13. cvičení z PSI

9.- 13. května 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

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

proti alternativní hypotéze:

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

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

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

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

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

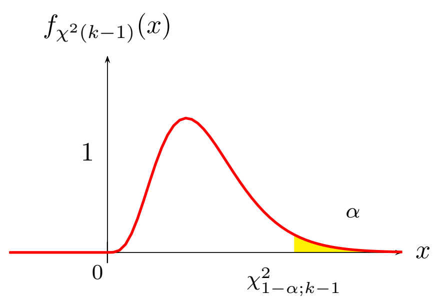
Hodnoty $n \cdot p_i$ se označují jako tzv. *teoretické četnosti*.

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 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 **H**$_0$!) je pak omezena hodnotou α.

13.1 *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

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

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

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:

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	8	6
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

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

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).

Poznámky k testu nezávislosti: Máme veličiny

- X s (různými) hodnotami $\{a_1, \ldots, a_k\}$ a
- Y s (různými) hodnotami $\{b_1, \ldots, b_\ell\}$

a chceme otestovat (na hladině α), hypotézu

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

proti alternativní hypotéze:

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

Při n pokusech s náhodným vektorem (X,Y) si pro $i = 1, \ldots, k$ označme veličiny

$$N_{i,j} = \text{"počet výskytů případu } (X,Y) = (a_i, b_j) \text{ při } n \text{ pokusech"}\,.$$

a opět máme náhodný vektor

$$\boldsymbol{N} = (N_{1,1}, \ldots, N_{k,\ell})$$

s multinomickým rozdělením. Marginální rozdělení jednotlivých veličin $N_{i,j}$ jsou opět binomická a *za předpokladu nezávislosti X a Y* mají střední hodnotu

$$E(N_{i,j}) = n \cdot P(X = a_i, Y = b_j) \overset{\text{(nezáv.)}}{=} n \cdot P(X = a_i) \cdot P(Y = b_j)\,.$$

Podobně jako v testu dobré shody by tyto střední hodnoty představovaly teoretické četnosti až na to, že pravděpodobnosti $P(X = a_i)$ a $P(Y = b_j)$ nemáme v hypotéze uvedeny. Proto je odhadneme jako

$$P(X = a_i) \doteq \frac{n_{i,\bullet}}{n} \quad \text{a} \quad P(Y = b_j) \doteq \frac{n_{\bullet,j}}{n}$$

kde $n_{i,\bullet}$ a $n_{\bullet,j}$ jsou naměřené hodnoty veličin

$$N_{i,\bullet} = \sum_{j=1}^{\ell} N_{i,j} = \text{"počet výskytů případu } X = a_i \text{ při } n \text{ pokusech"}$$

$$N_{\bullet,j} = \sum_{i=1}^{k} N_{i,j} = \text{"počet výskytů případu } Y = b_j \text{ při } n \text{ pokusech"}$$

což jsou tzv. *marginální četnosti.*

Z tohoto důvodu jako testovací veličinu volíme:

$$T = \sum_{i=1}^{k} \sum_{j=1}^{\ell} \frac{\left(N_{i,j} - \frac{N_{i,\bullet} \cdot N_{\bullet,j}}{n}\right)^2}{\frac{N_{i,\bullet} \cdot N_{\bullet,j}}{n}}$$

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

$$\frac{n_{i,\bullet} \cdot n_{\bullet,j}}{n} \geq 5 \quad \text{pro všechna} \quad i = 1,\dots,k \text{ a } j = 1,\dots,\ell.$$

Kritérium pro **ZAMÍTNUTÍ H**$_0$ (na hladině α) volíme podobně jako u testu dobré shody a sice

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

13.2 *Na $n = 100$ osobách byla pozorována barva očí a vlasů. Naměřeny byly následující sdružené četnosti:*

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

(a) Jsou barvy očí a vlasů nezávislé? Testujte na hladině 5%.

(b) Otestujte na hladině 5%, jestli je v populaci stejně tmavovlasých jako světlovlasých.

