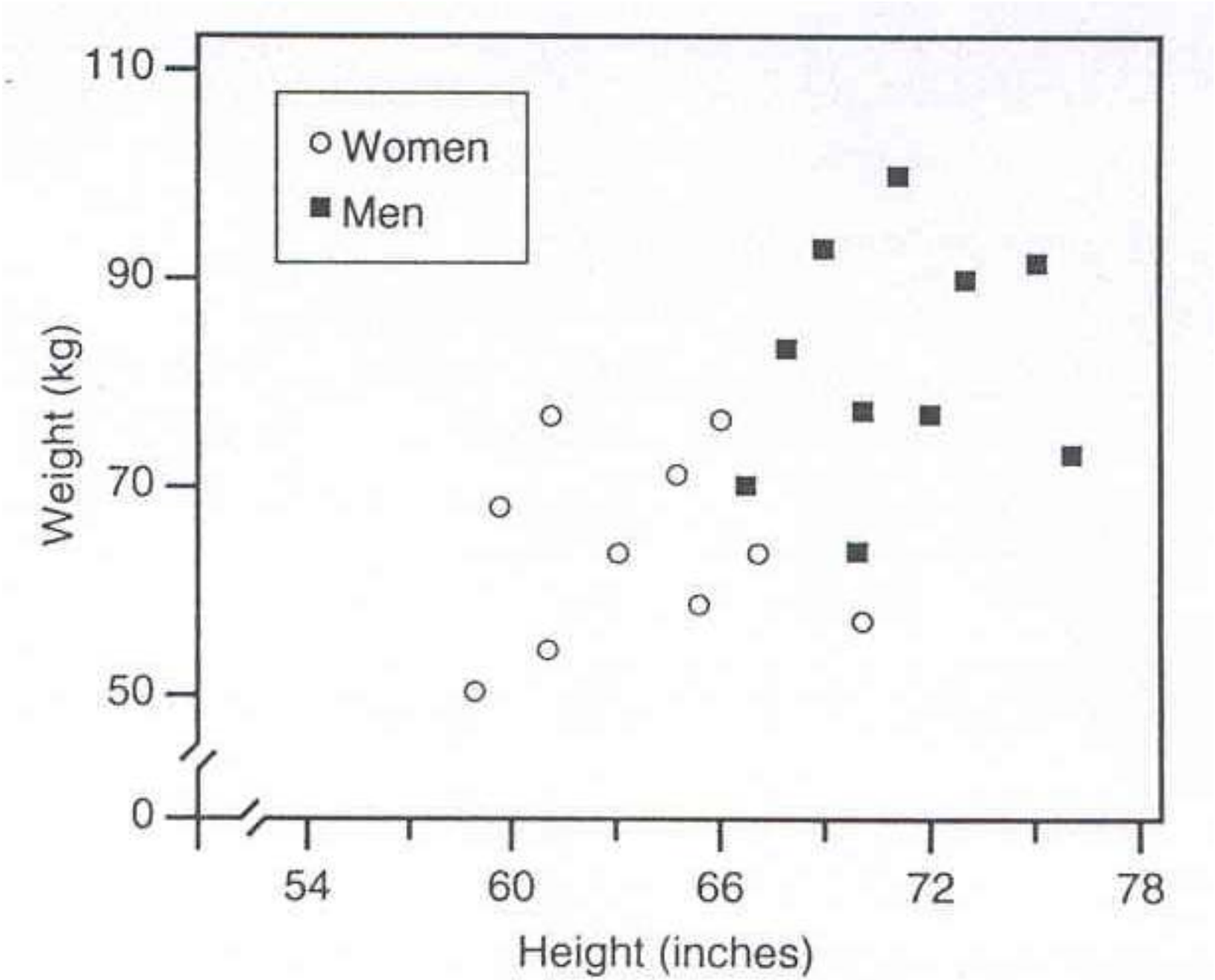
(L. E. Phant et al.: Big Animals, 2004;25:23-45)



(B. Hamster, P. Rat: Small Animals 1998;234:56-78)

