Cvičenie 12 – Závislosť / Nezávislosť

Zo všetkých kútov pozbierané a doteraz nedostatočne spomínané múdrosti o vzťahu dvoch náhodných veličín

Riešené príklady

a) Pearsonov test nezávislosti

Príklad 1

Na parkovisku vyčlenenom pre návštevníkov vianočných trhov agentúrni pracovníci kládli vodičom otázku, ktorý z typických gastro-produktov si kúpia a ochutnajú ako prvý – pečené gaštany (PG), lokše (LK) alebo cigánsku pečienku (CP). Zároveň (bez vedomia opýtaných) zaznamenávali región, odkiaľ pochádza ich auto (rozhoduje historické sídlo centrály celého koncernu, nie montážna linka).
Na vzorke 1528 respondentov boli v ankete zistené nasledujúce počty v kombináciách preferencií (*kontingenčná tabuľka*):

jedlo \ auto	Ázia	FR + ITA	Nemecko	Ostatné	spolu
PG	50	103	116	7	276
LK	106	157	310	54	627
CP	115	142	316	52	625
spolu	271	402	742	113	n=1528

Na hladine významnosti 0.05 testujte hypotézu H_0, že preferencia v oblasti gastrošpecialít nesúvisí s preferenciou v oblasti výberu značky auta, tj. že obe preferencie sú štatisticky nezávislé.

Riešenie:
Zostavme si tabuľku relatívnych početností (všetky čísla v tabuľke podelíme hodnotou n=1528). Na samotné riešenie príkladu to nie je nevyhnutné, ale pomôže nám to pochopiť súvislosti.

jedlo \ auto	Ázia	FR + ITA	Nemecko	Ostatné	spolu
PG	50/1528	103/1528	116/1528	7/1528	276/1528
LK	106/1528	157/1528	310/1528	54/1528	627/1528
CP	115/1528	142/1528	316/1528	52/1528	625/1528
spolu	271/1528	402/1528	742/1528	113/1528	1

Ako vyzará dokonalá nezávislosť dvoch veličín? Na úrovni relatívnych početností je to jednoduché, musí platiť

$$\text{RelPoč (jedlo I + auto J)} = \text{RelPoč(jedlo I)} * \text{RelPoč(auto J)}$$

Ak označíme hodnoty v jednotlivých políčkach n_{ij}, súčty po riadkoch n_i a po stĺpcoch n_j, môžeme vzorec písať takto:

$$n_{ij} / n = (n_i / n) * (n_j / n)$$

Na úrovni absolútnych početností musíme uvedený vzorec vynásobiť hodnotou n.

$$\text{Počet (jedlo I + auto J)} = \text{RelPoč(jedlo I)} * \text{RelPoč(auto J)} * n = \text{Poč(jedlo I)} * \text{Poč(auto J)} / n$$

$$n_{ij} = (n_i / n) * (n_j / n) * n = n_i * n_j / n$$

Podľa posledného vzorca zostavíme tabuľku ideálnych početností, aké by zodpovedali dokonalej nezávislosti veličín. Teda miesto hodnôt n_{ij} zapíšeme hodnoty $n_i * n_j / n$:

Očakávané hodnoty pri dokonalej nezávislosti veličín:

271*276/1528 = 48.95
271*627/1528 = 111.20
.....

jedlo \ auto	Ázia	FR + ITA	Nemecko	Ostatné	spolu
PG	48.95	72.61	134.03	20.41	276
LK	111.2	164.96	304.47	46.37	627
CP	110.85	164.43	303.5	46.22	625
spolu	271	402	742	113	n=1528

Suroviny sú nachystané, ideme variť. Nebude to nič zložité, len to nahádžeme do vzorca, aby sme získali hodnotu G, ktorú primeriame .... ale nepredbiehajme.

$$G = \Sigma_i \Sigma_j \left((n_{ij} - n_i n_j / n)^2 / (n_i n_j / n) \right) =$$

$$= (50-48.95)^2/48.95 + (106-111.2)^2/111.2 + ... = 30.41$$

Ak niekomu ten vzorec pripomína chí-kvadrát test dobrej zhody, môže sám seba pochváliť za dobrý postreh. Nie je to náhoda, princíp je podobný, aj keď tu máme údaje "dvojrozmerné".

V tabuľke je 3×4 údajov, ak každé z čísel odľahčíme o 1, dostaneme 2*3=6 .

