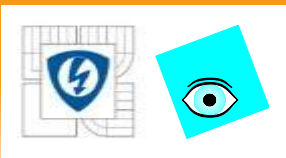


Statistická analýza jednorozměrných dat

Prof. RNDr. Milan Meloun, DrSc.
Univerzita Pardubice, Pardubice

31.ledna 2011

Tato prezentace je spolufinancována Evropským sociálním fondem a státním rozpočtem České republiky.



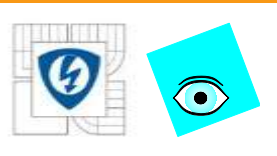
Kapitola 7.1

STATISTICKÁ ZÁVISLOST

31.1.2011

INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ





Statistická závislost

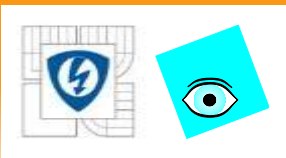
- **Korelace** popisuje vliv změny úrovně jednoho znaku na změnu úrovně jiných znaků a platí pro **kvantitativní (měřené) znaky**;
- **Kontingence** popisuje závislost **kvalitativních** (slovních, popisných) znaků, které mají více než dvě alternativy, tzv. **množných znaků** (např. druh dřeviny, národnost, apod.);
- **Asociace** popisuje závislost **kvalitativních** (slovních, popisných) znaků, které mají pouze dvě alternativy, tzv. **alternativních znaků** (např. pohlaví, odpovědi typu ano/ne, ...).



Korelace

- typy podle počtu korelovaných znaků

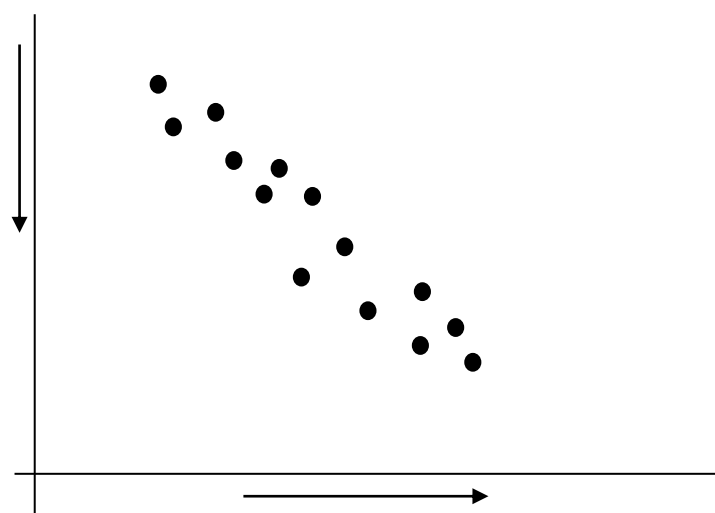
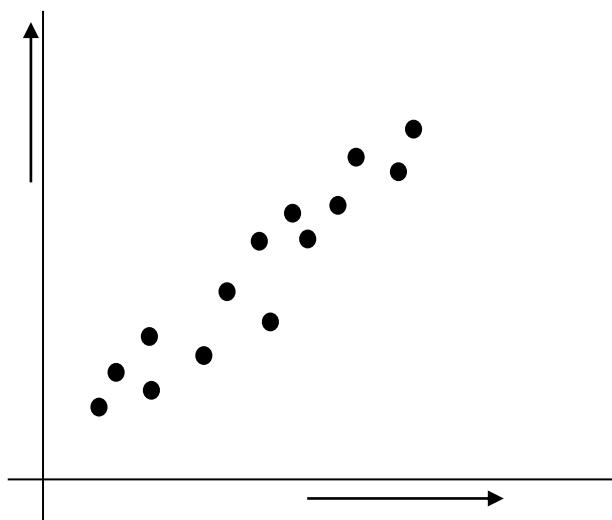
- **Jednoduchá** popisuje vztah dvou znaků,
- **Mnohonásobná** popisuje vztahy více než dvou znaků,
- **Parciální** popisuje závislost dvou znaků ve vícerozměrném statistickém souboru při vyloučení vlivu ostatních znaků na tuto závislost.

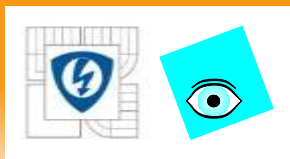


Korelace

- typy podle smyslu změny hodnot

- **Kladná** značí, že se zvyšováním hodnot jednoho znaku se zvyšují i hodnoty druhého znaku,
- **Záporná** značí, že se zvyšováním hodnot jednoho znaku se zmenšují hodnoty druhého znaku,

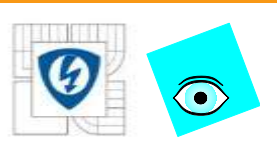




Korelace

- typy podle tvaru závislosti

- **Přímková (lineární)** značí, že grafickým obrazem závislosti je přímka (lineární trend),
- **Křivková (nelineární)** značí, že grafickým obrazem závislosti je křivka (nelineární trend).



