11 Další typy statistických hypotéz

Teorie: Chí kvadrát test dobré shody

H_0 pravděpodobnosti výsledků jsou $p_1, p_2, \ldots, p_k$ (data mají rozdělení daná těmito ppstmi)
Porovnáváme očekávané četnosti $n_i^O > 5$ určené pomocí ppsti p_i a naměřené četnosti n_i
Postup je následující:

- Rozdělíme obor hodnot náhodné veličiny X na k nepřekrývajících se tříd.
- Zjistíme, kolik hodnot realizovaného náhodného výběru se nachází v jednotlivých třídách. Počty prvků v jednotlivých třídách označíme n_i
- Pokud je $m > 0$, tj. některé parametry rozdělení jsou neznámé, tak je odhadneme (k dispozici máme tedy po tomto kroku pravděpodobnosti p_i dané tímto rozdělením).
- Pro každou třídu spočteme očekávaný počet hodnot v této třídě, ozn. n_i^O. Platí $n_i^O = np_i$.
- V případě, že je v některé třídě počet očekávaných hodnot menší než 5, pak musíme tuto třídu sdružit s jinou.
- Testovací statistika má tvar
$$\chi^2 = \sum_{i=1}^{k} \frac{(n_i - n_i^O)^2}{n_i^O}$$
- Testovací statistika má za platnosti nulové hypotézy asymptoticky χ^2-rozdělení se stupněm volnosti $\nu = k - 1 - m$.
- Hypotézu, že se náhodná veličina řídí předpokládaným modelem, zamítáme na hladině významnosti α, je-li
$$\chi^2 > \chi^2_{1-\alpha}(\nu),$$
kde $\chi^2_{1-\alpha}(\nu)$ je kvantil χ^2 rozdělení s volností $\nu = k - 1 - m$.

Excel: CHISQ.TEST(naměřené;očekávané) nebo CHITEST(naměřené;očekávané)

(11.1) Při 100 hodech kostkou byly zjištěny následující četnosti jednotlivých stran: $10, 18, 21, 17, 22, 12$. Rozhodněte, zda lze na hladině významnosti 5% považovat kostku za symetrickou.

(11.2) Výrobce prodává výrobek ve třech typech. Dosud bylo prodáno 28 výrobků typu A, 22 typu B a 20 typu C. Použitím chi kvadrát testu testujte H_0: Počet prodaných výrobků nezávisí na typu.

(11.3) Na základě dat v tabulce rozhodněte, zda lze na hladině významnosti 5% považovat počty gólů za výběr z Poissonova rozdělení. (průměr gólů 2.57)

počty gólů	0	1	2	3	4	5 a více
četnost	5	20	20	31	16	8

(11.4) Ověřte χ^2 - testem dobré shody při $\alpha = 5\%$, že výběrové hodnoty roztříděné do tříd s uvedenými experimentálními četnostmi odpovídají normálnímu rozdělení $N(\mu = 10.4; \sigma^2 = 0.5)$

do 9.62	8
9.63 - 10.12	33
10.13 - 10.62	60
10.63 - 11.12	31
11.13 a výše	9

Teorie: Chí kvadrát test nezávislosti (kontingenční tabulky)

Testujeme nezávislost dvou náhodných veličin X a Y, pro které máme k dispozici údaje o četnostech n_{ij} obsažené v kontingenční tabulce.

	y_1	y_2	...	y_J	součty
x_1	n_{11}	n_{12}	...	n_{1J}	$n_{1\bullet}$
x_2	n_{21}	n_{22}	...	n_{2J}	$n_{2\bullet}$
...	...	...	...	...	$n_{2\bullet}$
x_I	n_{I1}	n_{I2}	...	n_{IJ}	$n_{I\bullet}$
součty	$n_{\bullet 1}$	$n_{\bullet 2}$	...	$n_{\bullet J}$	n

Při platnosti nezávislosti lze očekávané četnosti vypočíst podle vztahu

$$n_{ij}^O = \frac{n_{i\bullet} . n_{\bullet j}}{n}$$

Testovací statistika má tvar:

$$\chi^2 = \sum_{i=1}^{I} \sum_{j=1}^{J} \frac{(n_{ij} - n_{ij}^O)^2}{n_{ij}^O}$$

Testovací statistika má za platnosti nulové hypotézy (nezávislost obou veličin) asymptoticky χ^2-rozdělení se stupněm volnosti $\nu = (I-1)(J-1)$.

Hypotézu, že se náhodné veličiny jsou nezávislé, zamítáme na hladině významnosti α, je-li

$$\chi^2 > \chi^2_{1-\alpha}(\nu),$$

kde $\chi^2_{1-\alpha}(\nu)$ je kvantil χ^2 rozdělení s volností $\nu = (I-1)(J-1)$.

(11.5) Byl získán náhodný výběr absolventů VŠ rozdělený následujícím způsobem.

	Bc.	Mgr.	Dr.
muži	534	144	22
žena	515	141	11

Rozhodněte, zda pohlaví a stupeň dosaženého vzdělání jsou nezávislé znaky.