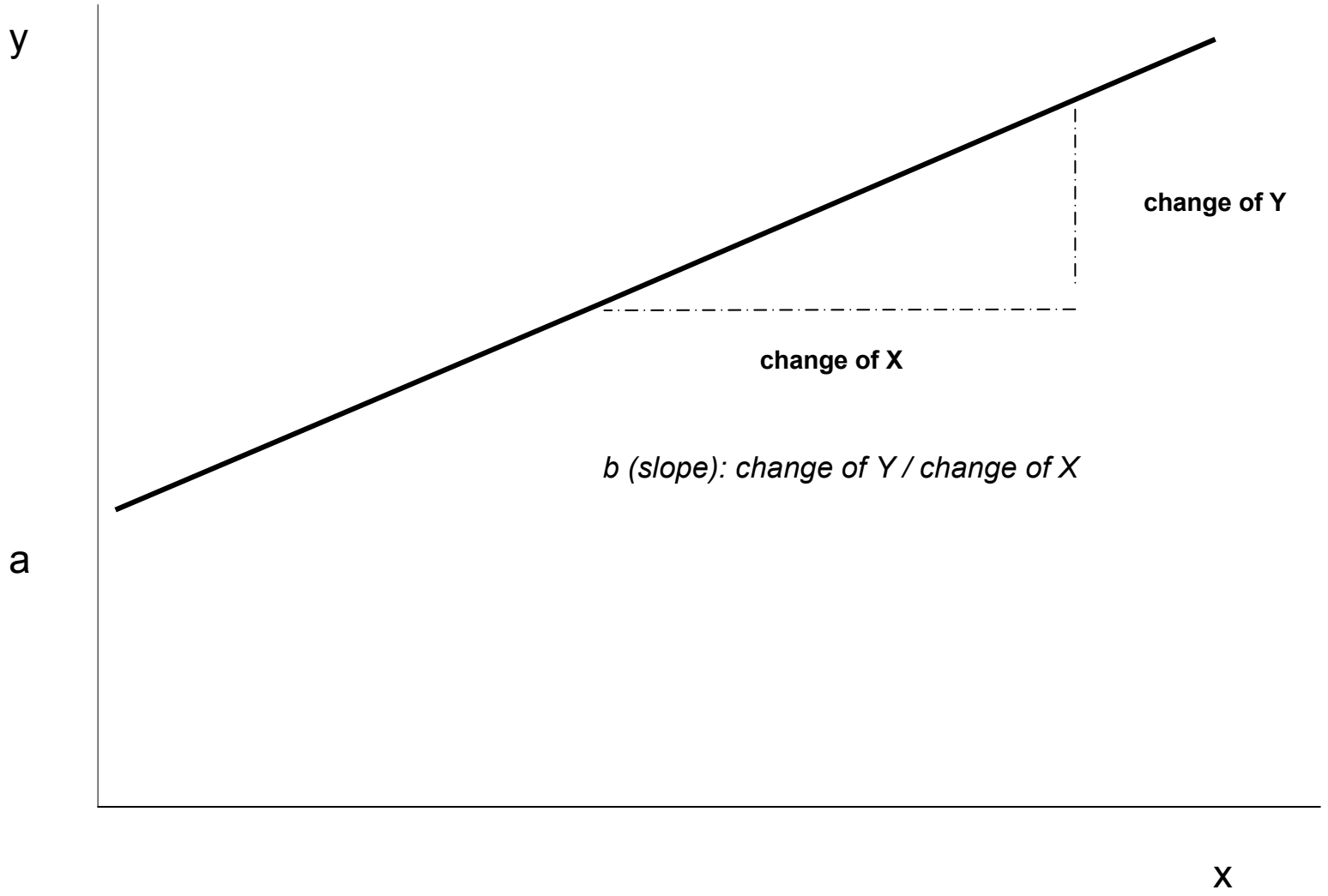


(G. Swine et al., unpublished)


110
90
70
50
0
Weight (kg)
Women
Men
54
60
66
72
78
Height (inches)

