

## 12. cvičení z STP

11. - 15. května 2020

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

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

proti alternativní hypotéze:

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

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

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

Máme tedy náhodný vektor

$$\mathbf{N} = (N_1, \dots, N_k)$$

a vztah  $N_1 + \dots + N_k = n$ . Už z toho vidíme, že veličiny  $N_i$  nejsou nezávislé, ale zase k té nezávislosti tak daleko nemají. Náhodný vektor  $\mathbf{N}$  má tzv. multinomické rozdělení a jednotlivá marginální rozdělení veličin jsou binomická, konkrétně  $N_i \sim \text{Bi}(n, p_i)$ . 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  $n \rightarrow \infty$ ) tzv.  $\chi^2$ -rozdělení s  $k - 1$  stupni volnosti. Pro praktické použití této asymptotiky se obvykle požaduje, aby platilo, že

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

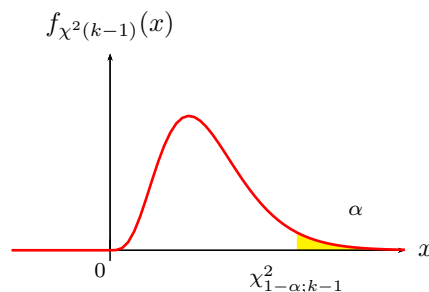
Hodnoty  $n \cdot p_i$  se označují jako tzv. *teoretické četnosti*.

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

$$P_{(\mathbf{H}_0 \text{ platí})} (T > \chi^2_{1-\alpha; k-1}) \doteq \alpha$$

kde  $\chi^2_{1-\alpha; k-1}$  je hodnota kvantilu pro  $\chi^2_{(k-1)}$  rozdělení (viz obrázek, kde  $\alpha$  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  $\alpha$ ) proto volíme jako

$$t > \chi_{1-\alpha; k-1}^2 \Leftrightarrow \text{zamítáme } H_0 \text{ (na hladině } \alpha \text{)} .$$

Z definice chyby 1. druhu, tj.

$$\text{nastává chyba 1. druhu} \Leftrightarrow (\text{hypotéza } H_0 \text{ platí} \ \& \ \text{my ji zamítáme})$$

pak totiž máme, že

$$\begin{aligned} P_{(H_0 \text{ platí})}(\text{nastává chyba 1. druhu}) &= P_{(H_0 \text{ platí})}(\text{zamítáme } H_0 \text{ (na hladině } \alpha \text{)}) = \\ &= P_{(H_0 \text{ platí})}(T > \chi_{1-\alpha; k-1}^2) \doteq \alpha \end{aligned}$$

neboli pravděpodobnost chyby 1. druhu (ovšem za předpokladu platnosti  $H_0$ !) je pak omezena hodnotou  $\alpha$ .

**Příklad 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

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

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

Nejdříve si potřebujeme zjistit, jaké rozdělení

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

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

$$H_0 : \text{veličina } X \text{ má rozdělení s pravděpodobnostmi } (p_1, p_2, p_3) = \left(\frac{1}{2}, \frac{1}{4}, \frac{1}{4}\right),$$

a alternativní hypotéza bude:

$$H_A : \text{veličina } X \text{ má rozdělení jiné než } \left(\frac{1}{2}, \frac{1}{4}, \frac{1}{4}\right).$$

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  $\geq 5$ , takže skutečně můžeme použít asymptotické přiblížení pro testovací statistiku  $T$  (ta tedy bude mít  $\chi^2$ -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  $\chi^2$ -rozdělení s  $k - 1 = 3 - 1 = 2$  stupni volnosti:

$$t = 1 \not\geq 5.99 \doteq \chi_{0.95; 2}^2 = \chi_{1-\alpha; k-1}^2$$

Protože zamítací kritérium NENÍ splněno, tak  $H_0$  **NEZAMÍTÁME** (na hladině  $\alpha$ ).

