11. cvičení z PSI

12. - 16. prosince 2016

11.1 (Test dobré shody - geometrické rozdělení)
Realizací náhodné veličiny X jsme dostali následující četnosti výsledků:

hodnota	0	1	2	3	4	5	6
pozorovaná četnost	29	15	10	5	3	0	2

Posuďte na hladině významnosti $\alpha = 0.05$ hypotézu, že náhodná veličina X má geometrické rozdělení s parametrem $q = 1/2$, tj. pravděpodobnostní funkce je

$$p_X(i) = q^i(1-q), \quad i \in \mathbb{N}_0 \ .$$

Řešení:
Veličina s geometrickým rozdělením nabývá nekonečně mnoha hodnot. Test dobré shody je ale možné dělat jen s veličinou s *konečně* mnoha hodnotami. Proto musíme některé hodnoty sloučit do jediné skupiny. Zde se přirozeně nabízí udělat to pro hodnoty 6 a výše. Pravděpodobnost pro tuto skupinu je pak součet pravděpodobností jednotlivých hodnot v této skupině. V našem případě je

$$P(X \geq 6) = 1 - P(X < 6) = 1 - \sum_{i=0}^{5} p_X(i) = 1 - \left(\frac{1}{2} + \frac{1}{4} + \frac{1}{8} + \frac{1}{16} + \frac{1}{32} + \frac{1}{64}\right) = \frac{1}{64} \ .$$

Při testu dobré shody porovnáváme naměřené četnosti s očekávanými četnostmi. Rozsah souboru (tj. počet měření) je $N = 29+15+10+5+3+0+2 = 64$. Naší tabulku tedy zpřesníme a doplníme o teoretické pravděpodobnosti p_i a teoretické (tj. očekávané) četnosti $N \cdot p_i$:

položka i	0	1	2	3	4	5	≥ 6
pozorovaná četnost n_i	29	15	10	5	3	0	2
teoretická pravděpodobnost p_i	1/2	1/4	1/8	1/16	1/32	1/64	1/64
teoretická četnost $N \cdot p_i$	32	16	8	4	2	1	1

Další podmínkou pro test dobré shody je to, aby jednotlivé položky měly TEORETICKÉ četnosti $N \cdot p_i \geq 5$. Pokud tomu tak není, je potřeba položky vhodně sloučit tak, abychom této hranice dosáhli. Zde se opět nabízí udělat to pro hodnoty $i \geq 3$.

Původní veličinu X tedy nakonec nahradíme veličinou X' popsanou následující tabulkou:

položka i	0	1	2	≥ 3
pozorovaná četnost n_i	29	15	10	10
teoretická pravděpodobnost p_i	1/2	1/4	1/8	1/8
teoretická četnost $N \cdot p_i$	32	16	8	8

Nyní už můžeme zformulovat naši nulovou hypotézu

$\mathbf{H}_0$: pro pravděpodobnosti hodnot veličiny X' platí $\left(p_0,\ p_1,\ p_2,\ p_{\geq 3}\right) = \left(\frac{1}{2}, \frac{1}{4}, \frac{1}{8}, \frac{1}{8}\right)$,

kterou budeme testovat proti alternativní hypotéze:

$\mathbf{H}_1$: pro pravděpodobnosti hodnot veličiny X' platí $(p_0,\ p_1,\ p_2,\ p_{\geq 3}) \neq \left(\frac{1}{2}, \frac{1}{4}, \frac{1}{8}, \frac{1}{8}\right)$.

Pro test dobré shody používáme určitou statistiku T, jejíž *realizace* t se počítá vzorcem

$$t = \sum_{i \in K} \frac{(n_i - N \cdot p_i)^2}{N \cdot p_i} \ ,$$

kde K je množina položek veličiny X a $k = |K|$ je jejich počet. Rozdělení statistiky T se pro $N \to \infty$ blíží k $\chi^2(k-1)$-rozdělení s $k-1$ stupni volnosti (právě kvůli přibližnosti jsme také potřebovali teoretické četnosti ≥ 5).

Kritérium pro **ZAMÍTNUTÍ** bude podobné jako u jednostranného testu rozptylu (protože jde opět o χ^2-rozdělení). Je tedy tvaru

$$t > q_{\chi^2(k-1)}\,(1-\alpha) \quad \Rightarrow \quad \text{zamítáme } \mathbf{H}_0 \text{ (na dané hladině } \alpha) \ .$$

Zdůvodnění tvaru zamítacího kritéria: Máme-li správné rozdělení, měly by být odchylky teoretických a naměřených četností malé a proto i hodnota statistiky T bude spíše menší. Jako kritický obor si tudíž volíme opět $W : (u_1, \infty)$, kde má platit, že $P(u_1 < T) = \alpha$. Dostaneme tak, že $u_1 = q_{\chi^2(k-1)}\,(1-\alpha)$, protože předpokládáme, že T má přibližně χ^2-rozdělení.

