Inovace bakalářského studijního oboru Aplikovaná chemie

http://aplchem.upol.cz

CZ.1.07/2.2.00/15.0247

Tento projekt je spolufinancován Evropským sociálním fondem a státním rozpočtem České republiky.

INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ

Korelace

Korelace

- závislost dvou náhodných veličin
- těsnost **lineární** závislosti
- kovarianční koeficient
- korelační koeficient
  - Pearsonův
  - Spearmannův

Výběrová kovariance

$$C_{XY} = \frac{1}{n-1}\sum^{n}(X_i - \overline{X})(Y_i - \overline{Y})$$

$$C_{X,Y} = C_{Y,X} \text{ a } C_(X,X) = s^2(X)$$

více – do kovarianční matice

$$C = \begin{pmatrix} s_1^2 & C_{12} & \cdots & C_{1n} \\ C_{21} & s_n^2 & & C_{2n} \\ \vdots & & \ddots & \vdots \\ C_{n1} & C_{n2} & \cdots & s_n^2 \end{pmatrix}$$

dále se používá u vícerozměrných metod

např. metody hlavních komponent (PCA)

Výběrová korelace - Pearson

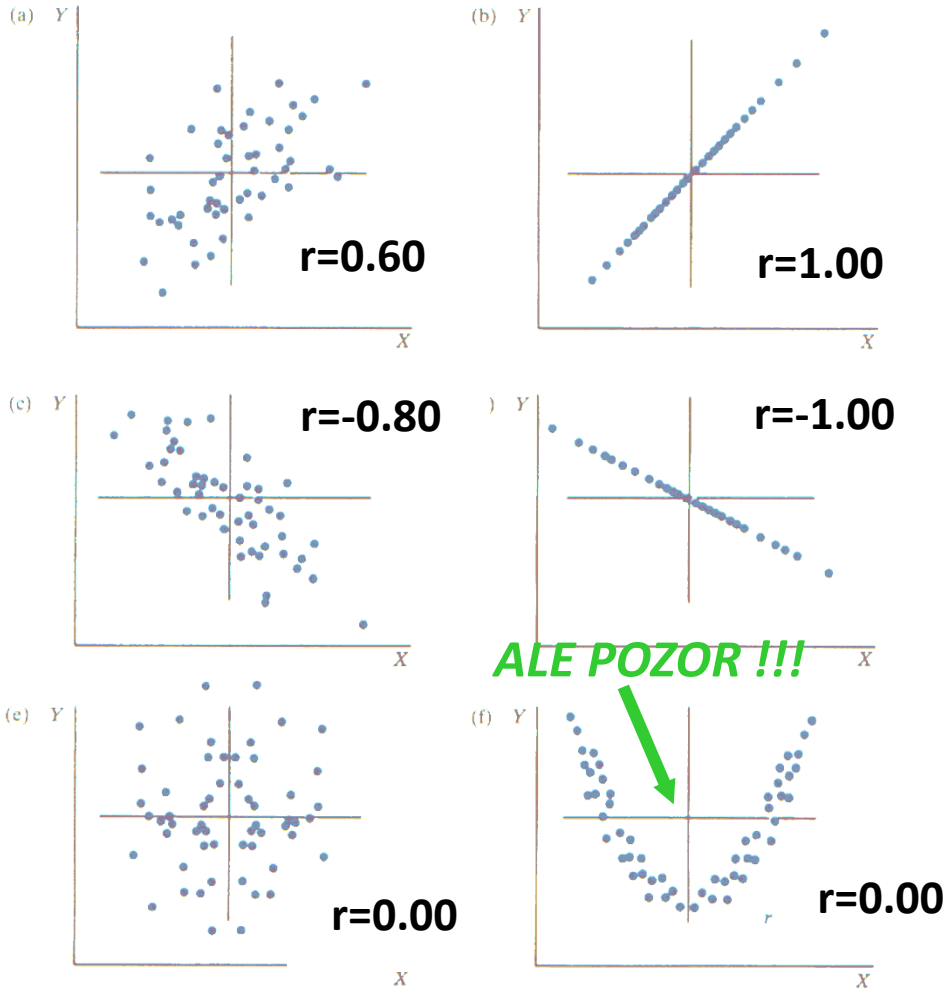
$$\langle -1, 1 \rangle \quad r_{XY} = \frac{\sum^n (X_i - \overline{X})(Y_i - \overline{Y})}{\sqrt{\sum^n (X_i - \overline{X})^2 \sum^n (Y_i - \overline{Y})^2}}$$