Řešení:

Označme si veličiny

X= *"barva očí daného člověka"*

Y= *"barva vlasů daného člověka"*

a dál budeme pracovat s náhodným vektorem (X,Y), tj. u daného člověka budeme zjišťovat barvu očí a barvu vlasů.

(a) 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 $\alpha = 5\%$.

Četnost případu $(X,Y) = (i,j)$ v tabulce označme jako $n_{i,j}$ a marginální četnosti pak budou

$n_{i,\bullet} = \sum_j n_{i,j}$ pro případ $X = i$

$n_{\bullet,j} = \sum_i n_{i,j}$ pro případ $Y = j$.

což jsou součty v řádcích a sloupcích tabulky:

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

Za předpokladu $\mathbf{H}_0$ pak jako teoretické četnosti budeme chápat hodnoty $\frac{n_{i,\bullet}\cdot n_{\bullet,j}}{n}$ v této tabulce:

$\frac{n_{i,\bullet}\cdot n_{\bullet,j}}{n}$ $(Y =)\ j$ / $(X =)\ i$	tmavé	světlé	$n_{i,\bullet}$
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
$n_{\bullet,j}$	60	40	

Podmínka na tyto teoretické (tj. očekávané) četnosti ≥ 5 je splněna, takže test nezávislosti můžeme použít. Pro hodnotu testovací statistiky dostaneme

$$t = \sum_{i,j} \frac{\left(n_{i,j} - \frac{n_{i,\bullet}\cdot n_{\bullet,j}}{n}\right)^2}{\frac{n_{i,\bullet}\cdot n_{\bullet,j}}{n}} =$$

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

Tuto hodnotu dále porovnáme s hodnotou kvantilu χ^2 pro $(k-1)(\ell-1)$ stupňů volnosti, kde k je počet položek veličiny X a ℓ je počet položek veličiny Y. Tento počet je nyní jiný, než by byl u "obvyklého" testu dobré shody s $k\cdot\ell$ položkami, protože data jsme použili k odhadu marginálních pravděpodobností.

Kritérium pro **ZAMÍTNUTÍ** $\mathbf{H}_0$ (na hladině α) bude tedy tvaru

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

Hledaný kvantil je

$$\chi^2_{1-\alpha;\ (3-1)(2-1)} = \chi^2_{0.95;\ 2} \doteq 5.992\ .$$

Protože

$$t \doteq 18.056 > 5.992 \doteq \chi^2_{0.95;\ 2}\ ,$$

hypotézu o nezávislosti **ZAMÍTÁME**.

- -

(b) V tomto případě budeme uvažovat pouze veličinu Y a testovat (na hladině $\alpha = 5\%$) hypotézu

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

proti alternativní hypotéze

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

Využijeme teď opět test dobré shody. Celkový počet měření je zase $n = 100$. Naměřené četnosti odpovídají už spočítaným marginálním četnostem pro hodnoty veličiny Y, tedy $n_i = n_{\bullet,i}$. Pro přehlednost si zase vypíšeme tabulku s jednotlivými četnostmi (pozorovanými i teoretickými):

i (barvy vlasů)	tmavé	světlé
n_i (pozorované četnosti)	60	40
p_i (teoretické pravděpodobnosti)	$\frac{1}{2}$	$\frac{1}{2}$
$n\cdot p_i$ (teoretické četnosti)	$100\cdot\frac{1}{2} = 50$	$100\cdot\frac{1}{2} = 50$

Vidíme, že všechny teoretické četnosti jsou ≥ 5, takže můžeme použít asymptotické přiblížení pro testovací statistiku T. Teď už si jen spočítáme hodnotu této statistiky

$$t = \sum_{i=1}^{2} \frac{(n_i - np_i)^2}{np_i} = \frac{(60-50)^2}{50} + \frac{(40-50)^2}{50} = 2 + 2 = 4$$

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

$$t = 4 > 3.84 \doteq \chi^2_{0.95;\,1} = \chi^2_{1-\alpha;\,k-1}$$

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

13.3 *V přímořském středisku probíhá kurz surfování a vodních lyží pro děti. Vybíráme 100 účastníků a sledujeme následující rozdělení sportů mezi chlapce a dívky:*

	surf	*vodní lyže*
chlapci	40	20
dívky	20	20

(a) Testujte na hladině $\alpha = 1\%$, zda je druh sportu nezávislý na tom, zda je zvolen chlapcem nebo dívkou.