REGRESSION

- Regression shows the relationship between two variables so that at the same time it also expresses the dependency rate of one variable (independent variable) to another variable (dependent variable), as well;
- linear and non-linear regression;
- simple and multiple regression;

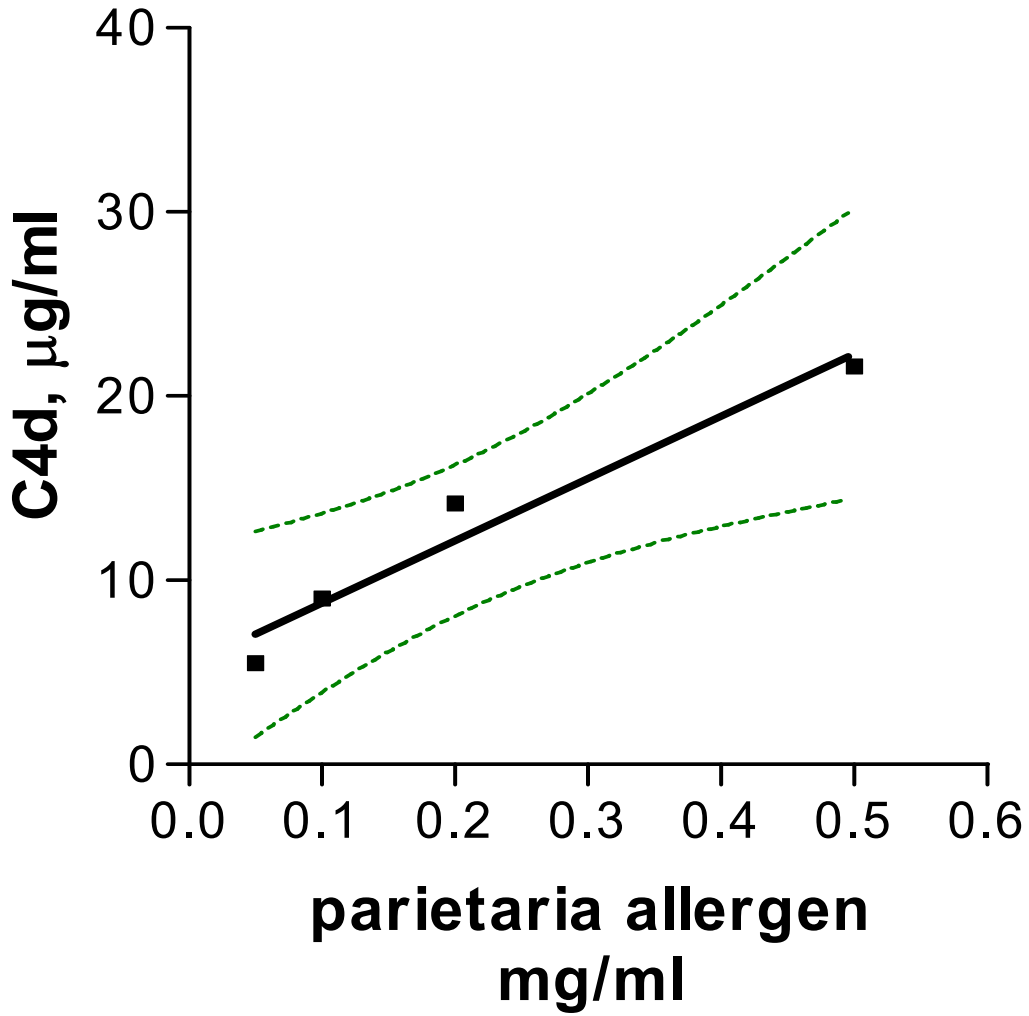

y
change of Y
change of X
b (slope): change of Y / change of X
a
x

- The fact of **linear regression analysis** is to draw such **a straight line, distance of which is the smallest from the measuring points**, i.e. it approximates them most closely (best fit regression line, i.e. **the least squares method**).

- The vertical distances between the points and the line are called **residuals**. The sum of the squares of the residuals is the **variance of the residuals**, the square root of which is the **standard deviation of the residuals**.

The **regression line** is the line for which **the standard deviation of residuals is minimum**.

