UNIVERSITA PALACKÉHO V OLOMOUCI

PŘÍRODOVĚDECKÁ FAKULTA

KATEDRA MATEMATICKÉ ANALÝZY A APLIKACÍ MATEMATIKY

školní rok 2009/2010

BAKALÁŘSKÁ PRÁCE

Testy dobré shody

Vedoucí diplomové práce:
RNDr. PhDr. Ivo Müller, Ph.D.
Rok odevzdání: 2010

Vypracovala:
Michaela Janebová
ME, 3. ročník

Prohlášení

Prohlašuji, že jsem diplomovou práci zpracovala samostatně pod vedením RNDr. PhDr. Iva Müllera, Ph.D., s použitím uvedené literatury.

V Olomouci dne 10. 4. 2010

.....

Poděkování

Děkuji RNDr. PhDr. Ivu Müllerovi, Ph.D. za čas, který věnoval konzultacím, za všestrannou pomoc, celkovou spolupráci, cenné rady a připomínky, jež mi v průběhu vzniku diplomové práce poskytoval.

OBSAH

Úvod

Tématem této bakalářské práce jsou některé z metod testování statistických hypotéz, a to testy pro ověřování shody typu rozdělení pravděpodobností vybraných dat s některým z teoretických modelů. Obecně jsou tyto testy shrnovány pod název testy dobré shody. Hlavním cílem této práce je nastudování teorie potřebné k provádění testů dobré shody, s podrobným odvozením důkazů, a následná aplikace postupů testování na praktických příkladech. Jednak jde o testy dobré shody použité v případě, kdy známe typ rozdělení, jež ověřujeme i s jeho parametry, a v druhé řadě jde o ověřování některých typů rozdělení, pro něž jen daný typ rozdělení předpokládáme, ale jeho parametry přímo neznáme.

Teoretická část se skládá ze tří kapitol, z nichž první se zabývá základem pro takováto testování, kterým je multinomické rozdělení vycházející z rozdělení binomického. Názorně je pak toto rozdělení demonstrováno na urně s barevnými kuličkami. Všímáme si zde jeho základních vlastností a charakteristik. Především nás zajímá, jak vypadá rozdělení podmíněné a marginální, a z charakteristik je rozebrána střední hodnota a varianční matice. Především je však v této kapitole odvozeno testové kritérium chí-kvadrát. Druhá kapitola této práce se zabývá principy teorie testování při známých parametrech a rozšířením odvozeného testového kritéria chí-kvadrát i pro neznámé parametry. Je zde také uvedena metoda pro odhad těchto parametrů. V kapitole třetí je pak na základě předešlé teorie uveden postup pro ověřování tří typů pravděpodobnostních rozdělení, a to diskrétního rozdělení Poissonova a spojitých rozdělení normálního a exponenciálního.

Teoretické poznatky jsou na závěr využity ve čtvrté části práce, která obsahuje dvě kapitoly, z nichž první se věnuje příkladům, kdy známe rozdělení i parametry a druhá je zaměřena na příklady pro ověřování Poissonova, normálního a exponenciálního rozdělení. Testována jsou jak data reálná, tak i data simulovaná generátorem pseudonáhodných číslic a data záměrně upravená. Přitom bychom měli dospět k výsledkům, které jsou v souladu s námi vyslovenou teorií a prokázat spolehlivost testovacích metod.

1. Multinomické rozdělení

V této kapitole ukážeme, jak je na základě multinomického rozdělení a jeho vlastností odvozeno testové kritérium chí-kvadrát, potřebné k teorii testování pomocí testů dobré shody. Teoretické zázemí je přitom čerpáno z literatury [2] a [3].

1.1 Binomické rozdělení jako základ pro rozdělení multinomické

Multinomické rozdělení zařazujeme mezi jedno z nejdůležitějších diskrétních mnohorozměrných rozdělení. Jeho odvození vychází z rozdělení binomického, které bývá nejčastěji uvažováno na sérii určitého většího počtu nezávislých pokusů, kde v každém pokusu mohou nastat pouze dva výsledky. V tomto případě výsledky označujeme jako úspěch a neúspěch. Pro popis náhodné veličiny s binomickým rozdělením jsou podstatné právě počty úspěchů v mnoha po sobě jdoucích sériích nezávislých pokusů.

Pro binomické rozdělení uvažujeme sérii n nezávislých pokusů spojenou s prostorem $\Omega = \Omega_1 \times \Omega_2 \times \cdots \times \Omega_n$, jehož prvky jsou jevy $\omega = (\omega_1, \omega_2, \cdots, \omega_n)$, kde $\omega_i = \{0,1\}$. Jde o označení úspěchu jako 1 s tím, že $\mathrm{P}(1) = p$, a označení neúspěchu jako 0 s tím, že $\mathrm{P}(0) = q$ a platí $p + q = 1$, tedy $q = 1 - p$. Pro pravděpodobnost platí $\mathrm{P}(\omega) = \prod_{i=1}^{n} \mathrm{P}(\omega_i) = \mathrm{P}(\omega_1)\mathrm{P}(\omega_2) \cdots \mathrm{P}(\omega_n)$.

Je-li v celé sérii pokusů právě m úspěchů, bude mezi ω_i jednička právě m-krát. Pro neúspěchy bude nula v ω_i tudíž $(n - m)$-krát. Pravděpodobnost takového výsledku je dána součinem $\mathrm{P}(\omega) = p^m q^{n-m}$.

K celé sérii takovýchto pokusů připojíme náhodnou veličinu β, která bude vyjadřovat počet úspěchů v celé sérii n pokusů. Náhodná veličina β nabývá hodnot $0,1,2,\cdots,n$ spolu s pravděpodobnostmi $\mathrm{P}(\beta = m) = \sum_{\omega:\, \beta(\omega)=m} \mathrm{P}(\omega) = \binom{n}{m} p^m q^{n-m}$.

Rozdělení veličiny β nazýváme binomické s parametry n a p. Charakteristiky binomického rozdělení střední hodnota a rozptyl jsou dány jako

$$\mathrm{E}\beta = \mathrm{E}b_1 + \mathrm{E}b_2 + \cdots + \mathrm{E}b_n = np,$$

$$\mathrm{var}\beta = \mathrm{var}b_1 + \cdots + \mathrm{var}b_n = npq,$$

kde b_i jsou nezávislé alternativní veličiny. Pro případ multinomického rozdělení uvažujeme podobné série nezávislých pokusů, jaké uvažujeme u rozdělení binomického, přičemž zároveň předpokládáme, že výsledků může nastat více než dva s tím, že pravděpodobnost každého z nich je pokus od pokusu stejná. Pro rozdělení náhodné veličiny s multinomickým rozdělením jsou pak podstatné souhrny jednotlivých výsledků v jednotlivých po sobě jdoucích sériích.

1.2 Model multinomického rozdělení na urně s barevnými kuličkami

Multinomické rozdělení bývá nejčastěji demonstrováno na urně s barevnými kuličkami, z níž vytahujeme několikrát po jedné kuličce s tím, že po každém vytažení kuličku do urny vrátíme. V sérii takto tažených kuliček je pak každá z barev vytažena několikrát. Souhrny kuliček stejné barvy považujeme za veličiny s multinomickým rozdělením. Pro další odvození a důkazy budeme používat níže uvedený pravděpodobnostní model.

Mějme urnu a v ní kuličky k různých barev. Nechť pravděpodobnost vytažení kuličky i-té barvy je rovna $p_i,\ i = 1, 2, \cdots, k$, přičemž

(1) $\quad 0 < p_i < 1, \qquad p_1 + \cdots + p_k = 1.$

Za těchto podmínek n-krát nezávisle na sobě vybereme (s vracením) po jedné kuličce. Označme X_i počet kuliček i-té barvy, které byly takto vybrány. Pak sdružené rozdělení pravděpodobnosti těchto náhodných veličin $X_1, \cdots, X_k$ je dáno vzorcem

(2) $$\mathrm{P}(X_1 = x_1, \cdots, X_k = x_k) = \frac{n!}{x_1! \cdots x_k!}\ p_1^{x_1} \cdots p_k^{x_k}$$

pro $x_i = 0,1, \cdots, n$ a $x_1 + \cdots + x_k = n,$ jinak je tato pravděpodobnost rovna nule.

Rozdělení dané vzorcem (2) se nazývá multinomické s parametry $n, p_1, \cdots, p_k$ a používá se pro něj označení $\mathrm{Mult}(n, p_1, \cdots, p_k)$.

Multinomické rozdělení se svým významem dá částečně přirovnat k mnohorozměrnému normálnímu rozdělení, jemuž se podobá některými svými vlastnostmi.

Nejprve se budeme zajímat o tzv. marginální podmíněné rozdělení některých jeho složek, tedy sledovat jen některé z $X_1, \cdots, X_k$ výše uvažovaných náhodných veličin tohoto rozdělení. Ukážeme, že všechna jeho marginální podmíněná rozdělení jsou opět multinomická a také, že všechna jeho jednorozměrná marginální rozdělení jsou navíc binomická.

1.3 Marginální a podmíněné rozdělení

Věta 1 (marginální rozdělení)

Nechť veličiny $X_1, \cdots, X_k$ mají multinomické rozdělení uvedené ve vzorci (2) a nechť r je celé číslo splňující podmínku $1 < r \leq k$. Pak marginální rozdělení veličin $X_r, \cdots, X_k$ je dáno vzorcem

(3) $\mathrm{P}(X_r = x_r, \cdots, X_k = x_k) =$

$$= \frac{n!}{x_r\,! \cdots\, x_k!(n - x_r - \cdots - x_k)!}\; p_r^{x_r} \cdots p_k^{x_k}(1 - p_r - \cdots - p_k)^{n - x_r - \cdots - x_k}.$$

Důkaz:

Přepišme pro r takové, že $1 < r \leq k$, $(X_1, \cdots, X_k) = (X_1, \cdots, X_{r-1}, X_r, \cdots, X_k)$, kde $X_1 = x_1, \cdots, X_k = x_k$ a $x_1 + \cdots + x_{r-1} + x_r + \cdots + x_k = n$.

Dále uvažujme, že prvních $r - 1$ barev zahrneme v jednu třídu a bude nás zajímat pravděpodobnost, že bylo vytaženo x_r kuliček r-té barvy až x_k kuliček k-té barvy a také $n - (x_r + \cdots + x_k) = n - x_r - \cdots - x_k$ ostatních barev.

Pro hodnoty $x_i = 0, 1, \cdots, n - x_r - \cdots - x_k$, kde $i = 1, \cdots, r-1$, je součet $x_1 + \cdots + x_{r-1} = n - x_r - \cdots - x_k$. Ze základní vlastnosti pro pravděpodobnost $\sum \mathrm{P}(x_i) = 1$ plyne rovnost $p_1 + \cdots + p_{r-1} = 1 - p_r - \cdots - p_k$. Vzorec (3) pak vznikne přímým dosazením těchto výrazů do vzorce (2). ■

Věta 2 (podmíněné rozdělení)

Nechť veličiny $X_1, \cdots, X_k$ mají multinomické rozdělení uvedené ve vzorci (2). Pak podmíněné rozdělení veličin $X_1, \cdots, X_{r-1}$ při pevných $X_r, \cdots, X_k$ je dáno vzorcem

(4) $\mathrm{P}(X_1 = x_1, \cdots, X_{r-1} = x_{r-1} | X_r = x_r, \cdots, X_k = x_k) =$

$$= \frac{(n - x_r - \cdots - x_k)!}{x_1! \cdots x_{r-1}!} \times \prod_{i=1}^{r-1} \left(\frac{p_i}{1 - p_r - \cdots - p_k} \right)^{x_i}.$$

Důkaz:

Pro odvození vztahu pro podmíněné multinomické rozdělení budeme vycházet z věty o podmíněné hustotě ([2], str. 53), která uvažuje dvě konečné σ-aditivní pravděpodobnostní míry, ε konečnou na $(R_r, \mathcal{B}_r)$ a ν konečnou na $(R_{k-r}, \mathcal{B}_{k-r})$, přičemž součinová míra φ konečná na $(R_k, \mathcal{B}_k)$ je dána jako $\varphi = \varepsilon \times \nu$. Dále je dána sdružená hustota $p(\boldsymbol{t}, \boldsymbol{u})$ náhodných vektorů $\boldsymbol{T}$ a $\boldsymbol{U}$ vzhledem k míře φ. Marginální hustota vektoru $\boldsymbol{T}$ je dána vztahem

$$q(\boldsymbol{t}) = \int_{R_{k-r}} p(\boldsymbol{t}, \boldsymbol{u}) d\,\nu(\boldsymbol{u}).$$

Pak podmíněná hustota vektoru $\boldsymbol{U}$ při pevném $\boldsymbol{T} = \boldsymbol{t}$ je rovna

$$r(\boldsymbol{u}|\boldsymbol{t}) = \begin{cases} \boldsymbol{p}(\boldsymbol{t}, \boldsymbol{u}) \,/\, q(\boldsymbol{t}) & \text{pro } q(\boldsymbol{t}) \neq 0, \\ 0 & \text{pro } q(\boldsymbol{t}) = 0. \end{cases}$$

Podmíněnou hustotou náhodného vektoru $\boldsymbol{U}$ při pevném $\boldsymbol{T}$ nazveme takovou nezápornou měřitelnou funkci $r(\boldsymbol{u}|\boldsymbol{t})$, která pro libovolné množiny $B \in \mathcal{B}_r$ a $C \in \mathcal{B}_{k-r}$ splňuje vztah $P(B \times C) = P(\boldsymbol{T} \in B, \boldsymbol{U} \in C) = \int_B \left[\int_C r(\boldsymbol{u}|\boldsymbol{t}) d\nu(\boldsymbol{u}) \right] q(\boldsymbol{t}) d\varepsilon(t)$, kde $q(\boldsymbol{t})$ je marginální hustota vektoru $\boldsymbol{T}$.

Diskrétní náhodný vektor $\boldsymbol{X} = (X_1, , \cdots, X_k)'$ má hustotu vzhledem k čítací míře, kterou můžeme považovat za míru φ. Položme vektory $\boldsymbol{T} = (X_1, , \cdots, X_r)'$ a $\boldsymbol{U} = (X_{r+1}, , \cdots, X_k)'$, kde $1 \leq r \leq k$. Budeme-li ve větě o podmíněné hustotě vektoru $\boldsymbol{U}$ při pevném $\boldsymbol{T}$ uvažovat nenáhodné proměnné $\boldsymbol{u} = (x_1, \cdots, x_{r-1})$ a $\boldsymbol{t} = (x_r, \cdots, x_k)$ a dosadíme-li na pravé straně vztahu pro podmíněnou hustotu za hustoty $r(\boldsymbol{u}|\boldsymbol{t})$, $p(\boldsymbol{t}, \boldsymbol{u})$ a $q(\boldsymbol{t})$ přímo výrazy pro pravděpodobnosti $\mathrm{P}(\boldsymbol{U} = \boldsymbol{u}|\boldsymbol{T} = \boldsymbol{t})$, $\mathrm{P}(\boldsymbol{t}, \boldsymbol{u})$ a $\mathrm{P}(\boldsymbol{T} = \boldsymbol{t}) \neq 0$, dostaneme vyjádření

$$\mathrm{P}(\boldsymbol{U} = \boldsymbol{u}|\boldsymbol{T} = \boldsymbol{t}) = \mathrm{P}(X_1 = x_1, \cdots, X_{r-1} = x_{r-1}|X_r = x_r, \cdots, X_k = x_k) =$$

$$= \frac{\mathrm{P}(\boldsymbol{U}=\boldsymbol{u}|\boldsymbol{T}=\boldsymbol{t})}{\mathrm{P}(\mathbf{T}=\boldsymbol{t})} = \frac{\mathrm{P}(X_1=x_1,\cdots,X_{r-1}=x_{r-1},X_r=x_r,\cdots,X_k=x_k)}{\mathrm{P}(X_r=x_r,\cdots,X_k=x_k)} = \frac{\mathrm{P}(X_1=x_1,\cdots,X_k=x_k)}{\mathrm{P}(X_r=x_r,\cdots,X_k=x_k)}.$$

Do tohoto vztahu dosadíme z výše uvedených vzorců (2) a (3) a upravíme vzniklý podíl použitím rovnosti $x_1 + \cdots + x_{r-1} = n - x_r - \cdots - x_k$ přímo na tvar

$$\frac{\mathrm{P}(X_1=x_1,\cdots,X_k=x_k)}{\mathrm{P}(X_r=x_r,\cdots,X_k=x_k)} =$$

$$= \frac{\frac{n!}{x_1! \cdots x_k!} p_1^{x_1} \cdots p_k^{x_k}}{\frac{n!}{x_r! \cdots x_k!(n-x_r-\cdots-x_k)!} p_r^{x_r} \cdots p_k^{x_k} (1-p_r-\cdots-p_k)^{n-x_r-\cdots-x_k}} =$$

$$= \frac{(n-x_r-\cdots-x_k)!}{x_1! \cdots x_{r-1}!} \cdot \frac{p_1^{x_1} \cdots p_k^{x_{r-1}}}{(1-p_r-\cdots-p_k)^{n-x_r-\cdots-x_k}} =$$

$$= \frac{(n-x_r-\cdots-x_k)!}{x_1! \cdots x_{r-1}!} \times \prod_{\mathrm{i}=1}^{\mathrm{r}-1} \frac{p_i^{x_i}}{(1-p_r-\cdots-p_k)^{x_i}}.$$

Vzorec (4) je takto odvozen dle obecné věty o podmíněné hustotě, přičemž vzorce (2) a (3) jsou také hustoty vzhledem k čítacím mírám. ■

Ve výše uvedených důkazech marginálního a podmíněného rozdělení jsme se zajímali především o poslední veličiny $X_r, \cdots, X_k$. Analogické tvrzení by však také platilo i pro jakoukoliv jinou neprázdnou vlastní podmnožinu multinomicky

rozdělených veličin $X_1, \cdots, X_k$. Z důkazů těchto vět je pak zřetelné, že všechna podmíněná marginální rozdělení jsou opět multinomická.

