KGG/STG Statistika pro geografy

9. Korelační analýza

Mgr. David Fiedor
20. dubna 2015

Analýza závislostí

- v řadě geografických disciplín studujeme jevy, u kterých vyšetřujeme nikoliv pouze jednu vlastnost (statistický znak), nýbrž znaků několik
- tyto znaky mohou být navzájem závislé
- cílem této části statistiky je vyšetřit, do jaké míry spolu dva či více statistických znaků souvisí (hodnota jednoho znaku podmiňuje hodnotu znaku druhého)

Analýza závislostí - příklady použití

- zkoumání vztahu mezi teplotou vzduchu a nadmořskou výškou
- množství srážek a velikost odtoku
- počet dojíždějících a vzdálenost od centra dojižďky
- mnoho dalších vztahů . . .

Analýza závislostí

- zabývat se budeme určením síly závislosti (korelační počet) a také druhu závislosti (regresní počet)
- je potřeba rozlišovat zkoumání závislosti pro různé typy statistických znaků (nominální, ordinální, intervalové a poměrové)

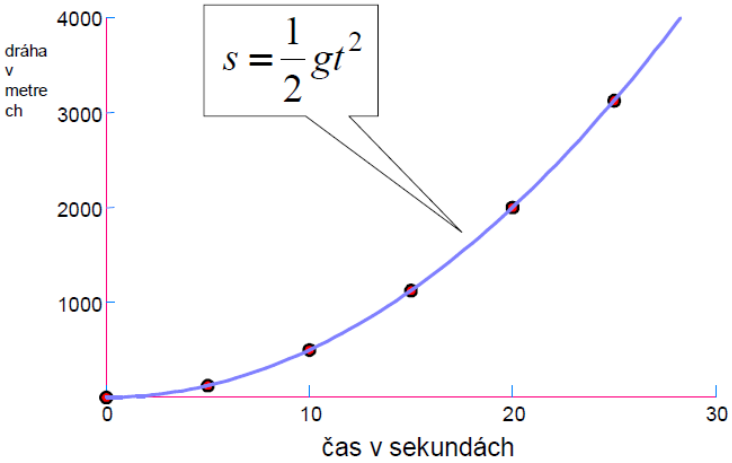
Vztahy mezi veličinami

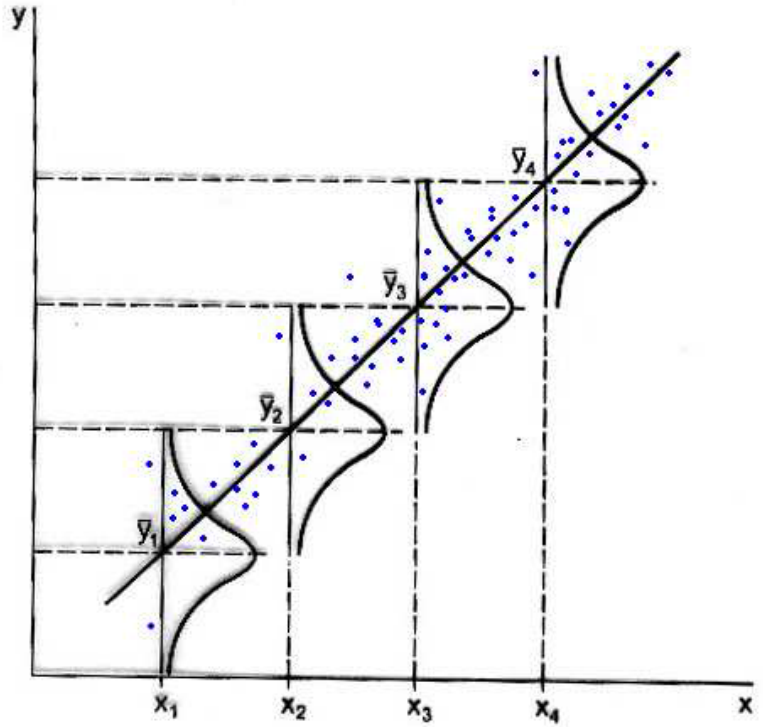
Vztahy jednostranné
Změna statistického znaku jednoho souboru náhodné veličiny (nezávisle proměnné) podmiňuje změnu statistického znaku souboru druhé náhodné veličiny (závisle proměnné).
Vztahy vzájemné
Nelze rozlišit mezi souborem závisle a nezávisle proměnné (např. vztah hodnot teplot vzduchu na dvou sousedních stanicích).

