Testy dobré shody při neznámých parametrech, ověřování exponenciálního rozdělení

Pavel Hellebrand

20.března 2008

Obsah

- Multinomické rozdělění
- Testy dobré shody při známých parametrech
- Testy dobré shody při neznámých parametrech
- Ověřování exponenciálního rozdělení

Multinomické rozdělení 1

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

$$(1) \qquad 0 < p_i < k \quad p_1 + \ldots + p_k = 1.$$

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

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

pro $x_i = 0, 1, \ldots, n \quad (i = 1, 2, \ldots, k), \quad x_1 + \ldots + x_k = n$

- Rozdělení dané vzorcem (2) se nazývá *multinomické s parametry* $n, p_1, \ldots, p_k$

Multinomické rozdělení 2

- Pro multinomické rozdělení platí, že všechna jeho marginální podmíněná rozdělení jsou opět multinomická.
- Všechna jednorozměrná marginální rozdělení jsou *binomická*. (X_i má binomické rozdělení s parametry *n* a p_i)
- $EX_i = np_i, \qquad \operatorname{var} X_i = np_i(1-p_i), \qquad 1 \leqslant i \leqslant k,$
  $\operatorname{cov}(X_i, X_j) = -np_i p_j, \qquad 1 \leqslant i \neq j \leqslant k.$
- Jestliže $\boldsymbol{X} = (X_i, \ldots, X_k)'$ má multinomické rozdělění (2), pak náhodná veličina

$$(3) \qquad \chi^2 = \sum_{i=1}^{k} \frac{(X_i - np_i)^2}{np_i}$$

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

Multinomické rozdělení 3

- Vzorec (3) lze snadno upravit na tvar

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

který je vhodnější pro výpočet, avšak v praxi se dává přednost vzorci (3), protože je z něj lépe vidět jakou měrou přispívá každý sčítanec k celkovému součtu χ^2.

Test dobré shody při známých parametrech 1

- Předpokládejme, že výsledky pozorování byly uspořádány do *k* tříd s četnostmi $X_1, \ldots, X_k$ (empirické četnosti)
- Dále předpokládejme, že teoretické modelové rozdělení četností je reprezentováno četnostmi np_i (očekávané četnosti)
- Potom shodu mezi empirickým a teoretickým rozdělením posuzujeme pomocí testovacího kriteria

$$(3) \qquad \chi^2 = \sum_{i=1}^{k} \frac{(X_i - np_i)^2}{np_i}$$

- Tzv. Pearsonův Chí-kvadrát test
- Pro to aby se tato veličina řídila asymproticky chí-kvadrát rozdělěním o *k-1* stupni volnosti je žádoucí aby *n* > *50*.

Test dobré shody při známých parametrech 2

- Dále je nutné, aby pro teoretické četnosti platilo $np_i \geqslant 5$
- Nevyhovují-li některé četnosti této podmínce, lze dosáhnout jejího splnění sloučením několika sousedních tříd. Tím se sníží počet stupňů volnosti, neboť *k* je rovno počtu tříd po sloučení.
- Pokud pro hodnotu testovací statistiky platí $\chi^2 < \chi^2_\alpha(k-1)$

  pak testovanou hypotézu nezamítáme na hladině významnosti alpha.

Chí-kvadrát test dobré shody – příklad se zmrzlinou

- *Řetězec cukráren, který nabízí 4 druhy zmrzliny otevřel provozovnu v nové lokalitě. Ve stávajících provozovnách řetězce byla dosud struktura prodeje podle druhů zmrzliny následující: vanilková 62%, čokoládová 18%, jahodová 12%, pistáciová 8%. Po otevření provozovny v nové lokalitě máme záznam o následujícím prodeji: vanilková 120, čokoládová 40 jahodová 18, pistáciová 22.*
- *Vyjádřete se pomocí statistického testu ke shodě či odlišnosti struktury prodeje v nové lokalitě oproti dosavadním prodejům řetězce.*

Příklad se zmrzlinou - řešení

- Pro získání očekávaných četností u prodeje zmrzliny (při platnosti stávající struktury prodeje pro novou lokalitu) aplikujeme dosavadní strukturu prodeje na celkové prodané množství v nové lokalitě (kde je prodáno celkem 200 kusů zmrzliny):

- Např. u vanilkové: očekávaná četnost při 200 prodaných kusech = 62% * 200 = 134 kusů

- Tyto očekávané četnosti konfrontujeme se skutečně pozorovanými (chí-kvadrát test dobré shody), výpočet testového kritéria:

$$\chi^2 = \sum_{i=1}^{k} \frac{(X_i - np_i)^2}{np_i} = \ldots = 4{,}32$$

Příklad se zmrzlinou – řešení

Výpočet je naznačen v následující tabulce:

	vanilková	čokoládová	jahodová	pistáciová	Σ
struktura prodeje	62%	18%	12%	8%	100%
nová provozovna	120	40	18	22	200
oč.při stejné struktuře	124	36	24	16	200
chi-square:	0,13	0,44	1,50	2,25	**4,32**

