12. cvičení z PSI

7. prosince 2022

Poznámky k testu dobré shody: Chceme otestovat (na hladině α), jestli daná veličina X s konečně mnoha (navzájem různými!) hodnotami $a_1, \ldots, a_k$ (ne nutně číselnými) má předepsané pravděpodobnosti $(p_1, \ldots, p_k)$, tedy nulovou hypotézu

$$\mathbf{H}_0 : P(X = a_i) = p_i \text{ pro všechna } i \in \{1, \ldots, k\}$$

proti alternativní hypotéze:

$$\mathbf{H}_A : P(X = a_{i_0}) \neq p_{i_0}, \text{ pro alespoň jedno } i_0 \in \{1, \ldots, k\}$$

Při n pokusech s veličinou X si pro $i = 1, \ldots, k$ označme veličiny

$$N_i = \text{“}\textit{počet výskytů případu } X = a_i \textit{ při } n \textit{ pokusech}\text{”} .$$

Máme tedy náhodný vektor

$$\boldsymbol{N} = (N_1, \ldots, N_k)$$

a vztah $N_1 + \cdots + N_k = n$. Už z toho vidíme, že veličiny N_i nejsou nezávislé, ale zase k té nezávislosti tak daleko nemají. Náhodný vektor $\boldsymbol{N}$ má tzv. multinomické rozdělení a jednotlivá marginální rozdělení veličin jsou binomická, konkrétně $N_i \sim \mathrm{Bi}(n, p_i)$. Speciálně tedy $E(N_i) = n \cdot p_i$.

Jako testovací veličinu zde používáme:

$$T = \sum_{i=1}^{k} \frac{(N_i - n \cdot p_i)^2}{n \cdot p_i}$$

která má asymptoticky (tj. pro $n \to \infty$) tzv. χ^2-rozdělení s $k-1$ stupni volnosti. Pro praktické použití této asymptotiky se obvykle požaduje, aby platilo, že

$$n \cdot p_i \geq 5 \quad \text{pro všechna} \quad i \in \{1, \ldots, k\} .$$

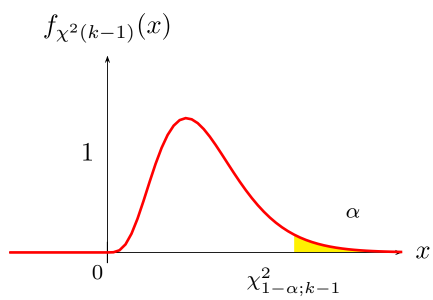
Hodnoty $n \cdot p_i$ se označují jako tzv. *teoretické četnosti*.

Pokud tedy platí nulová hypotéza $\mathbf{H}_0$, měly by být hodnoty veličiny T malé. Jestliže hodnoty T budou příliš velké, bude to důvod k zamítnutí nulové hypotézy.

Jak určit hranici, kde už nastane zamítnutí: veličina T má (přibližně) $\chi^2_{(k-1)}$ rozdělení, tedy platí

$$P_{(H_0 \text{ platí})}(T > \chi^2_{1-\alpha;\, k-1}) \doteq \alpha$$

kde $\chi^2_{1-\alpha;\, k-1}$ je hodnota kvantilu pro $\chi^2_{(k-1)}$ rozdělení (viz obrázek, kde α je velikost žluté plochy pod hustotou $f_{\chi^2(k-1)}(x)$ pro $\chi^2_{(k-1)}$ rozdělení).



Kritérium pro **ZAMÍTNUTÍ $\mathbf{H}_0$** (na hladině α) proto volíme jako

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

Z definice chyby 1. druhu, tj.

$$\text{nastává chyba 1. druhu} \quad \Leftrightarrow \quad (\text{hypotéza } \mathbf{H}_0 \text{ platí} \quad \& \quad \text{my ji zamítáme})$$

pak totiž máme, že

$$P_{(H_0 \text{ platí})}\Big(\text{nastává chyba 1. druhu}\Big) = P_{(H_0 \text{ platí})}\Big(\text{zamítáme } \mathbf{H}_0 \text{ (na hladině } \alpha)\Big) =$$

$$= P_{(H_0 \text{ platí})}\big(T > \chi^2_{1-\alpha;\, k-1}\big) \doteq \alpha$$

neboli pravděpodobnost chyby 1. druhu (ovšem za předpokladu platnosti $\mathbf{H}_0$!) je pak omezena hodnotou α.

12.1 *Firma má 3 pobočky. Dva roky bylo sledováno, která z nich zaznamenala nejvyšší měsíční výnos. Bylo zjištěno, že nejvýnosnější byla první pobočka $10\times$, druhá $6\times$ a třetí $8\times$.*
Je možné říct, že první pobočka je nejvýnosnější $2\times$ častěji než každá ze zbylých dvou? Testujte na hladině 5%.

Řešení:
Máme tedy veličinu

$$X = \text{"}\textit{číslo pobočky, která je zrovna (tj. v daném měsíci) nejvýnosnější}\text{"}$$

s $k = 3$ hodnotami {*první, druhá, třetí*}.