Druhy závislostí

1. **závislost funkční** - každé hodnotě znaku nezávisle proměnné náhodné veličiny x odpovídá vždy pouze jediná určitá hodnota závisle proměnné veličiny y
2. **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

Druhy závislostí





Kontingenční tabulka

Tabulka: Kontingenční tabulka

Úroveň spokojenosti	Nejvyšší dosažené vzdělání			$n_{i.}$
	ZŠ	SŠ	VŠ	
1	50	30	10	90
2	30	50	20	100
3	10	20	30	60
4	50	10	50	110
$n_{.j}$	140	110	110	360

- pozorované četnosti v jednotlivých „buňkách“ tabulky označuje obecně dvěma indexy n_{ij}
- marginální četnosti - ten index, přes který je sčítáno je označen tečkou

Čtyřpolní tabulka

- v případě alternativních znaků dostáváme čtyřpolní tabulku
  - alternativní znaky - mají pouze dvě možné varianty hodnot

Tabulka: Čtyřpolní tabulka

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

Měření závislosti kvalitativních znaků

- nulová hypotéza H_0: X, Y jsou nezávislé náhodné veličiny proti alternativě, že nejsou nezávislé
- test založen na porovnání zjištěných četností n_{ij} a tzv. teoretických četností $\frac{n_{i.}n_{.j}}{n}$, které by si měly být podobné
- testová statistika K (tzv. Pearsonova) má tvar, přičemž r, s jsou počty variant znaků

$$K = \sum_{i=1}^{r}\sum_{j=1}^{s} \frac{\left(n_{ij} - \frac{n_{i.}n_{.j}}{n}\right)^2}{\frac{n_{i.}n_{.j}}{n}}$$

Měření závislosti kvalitativních znaků

- měly by být splněny podmínky dobré aproximace, tj. teoretické četnosti $\frac{n_{i.} n_{.j}}{n}$ by měly aspoň v 80 % případů nabývají hodnoty větší nebo rovny pěti (a celkově neklesnout pod 2)
- při platnosti nulové hypotézy a splnění uvedené podmínky se testová statistika K řídí rozdělením $\chi^2((r-1)(s-1))$
- kritický obor $W = \langle \chi^2_{1-\alpha}((r-1)(s-1)), \infty)$

Cramérův koeficient

$$V = \sqrt{\frac{K}{n(m-1)}},$$

kde $m = \min\{r, s\}$

- koeficient nabývá hodnot mezi 0 a 1
- čím blíže k 1, tím je závislost těsnější; čím blíže je k 0, tím je závislost volnější

Význam hodnot Cramérova koeficientu

- mezi 0 až 0,1 $\Rightarrow$ zanedbatelná závislost
- mezi 0,1 až 0,3 $\Rightarrow$ slabá závislost
- mezi 0,3 až 0,7 $\Rightarrow$ střední závislost
- mezi 0,7 až 1 $\Rightarrow$ silná závislost

Příklad

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

Tabulka: Kontingenční tabulka

Televize	Sportování				$n_{i.}$
	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
$n_{.j}$	161	17	32	24	234

Řešení

- ověření podmínek dobré aproximace
- nulová hypotéza: neexistuje vztah mezi provozovaným sportem a sportem sledovaným v televizi
- výpočet testovacího kritéria K; $K = 273{,}3$
- zjistíme hodnotu příslušného kvantilu χ^2 rozdělení pro $(4-1)\cdot(4-1) = 9$ stupňů volnosti, tj. 16,9
- zamítáme proto nulovou hypotézu a tvrdíme, že vztah mezi sledovanými a provozovanými sporty je statisticky významný

Řešení

- pomocí výpočtu Cramérova koeficientu ještě můžeme kvantifikovat míru závislosti tohoto vztahu
- hodnotu určíme ze vztahu $V = \sqrt{\frac{K}{n(m-1)}}$, kde $m = \min\{r, s\}$
- $V = 0{,}62$, což považujeme za střední závislost

Řešení - postup v systému STATISTICA

- vytvoříme nový datový soubor o 3 proměnných (TV, Sportování, Četnost) a 16 případech (všechny varianty z tabulky)
- kontingenční tabulku vytvoříme takto: *Statistiky–Základní statistiky/tabulky–OK–Specif. Tabulky–List 1 TV, List 2 Sportování–OK* a zapneme proměnnou vah (četnost); na záložce *Možnosti* zaškrtneme *Procenta z počtu v řádku* a *Procenta z počtu ve sloupci–Výpočet*

Řešení - postup v systému STATISTICA

- ověříme podmínky dobré aproximace: *Statistiky–Základní statistiky/tabulky–Kontingenční tabulky–OK–Specif. Tabulky–List 1 TV, List 2 Sportování–OK*, zapneme proměnnou vah (četnost) *OK, Výpočet* a na záložce *Možnosti* zaškrtneme *Očekávané četnosti*
- hodnotu testové statistiky a Cramérův koeficient dostaneme tak, že na záložce *Možnosti* zaškrtneme *Pearsonův & M-V chí kvadrát a Cramérovo V*, na záložce *Detailní výsledky* vybereme *Detailní 2 rozm. tabulky*

Koeficient pořadové korelace

- Spearmanův koeficient r_s
- používá se k určení závislosti kvalitativních znaků ordinálního typu
- každé hodnotě x_i, 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

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

Vlastnosti Spearmanova koeficientu pořadové korelace

- platí vztah $-1 \leq r_s \leq 1$
- koeficient je rezistentní („odolný“) vůči odlehlým hodnotám
- používá se v situacích, kdy:
  - zkoumaná data mají aspoň ordinální charakter
  - nelze předpokládat, že vztah mezi veličinami X, Y je lineární
  - náhodný výběr nepochází z dvourozměrného normálního rozdělení