V našem případě máme $k = 4$. Hodnota statistiky je

$$t = \frac{(29-32)^2}{32} + \frac{(15-16)^2}{16} + \frac{(10-8)^2}{8} + \frac{(10-8)^2}{8} = 1.34375$$

a hodnota kvantilu je

$$q_{\chi(k-1)}(1-\alpha) = q_{\chi(3)}(0.95) \doteq 7.815 \ .$$

Protože

$$t = 1.34375 \not> 7.815 \doteq q_{\chi(3)}(0.95) \ ,$$

nulovou hypotézu $\mathbf{H}_0$ pro veličinu X' **NEZAMÍTÁME**. Tento výsledek interpretujeme tak, že hypotézu

$$X \textit{ má geometrické rozdělení s parametrem } q = 1/2,$$

rovněž **NEZAMÍTÁME**.

11.2 (Test dobré shody - rovnoměrné rozdělení)

Účastníci konference budou ubytováni ve čtyřpatrovém penziónu se 12 pokoji, v každém patře jsou tři pokoje se dvěma lůžky. Každý z $N = 20$ účastníků poslal organizátorům nezávisle svůj požadavek čísla pokoje, kde by chtěl být ubytovaný. Čísla byla následující

$$(8,\ 12,\ 5,\ 4,\ 3,\ 5,\ 6,\ 12,\ 11,\ 2,\ 6,\ 4,\ 2,\ 12,\ 11,\ 9,\ 6,\ 7,\ 9,\ 9).$$

Otestujte na hladině významnosti $\alpha = 0.05$ hypotézu

$\mathbf{H}_0$: rozdělení účastníků do pater je rovnoměrné

proti alternativě

$\mathbf{H}_1$: rozdělení účastníků do pater není rovnoměrné.

patro	1	2	3	4
čísla pokojů	$1-3$	$4-6$	$7-9$	$10-12$

Řešení:
Veličina X, která přiřazuje účastníkovi patro, ve kterém bude bydlet, má 4 položky. Požadavek rovnoměrného rozdělení znamená, že pravděpodobnosti p_i těchto položek (tj. patra očíslovaná pomocí $i = 1, \ldots, 4$) budou $(p_1, p_2, p_2, p_4) = \left(\frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4}\right)$. Hypotézu tedy vyjádříme konkrétně:

$\mathbf{H}_0$: pro pravděpodobnosti hodnot veličiny X platí $(p_1, p_2, p_2, p_4) = \left(\frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4}\right)$,

a alternativní hypotéza bude:

$\mathbf{H}_1$: pro pravděpodobnosti hodnot veličiny X platí $(p_1, p_2, p_2, p_4) \neq \left(\frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4}\right)$.

K rozhodování použijeme χ^2-test dobré shody. Setřídíme data do skupin a vypočteme empirické četnosti n_i, které zapíšeme spolu s teoretickými četnostmi $N \cdot p_i$ do tabulky. Ze zadání máme $N = 20$ počet dat, $k = 4$ počet tříd a $p_i = \frac{1}{4}$, $1 \leq i \leq 4$, pro rovnoměrné rozdělení.

číslo patra i	1	2	3	4
čísla pokojů	$1-3$	$4-6$	$7-9$	$10-12$
n_i	3	7	5	5
$N \cdot p_i$	5	5	5	5

Podmínka na teoretické četnosti ≥ 5 je splněna, takže položky nemusíme slučovat. Dosadíme do vzorce pro realizaci t statistiky T, která má v tomto případě rozdělení přibližně $\chi^2(k-1) = \chi^2(3)$

$$t = \sum_{i=1}^{4} \frac{(n_i - N \cdot p_i)^2}{N \cdot p_i} = \frac{1}{5}\left(2^2 + 2^2 + 0^2 + 0^2\right) = 1.6\,,$$

a porovnáme ji s kvantilem

$$q_{\chi^2(k-1)}(1-\alpha) = q_{\chi^2(3)}(0.95) \doteq 7.815\ .$$

Protože

$$t = 1.6 \not> 7.815 \doteq q_{\chi(3)}(0.95)\ ,$$

nulovou hypotézu $\mathbf{H}_0$, že rozdělení účastníků do pater je rovnoměrné, **NEZAMÍTÁME**.

11.3 (Test dobré shody - nezávislost veličin)
Na $N = 100$ lidech byla pozorována barva očí a vlasů. Data jsou shrnuta v tabulce. Na hladině $\alpha = 5\%$ testujte hypotézu o nezávislosti barvy očí a vlasů.

