Statistické metody a zpracování dat

VII. Korelační a regresní počet

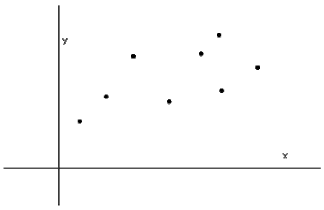
K čemu to je dobré?

Analýza závislostí

- V řadě geografických disciplín studujeme jevy, u kterých vyšetřujeme ne jednu jejich vlastnost (znak), ale znaků několik.
- Tyto znaky mohu být navzájem závislé.
- Cílem této části statistiky je vyšetřovat, do jaké míry spolu dva či více statistických znaků souvisí.
- Do jaké míry změna hodnoty jednoho znaku podmiňuje změnu hodnot znaku jiného.

Příklady použití

Př. *Vztah mezi teplotou vzduchu a nadmořskou výškou, mezi množstvím srážek a velikostí odtoku, mezi výnosy a hodnotami několika meteorologických prvků, mezi počtem dojíždějících a vzdáleností od centra dojížďky, …*



Analýza závislostí

- Předmětem statistické analýzy v tomto případě bude stanovení **síly závislosti** a **druhu závislosti**
- Analýzou síly závislosti statistických znaků se zabývá **korelační počet**
- Analýzou druhu závislosti statistických znaků se zabývá **regresní počet**
- Budeme tedy pracovat s dvourozměrnými soubory
- **Korelační i regresní počet** však lze využít i pro studium vícerozměrných souborů, pro studium znaků kvantitativních i kvalitativních.

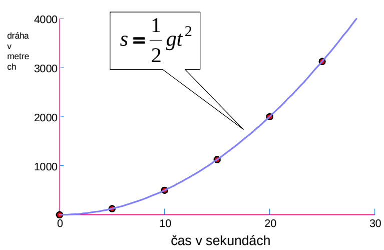
Druhy závislostí

- **Vztahy jednostranné:** Změna statistického znaku jednoho souboru náhodné veličiny - tzv. **nezávisle** proměnné (x) podmiňuje změnu statistického znaku souboru druhé náhodné veličiny - tzv. **závisle** proměnné (y).
- V tomto případě jde o vztahy příčiny a následku
- **Vztahy vzájemné**: Nelze rozlišit mezi souborem závisle a nezávisle proměnné (např. vztah hodnot teploty vzduchu na dvou sousedních stanicích)
- V geografii – tzv. **prostorová autokorelace**

Druhy závislostí:

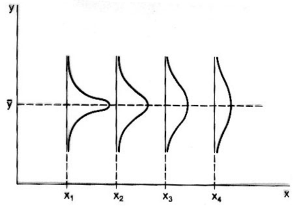
- Závislost funkční
- Závislost statistická
- Závislost korelační

Závislost funkční



Každé hodnotě znaku nezávisle proměnné náhodné veličiny *x* odpovídá vždy pouze jediná určitá hodnota závisle proměnné veličiny *y*

Závislost statistická

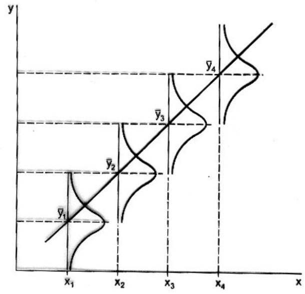


• Každé hodnotě znaku nezávisle proměnné náhodné veličiny *x* odpovídá více hodnot závisle proměnné veličiny *y*,

• Hodnoty *y* mají své rozdělení

• Při změně znaku nezávisle proměnné *x* mění podmíněná rozdělení relativních četností závisle proměnné *y*

Závislost korelační



Se změnou hodnoty znaku nezávisle proměnné *x* se mění podmíněná rozdělení relativních četností hodnoty znaku závisle proměnné *y* tak, že změna *x* podmiňuje změnu průměru $\bar{y}$ souborů hodnot *y*, odpovídajících daným hodnotám *x*.

Určení těsnosti korelační závislosti

• Úkolem korelačního počtu je vyjádřit tendenci změn hodnoty znaku závisle proměnné při změně hodnoty znaku nezávisle proměnné **matematickou funkcí**

• Tato funkce představuje tzv. **regresní čáru** a vyjadřuje, jaká hodnota znaku závisle proměnné odpovídá s největší pravděpodobností určité hodnotě znaku nezávisle proměnné.

• Odhad regresní závislosti je tím přesnější, čím větší je **těsnost korelační závislosti**.

• Určení těsnosti korelační závislosti je prvním krokem analýzy.

Charakteristiky korelační závislosti

Máme dva výběrové soubory náhodných veličin *X*, *Y*. Proměnlivost hodnot znaku obou výběrů můžeme vyjádřit odchylkami d_{xi} a d_{yi} prvků od jejich průměrů:

$$d_{xi} = x_i - \bar{x} \qquad d_{yi} = y_i - \bar{y}$$

Vzájemnou proměnlivost obou výběrových souborů charakterizuje součin odchylek :

$$(x_i - \bar{x}) \cdot (y_i - \bar{y})$$

Suma součinů odchylek vydělaná rozsahem výběrů n určuje tzv. **kovarianci** výběrových souborů s_{xy} – tedy první společnou charakteristiku proměnlivosti obou souborů:

$$s_{xy} = \frac{\sum (x_i - \bar{x}) \cdot (y_i - \bar{y})}{n-1}$$

Charakteristiky korelační závislosti

• Kovariance je obdobou rozptylu

• Omezenost - je mírou **absolutní** – nelze jí použít k porovnání těsnosti vztahu dvou či více dvojic výběrových souborů.