Confidence interval of the regression line can also be determiend.



Equation of the regression line

$$y = ax + b$$

$$a = r \cdot \frac{\sigma_y}{\sigma_x}$$

$$b = \bar{y} - a \cdot \bar{x}$$

Does the trend coefficient differ significantly from zero?

In order to answer this question, the trend coefficient received should be divided by the standard error of the sample elements. This ratio is the „A" probe statistics (see the formula below) that is necessary to carry out the t-test:

Standard error, SE (standard deviation of the mean):

$$SE = \frac{s}{\sqrt{n}}$$

$$A = \frac{a \cdot \sqrt{(n-2)\sum_{i=1}^{n}(x_i - \overline{X})^2}}{\sqrt{\sum_{i=1}^{n}(y_i - \overline{Y})^2}}$$

s = standard deviation of the sample;
n = element number of the sample;

During the analysis, the value of the „A" test statistic is compared to the critical value of the t-distribution with degrees of freedom of n-2 (n = sample size). The probability level of 0.95 is generally selected.

If $A > |t|_{0,95}^{n-2} \Rightarrow$ *a* (linear trend coefficient) significantly differs from 0;

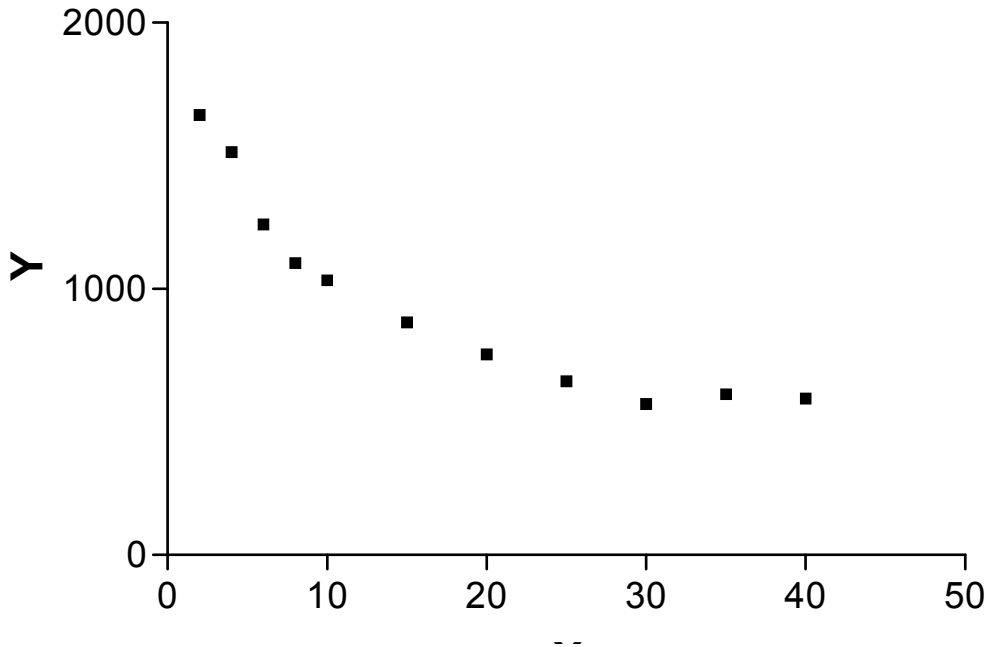
Table of the Student t-distribution

Sz.f.	0,55	0,6	0,7	0,75	0,8	0,9	0,95	0,975	0,99	0,995
1	0,158	0,325	0,727	1,000	1,376	3,078	6,314	12,706	31,821	63,656
2	0,142	0,289	0,617	0,816	1,061	1,886	2,920	4,303	6,965	9,925
3	0,137	0,277	0,584	0,765	0,978	1,638	2,353	3,182	4,541	5,841
4	0,134	0,271	0,569	0,741	0,941	1,533	2,132	2,776	3,747	4,604
5	0,132	0,267	0,559	0,727	0,920	1,476	2,015	2,571	3,365	4,032
6	0,131	0,265	0,553	0,718	0,906	1,440	1,943	2,447	3,143	3,707
7	0,130	0,263	0,549	0,711	0,896	1,415	1,895	2,365	2,998	3,499
8	0,130	0,262	0,546	0,706	0,889	1,397	1,860	2,306	2,896	3,355
9	0,129	0,261	0,543	0,703	0,883	1,383	1,833	2,262	2,821	3,250
10	0,129	0,260	0,542	0,700	0,879	1,372	1,812	2,228	2,764	3,169
11	0,129	0,260	0,540	0,697	0,876	1,363	1,796	2,201	2,718	3,106