### Příklad 12.2 (test dobré shody - geometrické rozdělení)

Při házení mincí jsme zaznamenávali, kdy nám poprvé padne líc. Veličina  $X$  měřící počet neúspěchů před prvním úspěchem měla 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  $\text{Geom}(\frac{1}{2})$ , tj.

$$P(X = i) = \left(\frac{1}{2}\right)^{i+1}, \quad i \in \mathbb{N}_0$$

### 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

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

kterou budeme testovat proti alternativní hypotéze:

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

Množina  $K$  všech položek (tj. hodnot) veličiny  $X'$  má tedy velikost  $k = 4$ . Pro ni si (pomocí *poslední* tabulky) spočítáme realizaci testovací statistiky

$$t = \sum_{i \in K} \frac{(n_i - n \cdot p_i)^2}{n \cdot p_i} = \frac{(29 - 32)^2}{32} + \frac{(15 - 16)^2}{16} + \frac{(10 - 8)^2}{8} + \frac{(10 - 8)^2}{8} \doteq 1.34$$

Kritérium pro **ZAMÍTNUTÍ** (viz opět poznámka výše) je tvaru

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

V našem případě je

$$\chi_{1-\alpha; k-1}^2 = \chi_{0.95; 3}^2 \doteq 7.815 .$$

Protože

$$t \doteq 1.34 \not> 7.815 \doteq \chi_{0.95; 3}^2 ,$$

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

*$X$  má geometrické rozdělení s parametrem  $q = 1/2$ ,*

rovněž **NEZAMÍTÁME**.

**Poznámky k testu nezávislosti:** Máme veličiny

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

a chceme otestovat (na hladině  $\alpha$ ), 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  $i = 1, \dots, k$  označme veličiny

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

a opět máme náhodný vektor

$$\mathbf{N} = (N_{1,1}, \dots, N_{k,\ell})$$

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

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

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  $n \rightarrow \infty$ ) opět  $\chi^2$ -rozdělení, tentokrát ale s  $(k-1) \cdot (\ell-1)$  stupni volnosti. Pro praktické použití této asymptotiky se obvykle opět požaduje, aby platilo, že

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

Kritérium pro **ZAMÍTNUTÍ  $H_0$**  (na hladině  $\alpha$ ) volíme podobně jako u testu dobré shody a sice

$$t > \chi_{1-\alpha; (k-1) \cdot (\ell-1)}^2 \Leftrightarrow \text{zamítáme } H_0 \text{ (na hladině } \alpha \text{)} .$$

**Příklad 12.3** Na  $n = 100$  osobách byla pozorována barva očí a vlasů. Naměřeny byly následující sdružené četnosti:

	Vlasy		
Oči		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 světlavlasých.

### Řešení:

Označme si veličiny

$X = \text{“barva očí daného člověka”}$

$Y = \text{“barva vlasů daného člověka”}$

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

(a) Budeme testovat hypotézu:

$H_0$  : rozdělení veličin  $X$  a  $Y$  jsou *nezávislá*

proti alternativní hypotéze:

$H_1$  : 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

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

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

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  $H_0$  pak jako teoretické četnosti budeme chápat hodnoty  $\frac{n_{i,\bullet} \cdot n_{\bullet,j}}{n}$  v této tabulce:

$\frac{n_{i,\bullet} \cdot n_{\bullet,j}}{n}$ ( $X =$ ) $i$ ( $Y =$ ) $j$	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  $\geq 5$  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  $\chi^2$  pro  $(k-1)(\ell-1)$  stupňů volnosti, kde  $k$  je počet položek veličiny  $X$  a  $\ell$  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  $H_0$**  (na hladině  $\alpha$ ) bude tedy tvaru

$$t > \chi_{1-\alpha; (k-1)(\ell-1)}^2 \Leftrightarrow \text{zamítáme } H_0 \text{ (na hladině } \alpha \text{)} .$$

Hledaný kvantil je

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

Protože

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

hypotézu o nezávislosti **ZAMÍTÁME**.