Vlasy \ Oči	tmavé	světlé
modré	10	20
šedé	10	10
hnědé	40	10

Řešení:
Označme si X veličinu, která přiřazuje danému člověku barvu vlasů a Y veličinu, která přiřazuje témuž člověku barvu očí. Budeme testovat hypotézu:

$\mathbf{H}_0$: rozdělení veličin X a Y jsou *nezávislá*

proti alternativní hypotéze:

$\mathbf{H}_1$: rozdělení veličin X a Y jsou *závislá*.

na hladině významnosti $\alpha = 5\%$. Označme si ještě pro jednoduchost obor hodnot veličiny X jako $A = \{$*modré, šedé, hnědé*$\}$ a obor veličiny Y jako $B = \{$*tmavé, světlé*$\}$. Četnosti pro $(i,j) \in A \times B$ z tabulky označme jako $n_{i,j}$.

Rozdělení veličin X ani Y neznáme a proto je odhadneme jako

$$p_X(i) = \frac{n_{i,*}}{N} \ ,$$

$$p_Y(j) = \frac{n_{*,j}}{N} \ ,$$

kde

$$n_{*,i} = \sum_{j \in B} n_{i,j} \quad \text{a} \quad n_{j,*} = \sum_{i \in A} n_{i,j}$$

jsou marginální četnosti.

$n_{i,j}$ $(Y =)\ j$ / $(X =)\ i$	tmavé	světlé	$n_{i,*}$
modré	10	20	30
šedé	10	10	20
hnědé	40	10	50
$n_{*,j}$	60	40	100

Za předpokladu nezávislosti veličin X a Y máme $p_{X,Y}(i,j) = p_X(i) \cdot p_Y(j)$. Hypotézu o nezávislosti tedy můžeme přeformulovat takto

$\mathbf{H}_0 : p_{X,Y}(i,j) = p_X(i) \cdot p_Y(i)$ pro všechna $(i,j) \in A \times B$,

a alternativní hypotézu jako

$\mathbf{H}_1 : p_{X,Y}(i,j) \neq p_X(i) \cdot p_Y(i)$ pro alespoň jedno $(i,j) \in A \times B$.

Otestování hypotézy $\mathbf{H}_0$ tak bude **TÉMĚŘ** odpovídat obvyklému testu dobré shody s předepsaným rozdělením (tentokrát pracujeme s diskrétním náhodným vektorem (X,Y)) ale s tím rozdílem, že počet stupňů volnosti bude (kvůli odhadu marginálních pravděpodobností) **JINÝ**, než by tomu bylo u obvyklého testu dobré shody se 6 položkami. Počet stupňů volnosti je v tomto případě

$$\big(|A| - 1\big) \cdot \big(|B| - 1\big) = (3-1) \cdot (2-1) = 2 \ .$$

Za předpokladu $\mathbf{H}_0$ pro očekávané četnosti pro jednotlivé hodnoty (i,j) náhodného vektoru (X,Y) pak bude platit, že

$$N \cdot p_{X,Y}(i,j) = N \cdot p_X(i) \cdot p_Y(j) = \frac{n_{i,*} \cdot n_{*,j}}{N} \ .$$

Tabulka pro pro tyto četnosti bude:

$N \cdot p_{X,Y}(i,j)$ $(Y =)\ j$ / $(X =)\ i$	tmavé	světlé	$n_{i,*}$
modré	$\frac{30 \cdot 60}{100} = 18$	$\frac{30 \cdot 40}{100} = 12$	30
šedé	$\frac{20 \cdot 60}{100} = 12$	$\frac{20 \cdot 40}{100} = 8$	20
hnědé	$\frac{50 \cdot 60}{100} = 30$	$\frac{50 \cdot 40}{100} = 20$	50
$n_{*,j}$	60	40	100

Podmínka na teoretické (tj. očekávané) četnosti ≥ 5 je splněna, takže položky nemusíme slučovat. Pro realizaci testovací statistiky dostaneme

$$t = \sum_{i,j} \frac{\Big(n_{i,j} - N \cdot p_{X,Y}(i,j)\Big)^2}{N \cdot p_{X,Y}(i,j)} =$$

$$= \frac{(10-18)^2}{18} + \frac{(10-12)^2}{12} + \frac{(40-30)^2}{30} + \frac{(20-12)^2}{12} + \frac{(10-8)^2}{8} + \frac{(10-20)^2}{20} = 18 + \frac{1}{18} \doteq 18.056$$

a porovnáme ji s hodnotou kvantilu χ^2 pro $(3-1)\cdot(2-1) = 2$ stupňů volnosti

$$q_{\chi^2(2)}(1-\alpha) = q_{\chi^2(2)}(0.95) \doteq 5.992 \ .$$

Protože

$$t \doteq 18.056 > 5.992 \doteq q_{\chi^2(2)}(0.95) \ ,$$

hypotézu o nezávislosti proto **ZAMÍTÁME**.