Relativní míra – kovariance dělená součinem směrodatných odchylek s_x a s_y obou výběrů - **korelační koeficient** r_{xy}:

$$r_{xy} = \frac{s_{xy}}{s_x \cdot s_y} = \frac{\frac{1}{n-1}\sum (x_i - \bar{x}) \cdot (y_i - \bar{y})}{\sqrt{\frac{1}{n-1}\sum (x_i - \bar{x})^2 \cdot \frac{1}{n-1}\sum (y_i - \bar{y})^2}}$$

Charakteristiky korelační závislosti

Úpravou výše uvedeného vztahu lze **korelační koeficient** r_{xy} vypočítat také podle následujícího vzorce:

$$r_{xy} = \frac{n\sum xy - \sum x \sum y}{\sqrt{[n\sum x^2 - (\sum x)^2] \cdot [n\sum y^2 - (\sum y)^2]}}$$

(vzorec je uveden pouze pro názornost výpočtu v následujícím příkladě)

Příklad

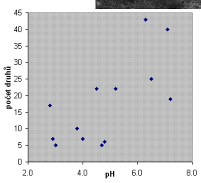
Jaká je závislost mezi pH půdy na výsypkách a počtem rostlinných druhů?

pH		počet druhů		
x	y	x^2	y^2	xy
2.8	17	7.8	289	47.6
2.9	7	8.4	49	20.3
3.8	10	14.4	100	38.0
4.5	22	20.3	484	99.0
7.1	40	50.4	1600	284.0
6.5	25	42.3	625	162.5
3.0	5	9.0	25	15.0
4.7	5	22.1	25	23.5
5.2	22	27.0	484	114.4
4.0	7	16.0	49	28.0
4.8	6	23.0	36	28.8
6.3	43	39.7	1849	270.9
7.2	19	51.8	361	136.8

$\sum x = 62{,}8$ $\sum y = 228$ $\sum xy = 1268{,}8$

$\sum x^2 = 332{,}3$ $\sum y^2 = 59676$

$\left(\sum x\right)^2 = 3943{,}84$ $\left(\sum y\right)^2 = 51984$



$$r_{xy} = \frac{n\sum xy - \sum x \sum y}{\sqrt{[n\sum x^2 - (\sum x)^2]\cdot[n\sum y^2 - (\sum y)^2]}}$$

$$r_{xy} = \frac{13*1268{,}8 - 62{,}8*228}{\sqrt{[13*332{,}3 - 3943{,}84]\cdot[13*5976 - 51984]}}$$

$$r_{xy} = 0{,}700$$

Příklad - pokračování

Interpretace: ze statistických tabulek zjistíme:

Hodnotě $r_{xy} = 0{,}700$ přísluší pro $v = n - 2 = 11$

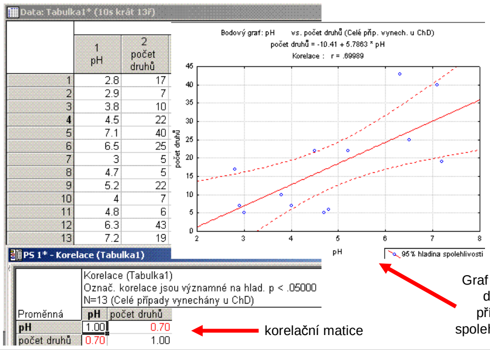
na hladině významnosti α = 0,05 kritická hodnota $r_{krit} = 0{,}553$

Závěr: prokázali jsme statisticky významný vztah mezi pH a množstvím rostlinných druhů rostoucích na výsypkách.

Příklad

Řešení v programu Statistica:

Statistika – Základní statistiky/tabulky – Korelační matice



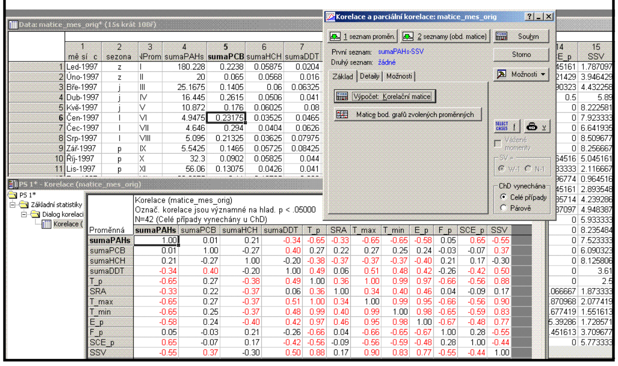
korelační matice

Graf korelačního pole doplněný regresní přímkou a intervaly spolehlivosti (viz. dále)

Příklad

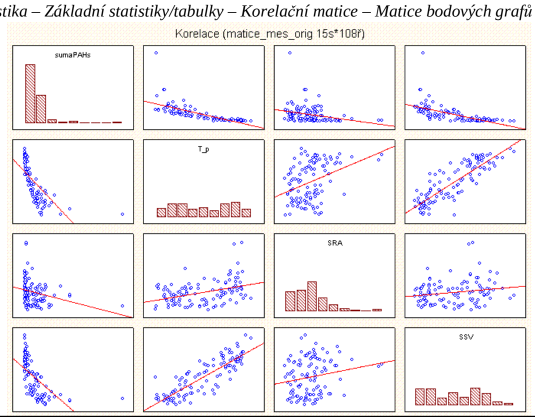
Korelační matice – r_{xy} mezi dvojicemi více proměnných

Statistika – Základní statistiky/tabulky – Korelační matice



Příklad

Statistika – Základní statistiky/tabulky – Korelační matice – Matice bodových grafů



Koeficient determinace

- Koeficient korelace se často ve výpočtech doplňuje hodnotou koeficientu determinace (r^2_{xy}).
- Jeho hodnota kolísá v intervalu 0 až 1
- Vynásoben 100 udává v procentech tu část rozptylu závisle proměnné *y*, která je vysvětlena (podmíněna) změnami hodnot nezávisle proměnné *x*.

V našem případě:

$$r_{xy} = 0{,}700 \longrightarrow r^2_{xy} = 0{,}49 = 49\%$$

Interpretace: Změna počtu druhů rostlin na výsypkách je z 49 % podmíněna změnami pH půdy na kterých tyto rostliny rostou.

Podmínky použitelnosti r_{xy}

Výpočet r_{xy} se opírá o rozptyl a směrodatnou odchylku

Jeho použití tedy předpokládá splnění tří následujících podmínek:

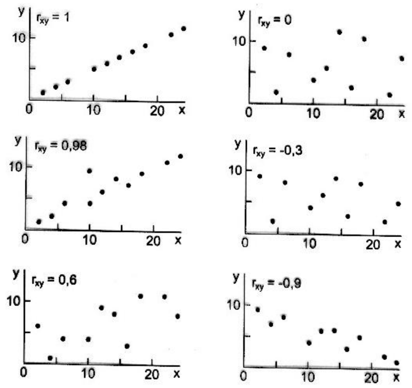
- normální rozdělení použitých výběrů
- dvojrozměrnost normálního rozdělení (každé hodnotě znaku veličiny x odpovídá soubor hodnot znaku y, který má normální rozdělení a naopak
- linearita vztahu hodnot x a y (regresní čára je přímka)

