Kapitola 8. A

Dvojrozmerný štatistický súbor, korelačná analýza

Dvojrozmerný štatistický súbor je súbor, ktorý obsahuje vždy dvojice hodnôt (meraní) týkajúcich sa sledovanej udalosti (napr. výška a hmotnosť človeka, prah počuteľnosti pravého a ľavého ucha, výška vodného toku na dvoch rôznych miestach a pod.). Ak sa zameriame iba na jedinú zložku meraní, môžeme údaje spracovávať ako jednorozmerný štatistický súbor a určiť všetky charakteristiky polohy a rozptylu. Navyše sa dajú vypočítať aj charakteristiky vlastné dvojrozmernej náhodnej veličine.

Kovariancia

Kovariancia sa ráta zo vzťahu:

$$\mathrm{cov}(X, Y) = \mathrm{E}[(X\text{-}\mathrm{E}(X))(Y\text{-}\mathrm{E}(Y))]$$

$$\mathrm{cov}(X, Y) = \mathrm{E}(XY)\text{-}\ \mathrm{E}(X)\mathrm{E}(Y)$$

Odhad kovariancie z nameraných dvojíc hodnôt vyrátame podľa vzorca:

$$s_{xy} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$$

$$s_{xy} = \frac{1}{n-1}(\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y})$$

Poznámka

Tento odhad je nevychýleným odhadom kovariancie, v praxi sa často používa aj odhad získaný metódou maximálnej vierohodnosti, ktorý sa líši od predošlého vzorca tým, že suma je delená počtom meraní n.

Kvantitatívnu mieru závislosti medzi dvomi zložkami určuje ***korelačný koeficient***

$$\rho_{xy} = \frac{\mathrm{cov}(X,Y)}{\sqrt{Var(X)Var(Y)}}$$

Korelačný koeficient je číslo z intervalu <-1, 1>. Ak sa korelačný koeficient rovná nule, zložky sú navzájom nezávislé, ak korelačný koeficient je v absolútnej hodnote blízky jednej, náhodne veličiny sa navzájom lineárne ovplyvňujú, ak sa korelačný koeficient rovná jednej alebo mínus jednej, je táto závislosť lineárna. V praxi korelačný koeficient iba odhadujeme z nameraných hodnôt. Odhad korelačného koeficientu (***výberový korelačný koeficient)*** rátame podľa vzorca

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

Vzorový príklad 1
V dvoch vodočetných staniciach na rieke sa sledovali prietoky vody (x_i, y_i) . Výsledky sú zaznamenané v tabuľke. Vypočítajte korelačný koeficient, určite mieru lineárnej závislosti.

x_i	23,1	12,8	17,8	21,3	18,5	93,5
y_i	25,9	15,1	20,4	23,5	21	105,9

Riešenie
Vypočítame výberový korelačný koeficient podľa predchádzajúceho vzorca, r_{xy}=0,999, to znamená, že medzi nameranými údajmi je silná lineárna závislosť
Niekedy je vhodnejšie na určenie miery závislosti použiť ***neparametrický (Spearmanov) korelačný koeficient***, napríklad vtedy, ak je porušený predpoklad normality nameraných údajov alebo máme k dispozícii len poradia. Postup pri výpočte tohoto koeficientu je nasledujúci: usporiadame vzostupne namerané hodnoty x_i a hodnoty y_i, nech R_i je poradie hodnoty x_i v usporiadaní a Q_i je poradie hodnoty y_i v usporiadanom poradí, pre rovnaké namerané hodnoty určíme poradie ako aritmetický priemer poradí rovnakých hodnôt. Spearmanov korelačný koeficient vyrátame podľa vzorca

$$r_s = 1 - \frac{6}{n(n^2-1)} \sum_{i=1}^{n} (R_i - Q_i)^2$$

Vzorový príklad 2
Vypočítajte Spearmanov korelačný koeficient pre hodnoty zo vzorového príkladu 1.

Riešenie
Usporiadame vzostupne a určíme poradia pre prietoky pre každú vodočetnú stanicu zvlášť.

x_i	23,1	12,8	17,8	21,3	18,5	93,5
R_i	5	1	2	4	3	6
y_i	25,9	15,1	20,4	23,5	21	105,9
Q_i	5	1	2	4	3	6

Keďže všetky poradia sa zhodujú, r_s =1.