(b) Testujte na hladině $\alpha = 5\%$, zda jsou počty chlapců a dívek účastnících se kurzu přibližně stejné.

Řešení:

Nechť veličina X označuje pohlaví daného dítěte a Y druh sportu, který si vybralo.

(a) 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 $\alpha = 5\%$.

Naměřené četnosti $n_{i,j}$ a marginální četnosti $n_{i,\bullet}$ a $n_{\bullet,j}$ pak budou:

$n_{i,j}$ $(Y =)\ j$ $(X =)\ i$	surf	vodní lyže	$n_{i,\bullet}$
chlapci	40	20	60
dívky	20	20	40
$n_{\bullet,j}$	60	40	

Celkový počet dětí je $n = \sum_{i,j} n_{ij} = 40 + 20 + 20 + 20 = 100$. Za předpokladu $\mathbf{H}_0$ pak teoretické četnosti $\frac{n_{i,\bullet} \cdot n_{\bullet,j}}{n}$ v této tabulce budou:

$\frac{n_{i,\bullet}\cdot n_{\bullet,j}}{n}$ $(Y=)\ j$ $(X=)\ i$	surf	vodní lyže	$n_{i,\bullet}$
chlapci	$\frac{60\cdot 60}{100}=36$	$\frac{40\cdot 60}{100}=24$	60
dívky	$\frac{60\cdot 40}{100}=24$	$\frac{40\cdot 40}{100}=16$	40
$n_{\bullet,j}$	60	40	

Podmínka na tyto teoretické (tj. očekávané) četnosti ≥ 5 je splněna, takže test nezávislosti můžeme použít. Pro hodnotu testovací statistiky dostaneme

$$t=\sum_{i,j}\frac{\left(n_{i,j}-\frac{n_{i,\bullet}\cdot n_{\bullet,j}}{n}\right)^2}{\frac{n_{i,\bullet}\cdot n_{\bullet,j}}{n}}=$$

$$=\frac{(40-36)^2}{36}+\frac{(20-24)^2}{24}+\frac{(20-24)^2}{24}+\frac{(20-16)^2}{16}\doteq 2.78\ .$$

Tuto hodnotu dále porovnáme s hodnotou kvantilu χ^2 pro $(k-1)(\ell-1)=1$ stupňů volnosti, kde $k=2$ je počet položek veličiny X a $\ell=2$ je počet položek veličiny Y.

Kritérium pro **ZAMÍTNUTÍ $\mathbf{H}_0$** (na hladině α) bude tedy tvaru

$$t>q_{\chi^2_1(1-\alpha)}\quad\Leftrightarrow\quad \text{zamítáme } \mathbf{H}_0 \text{ (na hladině } \alpha)\ .$$

Hledaný kvantil je

$$q_{\chi^2_1}(1-\alpha)=q_{\chi^2_1}(0.99)=6.63\ .$$

Protože

$$t\doteq 2.78\not> 6.63\doteq q_{\chi^2_1}(0.99)\ ,$$

hypotézu o nezávislosti **NEZAMÍTÁME**.

- -

(b) V tomto případě budeme uvažovat pouze veličinu X a testovat (na hladině $\alpha=5\%$) hypotézu

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

proti alternativní hypotéze

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

Využijeme teď opět test dobré shody. Celkový počet měření je zase $n=100$. Naměřené četnosti odpovídají už spočítaným marginálním četnostem pro hodnoty veličiny X, tedy $n_i=n_{i,\bullet}$. Pro přehlednost si zase vypíšeme tabulku s jednotlivými četnostmi (pozorovanými i teoretickými):

i	chlapci	dívky
n_i (pozorované četnosti)	60	40
p_i (teoretické pravděpodobnosti)	$\frac{1}{2}$	$\frac{1}{2}$
$n\cdot p_i$ (teoretické četnosti)	$100\cdot\frac{1}{2}=50$	$100\cdot\frac{1}{2}=50$