Hodnota r_{xy} nás informuje o druhu a těsnosti závislosti

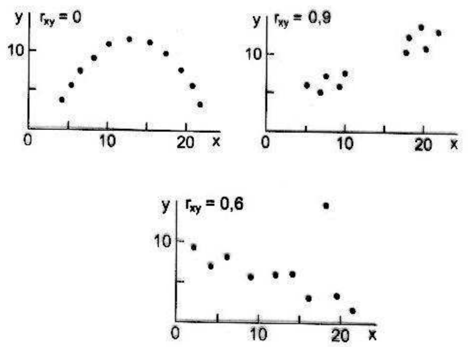
Dokonalá korelační závislost přímá $r_{xy} = 1$

Dokonalá korelační závislost nepřímá $r_{xy} = -1$

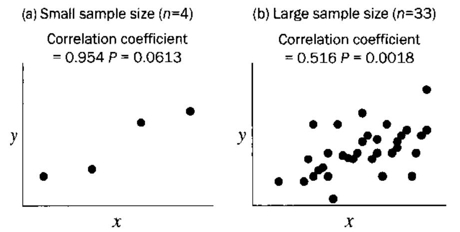
Graf korelačního pole pro různá r_{xy}



Graf korelačního pole pro různá r_{xy} ???



Hodnocení významnosti koeficientu korelace

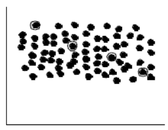


větší r_{xy} → menší *p hodnota* → větší pravděpodobnost zamítnutí nulové hypotézy

větší *n* → menší *p hodnota*

Hodnocení významnosti koeficientu korelace

- Významnost r_{xy} závisí na povaze řešeného problému
- Jeho hodnota je mírou relativní a posouzení těsnosti je do značné míry subjektivní.



Významnost r_{xy} lze též zjistit objektivně – testováním

Hodnocení významnosti koeficientu korelace

Při testování r_{xy} vycházíme z nulové hypotézy, která je $\rho = 0$ (tedy mezi dvěma základními soubory nepředpokládáme žádný korelační vztah).

Testovací kritérium se vypočte podle vztahu:

$$t = \frac{r_{xy}}{\sqrt{1 - r_{xy}^2}} \cdot \sqrt{n-2}$$

Přísluší mu *t*-rozdělení s $\nu = n - 2$ stupni volnosti.

S určitou pravděpodobností - tedy na určité hladině významnosti předpokládáme, že hodnota *t* nepřekročí kritickou hodnotu t_p (při správnosti nulové hypotézy).

V opačném případě zamítáme nulovou hypotézu – mezi výběry náhodných veličin vztah existuje.

Hodnocení významnosti koeficientu korelace - tabulky

- 177 -

Příloha VIII. Kritické hodnoty výběrového koeficientu korelace r_p za předpokladu, že $\rho = 0$, pro počet stupňů volnosti $\nu = n - 2$

ν	p 0,05	p 0,01	ν	p 0,05	p 0,01
1	0,9969	0,9999	16	0,4683	0,5897
2	0,9500	0,9900	17	0,4555	0,5751
3	0,8783	0,9587	18	0,4438	0,5614
4	0,8114	0,9172	19	0,4329	0,5487
5	0,7545	0,8745	20	0,4227	0,5368
6	0,7067	0,8343	25	0,3809	0,4869
7	0,6664	0,7977	30	0,3494	0,4487
8	0,6319	0,7646	35	0,3246	0,4182
9	0,6021	0,7348	40	0,3044	0,3932
10	0,5760	0,7079	45	0,2875	0,3721
11	0,5529	0,6835	50	0,2732	0,3541
12	0,5324	0,6614	60	0,2500	0,3248
13	0,5139	0,6411	70	0,2319	0,3017
14	0,4973	0,6226	80	0,2172	0,2830
15	0,4821	0,6055	100	0,1946	0,2540

Koeficient pořadové korelace (Spearmanův) (r_s)

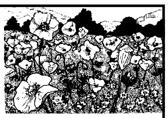
Používá se k určení závislosti **kvalitativních znaků.**

Každé hodnotě x_i a y_i přiřadíme pořadové číslo px_i a py_i podle velikosti hodnot x_i a y_i.

Určíme rozdíly D_i dvojic pořadových čísel odpovídajících si hodnot.

$$r_s = 1 - \frac{6\sum D_i^2}{n \cdot (n^2 - 1)}$$

Koeficient pořadové korelace - příklad



Příklad: Kvantifikujte vztah mezi dobou, po kterou jsou pole ponechána ladem a počtem rostlinných druhů (na m²).

Zjištěná data		Pořadová čísla		Diference	
Počet roků	Počet druhů	Počet roků	Počet druhů	D	D²
1	2	1	1	0	0
2	3	2	2	0	0
3	5	3	4	-1	1
4	4	4	3	-1	1
8	7	5	6,5	-1,5	2,25
10	6	6	5	1	1
> 10	7	7	6,5	0,5	0,25

$\sum D_i^2 = 5{,}5$

$$r_s = 1 - \frac{6\sum D_i^2}{n \cdot (n^2 - 1)} = 1 - \frac{6 \times 5{,}5}{7 \times (49 - 1)} = 0{,}902$$

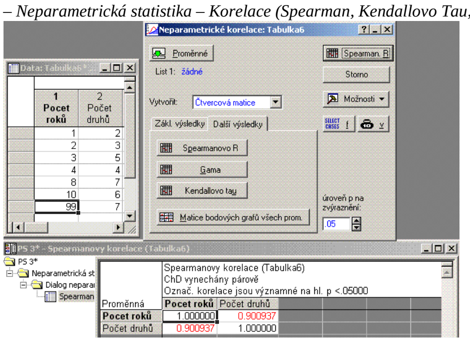
V tabulkách vyhledáme pro *n=7* a *α=0,05* kritickou hodnotu: $r_{krit}=0{,}786$

Závěr: Existuje statisticky významný vztah mezi dobou, po kterou jsou pole ponechána ladem a počtem rostlinných druhů, které se na nich vyskytují.

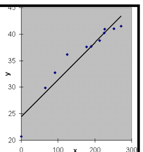
Koeficient pořadové korelace

Řešení v programu Statistica:

Statistika – Neparametrická statistika – Korelace (Spearman, Kendallovo Tau, Gama)



Nelineární závislost



V případě, kdy regresní čára není přímka, ale je vyjádřena složitější matematickou funkcí, se jako míry korelační závislosti používá tzv. korelační poměr (η_{yx}).