[Očekávané četnosti

	Bc.	Mgr.	Dr.
muži	537.16	145.94	16.90
žena	511.84	139.06	16.10

, hodnota kritéria $\chi^2 = 3.25$]

[kritická hodnota $\chi^2_{0.95} = 5.99$, není důvod zamítat nezávislost znaků]

(11.6) V tabulce jsou uvedeny četnosti barev očí a vlasů u 6800 mužů.

	Světlá	Kaštanová	Černá	Zrzavá
Modrá	1 768	807	189	47
Šedá nebo zelená	946	1 387	746	53
Tmavohnědá	115	438	288	16

Rozhodněte, zda barvy vlasů a očí jsou nezávislé znaky.

[hodnota kritéria $\chi^2 = 1073.51$, kritická hodnota $\chi^2_{0.95} = 12.59$]
[zamítáme hypotézu, že barvy očí a vlasů jsou nezávislé]

(11.7) Z průzkumu provedeného u 1 000 osob, který měl zjistit efektivnost očkování proti chřipce, byly získány tyto výsledky:

	Bez očkování	Jedno očkování	Dvě očkování	Celkem
Chřipka	24	9	13	46
Bez chřipky	289	100	565	954
Celkem	313	109	578	1 000

Testujte, zda má očkování a výskyt chřipky jsou nezávislé veličiny.

Teorie: Kovariance a korelace

Pro dvě náhodné veličiny X a Y lze lineární vazbu vyjádřit pomocí kovariance

$$\operatorname{cov}(X,Y) = \mathrm{E}\left((X-\mathrm{E}X)\cdot(Y-\mathrm{E}Y)\right) = \mathrm{E}\left(X\cdot Y\right) - \mathrm{E}\left(X\right)\cdot \mathrm{E}\left(Y\right)$$

nebo korelace

$$\operatorname{cor}(X,Y) = \rho = \frac{\operatorname{cov}(X,Y)}{\sqrt{\operatorname{var}(X)\cdot\operatorname{var}(Y)}}.$$

Korelace je míra lineární závislosti, pokud $\rho = 0$ jsou X, Y lineárně nezávislé, pokud $\rho = \pm 1$ jsou zcela lineárně závislé.
POZOR: pro $\rho = 0$ může nastat situace, že X, Y jsou závislé nelineárně.
Pro X, Y nezávislé je $\rho = 0$, ale opačně implikace obecně neplatí. (pro normální rozdělení platí)

Na základě dat $x_1, x_2, x_3, \ldots, x_n$ lze odhadnout výběrové charakteristiky. Výběrová kovariance je dána vztahem

$$s_{XY} = \frac{\sum(x_i-\overline{x})(y_i-\overline{y})}{n-1}$$

Výběrová korelace je dána vztahem $r = \dfrac{s_{XY}}{s_X s_Y}$, kde s_{XY} je výběrová kovariance a s_X, resp. s_Y jsou výběrové směrodatné odchylka.

Excel: COVARIANCE.P(data1;data2), COVARIANCE.S(data1;data2),
COVAR(data1;data2), CORREL(data1;data2)

Teorie: Test o koeficientu korelace

Testujeme lineární závislost dvou znaků.
Předpokládáme $H_0 : \rho = 0$ proti alternativní hypotéze $H_1 : \rho \neq 0$.
Kritérium je $T = \frac{r}{\sqrt{1-r^2}}\sqrt{n-2}$ a H_0 zamítáme, pokud $|T| > t_{1-\alpha/2}(\nu = n-2)$.

(11.9) Na základě údajů o 63 studentech bylo zjištěno, že výběrový korelační koeficient mezi body získanými za 1. a 2. písemkou je $r = 0.32$ Rozhodněte, zda výsledky z 1. a 2. testu jsou nekorelované.

[Výsledky jsou korelované, protože $T = 2.60$ a kritická hodnota je 2.00.]
[Výsledné body získané z 1. a 2.testu jsou závislé.]

Teorie: Párový t-test (pro normální rozdělení)

Porovnáváme střední hodnoty souvisejících veličin X a Y:
Předpokládáme $H_0 : \mu_x - \mu_y = d$, zavedeme $z_i = x_i - y_i$ a použijeme kritérium: $T = \frac{\overline{z} - d}{s_z}\sqrt{n}$,
pak zamítám H_0 pokud $|T| > t_{1-\alpha/2}(\nu = n-1)$

(11.10) Na základě údajů o 63 studentech bylo zjištěno, že průměrný rozdíl v získaných bodech za 1. a 2. zápočtovou písemku je 1.10, výběrová směrodatná odchylka je 1.35.

Je třeba ověřit, zda rozdíl v získaných bodech je významný.
[Je významný, protože $T = 6.48$ a kritická hodnota je 2.00.]

(11.11) Byla sledována váha osob před a po tréninku:

před tréninkem	99	62	74	59	70
po tréninku	93	58	66	59	70

Testujte hypotézu, že trénink vede ke změně váhy.

[P-hodnota testu 0.0876, na hladině $\alpha = 5\%$ nezamítám H_0.]