$$= \frac{C_{XY}}{\sqrt{s_X^2 s_Y^2}}$$

$$= \frac{\sum^n X_i Y_i - n\overline{XY}}{\sqrt{\sum^n (X_i - \overline{X})^2 \sum^n (Y_i - \overline{Y})^2}},$$

r^2 koeficient determinace

Korelační koeficient



Testování kor. koef.

$$t_e = \frac{r_{XY}}{\sqrt{1 - r_{XY}^2}} \sqrt{n-2}. \tag{6.3}$$

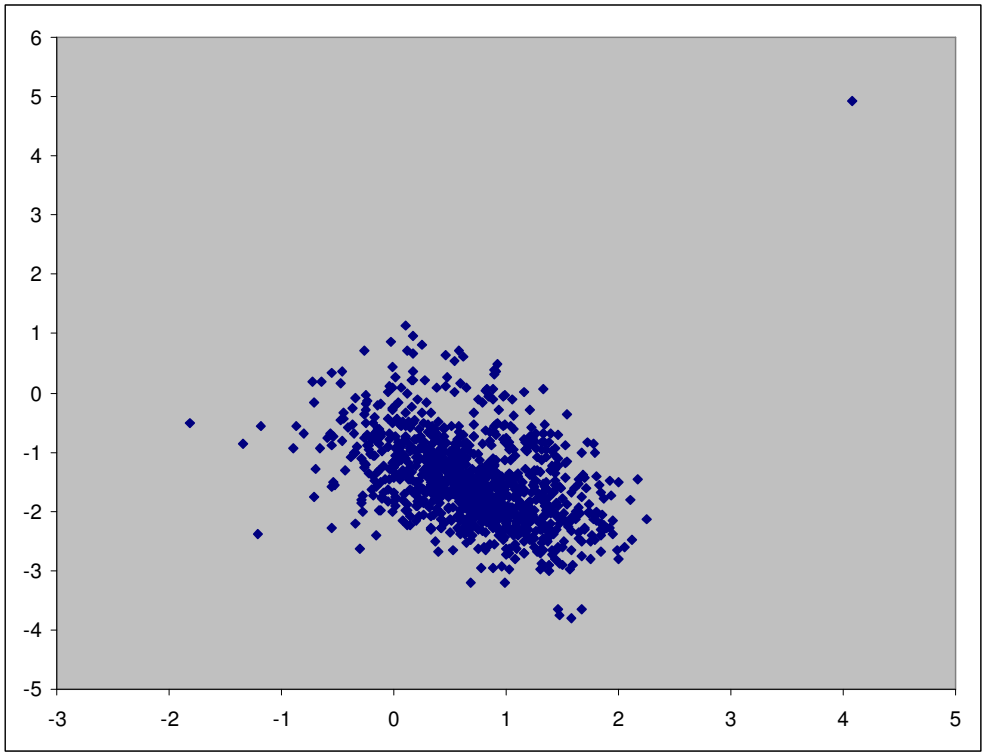
Nulová hypotéza $H_0 : \rho_{XY} = 0$ se na hladině α zamítá ve prospěch oboustranné hypotézy $H_1 : \rho_{XY} \neq 0$, právě když je $|t_e| \geq t_{n-2}(\alpha)$, pokud se nepoužije absolutní hodnota lze zamítnout ve prospěch levo- či pravostranné hypotézy.

Řešení v MS Excelu

CORREL(výběr1;výběr2)

COVAR(výběr1;výběr2)

graf – XY bodový



CORREL(výběr1;výběr2)

-0.45807

COVAR(výběr1;výběr2)

-0.19072

Řešení v MS Excelu

Analýza dat: korelace, kovariance
korelační matice, kovarianční matice

proč tu nic není?