Prvky výběru závisle proměnné y_i rozdělíme podle hodnot nezávisle proměnné x_i do skupin označených y_j a pro každou skupinu vypočteme průměr $\bar{y}_j$. Korelační poměr se vypočte podle vztahu:

$$\eta_{yx} = \sqrt{\frac{\sum (y_j - y) \cdot n_j}{\sum (y_i - y)^2}} = \sqrt{\frac{\sum (y_j n_j - ny)^2}{\sum y_i^2 - ny^2}}$$

V uvedeném vzorci je n_j četnost v y_j. Při výpočtu **záleží** na tom, kterou proměnou zvolíme za závislou a kterou za nezávislou.

Porovnání hodnot korelačního koeficientu a korelačního poměru lze použít jako kritéria linearity vztahu.

Pokud se hodnoty přibližně rovnají, jedná se o závislost lineární, pokud je r_{xy} výrazně větší, jde o závislost nelineární.

Koeficient mnohonásobné korelace (r_{xyz})

Vztah dvou proměnných je často ovlivněn dalšími proměnnými.

Používá se pro hodnocení korelační závislosti tří nebo více výběrů náhodných veličin.

Při jeho určení se vychází z jednotlivých korelačních koeficientů pro dva výběry (r_{xy}, r_{xz}, r_{yz}) a jejich hodnoty se dosazují do vzorce pro r_{xyz}:

$$r_{xyz} = \sqrt{\frac{r_{xz}^2 + r_{yz}^2 - 2r_{xy} \cdot r_{xz} \cdot r_{yz}}{1 - r_{xy}^2}}$$

Příklad – viz. vícerozměrná regrese

Dílčí (parciální) korelace:

Řeší otázku vlivu jedné nebo více nezávisle proměnných na závisle proměnnou při **vyloučení vlivu** zbývajících nezávisle proměnných, u nichž předpokládáme konstantní hodnotu.

Jedná se o zvláštní případ mnohonásobné korelace, kdy další proměnné považujeme za „**rušivé**“ (např. věk, počet obyvatel sídla, …).

Hodnota koeficientu dílčí korelace $r_{xy.z}$ se vypočte podle vztahu:

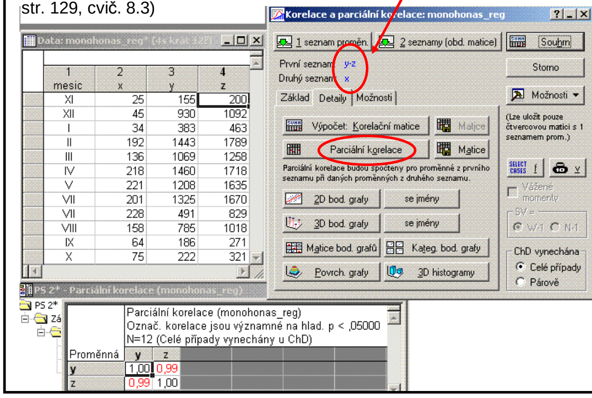
$$r_{xy \cdot z} = \frac{r_{xy} - r_{xz} \cdot r_{yz}}{\sqrt{(1 - r_{xz}^{2}) \cdot (1 - r_{yz})^{2}}}$$

Tečkou v indexu se označuje nezávisle proměnná, jejíž hodnotu považujeme za konstantní.

Parciální korelace

Příklad (viz. Brázdil a kol., 1995, str. 129, cvič. 8.3)

Způsob zadání proměnných (korelace mezi *y* a *z* při vyloučení vlivu *x*)



1 mesic	2 x	3 y	4 z
XI	25	155	200
XII	45	930	1092
I	34	383	463
II	192	1443	1789
III	136	1069	1258
IV	218	1460	1718
V	221	1208	1635
VII	201	1325	1670
VII	228	491	829
VIII	158	785	1018
IX	64	186	271
X	75	222	321

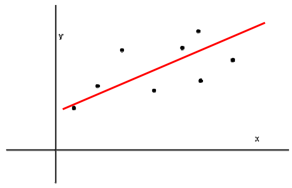
Proměnná	y	z
y	1,00	0,99
z	0,99	1,00

Poznámky k aplikaci korelačního počtu:

Použití korelačního počtu je nevhodné např. v těchto případech:

- Korelace je způsobena formálními vztahy mezi veličinami (hodnoty x a y se doplňují do 100%)
- Korelace je způsobena nehomogenitou studovaného materiálu (obsahuje tzv. subpopulace – viz. obr. bodového grafu)
- Korelace je výsledkem působení třetí veličiny (korelace mezi počtem lékařů a počtem nemocných, …)

Regresní analýza



Úkolem regresní analýzy je sestavit **vztah (model)** závislosti mezi závisle a nezávisle proměnnou.

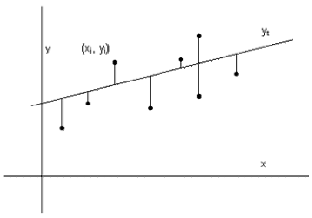
Regresní analýza řeší :

- odhady neznámých parametrů regresní funkce
- testování hypotéz o těchto parametrech
- ověřování předpokladů regresního modelu

Určení lineární regresní závislosti

Nejjednodušším případem regresní závislosti je případ, kdy regresní funkce je přímkou. Rovnice regresní přímky má tvar:

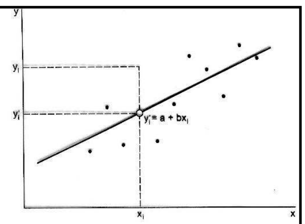
y’ = a + bx



Symbol *y’* se používá pro označení **nejpravděpodobnější teoretické hodnoty** y odpovídající danému x, která leží na regresní přímce a která se odlišuje od konkrétních hodnot y_i, které se nacházejí mimo ni.

