Poznámky k předmětu Aplikovaná statistika, 11. téma

Testy založené na χ^2–rozdělení

V přehledu významných rozdělení jsme si uvedli, že Poissonovým rozdělením se modeluje počet událostí, které nastanou v nějaké měrné jednotce. To ovšem platí jen za předpokladu, že události nastávají náhodně a nezávisle na sobě. V reálném životě ovšem nemůžeme mít jistotu, zda jsou předpoklady splněny a jestli se tedy jedná právě o Poissonovo rozdělení (nebo kterékoliv jiné). Následující odstavec se věnuje jednomu z testů rozdělení.

χ^2–test dobré shody

Podívejme se nejdříve na případ, kdy chceme testovat, zda náhodný výběr pochází z diskrétního rozdělení. Spojitý případ je pak jednoduchým zobecněním.

Příklad 1. *Provedli jsme* 600 *hodů šestistěnnou kostkou. Výsledky jsou uvedeny v tabulce četností* (n_i *je četnost počtu ok* x_i, $i = 1, \ldots, 6$):

i	1	2	3	4	5	6
počet ok x_i	1	2	3	4	5	6
četnost n_i	122	80	85	98	125	90

Otestujte na hladině významnosti 1 %*, zda je kostka spravedlivá, tj. zajímá nás, zdali počty ok pocházejí z diskrétního rovnoměrného rozdělení na množině* $\{1, \ldots, 6\}$.

Uvažujme náhodný výběr $X_1, \ldots, X_n$. Chtěli bychom otestovat, zda výběr pochází z rozdělení $\mathcal{L}$, **jehož parametry známe**. Nulová a alternativní hypotéza jsou tvaru

$$H_0 : \text{náhodný výběr pochází z rozdělení } \mathcal{L}$$
$$H_1 : \text{non } H_0 \text{ (náhodný výběr nepochází z rozdělení } \mathcal{L}).$$

Předpokládejme, že, stejně jako v Příkladě 1, máme realizaci náhodného výběru uspořádanou do tabulky četností. Definujme si teoretické četnosti jednotlivých tříd

$$N_i = n \cdot p(x_i), \quad i = 1, \ldots k,$$

kde $p(x_i) = \mathbb{P}(X = x_i)$ pro náhodnou veličinu X mající rozdělení $\mathcal{L}$. Teoretická četnost N_i odpovídá tomu, kolikrát by v n pokusech měla přibližně nastat možnost x_i, pokud výběr skutečně pochází z $\mathcal{L}$. Myšlenka χ^2–testu dobré shody je geniálně jednoduchá. Pokud výběr pochází z hypotetického rozdělení, pak by skutečné četnosti n_i jednotlivých tříd měly být přibližně stejné jako teoretické četnosti N_i. Testová statistika je založena na rozdílech $n_i - N_i$. Její přesný tvar je

$$R = \sum_{i=1}^{k} \frac{(n_i - N_i)^2}{N_i}$$

a za platnosti H_0 má asymptoticky χ^2–rozdělení o $k - 1$ stupních volnosti.

Požadavek, že testované rozdělení známe včetně jeho parametrů, není v běžném životě obvyklý (pokud zrovna nechceme testovat např. rovnoměrné rozdělení jako v Příkladě 1). Předpokládejme proto, že náhodný výběr pochází z rozdělení $\mathcal{L}(\theta)$, kde θ je neznámý parametr. Chceme-li využít výše uvedený postup k testování rozdělení, v němž se počítají teoretické četnosti N_i rozdělení $\mathcal{L}(\theta)$, je nutné parametr θ z dat nějak rozumně odhadnout. V Příkladě 2, v němž chceme testovat, že

Příklad 2. *Z nejmenovaného supermarketu jsme dostali informace o zásilce vajec. Přesněji, kolik rozbitých vajec bylo v jednotlivých (malých) platech pocházejících z této zásilky. Počty jsou uvedeny v tabulce četností, kde n_i značí počet plat, v nichž bylo právě x_i rozbitých vajec.*

i	1	2	3	4	5	6	7
x_i	0	1	2	3	4	5	6
n_i	28	15	2	4	1	0	0

Nás by zajímalo, zda počty rozbitých vajec v platu mají binomické rozdělení.

výběr pochází z $Bi(6, p)$, bychom mohli odhadnout střední hodnotu μ pomocí průměru $\bar{X}$ a parametr p pak z rovnosti $\mu = 6p$. Musíme však mít na paměti, že zamítneme-li nulovou hypotézu, zamítáme pouze, že výběr pochází z binomického rozdělení s daným (odhadnutým) parametrem p a nikoliv, že pochází z binomického rozdělení obecně.
Takto použitý test tedy **neumí testovat** typ rozdělení, pouze konkrétní rozdělení s konkrétní volbou parametrů. Tento nedostatek je však možné odstranit. Pokud odhadneme neznámý parametr θ rozdělení $\mathcal{L}(\theta)$ tzv. *modifikovanou metodou minimálního* χ^2, lze dokázat, že testová statistika

$$R = \sum_{i=1}^{k} \frac{(n_i - \hat{N}_i)^2}{\hat{N}_i} \ [1]$$

má za platnosti H_0 asymptoticky χ^2–rozdělení o $k - r - 1$ stupních volnosti, kde