	Sloupec 1	*Sloupec 2*	*Sloupec 3*	*Sloupec 4*
Sloupec 1	1			
Sloupec 2	-0.50655	1		
Sloupec 3	-0.06564	0.055605	1	
Sloupec 4	-0.01541	-0.01958	-0.02002	1

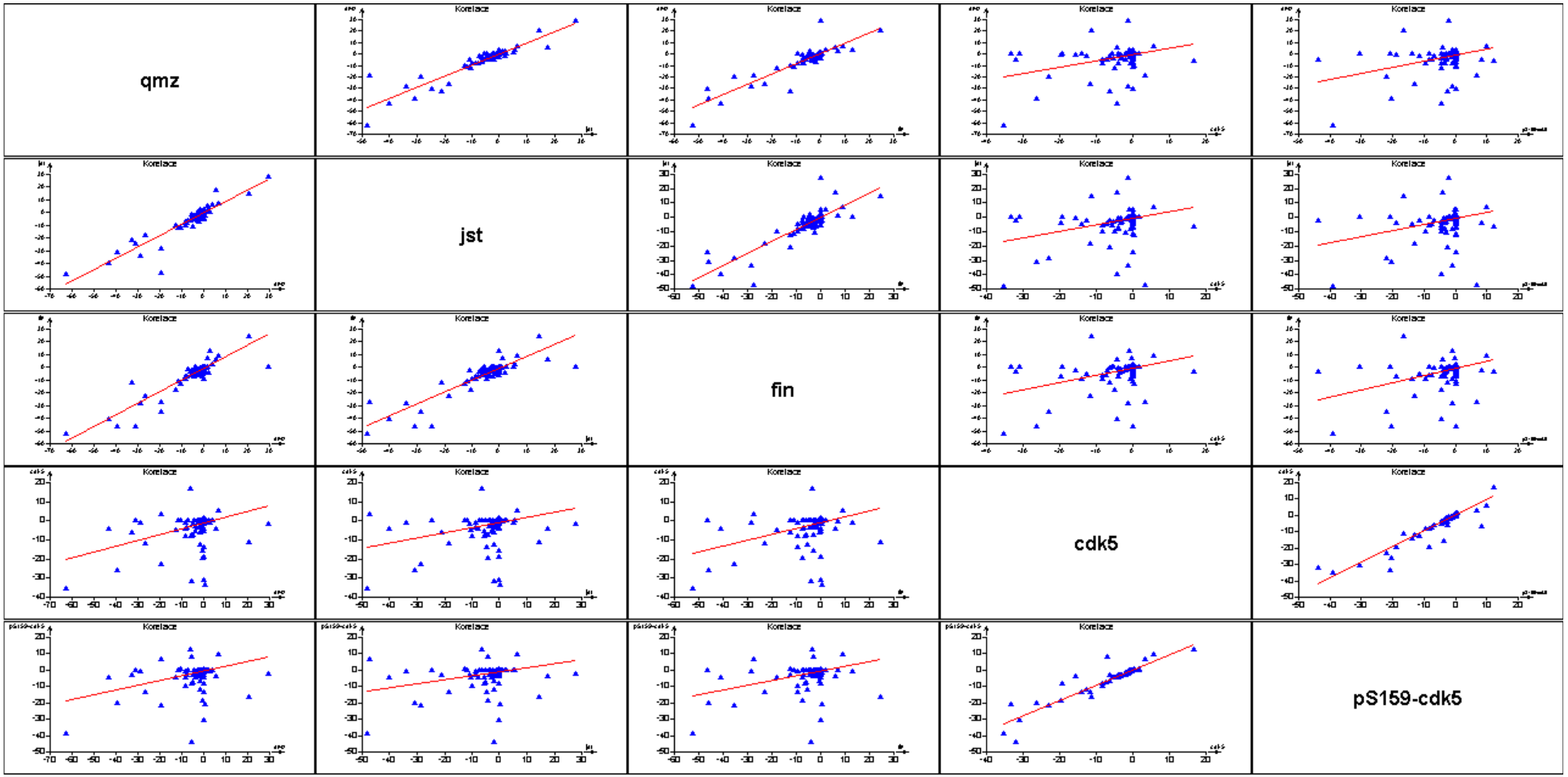
$$r_{XY} = r_{YX}$$

	Sloupec 1	*Sloupec 2*	*Sloupec 3*	*Sloupec 4*
Sloupec 1	1.000	-0.507	-0.066	-0.015
Sloupec 2	-0.507	1.000	0.056	-0.020
Sloupec 3	-0.066	0.056	1.000	-0.020
Sloupec 4	-0.015	-0.020	-0.020	1.000