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

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

proti alternativní hypotéze

$$\tilde{H}_A : \text{veličina } Y \text{ má rozdělení } \textit{jiné} \text{ než } \left(\frac{1}{2}, \frac{1}{2}\right).$$

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  $\geq 5$ , 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  $\chi^2$ -rozdělení s  $k - 1 = 2 - 1 = 1$  stupněm volnosti:

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

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

#### Příklad 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  $j$ -tém termínu”) := “zda uspěl/neuspěl”

$Y$  (“účast daného studenta na  $j$ -tém termínu”) :=  $j$

Pravděpodobnost úspěchu v  $j$ -tém termínu je dána jako  $P(X = \text{uspěl} \mid Y = j)$ .

Ukážeme, že tato hodnota bude nezávislá na  $j$ , právě když veličiny  $X$  a  $Y$  budou nezávislé.

**Důkaz:**

( $\Leftarrow$ ): Z nezávislosti  $X$  a  $Y$  ihned máme, že  $P(X = \text{uspěl} \mid Y = j) = P(X = \text{uspěl}) = \text{konst.}$

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

$$P(X = \text{uspěl} \cap Y = j) = c \cdot P(Y = j)$$

a sečtením přes všechna  $j$  dostaneme, že

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

neboli

$$P(X = \text{uspěl} \cap Y = j) = P(X = \text{uspěl}) \cdot P(Y = j).$$

Podobně z  $P(X = \text{neuspěl} \mid Y = j) = 1 - c$  pro všechna  $j$  odvodíme, že

$P(X = \text{neuspěl} \cap Y = j) = P(X = \text{neuspěl}) \cdot P(Y = j)$ . Tedy veličiny  $X$  a  $Y$  jsou nezávislé.

Úloha tedy ve skutečnosti znamená, že 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  $\alpha = 5\%$ .

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

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

kde

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

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



$\frac{n_{i\cdot} \cdot n_{\cdot j}}{n}$ ( $X =$ ) $i$ ( $Y =$ ) $j$	1	2	3	4	5	$n_{i\cdot}$
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
$n_{\cdot j}$	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

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

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

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

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

$$\chi_{1-\alpha; 4}^2 = \chi_{0.95; 4}^2 \doteq 9.49 .$$

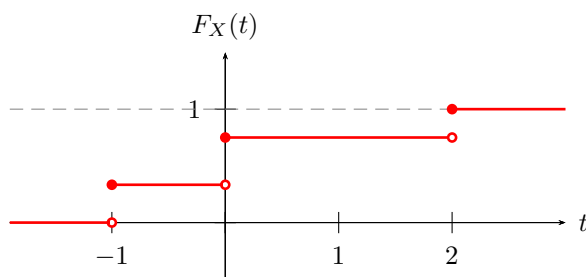
Protože

$$t \doteq 21.68 > 9.49 \doteq \chi_{0.95; 4}^2 ,$$

hypotézu o nezávislosti proto **ZAMÍTÁME**.

### Poznámky ke kvantilům:

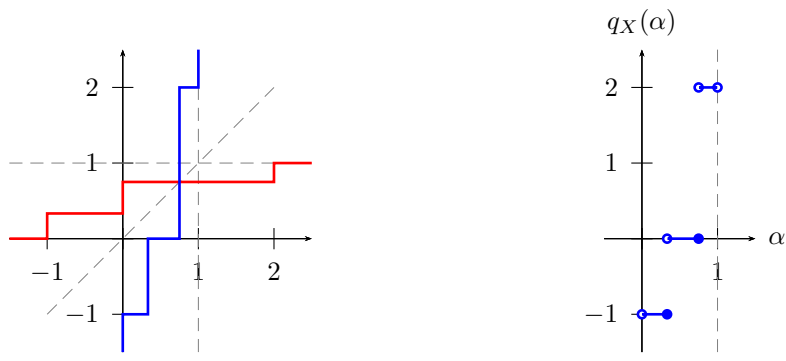
Kvantilovou funkci máme definovanou pro určitá spojitá rozdělení. Pro obecnou náhodnou veličinu  $X$  a pravděpodobnost  $\alpha \in (0, 1)$  bychom ale také potřebovali umět najít  $t \in \mathbb{R}$ , že  $P(X \leq t) = \alpha$ , tj.  $F_X(t) = \alpha$ . Takové  $t$  ovšem obecně nemusí existovat (např. když  $F_X$  má skoky) nebo nemusí být určeno jednoznačně (když  $F_X$  je místy konstantní). Například si vezměme tuto distribuční funkci:



