Marie Turčičová

Úvod do analýzy kategoriálních dat

V tomto textu si ozřejmíme základní postupy pro analýzu jedné či dvou kategoriálních veličin. Podíváme se jak otestovat, že pravděpodobnosti jednotlivých kategorií jsou rovny daným číslům a dále se podíváme na základní testy v kontingenčních tabulkách.

1. Multinomické rozdělení

S kategoriálními daty je neodmyslitelně spjato multinomické rozdělení. Vyskytuje se všude tam, kde máme nějakou veličinu, která nabývá konečného počtu kategorií.

Představme si, že máme n nezávislých pokusů a v každém z nich musí nastat právě jedna z k kategorií. Pravděpodobnosti jednotlivých kategorií si označme $(\pi_1, \pi_2, \ldots, \pi_k)$. (Samozřejmě předpokládejme, že tyto pravděpodobnosti se v průběhu těch n pokusů nemění). Počty pokusů, které spadly do jednotlivých kategorií si označme $(Y_1, Y_2, \ldots, Y_k)$. Náhodný vektor $(Y_1, Y_2, \ldots, Y_k)$ má potom multinomické rozdělení s parametry $n, \pi_1, \pi_2, \ldots, \pi_k$, což formálně zapíšeme jako

$$(Y_1, Y_2, \ldots, Y_k) \sim M(n, (\pi_1, \pi_2, \ldots, \pi_k)) \tag{1}$$

Zdůrazňuji, že se jedná o rozdělení celého vektoru $(Y_1, Y_2, \ldots, Y_k)$, tj. je to rozdělení mnohorozměrné. Jinými slovy by taky se dalo říct, že je to sdružené rozdělení veličin $Y_1, Y_2, \ldots, Y_k$. Samozřejmě musí platit, že

- $\pi_1 + \pi_2 + \ldots + \pi_k = 1$ (protože některá z k kategorií se vždy musí realizovat)
- $Y_1 + Y_2 + \ldots + Y_k = n$ (součet realizací jednotlivých kategorií mi musí dát celkový počet pokusů).

Z toho, že $Y_1 + Y_2 + \ldots + Y_k = n$ mimo jiné vidíme i to, že veličiny $Y_1, Y_2, \ldots, Y_k$ jsou závislé. Známe-li totiž hodnoty $k-1$ z nich, hodnotu zbývající veličiny snadno dopočteme z podmínky, že jejich součet je n. Tuto vlastnost popisujeme slovy tak, že k-tice náhodných veličiny $Y_1, Y_2, \ldots, Y_k$ má $k-1$ stupňů volnosti.

Možná jste si všimli, že v případě, kdy máme pouze dvě kategorie, tj. $k = 2$, tak se multinomické rozdělení shoduje s binomickým. Stejně tak všechna marginální rozdělení (tj. rozdělení jednotlivých Y_j ve vektoru $Y_1, Y_2, \ldots, Y_k$) jsou binomická, to jest $Y_j \sim Bi(n, \pi_j)$.

U všech diskrétních rozdělení jsme si vždy udávali vzoreček pro pravděpodobnost, že daná veličina nabyde nějakou konkrétní hodnotu. I zde se tedy podívejme, jaká je pravděpodobnost, že náhodný vektor $(Y_1, Y_2, \ldots, Y_k)$ nabyde hodnoty $(n_1, n_2, \ldots, n_k)$:

$$P(Y_1 = n_1, Y_2 = n_2, \ldots, Y_k = n_k) = \frac{n!}{n_1! n_2! \ldots n_k!} \pi_1^{n_1} \pi_2^{n_2} \ldots \pi_k^{n_k}, \tag{2}$$

kde vykřičník značí faktoriál daného čísla.

Na závěr si uveďme příklady, kde lze v běžném životě multinomické rozdělení potkat:

- Osudí a v něm koule k barev, provádíme n tahů s vracením[1]: pravděpodobnost vytažení kuličny i-té barvy by byla π_i a Y_i by byl počet kuliček i-té barvy, které byly takto vybrány.
- n hodů kostkou, kde si v každém hodu zaznamenáváme, která hodnota na kostce padla. Je-li kostka spravedlivá, pak $(Y_1, Y_2, \ldots, Y_6) \sim M(n, (\frac{1}{6}, \frac{1}{6}, \ldots, \frac{1}{6}))$.
- Krevní skupiny ($k = 4$): u n náhodně vybraných osob určíme jejich krevní skupiny, pak počty osob s jednotlivými skupinami (tj. vektor (Y_1, Y_2, Y_3, Y_4)) mají multinomické rozdělení.