Vidíme, že všechny teoretické četnosti jsou ≥ 5, takže můžeme použít asymptotické přiblížení pro testovací statistiku T. Teď už si jen spočítáme hodnotu této statistiky

$$t=\sum_i\frac{(n_i-np_i)^2}{np_i}=\frac{(60-50)^2}{50}+\frac{(40-50)^2}{50}=2+2=4$$

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

$$t = 4 > 3.84 \doteq q_{\chi^2_1}(0.95) = q_{\chi^2_1}(1-\alpha)$$

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

13.4 *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.

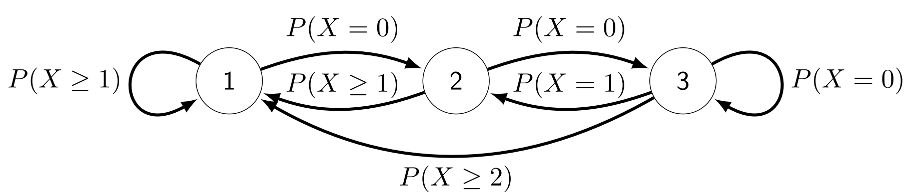
(a) Určete počáteční rozdělení a matici pravděpodobností přechodu.

(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

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

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

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

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

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 $\|\boldsymbol{p}\|_1 := \sum_i |p_i|$.

13.5 *Markovův řetězec má matici přechodu*

$$\begin{pmatrix} 0 & 0 & 0 & 1/5 & 4/5 \\ 0 & 1/2 & 0 & 1/2 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 1/3 & 0 & 2/3 & 0 \\ 1/3 & 0 & 1/3 & 0 & 1/3 \end{pmatrix}$$

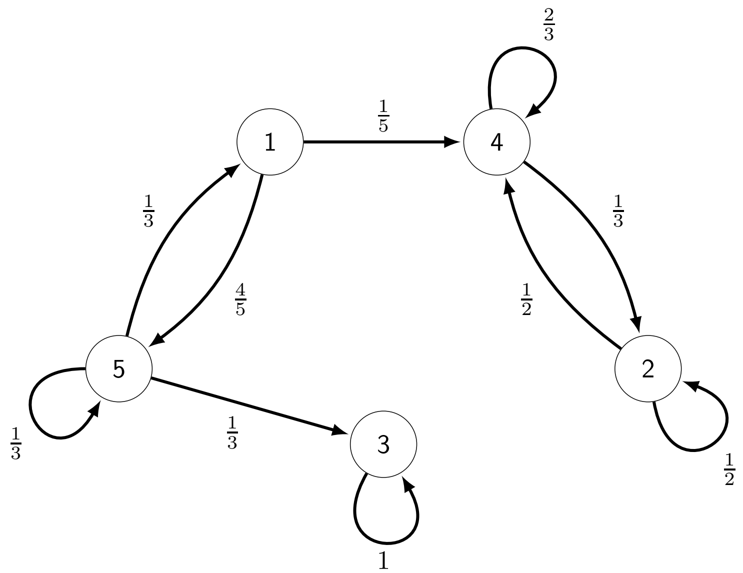
(a) Klasifikujte všechny stavy.

(b) Najděte všechny uzavřené množiny stavů.

(c) Najděte všechna stacionární rozdělení pravděpodobností a posuďte, zda řetězec k některému z nich konverguje.