Chtěli bychom tedy ideálně mít inverzní funkci k  $F_X$ , která ale obecně neexistuje. Přesto můžeme něco podobného i zde definovat a to díky tomu, že  $F_X$  je neklesající. Zavedeme si obecně *kvantilovou funkci*  $q_X : (0, 1) \rightarrow \mathbb{R}$  a to následujícím způsobem:

- graf  $F_X$  doplníme na "souvislou čáru", tj. případné skoky funkce  $F_X$  nahradíme spojitou svislou úsečkou,

- tento útvar převrátíme podle osy 1. a 3. kvadrantu (tj. podle přímky "x = y"),
- tam, kde převrácený útvar není funkcí (tj. obsahuje svislé čáry) tyto úseky odstraníme a nahradíme jedinou hodnotou, a sice limitou zleva (a případné krajní úseky v bodech 0 a 1 odstraníme úplně, protože tam se kvantil  $q_X$  nedefinuje)
- výsledným útvarem si pak definujeme graf funkce  $q_X$ .



Jak je tedy vidět, grafy funkcí  $F_X$  a  $q_X$  (po doplnění na souvislé čáry) si budou navzájem zrcadlovými obrazy (vzhledem k ose  $x = y$ ). Takováto definice kvantilu je sice názorná, ale chtělo by to i explicitní popis. Platí:

- $q_X(\alpha) = \min\{t \in \mathbb{R} \mid F_X(t) \geq \alpha\}$  pro všechna  $\alpha \in (0, 1)$
- $P(X \leq q_X(\alpha)) \geq \alpha$  pro všechna  $\alpha \in (0, 1)$
- $P(X < q_X(\alpha)) \leq \alpha$  pro všechna  $\alpha \in (0, 1)$
- $q_X$  je neklesající a zleva spojitá funkce
- Jestliže je  $F_X$  spojité a ostře rostoucí na nějakém intervalu  $J$ , pak  $q_X$  je inverzní funkcí k  $F_X$  na tomto intervalu  $J$ . Tedy speciálně se tato rozšířená definice  $q_X$  shoduje s dřívější definicí ve spojitých případech.

**Poznámky k empirickému rozdělení:**

Nechť  $x_1 \leq \dots \leq x_n$  jsou naměřené hodnoty (veličiny  $X$ ). Pro ně si můžeme přirozeně definovat empirickou náhodnou veličinu Emp s diskrétním rozdělením, oborem hodnot

$$A = \{a \in \mathbb{R} \mid a = x_i \text{ pro nějaké } i\}$$

a jejich pravděpodobnostmi

$$P(\text{Emp} = a) = \frac{\text{"počet výskytů a mezi hodnotami } x_1, \dots, x_n\text{"}}{n}.$$

Když si k této veličině zjistíme distribuční funkci, dostaneme známou empirickou distribuční funkci:

$$F_{\text{Emp}}(t) = P(\text{Emp} \leq t) = \frac{\#\{i \mid x_i \leq t\}}{n}$$

Od ní si pak vytvoříme kvantilovou funkci  $q_{\text{Emp}}$ , která má tvar

$$q_{\text{Emp}}(\alpha) = \min\{t \in \mathbb{R} \mid F_{\text{Emp}}(t) \geq \alpha\} = \min\{x_j \mid F_{\text{Emp}}(x_j) \geq \alpha\}.$$

a nakonec se dá přepsat jako

$$q_{\text{Emp}}(\alpha) = x_{\lceil n\alpha \rceil} \text{ pro } \alpha \in (0, 1)$$

kde  $\lceil u \rceil$  je horní celá část z  $u \in \mathbb{R}$ , tj. zaokrouhlení desetinných čísel nahoru. Speciální hodnoty se pak jmenují

- 1. kvartil =  $q_{\text{Emp}}(\frac{1}{4}) = x_{\lceil \frac{n}{4} \rceil}$
- 2. kvartil =  $q_{\text{Emp}}(\frac{2}{4}) = x_{\lceil \frac{n}{2} \rceil}$  (tzv. medián)
- 3. kvartil =  $q_{\text{Emp}}(\frac{3}{4}) = x_{\lceil \frac{3n}{4} \rceil}$

Podobným způsobem se kvartily definují pro libovolnou veličinu  $X$  (jako  $q_X(\frac{1}{4})$ ,  $q_X(\frac{1}{2})$  a  $q_X(\frac{3}{4})$ ).