[1] Rozmyslete si, že pokud bychom tahali bez vracení, už by to nebylo multinomické rozdělení, neboť pravděpodobnosti $(\pi_1, \pi_2, \ldots, \pi_k)$ by se v průběhu tahů měnily.

1.1 χ^2 test dobré shody pro multinomické rozdělení

Testy dobré shody tvoří skupinu testů, které zkoumají shodu teoretických četností/pravděpodobností s nějakými předepsanými hodnotami. Ověřují, jak dobře se naše data shodují s tím, co bychom očekávali za nulové hypotézy.

Asi nejlepším úvodem pro pochopení principu testů dobré shody je následující jednoduchý příklad. Pokusme se ověřit, zda je hrací kostka spravedlivá, tj. chceme testovat hypotézu, že pravděpodobnosti jednotlivých hodnot na kostce jsou vždy 1/6, to jest

$$H_0 : (\pi_1, \pi_2, \ldots, \pi_6) = \left(\frac{1}{6}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6}\right) \tag{3}$$

$$H_1 : (\pi_1, \pi_2, \ldots, \pi_6) \neq \left(\frac{1}{6}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6}\right). \tag{4}$$

Provedli jsme 100 hodů a zjistili jsme, kolikrát padly jednotlivé hodnoty:

hodnota	1	2	3	4	5	6
naměřená četnost	5	39	8	26	10	12

Pokud je kostka spravedlivá (což je naše nulová hypotéza), měly by mít všechny hodnoty shodnou pravděpodobnost 1/6, a tudíž bychom očekávali, že naměříme

hodnota	1	2	3	4	5	6
očekávaná četnost	$\frac{100}{6}$	$\frac{100}{6}$	$\frac{100}{6}$	$\frac{100}{6}$	$\frac{100}{6}$	$\frac{100}{6}$

Potřebovali bychom tedy nějakou statistiku (test), která bude umět porovnávat očekávané četnosti s naměřenými. A to přesně umí statistika χ^2-testu dobré shody.

Testujeme-li obecnou hypotézu

$$\begin{aligned} H_0 &: (\pi_1, \pi_2, \ldots, \pi_k) = \left(\pi_1^0, \pi_2^0, \ldots, \pi_k^0\right) \\ H_1 &: (\pi_1, \pi_2, \ldots, \pi_k) \neq \left(\pi_1^0, \pi_2^0, \ldots, \pi_k^0\right) \end{aligned} \tag{5}$$

kde $(\pi_1^0, \pi_2^0, \ldots, \pi_k^0)$ jsou daná čísla, jejichž shodu se skutečnými pravděpodobnostmi chceme otestovat (v příkladu s kostkou byly všechny $\pi_j^0 = \frac{1}{6}$), pak **testová statistika χ^2-testu dobré shody** má tvar

$$\mathsf{X}^2 = \sum_{j=1}^{k} \frac{(Y_j - n\pi_j^0)^2}{n\pi_j^0}. \tag{6}$$

Jak je patrné zejména z tvaru čitatele, tato statistika porovnává napozorované četnosti $(Y_1, Y_2, \ldots, Y_k)$ s četnostmi $\left(n\pi_1^0, n\pi_2^0, \ldots, n\pi_k^0\right)$, které bychom očekávali, že naměříme, pokud by platila nulová hypotéza. Statistika (6) má při platnosti H_0 asymptoticky (tj. pro velká n, v literatuře se většinou uvádí požadavek $n\pi_j^0 \geq 5$ pro všechny $j = 1, \ldots, k$) rozdělení chí-kvadrát s $k-1$ stupni volnosti, které se symbolicky zapisuje jako χ^2_{k-1}. Nulovou hypotézu (5) zamítáme pokud

$$\boxed{\text{statistika } \mathsf{X}^2 \geq q\chi^2_{k-1}(1-\alpha),}$$

kde α je zvolená hladina testu a $q\chi^2_{k-1}$ značí kvantil rozdělení χ^2_{k-1}.