12	0,128	0,259	0,539	0,695	0,873	1,356	1,782	2,179	2,681	3,055
13	0,128	0,259	0,538	0,694	0,870	1,350	1,771	2,160	2,650	3,012
14	0,128	0,258	0,537	0,692	0,868	1,345	1,761	2,145	2,624	2,977
15	0,128	0,258	0,536	0,691	0,866	1,341	1,753	2,131	2,602	2,947
16	0,128	0,258	0,535	0,690	0,865	1,337	1,746	2,120	2,583	2,921
17	0,128	0,257	0,534	0,689	0,863	1,333	1,740	2,110	2,567	2,898
18	0,127	0,257	0,534	0,688	0,862	1,330	1,734	2,101	2,552	2,878
19	0,127	0,257	0,533	0,688	0,861	1,328	1,729	2,093	2,539	2,861
20	0,127	0,257	0,533	0,687	0,860	1,325	1,725	2,086	2,528	2,845
21	0,127	0,257	0,532	0,686	0,859	1,323	1,721	2,080	2,518	2,831
22	0,127	0,256	0,532	0,686	0,858	1,321	1,717	2,074	2,508	2,819
23	0,127	0,256	0,532	0,685	0,858	1,319	1,714	2,069	2,500	2,807
24	0,127	0,256	0,531	0,685	0,857	1,318	1,711	2,064	2,492	2,797
25	0,127	0,256	0,531	0,684	0,856	1,316	1,708	2,060	2,485	2,787
26	0,127	0,256	0,531	0,684	0,856	1,315	1,706	2,056	2,479	2,779
27	0,127	0,256	0,531	0,684	0,855	1,314	1,703	2,052	2,473	2,771
28	0,127	0,256	0,530	0,683	0,855	1,313	1,701	2,048	2,467	2,763
29	0,127	0,256	0,530	0,683	0,854	1,311	1,699	2,045	2,462	2,756
30	0,127	0,256	0,530	0,683	0,854	1,310	1,697	2,042	2,457	2,750
40	0,126	0,255	0,529	0,681	0,851	1,303	1,684	2,021	2,423	2,704
60	0,126	0,254	0,527	0,679	0,848	1,296	1,671	2,000	2,390	2,660
100	0,126	0,254	0,526	0,677	0,845	1,290	1,660	1,984	2,364	2,626
120	0,126	0,254	0,526	0,677	0,845	1,289	1,658	1,980	2,358	2,617
50000	0,126	0,253	0,524	0,674	0,842	1,282	1,645	1,960	2,326	2,576

What to do, if the relationship between x and y is non-linear?

- 1. You have to try to transform the values to get a linear relationship;
- 2. If this is not possible, you should work with the non-linear regression;

x	y
2.00	1654.00
4.00	1515.00
6.00	1243.00
8.00	1098.00
10.00	1032.00
15.00	874.00
20.00	754.00
25.00	653.00
30.00	567.00
35.00	604.00
40.00	587.00



Spearman rank correlation coefficient (r_s)

