Cvičení 11 – Korelační analýza

Výběrový Spearmanův koeficient pořadové korelace: $r_S = 1 - \frac{6}{n(n^2-1)} \sum_{i=1}^{n} (R_i - Q_i)^2$,

kde n je rozsah dvourozměrného náhodného výběru (X_1, Y_1), ..., (X_n, Y_n), R_i pořadí náhodné veličiny X_i a Q_i pořadí náhodné veličiny Y_i, i = 1, ..., n.

Testování nezávislosti ordinálních veličin

Na hladině významnosti α testujeme H_0: X, Y jsou pořadově nezávislé náhodné veličiny proti H_1: X, Y jsou pořadově závislé náhodné veličiny

Jako testová statistika slouží Spearmanův koeficient pořadové korelace r_S.

H_0 zamítáme na hladině významnosti α, když $|r_S| \geq r_{S,1-\alpha/2}(n)$, kde $r_{S,1-\alpha/2}(n)$ je kritická hodnota, kterou pro α = 0,05 nebo 0,01 a n ≤ 30 najdeme v tabulkách.

Asymptotická varianta testu

Pro n > 20 lze použít testovou statistiku $T_0 = \frac{r_S\sqrt{n-2}}{\sqrt{1-r_S^2}} \approx t(n-2)$ za platnosti H_0.

Kritický obor $W = \left(-\infty, -t_{1-\alpha/2}(n-2)\right\rangle \cup \left\langle t_{1-\alpha/2}(n-2), \infty\right)$

Výběrový Pearsonův koeficient korelace: $R_{12} = \frac{S_{12}}{S_1 S_2}$ pro $S_1 S_2 > 0, = 0$ jinak

Testování nezávislosti intervalových a poměrových veličin

Důležitý předpoklad: daný náhodný výběr (X_1, Y_1), ..., (X_n, Y_n) pochází z dvourozměrného normálního rozložení s koeficientem korelace ρ. Pak jsou pojmy „stochastická nezávislost“ a „nekorelovanost“ ekvivalentní.

Na hladině významnosti α testujeme H_0: X, Y jsou stochasticky nezávislé náhodné veličiny (tj. ρ = 0) proti H_1: X, Y jsou stochasticky závislé náhodné veličiny (tj. ρ ≠ 0).

Testová statistika $T_0 = \frac{R_{12}\sqrt{n-2}}{\sqrt{1-R_{12}^2}} \sim t(n-2)$ za platnosti H_0.

Kritický obor: $W = \left(-\infty, -t_{1-\alpha/2}(n-2)\right\rangle \cup \left\langle t_{1-\alpha/2}(n-2), \infty\right)$

Porovnání koeficientu korelace s danou konstantou

Testujeme H_0: ρ = c proti H_1: ρ ≠ c.

Pro n ≥ 10: testová statistika $U = \left(Z - \frac{1}{2}\ln\frac{1+c}{1-c} - \frac{c}{2(n-1)} \right)\sqrt{n-3} \approx N(0,1)$ za platnosti H_0,

přičemž $Z = \frac{1}{2}\ln\frac{1+R_{12}}{1-R_{12}}$ je tzv. Fisherova Z-transformace.

Kritický obor: $W = \left(-\infty, -u_{1-\alpha/2}\right\rangle \cup \left\langle u_{1-\alpha/2}, \infty\right)$.

Porovnání dvou korelačních koeficientů

Máme dva nezávislé náhodné výběry o rozsazích n a n^* z dvourozměrných normálních rozložení s korelačními koeficienty ρ a ρ^*. Testujeme H_0: $\rho = \rho^*$ proti H_1: $\rho \neq \rho^*$.

Označme R_{12} výběrový korelační koeficient 1. výběru a R_{12}^* výběrový korelační koeficient 2. výběru. Položme $Z = \frac{1}{2}\ln\frac{1+R_{12}}{1-R_{12}}$ a $Z^* = \frac{1}{2}\ln\frac{1+R_{12}^*}{1-R_{12}^*}$.

Platí-li H_0, pak testová statistika $U = \frac{Z - Z^*}{\sqrt{\frac{1}{n-3}+\frac{1}{n^*-3}}}$ má asymptoticky rozložení N(0,1).

Kritický obor: $W = \left(-\infty, -u_{1-\alpha/2}\right\rangle \cup \left\langle u_{1-\alpha/2}, \infty\right)$.

Interval spolehlivosti pro koeficient korelace

$(d, h) = \operatorname{tgh}\left(Z \mp \frac{u_{1-\alpha/2}}{\sqrt{n-3}}\right)$, přičemž $\operatorname{tgh} x = \frac{e^x - e^{-x}}{e^x + e^{-x}}$.

Příklad 1.: (viz př. 11.5.2 ze skript) 12 různých softwarových firem nabízí speciální programové vybavení pro vedení účetnictví. Jednotlivé programy byly posouzeny odbornou komisí složenou z počítačových odborníků a komisí složenou z profesionálních účetních. Úkolem bylo doporučit vhodný program na základě stanovení pořadí jednotlivých programů. Výsledky posouzení:

Produkt firmy číslo	1	2	3	4	5	6	7	8	9	10	11	12
Pořadí dle odborníků	6	7	1	8	4	2,5	9	12	10	2,5	5	11
Pořadí dle účetních	4	5	2	10	6	1	7	11	8	3	12	9

Vypočtěte Spearmanův koeficient pořadové korelace a na hladině významnosti 0,05 testujte hypotézu, že hodnocení obou komisí jsou nezávislá.

Příklad 2.: Pět mužů, kteří bydlí v jednom panelovém domě, se rozhodlo zjistit a zapsat svou hmotnost [kg] a výšku [cm]. Zapsané hodnoty jsou: (76, 170), (86, 177), (73, 169), (84, 174), (79, 175). Najděte realizaci výběrového koeficientu korelace a na hladině významnosti 0,05 testujte hypotézu, že hmotnost a výška jsou nezávislé veličiny proti oboustranné alternativě. Pro úsporu času máte uvedeny tyto číselné realizace: $s_1^2 = 29{,}3,\ s_2^2 = 11{,}5,\ s_{12} = 16{,}5$.

Příklad 3.: V pedagogicko – psychologické poradně bylo vyšetřeno 12 hochů a 15 dívek. Mimo jiné byla zjišťována jejich verbální složka IQ a performační složka IQ. Předpokládáme, že v obou skupinách dětí se zjištěná data řídí dvourozměrným normálním rozložením, v prvním případě s koeficientem korelace ρ, ve druhém případě s koeficientem korelace ρ^*. Ve skupině hochů nabyl výběrový koeficient korelace mezi verbálním a performačním IQ hodnoty 0,6033, u dívek pak 0,5833.

a) Sestrojte 95% asymptotický interval spolehlivosti pro ρ.

b) Na hladině významnosti 0,05 testujte H_0: $\rho = \rho^*$ proti oboustranné alternativě H_0: $\rho \neq \rho^*$.