Na zvoleném modelu je každé tažení kuličky určité barvy náhodnou veličinou právě s jednorozměrným marginálním multinomickým rozdělením. Označíme-li vytažení kuličky určité barvy za úspěch a kterékoli jiné barvy za neúspěch, vidíme rovnou, že takto pojatá jednorozměrná marginální rozdělení jsou navíc přímo binomická s parametry n a p.

1.4 Charakteristiky multinomického rozdělení

Základními charakteristikami multinomického rozdělení jsou vektor středních hodnot $\mathrm{E}\boldsymbol{X} = (\mathrm{E}X_1, \mathrm{E}X_2, \cdots, \mathrm{E}X_k)'$ a varianční matice $\boldsymbol{V} = \mathrm{var}\,\boldsymbol{X}$. Tyto charakteristiky jsou popsány v následující větě.

Věta 3

Nechť veličiny $X_1, \cdots, X_k$ mají multinomické rozdělení dané vzorcem (2). Pak pro jeho charakteristiky platí vztahy:

(5) $\mathrm{E}X_i = np_i, \qquad \mathrm{var}X_i = np_iq_i, \qquad q_i = 1 - p_i, \qquad 1 \le i \le k,$

(6) $\mathrm{cov}(X_i, X_j) = -np_ip_j, \qquad 1 \le i \ne j \le k.$

Důkaz:

Jelikož je multinomické rozdělení odvozeno z rozdělení binomického, jehož střední hodnota je rovna p a rozptyl pq, jsou multinomické charakteristiky v podstatě n-rozměrným rozšířením charakteristik binomických. Tato skutečnost byla již odvozena výše, kde poukázala na to, že jednorozměrné marginální rozdělení X_i je binomické s parametry n a p. Tudíž vztahy v (5) jsou oním n-rozměrným rozšířením.

Pro odvození vzorce (6) zavedeme pomocné náhodné veličiny ξ_{hi} tak, že $\xi_{hi} = 1$, jestliže byla v h-tém tahu vytažena kulička i-té barvy, a $\xi_{hi} = 0$ v případě opačném. Pro takto zavedené veličiny platí vztahy $X_i = \sum_{h=1}^{n} \xi_{hi}$ a $\mathrm{E}\xi_{hi} = p_i$.

Podobně uvažujeme náhodné veličiny ξ_{mj} pro m-tý tah kuličky j-té barvy, takový že $h \neq m$. V případě, že jsou pomocné veličiny ξ_{hi} a ξ_{mj} nezávislé, je jejich kovariance nulová. Navíc je obecně kovariance náhodných veličin definována tehdy, pokud tyto veličiny mají konečné první a druhé momenty. V našem případě jsou tyto momenty dány ve vztazích (5) a kovariance je přímo definována vzorcem ([2], str. 27).

$$\mathrm{cov}(X_i, X_j) = \mathrm{E}\big[(X_i - \mathrm{E}X_i)(X_j - \mathrm{E}X_j)\big] = \mathrm{E}X_iX_j - \mathrm{E}X_i\mathrm{E}X_j \quad \text{pro} \quad 1 \leq i, j \leq k.$$

Podle tohoto vzorce a s užitím námi zavedených pomocných veličin ξ_{hi} a ξ_{mj} a vzorců (5) lze kovarianci upravit takto:

$$\mathrm{cov}(X_i, X_j) = \mathrm{cov}\Big(\sum_{h=1}^{n} \xi_{hi}\ , \sum_{m=1}^{n} \xi_{mj}\Big) = \sum_{h=1}^{n}\sum_{m=1}^{n} \mathrm{cov}(\xi_{hi}, \xi_{mj}) =$$

$$= \sum_{h=1}^{n} \mathrm{cov}(\xi_{hi}, \xi_{hj}) = \sum_{h=1}^{n}\big(\mathrm{E}\xi_{hi}\xi_{hj} - \mathrm{E}\xi_{hi}\mathrm{E}\xi_{hj}\big) = \sum_{h=1}^{n}\big(\mathrm{E}\xi_{hi}\xi_{hj} - p_ip_j\big),$$

tedy

$$\mathrm{cov}(X_i, X_j) = \sum_{h=1}^{n}\big(\mathrm{E}\xi_{hi}\xi_{hj} - p_ip_j\big). \tag{7}$$

Jelikož v libovolném daném pokusu nemůže být vybrána kulička i-té a j-té barvy současně, musí být alespoň jedna z veličin ξ_{hi} a ξ_{hj} rovna nule. Je tedy zřejmé, že platí rovnost $\mathrm{E}\xi_{hi}\xi_{hj} = 0$, a tedy ze vzorce (7) plyne rovnou po dosazení této skutečnosti vzorec (6). ■

1.5 Odvození testových kritérií

V následujících větách se nejprve zaměříme na vlastnosti varianční matice vektoru $\boldsymbol{X}$ a odvodíme tvar pseudoinverze. Poté ukážeme odvození testových kriterií

(výběrových funkcí), která je možno porovnávat s kritickými hodnotami normálního rozdělení a rozdělení chí-kvadrát. Základem bude (výše uvažovaná) náhodná vektorová veličina $\boldsymbol{X}$ mající multinomické rozdělení.

Věta 4 (o hodnosti varianční matice)

Nechť veličiny $X_1, \cdots, X_k$ mají multinomické rozdělení dané vzorcem (2). Označme vektor náhodných veličin $\boldsymbol{X} = (X_1, , \cdots, X_k)'$. Pak varianční matice $\boldsymbol{V} = \text{var}\, \boldsymbol{X}$ má hodnost $k - 1$ a platí

(8) $$\boldsymbol{V}^- = \text{diag}\left\{\frac{1}{np_1}, \cdots, \frac{1}{np_k}\right\}.$$

Důkaz:

Nechť σ_{ij} jsou prvky matice $\boldsymbol{V}$. Ve větě 3 bylo ukázáno, že pro $i = j$ jde o rozptyly (variance) dané vztahem $\sigma_{ii} = np_i(1 - p_i)$. V případě, že $i \neq j$, jsou prvky σ_{ij} kovariance dané vztahem $\sigma_{ij} = -np_ip_j$.

Pro další výpočty označme:

(9) $$\boldsymbol{p} = \left(\sqrt{p_1}, \cdots, \sqrt{p_k}\right)', \qquad \boldsymbol{Q} = \boldsymbol{I} - \boldsymbol{pp}',$$

(10) $$\boldsymbol{D} = \text{diag}\left\{\sqrt{np_1}, \cdots, \sqrt{np_k}\right\}.$$

Nyní ověřme platnost vztahu

(11) $$\boldsymbol{V} = \boldsymbol{DQD}.$$

Prvky matice $\boldsymbol{Q}$ vypadají takto:

$$\boldsymbol{Q} = \begin{pmatrix} 1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & 1 \end{pmatrix} - \begin{pmatrix} \sqrt{p_1p_1} & \cdots & \sqrt{p_1p_k} \\ \vdots & \ddots & \vdots \\ \sqrt{p_kp_1} & \cdots & \sqrt{p_kp_k} \end{pmatrix} = \begin{pmatrix} 1 - p_1 & \cdots & -\sqrt{p_1p_k} \\ \vdots & \ddots & \vdots \\ -\sqrt{p_kp_1} & \cdots & 1 - p_k \end{pmatrix}.$$

Po vynásobení maticí $\boldsymbol{D}$ zleva dostaneme:

$$\boldsymbol{DQ} = \begin{pmatrix} \sqrt{np_1} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \sqrt{np_k} \end{pmatrix} \cdot \begin{pmatrix} 1-p_1 & \cdots & -\sqrt{p_1 p_k} \\ \vdots & \ddots & \vdots \\ -\sqrt{p_k p_1} & \cdots & 1-p_k \end{pmatrix}.$$

Uvažujeme-li η_{ij} za prvky takto vzniklé matice, pak např.:

$$\eta_{11} = \left(\sqrt{np_1} \cdot (1-p_1)\right) + \left(0 \cdot \left(-\sqrt{p_2 p_1}\right)\right) + \cdots + \left(0 \cdot \left(-\sqrt{p_k p_1}\right)\right) = \sqrt{np_1} \cdot (1-p_1),$$

$$\eta_{32} = \left(0 \cdot \left(-\sqrt{p_1 p_2}\right)\right) + \left(0 \cdot (1-p_2)\right) + \left(\sqrt{np_3} \cdot \left(-\sqrt{p_3 p_2}\right)\right) + \cdots + \left(0 \cdot \left(-\sqrt{p_k p_2}\right)\right) =$$

$$= \left(\sqrt{np_3} \cdot \left(-\sqrt{p_3 p_2}\right)\right) = -p_3\sqrt{np_2},$$

přičemž podobně bychom vypočetli ostatní prvky matice. Tedy

$$\boldsymbol{DQ} = \begin{pmatrix} \sqrt{np_1}(1-p_1) & \cdots & -p_1\sqrt{np_k} \\ \vdots & \ddots & \vdots \\ -p_k\sqrt{np_1} & \cdots & \sqrt{np_k}(1-p_k) \end{pmatrix}.$$

Ukážeme, že dalším vynásobením součinu $\boldsymbol{DQ}$ maticí $\boldsymbol{D}$ zprava, tj.

$$(\boldsymbol{DQ})\boldsymbol{D} = \begin{pmatrix} \sqrt{np_1}(1-p_1) & \cdots & -p_1\sqrt{np_k} \\ \vdots & \ddots & \vdots \\ -p_k\sqrt{np_1} & \cdots & \sqrt{np_k}(1-p_k) \end{pmatrix} \cdot \begin{pmatrix} \sqrt{np_1} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \sqrt{np_k} \end{pmatrix},$$

dostaneme přímo varianční matici $\boldsymbol{V}$ s prvky σ_{ij}. Obdobně jako výše, provedeme podobný výpočet na dvou vybraných prvcích. Tedy např.:

$$\sigma_{11} = \left(\left(\sqrt{np_1}(1-p_1)\right) \cdot \sqrt{np_1}\right) + \left(\left(-p_1\sqrt{np_2}\right) \cdot 0\right) + \cdots + \left(\left(-p_1\sqrt{np_k}\right) \cdot 0\right) = np_1(1-p_1),$$

$$\sigma_{32} = \left(\left(-p_3\sqrt{np_1}\right) \cdot 0\right) + \left(\left(-p_3\sqrt{np_2}\right) \cdot \sqrt{np_2}\right) + \left(\left(\sqrt{np_3}(1-p_3)\right) \cdot 0\right) + \cdots +$$

$$+\left((-p_3\sqrt{np_k})\cdot 0\right)=\left((-p_3\sqrt{np_2})\cdot\sqrt{np_2}\right)=-np_2p_3.$$

Odtud je vidět, že

$$\boldsymbol{DQD}=\begin{pmatrix} np_1(1-p_1) & \cdots & -np_1p_k \\ \vdots & \ddots & \vdots \\ -np_kp_1 & \cdots & np_k(1-p_k)\end{pmatrix}=\boldsymbol{V},$$

tedy můžeme říci, že je ověřena platnost vztahu (11).

Dále se zaměříme na některé významné vlastnosti matice $\boldsymbol{Q}$. Je zřejmé, že $\boldsymbol{Q}$ je čtvercová a symetrická. Ukážeme, že je také idempotentní, což znamená, že je splněna vlastnost idempotence $\boldsymbol{Q}=\boldsymbol{Q^2}$. Tedy

$$\boldsymbol{Q^2}=\boldsymbol{Q}\cdot\boldsymbol{Q}=(\boldsymbol{I}-\boldsymbol{pp'})(\boldsymbol{I}-\boldsymbol{pp'})=$$

$$=\begin{pmatrix} 1-p_1 & \cdots & -\sqrt{p_1p_k} \\ \vdots & \ddots & \vdots \\ -\sqrt{p_kp_1} & \cdots & 1-p_k\end{pmatrix}\begin{pmatrix} 1-p_1 & \cdots & -\sqrt{p_1p_k} \\ \vdots & \ddots & \vdots \\ -\sqrt{p_kp_1} & \cdots & 1-p_k\end{pmatrix}=\boldsymbol{Q}$$

Označíme-li si prvky matice $\boldsymbol{Q^2}$ jako α_{ij} a prvky matice $\boldsymbol{Q}$ jako q_{ij}, pak např.:

$$\alpha_{11}=(1-p_1)(1-p_1)+\left(-\sqrt{p_1p_2}\right)\left(-\sqrt{p_2p_1}\right)+\left(-\sqrt{p_1p_3}\right)\left(-\sqrt{p_3p_1}\right)+\cdots+$$

$$+\left(-\sqrt{p_1p_k}\right)\left(-\sqrt{p_kp_1}\right)=1-2p_1+{p_1}^2+p_1p_2+p_1p_3+\cdots+p_1p_k=$$

$$=1-p_1\left(2-(p_1+p_2+p_3+\cdots+p_k)\right)=1-p_1(2-1)=1-p_1=q_{11}$$

a

$$\alpha_{32}=\left(-\sqrt{p_3p_1}\right)\left(-\sqrt{p_1p_2}\right)+\left(-\sqrt{p_3p_2}\right)(1-p_2)+(1-p_3)\left(-\sqrt{p_3p_2}\right)+$$

$$\left(-\sqrt{p_3p_4}\right)\left(-\sqrt{p_4p_2}\right)+\cdots+\left(-\sqrt{p_3p_k}\right)\left(-\sqrt{p_kp_2}\right)=p_1\sqrt{p_3p_2}+(1-p_2)\left(-\sqrt{p_3p_2}\right)+$$

$$+(1-p_3)\left(-\sqrt{p_3p_2}\right)+p_4\sqrt{p_3p_2}+\cdots+p_k\sqrt{p_3p_2}=$$

$$= -\sqrt{p_3 p_2}(-p_1 + 1 - p_2 + 1 - p_3 - p_4 - \cdots - p_k) =$$

$$= -\sqrt{p_3 p_2}\left(2 - (p_1 + \cdots + p_k)\right) = -\sqrt{p_3 p_2} = q_{32}$$

Podobným způsobem s využitím vlastnosti $p_1 + \cdots + p_k = 1 = \boldsymbol{p'p}$ bychom dostali ostatní prvky matice $\boldsymbol{Q^2}$. Můžeme tedy říci, že matice $\boldsymbol{Q}$ je idempotentní.

Dle věty o hodnosti idempotentní matice je hodnost rovna stopě ([3], str. 220). Užitím stejných vlastností a podobných úprav jako v důkazu idempotence matice $\boldsymbol{Q}$ lze zjistit, že i matice $\boldsymbol{I - Q = pp'}$ je idempotentní. Víme, že platí $\boldsymbol{Q^2 = Q}$. Roznásobením matic ověříme idempotenci $\boldsymbol{I - Q}$:

$$\boldsymbol{(I - Q)(I - Q) = I - Q - Q + Q^2 = I - Q - Q + Q = I - Q}.$$

Pak hodnost matice $\boldsymbol{Q}$ můžeme psát jako:

(12) $$\mathrm{h}(\boldsymbol{Q}) = \mathrm{Tr}(\boldsymbol{I - pp'}) = \mathrm{Tr}\,\boldsymbol{I} - \mathrm{Tr}\,\boldsymbol{pp'} = \mathrm{Tr}\,\boldsymbol{I} - \mathrm{h}(\boldsymbol{pp'}) = k - 1.$$

Při ověřování vztahu $\boldsymbol{V = DQD}$ jsme použili násobení regulární maticí $\boldsymbol{D}$ zleva i zprava, čímž se však hodnost matice $\boldsymbol{Q}$ nezměnila, tedy $\mathrm{h}(\boldsymbol{V}) = k - 1$.

Ze vztahu pro idempotenci $\boldsymbol{Q^2 = Q}$ a ze vzorce (11) dostaneme přímo vztah pro pseudoinverzi $\boldsymbol{V^- = D^{-2}}$, neboť

$$\boldsymbol{VV^-V = DQDD^{-2}DQD = DQ^2D = DQD = V}.$$

Pseudoinverzní matici $\boldsymbol{V^-}$ můžeme vyjádřit ve tvaru

$$\boldsymbol{V^- = D^{-2}} = \begin{pmatrix} np_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & np_k \end{pmatrix}^{-1} = \begin{pmatrix} \frac{\mathbf{1}}{np_1} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \frac{\mathbf{1}}{np_k} \end{pmatrix},$$

což je dokazovaný vzorec (8). ■

Už výše bylo zmíněno, že se multinomické rozdělení podobá mnohorozměrnému normálnímu rozdělení. V následujících dvou větách ukážeme, že vhodně transformovaný náhodný vektor, s multinomickým rozdělením, má jednak asymptoticky pro $n \to \infty$ normální rozdělení, a dále, že kvadratická forma tohoto vhodně transformovaného vektoru má navíc asymptoticky rozdělení chí-kvadrát o $k-1$ stupních volnosti.

