# 12. cvičení z PST

19. prosince 2018

**12.1** (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 | $\geq 6$ |
|---|---|---|---|---|---|---|---|
| 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 | $\geq 3$ |
|---|---|---|---|---|
| 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 $(p_0,\ p_1,\ p_2,\ p_{\geq 3}) = (\frac{1}{2}, \frac{1}{4}, \frac{1}{8}, \frac{1}{8})$,

kterou budeme testovat proti alternativní hypotéze:

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

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 $\geq 5$).

Kritérium pro **ZAMÍTNUTÍ** bude podobné jako u jednostranného testu rozptylu (protože jde opět o $\chi^2$-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 $\chi^2$-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**.

**12.2** (test dobré shody - rozdělení dané geometrickou pravděpodobností)

Chceme zjistit, zda si jistý druh ptáka buduje hnízda rovnoměrně po krajině. K tomu jsme rozdělili testovací region na 6 souvislých částí, jejichž rozlohy v km$^2$ jsou uvedeny v tabulce. Tabulka udává i počet hnízd nalezených v dané části regionu. Za hladinu významnosti považujte $\alpha = 5$ %.

| oblast | A | B | C | D | E | F |
|---|---|---|---|---|---|---|
| rozloha (v km$^2$) | 5 | 10 | 10 | 5 | 15 | 15 |
| počet hnízd | 14 | 22 | 28 | 12 | 40 | 34 |

**Řešení:**

Využijeme test dobré shody. Počet hnízd v regionu je $n = 14 + 22 + 28 + 12 + 40 + 34 = 150$. Označíme $X_i$, $i \in \{1, ..., n\}$, veličinu nabývající hodnot $A, ..., F$, v závislosti na tom, v které části regionu leží $i$-té hnízdo. Neboli máme veličinu

$$X(\textit{“dané hnízdo”}) = \textit{“oblast, ve které hnízdo leží”}$$

a $X_i$ jsou její kopie v jednotlivých pokusech (tedy jsou to nezávislé veličiny se stejným rozdělením).

Celkově rozloha testovacího regionu je $S = 5 + 10 + 10 + 5 + 15 + 15 = 60$ km$^2$. Protože předpokládáme rovnoměrné rozdělení po krajině, bude pravděpodobnost $p_j$ nalezení hnízda v dané oblasti $j \in \{A, ..., F\}$ úměrná její velikosti, tj. daná geometrickou pravděpodobností jako

$$p_j = \frac{\text{“rozloha oblasti } j\text{”}}{S} \ .$$

Můžeme tedy doplnit tabulku následovně:

| oblast | A | B | C | D | E | F |
|---|---|---|---|---|---|---|
| rozloha (v km$^2$) | 5 | 10 | 10 | 5 | 15 | 15 |
| počet hnízd $n_j$ | 14 | 22 | 28 | 12 | 40 | 34 |
| teoretická pravděpodobnost $p_j$ | $\frac{5}{60} = \frac{1}{12}$ | $\frac{10}{60} = \frac{1}{6}$ | $\frac{10}{60} = \frac{1}{6}$ | $\frac{5}{60} = \frac{1}{12}$ | $\frac{15}{60} = \frac{3}{12}$ | $\frac{15}{60} = \frac{3}{12}$ |
| teoretický počet hnízd $np_j$ | 12.5 | 25 | 25 | 12.5 | 37.5 | 37.5 |
| příspěvek ke statistice $\frac{(n_j - np_j)^2}{np_j}$ | 0.18 | 0.36 | 0.36 | 0.02 | 0.167 | 0.327 |

Hypotézu tedy vyjádříme konkrétně:

$\mathbf{H}_0$ : pro pravděpodobnosti hodnot veličiny $X$ platí $(p_A, \dots, p_F) = \left(\frac{1}{12}, \dots, \frac{3}{12}\right)$,

a alternativní hypotéza bude:

$\mathbf{H}_1$ : pro pravděpodobnosti hodnot veličiny $X$ platí $(p_A, \dots, p_F) \neq \left(\frac{1}{12}, \dots, \frac{3}{12}\right)$.

Testovací statistika

$$T = \sum_{j=A}^{F} \frac{(n_j - np_j)^2}{np_j}$$

má za předpokladu rovnoměrného rozdělení hnízd v regionu $\chi^2$-rozdělení o 5 stupních volnosti. Z našich realizací získáme $t \doteq 1.413$, což porovnáme s tabulkovou hodnotou kvantilu $q_{\chi^2(5)}(1-\alpha) \doteq 11.1 > t$. Na hladině významnosti 5 % tedy nemůžeme zamítnout, že si pták buduje hnízda rovnoměrně po krajině.

**12.3** (test nezávislosti 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.

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

**Řešení:**

Označme si $X$ veličinu, která přiřazuje danému člověku barvu očí a $Y$ veličinu, která přiřazuje témuž člověku barvu vlasů. 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.

| $n_{i,j}$ $(X =)\ i$ \ $(Y =)\ j$ | 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 $\geq 5$ 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 $\chi^2$ 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**.