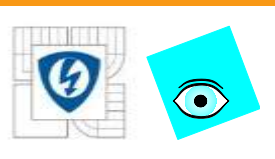
Korelační počet

Korelační analýza

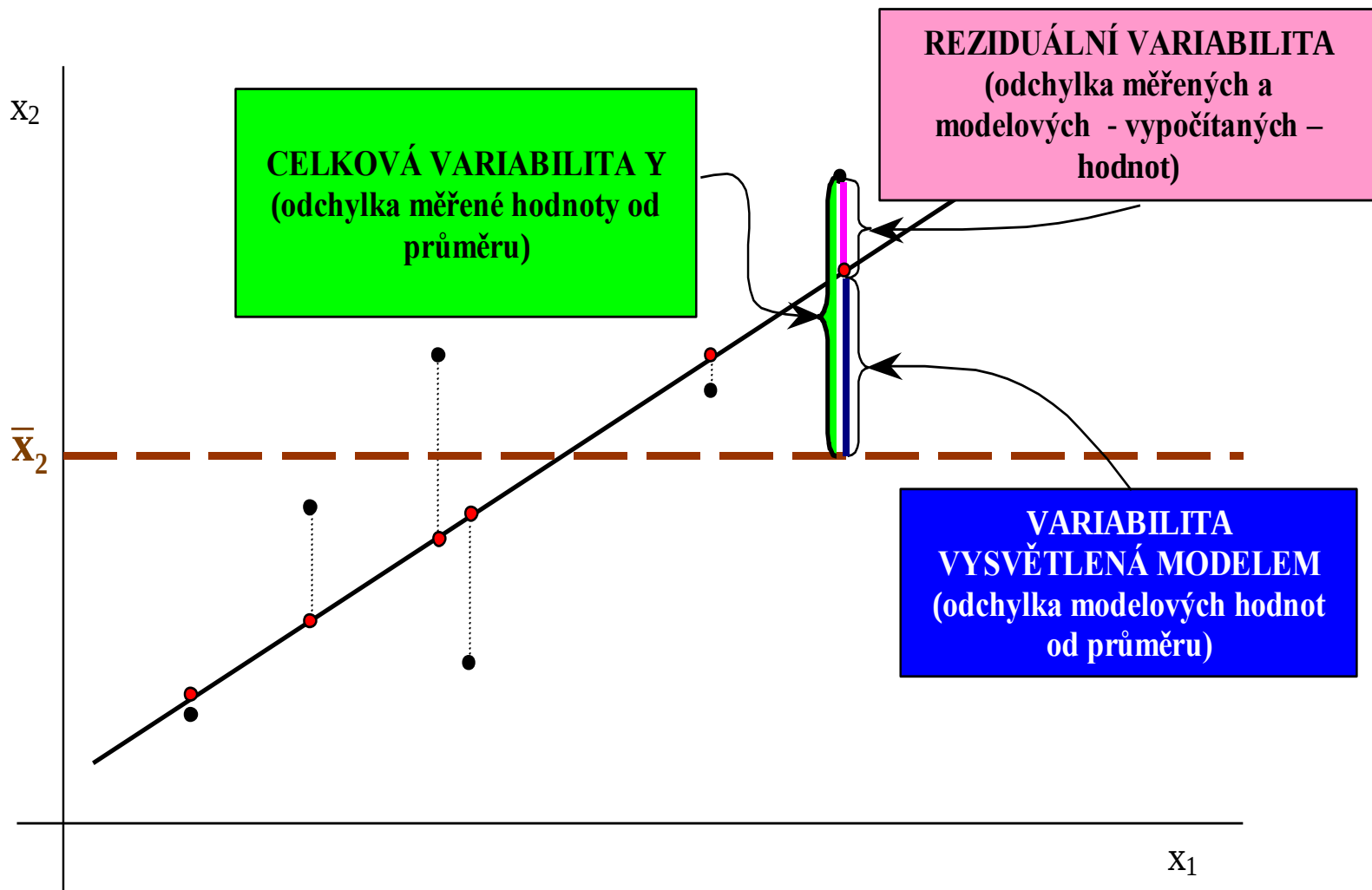
- zjišťuje **existenci závislosti** a její druhy,
- měří **těsnost závislosti**,
- ověřuje **hypotézy o statistické významnosti** závislosti;

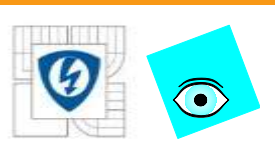
Regresní analýza

- zabývá se **vytvořením vhodného matematického modelu** závislosti,
- stanoví **parametry** tohoto modelu,
- ověřuje **hypotézy o vhodnosti a důležitých vlastnostech modelu**.