Pro libovolnou veličinu  $X$  (a speciálně pro  $X = \text{Emp}$ ) platí:

$$P\left(X \leq 1. \text{ kvartil} \right) \geq \frac{1}{4}$$
$$P\left(1. \text{ kvartil} \leq X \leq 3. \text{ kvartil} \right) \geq \frac{1}{2}.$$
$$P\left(X \geq 3. \text{ kvartil} \right) \geq \frac{1}{4}$$

-----

**Příklad 12.5** Uvažujme následující data:

(1) počty výskytů jistého druhu rostliny na ploše  $1 \text{ m}^2$ :

0, 2, 1, 4, 4, 5, 2, 3, 7

(2) časy (v sekundách) mezi impulzy v mozku:

4.25, 0.65, 1.35, 0.20, 0.55, 6.63, 1.38, 0.22, 0.27

(3) venkovní teploty naměřené v různých letech při pravidelné podzimní akci:

8.07, 19.23, 9.27, 5.71, 12.62, 11.24, 11.92, 17.30, 14.87

Nakreslete pro tato data

- (a) histogramy
- (b) boxploty
- (c) empirickou distribuční funkci

a odhadněte, z jakého rozdělení mohou tato data pocházet.

### Řešení:

Histogram (pro četnosti): Naměřená data si rozdělíme do disjunktních intervalů  $I_i$  (stejně délky) pro  $i = 1, \dots, k$ , které na sebe budou navazovat. Nad  $I_i$  nakreslíme sloupec výšky  $m_i$ , která znamená četnost dat, jež spadnou do  $I_i$ . Abychom z histogramu něco mohli vyčíst a uměli ho (ručně) nakreslit, volíme “rozumný” počet sloupců (např. něco mezi 5 a 15).

Boxplot (neboli krabicový graf): Na rozdíl od histogramu je vždy definován stejně. Krajní vousy (“whiskers”) jsou dány krajními naměřenými hodnotami a krabice (“box”) uprostřed je pak určena hodnotami jednotlivých kvartilů.

Počet měření je zde ve všech případech stejný:  $n = 9$ . Při uspořádaných datech  $x_1 \leq \dots \leq x_9$  tak budou hodnoty kvartilů tyto:

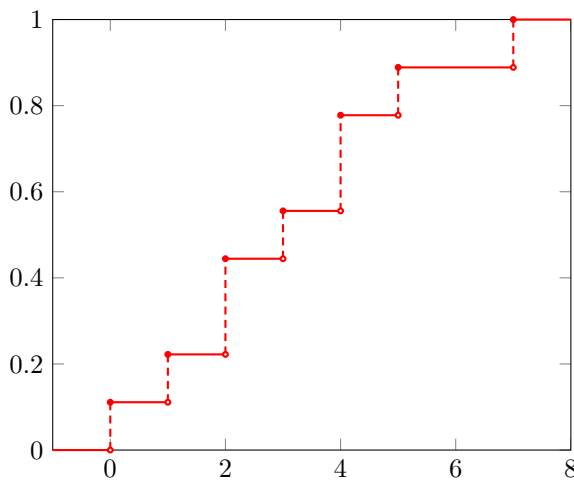
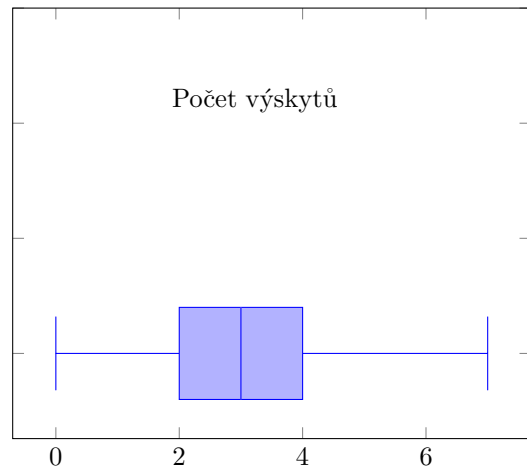
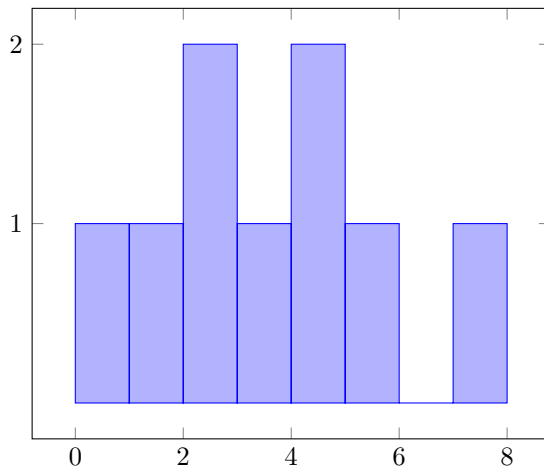
- 1. kvartil =  $x_{\lceil \frac{9}{4} \rceil} = x_3$
- medián =  $x_{\lceil \frac{9}{2} \rceil} = x_5$
- 3. kvartil =  $x_{\lceil \frac{3 \cdot 9}{4} \rceil} = x_7$

Medián je (v rámci uspořádání podle indexu) tedy přibližně uprostřed naměřených hodnot a podobně je to s okolními kvartily. Data si tudíž před výpočtem vždy uspořádáme.

(1) Uspořádaná data:

0, 1, 2, 2, 3, 4, 4, 5, 7  
 $x_1$       1.kvar.      med.      3.kvar.       $x_n$

Rozdíl mezi největší a nejmenší hodnotou je  $x_n - x_1 = 7 - 0 = 7$ . Tuto délku tedy budeme potřebovat pokrýt několika disjunktními intervaly a protože se zde jedná o diskrétní veličinu (počty výskytů), bude vhodné si zvolit šířku sloupce rovnou 1. Intervaly pak budou  $\langle 0, 1 \rangle, \langle 1, 2 \rangle, \dots, \langle 7, 8 \rangle$ .