- It is known since the beginning of the 20th century, this is used most often;
- One or both variables are ordinal variables (e.g. the relationship between the taste and the colour of apple);
- Interpretation range: [-1, +1];
  - If $r_s = 1 \Rightarrow$ the two rankings are the same;
  - If $r_s = 0 \Rightarrow$ two rankings are independent from each other, and
  - if $r_s = -1 \Rightarrow$ the two rankings are reversals of each other.
- Indication: r_s ;
- Formula:

$$r_s = 1 - \left[6 \sum_{i=1}^{n} \frac{d_i^2}{n(n^2 - 1)} \right]$$

- ✓ the chronologically given sample elements x_i and y_i with the same element number are indiceted by rank numbers between 1 and n. Namely: the smallest x_i and y_i are indicated by 1, the second smallest x_i and y_i are indicated by 2, …. , the highest x_i, and y_i are indicated by n, namely the highest rank number;
- ✓ for identical values, the average rank number is written to them. In both data sets at most the one-fifth of the observations can be of the same rank number.
- ✓ form the difference of the rank numbers of the (x_i, y_i) pairs that are indicated by d_i.
- ✓ take the squares of the differences (d_i) of the rank numbers (serial numbers) of the (x_i, y_i) pairs:

$$\sum_i (\text{rang}(x_i) - \text{rang}(y_i))^2 = \sum_i d_i^2$$

- ✓ n: number of element pairs;

10 essays are ranked by two reviewers, as follows:

	Reviewers									
	A	1	2	3	4	5	6	7	8	9	10
	B	2	3	1	4	6	5	8	9	7	10
difference	(d)	-1	-1	2	0	-1	1	-1	-1	2	0
	d^2	1	1	4	0	1	1	1	1	4	0

$$r_s = 1 - \frac{6 \times 14}{1000 - 10} = 1 - \frac{84}{990} = 0{,}915$$

Ranking of the two reviewers shows high similarity.

(-1: completely opposite, +1: completely agrees)

Rank correlation

Management data of companies in a region

Region	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
Turnover (M HUF)	34	30	25	22	21	10	12	8	31	20
Profit (M HUF)	16	10,5	10	12	7	4	2	1	9	11
x	10	8	7	6	5	2	3	1	9	4
y	10	7	6	9	4	3	2	1	5	8
d	0	1	1	-3	1	-1	1	0	4	-4
d^2	0	1	1	9	1	1	1	0	16	15

$$r_s = 1 - \frac{6\sum d_i^2}{n\ (n^2 - 1)} = 1 - \frac{6 \cdot 45}{10\left(10^2 - 1\right)} = 0{,}7273$$

Interpretation: there is a stronger than medium positive relationship between the sales and profits of the companies.

Relationship between the diameter of apple blossom and the weight of apple. An example to illustrate the principle of the rank correlation method.

Serial number of blossom-apple pairs	Diameter of the blossom, mm x_i	Rank number $r(x_i)$	Weight of apple, g y_i	Rank number $r(y_i)$	Rank no. difference $d_i=r(x_i)-r(y_i)$
1	32	3,5	210	4	-0.5
2	18	1	150	1	0
3	36	5	235	6	-1
4	32	3,5	205	3	0.5
5	39	7	220	5	2
6	37	6	256	7	-1
7	30	2	190	2	0
8	42	8	300	8	0
Spearman correlation coefficient				r=0.9222	

Is significant this correlation coefficient?

Does the correlation coefficient differ significatly from 0?

✓ probe statistics of the correlation coeffiicent: is of Student distribution, with degrees of freedom: *n-2* ;

✓ *t*-test can be carried out for accepting or rejecting H_0;

H_0 hypothesis: $r = r_0$

H_0: in case of $r = 0$

$$t\ (r) = r\sqrt{\frac{n-2}{1-r^2}} \quad \Rightarrow \text{ the critical r-value: } \quad r\,crit = \frac{t}{\sqrt{n-2+t^2}}$$

In the formula of r_{crit} t is determined from the table of the Student t-distribution, namely: $t \rightarrow |t|^{n-2}_{0,95}$.