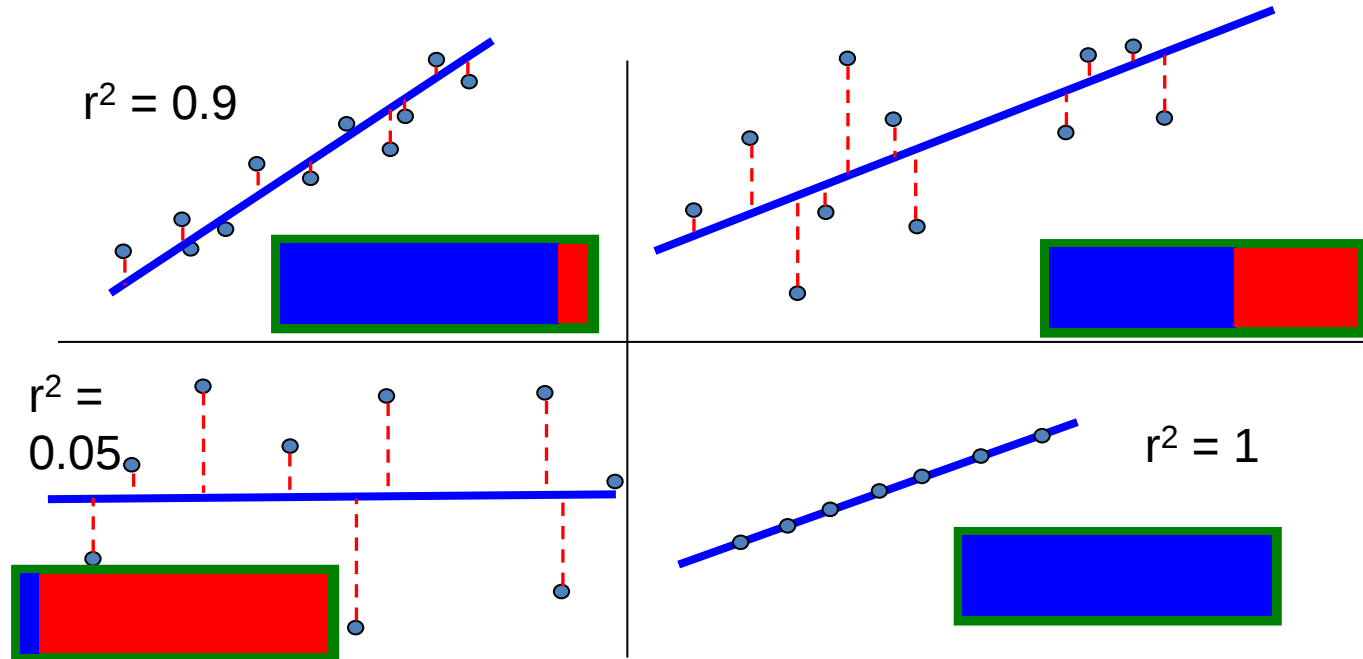
Míra korelační závislosti

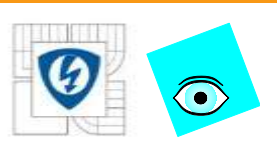




Koeficient determinace

- vyjadřuje, jakou **část celkové variability** závisle proměnné (vysvětlované proměnné) objasňuje regresní model.



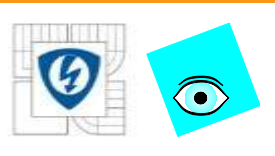


Korelační koeficient

Pro jednoduchou korelaci:

Párový představuje zvláštní případ vícenásobného korelačního koeficientu, kdy vyjadřuje míru lineární stochastické závislosti mezi náhodnými veličinami x_i a x_j ,

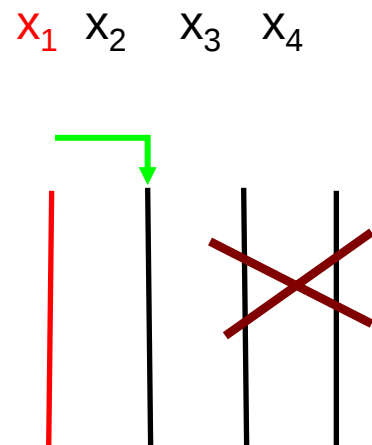
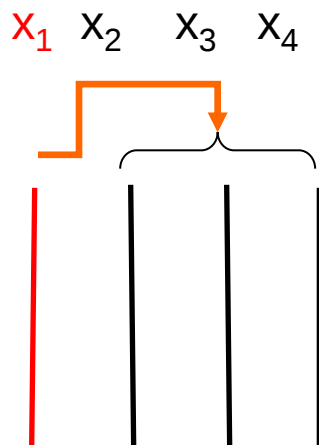
- **Pearsonův**
- **Spearmanův** (korelace pořadí)

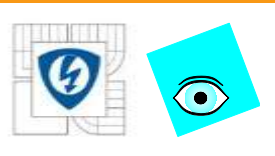


Korelační koeficient

Pro vícenásobnou korelaci:

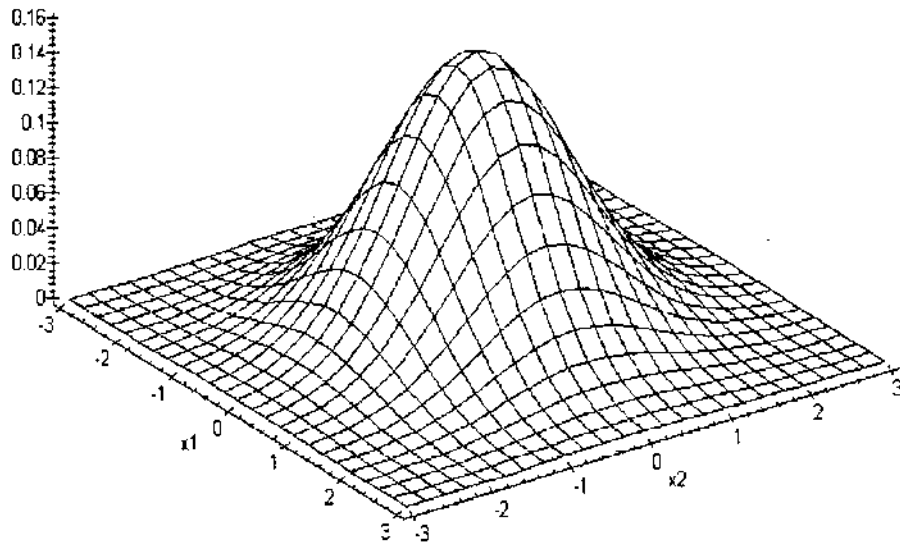
- **Vícenásobný** definuje míru lineární stochastické závislosti mezi náhodnou veličinou x_1 a nejlepší lineární kombinací složek $x_2, x_3, \dots, x_m$ náhodného vektoru x
- **Parciální** definuje míru lineární stochastické závislosti mezi náhodnými veličinami x_i a x_j při skonstruktivní konstantní ostatních složek vektoru x





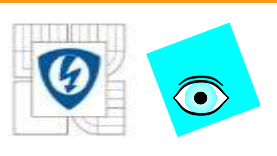
Pearsonův korelační koeficient r

- Podmínkou je dodržení **dvourozměrného normálního rozdělení**



$$r_{x_1 x_2} = r_{x_2 x_1} = \frac{\text{COV}(x_1, x_2)}{S_{x_1} \cdot S_{x_2}}$$

$\text{COV}(x_1, x_2)$ **je normovaná kovariance**

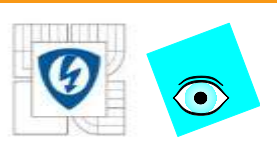


Pearsonův korelační koeficient r

Kovariance:

- **míra intenzity vztahu** mezi složkami vícerozměrného souboru
- je mírou intenzity **lineární** závislosti
- je vždy **nezáporná**
- její **limitou je součin směrodatných odchylek**
- je **symetrickou funkcí** svých argumentů
- její **velikost je závislá na měřítku argumentů** → **nutnost normování**

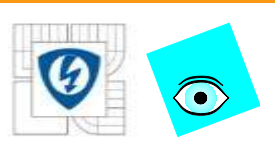
$$\text{cov}(x_1, x_2) = \frac{1}{n} \sum_{i=1}^n (x_{1i} - \bar{x}_1) (x_{2i} - \bar{x}_2)$$



Pearsonův korelační koeficient r

Základní vlastnosti Pearsonova korelačního koeficientu:

- je to **bezrozměrná** míra lineární korelace;
- nabývá hodnoty **0 – 1 pro kladnou korelaci, 0 – (-1) pro zápornou korelaci**;
- hodnota **0** znamená, že mezi posuzovanými veličinami **není žádný lineární vztah** (může být nelineární) nebo tento vztah zůstal na základě dat, které máme k dispozici, neprokázán;
- hodnota **1** nebo **(-1)** indikuje **funkční závislost**;
- hodnota korelačního koeficientu je stejná pro závislost x_1 na x_2 i pro opačnou závislost x_2 na x_1 .



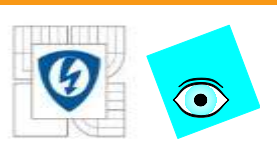
Spearmanův korelační koeficient

Neparametrický korelační koeficient, vycházející nikoli z hodnot, ale z jejich pořadí.

Používá se tehdy, nejsou-li **závažným způsobem** splněny předpoklady pro použití Pearsonova korelačního koeficientu.

Diference mezi pořadími hodnot x a y v jednom řádku

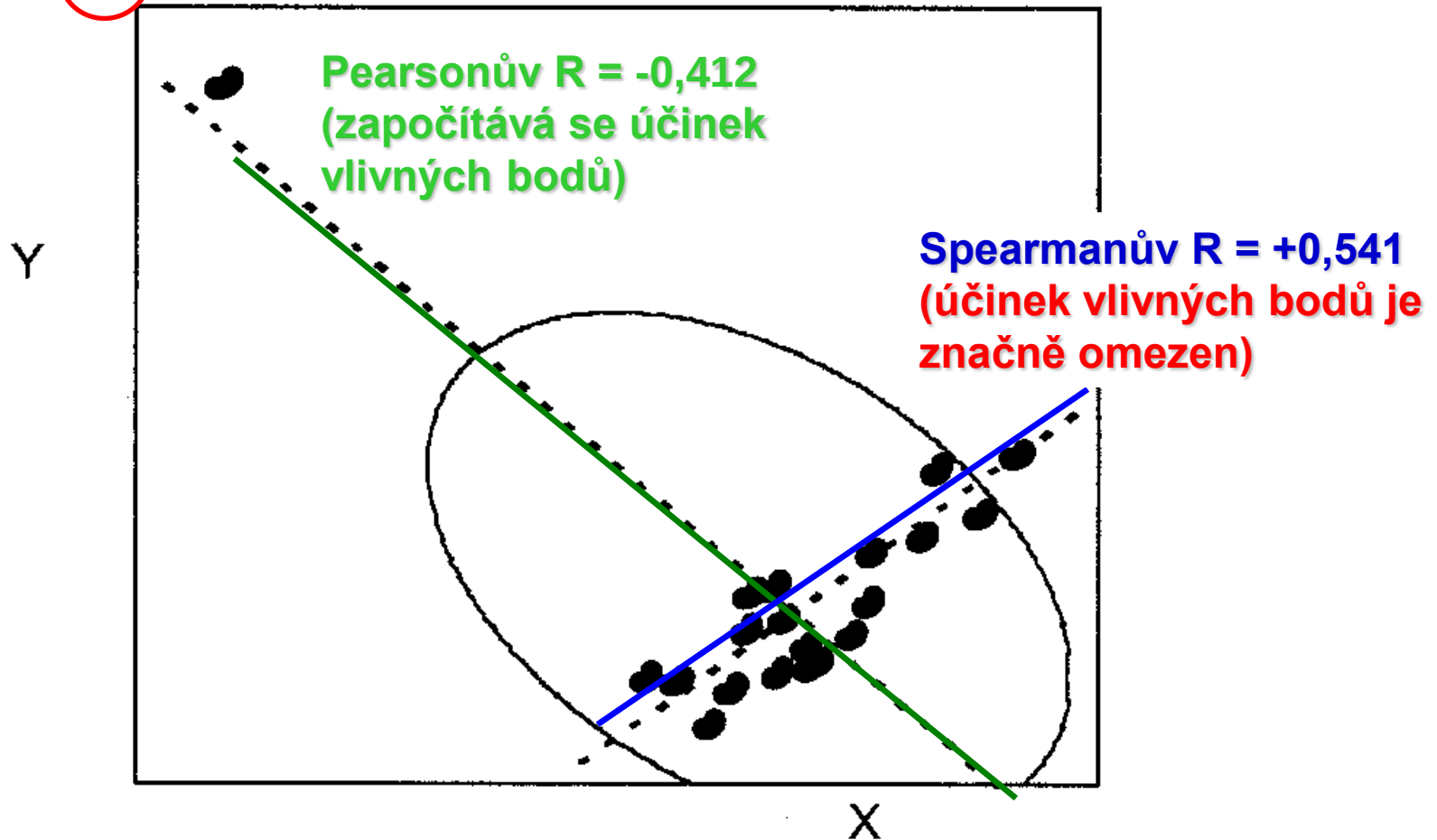
$$r_s = 1 - \frac{6 \cdot \sum_{i=1}^n d_i^2}{n^3 - n}$$