Příklad se zmrzlinou - řešení

- Spočtenou hodnotu testového kritéria porovnáme s příslušným kvantilem rozdělení χ^2 s (k-1), tedy se 3 stupni volnosti. Pro 5% hladinu významnosti půjde o kvantil $\chi^2_{(1-\alpha)}$, tedy o kvantil $\chi^2_{0,95} = 7,82$

- Spočtená hodnota testového kritéria (4,32) nepřekračuje mez vymezující kritický obor (7,82), nachází se v oboru přijetí a na zvolené 5%ní hladině významnosti hypotézu o shodě struktury prodeje nezamítáme.

Testy dobré shody při neznámých parametrech

- V praxi se často stává, že pravděpodobnosti $p_1, \dots, p_k$ uvažovaného multinomického rozdělení závisejí na nějakých neznámých parametrech $a_1, \dots, a_m$. Vzniká pak problém jak pozměnit testovací kritérium, aby se hodilo i na tento příklad.
- Nabízí se možnost tyto parametry odhadnout, odtud získat i odhady pro pravděpodobnosti $p_1, \dots, p_k$ a do vzorce (3) pak dosadit tyto odhady.
- Lze očekávat, že pak rozdělení chí-kvadrát bude mít o tolik stupňů volnosti méně, kolik parametrů jsme museli odhadovat.

Testy dobré shody při neznámých parametrech

- Označme $\boldsymbol{a}=(a_1,\ldots,a_m)'$. Předpokládejme, že $p_1=p_1(\boldsymbol{a}),\ldots,p_k=p_k(\boldsymbol{a})$ jsou dostatečně hladké funkce proměnné ***a*** . Protože platí pro každé ***a***

$$(4)\qquad p_1(\boldsymbol{a})+\ldots+P_k(\boldsymbol{a})=1,$$

derivováním dostaneme

$$(5)\qquad \frac{\partial p_1(\boldsymbol{a})}{\partial a_j}+\ldots+\frac{\partial p_k(\boldsymbol{a})}{\partial a_j}=0 \quad j=1,2,\ldots,m.$$

- Nyní vzniká problém, jakým způsobem pořídit odhady parametru ***a***. Jedna možnost spočívá v tom, že se za odhad vezme ta hodnota ***a***, která při daných veličinách $X_1,\ldots,X_k$ minimalizuje χ^2 ve vzorci (3)
- Jedná se o jakousi analogii metody nejmenších čtverců. Říkáme, že jde o odhad parametru ***a*** pořízený metodou minimálního χ^2

Testy dobré shody při neznámých parametrech

- Po derivaci vzorce (3) dostaneme tuto soustavu rovnic

$$(6) \qquad -\frac{1}{2}\frac{\partial \chi^2}{\partial a_j} = \sum_{i=1}^{k} \left\{ \frac{X_i - np_i(\boldsymbol{a})}{p_i(\boldsymbol{a})} + \frac{[X_i - np_i(\boldsymbol{a})]^2}{2\mathrm{np}_i^2(\boldsymbol{a})} \right\} \frac{\partial\, p_i(\boldsymbol{a})}{\partial a_j} = 0, \qquad j = 1,2,\ldots,m.$$

- Ukazuje se, že vliv druhého členu v (6) při velkém *n* není příliš podstatný, takže řešení soustavy rovnic

$$(7) \qquad \sum_{i=1}^{k} \frac{X_i - np_i(\boldsymbol{a})}{p_i(\boldsymbol{a})} \frac{\partial\, p_i(\boldsymbol{a})}{\partial a_j} = 0, \qquad j = 1,2,\ldots,m,$$

se příliš neliší od řešení soustavy (6). Vzhledem k (5) lze soustavu (7) upravit na tvar

$$(8) \qquad \sum_{i=1}^{k} \frac{X_i}{p_i(\boldsymbol{a})} \frac{\partial\, p_i(\boldsymbol{a})}{\partial a_j} = 0, \qquad j = 1,2,\ldots,m.$$

Testy dobré shody při neznámých parametrech

- Řešením soustavy (8) je tzv. odhad parametru ***a*** modifikovanou metodou minimálního χ^2
- **Věta:** *Budiž dáno k funkcí* $p_1(\boldsymbol{a}), \ldots, p_k(\boldsymbol{a}), \quad \text{kde} \quad a=(a_1, \ldots, a_m)'$. *Předpokládejme, že* $m<k-1$. *Nechť pro všechny body* ***a*** *nedegenerovaného konečného uzavřeného intervalu A z* R_m *platí:*