MNČ

Průběh regresní přímky je určen tzv. **metodou nejmenších čtverců**, kdy musí být splněna podmínka takového průběhu přímky, při kterém je součet čtverců vzdálenosti všech bodů pole od přímky minimální, tedy platí:



$$\sum (y_i - y_i^{'})^2 = \min$$

Výpočet vertikální vzdálenosti bodů korelačního pole od regresní přímky se provádí podle uvedeného obrázku. Z něho je zřejmé, že pro vzdálenost konkrétní hodnoty závisle proměnné y_i od bodu regresní přímky y_i’ musí platit:

$$y_i - y_i^{'} = y_i - (a + bx_i) = y_i - a - bx_i$$

Součet čtverců svislých vzdáleností y_i od regresní přímky je potom:

$$\sum (y_i - y_i^{'})^2 = \sum (y_i - a - bx_i)^2 = A$$

Pro MNČ musí platit

$$A = \sum (y_i - a - bx_i)^2 = \min$$

Výpočet koeficientů regresní přímky

Z výše uvedených vztahů lze následnými úpravami obdržet výrazy pro výpočet koeficientů regresní přímky *a,b*

$$b = \frac{\sum x_i y_i - n\bar{x}\bar{y}}{\sum x_i^2 - n\bar{x}^2} \qquad a = \bar{y} - b\bar{x}$$

Koeficient b (angl. slope) se označuje jako koeficient regrese a je směrnicí regresní přímky (tangentou úhlu, který přímka a svírá s osou x). Je-li b>0, mluvíme o regresi pozitivní, je-li b<0 o regresi negativní**.**

Výpočet koeficientů regresní přímky

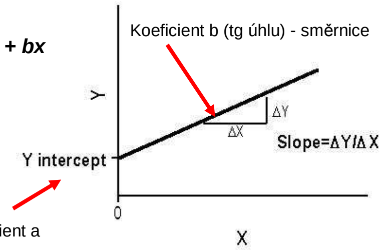
Vzorec pro výpočet koeficientu b lze zjednodušit pomocí vztahů pro kovarianci s_{xy} a směrodatnou odchylku s_x, tedy:

$$b = \frac{s_{xy}}{s_x^2}$$

Hodnota **koeficientu *a*** (angl. intercept) představuje y-ovou souřadnici průsečíku regresní přímky s osou y (tedy při x=0).

Koeficienty lineární regresní závislosti

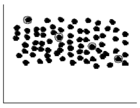
y' = a + bx



Koeficient a

Hranice (EXCEL)

Abs. člen (Statistica)



Koeficienty (parametry) jsou bodovými odhady !

Intervaly a pásy spolehlivosti lineární regresní závislosti

- Konstrukci regresní přímky provádíme na základě výběrových souborů.
- Proto se její rovnice může u různých výběrů ze stejných základních souborů lišit.
- Z tohoto důvodu je potřebné doplnit průběh regresní přímky také tzv. **intervaly spolehlivosti.**
- Výpočtem intervalů spolehlivosti určujeme pro vybraná x interval, v němž se mohou s určitou pravděpodobností vyskytovat hodnoty y s tím, že jejich nejreprezentativnější hodnota je y'.

Intervaly a pásy spolehlivosti

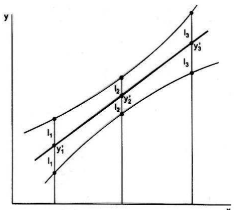
Nejprve je zapotřebí zvolit hladinu spolehlivosti – tedy pravděpodobnost, s níž očekáváme výskyt hodnot *y* v určených mezích *1-p (p=0,05 či 0,01)*. Poloviční šířka intervalu spolehlivosti *l* je dána výrazem:

$$l = t_{1-p} \cdot \frac{h\sqrt{A}}{\sqrt{n-2}} \qquad h = \sqrt{\frac{1}{n} + \frac{(x-\bar{x})^2}{(n-1)s_x^2}}$$

Hodnota t_p je kritická hodnota rozdělení pro *n-2* stupňů volnosti a hladinu významnosti *p*. Meze intervalů spolehlivosti určíme pomocí hodnot *y'* z rovnice $y' - \bar{y} = b(x - \bar{x})$

horní mez: *y' + l*

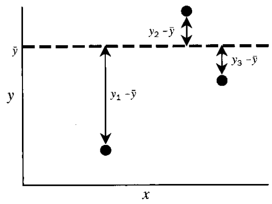
dolní mez: *y' - l*



Pásy spolehlivosti vzniknou spojením krajních bodů intervalů spolehlivosti.

Testování významnosti regresní závislosti

- K testování významnosti zjištěné regresní závislosti lze využít **t-testu**, kterým lze zjistit, zda se směrnice významně liší od nuly
- Nejčastěji se k testování používá **analýzy rozptylu (ANOVA).**
- **Princip:** Zjistíme celkovou proměnlivost hodnot *y* a následně vypočteme, z jaké části je tato celková variabilita objasněna proměnlivostí v hodnotách *x*.



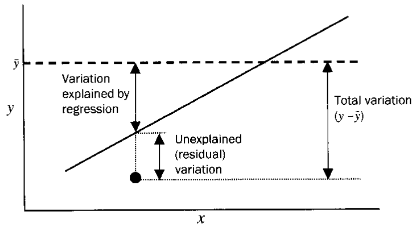
SS_{total} - **celková variabilita**: celková suma čtverců: od každé hodnoty y odečteme průměr, výsledek povýšíme na druhou a sečteme pro všechna y.

Testování významnosti regresní závislosti

Celkovou variabilitu SS_{total} lze rozdělit na dvě části:

$SS_{regrese}$ - variabilitu **vysvětlenou** regresní čarou

$SS_{reziduální}$ – zbytková variabilita **nevysvětlená** regresním modelem



$$SS_{reziduální} = SS_{total} - SS_{regrese}$$

Testování významnosti regresní závislosti

Tabulka ANOVA

Variabilita	stupně volnosti (df)	Suma čtverců (SS)	Průměr sumy čtverců (MS)	F hodnota	p hodnota
Regresní	1	SS_{regres}	$\frac{SS_{regres}}{1}$	$\frac{MS_{regres}}{MS_{resid}}$	F hodnota pro df = 1 a n-2
Reziduální	n-2	$SS_{rezidual}$	$\frac{SS_{residual}}{n-2}$		
Celková	n-1	SS_{total}			

Koeficient determinace regresní závislosti:

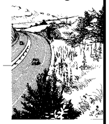
$$r^2 = \frac{SS_{regres}}{SS_{total}}$$

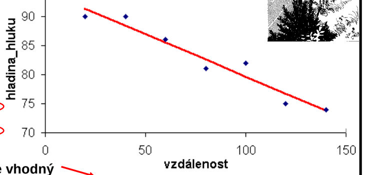
Příklad regresní analýzy v EXCELu

Nástroje – Analýza dat - Regrese

vzdálenost	hladina_hluku
20	90
40	90
60	86
80	81
100	82
120	75
140	74

Zjistěte, jak souvisí hladina hluku se vzdáleností od komunikace.





