MASARYKOVA UNIVERZITA

Z1069 Statistické metody a zpracování dat

VI. Korelační a regresní počet

INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ

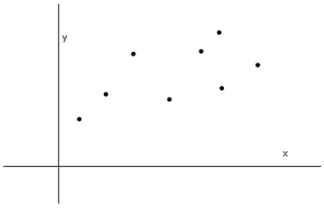
K čemu to je dobré?

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

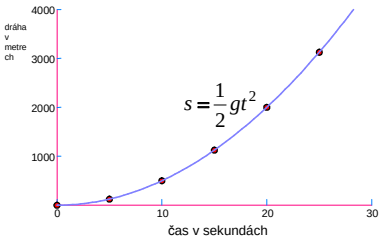
- Předmětem statistické analýzy v tomto případě bude stanovení **síly závislosti** a **druhu závislosti**
- Analýzou síly závislosti statistických znaků se zabývá **korelační počet**
- Analýzou druhu závislosti statistických znaků se zabývá **regresní počet**
- Budeme tedy pracovat s dvourozměrnými soubory
- **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)

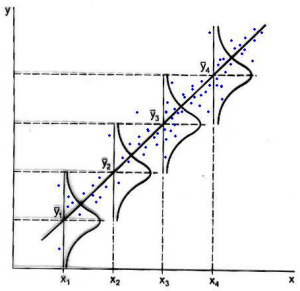
Druhy závislostí:
- závislost funkční
- závislost korelační

Závislost funkční



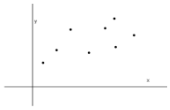
Každé hodnotě znaku nezávisle proměnné náhodné veličiny *x* odpovídá vždy pouze jediná určitá hodnota závisle proměnné veličiny *y*

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*.

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:

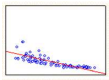
$$r_{xy} = \frac{n\sum xy - \sum x \sum y}{\sqrt{[n\sum x^2 - (\sum x)^2] \cdot [n\sum y^2 - (\sum y)^2]}}$$

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

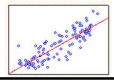
Interpretace r_{xy}

Hodnota korelačního koeficientu kolísá v intervalu od -1 do 1

- r_{xy} -> 0 nezávislost
- r_{xy} -> -1 nepřímá závislost



- r_{xy} -> 1 přímá závislost



Příklad

Ze 13 různých lokalit na výsypkách máme k dispozici měření pH půdy a údaje o počtu rostlinných druhů. Máme zjistit, zda existuje závislost mezi pH a počtem rostlinných druhů?

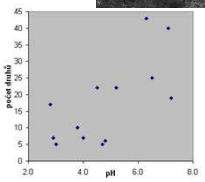
pH počet druhů

x	y	x²	y²	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

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

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

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



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

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

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

Příklad - pokračování

- Je zjištěný vztah statisticky významný?
- (H0: r_{xy} se významně neliší od nuly – viz. dále)

Ze statistických tabulek zjistíme:

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

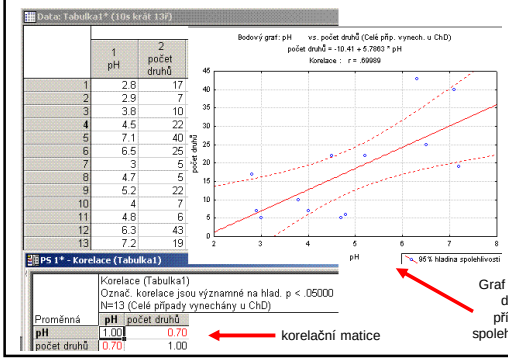
na hladině významnosti $\alpha = 0{,}05$ kritická hodnota $r_{krit} = 0{,}553$

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



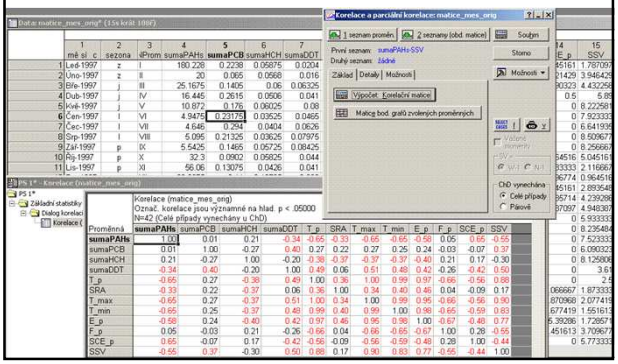
korelační matice

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

Příklad

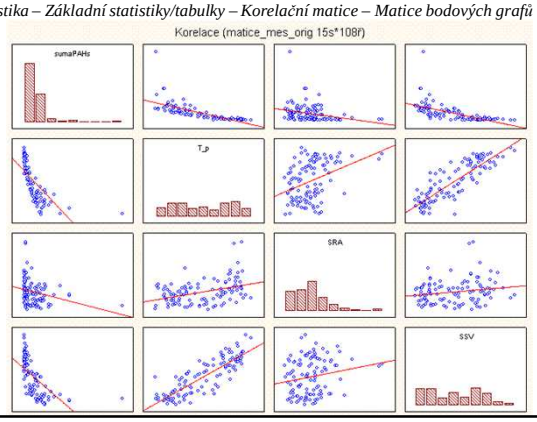
Korelační matice – r_{xy} mezi dvojicemi více proměnných