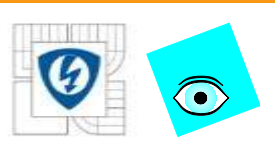


Spearmanův korelační koeficient



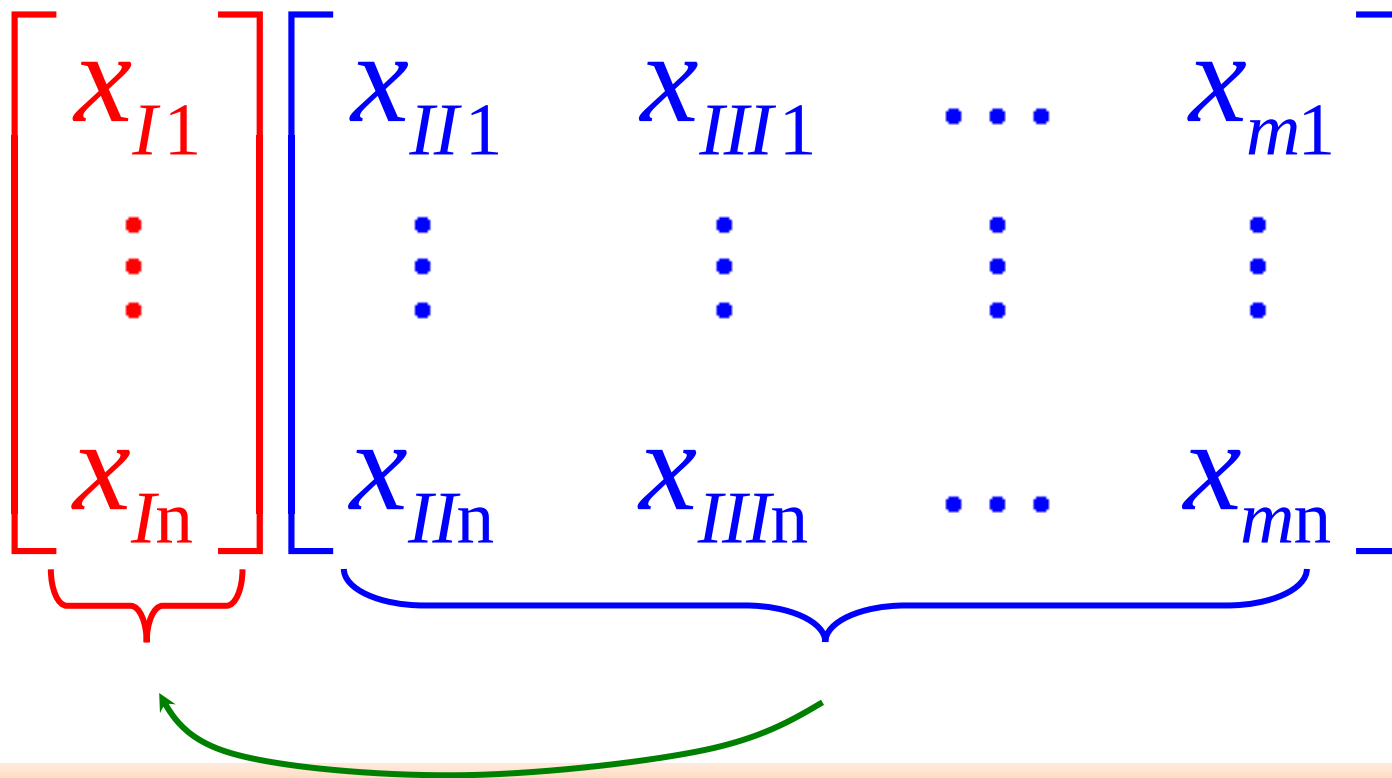
vlivné body

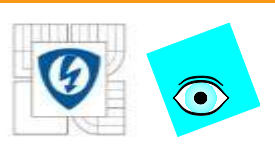




Mnohonásobný korelační koeficient

vyjadřuje sílu závislosti **jedné proměnné** na **dvou a více** jiných proměnných



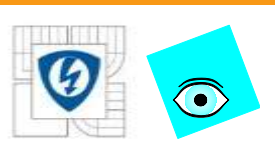


Mnohonásobný korelační koeficient

[?] Základní vlastnosti:

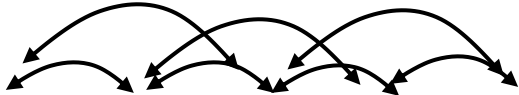
- a) $0 \leq R \leq 1$
- b) Pokud je $R = 1$, znamená to, že závisle proměnná x_1 je přesně lineární kombinací veličin $x_2, \dots, x_m$.
- c) Pokud je $R = 0$, potom jsou také všechny párové korelační koeficienty nulové.
- d) S růstem počtu vysvětlujících (nezávislých) proměnných hodnota vícenásobného korelačního koeficientu neklesá, tj. platí

$$R_{1(2)} \leq R_{1(2,3)} \leq \dots \leq R_{1(2,\dots,m)}.$$



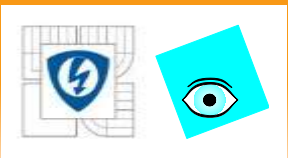
Parciální korelační koeficient

Používá se k posouzení síly závislosti **dvou veličin** ve vícerozměrném souboru **při vyloučení vlivu ostatních veličin**.