VÝSLEDEK

Regresní statistika	
Násobné R	0,969074891
Hodnota spolehlivosti R	0,939106145
Nastavená hodnota spolehlivosti R	0,926927374
Chyba stř. hodnoty	1,764733893
Pozorování	7

ANOVA

model je vhodný

	Rozdíl	*SS*	*MS*	*F*	*Významnost F*
Regrese	1	240,1428571	240,1429	77,11009	0,000317686
Rezidua	5	15,57142857	3,114286		
Celkem	6	255,7142857			

95% int. odhad hladiny hluku ve vzdálenosti 0 metrů

95% int. odhad poklesu hl. hluku na každý metr

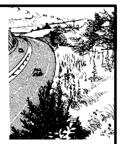
	Koeficienty	*Chyba stř. hodnoty*	*t stat*	*Hodnota P*	*Dolní 95%*	*Horní 95%*	*Dolní 95,0%*	*Horní 95,0%*
Hranice	94,2857	1,4915	63,2165	0,0000	90,4518	98,1197	90,4518	98,1197
vzdálenost	-0,1464	0,0167	-8,7812	0,0003	-0,1893	-0,1036	-0,1893	-0,1036

y' = 94,2857 - 0,1464x

pokles hl. hluku na každý metr

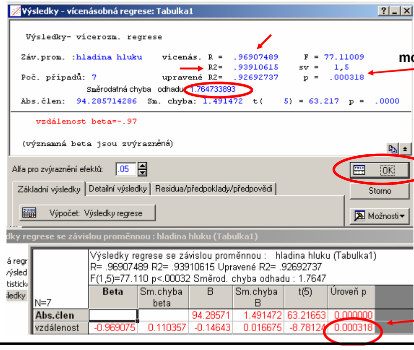
Existuje **signifikantní pokles hladiny hluku** se vzdáleností od komunikace. Lineární regresní model vysvětluje 93,9 % variability hodnot hladiny hluku.

Řešení v programu Statistica



1) Statistika – Vícerozměrná regrese

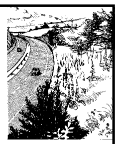
(zvolení závisle a nezávisle proměnné)



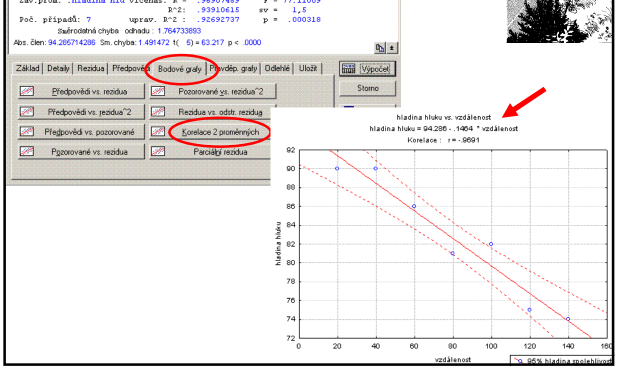
model je vhodný

pokles je statisticky významný

Řešení v programu Statistica

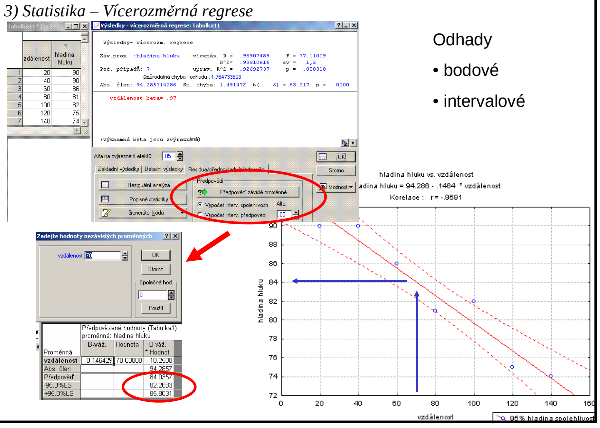


2) OK – Bodové grafy – Korelace 2 proměnných



Výpočet neznámé hodnoty - předpovědi

3) Statistika – Vícerozměrná regrese



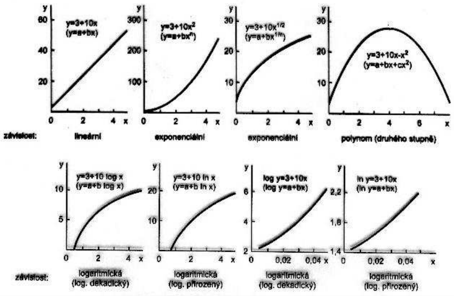
Odhady

- bodové
- intervalové

Další typy regresních funkcí

Regresní vztah dvou proměnných často nelze vhodně vyjádřit přímkou – jiné typy funkcí.

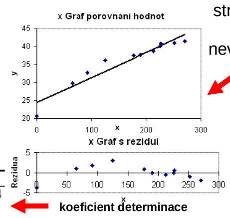
Může mít tvar např. logaritmických či exponenciálních funkcí a nebo je vztah vyjádřen rovnicí polynomu m-tého stupně.



Regresní závislost není přímka

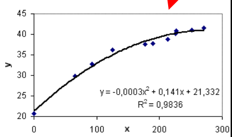
Příklad (viz. Brázdil a kol., 1995, str. 139, cvič. 8.5)

Stanice	x	y
Ivančice	0	20,7
Sen. Mlýn	65	29,9
Jakub	92	32,7
Senorady	125	36,2
Zastávka	177	37,6
Kratochvilka	190	37,7
Příbram	213	38,8
Bučín	226	40,2
Hlína	227	40,9
Ketkovice	252	41
Vys. Popovice	271	41,5



nevhodný model

vhodný model



VÝSLEDEK

Regresní statistika	
Násobné R	0,956724
Hodnota spc	0,915321
Nastavená h	0,905912
Chyba stř. h	1,921192
Pozorování	11

koeficient determinace

směrodatná chyba odhadu

ANOVA

	Rozdíl	SS	MS	F	Významnost F
Regrese	1	359,0703	359,0703	97,28322	4,01237E-06
Rezidua	9	33,21881	3,690979		
Celkem	10	392,2891			

	Koeficienty	ba stř. hod.	t stat	Hodnota P	Dolní 95%	Horní 95%	Dolní 95,0%	Horní 95,0%
Hranice	24,43416	1,31782	18,54135	1,77E-08	21,45303879	27,41527	21,45304	27,41527
x	0,069872	0,007084	9,863225	4,01E-06	0,053846464	0,085897	0,053846	0,085897

EXCEL:

pravým tlačítkem myši na graf –

přidat spojnici trendu

Hledání vhodného regresního modelu

Lze postupovat dvěma způsoby:

1. Volba vhodného modelu na základě praktické zkušenosti či teoretických předpokladů
2. Posouzením bodového grafu a interpretací nástrojů regresní analýzy

Způsoby hodnocení vhodnosti regresního modelu

- analýza reziduálních hodnot
- výpočet směrodatné chyby odhadu (s_e)
- výpočet koeficientu determinace (r^2_{xy}).