The transformation makes t-distribution from the distribution of r

$\Rightarrow$ if $r > r_{crit}$ $\Rightarrow$ r (linear correlation coefficient) significantly differs from 0;

Table of the Student t-distribution

Sz.f.	0,55	0,6	0,7	0,75	0,8	0,9	0,95	0,975	0,99	0,995
1	0,158	0,325	0,727	1,000	1,376	3,078	6,314	12,706	31,821	63,656
2	0,142	0,289	0,617	0,816	1,061	1,886	2,920	4,303	6,965	9,925
3	0,137	0,277	0,584	0,765	0,978	1,638	2,353	3,182	4,541	5,841
4	0,134	0,271	0,569	0,741	0,941	1,533	2,132	2,776	3,747	4,604
5	0,132	0,267	0,559	0,727	0,920	1,476	2,015	2,571	3,365	4,032
6	0,131	0,265	0,553	0,718	0,906	1,440	1,943	2,447	3,143	3,707
7	0,130	0,263	0,549	0,711	0,896	1,415	1,895	2,365	2,998	3,499
8	0,130	0,262	0,546	0,706	0,889	1,397	1,860	2,306	2,896	3,355
9	0,129	0,261	0,543	0,703	0,883	1,383	1,833	2,262	2,821	3,250
10	0,129	0,260	0,542	0,700	0,879	1,372	1,812	2,228	2,764	3,169
11	0,129	0,260	0,540	0,697	0,876	1,363	1,796	2,201	2,718	3,106
12	0,128	0,259	0,539	0,695	0,873	1,356	1,782	2,179	2,681	3,055
13	0,128	0,259	0,538	0,694	0,870	1,350	1,771	2,160	2,650	3,012
14	0,128	0,258	0,537	0,692	0,868	1,345	1,761	2,145	2,624	2,977
15	0,128	0,258	0,536	0,691	0,866	1,341	1,753	2,131	2,602	2,947
16	0,128	0,258	0,535	0,690	0,865	1,337	1,746	2,120	2,583	2,921
17	0,128	0,257	0,534	0,689	0,863	1,333	1,740	2,110	2,567	2,898
18	0,127	0,257	0,534	0,688	0,862	1,330	1,734	2,101	2,552	2,878
19	0,127	0,257	0,533	0,688	0,861	1,328	1,729	2,093	2,539	2,861
20	0,127	0,257	0,533	0,687	0,860	1,325	1,725	2,086	2,528	2,845
21	0,127	0,257	0,532	0,686	0,859	1,323	1,721	2,080	2,518	2,831
22	0,127	0,256	0,532	0,686	0,858	1,321	1,717	2,074	2,508	2,819
23	0,127	0,256	0,532	0,685	0,858	1,319	1,714	2,069	2,500	2,807
24	0,127	0,256	0,531	0,685	0,857	1,318	1,711	2,064	2,492	2,797
25	0,127	0,256	0,531	0,684	0,856	1,316	1,708	2,060	2,485	2,787
26	0,127	0,256	0,531	0,684	0,856	1,315	1,706	2,056	2,479	2,779
27	0,127	0,256	0,531	0,684	0,855	1,314	1,703	2,052	2,473	2,771
28	0,127	0,256	0,530	0,683	0,855	1,313	1,701	2,048	2,467	2,763
29	0,127	0,256	0,530	0,683	0,854	1,311	1,699	2,045	2,462	2,756
30	0,127	0,256	0,530	0,683	0,854	1,310	1,697	2,042	2,457	2,750
40	0,126	0,255	0,529	0,681	0,851	1,303	1,684	2,021	2,423	2,704
60	0,126	0,254	0,527	0,679	0,848	1,296	1,671	2,000	2,390	2,660
100	0,126	0,254	0,526	0,677	0,845	1,290	1,660	1,984	2,364	2,626
120	0,126	0,254	0,526	0,677	0,845	1,289	1,658	1,980	2,358	2,617
50000	0,126	0,253	0,524	0,674	0,842	1,282	1,645	1,960	2,326	2,576

Two-sided table!

Namely, if e.g. the t-threshold belonging to the 95% probability level is looked for, then under a given degree of freedom the table value of 0.975 belongs to this.

$$r = \frac{t}{\sqrt{n-2+t^2}}$$

We finished for today, goodbye!