QCExpert

vícerozměrné metody - korelace

vizuální analýza: korelační grafy



evropský sociální fond v ČR
EVROPSKÁ UNIE
MINISTERSTVO ŠKOLSTVÍ, MLÁDEŽE A TĚLOVÝCHOVY
OP Vzdělávání pro konkurenceschopnost
OKRESNÍ HOSPODÁŘSKÁ KOMORA OLOMOUC
INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ

Inovace bakalářského studijního oboru Aplikovaná chemie

Autokorelace

6.1.1 Autokorelace

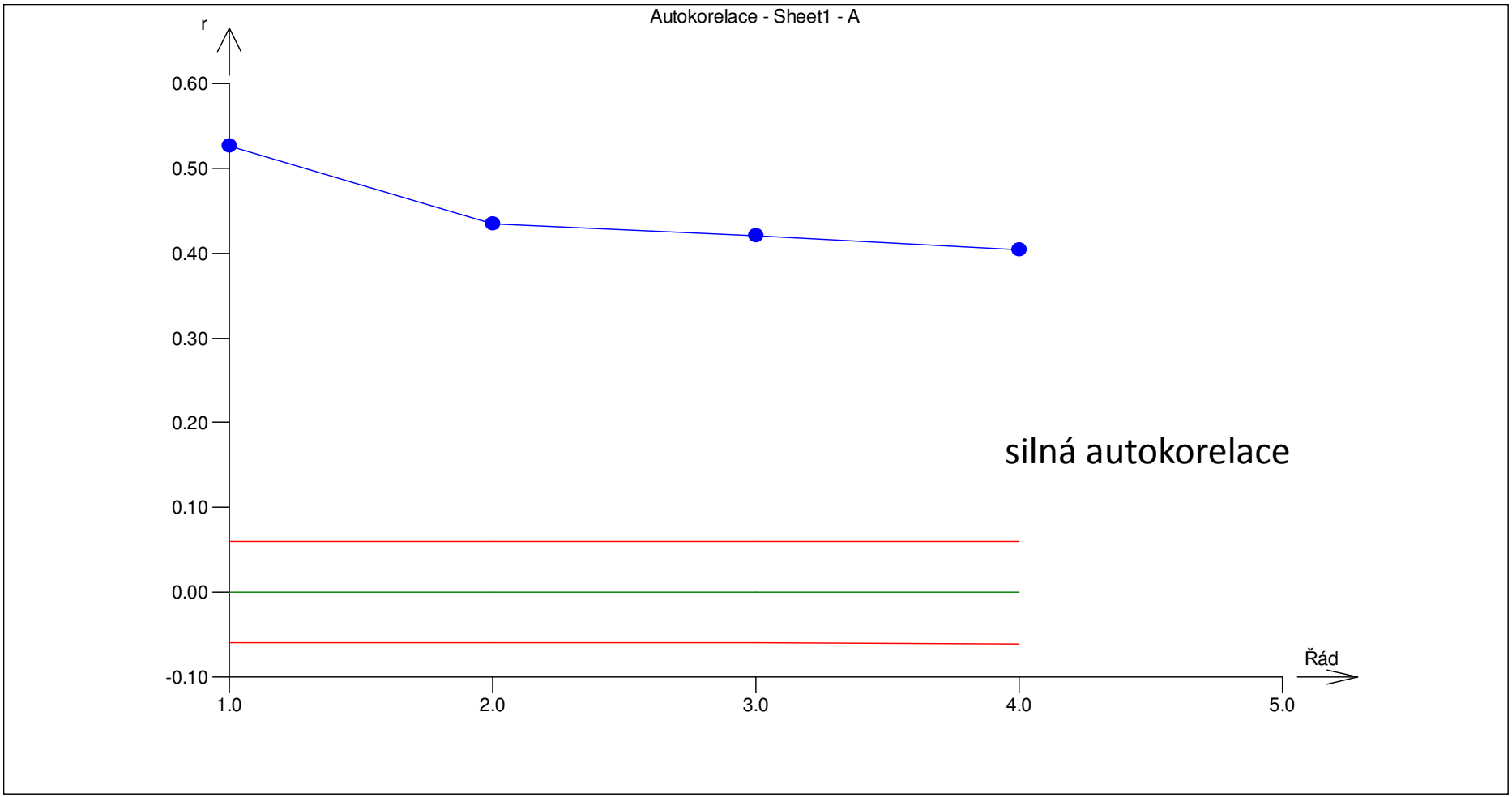
Autokorelační koeficient 1. řádu r_1 lze chápat jako „korelaci“ 1. hodnoty se 2., 2. se 3., 3. se 4. atd. Obecněji je autokorelační koeficient k-tého řádu r_k je definován jako