- r je počet odhadnutých parametrů rozdělení,
- $\hat{N}_i = n\hat{p}(x_i)$, kde $\hat{p}(x_i) \equiv p(x_i, \hat{\theta})$ jsou odhady $p(x_i)$ založené na modifikované metodě minimálního χ^2.

Jelikož je obvykle nemožné získat $\hat{p}(x_i)$ modifikovanou metodou minimálního χ^2 přesně (jde o řešení složité soustavy rovnic), jako odhady pravděpodobností $p(x_i)$ se běžně berou rozumné odhady spočítané z dat (viz komentář k Příkladu 2). Tímto postupem se v praxi testuje nulová hypotéza, že data pocházejí z daného typu rozdělení $\mathcal{L}(\theta)$ proti alternativě, že tomu tak není.

Poznámka 1. *Kvůli přibližnému charakteru χ^2–testu dobré shody je nezbytné klást požadavky na četnosti jednotlivých tříd. Je třeba, aby $\hat{N}_i \geq 1$, $\forall i$, a $\hat{N}_i \geq 5$ pro alepoň 80% tříd. Pokud nejsou tyto předpoklady splněny, je třeba vhodně sloučit některé třídy (anebo si obstarat další data). Pokud to není možné, nelze test použít.*

[1] Po úpravách lze získat pro numerické výpočty vhodnější tvar

$$R = \sum_{i=1}^{k} \frac{n_i^2}{\hat{N}_i} - n,$$

i když z něj není patrné, jak testová statistika vznikla.

Pro spojitý případ si opět uvedeme ilustrační příklad. Mějme realizaci náhodného výběru ze spojitého rozdělení danou následující tabulkou četností:

zástupce	6	7	8	9	10	38
x	$-\infty$–6, 5	6, 5–7, 5	7, 5–8, 5	8, 5–9, 5	9, 5–10, 5	37, 5–∞
n_i	3	7	20	10	9	1

Teoretické četnosti jednotlivých tříd (či jejich odhady) se spočítají analogicky jako v případě diskrétním, ale pravděpodobnosti $p(x_i)$ jsou nahrazeny pravděpodobnostmi, že náhodná veličina s rozdělením $\mathcal{L}(\theta)$ padne do daného intervalu. Tyto pravděpodobnosti spočteme z hustoty nebo z tabulek v případě, že známe všechny parametry rozdělení. Pokud některé parametry neznáme, odhadneme je z dat.

Testy nezávislosti v kontingenčních tabulkách

Nyní bychom rádi zkoumali závislost barvy očí a barvy vlasů. V předchozí kapitole jsme se seznámili s testem nezávislosti založeným na výběrovém korelačním koeficientu. Pro případ barvy očí a vlasů je ovšem tento test nepoužitelný. Nejedná se totiž o přirozené číselné veličiny, ale o tzv. **kvalitativní veličiny**. Můžeme jim sice přiřadit (pro snadnou manipulaci) číselné hodnoty, ale nemůžeme zavést žádné uspořádání – nemá smysl diskutovat, zda jsou modré oči „víc“ než oči zelené. Pro takový případ je vhodné použít tzv. **kontingenční tabulky**. Jedná se o tabulku četností, jejíž řádky udávají četnosti veličiny X (barva vlasů) a sloupce četnosti veličiny Y (barva očí).

	zelené	modré	hnědé	černé
blond	6	12	4	0
hnědé	11	7	13	3
černé	1	0	9	17
zrzavé	15	8	6	2

Chceme testovat, zda jsou veličiny X a Y nezávislé. Uvažujme obecně kontingenční tabulku (n_{ij}) o r řádcích a c sloupcích. Označme $n = \sum_{i=1}^{r} \sum_{j=1}^{c} n_{ij}$ a dále

$$n_{i\cdot} = \sum_{j=1}^{c} n_{ij},$$

$$n_{\cdot j} = \sum_{i=1}^{r} n_{ij}.$$

Definujme

$$\hat{N}_{ij} = \frac{n_{i\cdot} n_{\cdot j}}{n}. \tag{1}$$

Potom testová statistika pro nulovou hypotézu

$$H_0 : \text{veličiny } X \text{ a } Y \text{ jsou nezávislé},$$

oproti alternativě, že tomu tak není, má tvar

$$R = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(n_{ij} - \hat{N}_{ij})^2}{\hat{N}_{ij}}, \ ^2$$

a za platnosti H_0 má asymptoticky χ^2–rozdělení o $(r-1)(c-1)$ stupních volnosti.