To znamená, že získanú hodnotu G budeme primeriavať k hodnotám rozdelenia chí-kvadrát so stupňom voľnosti 6 (6-ty riadok):

	$\chi^2(v)$									
v	0,005	0,01	0,025	0,05	0,1	0,9	0,95	0,975	0,99	0,995
6	0,68	0,87	1,24	1,64	2,20	10,64	12,59	14,45	16,81	18,55

Pri požadovanej hladine 5% by sa mal výsledok G zmestiť do bezpečnej oblasti od 0 do 0.95.
Ak však vidíme, pri 0.95 svieti 12.59.

$$12.59 < 30.41$$

Vypočítané G je výrazne väčšie, nachádza sa dosť mimo bezpečia, teda je hlboko v kritickej oblasti. Hypotézu H_0 preto zamietame na požadovanej hladine 0.05, ale ako vidno, zamietnuť by sa to dalo aj na postatne kvalitnejších hladinách. Vzťah medzi preferenciami šoférov v dvoch naoko nesúvisiacich oblastiach sa v skúmanom prípade nedá nazvať nezávislosťou.

Poznámka: Na to, aby bol test nezávislosti spoľahlivý, musia mať všetky údaje v tabuľke hodnotu viac ako 2 (bolo splnené) a aspoň 80 percent údajov musí byť väčších ako 5.

Iný príklad:

https://portal.matematickabiologie.cz/index.php?pg=aplikovana-analyza-klinickych-a-biologickych-dat--analyza-a-management-dat-pro-zdravotnicke-obory--testovani-hypotez-o-kvalitativnich-promennych--analyza-kontingencnich-tabulek--testovani-nezavislosti-pearsonuv-chi-kvadrat-test

b) Gymnastika s varianciou a kovarianciou

Niekoľko užitočných vzorcov, pomocou ktorých sa dá cvičiť s rozptylom (disperziou) a kovarianciou. Niektoré už poznáme, ostatné sa dajú ľahko odvodiť.

$cov(X,Y) = E(XY) - E(X)E(Y)$	$cov(aX+b, Y) = a\ cov(X,Y)$
$cov(X,X) = D(X)$	$cov(X+Y, Z) = cov(X,Z) + cov(Y,Z)$
$cov(X,Y) = cov\ (Y,X)$	$D(X+Y)=D(X) + 2cov(X,Y) + D(Y)$

Príklad 2

Je daná dvojica náhodných veličín X, Y.
Vieme, že

$E(X)=1.7 \quad D(X)=1.21$
$E(Y)=0.2 \quad D(Y)=0.96$
$cov(X,Y) = 1$

Vypočítajte, čomu sa rovná

$D(X+Y)$
$D(2X-Y)$
$cov(2X+Y, X-2Y)$

Riešenie:
Postupujeme podľa vzorcov.

$D(X+Y) = D(X) + 2cov(X,Y) + D(Y) = 1.21+2*1+0.964.17$

$D(2X-Y) = cov(2X-Y,2X-Y) = 4D(X) - 4cov(X,Y) + D(Y) = 4.84 - 4 + 0.96 = 1.8$

$cov(2X+Y, X-2Y) = 2cov(X,X-2Y) + cov(Y,X-2Y) = 2D(X) - 3cov(X,Y) - 2D(Y) =$
$= 2.42 - 3*1 - 1.92 = -2.5$

Neriešené príklady

1. 200 študentov sa (ešte v nepandemických časoch) pýtali, akým spôsobom sa prevažne dopravujú do školy. Zistené údaje spárovali s ich známkami z predmetu Matematická štatistika. Výsledky sú uvedené v tabuľke:

cesta\známka	FX	E	D	C	B	A
pešo	6	5	7	7	5	3
verejná doprava	11	15	10	8	7	6
na 1-2 kolesách	6	6	7	11	12	6
na aute	18	16	11	7	7	3

Na hladine 0.05 (a iných) testujte hypotézu o nezávislosti oboch skúmaných veličín.

2. Riešte predošlý príklad v zjednodušenej podobe – rozlišujme len to, či prichádzajú vlastným dopravným prostriedkom alebo nie, a či skúšku spravili alebo nie.

3. Je daná dvojica náhodných veličín X, Y.
Vieme, že

$$E(X)=3 \qquad D(X)=4$$
$$E(Y)=2 \qquad D(Y)=9$$
$$\mathrm{cov}(X,Y) = 3$$

Vypočítajte, čomu sa rovná

$D(2X+3Y)$
$D(7Y-2X)$
$\mathrm{cov}(2X-3Y, 3X+2Y)$

4. Je daná trojica náhodných veličín X, Y, Z. Vieme, že

$$E(X)=3 \qquad D(X)=4$$
$$E(Y)=2 \qquad D(Y)=9$$
$$E(Z)=4 \qquad D(Z)=12$$
$$\mathrm{cov}(X,Y) = 3$$
$$\mathrm{cov}(X,Z) = 5$$
$$\mathrm{cov}(Y,Z) = 8$$

Vypočítajte, čomu sa rovná

$D(X+Y+Z)$
$D(2X-3Y-Z)$
$\mathrm{cov}(2X-3Y, Y+3Z)$
$\mathrm{cov}(X+2Z, 3X+Y-2Z)$
$\mathrm{cov}(3X-2Y-Z, -X+3Y+4Z)$