$$r_k = \frac{\sum_{i=1}^{N-k}(y_i - \overline{y})(y_{i+k} - \overline{y})}{\sum_{i=1}^{N}(y_i - \overline{y})^2}. \qquad (6.4)$$

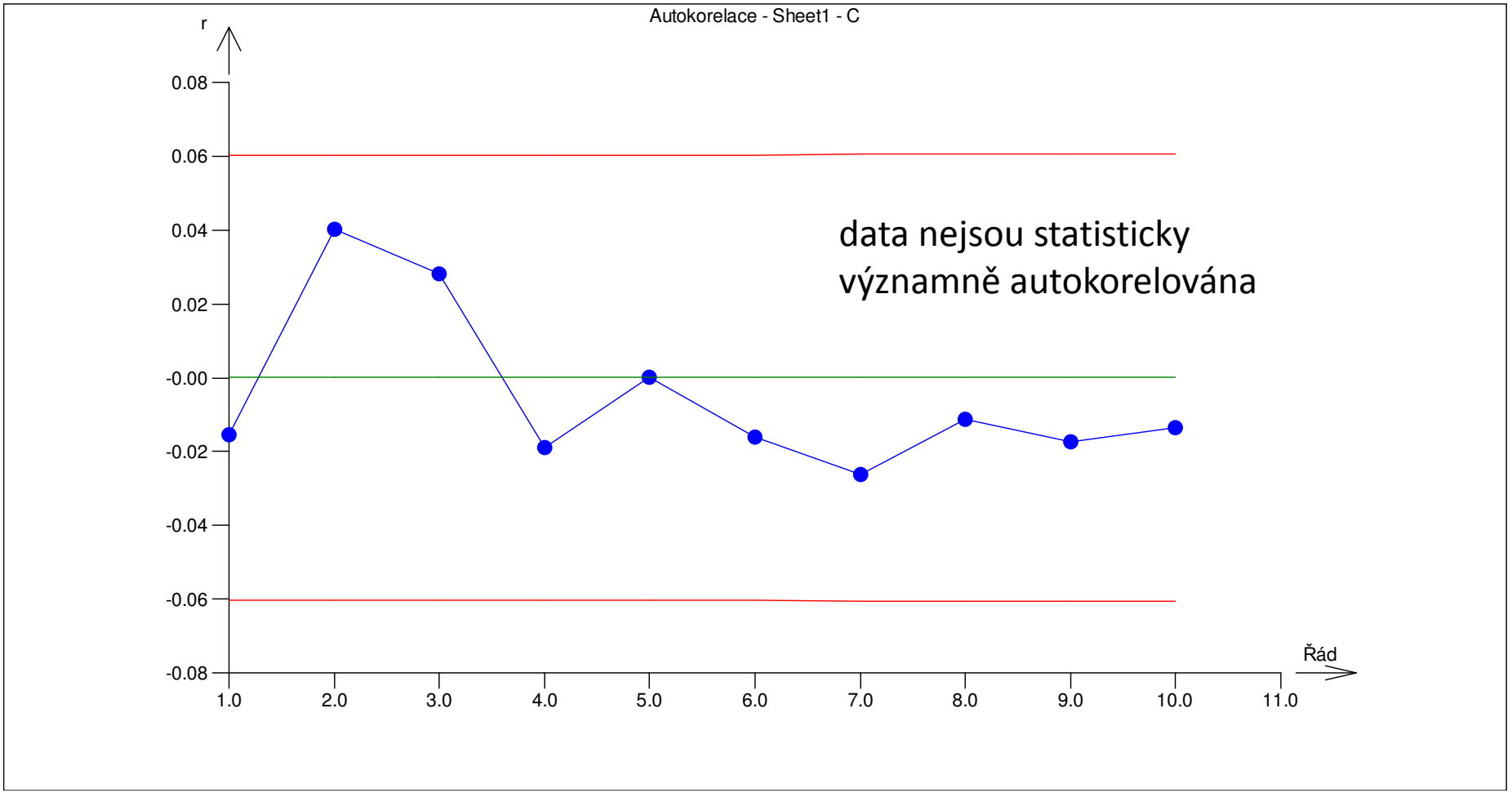
závislost *n*-té a *n*+*k*-té hodnoty, kde *k* je řád autokorelace