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:

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

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:

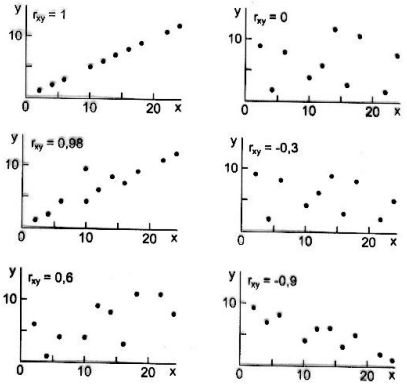
- normální rozdělení použitých výběrů
- dvojrozměrnost normálního rozdělení (každé hodnotě znaku veličiny x odpovídá soubor hodnot znaku y, který má normální rozdělení a naopak
- linearita vztahu hodnot x a y (regresní čára je přímka)

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

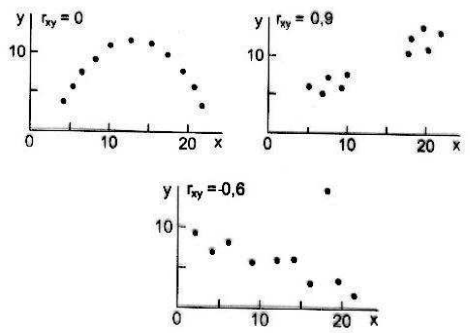
Dokonalá korelační závislost přímá $r_{xy} = 1$

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

Graf korelačního pole pro různá r_{xy}



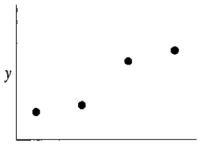
Graf korelačního pole pro různá r_{xy} ???



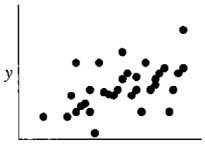
Důležitá role explorační (průzkumové) analýzy dat

Hodnocení významnosti koeficientu korelace

- malý rozsah souboru (n = 4)
- rxy = 0.95
- p = 0.0613



- velký rozsah souboru (n = 33)
- rxy = 0.52
- p = 0.0018



• Při velkém rozsahu souboru (*n*) roste pravděpodobnost, že i relativně malá hodnota korelačního koeficientu (r_{xy}) nám vyjde jako statisticky významná – tedy zamítneme nulovou hypotézu, že se r_{xy} neliší od nuly.

• **K vyhodnocení míry závislosti nelze přistupovat formálně**

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

r_{xy} – korelační koeficient mezi dvěma výběrovými soubory hodnot x a y
ρ – korelační koeficient mezi dvěma základními soubory hodnot x a y
Hodnota r_{xy} je odhadem hodnoty ρ

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		ν	p	
	0,05	0,01		0,05	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 $\alpha=0{,}05$ kritickou hodnotu:

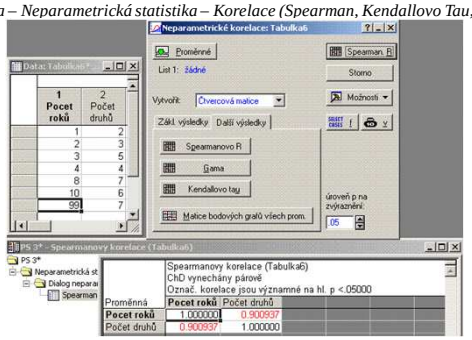
$r_{krit}=0{,}786$

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

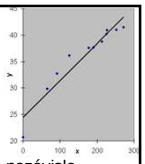
Koeficient pořadové korelace

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_j 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:

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

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:

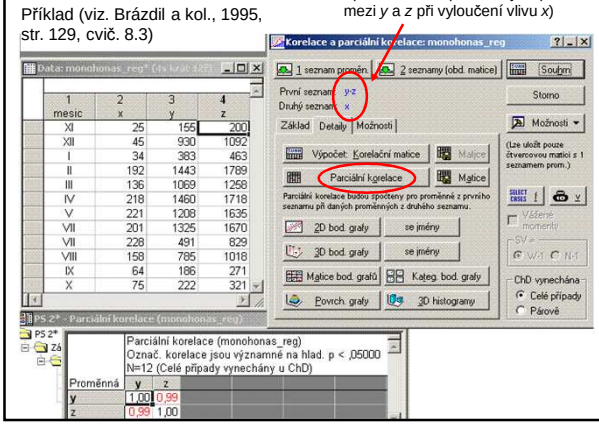
$$r_{xy\cdot z} = \frac{r_{xy} - r_{xz}\cdot r_{yz}}{\sqrt{(1-r_{xz}^2)\cdot(1-r_{yz})^2}}$$

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*)



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, …)