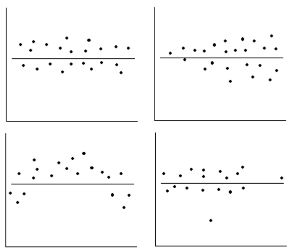
Hledání vhodného regresního modelu

Analýza reziduálních hodnot

Rezidua jsou vzdálenosti skutečných hodnot y_i od modelem odhadnutých hodnot y_i`

Zvolený regresní model považujeme za vhodný, pokud reziduální hodnoty splňují všechny následující podmínky:

- rezidua jsou ***náhodná*** a ***nezávislá***
- mají ***normální rozdělení*** s nulovým průměrem a konstantním rozptylem
- rozptyl reziduí je ***konstantní***.



Hledání vhodného regresního modelu

Směrodatná chyba odhadu – je vyjádřením směrodatné odchylky resp. rozptylu reziduálních hodnot a vhodnou mírou pro posouzení vhodnosti použité regresní závislosti

$$s_e = \sqrt{\frac{\sum_{i=1}^{n}(y_i - y_i^`)^2}{n-2}}$$

Čím je hodnota reziduálního rozptylu nižší, tím je model vhodnější.

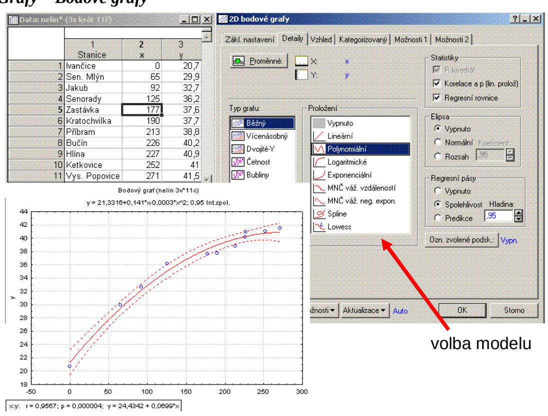
Koeficient determinace (r^2_{xy}) – viz. Korelační počet

$$r^2 = \frac{SS_{regres}}{SS_{total}}$$

Čím je hodnota koeficientu determinace větší, tím je model vhodnější.

Hledání vhodného regresního modelu

Grafy – Bodové grafy



volba modelu

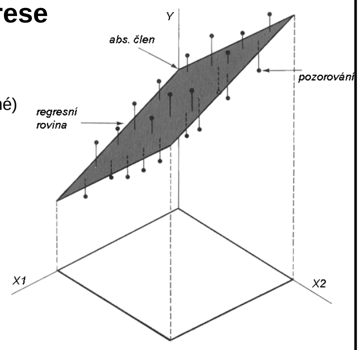
Vícerozměrná regrese

Popisuje závislost více proměnných z nichž více je příčinami (vysvětlující proměnné) a jen jedna je důsledek (vysvětlovaná proměnná).

Jsou–li dvě vysvětlující proměnné regresní model je rovina

Odhad parametrů se provádí MNČ

Výstupy a interpretace jsou „**obdobné**“ jako u modelů jednorozměrné regrese



$$y' = a + b_1x_1 + b_2x_2 + \ldots$$

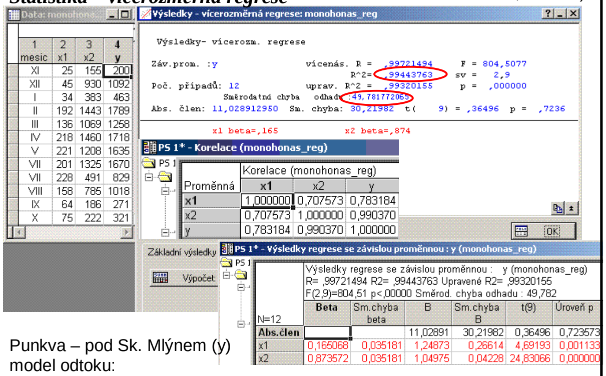
Např:

*Úhrn_srážek = 345,6 + 0,45*zem_délka + 1,23*nadm_výška*

Vícerozměrná regrese

(data viz. Brázdil a kol., 1995, str. 129, cvič. 8.3)

Statistika – vícerozměrná regrese



1 mesic	2 x1	3 x2	4 y
XI	25	155	200
XII	45	930	1092
I	34	383	463
II	192	1443	1789
III	136	1069	1258
IV	218	1460	1718
V	221	1208	1635
VII	201	1325	1670
VII	228	491	829
VIII	158	785	1018
IX	64	186	271
X	75	222	321

Výsledky - vícerozměrná regrese: monohonas_reg

Výsledky- vícerozm. regrese

Záv.prom. :y vícenás. R = ,99721494 F = 804,5077
R^2= ,99443763 sv = 2,9
Poč. případů: 12 uprav. R^2 = ,99320155 p = ,000000
Směrodatná chyba odhadu :49,78177206
Abs. člen: 11,028912950 Sm. chyba: 30,21982 t(9) = ,36496 p = ,7236

x1 beta=,165 x2 beta=,874

PS 1* - Korelace (monohonas_reg)

Proměnná	x1	x2	y
x1	1,000000	0,707573	0,783184
x2	0,707573	1,000000	0,990370
y	0,783184	0,990370	1,000000

PS 1* - Výsledky regrese se závislou proměnnou : y (monohonas_reg)

Výsledky regrese se závislou proměnnou : y (monohonas_reg)
R= ,99721494 R2= ,99443763 Upravené R2= ,99320155
F(2,9)=804,51 p<,00000 Směrod. chyba odhadu : 49,782

N=12	Beta	Sm.chyba beta	B	Sm.chyba B	t(9)	Úroveň p
Abs.člen			11,02891	30,21982	0,36496	0,723573
x1	0,165068	0,035181	1,24873	0,26614	4,69193	0,001133
x2	0,873572	0,035181	1,04975	0,04228	24,83066	0,000000

Punkva – pod Sk. Mlýnem (y) model odtoku:

$$y = 11{,}0289 + 0{,}165x_1 + 0{,}874x_2$$

Měření závislosti kvalitativních znaků

- Kvalitativní znaky mají slovní charakter a získáváme je v sociologických průzkumech, při terénním šetření apod.
- K charakterizování závislostí kvalitativních znaků slouží tzv. **kontingenční tabulky**