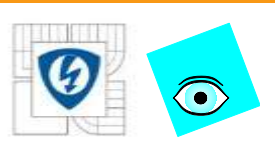
	A	B	C	D	E
1	X1	X2	X3	X4	X5
2	5	5	8	7	7
3	2	2	9	8	8
4	4	5	5	9	9
5	5	1	2	5	5
6	6	5	3	4	4
7	2	4	1	1	2
8	4	1	4	2	1

Podle počtu „vyloučených“ proměnných se stanovují řády parciálního R v příkladu vlevo to je parciální korelace III. řádu (3 „vyloučené“ proměnné)



Testy významnosti v korelační a regresní analýze

- test významnosti korelačního koeficientu
 - test významnosti modelu jako celku
 - test významnosti jednotlivých regresních parametrů
 - test shody lineárních regresních modelů
- a mnoho dalších



Test významnosti R



Test významnosti odpoví, zda je korelace mezi výběrovými proměnnými R natolik silná, abychom ji mohli považovat za dostatečně prokázanou i pro základní soubor ρ .

Pro párový R :

$$t_R = \frac{R \cdot \sqrt{n-2}}{\sqrt{1-R^2}}$$

KH

$t_{\alpha, n-2}$

n – počet
hodnot výběru

Pro násobný R :

$$F_R = \frac{R^2(n-m)}{(1-R^2)(m-1)}$$

$t_{\alpha, n-m}$

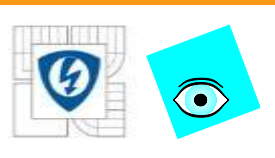
m – počet
proměnných

Pro parciální R :

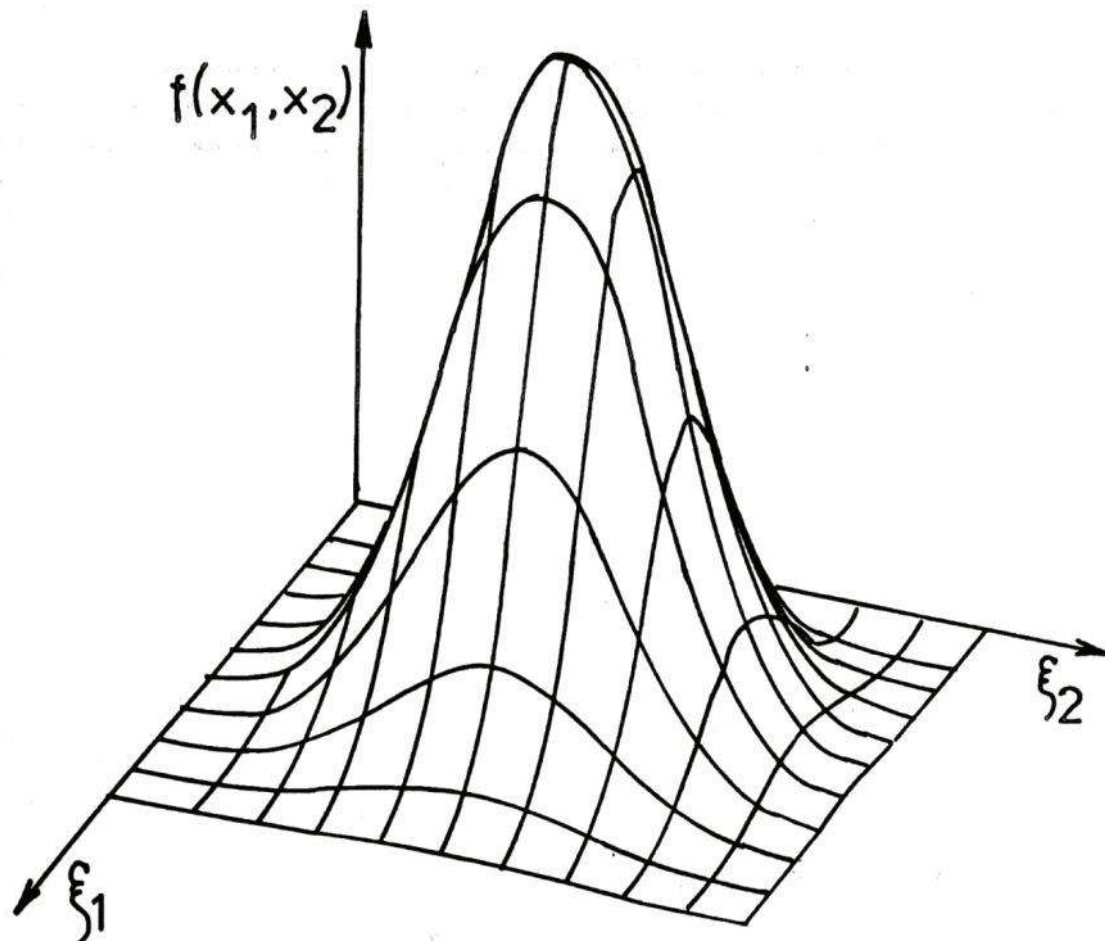
$$t_R = \frac{R \cdot \sqrt{n-k-2}}{\sqrt{1-R^2}}$$

