Test nezávislosti kardinálních veličin

Všichni úspěšní studenti VŠCHT musí během svého studia absolvovat povinný předmět Matematika I (respektive Matematika A). V souboru BODY.csv jsou uloženy celkové výsledky paralelkových testů jednotlivých studentů všech skupin od téhož cvičího za zimu roku 2013 a jejich docházka v procentech. Nás teď zajímá, zda výsledky paralelkových testů závisí na docházce.

A) Otestujte, zda počet bodů z paralelkových testů závisí na docházce.

B) Otestujte, že dobrá docházka má pozitivní vliv na výsledky studentů v testech.

Řešení - teorie

Náhodná veličina X „celkový počet bodů“ a náhodná veličina Y „docházka v %“ jsou tzv. kardinální veličiny.

Kardinálními náhodnými veličinami nazýváme číselné náhodné veličiny, u nichž má jejich číselná hodnota reálný význam a má tedy smysl je navzájem porovnávat. Takovými veličinami jsou například výška obyvatel, IQ populace, počet aut, které projedou křižovatkou za hodinu atp.

Veličiny, jimž sice můžeme přiřadit číselnou hodnotu, ale navzájem nedává smysl je porovnávat, nazýváme **nominální veličiny**. Například, budeme-li zjišťovat barvu vlasů u populace, můžeme pro jednodušší manipulaci označit jednotlivé barvy čísly. Například blond vlasy číslem 1, černé číslem 2 atd. Tyto veličiny jsou potom číselné, ale nemá v jejich případě smysl tvrdit, že $2 > 1$, respektive, že černá je „víc“ než blond. Další - velmi přirozenou - nominální veličinou je například pohlaví. Závislostí nominálních veličin (respektive závislostí kardinální a ordinální veličiny) se budeme zabývat v úloze Testů nezávislosti v kontingenčních tabulkách.

Pro kardinální náhodné veličiny X a Y s konečným nenulovým rozptylem definujeme Pearsonův korelační koeficient předpisem

$$\rho(X,Y) = \frac{\mathrm{cov}(X,Y)}{\sqrt{\mathrm{var}X}\sqrt{\mathrm{var}Y}} \tag{1}$$

Tento korelační koeficient udává, vágně řečeno, míru lineární závislosti náhodných veličin X a Y. Čím bližší je ± 1, tím těsnější je lineární závislost. Je-li $\rho(X,Y) = 0$, nazýváme veličiny X a Y **nekorelované**. Dá se o nich předpokládat, že nejsou lineárně závislé, ale v žádném případě z nekorelovanosti nemůžeme usuzovat na obecnou nezávislost těchto veličin.

Uvažujme náhodné výběry $X_1, \ldots, X_n$ a $Y_1, \ldots, Y_n$. Pearsonův korelační test (test lineární ne/závislosti) veličin X a Y je založen na výběrovém korelačním koeficientu

$$r(X,Y) = \frac{S_{XY}}{S_X S_Y},$$

kde S_{XY} je tzv. **výběrová kovariance**, neboli bodový odhad kovariance, definovaný vztahem

$$S_{XY} \equiv \widehat{\mathrm{cov}}(X,Y) = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y}) = \frac{1}{n-1}\sum_{i=1}^{n}\left(X_iY_i - n\bar{X}\bar{Y}\right).$$

Pro úlohu **A)** budeme testovat nulovost korelačního koeficientu oproti oboustranné alternativě. Nulovou a alternativní hypotézu proto formulujeme následovně:

$$\begin{aligned} H_0 &: \quad \rho(X,Y) = 0, \\ H_1 &: \quad \rho(X,Y) \neq 0. \end{aligned}$$

Pokud by veličiny X a Y pocházely z dvourozměrného normálního rozdělení, má testová statistika

$$R = \frac{r}{\sqrt{1-r^2}}\sqrt{n-2}$$

Studentovo $t-$rozdělení o $n-2$ stupních volnosti. V případě, že data nejsou normálně rozdělená, bude se jednat o asymptotický test.

V úloze **B)** se nám jedná vlastně o potvrzení hypotézy o kladné korelovanosti jednotlivých veličin. Kladná korelovanost totiž znamená, že se zvyšujícími hodnotami jedné veličiny má i druhá veličina tendenci růst. Jak již víme, chceme-li nějakou hypotézu potvrdit, je vhodné ji dát do alternativní hypotézy, neboť zamítáme-li nulovou hypotézu, je to vždy ve prospěch alternativy. To znamená, že na dané hladině α statisticky významně tvrdíme, že alternativa platí. Zatímco dáme-li naši hypotézu testovat jako nulovou, nikdy se nedočkáme potvrzení. V této úloze tedy formulujeme nulovou a alternativní hypotézu následovně

$$\begin{aligned} H_0 &: \quad \rho(X,Y) = 0, \\ H_1 &: \quad \rho(X,Y) > 0. \end{aligned}$$

Testová statistika i její rozdělení za platnosti H_0 je stejné jako při oboustranném testu, stejně tak i případná asymptotika.

Poznámky

(*i*) Nezamítnutí nulové hypotézy, jak už víme, neznamená její přijetí. Dokonce ani skutečná platnost hypotézy H_0 by neznamenala nezávislost veličin X a Y, neboť Pearsonův korelační koeficient popisuje pouze míru lineární závislosti, nikoliv obecné!

(*ii*) Kromě Pearsonova korelačního koeficientu existují ještě další typy korelačních koeficientů. Za všechny zmiňme Spearmanův a Kendallův koeficient. Místo na hodnotách veličin X a Y jsou tyto koeficienty založeny na pořadí jejich hodnot ve výběru. Spearmanův korelační koeficient neindikuje přímou či nepřímou lineární závislost, nýbrž pouze monotónní vztah mezi veličinami.