Věta 5

Nechť náhodný vektor $\boldsymbol{X} = (X_1, , \cdots, X_k)'$ má multinomické rozdělení dané vzorcem (2). Položme

(13) $Y_i = \frac{X_i - np_i}{\sqrt{np_i}}$ pro $i = 1, \cdots, k.$

Pak rozdělení náhodného vektoru $\boldsymbol{Y} = (Y_1, , \cdots, Y_k)'$ pro $n \to \infty$ konverguje ke k-rozměrnému normálnímu rozdělení $\mathrm{N_k}(\mathbf{0}, \boldsymbol{Q})$, kde matice $\boldsymbol{Q}$ je dána vzorcem (9).

Důkaz:

Pro důkaz této věty budeme opět používat pomocné veličiny ξ_{hi} podobně, jako jsme je již použili v důkazu věty 3. Tedy $\xi_{hi} = 1$, jestliže v h-tém tahu byla vybrána kulička i-té barvy, a nebo $\xi_{hi} = 0$ v případě opačném. Dále budeme uvažovat, že náhodné vektory $\xi_h = (\xi_{h1}, \cdots, \xi_{hk})'$ pro $h = 1,2, \cdots, n$ jsou nezávislé a každý z nich má $\mathrm{Mult}(n = 1, p_1, \cdots, p_k)$.

Vektor $\boldsymbol{X}$ je tedy uvažován ve tvaru:

(14) $\boldsymbol{X} = \xi_1 + \xi_2 + \cdots + \xi_n = \sum_{h=1}^{n} \xi_h.$

Podle věty 3 a dle vztahu pro varianční matici v důkazu věty 4 platí pro charakteristiky takovýchto náhodných vektorů vztahy:

(15) $\mathrm{E}\xi_h = \boldsymbol{\mu} = (p_1, \cdots, p_k)'$, $\mathrm{var}\ \xi_h = \boldsymbol{V} = \boldsymbol{D_1 Q D_1}$,

kde $\boldsymbol{D_1} = \mathrm{diag}\{\sqrt{p_1}, \cdots, \sqrt{p_k}\}$,

což je matice ze vzorce (10) pro $n = 1$.

Budeme uvažovat, že ξ_h jsou nezávislé náhodné k-rozměrné vektory se stejným rozdělením, v našem případě multinomickým, jež má střední hodnotu $\boldsymbol{\mu}$ a konečnou varianční matici $\boldsymbol{V}$ danou vzorcem (15). Takovéto náhodné vektory splňují předpoklady mnohorozměrné verze centrální limitní věty ([2], str. 185), podle níž, označíme-li

$$\boldsymbol{\Psi} = \frac{1}{\sqrt{n}}(\xi_1 - \boldsymbol{\mu}) + \cdots + \frac{1}{\sqrt{n}}(\xi_n - \boldsymbol{\mu}),$$

pak náhodný vektor $\boldsymbol{\Psi}$ pro $n \to \infty$ konverguje v distribuci k normálnímu rozdělení $\mathrm{N}_k(\boldsymbol{0}, \boldsymbol{V})$, kde $\boldsymbol{V} = \boldsymbol{D_1 Q D_1}$. Vyjdeme-li ze vztahu (14) pro vektor $\boldsymbol{X}$, můžeme tento vektor dále přepsat do tvaru

$$\boldsymbol{\Psi} = \frac{1}{\sqrt{n}}\sum_{k=1}^{n}(\xi_k - \boldsymbol{\mu}) = \frac{1}{\sqrt{n}}(\boldsymbol{X} - n\boldsymbol{\mu}) = \boldsymbol{D_1 Y}.$$

Pro vektor $\boldsymbol{Y}$ je pak zřejmé, že má také asymptoticky k-rozměrné normální rozdělení s varianční maticí $\mathrm{var}\,\boldsymbol{Y} = \mathrm{var}\,\boldsymbol{D_1}^{-1}\boldsymbol{\Psi} = \boldsymbol{D_1}^{-1}\boldsymbol{D_1 Q D_1 D_1}^{-1} = \boldsymbol{Q}$, tj. $\mathrm{N}_k(\boldsymbol{0}, \boldsymbol{Q})$. Hodnost matice $\boldsymbol{Q}$ se dle vztahu (12) ve větě 4 rovná $k - 1$, z čehož plyne, že rozdělení $\mathrm{N}_k(\boldsymbol{0}, \boldsymbol{Q})$ je singulární. ■

Věta 6

Jestliže $\boldsymbol{X}$ má multinomické rozdělení dané vzorcem (2), pak náhodná veličina

(17) $$\boldsymbol{\chi^2} = \sum_{i=1}^{k}\frac{(X_i - np_i)^2}{np_i},$$

má pro $n \to \infty$ asymptoticky rozdělení χ^2_{k-1}.

Důkaz:

Mějme náhodný vektor $\boldsymbol{Y}$ daný v předchozí větě vzorcem (13), jež má asymptoticky normální rozdělení $\mathrm{N}_k(\boldsymbol{0}, \boldsymbol{Q})$.

Víme, že pokud má náhodná vektorová veličina $\boldsymbol{X}$ normální rozdělení $\mathrm{N}_k(\boldsymbol{\mu}, \boldsymbol{V})$ s varianční maticí, jejíž hodnost je větší nebo rovna jedné, má náhodná veličina $(\boldsymbol{X} - E\boldsymbol{X})' \boldsymbol{V}^{-}(\boldsymbol{X} - E\boldsymbol{X})$ rozdělení chí-kvadrát o tolika stupních volnosti, kolik je hodnost varianční matice, a to při libovolné volbě pseudoinverzní matice $\boldsymbol{V}^{-}$ ([2], str. 79).

Ukažme tedy, že veličina chí-kvadrát je v podstatě dána součtem kvadrátů veličin Y_i, pro které bylo v minulé větě dokázáno, že mají sdružené normální rozdělení $\mathrm{N}_k(\boldsymbol{\mu}, \boldsymbol{V})$, kde $\boldsymbol{\mu} = \boldsymbol{0}$ (díky normování) a $\boldsymbol{V} = \boldsymbol{Q}$. Tedy že veličina $\boldsymbol{Y}' \boldsymbol{Q}^{-} \boldsymbol{Y} = \boldsymbol{Y}'\boldsymbol{Y}$ má rozdělení χ^2_{k-1}.

Vezměme náhodnou veličinu $(\boldsymbol{X} - n\boldsymbol{p})' \boldsymbol{V}^{-}(\boldsymbol{X} - n\boldsymbol{p})$, kde $\boldsymbol{X}$ je multinomický náhodný vektor a $\boldsymbol{V}^{-}$ je pseudoinverzní matice daná ve vztahu (8). Tato veličina má pro zřetelnost tvar:

$$(X_1 - np_1, \cdots, X_k - np_k) \cdot \begin{pmatrix} \frac{1}{np_1} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \frac{1}{np_k} \end{pmatrix} \cdot \begin{pmatrix} X_1 - np_1 \\ \vdots \\ X_k - np_k \end{pmatrix},$$

který lze snadno upravit na součet:

$$= \left(\frac{X_1 - np_1}{np_1}, \cdots, \frac{X_k - np_k}{np_k}\right) \cdot \begin{pmatrix} X_1 - np_1 \\ \vdots \\ X_k - np_k \end{pmatrix} =$$

$$= \frac{(X_1 - np_1)^2}{np_1} + \cdots + \frac{(X_k - np_k)^2}{np_k} = \left(\frac{X_1 - np_1}{\sqrt{np_1}}\right)^2 + \cdots + \left(\frac{X_k - np_k}{\sqrt{np_k}}\right)^2 = \boldsymbol{\chi^2}.$$

Odtud vidíme, že jde o součet kvadrátů asymptoticky normálně rozdělených veličin Y_i daných v předchozí větě vzorcem (13). Za pseudoinverzní matici vezměme matici $\boldsymbol{Q}^{-}$. Tuto volíme neboť $\boldsymbol{Q}$ je idempotentní z čehož $\boldsymbol{Q}^{-} = \boldsymbol{I}$. Tedy platí vztah:

$$\boldsymbol{Y'\,Q^{-}\,Y = Y'Y} = \sum_{i=1}^{k} Y_i^{\,2} = \left(\frac{X_1-np_1}{\sqrt{np_1}},\cdots,\frac{X_k-np_k}{\sqrt{np_k}}\right)\cdot\begin{pmatrix}\frac{X_1-np_1}{\sqrt{np_1}}\\ \vdots \\ \frac{X_k-np_k}{\sqrt{np_k}}\end{pmatrix} = \sum_{i=1}^{k}\frac{(X_i-np_i)^2}{np_i} = \boldsymbol{\chi^2}.$$

Jelikož jsme ve větě 4 ukázali, že hodnost varianční matice $\boldsymbol{Q}$ je $k-1$, můžeme říci, že veličina $\boldsymbol{\chi^2}$ má opravdu při libovolné volbě pseudoinverzní matice rozdělení χ^2_{k-1}.

■

Pro praktické výpočty se dá vzorec (17) upravit na jednodušší tvar:

(18) $$\boldsymbol{\chi^2} = \sum_{i=1}^{k}\frac{X_i^2}{np_i} - n.$$

Při úpravách využijeme vlastnost pro pravděpodobnost $\sum_{i=1}^{k} p_i = 1$ a vlastnost $\sum_{i=1}^{k} X_i = n$ pro multinomické rozdělení. Ve vzorci (17) je navíc rovnou vidět, jakou hodnotou přispívá každý sčítanec k celkovému vzorci $\boldsymbol{\chi^2}$. Tedy:

$$\boldsymbol{\chi^2} = \sum_{i=1}^{k} Y_i^{\,2} = \sum_{i=1}^{k}\left(\frac{X_i-np_i}{\sqrt{np_i}}\right)^2 = \sum_{i=1}^{k}\left(\frac{X_i}{\sqrt{np_i}} - \frac{np_i}{\sqrt{np_i}}\right)^2 =$$

$$= \sum_{i=1}^{k}\left(\frac{X_i^2}{np_i} - 2\frac{X_i}{\sqrt{np_i}}\frac{np_i}{\sqrt{np_i}} + \frac{n^2 p_i^{\,2}}{np_i}\right) = \sum_{i=1}^{k}\frac{X_i^2}{np_i} - 2\sum_{i=1}^{k} X_i + n\sum_{i=1}^{k} p_i =$$

$$= \sum_{i=1}^{k}\frac{X_i^2}{np_i} - 2n + n = \sum_{i=1}^{k}\frac{X_i^2}{np_i} - n.$$

2. Testy dobré shody

Na základě odvození testového kritéria chí-kvadrát, ukážeme v této kapitole, jak jsou prováděny test dobré shody pro známé i neznámé parametry. Přitom v této kapitole zmiňujeme také použití podmínky pro rozsah výběru $np_i \geq 5$. Čerpáno je v této kapitole především z literatury [2],[4],[5],[8] a [9].

2.1 Teorie pro testování hypotéz o shodě při známých parametrech

V této kapitole ukážeme postup, kterým lze ověřit shodu v rozdělení empirických a teoretických veličin. Podle tvrzení věty 6 první kapitoly lze totiž lehce ověřovat, zda jsou teoretické modely v souladu se získanými výsledky (daty). Takováto známá data jsou v našem případě empirické četnosti $X_1, \cdots, X_k$, k nimž předpokládáme existenci četností teoretických $np_1, \cdots, np_k$ definovaných modelem.

Budeme uvažovat, že základní soubor má určité rozdělení. Naším cílem bude pomocí testu s námi zvolenou hladinou významnosti rozhodnout, zda náhodný výběr (data) ze základního souboru má skutečně určitý typ rozdělení. Přičemž podle zákona velkých čísel se empirické rozdělení relativních četností výběru blíží teoretickému pravděpodobnostnímu rozdělení základního souboru, za předpokladu velkého počtu měření.

Výběrový soubor je dán tak, že naměřené hodnoty sledovaného znaku jsou rozčleněny podle velikosti do vhodného počtu tříd a pro tyto třídy známe četnosti empirické. Tímto vzniká multinomické rozdělení, jež jsme podrobněji rozebrali v první kapitole. Testování pak spočívá v tom, že srovnáváme, zdali se tyto empirické četnosti rovnají předpokládaným četnostem teoretickým. Vztahy pro teoretické četnosti jsou přitom stanoveny v kapitole předešlé. Je patrné, že vycházejí z charakteristik právě multinomického rozdělení, a to především ze střední hodnoty.

Vyjdeme z náhodné vektorové veličiny $\boldsymbol{X} = (X_1, , \cdots, X_k)'$ mající multinomické rozdělení. Na základě teoretických četností, jež jsou odvozeny z pravděpodobností, spočítáme hodnotu testového kritéria $\boldsymbol{\chi^2}$

$$\boldsymbol{\chi^2} = \sum_{i=1}^{k} \frac{(X_i - np_i)^2}{np_i} = n \sum_{i=1}^{k} \frac{(X_i/n - p_i)^2}{p_i} = \sum_{i=1}^{k} (Y_i)^2,$$

které je dáno součtem kvadrátů asymptoticky normálně rozdělených náhodných veličin $Y_i = \frac{X_i - np_i}{\sqrt{np_i}}$. Přitom skutečnost, že rozdělení náhodného vektoru $\boldsymbol{Y} = (Y_1, , \cdots, Y_k)'$ asymptoticky konverguje k normálnímu rozdělení byla ukázána v důkazu věty 5.

Pomocí testovací statistiky $\boldsymbol{\chi^2}$ srovnáme empirické a teoretické četnosti v jednotlivých třídách, kde X_i jsou naměřené četnosti výskytu v jednotlivých třídách, k je počet uvažovaných tříd, n je celkový počet pozorování a np_i jsou teoretické četnosti pro jednotlivé třídy.

Při praktickém testování je na teoretické četnosti kladena podmínka, která se týká dostatečného rozsahu výběru. V případě, že je tato podmínka splněna, je přesné rozdělení testového kriteria za platnosti nulové hypotézy dobře aproximováno chí-kvadrát rozdělením. Pro testování je třeba volit rozsah výběru n dostatečně velký, tak aby bylo zajištěno dostatečné obsazení všech tříd, do nichž byl soubor roztřízen tj. $np_i \leq 5$ pro $i = 1, \cdots, k$. Při malém počtu prvků v některé ze tříd je možné tyto třídy slučovat, přitom slučujeme třídy nějak příbuzné, nějak spolu věcně související. Nejčastěji jsou slučovány třídy okrajové a to se sousedními. Přitom slučujeme právě ty třídy, které mají $np_i \leq 5$. Sloučením počtu tříd se mění i počet stupňů volnosti aproximovaného chí-kvadrátu. V praxi je tato podmínka někdy opomíjena uvádí se i jiné podmínky např. jen $np_i \leq 1$. Dodržení podmínky totiž vede na sloučení do menšího počtu tříd, než je potřebný pro provedení testu. Při odstoupení od podmínky slučování pro $np_i \leq 5$, lze test dokončit vždy, otázkou je, jak je rozhodnutí testu důvěryhodné. Obecně je totiž nemožnost testem rozhodnout, při dodržení této podmínky, bráno, jako varovný signál k tomu, že test provádíme při nedostatečném rozsahu výběru. Při takovémto varování je nejlepší variantou daný soubor rozšířit, pokud je to nějakým způsobem možné.

Pokud platí testovaný předpoklad o typu rozdělení a výběr má dostatečný počet prvků, pak podle věty 6 můžeme říci, že testovací statistika $\boldsymbol{\chi^2}$ má asymptoticky rozdělení chí-kvadrát o $k - 1$ stupních volnosti. Nulovou hypotézu, že získaný soubor

má dané rozdělení, zamítáme, pokud $\boldsymbol{\chi^2} \geq \chi^2_{k-1}(\alpha)$. Přitom kritickou hodnotu hledáme jako příslušný $(1-\alpha)$-kvantil rozdělení chí-kvadrát o $k-1$ stupních volnosti. Pokud tedy hodnota $\boldsymbol{\chi^2}$ padne do kritického oboru, zamítáme nulovou hypotézu na hladině významnosti testu $(1-\alpha)$ ve prospěch alternativy, která říká, že výběrový soubor předpokládaný typ rozdělení nemá. V případě opačném, kdy testová statistika nedosáhne kritické hodnoty, nelze tzv. nulovou hypotézu o tom, že výběr pochází z uvažovaného typu rozdělení, zamítnout na hladině testu α. Přitom α (tzv. riziko) je vlastně pravděpodobnost toho, že se mýlíme, když nulovou hypotézu zamítáme.

Předmětem nulové hypotézy v testech dobré shody může být např. rovnoměrnost v porodnosti, mezi pravděpodobností narození děvčete a pravděpodobností narození chlapce, za určité období. Zde je jasné, že rovnoměrnost bude znamenat shodnost pravděpodobností rovnajících se $^1/_2$. Podobným způsobem lze ověřovat rovnoměrnost číslic padlých při větším počtu hodů hrací kostkou nebo rovnoměrnost počtu dětí narozených za měsíce v daném roce. Známým příkladem je také ověřování rovnoměrného zastoupení číslic $0, \cdots, 9$ při náhodném vygenerování určitého počtu číslic, kde rovnoměrností rozumíme, že každá z těchto číslic se generuje s pravděpodobností $^1/_{10}$.

Pro testy dobré shody, kde parametry testovaného rozdělení známe, můžeme přímo spočítat testové kriterium a rozhodnout na základě výše shrnuté teorie testování statistických hypotéz. Aplikace takto vyslovené teorie je ukázána v kapitole 4 na příkladech 1 a 2. V praxi se ovšem také setkáme s případy, že parametry testovaného rozdělení neznáme.