$t_{\alpha, n-k-2}$

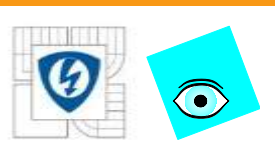
k – počet
„vyloučených“
proměnných



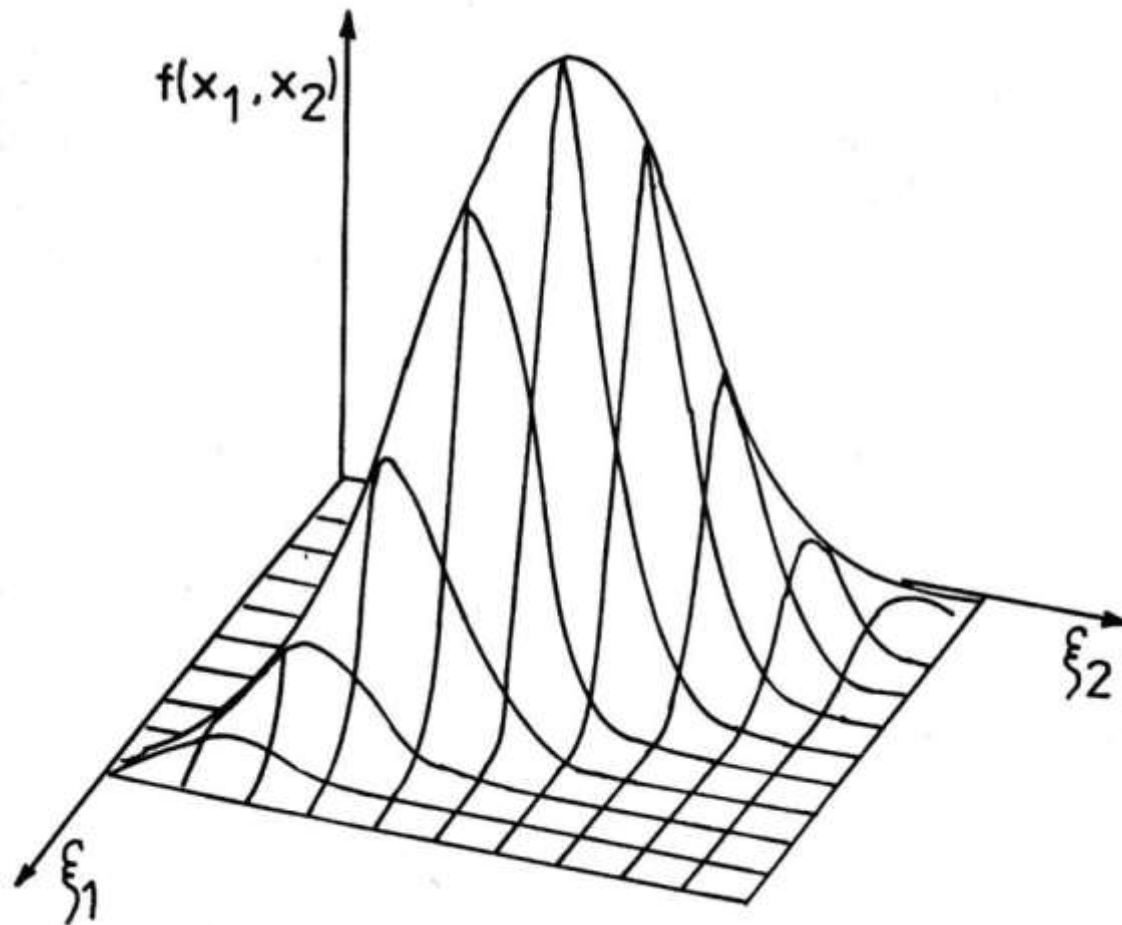
Povrch simultánní hustoty pravděpodobnosti pro $\rho_{12} = 0$



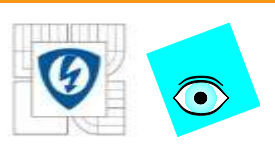
Obr. 4.3 Povrch simultánní hustoty pravděpodobnosti pro $\rho_{12} = 0$



Povrch simultánní hustoty pravděpodobnosti pro $\rho_{12} = 0.9$

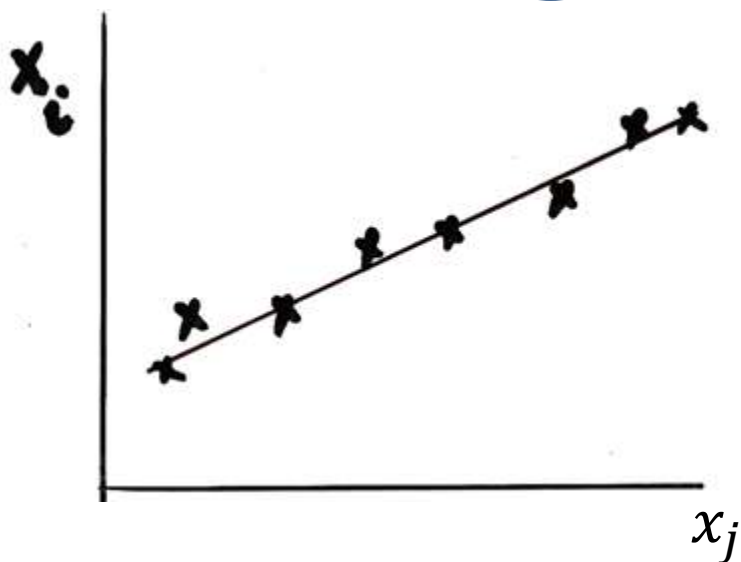


Obr. 4.4 Povrch simultánní hustoty pravděpodobnosti pro $\rho_{12} = 0.9$

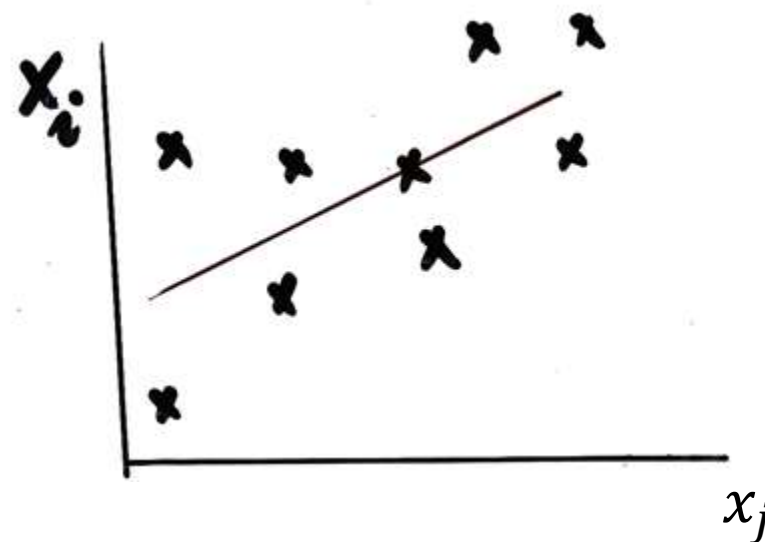


Korelační charakteristiky

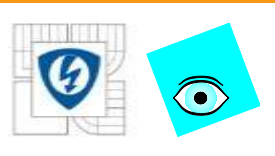
Míry lineární závislosti mezi dvěma či více náhodnými veličinami