Připomeňme ještě terminologii:

$$\begin{aligned} Y_1, Y_2, \ldots, Y_k &= \text{ empirické (napozorované) četnosti} \\ n\pi_1^0, n\pi_2^0, \ldots, n\pi_k^0 &= \text{ teoretické (očekávané) četnosti (očekávané za } H_0). \end{aligned}$$

Poznámka 1 *Testová statistika (6) se ze zvyku označuje velkým "X", neboť většina softwaru (a dříve psacích strojů) neumí napsat řecké chí: χ.*

Poznámka 2 *Tvar statistiky (6) je univerzální a jsou na něm založeny všechny testy dobré shody, které uvidíme dále. Klíčovým prvkem u každé hypotézy je pouze odvození očekávaných četností, které bychom za platnosti H_0 měli naměřit.*

2. Kontingenční tabulky

Kontingenční tabulka je tabulka obsahující napozorované četnosti nějakých dvou kategoriálních znaků. Označíme-li si tyto znaky X (má I kategorií) a Y (má J kategorií) a je-li n_{ij} napozorovaná četnost jevu $[X = i, Y = j]$, pak příslušná kontingenční tabulka vypadá následovně:

		X					
		1	2	3	...	I	
	1	n_{11}	n_{12}	n_{13}	...	n_{1I}	$n_{1\cdot}$
	2	n_{21}	n_{22}	n_{23}	...	n_{2I}	$n_{2\cdot}$
Y	3	n_{31}	n_{32}	n_{33}	...	n_{3I}	$n_{3\cdot}$
	⋮						
	J	n_{J1}	n_{J2}	n_{J3}	...	n_{JI}	$n_{J\cdot}$
		$n_{\cdot 1}$	$n_{\cdot 2}$	$n_{\cdot 3}$	...	$n_{\cdot I}$	n

Tabulka 1: Tabulka napozorovaných četností

kde $n_{i\cdot} = \sum_{j=1}^{J} n_{ij}$ a $n_{\cdot j} = \sum_{i=1}^{I} n_{ij}$ jsou tzv. **marginální četnosti**. Celkový počet měření je označen n. Takovým teoretickým protějškem Tabulky 1 je tabulka obsahující pravděpodobnosti $\pi_{ij} = P([X = i, Y = j])$ v populaci. Tato tabulka má analogický tvar

		X					
		1	2	3	...	I	
	1	π_{11}	π_{12}	π_{13}	...	π_{1I}	$\pi_{1\cdot}$
	2	π_{21}	π_{22}	π_{23}	...	π_{2I}	$\pi_{2\cdot}$
Y	3	π_{31}	π_{32}	π_{33}	...	π_{3I}	$\pi_{3\cdot}$
	⋮						
	J	π_{J1}	π_{J2}	π_{J3}	...	π_{JI}	$\pi_{J\cdot}$
		$\pi_{\cdot 1}$	$\pi_{\cdot 2}$	$\pi_{\cdot 3}$	...	$\pi_{\cdot I}$	1

Tabulka 2: Tabulka teoretických pravděpodobností

kde opět $\pi_{i\cdot} = \sum_{j=1}^{J} \pi_{ij}$ a $\pi_{\cdot j} = \sum_{i=1}^{I} \pi_{ij}$ jsou **marginální pravděpodobnosti**.

Analýza kontingenční tabulky nám může pomoci rozhodnout o nezávislosti daných dvou znaků, nebo třeba o symetrii jejich rozdělení. Nejtěžším úkolem při konstrukci takových testů je určit to, jak by měla za platnosti nulové hypotézy vypadat tabulka teoretických pravděpodobností. Zbytek už je pak snadný, protože stačí pomocí testové statistiky (6) porovnat napozorované četnosti n_{ij} s očekávanými četnostmi $n\pi_{ij}$.

Testy dobré shody v kontingenčních tabulkách

2.1 Test nezávislosti

Chceme testovat hypotézu

$$H_0 : \text{znaky } X \text{ a } Y \text{ jsou nezávislé}$$
$$\text{proti } H_1 : \text{znaky } X \text{ a } Y \text{ jsou závislé.}$$

Máme napozorované četnosti n_{ij} jevů $[X = i, Y = j]$, tj. máme k dispozici tabulku napozorovaných četností, a chceme použít test dobré shody. Jaké budou ale očekávané četnosti? Jak se do tabulky teoretických pravděpodobností promítne nezávislost?