2.2 Rozšíření tesového kriteria chí-kvadrát pro neznámé parametry

V této kapitole se zaměříme na testy dobré shody, kde jsou některé nebo všechny parametry rozdělení neznámé. V praxi tento případ nastává dokonce častěji než případ, že parametry známe. Znamená to, že pravděpodobnosti $p_1, \cdots, p_k$ uvažovaného multinomického rozdělení závisejí na některých neznámých parametrech $a_1, \cdots, a_m$. Ukažme tedy, jak vypadá testové kriterium, jak odhadnout neznámé parametry a jak rozhodnout o hypotézách.

Půjde o to upravit větu 6 tak, aby vyhovovala i pro takovýto případ. Nejprve musíme parametry a_i odhadnout, dle těchto spočítat odhady pro pravděpodobnosti $p_1, \cdots, p_k$, které poté můžeme přímo dosadit do vzorce (17) respektive (18). Rozdělení chí-kvadrát pak bude mít o tolik stupňů volnosti méně, kolik parametrů budeme odhadovat. Tato skutečnost plyne z určité analogie s teorií regrese, jejíž nejznámější metodu nejmenších čtverců použijeme v jisté modifikaci pro odhad našich neznámých parametrů.

Neznámé parametry označme $\boldsymbol{a} = (a_1, \cdots, a_m)'$ a pro pravděpodobnosti předpokládejme, že $p_1 = p_1(\boldsymbol{a}), \cdots, p_k = p_k(\boldsymbol{a})$ jsou dostatečně hladkými funkcemi proměnné $\boldsymbol{a}$ a platí pro ně základní vlastnost

(19) $$p_1(\boldsymbol{a}) + \cdots + p_k(\boldsymbol{a}) = 1.$$

Budeme-li tento vztah derivovat po částech, získáme rovnost, kterou velmi často využijeme v následných úpravách vedoucích k odhadům parametru $\boldsymbol{a}$. Parciálně derivovaná rovnost (19) je tedy tvaru

(20) $$\frac{\partial p_1(\boldsymbol{a})}{\partial a_j} + \cdots + \frac{\partial p_k(\boldsymbol{a})}{\partial a_j} = 0\,, \qquad j = 1, 2, \cdots, m.$$

Nejčastěji využívanou metodou pro odhad $\boldsymbol{a} = (a_1, \cdots, a_m)'$ je metoda minimálního chí-kvadrátu. Jde o obdobu regresní metody nejmenších čtverců, spočívající v tom, že za odhad bereme tu hodnotu $\boldsymbol{a}$, která při daných multinomických veličinách $X_1, \cdots, X_k$ minimalizuje $\boldsymbol{\chi^2}$ ve vzorcích (17) resp. (18). Podobně jako u metody nejmenších čtverců získáme soustavu normálních rovnic pro výpočet odhadu parametru a. Derivujme tedy po částech upravený vzorec testového kriteria (18) dle jednotlivých a_j. Počet normálních rovnic bude přitom roven počtu odhadovaných parametrů. Dostaneme vztah

(21) $$\frac{\partial \boldsymbol{\chi^2}}{\partial a_j} = -\sum_{i=1}^{k} \frac{X_i^2}{n p_i^2(\boldsymbol{a})} \cdot \frac{\partial p_i(\boldsymbol{a})}{\partial a_j} = 0\,, \qquad j = 1, 2, \cdots, m.$$

Jelikož dle důkazu věty 6 víme, že se vzorce (17) a (18) rovnají, je jasné, že podobným derivováním bychom měli dostat soustavu normálních rovnic i ze vztahu (17). Derivujme proto po částech testové kriterium $\boldsymbol{\chi^2} = \sum_{i=1}^{k} \frac{(X_i - np_i(\boldsymbol{a}))^2}{np_i(\boldsymbol{a})}$ podle p_i,

$$\sum_{i=1}^{k} \frac{(-2)n^2(X_i - np_i(\boldsymbol{a}))p_i(\boldsymbol{a}) - n(X_i - np_i(\boldsymbol{a}))^2}{n^2 p_i^2(\boldsymbol{a})} = (-2)\sum_{i=1}^{k} \left\{ \frac{X_i - np_i(\boldsymbol{a})}{p_i(\boldsymbol{a})} + \frac{[X_i - np_i(\boldsymbol{a})]^2}{2np_i^2(\boldsymbol{a})} \right\}.$$

Soustavu normálních rovnic, pak můžeme napsat ve tvaru:

$$(22) \qquad -\frac{1}{2}\frac{\partial \chi^2}{\partial a_j} = \sum_{i=1}^{k} \left\{ \frac{X_i - np_i(\boldsymbol{a})}{p_i(\boldsymbol{a})} + \frac{[X_i - np_i(\boldsymbol{a})]^2}{2np_i^2(\boldsymbol{a})} \right\} \cdot \frac{\partial p_i(\boldsymbol{a})}{\partial a_j} = 0, \qquad j = 1, 2, \cdots, m.$$

Můžeme říci, že soustavy (21) a (22) se sobě rovnají. Budeme dále uvažovat, že druhý členu součtu ve vztahu (22) je limitně pro $n \to \infty$ roven nule, tedy jeho vliv není příliš podstatný, a soustavu (22) můžeme psát ve tvaru, který se od (22) příliš neliší:

$$(23) \qquad \sum_{i=1}^{k} \frac{X_i - np_i(\boldsymbol{a})}{p_i(\boldsymbol{a})} \cdot \frac{\partial p_i(\boldsymbol{a})}{\partial a_j} = 0, \qquad j = 1, 2, \cdots, m.$$

Na základě vlastnosti pro parciální derivování ve vztahu (20) lze soustavu (23) upravit na zjednodušený tvar:

$$(24) \qquad \sum_{i=1}^{k} \frac{X_i}{p_i(\boldsymbol{a})} \cdot \frac{\partial p_i(\boldsymbol{a})}{\partial a_j} = 0, \qquad j = 1, 2, \cdots, m.$$

Řešením rovnic daných ve (24) dostaneme odhad parametru $\boldsymbol{a}$ tzv. modifikovanou metodou minimálního $\boldsymbol{\chi^2}$.

Uvažujeme-li, že jsme schopni odhadnout neznámé parametry touto metodou, tedy v podstatě, že parametry rozdělení již známe, dostáváme se tak k teorii odhadu, která byla ukázána výše pro testy dobré shody při známých parametrech. Můžeme říci, že odhadem parametrů, při dodržení určitých technických předpokladů, se testové kriterium pro chí-kvadrát nepokazí a veličina daná úpravou testového kriteria (17) na vzorec

(25) $\chi^2 = \sum_{i=1}^{k} \frac{\left(X_i - np_i(\boldsymbol{a})\right)^2}{np_i(\boldsymbol{a})}$,

má asymptoticky pro $n \to \infty$ rozdělení chí-kvadrát o $k - m - 1$ stupních volnosti. Přitom m je počet odhadovaných parametrů, tedy na každý odhadovaný parametr odečítáme jeden stupeň volnosti. V podstatě jde o zobecnění věty 6 pro případ, kdy parametry neznáme.([2], str. 197).

3. Ověřování jednotlivých typů rozdělení

Základním cílem všech tří následujících podkapitol bude ověřit, zdali náhodný výběr pochází z určitého známého typu rozdělení. Jde o odvození pravidla, na jehož základě, budeme schopni rozhodnout o statistických hypotézách. Přitom základem je zde testové kritérium chí-kvadrát a teorie testování statistických hypotéz. Za nulovou hypotézu u testů dobré shody, je vždy pokládán původ náhodného výběru z určitého známého druhu rozdělení, přičemž alternativa znamená, že náhodný výběr určitý druh rozdělení nemá. V našem případě jsme si vybrali z diskrétních rozdělení Poissonovo a ze spojitých rozdělení normální a exponenciální. Přitom pro odvození jednotlivých testování v následujících podkapitolách je čerpáno především z literatury [1] a [2].

3.1 Ověřování Poissonova rozdělení

Jako první začneme ověřením Poissonova rozdělení. Budeme uvažovat náhodný výběr pocházející z tohoto diskrétního rozdělení. Omezíme navíc obor hodnot, jež bude takováto náhodná veličina nabývat, pouze na nezáporné a celočíselné hodnoty. Budeme ověřovat nulovou hypotézu, že náhodný výběr pochází právě z Poissonova rozdělení, přičemž parametr tohoto rozdělení bude neznámý.

Uvažujme náhodný výběr $Z_1, \cdots, Z_n$ pocházející z Poissonova rozdělení s parametrem λ. Princip testování spočívá v rozdělení náhodného výběru do tříd podle určitých znaků. Pro takovéto rozdělení si vhodně zvolíme celá čísla $r \geq 0$ a $k \geq 3$. Třídy vytvoříme tak, že do první z nich zařadíme výběrové hodnoty, které jsou menší nebo rovny číslu r. Další budou po řadě tvořeny hodnotami $r+1,\ r+2, \cdots, r+k-2$ a poslední třída bude obsahovat hodnoty větší nebo rovny číslu $r+k-1$. Celkově takto vytvoříme k tříd s tím, že jejich četnosti označíme $X_r,\ X_{r+1}, \cdots,\ X_{r+k-2},\ X_{r+k-1}$.

Pravděpodobnost, že náhodná veličina z původního výběru majícího Poissonovo rozdělení s parametrem λ, nabývá hodnoty i, je dána vztahem:

(26) $\quad q_i = \mathrm{P}(Z_j = i) = \frac{\lambda^i}{i!} e^{-\lambda}$ pro $i = 0,1,2, \cdots$ a $j = 1,2, \cdots, n$

Pro součet četností tříd, po námi provedeném rozdělení, platí vztah

(27) $n = X_r + X_{r+1} + \cdots + X_{r+k-2} + X_{r+k-1},$

tedy součet četností je celkovým rozsahem výběru.

Přitom pravděpodobnosti náhodných veličin námi přerozděleného výběru jsou dány vztahy

(28) $p_r = q_0 + \cdots + q_r, p_{r+1} = q_{r+1}, \cdots, p_{r+k-2} = q_{r+k-2}, p_{r+k-1} = q_{r+k-1} + q_{r+k} + \cdots .$

Toto přerozdělení výběru pomocí zvoleného $r \geq 0$ a $k \geq 3$ je nejčastěji provedeno sloučením okrajových tříd s malými empirickými četnostmi nebo z důvodu dodržování pravidla pro teoretické četnosti $np_i \geq 5$.

Jelikož parametr λ je pro nás neznámý, budeme se ho nyní snažit odhadnout. Pro odhad použijeme metodu minimálního chí-kvadrátu. Derivujeme q_i podle λ:

$$\frac{\partial q_i}{\partial \lambda} = \frac{\partial}{\partial \lambda}\frac{\lambda^i e^{-\lambda}}{i!} = \frac{1}{i!}\left(i\lambda^{i-1}e^{-\lambda} + \lambda^i(-1)e^{-\lambda}\right) = \frac{\lambda^i e^{-\lambda}}{i!}\left(\frac{i}{\lambda} - 1\right) = \left(\frac{i}{\lambda} - 1\right) q_i.$$

Pak soustava normálních rovnic dle (24), rovna $\sum_{i=1}^{k} \frac{X_i}{p_i(\boldsymbol{a})} \cdot \frac{\partial p_i(\boldsymbol{a})}{\partial a_j} = 0$, má pro jeden parametr $\boldsymbol{a} = \lambda$ pouze jednu rovnici. Tuto lze s hledem na vztah (27) rozepsat do postupných součtů takto

$$X_r \frac{\sum_{i=0}^{r}\left(\frac{i}{\lambda}-1\right)q_i}{\sum_{i=0}^{r} q_i} + \sum_{i=r+1}^{r+k-2}\left(\frac{i}{\lambda} - 1\right)X_i + X_{r+k-1}\frac{\sum_{i=r+k-1}^{\infty}\left(\frac{i}{\lambda}-1\right)q_i}{\sum_{i=r+k-1}^{\infty} q_i} = 0,$$

přitom řešením této soustavy je námi odhadovaný parametr λ. Soustavu rozšíříme vynásobením λ a upravíme

$$X_r \frac{\sum_{i=0}^{r}(iq_i - \lambda q_i)}{\sum_{i=0}^{r} q_i} + \sum_{i=r+1}^{r+k-2}(iX_i - \lambda X_i) + X_{r+k-1}\frac{\sum_{i=r+k-1}^{\infty}(iq_i - \lambda q_i)}{\sum_{i=r+k-1}^{\infty} q_i} = 0,$$

$$X_r \frac{\sum_{i=0}^{r} iq_i - \lambda \sum_{i=0}^{r} q_i}{\sum_{i=0}^{r} q_i} + \sum_{i=r+1}^{r+k-2} iX_i - \lambda \sum_{i=r+1}^{r+k-2} X_i + X_{r+k-1} \frac{\sum_{i=r+k-1}^{\infty} iq_i - \lambda \sum_{i=r+k-1}^{\infty} q_i}{\sum_{i=r+k-1}^{\infty} q_i} = 0,$$

$$X_r \frac{\sum_{i=0}^{r} iq_i}{\sum_{i=0}^{r} q_i} - \lambda + \sum_{i=r+1}^{r+k-2} iX_i - \lambda(n-2) + X_{r+k-1} \frac{\sum_{i=r+k-1}^{\infty} iq_i}{\sum_{i=r+k-1}^{\infty} q_i} - \lambda = 0,$$

(29) $$\lambda = \frac{1}{n}\left[X_r \frac{\sum_{i=0}^{r} iq_i}{\sum_{i=0}^{r} q_i} + \sum_{i=r+1}^{r+k-2} iX_i + X_{r+k-1} \frac{\sum_{i=r+k-1}^{\infty} iq_i}{\sum_{i=r+k-1}^{\infty} q_i}\right].$$

Nyní si všimněme, že prostřední člen v součtu (29) se po dělení $1/n$ přibližně rovná výběrovému průměru

(30) $$\bar{X} = \frac{1}{n}\sum_{i=r}^{r+k-1} iX_i,$$

chybí v něm pouze první a poslední člen. Vezmeme-li tento výběrový průměr za námi hledaný odhad parametru λ, je otázkou, o jak přesný odhad parametru jde. Pro upřesnění tohoto odhadu použijeme iterační metodu postupných aproximací λ. Výběrový průměr (30) vezmeme za počáteční aproximaci λ_0, dosadíme jej do vztahu (26) a vypočteme pravděpodobnosti q_i, ty dosadíme do vzorce (29) a vypočteme tak další aproximaci λ_1. Tento postup je opakován až do chvíle, kdy rozdíl mezi dvěma následujícími odhady je menší než zadaná tolerance, kterou si sami určíme.

Vzorec pro testové kriterium je pak dán vztahem

(31) $$\boldsymbol{\chi^2} = \sum_{i=\mathrm{r}}^{r+k-1} \frac{(X_i - np_i)^2}{np_i}$$

a platí-li hypotéza o tom, že výběr pochází z Poissonova rozdělení, má veličina $\boldsymbol{\chi^2}$ asymptoticky chí-kvadrát rozdělení o $k-2$ stupních volnosti. Padne-li pak hodnota testového kriteria do kritického oboru určeného kritickou hodnotou $\chi^2_{k-2}(\alpha)$, zamítáme nulovou hypotézu na hladině významnosti testu α ve prospěch alternativy, že daný náhodný výběr z Poissonova rozdělení nepochází.

Pro ověřování Poissonova rozdělení se používá i jiných testů, uveďme např. ještě test, jež vychází ze skutečnosti, že střední hodnota i rozptyl Poissonova rozdělení

jsou rovny λ. A využívá toho, že výběrový průměr $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} iZ_i$ je nestranným odhadem střední hodnoty a výběrový rozptyl $S^2 = (n-1)^{-1}\sum_{i=1}^{n}(Z_i - \bar{X})^2$ je nestranným odhadem rozptylu. Vzhledem k tomu, že se obě tyto charakteristiky u Poissonova rozdělení rovnají, očekává se, že jejich podíl bude přímo nebo alespoň přibližně roven jedné. Místo takto myšleného podílu se v praxi pro test uvažuje veličina s Poissonovým rozdělením kolísající okolo hodnoty $(n-1)$ daná vztahem $Q = \frac{\sum_{i=1}^{n}(Z_i - \bar{X})^2}{\bar{X}}$. Roli teoretických četností hraje v tomto testu výběrový průměr $\bar{X}$.