QCExpert – základní statistika
EXCEL – posunuté sloupce, a fce. CORREL

Autokorelační graf



Autokorelační graf



Spearmannův korelační koeficient

pořadová korelace

$$r_S = 1 - \frac{6\sum^n q_i^2}{n^3 - n},$$

rozdíl v pořadí

Příklad 6.1 Máme n různých chemických látek a máme rozhodnout, zdali existuje korelace mezi interakční energií enzym-inhibitor, vypočtenou na základě molekulové mechaniky, a inhibičním účinkem (IC_{50}) těchto látek na příslušný enzym.

látka n	energie kcal·mol^{-1}	IC_{50} μM	pořadí podle E k	pořadí podle IC_{50} l	rozdíl pořadí q^2
A	−25	12	1	1	0
B	−26	11	2	2	0
C	−28	10	3	3	0
D	−30	9	4	4	0
E	−31	7	5	6	1
F	−32	8	6	5	1
G	−33	1	7	8	1
H	−35	5	8	7	1

Odtud plyne, že $r_S = 1 - \frac{24}{504} = 0,9524$. Pro zajímavost uvádíme i $r_{XY} = 0,8449$. Na základě srovnání s kritickou hodnotou Spearmanova korelačního koeficientu, na hladině významnosti $\alpha = 0,05$, $r_S = 0,6905$ plyne, že existuje korelace mezi vypočtenou interakční energií a inhibičním účinkem. □

Závislost nominálních veličin

Kontingenční tabulky

$$N_{i.} = \sum_{j}^{b} N_{ij}, \qquad N_{.j} = \sum_{i}^{a} N_{ij},$$

marginální četnosti

$$X^2 = \sum_i \sum_j \frac{(N_{ij} - N_{i.}N_{.j}/n)^2}{N_{i.}N_{.j}/n}. \tag{6.7}$$

Nulovou hypotézu pak budeme zamítat na hladině α, pokud bude platit

$$X^2 \geq \chi^2_{(a-1)(b-1)}(\alpha). \tag{6.8}$$

Kontingenční tabulka

Příklad 6.32. Rozhodněte zda je ve studované populaci barva lidských vlasů závislá na pohlaví.