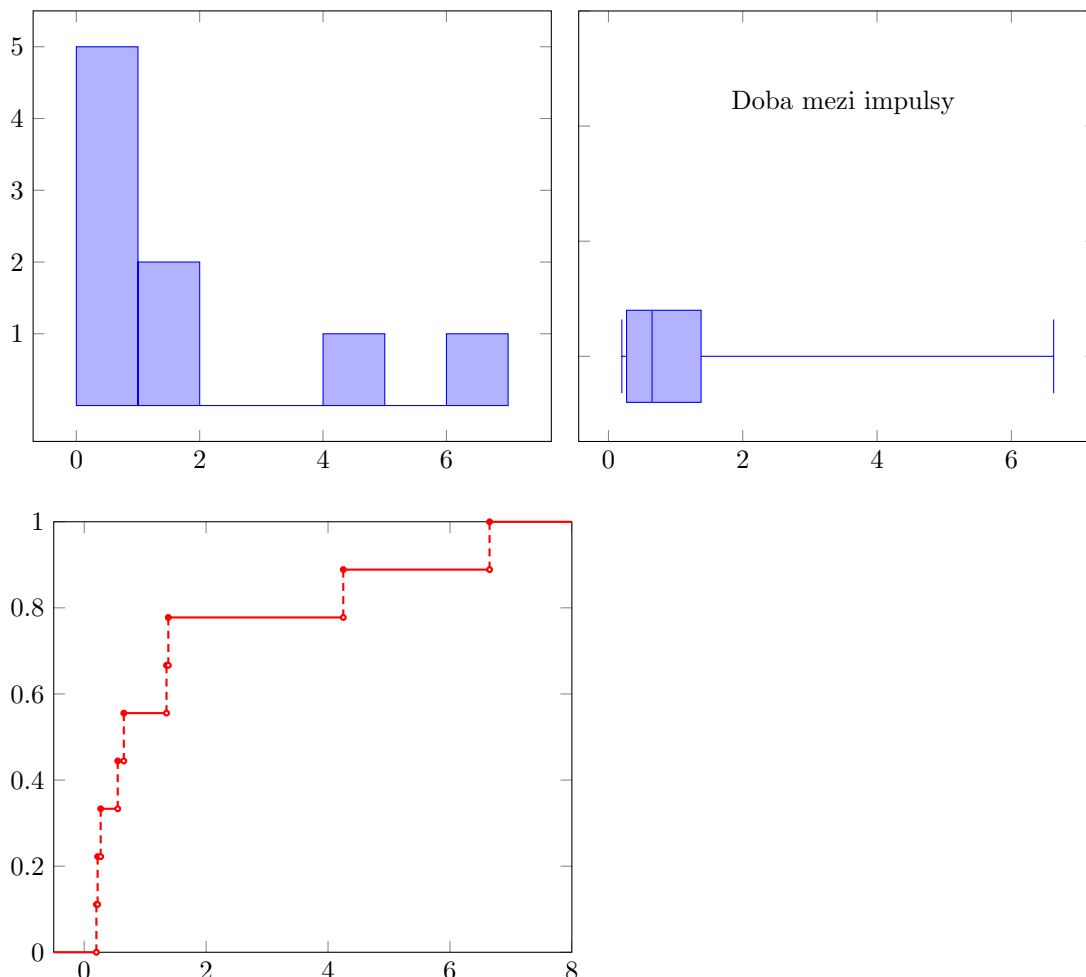


Vzhledem k popisu dat (počty výskytů na dané ploše) to vypadá na Poissonovo rozdělení. Tomu také zhruba odpovídají i grafická znázornění (histogram, boxplot, emp. distr. funkce).

(2) Uspořádaná data:

0.20, 0.22, 0.27, 0.55, 0.65, 1.35, 1.38, 4.25, 6.63  
 $x_1$       1.kvar.      med.      3.kvar.       $x_n$

Rozdíl mezi největší a nejmenší hodnotou je  $x_n - x_1 = 6.63 - 0.2 = 6.42$ . Tuto délku budeme zase potřebovat pokrýt několika disjunktními intervaly. Zkusíme si opět vzít šířku sloupce rovnou 1. Intervaly si pro změnu zvolíme jako  $(0, 1)$ ,  $(1, 2)$ ,  $\dots$ ,  $(6, 7)$ . Výběr toho, do kterého z intervalů přiřadíme dělicí body, není podstatný. Zde jsme si to takto zvolili čistě jen proto, že hodnoty čekací doby jsou vždy nenulové (tj. první interval by ideálně neměl začínat nulou).



Vzhledem k popisu dat (doba čekání na další událost) to vypadá na exponenciální rozdělení. Tomu také zhruba odpovídají i grafická znázornění, kde boxplot je hodně posunutý doleva a empirická distribuční funkce připomíná exponenciálu.

(3) Uspořádaná data:

5.71,	8.07,	9.27,	11.24,	11.92,	12.62,	14.87,	17.30,	19.23
$x_1$		1.kvar.		med.		3.kvar.		$x_n$

Rozdíl mezi největší a nejmenší hodnotou je  $x_n - x_1 = 19.23 - 5.71 = 13.52$  a tuto délku potřebujeme pokrýt několika disjunktními intervaly. Tady se nabízí vzít si větší (ideálně celočíselnou šířku), takže zkusíme šířku sloupce rovnou 2. Intervaly si zvolíme např.  $\langle 4, 6 \rangle$ ,  $\langle 6, 8 \rangle$ ,  $\dots$ ,  $\langle 18, 20 \rangle$ .