3.2 Ověřování normálního rozdělení

V této podkapitole se pro nás stane předmětem hypotéza o výběru z rozdělení normálního. Tedy, že hodnoty náhodného výběru pocházejí z $N(\mu, \sigma^2)$, přičemž parametry μ a σ^2 nejsou známy. Budeme postupovat tak, že hodnoty výběru rozdělíme do intervalů, neboli vytvoříme celkem k tříd $(-\infty, b_1), \langle b_1, b_2), \langle b_2, b_3), \cdots, \langle b_{k-2}, b_{k-1}), \langle b_{k-1}, b_\infty)$, kde $k \geq 4$. Konstanty c_i zavedeme takto:

$$c_1 = b_1 - \frac{b_2 - b_1}{2},\ c_2 = \frac{b_1 + b_2}{2}, \cdots, c_{k-1} = \frac{b_{k-2} + b_{k-1}}{2}, c_k = b_{k-1} + \frac{b_{k-1} - b_{k-2}}{2},$$

přičemž $c_2, \cdots, c_{k-1}$ jsou voleny jako středy tříd a krajní hodnoty jsou voleny tak, aby c_1 ležela v první třídě a měla stejnou vzdálenost k b_1, jako je vzdálenost b_1 k c_2. Krajní hodnotu c_k volíme analogicky. Pro přehlednost označíme i-tou třídu symbolem J_i, kde $i = 1,2, \cdots, k$. Nechť pro výběrové četnosti jednotlivých tříd X_i platí vztah $n = X_1 + \cdots + X_k$ a navíc označme jako $f(x)$ hustotu normálního rozdělení $N(\mu, \sigma^2)$

$$(32) \quad f(x) = \frac{1}{\sigma\sqrt{2\pi}}\, e^{-\frac{(x-\mu)^2}{2\sigma^2}}.$$

Přitom pochází-li výběr z $N(\mu, \sigma^2)$, pak pravděpodobnost, že jednotka padne do i-té třídy je rovna

(33) $p_i = P(x \in J_i) = \int_{J_i} f(x)dx = \int_{J_i} \frac{1}{\sigma\sqrt{2\pi}}\, e^{-\frac{(x-\mu)^2}{2\sigma^2}} dx \quad$ pro $\quad i = 1,2,\cdots,k.$

Pro odhad neznámých parametrů μ a σ^2 použijeme, podobně jako u odhadu parametru z Poissonova rozdělení, modifikovanou metodou minimálního $\boldsymbol{\chi^2}$**.** Je tedy třeba získat soustavu normálních rovnic. Nejprve parciálně derivujme pravděpodobnosti p_i dle jednotlivých parametrů, přičemž víme, že v případě normálního rozdělení je možno derivovat za integrálem. Tedy je splněna podmínka regularity pro prohození derivace a integrálu.

$$\frac{\partial p_i}{\partial \mu} = \frac{1}{\sigma\sqrt{2\pi}} \int_{J_i} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \frac{1}{\sigma^2}(x-\mu)dx = \frac{1}{\sigma^2}\int_{J_i}(x-\mu)f(x)dx,$$

$$\frac{\partial p_i}{\partial \sigma} = \frac{1}{\sqrt{2\pi}} \int_{J_i} \left(-\frac{1}{\sigma^2} e^{-\frac{(x-\mu)^2}{2\sigma^2}} + \frac{1}{\sigma^4} e^{-\frac{(x-\mu)^2}{2\sigma^2}} (x-\mu)^2 \right) dx = \frac{1}{\sigma^3}\int_{J_i}(x-\mu)^2 f(x)dx - \frac{p_i}{\sigma}.$$

Normální rovnice (24) pak mají tvar

$$\sum_{i=1}^{k} \frac{X_i}{p_i} \frac{1}{\sigma^2} \int_{J_i}(x-\mu)f(x)dx = 0,$$

$$\sum_{i=1}^{k} \frac{X_i}{p_i} \left[\frac{1}{\sigma^3} \int_{J_i}(x-\mu)^2 f(x)dx - \frac{p_i}{\sigma} \right] = 0.$$

Pomocí úprav první rovnice s využitím vztahu (33) a vlastnosti $\sum_{i=1}^{k} X_i = n$ dostaneme odhad střední hodnoty:

$$\frac{1}{\sigma^2}\sum_{i=1}^{k} \frac{X_i}{p_i} \int_{J_i} xf(x)dx = \frac{\mu}{\sigma^2}\sum_{i=1}^{k} \frac{X_i}{p_i} \int_{J_i} f(x)dx,$$

$$\sum_{i=1}^{k} \frac{X_i}{p_i} \int_{J_i} xf(x)dx = \mu \sum_{i=1}^{k} \frac{X_i}{p_i} p_i,$$

(34) $$\mu = \frac{1}{n}\sum_{i=1}^{k} \frac{X_i}{p_i} \int_{J_i} xf(x)dx.$$

Z druhé rovnice lze pak odhad rozptylu získat podobně:

$$\frac{1}{\sigma^3}\sum_{i=1}^{k}\frac{X_i}{p_i}\int_{J_i}(x-\mu)^2 f(x)dx = \sum_{i=1}^{k}\frac{X_i}{p_i}\frac{p_i}{\sigma},$$

(35) $$\sigma^2 = \frac{1}{n}\sum_{i=1}^{k}\frac{X_i}{p_i}\int_{J_i}(x-\mu)^2 f(x)dx.$$

Tyto rovnice se řeší iteračně s tím, že za první aproximace volíme

(36) $$\mu_0 = \frac{1}{n}\sum_{i=1}^{k} X_i c_i \quad \text{a} \quad \sigma_0^2 = \frac{1}{n}\sum_{i=1}^{k} X_i {c_i}^2 - \mu_0^2.$$

Spočítáme pravděpodobnosti p_i pro tyto parametry podle vzorce (33) a dosadíme je do pravých stran rovnic (34) a (35) k získání dalších aproximací μ_1 a σ_1^2. Pro výpočet těchto pravých stran si navíc označme body $b_0 = -\infty$ a $b_k = \infty$ a použijeme upraveného vzorce pro pravděpodobnost pomocí distribučních funkcí

(37) $$p_i = \int_{b_{i-1}}^{b_i}\frac{1}{\sigma\sqrt{2\pi}}\, e^{-\frac{(x-\mu)^2}{2\sigma^2}}dx = \left|sub.: y = \frac{x-\mu}{\sigma}, dy = \frac{1}{\sigma}dx\right| =$$

$$\int_{\frac{b_{i-1}-\mu}{\sigma}}^{\frac{b_i-\mu}{\sigma}}\frac{1}{\sqrt{2\pi}}e^{-\frac{y^2}{2}}dy = \Phi\left(\frac{b_i-\mu}{\sigma}\right) - \Phi\left(\frac{b_{i-1}-\mu}{\sigma}\right)$$

a vzorců

$$\int_{J_i} xf(x)dx = \mu p_i + \frac{1}{\sqrt{2\pi}}\left[e^{-\frac{(b_{i-1}-\mu)^2}{2\sigma^2}} - e^{-\frac{(b_i-\mu)^2}{2\sigma^2}}\right],$$

$$\int_{J_i}(x-\mu)^2 f(x)dx = \sigma^2 p_i + \frac{\sigma^2}{\sqrt{2\pi}}\left[\frac{b_{i-1}-\mu}{\sigma}e^{-\frac{(b_{i-1}-\mu)^2}{2\sigma^2}} - \frac{b_i-\mu}{\sigma}e^{-\frac{(b_i-\mu)^2}{2\sigma^2}}\right].$$

Máme-li iteračně vypočteno řešení, tedy známe-li odhady neznámých parametrů μ a σ^2 a jim příslušející pravděpodobnosti p_i, můžeme spočítat hodnotu testového kritéria dle vztahu

(38) $$\boldsymbol{\chi^2} = \sum_{i=1}^{k}\frac{(X_i - np_i)^2}{np_i}$$

a porovnat ji s kritickou hodnotou $\chi^2_{k-3}(\alpha)$. Je-li hodnota testového kritéria chí-kvadrát větší než hodnota kritická, tedy padne-li hodnota chí kvadrát do kritického oboru, zamítáme hypotézu o normálním rozdělení na hladině, která je asymptoticky rovna α.

Pro ověření normálního rozdělení existují i jiné testy. Je to např. test, jež nejprve definuje výběrový obecný i centrální moment k-tého řádu a pomocí nich výběrovou šikmost a špičatost. A jelikož je známo, že výběrová šikmost normálního rozdělení je rovna nule a výběrová špičatost je rovna třem, předpokládá se, že výběrové veličiny budou těmto hodnotám blízké. Jde tedy o testy založené na charakteristikách šikmosti a špičatosti. Zmiňme ještě např. jeden z mnoha testů normality a to, Shapirův-Wilksův test, který používá uspořádaného náhodného výběru z normálního rozdělení $N(\mu, \sigma^2)$, přičemž pro odhad rozptylu $\hat{\sigma}^2$ pro σ^2 se používá speciálních tabulek a testové kritérium je pak založeno na podílu odhadovaného rozptylu a rozptylu výběrového $\hat{\sigma}^2/S^2$.

3.3 Ověřování exponenciálního rozdělení

Pro tuto podkapitolu vezměme jedno z dalších známých spojitých rozdělení, které je typické exponenciálním průběhem své hustoty. Budeme tedy chtít ověřit, zdali daný výběr pochází právě z exponenciálního rozdělení. Vezměme vztah pro vyjádření jeho hustoty

(39) $\quad f(x) = \frac{1}{\lambda}\, e^{-\frac{x}{\lambda}} \quad \text{pro} \quad x > 0,$

kde $\lambda > 0$ je neznámý parametr tohoto rozdělení.

Dále budeme postupovat podobně jako u předešlých ověřování, tedy rozdělíme výběr do k-tříd s krajními body b_i. Nechť délka třídy je rovna $h > 0$, pak položme

$$b_i = ih,$$

kde $i = 0,1,\cdots,k-1$. Položíme-li navíc $b_k = \infty$, vypadá rozdělení na třídy takto:

$$\langle 0, b_1), \langle b_1, b_2), \langle b_2, b_3), \cdots, \langle b_{k-2}, b_{k-1}), \langle b_{k-1}, \infty).$$

Nechť četnosti těchto tříd jsou X_i a $\sum_{i=1}^{k} X_i = n$. Distribuční funkce exponenciálního rozdělení je dána vztahem

$$F(x) = 1 - e^{-\frac{x}{\lambda}}.$$

Tedy za předpokladu exponenciálního rozdělení je pravděpodobnost, že jednotka padne do i-té třídy

$$(40) \quad p_i = F(b_i) - F(b_{i-1}) = 1 - e^{-\frac{b_i}{\lambda}} - 1 + e^{-\frac{b_{i-1}}{\lambda}} = e^{-\frac{b_{i-1}}{\lambda}} - e^{-\frac{b_i}{\lambda}}$$

pro $i = 1,2, \cdots, k.$

Pro odhad parametru použijeme již známou metodu minimálního chí-kvadrátu. Pro odvození budeme formálně uvažovat rovnost $\infty e^{-\infty} = 0$. Derivace je pak rovna

$$\frac{\partial p_i}{\partial \lambda} = \frac{1}{\lambda^2}\left(b_{i-1}\, e^{-\frac{b_{i-1}}{\lambda}} - b_i\, e^{-\frac{b_i}{\lambda}}\right) \quad \text{pro} \quad i = 1,2, \cdots, k\,,$$

a normální rovnice je dle (24) dána vztahem

$$\frac{1}{\lambda^2}\sum_{i=1}^{k} X_i \frac{b_{i-1}\, e^{-\frac{b_{i-1}}{\lambda}} - b_i\, e^{-\frac{b_i}{\lambda}}}{e^{-\frac{b_{i-1}}{\lambda}} - e^{-\frac{b_i}{\lambda}}} = 0.$$

Vynásobíme-li obě strany λ^2 a vytkneme první exponencielu, dostaneme ekvivalentní vyjádření

$$\sum_{i=1}^{k} X_i \frac{b_{i-1} - b_i\, e^{\frac{(b_{i-1}-b_i)}{\lambda}}}{1 - e^{\frac{(b_{i-1}-b_i)}{\lambda}}} = 0.$$

Nyní do tohoto vztahu dosaďme za $b_i = ih$ přičemž $i = 0,1,\cdots,k-1$ a $b_k = \infty$. Součet pak lze přepsat do tvaru

$$\sum_{i=1}^{k-1} X_i \frac{(i-1)h - ihe^{-\frac{h}{\lambda}}}{1-e^{-\frac{h}{\lambda}}} + \frac{X_k(k-1)h-\infty\cdot 0}{1-0} = 0,$$

$$\frac{1}{1-e^{-\frac{h}{\lambda}}} \sum_{i=1}^{k-1} \left[ihX_i \left(1 - e^{-\frac{h}{\lambda}}\right) - hX_i \right] + X_k(k-1)h = 0$$

$$\sum_{i-1}^{k-1} ihX_i - \frac{\sum_{i=1}^{k-1} hX_i}{1-e^{-\frac{h}{\lambda}}} + X_k(k-1)h = 0,$$

$$\sum_{i=1}^{k-1} ihX_i + X_k(k-1)h = \frac{\sum_{i=1}^{k-1} hX_i}{1-e^{-\frac{h}{\lambda}}},$$

a odtud je

$$e^{-\frac{h}{\lambda}} = 1 - \frac{\sum_{i=1}^{k-1} hX_i}{\sum_{i=1}^{k-1} ihX_i + X_k(k-1)h} = \frac{\sum_{i=1}^{k-1}(i-1)hX_i + (k-1)hX_k}{\sum_{i=1}^{k-1} ihX_i + (k-1)hX_k} = \frac{hX_2 + 2hX_3 + \cdots + (k-1)hX_k}{hX_1 + 2hX_2 + \cdots + khX_k - hX_k}.$$

Položíme-li nyní výběrový průměr roven

$$\bar{X} = \frac{hX_1 + 2hX_2 + \cdots + khX_k}{n}, \tag{41}$$

můžeme psát

$$e^{-\frac{h}{\lambda}} = \frac{(n\bar{X} - nh)}{(n\bar{X} - hX_k)}$$

a hodnotu parametru λ najdeme zlogaritmováním obou stran ve tvaru

$$\lambda = -\frac{h}{ln\left(\frac{n\bar{X}-nh}{n\bar{X}-hX_k}\right)}. \tag{42}$$

Při rozdělování hodnot do jednotlivých tříd, volíme třídy tak, aby $^{h}/_{\bar{X}} \ll 1$ a také aby $^{X_k}/_{n} \ll 1$. Parametr λ lze přibližně aproximovat výběrovým průměrem $\bar{X}$, což zjistíme, upravíme-li zlomek ze vzorce (42) na tvar

$$\frac{n\bar{X}-nh}{n\bar{X}-hX_k} = \frac{n\bar{X}\left(1-\frac{nh}{n\bar{X}}\right)}{n\bar{X}\left(1-\frac{hX_k}{n\bar{X}}\right)} = \frac{1-\frac{h}{\bar{X}}}{1-\left(\frac{X_k}{n}\right)\left(\frac{h}{\bar{X}}\right)}.$$

Dále si povšimneme, že jmenovatel tohoto zlomku pro velká n můžeme pominout a psát

$$\frac{1-\frac{h}{\bar{X}}}{1-\left(\frac{X_k}{n}\right)\left(\frac{h}{\bar{X}}\right)} \doteq 1-\frac{h}{\bar{X}}.$$

Dle Taylorova rozvoje pro $\ln y$ je v $y_0 = 0$

$$ln\,(1-y) \doteq ln\,(1-y_0) + \frac{(-1)}{1-y_0}(y-y_0) = 0 - y = -y,$$

pak logaritmu ze vztahu (42) můžeme přepsat do tvaru $ln\left(1-\frac{h}{\bar{X}}\right)$ a provedeme-li substituci $y = \frac{h}{\bar{X}}$ je

$$ln\left(1-\frac{h}{\bar{X}}\right) \doteq -\frac{h}{\bar{X}}.$$

Po dosazení této rovnosti do vztahu (42) dostáváme vztah

(43) $$\lambda \doteq -\frac{h}{\frac{h}{\bar{X}}} = \bar{X}.$$

Za odhad parametru λ bereme tedy výběrový průměr

(44) $$\bar{X} = \frac{1}{n}\sum_{i=1}^{k} ih\,X_i.$$

Hodnotu odhadnutého parametru pak použijeme při výpočtu pravděpodobností daných vztahem (40). Poté můžeme určit hodnotu testového kriteria $\boldsymbol{\chi^2}$ dle vztahu (38) a porovnat ji s kritickou hodnotou $\chi^2_{k-2}(\alpha)$. V případě, že $\boldsymbol{\chi^2} \geq \chi^2_{k-2}(\alpha)$ zamítáme hypotézu o exponenciálním rozdělení na hladině, která je asymptoticky rovna α.

4. Aplikace testů dobré shody v příkladech

Teorie odvozená v minulých dvou kapitolách je následně použita v aplikacích na konkrétní příklady, jde o testy dobré shody pro rozdělení se známými parametry i pro rozdělení, jehož parametry neznáme. Půjde o to rozhodnout pomocí teorie testování statistických hypotéz o tom, zda námi sebraná data, popřípadě data simulovaná, mají či nemají námi předpokládaný známý druh rozdělení pravděpodobnosti. V našem případě rozdělení se známými parametry ověřujeme rozdělení rovnoměrné a rozdělení jenž je tvořeno na základě procentuální skladby sortimentu. Z rozdělení, kde parametry přímo neznáme, je ověřováno rozdělení Poissonovo, normální a exponenciální. Pro výpočty a jednotlivá ověřování je použit postup odvozený a uvedený v předešlých dvou kapitolách. Pro ověření správnosti postupů, pro kontrolu a získání teoretického zázemí bylo využito literatury [2], [6] a [7].

4.1 Testy dobré shody pro známé parametry

Příklad 1: (hrací kostka)

Jako první příklad si pro jednoduchost ověřme hrací kostku. Provedeme 360 hodů hrací kostkou. Přitom počty (četnosti) hození jednotlivých stran kostky a hodnotu testového kriteria udává tabulka 1.

Naším úkolem je rozhodnout, zda je hrací kostka homogenní. Ověřit tedy na hladině významnosti $\alpha = 0{,}05$, zda rozdělení hodnot, které padají na kostce, je rovnoměrné s pravděpodobností $p_i = 1/6 = 0{,}1667$ pro každé celé číslo od jedné do šesti. Teoretické četnosti jsou v tomto případě všechny rovny 60. Předmětem nulové hypotézy je tedy homogenita kostky. Alternativa pak tvrdí, že kostka homogenní není.