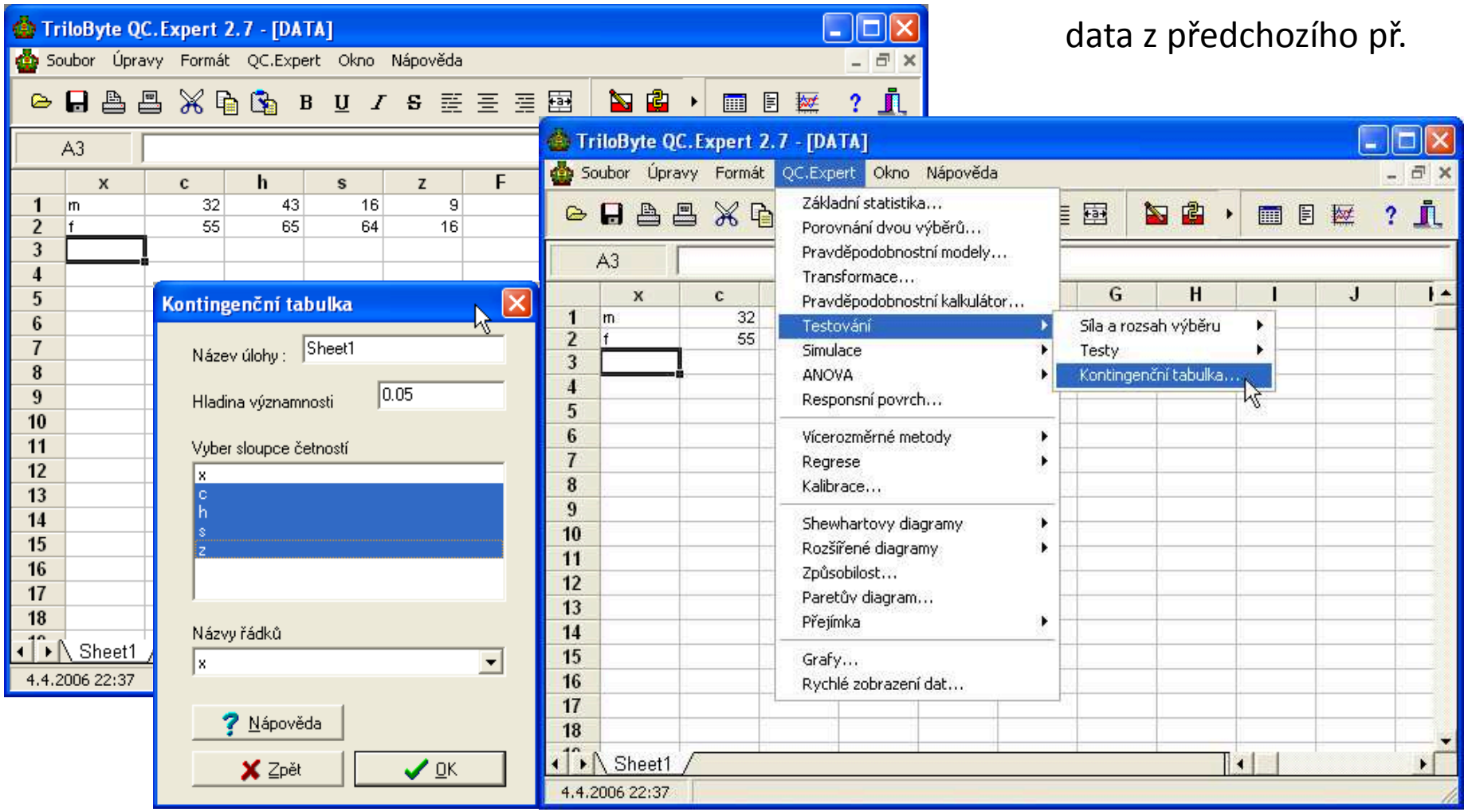
pohlaví	barva vlasů				
	černá	hnědá	světlá	zrzavá	celkem
mužské	32	43	16	9	100
ženské	55	65	64	16	200
celkem	87	108	80	25	300