Velká korelace $r_{ij} \approx 1$

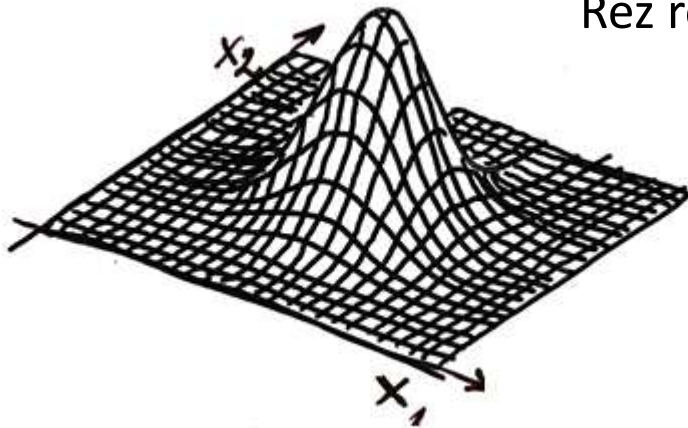


Malá korelace $r_{ij} \approx 0.2$

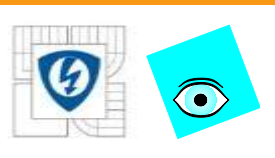


Dvourozměrné normální rozdělení

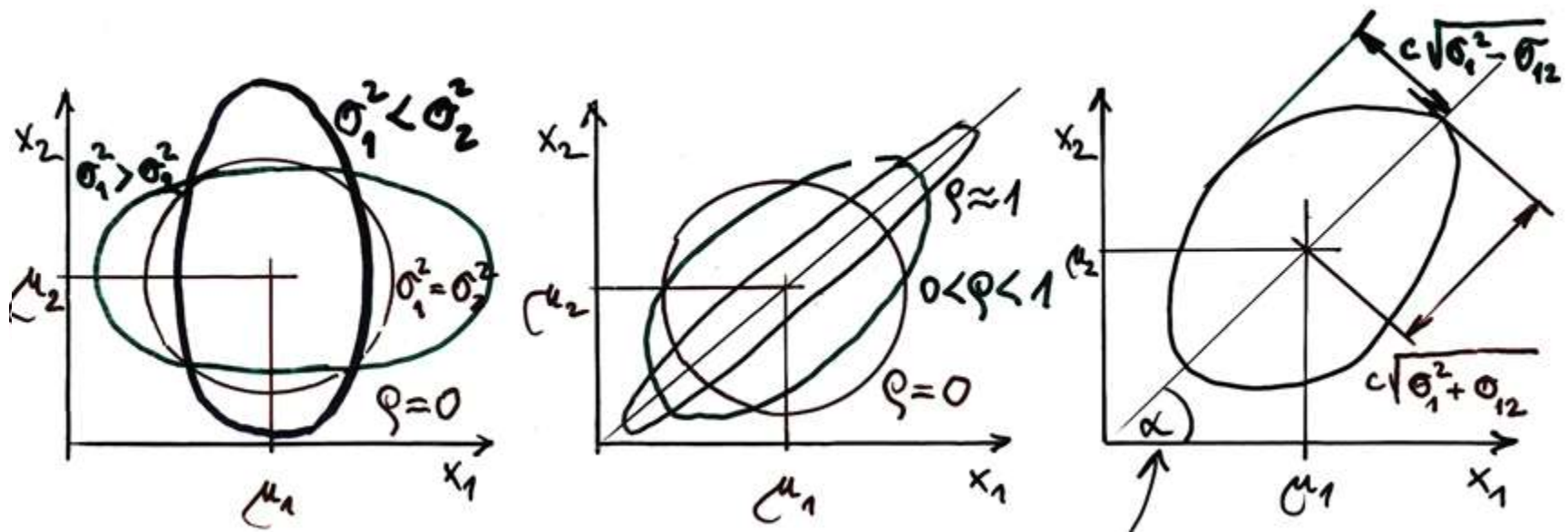
Dvě náhodné veličiny x_1 a x_2 : střední hodnoty μ_1 a μ_2 , rozptyly σ_1^2 a σ_2^2 , kovariance σ_{12} , korelační koeficient r_{12} .



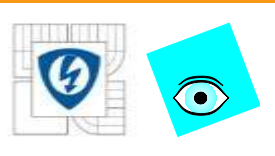
Řez rovinou dá elipsy konstantní hustoty.



Dvourozměrné normální rozdělení



Úhel charakterizuje stupeň lineární závislosti mezi x_1 a x_2



Dvourozměrné normální rozdělení

Postačuje nahradit střední hodnoty μ_1 a μ_2 aritmetickými průměry $\bar{x}_1$ a $\bar{x}_2$, dále rozptyly σ_1^2 a σ_2^2 výběrovými rozptyly s_1^2 a s_2^2 a konečně korelační koeficient ρ výběrovým korelačním koeficientem

$$R = \sqrt{\frac{\sum_{i=1}^n (x_{1i} - \bar{x}_1) (x_{2i} - \bar{x}_2)}{\sum_{i=1}^n (x_{1i} - \bar{x}_1)^2 \sum_{i=1}^n (x_{2i} - \bar{x}_2)^2}}$$

Směrnice b_1 a úsek b_2 regrese $E(\xi_2/x_1)$ odpovídají odhadům určeným metodou nejmenších čtverců a $D(\xi_2/x_1)$ odpovídá reziduálnímu součtu čtverců odchylek.