Hodnota testového kriteria $\boldsymbol{\chi^2}$ je podle vzorce (17) rovna $6{,}1$. Kritická hodnota kvantilu $\chi^2_{5;0,95}$ je rovna $11{,}070$, a protože zjištěná hodnota testového kriteria je menší než kritická hodnota, tedy nepadne do oboru pro zamítnutí nulové hypotézy ve prospěch

alternativy, nemůžeme na hladině $\alpha = 0{,}05$ nulovou hypotézu zamítnout. Naše kostka je homogenní.

tabulka1: homogenní hrací kostka

strana kostky	**počet hození**	χ^2
1	55	0,4167
2	68	1,0667
3	47	2,8167
4	58	0,0667
5	62	0,0667
6	70	1,6667
∑	**360**	**6,1000**

Hodnota testového kriteria χ^2 je podle vzorce (17) rovna $6{,}1$. Kritická hodnota kvantilu $\chi^2_{5;0,95}$ je rovna $11{,}070$, a protože zjištěná hodnota testového kriteria je menší než kritická hodnota, tedy nepadne do oboru pro zamítnutí nulové hypotézy ve prospěch alternativy, nemůžeme na hladině $\alpha = 0{,}05$ nulovou hypotézu zamítnout. Naše kostka je tedy homogenní.

Uvažujme nyní kostku falešnou, ovlivněnou zátěží, tak aby padala šestka. Takovouto kostkou byly naházeny tři série vždy po sto hodech, tak jak to je zaznamenáno v tabulce 2. Ověřme nyní znovu rovnoměrnost kostky s tím, že předpokládáme, že homogenní spíše nebude. Přitom rozdělení padajících hodnot je opět rovnoměrné s pravděpodobností $p_i = 1/6 = 0{,}1667$. Teoretické četnosti jsou pak každá rovna $16{,}6667$.

tabulka 2: falešná hrací kostka

strana kostky	**1.série**	**2.série**	**3.série**	$\chi^2(\mathbf{1}.)$	$\chi^2(\mathbf{2}.)$	$\chi^2(\mathbf{3}.)$
1	4	10	9	9,6267	2,6667	3,5267
2	8	13	21	4,5067	0,8067	1,1267
3	16	14	17	0,0267	0,4267	0,0067
4	18	15	8	0,1067	0,1667	4,5067
5	26	19	18	0,5352	0,3267	0,1067
6	28	29	27	7,7067	9,1267	6,4067
∑	**100**	**100**	**100**	**22,5085**	**13,5200**	**15,6800**

Ve všech třech případech padne hodnota testového kritéria do kritického oboru, určeného hodnotou $\chi^2_{5;0,95} = 11{,}070$, tedy zamítáme hypotézu o rovnoměrnosti kostky ve prospěch alternativy, že kostka homogenní není. Potvrdil se tedy námi vyslovený předpoklad, který vycházel z toho, že víme o falešnosti kostky. Tedy že vlivem zatížení, padá více šestek, což její homogenitu porušuje.

Příklad 2: (frgály)

Uvažujeme obchodní síť prodejen prodávajících koláče nazývané Valašské frgály. Tyto specializované prodejny, kterých je na Valašsku jen několik, prodávají pět druhů těchto koláčů, a to: tvarohový, makový, povidlový, hruškový a borůvkový (tzv. haferák). Jelikož mají na trhu dlouholetou tradici, je také známa struktura odbytu prodeje jednotlivých druhů (% skladba), která byla obchodníky prozrazena tak, jak je uvedena v tabulce 3 popř. tabulce 4.

Prodej frgálů, kromě prodeje ve specializovaných prodejnách, probíhá také hojně na různých kulturních akcích, týkajících se především valašského folklóru. Pozornost přitom budeme věnovat akcím Masopust a Valašské folkrokování, konané každoročně ve Valašském muzeu v přírodě v Rožnově pod Radhoštěm. Z pohledu prodejce nás bude zajímat, zdali druhová skladba prodeje těchto koláčů, na Masopustu, měla klasickou strukturu prodeje známou z tradice prodejen, přičemž příslušná data a výpočty najdeme v tabulce 3. A navíc nás bude zajímat, zda složení druhů sedmi seti kusů koláčů vybraných na Valašské folkrokování, tak jak je uvedeno v tabulce 4, je v souladu se skladbou tradičního prodeje. V tomto případě je třeba podotknout, že počty kusů jednotlivých druhů jsou voleny na základě očekávaného prodeje, který vychází zásadně z poznatků prodeje v letech minulých.

tabulka 3: Masopust

druh frgálu	**tvarohový**	**makový**	**povidlový**	**hruškový**	**borůvkový**	∑
počet koláčů	99	103	86	161	51	**500**
% skladba	21	20	19	34	6	**100**
pravděpodobnosti	0,21	0,2	0,19	0,34	0,06	**1**
teoretické četnosti	105	100	95	170	30	**500**
χ^2	0,3429	0,0900	0,8526	0,4765	14,7000	**16,4620**

Z tabulky 2 vidíme, že hodnota χ^2 je v případě akce Masopust, konané v lednu letošního roku, rovna $16{,}4620$. Z tabulek je kritická hodnota $\chi^2_{4;0,95}$ dána hodnotou $9{,}488$. Jelikož $9{,}488 < 16{,}4620$ což znamená, že hodnota testového kriteria padne do kritického oboru, zamítáme hypotézu o tom, že je struktura prodeje koláčů na Masopustu stejná jako tradiční struktura, ve prospěch alternativy, že struktura stejná není.

tabulka 4: Valašské folkrokování

druh frgálu	**tvarohový**	**makový**	**povidlový**	**hruškový**	**borůvkový**	∑
počet koláčů	150	140	140	220	50	**700**
% skladba	21	20	19	34	6	**100**
pravděpodobnosti	0,21	0,2	0,19	0,34	0,06	**1**
teoretické četnosti	147	140	133	238	42	**700**
χ^2	0,0612	0,0000	0,3684	1,3613	1,5238	**3,3148**

Z hodnot tabulky 4 lze říci, že hodnota χ^2 je v případě akce Valašské folkrokování rovna $3{,}3148$. Přitom při porovnání s kritickou hodnotou $9{,}488$ můžeme říci, že hodnota testového kriteria neleží v kritické oblasti pro zamítnutí nulové hypotézy. Tedy hypotézu o stejné struktuře prodeje frgálů na akci Valašské folkrokování nelze zamítnout a můžeme říci, že takovýto prodej má strukturu shodnou se strukturou tradičního prodeje specializovaných prodejen.

V prvních dvou řešených příkladech jsme ukázali aplikaci teorie odvozené v prvních dvou kapitolách. V jednoduchém příkladě s hrací kostkou bylo poukázáno na vznik teoretických četností při rovnoměrném rozdělení pravděpodobností (po šesti). V příkladě s frgály byly teoretické četnosti dány z procentuální skladby sortimentu.

4.2 Testy dobré shody pro neznámé parametry

Příklad 3: (hokejová extraliga)

V tomto příkladě se zaměříme na aplikaci teorie pro ověření Poissonova rozdělení uvedenou v podkapitole 3.1. Předmětem našeho zájmu budou nyní data z hokejových sportovních utkání. Zkusme ověřit, zda počet branek, vstřelených v jednotlivých zápasech extraligy, se řídí Poissonovým rozdělením. Máme k dispozici

výsledky 279 hokejových utkání, sehraných v základní části Slovnaft extraligy 2008/2009. Utkání roztřídíme podle počtu vstřelených branek, přičemž empirické četnosti X_i námi vytvořených čtrnácti tříd jsou dány v tabulce 5.

Dle teorie odvozené pro ověření Poissonova rozdělení uvedené v podkapitole 3.1 je nejprve potřeba se rozhodnout, zdali se budeme snažit dodržet podmínku pro rozsah výběru, podle níž musí pro teoretické četnosti platit $np_i \geq 5$, což většinou vede ke slučování málo četných tříd. Jelikož jsou navíc pravděpodobnosti Poissonova rozdělení definovány do nekonečna, je třeba slučovat vždy některé třídy poslední, a to právě ty, jejichž pravděpodobnosti jsou velmi malé až nulové. V tomto příkladě budeme na podmínku pro rozsah výběru dbát.

tabulka 5: hokejová extraliga

	počet vstřelených branek i	**empirická četnost utkání X_i**	**pravděpo-dobnosti q_i**	**empirická četnost tříd X_i**	**pravděpo-dobnosti p_i**	**teoretické četnosti np_i**	**testové kritérium chí-kvadrát**
	0	0	0,0020				
	1	4	0,0122				
r	2	4	0,0381	8	0,0523	14,5926	2,9784
	3	30	0,0792	30	0,0792	22,0974	2,8262
	4	22	0,1234	**22**	0,1234	34,4393	**4,4930**
	5	62	0,1539	**62**	0,1539	42,9400	**8,4607**
	6	24	0,1599	**24**	0,1599	44,6148	**9,5253**
	7	55	0,1424	**55**	0,1424	39,7333	**5,8660**
	8	32	0,1110	32	0,1110	30,9626	0,0348
	9	24	0,0769	24	0,0769	21,4471	0,3039
	10	13	0,0479	13	0,0479	13,3703	0,0103
	11	5	0,0272	9	0,0531	14,8029	2,2748
	12	1	0,0141				
	13	0	0,0068				
	14	3	0,0030				
	15		0,0013				
	16		0,0005				
	17		0,0002				
	18		0,0000				
	19		0				
	20		0				
∑		**279**	**1**	**279**	**1**	**279**	**36,7733**

Nejprve vezmeme četnosti všech patnácti tříd. Přitom třídy nijak neslučujeme, volíme $r = 0$ a $k = 15$. Pro odhad parametru použijeme výběrový průměr daný vzorcem (30), jehož hodnota je $\lambda = 6{,}2366$. Dále je třeba spočítat pravděpodobnosti Poissonova rozdělení q_i, pro které platí vzorec (26), přitom si tyto pravděpodobnosti spočítáme alespoň pro $i = 20$. Z tabulky 5 je pak vidět, že od $i = 18$ jsou tyto pravděpodobnosti nulové. Jelikož víme, že pravděpodobnosti p_i jsou z pravděpodobností q_i tvořeny dle vztahu (28), který vychází právě z onoho slučování, překontrolujeme nyní, zda všechny $nq_i \geq 5$. Zjistíme, že teoretické četnosti nq_i nevyhovují naší podmínce v případě prvních dvou tříd a posledních tříd braných od $i = 12$ do nekonečna. Proto tyto sloučíme s nejbližšími sousedními třídami.

Sloučením dodržíme naši podmínku pro rozsah výběru $n = 279$, přičemž je nyní $r = 2$ a počet tříd se zmenšil na $k = 10$. Hodnoty empirických četností po sloučení udává tabulka 5. Pro takto sloučené je třeba odhadnout znovu neznámý parametr λ. Dle vzorce (30) vezmeme za počáteční aproximaci, uvedené iterační metody odhadu neznámého parametru λ, výběrový průměr, jehož hodnota je $\lambda_0 = 6{,}2151$. Další aproximace dostaneme tak, že dle vztahu (26) spočteme pravděpodobnosti q_i odpovídající tomuto λ_0, po jejichž dosazení do vzorce (29) dostáváme další aproximaci $\lambda_1 = 6{,}2085$. Stejným způsobem počítáme aproximace další, až do té doby, než se budou rovnat na čtyři desetinná místa, což jsme si určili pro ukončení iterační metody. Takto dostáváme $\lambda_2 = 6{,}2339$, $\lambda_3 = 6{,}2341$ a $\lambda_4 = 6{,}2341$. Jelikož třetí a čtvrtá aproximace se již neliší, bereme za parametr aproximaci $\lambda_3 = 6{,}2341$ a tomuto parametru příslušné pravděpodobnosti q_i. Na základě pravděpodobností q_i dostaneme dle vztahu (28) pravděpodobnosti p_i. Spočítáme-li nyní teoretické četnosti np_i zjistíme, že jsme volbou $r = 2$ a $k = 10$, při níž došlo ke sloučení prvních tří tříd a čtyř tříd posledních, což je také zřetelné z tabulky 5, dodrželi podmínku pro dostatečný rozsah výběru, tedy pro všechny teoretické četnosti platí $np_i \geq 5$. Pro sloučených deset tříd je pak hodnota testového kritéria $\boldsymbol{\chi^2}$ dle vzorce (31) rovna 36,7733. Tu porovnáváme s hodnotou kvantilu chí-kvadrát $\chi^2_{8;0,95} = 15{,}5070$ a jelikož hodnota testového kritéria padne do kritického oboru vymezeného tímto kvantilem, zamítáme nulovou hypotézu ve prospěch alternativy, tedy počty vstřelených branek v 279-ti zápasech Poissonovo rozdělení nemají. Tyto data Poissonovo rozdělení nemají zřejmě z důvodu četností třetí až šesté třídy (uvažovaných po sloučení empirických četností do deseti tříd), kde jim

příslušející sčítance pro testové kriterium chí-kvadrát dosahují největších hodnot a narušují tak zřejmě pravidelný tvar Poissonova rozdělení. V tabulce 5 je tato skutečnost vyznačena tučně.

Pokud bychom si při odhadu parametru v tomto příkladě stanovili podmínku pro ukončení iterační metody, jako rovnost aproximací na jedno desetinné místo, mohli bychom za odhad parametru brát již počáteční aproximaci $\lambda_0 = 6{,}2151$ rovnu výběrovému průměru dle vzorce (30). Stejným způsobem jako je popsáno výše, dospějeme k hodnotě testového kritéria $36{,}7588$, která se na jedno desetinné místo rovná hodnotě $36{,}7733$. Rozhodnutí dle kvantilu $\chi^2_{8;0,95} = 15{,}507$ vede rovněž k zamítnutí nulové hypotézy. Je tedy zřejmé, že pokud vezmeme za odhad parametru počáteční aproximaci, danou výběrovým průměrem ze vzorce (30), nedopouštíme se při našem rozsahu výběru $n = 279$ chyby, která by vedla ke špatnému rozhodnutí. V našem případě se hodnoty testových kriterií dokonce při zaokrouhlení na jedno desetinné místo rovnají. Můžeme říci, že vylepšení parametru iterační metodou, nebylo pro rozhodnutí testu příliš podstatné.

Nyní ještě uvažujme o vypuštění podmínky pro dostatečnost rozsahu výběru z důvodu, že se nám náš výběr zdá dosti velký. V tomto případě žádné s četností patnácti tříd neslučujeme, za odhad parametru vezmeme výběrový průměr dle (30) roven $\lambda = 6{,}2366$ a pravděpodobnosti p_i se budou rovnat q_i počítaných dle vzorce (26) s tím, že q_i pro $i = 14$ bude rovna součtu všech q_i pro $i = 14, \cdots$. V tomto případě je hodnota testového kritéria rovna $43{,}086$ a dle kvantilu $\chi^2_{13;0,95} = 22{,}362$ zamítáme nulovou hypotézu o původu z Poissonova rozdělení.

Pro porovnání zkusme nyní otestovat data vygenerovaná generátorem pseudonáhodných číslic. Půjde o výběr o rozsahu $n = 279$, jež má Poissonovo rozdělení s parametrem $\lambda = 6{,}2$. Tato data s daným rozsahem a parametrem byla generována proto, aby korespondovala s daty reálnými, se kterými jsme již počítali. Přitom ale nyní víme, že data Poissonovo rozdělení mají. Četnosti takto vytvořeného výběru jsou uvedeny v tabulce 6 v její levé části. Prováděný test by měl dojít k závěru, že data generovaná opravdu z Poissonova rozdělení pocházejí. Přitom budeme dbát na podmínku pro rozsah výběru. Odhadneme parametr výběrovým průměrem dle (30).

Pro $r = 0$ a $k = 15$ je roven $\lambda = 6{,}2652$. Při výpočtu pravděpodobností dle (26) dojdeme ke stejnému sloučení tříd jako u příkladu hokejové extraligy. Pro výpočet si tedy stanovíme podobně jako v předchozím příkladě $r = 2$ a $k = 10$, čímž vyhovíme podmínce $nq_i \geq 5$.

tabulka 6: simulovaná hokejová extraliga

Poissonovo rozdělení					**záměrně porušená data**				
	počet vstřelených branek i	**empirická četnost X_i**	**empirická četnost tříd X_i**	**chí-kvadrát**		**počet vstřelených branek i**	**empirická četnost X_i**	**empirická četnost tříd X_i**	**chí-kvadrát**
	0	1				0	1		
	1	3			r	1	3	4	2,4534
r	2	6	10	1,4517		2	**60**	60	**78,9256**
	3	26	26	0,8206		3	26	26	2,6580
	4	32	32	0,1280		4	32	42	6,6009
	5	56	56	4,1453		5	56	48	0,0029
	6	39	39	0,6948		6	39	40	3,6118
	7	32	32	1,5533		7	32	30	2,5431
	8	27	27	0,5680		8	27	18	0,2561
	9	24	24	0,2399		9	24	12	1,5837
	10	20	20	3,0111		10	20	2	8,7666
	11	6	13	0,2162		11	6	9	1,2204
	12	2				12	2		
	13	3				13	3		
	14	2				14	2		
∑		**279**	**279**	**12,8288**			**333**	**333**	**109,4769**

Jako odhad parametru λ bereme přímo počáteční aproximaci dle vzorce (30) rovnu $\lambda_0 = 6{,}233$, neboť iterační metoda na upřesnění parametru nemá v tomto případě podstatný vliv na hodnotu testového kritéria ani na rozhodnutí testu. Pravděpodobnosti q_i dostaneme dle vztaho (26) a jim odpovídající pravděpodobnosti p_i dle vztahu (28). Hodnoty teoretických četností splňují podmínku $np_i \geq 5$. Vypočteme-li již známým způsobem dle (31) testové kriterium, dojdeme k hodnotě $12{,}8288$, která v porovnání s kritickou hodnotou $\chi^2_{8;0,95} = 15{,}507$ nepadne do kritického oboru pro zamítnutí. Tedy přijímáme nulovou hypotézu o tom, že takto vygenerovaný soubor dat pochází z Poissonova rozdělení. Potvrdil se tedy námi vyslovený předpoklad, že test data generovaná s Poissonovým rozdělením daty Poissonovými potvrdí.