Poznámka 2. *Ačkoliv se mohou oba výše zmiňované testy zdát na první pohled absolutně odlišné, jedná se v podstatě o jeden a tentýž test. Stačí si uvědomit definici nezávislosti náhodných veličin, tedy že (diskrétní) veličiny X a Y jsou nezávislé právě tehdy, když*

$$\mathbb{P}(X = x_i, Y = y_j) = \mathbb{P}(X = x_i)\mathbb{P}(Y = y_j), \ \forall x_i, y_j,$$

zkráceně

$$p_{ij} = p_{i\cdot} \cdot p_{\cdot j}, \ \forall i, j.$$

Jde o testování $H_0: \ p_{ij} = p_{i\cdot} \cdot p_{\cdot j}, \ \forall i, j$, proti alternativě, že tomu tak není. Jedná se tedy o test shody dat se součinovým rozdělením, za platnosti H_0 by pro teoretické četnosti N_{ij} platilo

$$N_{ij} = n \cdot p_{ij} = n \cdot p_{i\cdot} \cdot p_{\cdot j}.$$

Jelikož marginální pravděpodobnosti $p_{i\cdot}$, $p_{\cdot j}$ obecně závisí na parametru θ rozdělení náhodného vektoru (X, Y), je třeba je odhadnout z dat

$$\hat{p}_{i\cdot} = \frac{n_{i\cdot}}{n}, \ \hat{p}_{\cdot j} = \frac{n_{\cdot j}}{n},$$

kde $n_{i\cdot}$, $n_{\cdot j}$ jsou skutečné marginální četnosti. Pak

$$\hat{N}_{ij} = n \cdot \hat{p}_{i\cdot} \cdot \hat{p}_{\cdot j} = n \cdot \frac{n_{i\cdot}}{n} \cdot \frac{n_{\cdot j}}{n} = \frac{n_{i\cdot} n_{\cdot j}}{n}$$

(srovnej s (1)).

CVIČENÍ K TOMUTO TÉMATU

(i) Nechť náhodná veličina X udává procentuální obsah bílkovin v zrnech pšenice. Testujte (na hladině 1 %), zdali má X normální rozdělení, pokud máte k dispozici data:

x_i	do $8,5$	$8,5-9,5$	$9,5-10,5$	$10,5-11,5$	přes $11,5$
n_i	2	10	10	35	33

(ii) Zkoumá se, zdali jsou rodinný stav ženicha (svobodný, rozvedený, vdovec) a nevěsty (svobodná, rozvedená, vdova) na sobě nezávislé na hladině významnosti 5 %. K dispozici jsou počty sňatků mezi jednotlivými skupinami za jeden rok:

[2] Viz poznámku pod čarou 1,

$$R = n \left(\sum_{i=1}^{r} \sum_{j=1}^{c} \frac{n_{ij}^2}{n_{i\cdot} n_{\cdot j}} - 1 \right).$$

Ženich	**Nevěsta**		
	svobodná	rozvedená	vdova
svobodný	7220	190	642
rozvedený	230	20	29
vdovec	760	29	45

(iii) Na základě ankety vykonané u studentů Aplikované statistiky z loňského roku byla statisticky odvozena oblíbenost kofoly (K) a piva (P). Pravěpodobnostní rozdělení je uvedeno v tabulce (symbol $\neg K$ značí negaci ke K, tedy „nepití kofoly“):

$P \wedge K$	$P \wedge \neg K$	$\neg P \wedge K$	$\neg P \wedge \neg K$
$0,35$	$0,3$	$0,25$	$0,1$
63	15	19	9

Poslední řádek udává četnosti jednotlivých možností získaných průzkumem ve vašich paralelkách. Otestujte na hladině 5 % hypotézu, zda jsou vaše chutě stejné jako chutě loňských studentů.

(iv) Byly zjištěny následky 50 lidí ušknutých vzácným druhem hada a také informace, zdali ušknutí lidé užívali léky proti vysokému tlaku:

Užívali léky	**Následky**		
	smrt	silné křeče, přežili	žádné následky
ano	14	6	7
ne	9	3	11

Rozhodněte, zda jsou následky uštknutí významně závislé na tom, zda uštknutí lidé používali léky proti vysokému tlaku (na hladině významnosti 10 %).

(dcv) V jedné lokální fotbalové lize se sledoval počet vstřelených branek za jednu sezónu:

počet branek	0	1	2	3	4 a více
četnost	19	30	17	10	8

Zjistěte na hladině významnosti $\alpha = 5$ %, zda počet vstřelených branek pochází z Poissonova rozdělení.