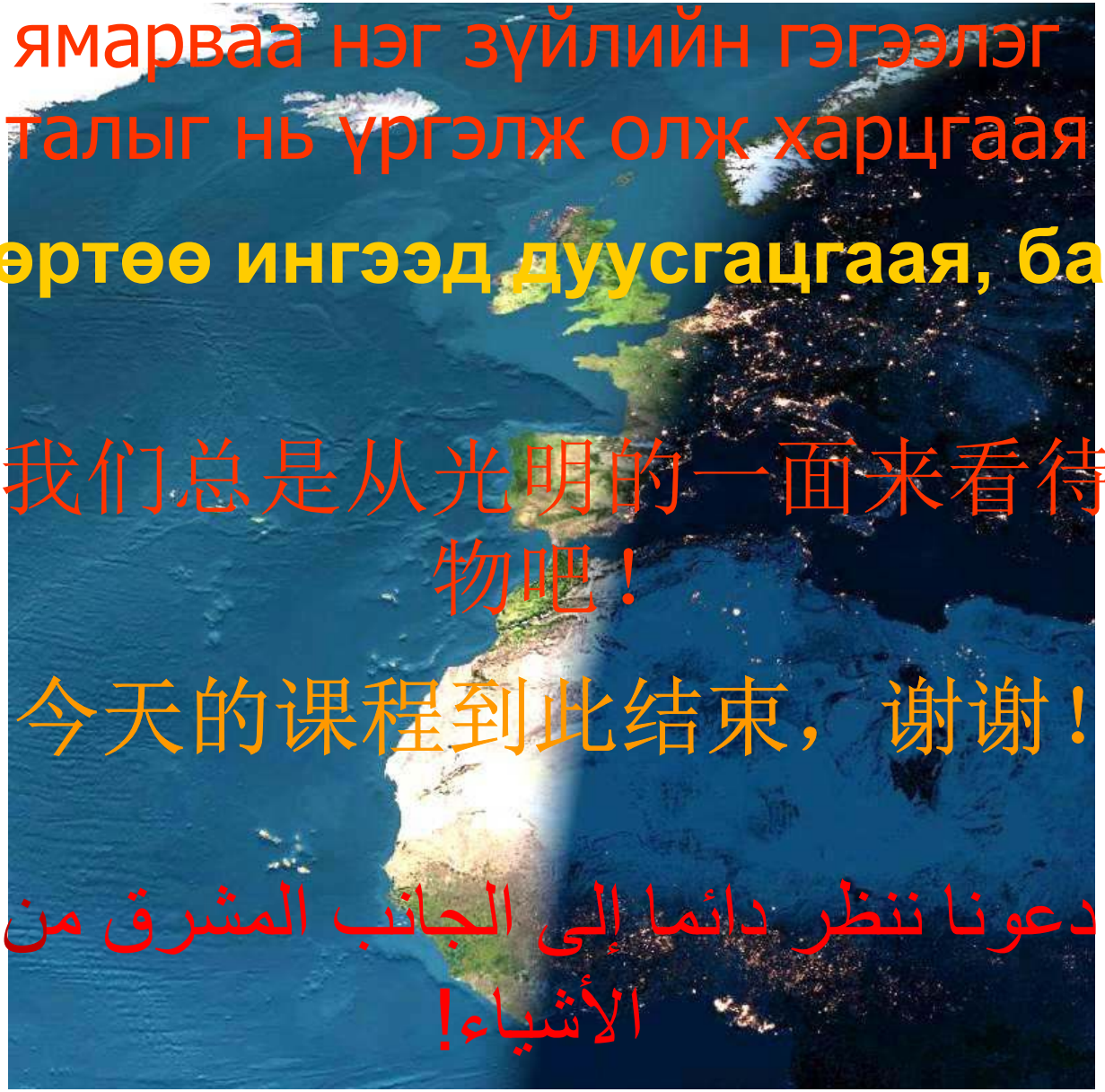
ямарваа нэг зүйлийн гэгээлэг
талыг нь үргэлж олж харцгаая

өнөөдөртөө ингээд дуусгацгаая, баяртай

让我们总是从光明的一面来看待事物吧！

今天的课程到此结束，谢谢！

دعونا ننظر دائما إلى الجانب المشرق من الأشياء!



انتهينا لهذا اليوم، وداعا!