**12.4** (test nezávislosti veličin)

Úspěšnost u zkoušek ve vztahu k počtu přítomných studentů udává tabulka:

| termín | 1. | 2. | 3. | 4. | 5. |
|---|---|---|---|---|---|
| počet přítomných | 20 | 30 | 40 | 60 | 50 |
| počet úspěšných | 11 | 8 | 14 | 43 | 24 |

Otestujte na hladině významnosti 5 % hypotézu, že pravděpodobnost úspěchu byla u všech zkouškových termínů stejná.

**Řešení:**

Označme si veličiny

$X($ *"účast daného studenta na i-tém termínu"* $) :=$ *"zda uspěl (1)/ nebo ne (0)"*

$Y($ *"účast daného studenta na i-tém termínu"* $) :=$ $i$

Pravděpodobnost úspěchu v $i$-tém termínu je dána jako $P(X = 1|\ Y = j)$. Ukážeme, že tato hodnota bude nezávislá na $j$ právě když veličiny $X$ a $Y$ budou nezávislé.

($\Leftarrow$): Z nezávislosti $X$ a $Y$ ihned máme, že $P(X = 1|\ Y = j) = P(X = 1) = \mathit{konst.}$

($\Rightarrow$): Naopak, nechť $c = P(X = 1|\ Y = j) = \frac{P(X=1\ \cap\ Y=j)}{P(Y=j)}$ pro všechna $j$. Tedy

$$P(X = 1\ \cap\ Y = j) = c \cdot P(Y = j)$$

a sečtením přes všechna $j$ dostaneme, že

$$P(X=1)=\sum_j P(X=1\ \cap\ Y=j)=\sum_j c\cdot P(Y=j)=c\cdot\underbrace{\sum_j P(Y=j)}_{=1}=c$$

neboli

$$P(X=1\ \cap\ Y=j)=P(X=1)\cdot P(Y=j)\ .$$

Podobně z $P(X=0|\ Y=j)=1-c$ pro všechna $j$ odvodíme, že $P(X=0\ \cap\ Y=j)=P(X=0)\cdot P(Y=j)$ . Tedy veličiny $X$ a $Y$ jsou nezávislé.

Budeme tedy 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 $\alpha=5\%$.

Četnosti $n_{ij}$ jednotlivých případů pro $X=i$ a $Y=j$ přepíšeme pomocí tabulky

| $n_{i,j}$ $(Y=)\ j$ / $(X=)\ i$ | 1 | 2 | 3 | 4 | 5 | $n_{i,*}$ |
|---|---|---|---|---|---|---|
| uspěl | 11 | 8 | 14 | 43 | 24 | 100 |
| neuspěl | 9 | 22 | 26 | 17 | 26 | 100 |
| $n_{*,j}$ | 20 | 30 | 40 | 60 | 50 | 200 |

kde

$$n_{*,i}=\sum_j n_{i,j}\quad\text{a}\quad n_{j,*}=\sum_i n_{i,j}$$

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 $N=\sum_{i,j} n_{i,j}=200$ a hypotézy tak můžeme vyjádřit jako

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

a alternativní hypotézu jako

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

Tabulka pro teoretické četnosti

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

pak bude:

| $N\cdot p_{X,Y}(i,j)$ $(Y=)\ j$ / $(X=)\ i$ | 1 | 2 | 3 | 4 | 5 | $n_{i,*}$ |
|---|---|---|---|---|---|---|
| uspěl | $\frac{20\cdot 100}{200}=10$ | $\frac{30\cdot 100}{200}=15$ | 20 | 30 | 25 | 100 |
| neuspěl | $\frac{20\cdot 100}{200}=10$ | 15 | 20 | 30 | 25 | 100 |
| $n_{*,j}$ | 20 | 30 | 40 | 60 | 50 | 200 |

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

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

$$=\frac{(11-10)^2}{10}+\frac{(9-10)^2}{10}+\frac{(8-15)^2}{15}+\frac{(22-15)^2}{15}+\frac{(14-20)^2}{20}+\frac{(26-20)^2}{20}+$$

$$+\frac{(43-30)^2}{30}+\frac{(17-30)^2}{30}+\frac{(24-25)^2}{25}+\frac{(26-25)^2}{25}=21.68$$

a porovnáme ji s hodnotou kvantilu $\chi^2$ pro $(5-1)\cdot(2-1)=4$ stupně volnosti

$$q_{\chi^2(4)}(1-\alpha) = q_{\chi^2(4)}(0.95) \doteq 9.49 \ .$$

Protože

$$t \doteq 21.68 > 9.49 \doteq q_{\chi^2(4)}(0.95) \ ,$$

hypotézu o nezávislosti proto **ZAMÍTÁME**.