		Inteligenční kvocient – IQ							Součet Σ
		95	98	100	105	110	112	120	
Obratnost	15		1						1
	16		1						1
	18	1	1						2
	20	1	2	3	4				10
	22			3	3	1			7
	25				1	3	1		5
	26				1	2	1		4
	32						1		1
	37							1	1
Σ		2	5	6	9	6	3	1	32

- Z kontingenční tabulky lze určit intenzitu závislosti ve dvojici slovních znaků.
- Máme-li dva alternativní znaky dostaneme tzv. **čtyřpolní tabulku**.

	praváci	leváci	celkem
muži	43	9	52
ženy	44	4	48
celkem	87	13	100

Měření závislosti kvalitativních znaků

Obecně může mít každý kvalitativní znak A *r* tříd a znak B *s* tříd. Výsledky šetření potom sestavujeme do kontingenční tabulky *r x s*.

Pozorované četnosti v jednotlivých buňkách označujeme dvěma indexy – obecně n_{ij}.

Také marginální četnosti mají dva indexy.

Ten, přes který je sčítáno je označen hvězdičkou – tedy n_{2*} značí součet četností v druhé řádce, n_{*1} značí součet četností v prvním sloupci.

Tabulka bývá doplněna hodnotami procentuálních (relativních) četností. Častým požadavkem je konstantní délka intervalů tvořících třídy.

Stejně jako v případě kvantitativních znaků ověřujeme i zde existenci vztahu testy významnosti a hodnotíme ho vhodnou mírou závislosti.

Kontingenční tabulka typu r x s

Tříděný znak		Znak B						Součet
		b_1	b_2	…	b_j	…	b_s	
Znak A	a_1	n_{11}	n_{12}				n_{1s}	n_{1*}
	a_2	n_{21}						n_{2*}
	⋮							⋮
	a_i				n_{ij}			n_{i*}
	⋮							⋮
	a_r	n_{r1}					n_{rs}	n_{r*}
Součet		n_{*1}	n_{*2}	…	n_{*j}	…	n_{*s}	$n_{**} = n$

Posuzování závislosti v kontingenčních tabulkách

Podmíněné četnosti uvnitř kontingenční tabulky mají podobný význam jako body korelačního diagramu — jejich rozmístění umožňuje usuzovat na charakter závislosti tříděných znaků.

Pro posouzení nezávislosti obou znaků můžeme vedle pozorovaných četností stanovit pro jednotlivá pole také očekávané (teoretické) četnosti :

$$n'_{ij} = \frac{n_{i*}n_{*j}}{n}$$

tedy jako součin okrajových četností příslušného řádku a sloupce dělený rozsahem souboru.

Pro každé pole kontingenční tabulky existuje dvojice četností - četnost pozorovaná a četnost vypočtená.

Hypotéza nezávislosti

Ukazatel, který pro tabulku jako celek měří rozdílnost pozorovaných a vypočtených četností v jednotlivých polích tabulky se nazývá čtvercová kontingence χ^2

$$\chi^2 = \sum_{i=1}^{r}\sum_{j=1}^{s}\frac{(n_{ij}-n'_{ij})^2}{n'_{ij}}$$

Je to bezrozměrná hodnota a platí: $\chi^2 \geq 0$

Hodnoty nula nabývá pouze v případě, že znaky v kontingenční tabulce jsou nezávislé.

Vypočtená hodnota χ^2 se porovnává na zvolené hladině významnosti α s kritickou hodnotou χ^2 rozdělení pro *(r-1)(s-1)* stupňů volnosti.

Hypotézu zamítáme, jestliže vypočtená hodnota je větší než tabulková, případně, když jí příslušející *p-hodnota* je menší než zvolená hladina významnosti.

Příklad analýzy závislosti v tabulce r x s

Pro výběr 234 studentů zjišťujeme, zda existuje vztah mezi sportem, který provozují a sportovními pořady, které sledují v televizi.

Sestavíme tabulku typu 4 x 4:

Oblíbenost při sledování televize	Oblíbenost při sportování				Řádkové součty
	hry	atletika	gymnastika	plavání	
hry	133	6	2	4	145
atletika	15	10	4	3	32
gymnastika	4	1	25	0	30
plavání	9	0	1	17	27
Sloupcové součty	161	17	32	24	234

Hypotéza nezávislosti H_0: Neexistuje vztah mezi provozovaným sportem a sportem sledovaným v TV.

Vypočtená hodnota testovacího kritéria $\chi^2 = 273{,}3$

Kritická hodnota z tabulek pro *p=0,05* a *(4-1)x(4-1)*=9 stupňů volnosti:

$\chi^2 = 16{,}9$

Závěr: H_0 zamítáme, existuje významný vztah.

Testování nezávislosti v tabulce 2 x 2

	Zájem o statistiku		řádkové součty
	ano	ne	
chlapci	a	b	a + b
dívky	c	d	c + d
sloupcové součty	a + c	b + d	n

Pro výpočet testovacího kritéria χ^2 v tabulce 2 x 2 můžeme využít zjednodušený vzorec:

$$\chi^2 = \frac{n(ad-bc)^2}{(a+b)(c+d)(a+c)(b+d)}$$

Protože v 2x2 tabulce můžeme uvažovat i směr poruchy nulové hypotézy – proto musíme rozhodnout, zda použijeme test jednostranný či dvoustranný.

Kritické hodnoty jsou uvedeny v tabulce χ^2- rozdělení o jednom stupni volnosti.

Příklad analýzy závislosti v tabulce 2 x 2

	Zájem o statistiku		řádkové součty
	ano	ne	
chlapci	30	36	66
dívky	11	63	74
sloupcové součty	41	99	140

Hypotéza nezávislosti H_0: Relativní četnost studentů se zájmem o statistiku je nezávislá na pohlaví.

Vypočtená hodnota testovacího kritéria: $\chi^2 = \frac{140(30\times 63-11\times 36)^2}{41\times 99\times 66\times 74} = 15{,}8$

Kritická hodnota χ^2-rozdělení z tabulek pro α=*0,05:* 3,84

Závěr: H_0 zamítáme, existuje významný rozdíl.

Zájem u chlapců: 30/66 = 0,45

Zájem u dívek: 11/74 = 0,14

Chlapci mají zhruba 3x větší zájem o statistiku než dívky.

… to byl jen smyšlený příklad