Test nulovosti korelačného koeficienta
Ak máme normálne rozdelený dvojrozmerný štatistický súbor, ktorého korelačný koeficient sa rovná nule, zložky súboru sú navzájom nezávislé. Predpoklad nezávislosti je v praxi dôležitý, preto má zmysel testovať či sa výberový (odhadnutý z nameraných údajov) korelačný koeficient rovná nule. Testujme hypotézu

H_0: $\rho=0$ oproti alternatíve H_1: $\rho\neq 0$

na príslušnej hladine významnosti α. Testovacia štatistika $T=\frac{r}{\sqrt{1-r^2}}\sqrt{n-2}$ má ***Studentovo t-rozdelenie*** s (n-2) stupňami voľnosti. Vypočítanú štatistiku T porovnáme s tabuľkovou hodnotou príslušného kvantilu t-rozdelenia (pozri aj kapitolu 7.A, Testy hypotéz). H_0 nezamietame na príslušnej hladine významnosti α, ak

$$-t_{n-2}(1-\frac{\alpha}{2}) \le T \le t_{n-2}(1-\frac{\alpha}{2})$$

Vzorový príklad 3
Testujte na hladine významnosti α=0,1 , že korelačný koeficient zo vzorového príkladu 1. je rovný nule.

Riešenie

Vypočítame štatistiku T a vypočítanú hodnotu porovnáme s $t_4(0,95)$, keďže T =44,68 > $t_4(0,95)$= 2,13 zamietame nulovú hypotézu na príslušnej hladine významnosti.

Nové pojmy a definície kapitoly 8. A

- *dvojrozmerný štatistický súbor*
- *kovariancia*
- *korelácia, korelačný koeficient*
- *Spearmanov korelačný koeficient*
- *test nulovosti korelačného koeficientu*

Cvičenia

V príkladoch 1. až 4. určite základné charakteristiky dvojrozmerného štatistického súboru, vypočítajte korelačný a Spearmanov korelačný koeficient, na hladine významnosti α=0,05 testujte hypotézu, že výberový korelačný koeficient sa rovná nule.

Príklad 1
V n=20 vzorkách rudy sa sledoval obsah striebra X a olova Y. Výsledky sú uvedené v tabuľke

X	5	6	11	13	20	21	22	22	25	26	27	31	34	37	37	41	42	44	49	50
Y	6	7	6	8	13	20	21	23	15	20	22	17	30	31	33	14	39	31	39	41

Z vypočítaných štatistík dvojrozmerného súboru zistite, či obsah striebra a olova v rude sú navzájom závislé.

Príklad 2
Osem vzoriek chemickej látky sme analyzovali titračne a polygraficky. Výsledky sú uvedené v tabuľke

X	18,6	27,6	27,5	25	24,5	26,8	29,7	26,5
Y	18,58	27,37	27,27	24,64	24,1	26,33	29,33	26,63

Z vypočítaných štatistík dvojrozmerného súboru zistite, či výsledky analýz spolu súvisia.

Príklad 3
Bolo potrebné zistiť, či špeciálny prípravok pridávaný do ocele zvyšuje jej pevnosť. Výsledky meraní pred pridaním prípravku a po pridaní prípravku sú uvedené v tabuľke

X	5,6	5,9	5,9	5,7	5,8	5,7	6	5,5	5,7	5,5
Y	6,2	5,7	5,6	6	6,3	5,8	5,7	6	6	5,8

Z vypočítaných štatistík dvojrozmerného súboru zistite, či výsledky meraní spolu súvisia.

Príklad 4
V šestnástich stavbárskych čatách sa počas roka sledoval počet menších úrazov (X) v závislosti od počtov pracovníkov (Y). Výsledky sú uvedené v tabuľke

X	18	19	20	21	22	22	25	26	26	26	27	28	29	30	31	33
Y	26	23	29	27	31	25	22	32	32	33	38	29	36	37	41	42

Z vypočítaných štatistík dvojrozmerného súboru zistite, či počty pracovníkov a počty menších úrazov spolu súvisia.

Riešenia

Príklad 1
$r_{xy}=0,876$, $r_s=0,8522$, $T=7,728$, $t_{18}(0,975)=2,1$, $T> t_{18}(0,975)$, zamietam hypotézu H_0 o nulovosti korelačného koeficienta

Príklad 2
$r_{xy}=0,998$, $r_s=0,968$, $T=41,48$, $t_6(0,975)=2,44$, $T> t_6(0,975)$, zamietam hypotézu H_0 o nulovosti korelačného koeficienta

Príklad 3
$r_{xy}=-0,437$, $r_s=-0,46$, $T=-1,375$, $t_8(0,975)=2,3$, $|T|< t_8(0,975)$, nezamietam hypotézu H_0 o nulovosti korelačného koeficienta

Príklad 4
$r_{xy}=0,815$, $r_s=0,804$, $T=5,272$, $t_{14}(0,975)=2,14$, $T> t_{14}(0,975)$, zamietam hypotézu H_0 o nulovosti korelačného koeficienta