Testování pořadové nezávislosti ordinálních veličin

- nulová hypotéza: X, Y jsou pořadově nezávislé náhodné veličiny
- jako testovací kritérium slouží Spearmanův koeficient pořadové korelace r_s
- kritický obor
  $W = \langle -1, -r_s, 1 - \alpha/2(n) \rangle \cup \langle r_s, 1 - \alpha/2(n), 1 \rangle$,
  přitom $r_s, 1 - \alpha/2(n)$ je kritická hodnota, kterou najdeme v tabulkách

Příklad

Kvantifikujte vztah mezi dobou, po kterou jsou pole ponechána ladem a počtem rostlinných druhů (na m^2).

Tabulka: Zjištěná data

Počet roků	Počet druhů
1	2
2	3
3	5
4	4
8	7
10	6
¿ 10	7

Řešení

- přiřadíme hodnotám jejich pořadí
- určíme diference mezi odpovídajícími si pořadovými daty
- vypočítáme Spearmanův koeficient - $r_s = 0{,}902$
- v tabulkách najdeme pro $n = 7$ a pro $\alpha = 0{,}05$ kritickou hodnotu 0,786
- existuje statisticky významný vztah mezi dobou, po kterou jsou pole ponechána ladem a počtem rostlinných druhů, které se na nich vyskytují

Řešení - pomocí systému STATISTICA

- *Statistiky–Neparametrické statistiky–Korelace–OK*
- vybereme *Vytvořit detailní report–Proměnné X, Y–OK–Spearmanův koef. R.*
- jelikož počet dvojic není dostatečný, můžeme z této tabulky použít hodnotu testovacího kritéria, ale neměli bychom používat k interpretaci *p*-hodnotu
- v tabulkách najdeme kritickou hodnotu, vytvoříme kritický obor a můžeme učinit závěr o testu

Kovariance

Kovariancí dvou náhodných veličin X, Y rozumíme střední hodnotu součinu centrovaných veličin, tj. číslo

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

- obdoba rozptylu
- charakterizuje proměnlivost realizací náhodných veličin X, Y kolem jejich středních hodnot
- kladná hodnota kovariance znamená, že veličiny X, Y spolu buď rostou nebo klesají; záporná hodnota znamená, že jedna z veličin roste a druhá klesá
- omezenost spočívá v tom, že se jedná o míru absolutní - nelze ji použít k porovnání těsnosti vztahu dvou či více dvojic výběrových souborů

Kovariance

Návod na výpočet kovariance v systému STATISTICA:
Statistiky–Vícenásobná regrese–Proměnné Nezávislá X, Závislá Y–OK–OK–Residua/předpoklady/předpovědi–Popisné statistiky–Další statistiky–Kovariance

- ve výstupu obdržíme tabulku, kde na hlavní diagonále jsou rozptyly proměnných X, Y
- mimo hlavní diagonálu jsou kovariance

Koeficient korelace r_{xy}

- nejpoužívanější charakteristika k vyjádření závislostí mezi dvěma veličinami
- narozdíl od kovariance se jedná o relativní míru závislosti
- určíme ji jako podíl kovariance a součinu směrodatných odchylek s_x a s_y obou výběrů

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

Koeficient korelace

Hodnota korelačního koeficientu kolísá v intervalu od -1 do 1:

- $r_{xy} = 0$ nezávislost
- $r_{xy} \Rightarrow -1$ nepřímá závislost
- $r_{xy} = \Rightarrow 1$ přímá závislost

Testování významnosti korelačního koeficientu

- pro testování významnosti korelačního koeficientu je zapotřebí ověřit normalitu obou proměnných
- další podmínkou použití jsou dvojrozměrnost normálního rozdělení (každé hodnotě znaku veličiny x odpovídá soubor hodnot znaku y, který má normální rozdělení a naopak)
- poslední podmínkou je linearita vztahu hodnot x a y (regresní čára je přímka)
- nulová hypotéza tvrdí, že se korelační koeficient neliší od nuly (tedy žádná závislost)
- tvar testovacího kritéria nebudeme uvádět

Příklad

Jaká je závislost mezi pH půdy na výsypkách a počtem rostlinných 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

Řešení

- vypočítáme hodnotu korelačního koeficientu
- *Statistiky–Základní statistiky/tabulky–Korelační matice–OK–1 seznam proměnných–X, Y–OK* a na záložce *Možnosti* zrušíme volbu *Včetně průměrů a sm. odch.–Výpočet*
- na záložce *Možnosti* zvolíme *Zobrazit r, p-hodnoty a N–Výpočet*

Koeficient determinace

- často se koeficient korelace ve výpočtech doplňuje hodnotou koeficientu determinace r_{xy}^2
- tato hodnota kolísá v intervalu 0 až 1
- vynásoben stem udává v procentech tu část rozptylu závisle proměnné y, která je vysvětlena (podmíněna) změnami hodnot nezávisle proměnné x

Děkuji za pozornost...