Z definice nezávislosti náhodných veličin víme, že jsou-li X a Y nezávislé, tak platí

$$P(X = i, Y = j) = P(X = i)P(Y = j)$$

což v našem značení je: $\pi_{ij} = \pi_{i\cdot}\pi_{\cdot j}$

Očekávané četnosti za nulové hypotézy tedy budou

$$o_{ij} = n\pi_{ij} = n\pi_{i\cdot}\pi_{\cdot j}.$$

Bohužel ale marginální pravděpodobnosti $\pi_{i\cdot}$ a $\pi_{\cdot j}$ neznáme, a musíme si je tudíž odhadnout z marginálních četností:

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

kde stříška značí „odhad“. Odhady četností, které bychom za H_0 očekávali, tedy jsou:

$$\hat{o}_{ij} = n\hat{\pi}_{ij} = n\hat{\pi}_{i\cdot}\hat{\pi}_{\cdot j} = n\frac{n_{i\cdot}}{n}\frac{n_{\cdot j}}{n} = \frac{n_{i\cdot}n_{\cdot j}}{n}. \tag{7}$$

K porovnání očekávaného a napozorovaného opět použijeme χ^2-statistiku

$$\mathsf{X}^2 = \sum_{i=1}^{I}\sum_{j=1}^{J}\frac{(n_{ij} - \hat{o}_{ij})^2}{\hat{o}_{ij}}. \tag{8}$$

V každém políčku tabulky porovnáme četnost napozorovanou s odhadem četnosti očekávané a výsledek posčítáme přes všechna políčka tabulky. Pokud H_0 platí, tak statistika (8) má asymptoticky χ^2-rozdělení s $(I-1)(J-1)$ stupni volnosti. Na hladině α tedy zamítáme nulovou hypotézu o nezávislosti pokud

$$\boxed{\mathsf{X}^2 \geq q\chi^2_{(I-1)(J-1)}(1-\alpha).}$$

Zdůrazněme ještě jednou, že test je pouze asymptotický a k jeho korektnímu použití je potřeba, aby všechny očekávané četnosti byly aspoň 5, tj. u všech políček tabulky musíme zkontrolovat, že $\hat{o}_{ij} \geq 5$.

Poznámka 3 *Všimněte si, že u testu dobré shody pro multinomické rozdělení (statistika (6)) jsme očekávané četnosti za H_0 znali přesně, protože hodnoty pravděpodobností π_{ij} byly nulovou hypotézou explicitně dané. U testu nezávislost tyto pravděpodobnosti neznáme (víme jen, že π_{ij} by mělo být součinem marginálních pravděpodobností), a musíme je tedy odhadnout.*

Poznámka 4 *V určitém kontextu je test nezávislosti nazýván **testem homogenity**. Typicky se jedná o situace, kdy jedna veličina definuje kategorie zkoumaného znaku a druhá veličina definuje např. oblasti, časová období, nebo nějakým jiným způsobem strukturuje populaci. Příkladem může být situace, kdy veličina X představuje krevní skupiny a veličina Y čtyři zeměpisné oblasti. Otázkou pak je, zda se rozložení krevních skupin v jednotlivých oblastech liší. Náhodný vektor (X_A, X_B, X_0, X_{AB}) představující četnosti krevních skupin mezi zkoumanými jedinci v jedné dané oblasti má multinomické rozdělení s pravděpodobnostmi $(\pi_A, \pi_B, \pi_0, \pi_{AB})$. Shodné rozložení krevních skupin v jednotlivých oblastech odpovídá tomu, že vektor pravděpodobností $(\pi_A, \pi_B, \pi_0, \pi_{AB})$ je pro všechny čtyři oblasti stejný. Nicméně tento test je zřejmě ekvivalentní tomu, že veličiny „krevní skupina“ a „oblast“ jsou nezávislé. Proto také testová statistika vypadá stejně. Rozlišení mezi testem nezávislosti a homogenity je tedy spíše filozofické.*

2.2 Test nezávislosti - Fisherův faktoriálový test

Test nezávislosti pro kontingenční tabulky založený na statistice X^2 má dvě základní nevýhody:

- musejí se odhadnout pravděpodobnosti π_{ij}
- u testové statistiky je známo pouze asymptotické rozdělení a test tedy není vhodný pro malé výběry, pro které tato asymptotika nefunguje.