Vezměme znovu již vygenerovaná data s rozsahem $n = 279$ pro parametr $\lambda = 6{,}2$. Tato data, jak už víme, mají skutečně Poissonovo rozdělení. Zkusme tato data nyní záměrně pokazit a zjistěme, zdali námi prováděný test pozná, že porušená data se od dat s Poissonovým rozdělením značně liší. Změněná data a hodnota testového kritéria jsou uvedena v tabulce 6 na pravé straně. Jak z této tabulky vidíme, byla změněna četnost pro $i = 2$ a to z šesti na šedesát. Takováto změna by se dala uvažovat např., jako chyba vzniklá při přepisování dat. Celkově takto dostáváme výběr o novém rozsahu $n = 333$. Nejprve dle podmínky pro rozsah výběru odhadneme parametr dle (30) pro $r = 0$ a $k = 15$ roven $\lambda = 5{,}5736$. Po výpočtu q_i dle (26) a kontrole podmínky $nq_i \geq 5$ dojdeme k závěru, že je třeba sloučit první dvě třídy a poslední čtyři. Pro sloučené třídy je pak $r = 1$ a $k = 11$. Za odhad parametru λ vezmeme počáteční aproximaci dle (30) rovnu $\lambda_0 = 5{,}5345$. Tomuto parametru odpovídají pravděpodobnosti q_i dle (26) a na jejich základě počítané pravděpodobnosti p_i dle vztahu (28). Všechny teoretické četnosti v tomto případě splňují podmínku pro dostatečný výběr $np_i \geq 5$. Spočítáme-li testové kritérium dle (31) dospějeme k hodnotě $109{,}4769$, která nás na základě kritické hodnoty $\chi^2_{9;0,95} = 16{,}919$ jasně opravňuje k zamítnutí nulové hypotézy ve prospěch alternativy. Námi prováděný test jasně odhalil, že na hladině $\alpha = 0{,}05$ takto záměrně pozměněný soubor dat Poissonovo rozdělení nemá.

Jelikož jsme doposud prováděli testování na poměrně velkých souborech dat, zkusme nyní provést testování pro rozsahy mnohem menší. Vezměme generovaná data s Poissonovým rozdělením o parametru $\lambda = 3$ po řadě o rozsahu, $n = 25$, $n = 28$, $n = 29$, $n = 30$, $n = 31$, $n = 32$ a zjistěme, zdali námi prováděný test pozná, že tato data poissonovo rozdělení nesou, a zda je tento test v některých případech vůbec možno provést v případě budeme li dbát na podmínku $np_i \geq 5$.

Provedením testování na takto generovaných datech jsme zjistili, že pro rozsahy výběru $n = 25$ až $n = 29$ došlo při provedení stejného principu testování, jako jsme používali výše, při dodržení podmínky $nq_i \geq 5$ ke sloučení empirických četností do jedné či dvou tříd, až na jednu vyjímku pro $n = 29$, což je nedostatečný počet tříd k dokončení testu, pro jehož dokončení jsou potřeba třídy alespoň tři. Sloučení do nedostatečného počtu tříd pak můžeme brát za varovný signál toho, že test provádíme

s malým rozsahem výběru. Pokud však toto varování nebudeme akceptovat a test dokončíme tak, že třídy nesloučíme, dojdeme v našem případě vždy k výsledku, který potvrzuje Poissonovo rozdělení generovaných dat.

Pro výběry generované o rozsahu $n = 29, 30, 31$ a $n = 32$, tak jak je to ukázáno v tabulce 7, došlo vždy, až na jednu výjimku, ke sloučení do tří tříd. Přitom odhad parametru je vždy počítán jako výběrový průměr, není upřesňován iterační metodou. Za zlomovou hodnotu rozsahu výběru, od které je test proveditelný při splnění podmínky $np_i \geq 5$ můžeme zřejmě považovat $n = 30$ a všechny rozsahy výběru, které budou větší. V případě upuštění od této podmínky a dokončení testu, došlo vždy k potvrzení Poissonova rozdělení generovaných dat. Pro případy generovaných výběrů uvedených v tabulce 7 slučovaných do tří tříd je hodnota kvantilu $\chi^2_{1;0,95} = 3{,}841$. Jen pro $n = 29$ sloučené do dvou tříd neexistuje kvantil pro rozhodnutí. Zároveň všechny hodnoty testových kritérií chí-kvadrát potvrzují nulovou hypotézu o původu výběru z Poissonova rozdělení, což jsme předpokládali, neboť data byla záměrně generována s tímto rozdělením, abychom mohli vyzkoušet spolehlivost testu.

tabulka 7: generované výběry s Poissonovým rozdělením

n	**29**		**29**		**30**		**30**		**31**		**32**	
λ	3,31	3,07	3,03	2,59	2,6	2,13	2,73	2,17	3,29	2,94	3,22	3,03
i	$\mathbf{X_i}$	**sloučené $\mathbf{X_i}$**	$\mathbf{X_i}$	**sloučené $\mathbf{X_i}$**	$\mathbf{X_i}$	**sloučené $\mathbf{X_i}$**	$\mathbf{X_i}$	**sloučené $\mathbf{X_i}$**	$\mathbf{X_i}$	**sloučené $\mathbf{X_i}$**	$\mathbf{X_i}$	**sloučené $\mathbf{X_i}$**
0	2		1		2		1		1		2	
1	3		2		6	8	7	8	8		3	
2	6	11	9	12	10	10	9	9	5	14	6	11
3	5	15	7	17	6	12	6	13	5	5	9	9
4	3	13	5		3		3		3	12	4	12
5	6		3		0		1		5		5	
6	4		2		0		1		0		1	
7	0		0		2		1		2		2	
8	0		0		1		0		0		0	
9	0		0		0		1		2		0	
10	0		0		0		0		0		0	
χ^2	**0,9059**		**1,3583**		**1,4731**		**1,2158**		**0,7740**		**0,8926**	

Příklad 4: (výšky studentů)

Vezměme tělesnou výšku měřenou v centimetrech u dvou ročníků studentů osmiletého gymnázia. Jde o 180 studentů ve věku mezi 12-ti až 15-ti lety. Nasbíraná data přitom pocházejí s Čajkova gymnázia v Olomouci ze školního roku 2008/2009. Úkolem je zjistit, zdali výška studentů má normální rozdělení. Pořízené hodnoty jsou nejprve roztříděny do tříd po třech centimetrech od velikosti 157 cm nahoru i dolů. Přitom tato hodnota byla spočtena zhruba jako aritmetický průměr výšek. Údaje o krajních hodnotách intervalů a empirických četnostech námi rozdělených tříd, tak jak jsme uvedli, najdeme v tabulce 8.

Pro takto rozdělená data do 13-ti tříd je nejprve potřeba spočítat odhady střední hodnoty a rozptylu dle iterační metody uvedené v podkapitole 3.2. Počáteční aproximace spočteme dle vztahu (36). Dostaneme $\mu_0 = 157{,}2667$ a $\sigma_0^2 = 34{,}5456$. Přitom první aproximace dostaneme tak, že na základě aproximací počátečních dopočítáme pravděpodobnosti p_i pomocí upraveného vzorce (37) a ty dosadíme do vzorců pro parametry (34) a (35). Výpočtem jsme zjistili, že $\mu_1 = 157{,}2660$ a $\sigma_0^2 = 34{,}5360$. Pokud aproximace zaokrouhlíme na dvě desetinná místa, vidíme, že se rovnají. Na základě tohoto poznatku a celkem zdlouhavých výpočtů v iterační metodě, budeme při dalších výpočtech brát za odhady parametrů aproximace počáteční. Během dalších výpočtů budeme dbát na dodržení pravidla pro dostatečný rozsah výběru $np_i \geq 5$, pokud tomu nebude uvedeno jinak.

Na základě provedeného odhadu parametrů spočítáme dle (37) pomocí distribučních funkcí normálního rozdělení pravděpodobnosti p_i. Z těchto dále teoretické četnosti, které zkontrolujeme, zdali splňují podmínku $np_i \geq 5$, díky tomu dochází ke sloučení krajních tříd s okolními, a to konkrétně prvních tří a posledních tří. Pro osm takto sloučených tříd je třeba znovu odhadnout parametry. Ty odhadneme opět dle vztahu (36). Jejich hodnota je nyní $\mu_0 = 157{,}25$ a $\sigma_0^2 = 27{,}9875$. Na základě těchto parametrů spočítáme pravděpodobnosti p_i ze vztahu (37) a z nich určíme teoretické četnosti. Ty zkontrolujeme, zda vyhovují podmínce $np_i \geq 5$. Zjistíme, že tomu tak ještě není v případě první třídy. Tu sloučíme dohromady s druhou a přepočítáme parametry pro těchto sloučených sedm tříd. Jejich hodnoty jsou dle (36) $\mu_0 = 157{,}4167$ a $\sigma_0^2 = 24{,}5431$. Při výpočtu pravděpodobností dle (37) a jim příslušných teoretických

četností zjistíme, že již všechny vyhovují podmínce $np_i \geq 5$. Nyní můžeme dopočítat testové kritérium dle (38). Toto má hodnotu $15{,}3663$, která padne do kritického oboru vymezeného kvantilem $\chi^2_{4;0,95} = 9{,}488$. Tedy zamítáme nulovou hypotézu o normálním rozdělení. Výšky studentů normální rozdělení nemají.

tabulka 8: výšky studentů gymnázia

i	**intervaly**		**empirické četnosti X_i**	**empirické četnosti tříd X_i**	**pravděpo-dobnosti p_i**	**teoretické četnosti np_i**	**testové kritérium chí-kvadrát**
	b_{i-1}	**b_i**					
1	131	141	2				
2	141	144	3				
3	144	147	5				
4	147	150	11	**21**	0,0672	12,0936	**6,5592**
5	150	153	15	15	0,1191	21,4450	1,9370
6	153	156	24	24	0,2011	36,2033	4,1134
7	156	159	49	49	0,2379	42,8237	0,8908
8	159	162	42	42	0,1972	35,4950	1,1922
9	162	165	17	17	0,1145	20,6139	0,6336
10	165	168	7	12	0,0629	11,3255	0,0402
11	168	171	3				
12	171	174	1				
13	174	183	1				
∑			**180**	**180**	**1**	**180**	**15,3663**

Pro zajímavost uveďme ještě variantu, pokud bychom se rozhodli opomenout pravidlo $np_i \geq 5$ a spočítali testové kritérium pro všech třináct tříd. Odhady parametrů již máme, máme i hodnoty pravděpodobností a teoretických četností. Testové kriterium dle (38) má hodnotu $22{,}1209$, která v porovnání s kvantilem $\chi^2_{10;0,95} = 18{,}307$ padne do kritického oboru pro zamítnutí nulové hypotézy. Tedy i při nedodržení této podmínky rozhodl test stejně, zamítl normalitu výšek studentů. Výšky studentů nemají normální rozdělení zřejmě proto, že nesou jistou známku zešikmení a nesymetrie. Přitom symetrie hustoty normálního rozdělení dle středu, je jednou ze základních vlastností tohoto rozdělení. Navíc by dalším prvkem, který normalitu výšek porušuje, mohly být některé neobvyklé hodnoty.

Jelikož je z praxe známo, že veličina udávající tělesnou výšku se často normálním rozdělením řídí, byl náš předpoklad také takový. Je možné, že naše data by normální rozdělení mít mohla v případě, když z nich vyloučíme nějak neobvyklé

hodnoty, jež jsou v našem souboru dat představovány jedním velkým chlapcem (183 cm) a jednou velmi malou dívkou (131 cm). Odeberme tedy tyto dvě hodnoty a zkusme znovu ověřit hypotézu o původu takto upraveného souboru dat z normálního rozdělení.

Vezměme znovu původní data výšek studentů gymnázia. Z nich jsme odstranili právě obě neobvyklé hodnoty popsané výše. Data jsme znovu setřídili okolo hodnoty, určené jako aritmetický průměr, rovné 155,5 cm, do devíti tříd po čtyřech centimetrech. Rozsah výběru se zmenšil na $n = 178$. Odhady parametrů dle vztahu (36) jsou rovny $\mu = 158{,}4045$ a $\sigma^2 = 31{,}8813$. Při výpočtu pravděpodobností z (37) a teoretických četností, je pro dodržení pravidla $np_i \geq 5$ potřeba sloučit první dvě a poslední jednu třídu se sousedními. Dostáváme tak šest tříd, pro něž jsou odhady parametrů dle (36) rovny $\mu = 158{,}4944$ a $\sigma^2 = 26{,}2275$. Hodnoty pravděpodobností dle (37), teoretických četností a testového kritéria dle (38) najdeme v tabulce 9. Hodnota kvantilu $\chi^2_{3;0,95} = 7{,}814$. Jelikož hodnota testového kritéria $14{,}2637$ padne do kritického oboru, zamítáme hypotézu o normalitě dat. Náš předpoklad se, ani při tomto přístupu k řešení, nepotvrdil.

tabulka 9: výšky studentů při odstranění dvou neobvyklých hodnot

i	**intervaly**		**empirické četnosti X_i**	**empirické četnosti tříd X_i**	**pravděpo-dobnosti p_i**	**teoretické četnosti np_i**	**testové kritérium chí-kvadrát**
	b_{i-1}	**b_i**					
1	138	142	1				
2	142	146	5				
3	146	150	8	14	0,0486	8,6497	3,3095
4	150	154	21	21	0,1415	25,1851	0,6955
5	154	158	34	34	0,2715	48,3207	4,2442
6	158	162	69	**69**	0,2916	51,9099	**5,6265**
7	162	166	29	29	0,1754	31,2286	0,1590
8	166	170	8	11	0,0714	12,7060	0,2291
9	170	174	3				
∑			**178**	**178**	**1**	**178**	**14,2637**

Zkusme naposledy vzít výchozí data, ale rozdělit je do tříd po pěti centimetrech okolo aritmetického průměru 157 cm. Jde tedy o podobné setřídění, jako bylo provedeno poprvé, jen jsme zvětšili délku třídy. Při tomto setřídění padne do první třídy jen hodnota výšky jedné velmi malé dívky a do druhé dokonce hodnota žádná. Zkusme

proto tyto dvě třídy úplně odebrat, přitom tak odstraníme jen jednu hodnotu z výběru. Rozsah testovaných hodnot nyní bude $n = 179$ a vzniklo nám takto devět tříd. Jako nulovou hypotézu opět položme původ dat z normálního rozdělení. Odhadem parametrů dle (36) dostaneme $\mu = 157{,}1648$ a $\sigma^2 = 36{,}4798$. Spočítáme pravděpodobnosti dle (37) a jim odpovídající teoretické četnosti. Pro vyhovění podmínce $np_i \geq 5$ dojde ke sloučení prvních dvou a posledních dvou tříd s okolními. Pro výsledných pět tříd mají odhady parametrů dle (36) hodnotu $\mu = 157{,}2486$ a $\sigma^2 = 28{,}2888$. Hodnoty pravděpodobností příslušejících těmto parametrům spočítáme opět dle (37). Jim odpovídající teoretické četnosti splňují podmínku pro rozsah výběru, tedy spočítáme testové kriterium dle (38) jehož hodnota je $4{,}791$, jež nespadá do kritického oboru pro zamítnutí daného hodnotou $\chi^2_{2;0,95} = 5{,}991$. Můžeme tedy říci, že takto námi upravená data, při rozdělení do tříd po pěti, mají normální rozdělení. Při tomto způsobu řešení se tedy potvrdil náš předpoklad o původu výběru z normálního rozdělení. Přičemž můžeme říci, že volba tříd a odstranění vybočujících hodnot z výběru má na výsledek testu podstatný vliv.

Kdybychom u této varianty upustili na začátku od podmínky pro dostatečný rozsah výběru a třídy nijak neslučovali, dostaneme dle vzorce (38) pro všech devět tříd hodnotu testového kritéria $11{,}908$. Hodnota kvantitu vymezujícího kritický obor je v tomto případě rovna $\chi^2_{6;0,95} = 12{,}592$. Hodnota testového kritéria do kritického oboru pro zamítnutí nespadá, tedy přijímáme hypotézu o normalitě.