$$X^2 = \sum_i \sum_j \frac{(N_{ij} - N_{i.}N_{.j}/n)^2}{N_{i.}N_{.j}/n}.$$

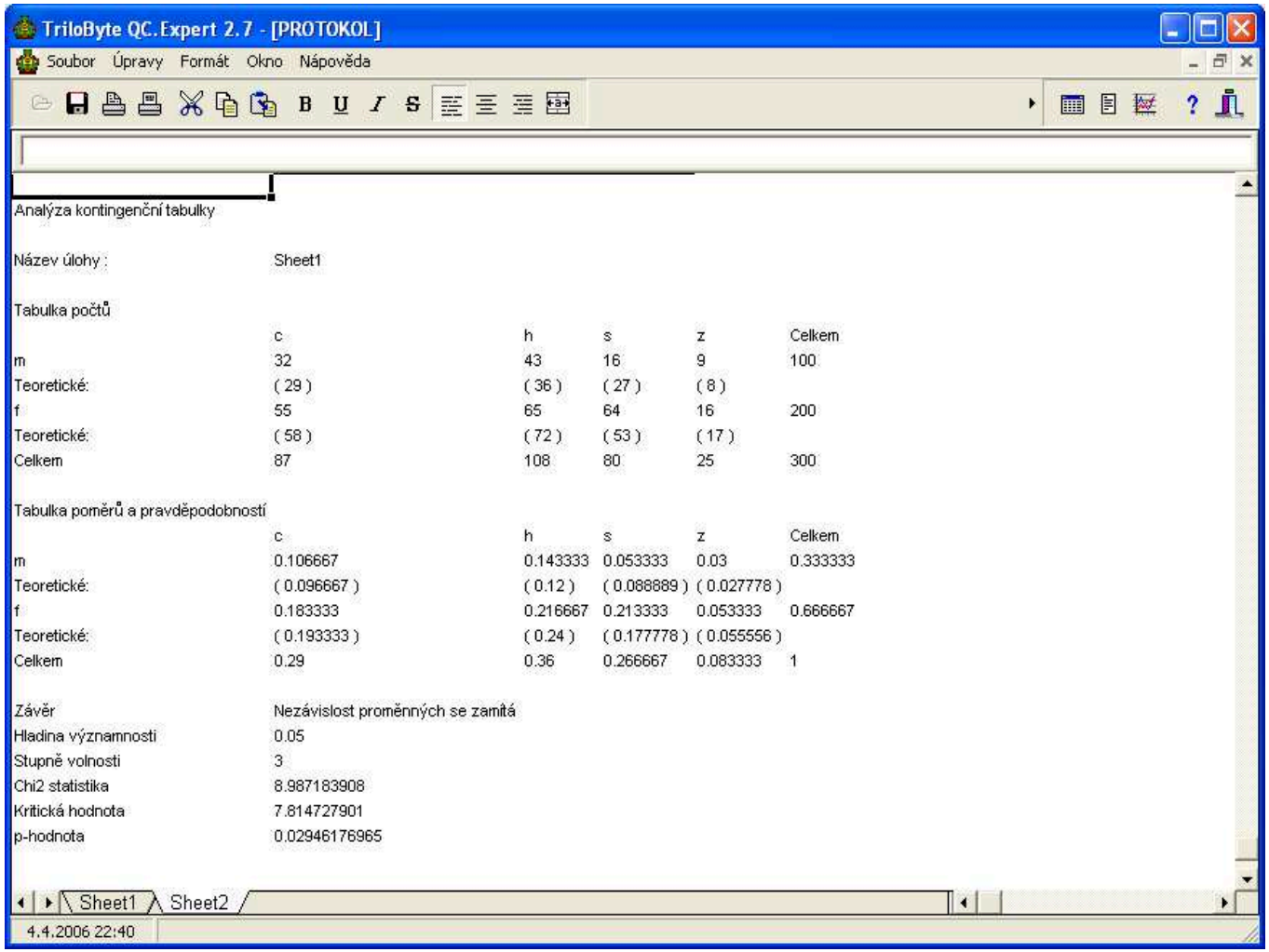
Nejprve vyčíslíme marginální četnosti $N_{M.} = 100$, $N_{Z.} = 200$, $N_{.c} = 87$, $N_{.h} = 108$, $N_{.s} = 80$, $N_{.z} = 25$. Následně vypočteme testovací statistiku $X^2 = \frac{(32-100\cdot 87/300)^2}{100\cdot 87/300} + \ldots = 8,987$. Po porovnání s kritickou hodnotou $\chi^2_{(3)}(0,05) = 7,815$ na hladině $\alpha = 0,05$ zamítáme nulovou hypotézu o nezávislosti barvy vlasů na pohlaví v dané populaci. □

QCExpert

data z předchozího př.



QCExpert


TriloByte QC.Expert 2.7 - [PROTOKOL]
Soubor Úpravy Formát Okno Nápověda
Analýza kontingenční tabulky
Název úlohy : Sheet1
Tabulka počtů
c h s z Celkem
m 32 43 16 9 100
Teoretické: (29) (36) (27) (8)
f 55 65 64 16 200
Teoretické: (58) (72) (53) (17)
Celkem 87 108 80 25 300
Tabulka poměrů a pravděpodobností
c h s z Celkem
m 0.106667 0.143333 0.053333 0.03 0.333333
Teoretické: (0.096667) (0.12) (0.088889) (0.027778)
f 0.183333 0.216667 0.213333 0.053333 0.666667
Teoretické: (0.193333) (0.24) (0.177778) (0.055556)
Celkem 0.29 0.36 0.266667 0.083333 1
Závěr Nezávislost proměnných se zamítá
Hladina významnosti 0.05
Stupně volnosti 3
Chi2 statistika 8.987183908
Kritická hodnota 7.814727901
p-hodnota 0.02946176965
Sheet1 Sheet2
4.4.2006 22:40