  1. $p_1(\boldsymbol{a})+\ldots+p_k(\boldsymbol{a})=1$
  2. *Existuje takové c > 0, že* $p_i(\boldsymbol{a})>c \quad \text{pro} \quad i=1,2,\ldots,k.$
  3. *Každá funkce* $p_i(\boldsymbol{a})$ *má spojité derivace*

     $$\partial p_i(\boldsymbol{a})/\partial a_j \quad a \quad \partial^2 p_i(\boldsymbol{a})/\partial a_j \partial a_s \quad (j,s=1,2,\ldots,m)$$

  4. *Matice* $(\partial p_i(\boldsymbol{a})/\partial a_j)_{i=1,j=1}^{k,m}$ *má hodnost m.*

  Nechť $\boldsymbol{a}^0$ *je vnitřním bodem A. Označme* $p_i^0=p_i(\boldsymbol{a}^0)$.

Testy dobré shody při neznámých parametrech

Nechť $\boldsymbol{X}=(X_1,\ldots,X_k)'$ *má multinomické rozdělení s parametry* $n, p_1^{0,}\ldots, p_k^0$.

Pak soustava rovnic (8) má právě jeden kořen ***a*** *takový, že* ***a*** *konverguje k* $\boldsymbol{a}^0$ *podle pravděpodobnosti při* $n\to\infty$. *Dosadíme-li tento kořen do výrazu*

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

má veličina χ^2 při $n\to\infty$ *asymptoticky rozdělení chí-kvadrát s k – m - 1 stupni volnosti.*

Důkaz: například v Andělovi (Matematická statistika)

- Tato věta se používá hlavně k ověřování typu rozdělění a při hodnocení kontigenčních tabulek.

Ověřování exponenciálního rozdělení

- Chceme ověřit, zda daný výběr pochází z exponenciálního rozdělení s hustotou

$$(9) \qquad f(x)=\frac{1}{\lambda}\mathrm{e}^{-x/\lambda} \qquad \text{pro} \quad x>0,$$

kde $\lambda > 0$ je neznámý parametr. Položme $b_i=ih\,, i=0,1,\ldots,k-1 \,\text{kde}\, h>0$ je vhodně zvolená délka třídy. Nechť $b_k=\infty$. Máme třídy

$$(10) \qquad \langle 0,b_1),\langle b_1,b_2),\ldots,\langle b_{k-2},b_{k-1}),\langle b_{k-1},\infty)$$

jejichž četnosti nechť jsou $X_1,\ldots,X_k$; opět položíme $X_1+\ldots+X_k=n.$

- Pravděpodobnost, že jednotka padne do *i*-té třídy, je rovna

$$(11) \qquad p_i=\mathrm{e}^{-b_{i-1}/\lambda}-\mathrm{e}^{-b_i/\lambda}, \qquad i=1,2,\ldots,k.$$

Ověřování exponenciálního rozdělění

- Položíme- li $\infty \mathrm{e}^{-\infty} = 0$, máme

$$\frac{\mathrm{d}\, p_i}{\mathrm{d}\,\lambda} = \frac{1}{\lambda^2}\left(b_{i-1}\mathrm{e}^{-b_{i-1}/\lambda} - b_i \mathrm{e}^{-b_i/\lambda}\right), \qquad i = 1, 2, \ldots, k.$$

- Rovnice (8) má pak tvar

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

Odtud

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

Ověřování exponenciálního rozdělení

- Dosazením za b_i dostaneme

$$\mathrm{e}^{-h/\lambda} = \frac{hX_2 + 2\mathrm{h}X_3 + \ldots + (k-1)hX_k}{hX_1 + 2\mathrm{h}X_2 + \ldots + khX_k - hXk}.$$

- Označme

$$(12) \qquad \bar{X} = (hX_1 + 2\mathrm{h}X_2 + \ldots + khX_k)/n.$$

- Pak máme

$$\mathrm{e}^{-h/\lambda} = \frac{n\bar{X} - nh}{n\bar{X} - hX_k},$$

- Takže

$$(13) \qquad \lambda = -h/\ln\frac{n\bar{X} - nh}{n\bar{X} - hX_k}.$$

Ověřování exponenciálního rozdělení

- Hodnotu λ z (13) dosadíme do (11) a vypočteme

$$(14) \qquad \chi^2 = \sum_{i=1}^{k} \frac{(X_i - np_i)^2}{np_i}$$

- V případě $\chi^2 \geq \chi^2_{k-2}(\alpha)$ zamítneme hypotézu o exponenciálním rozdělení na hladině, která je asymptoticky rovna α

Literatura

- Anděl J., Matematická statistika, SNTL, Praha, 1978
- Zvára K., Štěpán J., Pravděpodobnost a matematická statistika, MATFYZPRESS,Praha, 2001