Nyní ještě proveďme test na datech, o kterých víme, že pocházejí z normálního rozdělení o rozsahu $n = 180$, střední hodnotou $\mu = 157$ a směrodatnou odchylkou $\sigma = 5$. Jde o data nasimulovaná generátorem pseudonáhodných čísel, která mají podobnou povahu a strukturu jako data, která jsme již výše vyšetřovali. Ověřme tedy, zdali test jednoznačně přijme nulovou hypotézu. Přitom údaje o četnostech a testovém kritériu jsou dány v tabulce 10. Data jsme přitom roztřídili okolo hodnoty 157 do devíti tříd o délce tři. Při výpočtech budeme postupovat stejně, jako u testování reálných dat, dle teorie odvozené v podkapitole 3.2. Nejprve odhadneme parametry dle (36), jejich hodnota je $\mu = 157{,}4833$ a $\sigma^2 = 26{,}5497$. Pravděpodobnosti spočítáme dle vztahu (37) a jim odpovídající teoretické četnosti zkontrolujeme, zdali splňují podmínku $np_i \geq 5$. Zjistíme, že je třeba sloučit první a poslední třídu se sousedními. Pro takto

sloučených sedm tříd spočítáme opět parametry dle (36). Jejich hodnota je $\mu = 157{,}5$ a $\sigma^2 = 24{,}1$. Teoretické četnosti spočítaných s pravděpodobností již všechny splňují naši podmínku, tedy spočítáme testové kritérium dle vzorce (38). Jeho hodnota je $3{,}222$, která jasně nepadne do kritického oboru pro zamítnutí, vymezeného kritickou hodnotou kvantilu $\chi^2_{4;0,95} = 9{,}488$. Přijímáme tedy nulovou hypotézu o původu generovaných dat z normálního rozdělení na hladině testu $\alpha = 0{,}05$. Na takto uměle vytvořených datech, na která byl předem stanoven požadavek, aby měla, normální rozdělení s konkrétními parametry, byla pěkně ukázána spolehlivost testu, jež tuto skutečnost měla potvrdit, a také ji potvrdila.

tabulka 10: simulovaná data s normálním rozdělením pro $n = 100$

i	**intervaly**		**empirické četnosti X_i**	**empirické četnosti tříd X_i**	**pravděpo-dobnosti p_i**	**teoretické četnosti np_i**	**testové kritérium chí-kvadrát**
	b_{i-1}	**b_i**					
1	144	147	4				
2	147	150	12	16	0,0633	11,39168	1,8642
3	150	153	19	19	0,1164	20,94762	0,1811
4	153	156	30	30	0,2003	36,05593	1,0172
5	156	159	45	45	0,2401	43,20953	0,0742
6	159	162	36	36	0,2003	36,05593	0,0001
7	162	165	22	22	0,1164	20,94762	0,0529
8	165	168	9	12			
9	168	171	3				
∑			**180**	**180**	**1**	**180**	**3,222**

Pokud bychom v tomto případě upustili od podmínky $np_i \geq 5$. Zjistíme, že test daná data normálními opět potvrdí. Testové kriterium chí-kvadrát dle (38) má pro všech devět původně nesloučených tříd hodnotu $2{,}2743$, která nepadne do kritického oboru vymezeného kvantilem $\chi^2_{6;0,95} = 12{,}592$. Pak přijímáme hypotézu o původu generovaných dat s normálního rozdělení.

Jelikož rozsahy výběrů při takto provedených testech byly opět dostatečně veliké, zkusme nyní, zdali test funguje i pro hodnoty výběru mnohem menší. Pro zajímavost otestujme vygenerovaný výběr s normálním rozdělením o rozsahu $n = 30$ a parametry $\mu = 157$ a $\sigma = 5$, který v případě ověřování Poissonova rozdělení stačil pro možnost dokončení testu i za dodržení podmínky $np_i \geq 5$. Pro dokončení testu

s podmínkou je třeba, aby došlo ke sloučení minimálně do čtyř tříd a více. Data jsme rozdělili do osmi tříd o délce tři, jak je to uvedeno v tabulce 11. Podle (36) spočítáme odhady parametrů, které jsou $\mu = 156{,}9$ a $\sigma^2 = 32{,}84$. Pravděpodobnosti vypočteme dle (37) a zjistíme, že teoretické četnosti vedou ke sloučení do výsledných tří tříd což je nedostatečné pro dokončení testu.

Pokud podmínku vypustíme a spočítáme testové kritérium dle (38) dostaneme hodnotu $4{,}7031$, která v porovnání s kvantilem $\chi^2_{5;0,95} = 11{,}07$ nepadne do kritického oboru pro zamítnutí, můžeme tedy říci, že nasimulovaná data mají normální rozdělení. Přitom nemožnost test dokončit při dodržení podmínky $np_i \geq 5$ je signalizací k tomu, že testování provádíme na malém rozsahu výběru. V případě reálných dat by bylo třeba soubor dat nějakým způsobem rozšířit.

tabulka 11: simulovaná data s normálním rozdělením pro $n = 30$

i	**intervaly**		**empirické četnosti X_i**	**pravděpo-dobnosti p_i**	**teoretické četnosti np_i**	**empirické četnosti tříd X_i**	**testové kritérium chí-kvadrát**
	b_{i-1}	**b_i**					
1	146	149	3	0,0840	2,5205		0,0912
2	149	152	2	0,1122	3,3673		0,5552
3	152	155	7	0,1739	5,2156	12	0,6105
4	155	158	8	0,2060	6,1799	8	0,5361
5	158	161	2	0,1867	5,6018	10	2,3158
6	161	164	4	0,1295	3,8845		0,0034
7	164	167	2	0,0687	2,0605		0,0018
8	167	170	2	0,0390	1,1699		0,5891
∑			**30**	**1**	**30**	**30**	**4,7031**

Příklad 5: (pošta)

V tomto příkladě se budeme zabývat ověřením exponenciálního rozdělení. Vezměme soubor dat, která byla získána sledováním lidí využívajících služeb České pošty. Sledovány byly časové intervaly mezi příchody jednotlivých lidí na poštu k automatu pro přidělení čísla do fronty. Sledování bylo provedeno mezi třetí a čtvrtou hodinou odpoledne, přitom frekventovanost příchodů byla celkem velká, během této hodiny poštu navštívilo 85 lidí. Délky časů mezi jednotlivými příchody byly setříděny do jedenácti tříd po půl minutě, empirické četnosti při takto zvolených intervalech o délce $h = 0{,}5$ minuty jsou uvedeny v tabulce 12. Ověřme nyní, zdali délky časových

intervalů mezi jednotlivými příchody lidí na poštu mají exponenciální rozdělení. Z praxe přitom víme, že data udávající např. intervaly mezi jednotlivými příchody požadavků do fronty či intervaly mezi selháváním součástek apod. se řídí nejčastěji právě exponenciálním rozdělením.

tabulka 12: lidé přicházející na poštu

i	**intervaly**		**empirické četnosti X_i**	**empirické četnosti tříd X_i**	**pravděpo-dobnosti p_i**	**teoretické četnosti np_i**	**testové kritérium chí-kvadrát**
	b_{i-1}	**b_i**					
1	0	0,5	32	32	0,3291	27,9693	0,5809
2	0,5	1	19	19	0,2208	18,7660	0,0029
3	1	1,5	11	11	0,1481	12,5910	0,2010
4	1,5	2	5	5	0,0994	8,4479	1,4072
5	2	2,5	8	18	0,2027	17,2258	0,0348
6	2,5	3	4				
7	3	3,5	1				
8	3,5	4	3				
9	4	4,5	0				
10	4,5	5	1				
11	5	5,5	1				
∑			**85**	**85**	**1**	**85**	**2,2269**

Nejprve je dle podkapitoly 3.3 nutné odhadnout parametr exponenciálního rozdělení, tento odhadneme pomocí výběrového průměru daného vztahem (40). Jeho hodnota je $\lambda = 1{,}4059$. Dále dopočítáme pravděpodobnosti dle vztahu (39) a z nich teoretické četnosti. Budeme se řídit pravidlem pro slučování tříd $np_i \geq 5$. Díky tomuto dochází ke sloučení posledních sedmi tříd do jedné. Pro takto sloučené empirické četnosti je nyní hodnota parametru dle vztahu (40) rovna $\lambda = 1{,}2529$. Hodnota pravděpodobností, teoretických četností a testového kritéria chí-kvadrát po sloučení, je uvedena v tabulce 12. Jelikož došlo ke sloučení do pěti tříd, srovnáváme hodnotu testového kritéria spočítanou dle vzorce (44) s hodnotou kvantilu $\chi^2_{3;0,95} = 7{,}814$. Neboť $2{,}2269$ nepadne do kritického oboru pro zamítnutí, přijímáme nulovou hypotézu o původu dat z exponenciálního rozdělení. Můžeme tedy říci, že intervaly mezi jednotlivými příchody lidí na poštu se řídí exponenciálním rozdělením, jak bylo předpokládáno.

Kdybychom u tohoto příkladu pominuli pravidlo pro slučování $np_i \geq 5$ dostaneme dle vzorce (44) hodnotu testového kritéria, která je pro jedenáct tříd rovna $9{,}1470$. Kritická hodnota vymezující obor pro zamítnutí nulové hypotézy je $\chi^2_{9;0,95} = 16{,}919$. Jelikož hodnota testového kritéria nepadne do kritického oboru, přijímáme nulovou hypotézu o původu dat z exponenciálního rozdělení i v případě nesloučení málo četných tříd. Potvrdil se tedy předpoklad o tom, že takovýto soubor dat z exponenciálního rozdělení pochází.

Nyní vezměme data generovaná s exponenciálním rozdělením o rozsazích $n = 20,\ n = 30,\ n = 50$ a $n = 100$. Data jsou přitom vygenerována s parametrem $\lambda = 2$ a jsou roztřízena do tříd o délce $h = 1$. Empirické četnosti pro jednotlivé rozsahy výběrů jsou dány v tabulce 13.

tabulka 13: generovaná data s exponenciálním rozdělením

i	**intervaly**		**empirické četnosti X_i**	**empirické četnosti X_i**	**empirické četnosti X_i**	**empirické četnosti X_i**
	b_{i-1}	**b_i**				
1	0	1	9	15	20	41
2	1	2	3	6	11	23
3	2	3	4	3	5	14
4	3	4	2	4	5	12
5	4	5	0	1	4	5
6	5	6	1	0	0	3
7	6	7	0	1	1	2
8	7	8	1		2	
9	8	9			2	
∑			**20**	**30**	**50**	**100**

V první řadě vezměme simulovaná data s rozsahem výběru $n = 20$. Hodnota parametru odhadovaná pomocí výběrového průměru ze vztahu (40) je rovna $\lambda = 2{,}45$. Pravděpodobnosti spočítáme opět dle (39) a z nichž určíme hodnoty teoretických četností, přičemž zjistíme, že dle pravidla pro slučování $np_i \geq 5$ bychom museli slučovat do jedné třídy což je málo pro provedení testu, které vyžaduje alespoň třídy tři. Upustíme-li od pravidla pro slučování a spočítáme testové kritérium chí-kvadrát dle vzorce (44) pro všech osm tříd. Jeho hodnota je $3{,}554$. Jelikož nepadne do kritického oboru pro zamítnutí, vymezeného kritickou hodnotou $\chi^2_{6;0,95} = 12{,}592$, přijímáme nulovou hypotézu o původu dat z exponenciálního rozdělení na hladině testu $\alpha = 0{,}05$.

I při upuštění od požadavku na teoretické četnosti rozhodl test ve prospěch předpokladu původu dat z exponenciálního rozdělení, se kterým byla tato data vygenerována.

Pro generovaná data z rozsahem $n = 30$ má odhad parametru podle vzorce (40) hodnotu $\lambda = 2{,}13$. V případě, že dopočítáme pravděpodobnosti a z nich teoretické četnosti zjistíme, že dle pravidla pro slučování bychom sloučili empirické četnosti do dvou tříd, což je pro dokončení testu málo. Nebudeme tedy třídy slučovat a spočítáme hodnotu testového kritéria, ta má v tomto případě hodnotu $4{,}1650$, která nepadne do kritického oboru pro zamítnutí vymezeného hodnotou $\chi^2_{5;0,95} = 11{,}07$. Na hladině testu $\alpha = 0{,}05$ přijímáme hypotézu o původu dat z exponenciálního rozdělení. Tedy prováděný test opět data úspěšně identifikoval jako exponenciální.

V třetím případě otestujme data generovaná z rozsahem $n = 50$. Odhad parametru dle (40) je $\lambda = 2{,}76$. Po výpočtu pravděpodobností a teoretických četností sloučíme posledních pět tříd do jedné a dostaneme tím čtyři třídy, pro které je nyní odhad parametru roven $\lambda = 2{,}26$. Testové kriterium má pak po výpočtu pravděpodobností a teoretických četností hodnotu $1{,}0808$, která nepadne do kritického oboru vymezeného hodnotou $\chi^2_{2;0,95} = 5{,}991$. Na hladině testu $\alpha = 0{,}05$ pak přijímáme nulovou hypotézu o původu dat z exponenciálního rozdělení.

Naposledy ještě otestujme data o rozsahu $n = 100$. Odhad parametru je při použití stejného postupu jako výše roven $\lambda = 2{,}34$. Podle pravidla pro slučování dojde ke sloučení posledních dvou tříd. Pak pro takto vzniklých šest tříd je odhad parametru roven $\lambda = 2{,}32$. Hodnota testového kriteria je rovna $5{,}6555$. Kritická hodnota je $\chi^2_{4;0,95} = 9{,}488$. Jelikož $5{,}6555 < 9{,}488$, přijímáme nulovou hypotézu o původu simulovaných dat z exponenciálního rozdělení.

V druhé části příkladu pro data generovaná s exponenciálním rozdělením vždy test správně rozhodl, že data exponenciální rozdělení mají. Navíc je zřejmé, že testování prováděné při dodržování podmínky pro slučování je možné až pro větší rozsahy výběrů.

Závěr

Bakalářská práce na téma „Testy dobré shody“ mě nejprve vedla k seznámení s potřebným teoretickým zázemím. Následovalo podrobné odvození teorie, na které je založeno testování statistických hypotéz typu: „náhodná veličina má exponenciální rozdělení“, „kostka je homogenní“ či „výška studentů má normální rozdělení“, jež jsem učinila a uvedla v prvních třech kapitolách práce, což bylo pro mě důležité pro pochopení jak, a na jakém principu, testy dobré shody fungují. Přínosnější však pro mě byla praktická část výpočtů na jednotlivých příkladech, jež se pro mě stala vodítkem k provázání získaných teoretických poznatků s praktickým výpočtem na reálně situovaných příkladech. Rozšířila jsem si navíc přehled o tom, co je možno pomocí těchto testů rozhodovat.

Během výpočtů jsem se setkala s nutností dodržovat určité postupy a stanovená pravidla vedoucí ke správnému rozhodnutí testu. V příkladech jsem používala jednu z nejznámějších podmínek, kladenou na teoretické četnosti, která se používá proto, aby došlo k zaručení dostatečného rozsahu výběru pro prováděný test. Zjistila jsem však, že tato podmínka není přímo nutná k dokončení testu, ale je jen signalizátorem k tomu, že provádíme testování na malém rozsahu výběru. V každém případě, při nedodržení této podmínky a dokončení testu, rozhodl prováděný test vždy ve prospěch mého předpokladu. Chci ale poznamenat, že jsem povahu dat většinou znala. Takovýto výsledek testu bez dodržení zmíněné podmínky, by však mohl být zpochybnitelný v případě, že bychom povahu testovaných dat vůbec neznali. Přitom hodnotit správnost rozhodnutí prováděného testování mi umožnila právě data simulovaná, o kterých jsem věděla, že určitý typ rozdělení mají.

Také jsem si potvrdila, že dostatečný rozsah výběru je velmi důležitý. S ohledem na dodržení podmínky pro teoretické četnosti, kdy bylo prováděno testování při velkém rozsahu výběru, nikdy nedošlo k situaci, že by test nebylo možné dokončit. V případě malých rozsahů výběru, jsem ale několikrát dospěla k závěru, že při dodržení podmínky rozsahu, nešlo test dokončit. Domnívám se, že v případě Poissonova rozdělení je hranice možného provedení testu, za dodržení podmínky, někde okolo rozsahu výběru daného třiceti a více daty. Pro spojitá rozdělení normální a exponenciální jsem ověřila

proveditelnost testu za dodržení podmínky pro rozsah výběru, jež je dán pomocí padesáti a více dat.

Pro výpočty jsem používala program Excel, pomocí kterého jsem získala i data simulovaná generátorem pseudonáhodných čísel. Přitom vzhledem k výpočtům se mi zdálo nejsložitější ověřování normálního rozdělení. Složitost výpočtů u něj zřejmě plyne z toho, že má dva parametry a hustotu, jež nelze vyjádřit explicitně. Celkově hodnotím testy dobré shody jako silné pravidlo pro rozhodnutí statistické hypotézy. Ověřila jsem si, že i při menších odchylkách v hodnotách parametrů, při záměrném porušení dat dokonce i při nedodržování některých podmínek používaných při testování, test spolehlivě a důvěryhodně rozhoduje.

Práce se stala přínosem pro rozšíření mých znalostí z matematické statistiky. Nejvíce v oblasti aplikace teorií na reálné situace. Proto věřím, že nabyté znalosti a zkušenosti budu moci využít i ve své další práci a životě.

Použitá literatura

[1] Anděl, J., *Statistické metody*, 3.vydání. Praha: Matfyzpress, 2003.

[2] Anděl, J., *Matematická statistika*, 2.vydání. Praha: SNTL, 1985.

[3] Hebák, P., a kol., *Vícerozměrné statistické metody (1)*, 1. vydání. Praha: Informatorium, 2004.

[4] Hindls, R., a kol., *Statistika pro ekonomy*, 5.vydání. Praha: Professional Publishing, 2004.

[5] Kunderová, P., *Úvod do teorie pravděpodobnosti a matematické statistiky*, 2. nezměněné vydání. Olomouc: Universita Palackého v Olomouci, 2004.

[6] Marek, L., a kol., *Statistika pro ekonomy – aplikace*, 1. Vydání. Praha: Professional Publishing, 2005.

[7] Likeš, J., Machek, J., *Matematická statistika*, 2. vydání. Praha: SNTL, 1988.

[8] Reif, J., *Metody matematické statistiky*, 2. upravené vydání. Plzeň: Západočeská universita v Plzni, 2004.

[9] Seger, J., Hindls, R., *Statistické metody v tržním hospodářství*, 1. vydání. Praha: Victoria Publishing, 1995.