Poznámky k předmětu Aplikovaná statistika, 10. téma

10. Test nezávislosti náhodných veličin

Mohlo by vás např. zajímat, jestli jsou průměry známek z chemie a matematiky ze středoškolského studia studentů VŠCHT nezávislé, anebo jestli tyto průměry spolu nějak souvisí. Tuto úlohu a další podobné se naučíme řešit v tomto tématu.
Mějme dvourozměrný náhodný výběr $(X_1, Y_1), \ldots, (X_n, Y_n)$ z rozdělení náhodného vektoru (X, Y). Test nezávislosti náhodných veličin X, Y je za předpokladu normality v podstatě test nulovosti korelačního koeficientu, jak uvidíme níže. Nejprve si připomeneme definici **korelačního koeficientu** $\rho(X, Y)$ (z Tématu 4)

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

kde $\mathrm{cov}(X, Y) = \mathbb{E}(X - \mathbb{E}X)(Y - \mathbb{E}Y)$, a jeho nejdůležitější vlastnosti:

- $-1 \leq \rho(X, Y) \leq 1$,
- $\rho(X, Y) = 0$ znamená, že náhodné veličiny X, Y jsou nekorelované, tj. nejsou statisticky lineárně závislé.

Pokud ovšem předpokládáme, že (X, Y) pochází z **dvourozměrného normálního rozdělení**, pak platí

$$\rho(X, Y) = 0 \iff X, Y \text{ jsou nezávislé.}$$

Této vlastnosti se využívá při konstrukci testové statistiky založené na korelačním koeficientu. Nejprve bodově odhadneme $\rho(X, Y)$:

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

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

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

Odhad $r(X, Y)$ korelačního koeficientu se nazývá **výběrový korelační koeficient** a platí pro něj

$$-1 \leq r(X, Y) \leq 1.$$

TEST NEZÁVISLOSTI
Nechť náhodný výběr $(X_1, Y_1), \ldots, (X_n, Y_n)$ pochází z dvourozměrného normálního rozdělení. Hypotézu

$$H_0 : X, Y \text{ jsou nezávislé}$$

lze proto ekvivalentně formulovat ve tvaru

$$H_0 : \rho(X, Y) = 0.$$

Příslušná testová statistika

$$R = \frac{r(X,Y)}{\sqrt{1-r^2(X,Y)}}\sqrt{n-2}$$

má za platnosti nulové hypotézy t–rozdělení o $n-2$ stupních volnosti. Informace o kritickém oboru najdete v tabulkách.

Příklad 1. *U sta studentů VŠCHT učastnících se prvního zkouškového termínu z Matematiky I v zimním semestru akademického roku* 2015/2016 *byla zjištěna průměrná známka z matematiky a chemie z jejich středoškolského studia. Výsledky jsou zaznamenány v tabulce*

M	1	2	3	2,25	2,25	1,5	2,5	1	3	1,75	1,75	2	2,5
CH	1,25	2,5	2,5	2	1,5	1,5	1,5	1,25	2	2,5	1,75	2	2,75
M	2,25	1	2,5	1,5	1	1,5	2	2	2,5	1,25	1	1,5	1
CH	2,6	1,5	1,75	1,25	1	1,75	2	1	1,75	1,33	1	2	1
M	2	2,75	2,25	1,5	1,75	2,25	2,25	1,5	2	1	2	2,75	2,5
CH	1	2	2,5	2	1,5	2,25	1,75	1	2	1	1,25	1,33	1,5
M	2,25	1	2	2,25	2,25	2,25	1	2	1,75	1,75	1,75	1	1
CH	2	1	1,75	1,75	2,25	1,5	1	2,25	1,75	2,75	1	1,4	1
M	1	2,25	2,75	1,2	2,5	1,75	2	2	1,75	2	1	1	2,25
CH	1,5	1,5	1,5	1,4	1,5	2	1,25	1	2	1,25	1	1	1,33
M	1,25	1,75	3	1,75	2	1,5	1,75	2,5	1	1,6	2,25	2	1
CH	1,25	1,75	3	2,75	1,5	1	1	1,75	1	1	2	1,5	1
M	1,5	1,25	1,5	1,25	1	1,75	2,75	1,6	2,25	1	3,25	1	1,25
CH	1,25	1,75	1	1,75	1	1	1,5	1,2	1,5	1,75	3	1	1,75
M	2,5	2	2,25	2	1,5	2,75	1,5	1	2,75				
CH	1,4	2	2	1,25	1,25	2,25	1,25	1	2,2				

Zjistěte, zdali spolu průměry známek z matematiky a chemie statisticky významně souvisí. Volte hladinu významnosti 5%.

Poznámka 1 (O interpretaci získaných výsledků)**.** *V Příkladě 1 by se v případě prokázané závislosti průměrných známek z matematiky a chemie dala uvažovat jak závislost známky z matematiky na známce z chemie, tak závislost známky z chemie na známce matematice. Každou z proměnných totiž můžete volit jako nezávisle proměnnou a úloha dává smysl (tato úloha je „symetrická“). Jinak je tomu ovšem ve Cvičení (i). Tam by se případné prokázání závislosti výnosu česneku na počtu slunečných dnů muselo interpretovat jediným způsobem. Výnos česneku by byl závislý na počtu slunečných dnů, tj. počet slunečných dnů je nezávisle proměnná a výnos česneku závisle proměnná (prohození role závisle a nezávisle proměnné by v tomto případě nedávalo smysl!).*

CVIČENÍ K TOMUTO TÉMATU

(i) V průběhu osmi let se zaznamenávaly výnosy česneku (v kg/m^2) a průměrné denní doby slunečního svitu (v hodinách) ve vegetačním období česneku

rok	1.	2.	3.	4.	5.	6.	7.	8.
svit	4,1	3,9	3,5	3,8	4,2	4,1	4,1	3,9
výnos	1,2	1,1	0,5	0,9	1,0	1,3	1,0	1,0

.

Je možné na základě těchto dat považovat za prokázané ($\alpha = 10$ %), že se zvětšující se dobou slunečního svitu se zvětšuje výnos česneku? Předpokládáme dvourozměrné normální rozdělení doby svitu a výnosu česneku. Nadto určete přibližně dosaženou hladinu testu.

(dcv) Zemědělci oseli 6 polí ošetřených hnojivem A a 5 polí ošetřených hnojivem B. Po sklizni byly zjištěny průměrné výnosy z každého pole (v t/ha)

hnojivo A	62	54	55	60	53	58
hnojivo B	52	56	49	50	51	

.

Zjistěte na hladině významnosti $\alpha = 5$ %, zdali jsou hnojiva stejně efektivní. Dále určete přibližně p–hodnotu testu.