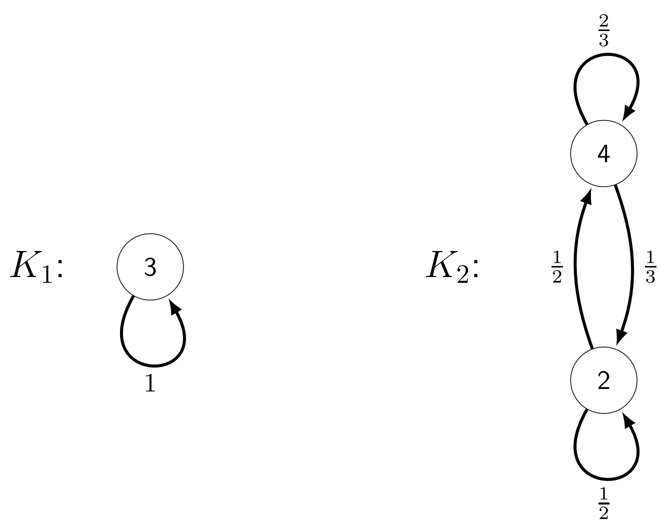
Řešení:

(a) Nakreslíme si příslušný orientovaný graf přiřazený této matici:



Z něj už je snadno vidět, že

- stav 3 je trvalý, dokonce absorpční (tedy tvoří komponentu $K_1 = \{3\}$),
- stavy 2 a 4 jsou trvalé a vzájemně propojené (takže tvoří komponentu $K_2 = \{2, 4\}$) a
- stavy 1 a 5 jsou přechodné.



Obě komponenty mají periodu 1, např. proto, že obsahují smyčku (tj. uzavřenou cestu délky 1) a tedy je $\pi(K_i) = \gcd(\textit{délky uzavřených cest v } K_i) = \gcd(1, \dots) = 1$.

(b) Všechny uzavřené množiny trvalých stavů (tj. množiny trvalých stavů, ze kterých nevedou ven žádné šipky) jsou

$$\emptyset,\ \{3\},\ \{2,4\},\ \{2,3,4\}$$

neboli (kromě prázdné množiny) to jsou vždy sjednocení nějakých komponent.

(c) Podle Věty o stacionárních rozděleních (viz Poznámky k Markovovým řetězcům), jsou všechna stacionární rozdělení na celém řetězci konvexní kombinace (jediných) stacionárních rozdělení příslušných jednotlivým komponentám (tj. takovéto rozdělení je vždy nulové mimo danou komponentu).

Pro $K_1 = \{3\}$ je příslušné stacionární řešení zřejmě $\boldsymbol{p}_1 = (0, 0, 1, 0, 0)$. Pro komponentu $K_2 = \{2, 4\}$ si napíšeme matici přechodů (pro pořadí stavů $2, 4$):

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

a najdeme řešení $\boldsymbol{p} = (p_2, p_4)$ soustavy $(\mathbf{P}^T - \mathbf{I}_2^T) \cdot \boldsymbol{p}^T = 0$ reprezentované maticí

$$\mathbf{P}^T - \mathbf{I}_2^T = \begin{pmatrix} -1/2 & 1/3 \\ 1/2 & -1/3 \end{pmatrix} \sim \cdots \sim \begin{pmatrix} -3 & 2 \\ 0 & 0 \end{pmatrix}.$$

Tato soustava má hodnost 1, tedy má pouze $2 - 1 = 1$ lineárně nezávislých řešení, např. $(2, 3)$. Po jeho "znormování" (tj. vydělení číslem $2 + 3 = 5$) pak dostaneme

$$\boldsymbol{p} = (p_2, p_4) = \left(\frac{2}{5}, \frac{3}{5}\right) \ .$$

Když toto řešení zapíšeme pro původní řetězec s pěti stavy, tak pro komponentu K_2 máme stacionární řešení $\boldsymbol{p}_2 = \left(0, \frac{2}{5}, 0, \frac{3}{5}, 0\right)$.

Každé stacionární řešení pro původní řetězec je tedy tvaru

$$\boldsymbol{p}_{stac} = (1 - t) \cdot \boldsymbol{p}_1 + t \cdot \boldsymbol{p}_2 = \left(0,\ \tfrac{2}{5}t,\ 1 - t,\ \tfrac{3}{5}t,\ 0\right)$$

kde $0 \leq t \leq 1$.

Protože všechny komponenty mají periodu 1, tak libovolné počáteční rozdělení konverguje k nějakému stacionárnímu rozdělení.