Nejdříve si potřebujeme zjistit, jaké rozdělení

$$P(X = \textit{první}) = p_1, \quad P(X = \textit{druhá}) = p_2, \quad P(X = \textit{třetí}) = p_3$$

vlastně předpokládáme. Z požadavku máme

$$p_1 = 2 \cdot p_2, \quad p_1 = 2 \cdot p_3, \quad p_1 + p_2 + p_3 = 1$$

z čehož dostáváme

$$(p_1, p_2, p_3) = \left(\frac{1}{2}, \frac{1}{4}, \frac{1}{4}\right)$$

Naše hypotéza tedy je

$\mathbf{H}_0$: veličina X má rozdělení s pravděpodobnostmi $(p_1, p_2, p_3) = \left(\frac{1}{2}, \frac{1}{4}, \frac{1}{4}\right)$,

a alternativní hypotéza bude:

$\mathbf{H}_A$: veličina X má rozdělení *jiné* než $\left(\frac{1}{2}, \frac{1}{4}, \frac{1}{4}\right)$.

Využijeme test dobré shody. Celkový počet měření (tj. počet měsíců) je $n = 10 + 6 + 8 = 24$. Pro přehlednost si vypíšeme tabulku s jednotlivými četnostmi (pozorovanými i teoretickými):

i (pobočky)	první	druhá	třetí
n_i (pozorované četnosti)	10	6	8
p_i (teoretické pravděpodobnosti)	$\frac{1}{2}$	$\frac{1}{4}$	$\frac{1}{4}$
$n \cdot p_i$ (teoretické četnosti)	$24 \cdot \frac{1}{2} = 12$	$24 \cdot \frac{1}{4} = 6$	$24 \cdot \frac{1}{4} = 6$

Vidíme, že všechny teoretické četnosti jsou ≥ 5, takže skutečně můžeme použít asymptotické přiblížení pro testovací statistiku T (ta tedy bude mít χ^2-rozdělení). Teď už si jen spočítáme hodnotu této statistiky

$$t = \sum_{i=1}^{3} \frac{(n_i - np_i)^2}{np_i} = \frac{(10-12)^2}{12} + \frac{(6-6)^2}{6} + \frac{(8-6)^2}{6} = \frac{1}{3} + 0 + \frac{2}{3} = 1$$

a porovnáme s kvantilem χ^2-rozdělení s $k - 1 = 3 - 1 = 2$ stupni volnosti:

$$t = 1 \not> 5.99 \doteq \chi^2_{0.95;\, 2} = \chi^2_{1-\alpha;\, k-1}$$

Protože zamítací kritérium NENÍ splněno, tak $\mathbf{H}_0$ **NEZAMÍTÁME** (na hladině α).

12.2 *Pro pojištění motorových vozidel používá pojišťovna tři kategorie pojistného:*

1 - základní pojistné,

2 - bonus 30%,

3 - bonus 50%.

V prvním roce platí pojištěný základní pojistné. Jestliže má rok bezeškodní průběh, je pojištěný v dalším roce zařazen o třídu výše (pokud je to možné), pokud ale uplatní jeden pojistný nárok, je v příštím roce zařazen o kategorii níže (pokud je to možné) a při uplatnění více než jednoho nároku o dvě kategorie níže (pokud je to možné). Počty výskytů pojistné události v jednotlivých letech jsou nezávislé, stejně rozdělené náhodné veličiny s Poissonovým rozdělením s parametrem λ.

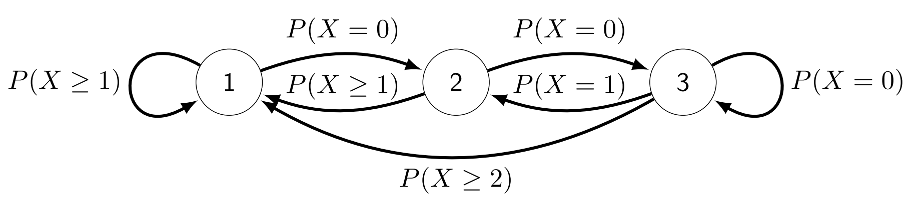
(a) *Určete počáteční rozdělení a matici pravděpodobností přechodu.*

(b) *Najděte stacionární rozdělení pro případ, kdy každý řidič hlásí škodu průměrně jednou za 10 let (hodnoty v matici pravděpodobností přechodu zaokrouhlete na jedno desetinné místo).*

Řešení:
Veličina

$$X = \textit{počet pojistných události pojištěnce za jeden rok}$$

má Poissonovo rozdělení $Poiss(\lambda)$, s konkrétní volbou $\lambda = \frac{1}{10} = 0.1$. Graf Markovova řetězce s pravděpodobnostmi přechodu bude následující



Konkrétní pravděpodobnosti jsou

$$P(X = 0) = e^{-\lambda} = e^{-0.1} \doteq 0.9$$

$$P(X = 1) = \lambda e^{-\lambda} = 0.1 e^{-0.1} \doteq 0.1$$

$$P(X \geq 1) = 1 - e^{-\lambda} = 1 - e^{-0.1} \doteq 0.9$$

$$P(X \geq 2) = 1 - (1+\lambda)e^{-\lambda} = 1 - 1.1e^{-0.1} \doteq 0 \ .$$

Počáteční rozdělení vycházející ze stavu 1 bude $\boldsymbol{p}_0 = (1,0,0)$. Matice přechodu pak je

$$\mathbf{P} = \begin{pmatrix} 1-e^{-\lambda} & e^{-\lambda} & 0 \\ 1-e^{-\lambda} & 0 & e^{-\lambda} \\ 1-(1+\lambda)e^{-\lambda} & \lambda e^{-\lambda} & e^{-\lambda} \end{pmatrix} \doteq \begin{pmatrix} 0.1 & 0.9 & 0 \\ 0.1 & 0 & 0.9 \\ 0 & 0.1 & 0.9 \end{pmatrix}$$

Stacionární rozdělení pravděpodobnosti $\boldsymbol{p} = (p_1, p_2, p_3) \in \mathbb{R}^3$ je takové, které se nemění při jednom kroku. Tedy platí

$$\boldsymbol{p} = \boldsymbol{p} \cdot \mathbf{P} \quad \text{neboli} \quad \boldsymbol{p} \cdot (\mathbf{P} - \mathbf{I}_3) = 0 \ .$$

Ekvivalentní zápis je

$$(\mathbf{P}^T - \mathbf{I}_3^T) \cdot \boldsymbol{p}^T = 0^T \ .$$

(A pochopitelně musí být $p_1, p_2, p_3 \geq 0$ a $\sum_{j=1}^{3} p_j = 1$, protože se jedná o rozdělení pravděpodobnosti.)

Pozor! Přechod k transponované matici je nezbytný, pokud budeme při Gaussově eliminaci používat ekvivalentní ŘÁDKOVÉ úpravy soustavy $(\mathbb{A}|0)$. Ta totiž odpovídá rovnici $\mathbb{A}x = 0$, kdy x je SLOUPCOVÝ vektor.

Pomocí Gaussovy eliminace najdeme tudíž řešení soustavy reprezentované maticí

$$\mathbf{P}^T - \mathbf{I}_3^T = \begin{pmatrix} 0.1 & 0.1 & 0 \\ 0.9 & 0 & 0.1 \\ 0 & 0.9 & 0.9 \end{pmatrix} - \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} = \begin{pmatrix} -0.9 & 0.1 & 0 \\ 0.9 & -1 & 0.1 \\ 0 & 0.9 & -0.1 \end{pmatrix} \sim \begin{pmatrix} -0.9 & 0.1 & 0 \\ 0 & 0.9 & -0.1 \end{pmatrix}.$$

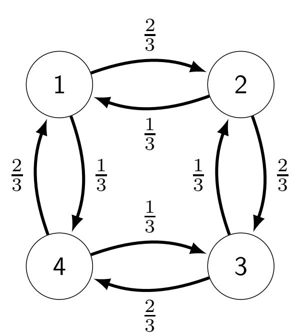
Tato soustava má hodnost 2, tedy má pouze $3 - 2 = 1$ lineárně nezávislých řešení, např. $(\frac{1}{9}, 1, 9)$. Po jeho "znormování" (tj. vydělení číslem $\frac{1}{9} + 1 + 9 = \frac{91}{9}$) pak dostaneme

$$\boldsymbol{p} = \left(\frac{1}{91}, \frac{9}{91}, \frac{81}{91} \right) \ .$$

Pozor! NEJEDNÁ se o eukleidovskou normu $\|\boldsymbol{p}\|_2 := \left(\sum_i |p_i|^2\right)^{\frac{1}{2}}$, ALE o normu tvaru $\|\boldsymbol{p}\|_1 := \sum_i |p_i|$.

Poznámka: Postupné změny rozdělení pravděpodobnosti V Markovově řetězci během jednotlivých kroků si můžeme představovat tak, že máme k dispozici dané množství kapaliny (o objemu 1), které se přelévá mezi jednotlivými stavy. Přechodné stavy se pak vyznačují tím, že to, co z nich "odteče" do trvalých stavů, se už do nich nevrátí, takže celkové množství kapaliny v těchto stavech se postupně v limitě sníží až k nule.

12.3 Markovovův řetězec je dán grafem:



(a) Napište matici přechodu, klasifikujte stavy a stanovte všechny komponenty.

(b) Najděte stacionární rozdělení.

Řešení:

(a) Odpovídající matice přechodu je

$$\mathbf{P} = \begin{pmatrix} 0 & 2/3 & 0 & 1/3 \\ 1/3 & 0 & 2/3 & 0 \\ 0 & 1/3 & 0 & 2/3 \\ 2/3 & 0 & 1/3 & 0 \end{pmatrix} .$$

Všechny stavy jsou trvalé a navzájem propojené cestami. Tvoří tedy jedinou komponentu.

(b) Řetězec má jen jednu komponentu a stacionární rozdělení je tak jen jedno. Protože všechny vrcholy v grafu mají stejné vlastnosti, můžeme toto rozdělení uhádnout

$$\mathbf{p}_{stac} = \left(\frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4}\right)$$

(a pak jen snadno ověřit), že to skutečně stacionární rozdělení je.