Oba tyto problémy odstranil Fisherů faktoriálový (nebo též exaktní) test. Ten je určen zejména pro malé četnosti (tj. pokud není splněno, že $\hat{o}_{ij} \geq 5\ \forall i, j$) a vystupují v něm faktoriály napozorovaných četností n_{ij}. Netradiční na tomto testu je to, že nepočítá žádnou testovou statistiku, ale pomocí pravděpodobnosti realizace naší konkrétní datové tabulky při platnosti hypotézy nezávislosti počítá přímo p-hodnotu. Více si o něm povíme až u „čtyřpolních“ tabulek, ale v `R` je dostupný pro kontingenční tabulky libovolného rozměru.

2.3 Test symetrie - Bowkerův

Další věcí, kterou lze pomocí kontingenčních tabulek zkoumat, je to, zda je sdružené rozdělení vektoru (X, Y) symetrické, tj. zda $P(X = i, Y = j) = P(X = j, Y = i)$. Tento test je přirozeně určený pouze pro čtvercové tabulky, tj. když $I = J$.

Testovaná dvojice hypotéz má tvar:

$$H_0 : \pi_{ij} = \pi_{ji} \text{ pro všechny dvojice } (i, j)$$
$$\text{proti: } H_1 : \pi_{ij} \neq \pi_{ji} \text{ pro některou dvojici } (i, j).$$

Testujeme-li shodu π_{ij} a π_{ji}, přišlo by nám logické to provést prostřednictvím porovnání n_{ij} s n_{ji} . Není tedy překvapením, že testová statistika má tvar

$$Q = \sum_{i=1}^{I} \sum_{j=1}^{i-1} \frac{(n_{ij} - n_{ji})^2}{n_{ij} + n_{ji}} \tag{9}$$

a za platnosti H_0 má asymptoticky rozdělení $\chi^2_{I(I-1)/2}$. Příslušný test se nazývá Bowkerův a nulovu hypotézu zamítáme na hladině α, pokud

$$\boxed{Q \geq q\chi^2_{I(I-1)/2}(1-\alpha),}$$

kde $q\chi^2_{I(I-1)/2}(1-\alpha)$ označuje $(1-\alpha)\cdot 100\%$ kvantil χ^2-rozdělení s $I(I-1)/2$ stupni volnosti.

Poznámka 5 *(pro zájemce) Statistika (9) neuvažuje počty případů, kdy oba znaky nabyly stejné hodnoty (tj. vynechává četnosti n_{ii}), protože tato informace je pro test symetrie bezcenná. Dále si všimněte, že sčítání ve statistice Q probíhá pouze přes $i < j$. Je to proto, že $(n_{ij} - n_{ji})^2 = (n_{ji} - n_{ij})^2$.*

Poznámka 6 *V `R` najdete Bowkerův test pod názvem **McNemarův**, což není příliš šťastné, protože McNemarův test je pouze speciálním případem uvedeného testu symetrie, který odvodil Bowker v roce 1948.*

3. Testy o parametrech binomického rozdělení

3.1 Pro jedno binomické rozdělení

Představme si, že máme veličinu Y, která má binomické rozdělení $Bi(n, \pi)$ a chceme testovat hypotézu

$$H_0 : \pi = \pi_0 \tag{10}$$
$$\text{proti } H_1 : \pi \neq \pi_0,$$

kde π_0 je hodnota pravděpodobnosti, kterou chceme testovat. Například máme minci a chceme zjistit, zda rub i líc padají se stejnou pravděpodobností 1/2. Veličina Y představuje počet líců v n hodech touto mincí a je to náhodná veličina s rozdělením $Bi(n, \pi)$. Chceme-li ověřit, že mince je spravedlivá, potřebujeme otestovat hypotézu $H_0 : \pi = \frac{1}{2}$ proti $H_1 : \pi \neq \frac{1}{2}$. Tedy v obecné formulaci (10) bereme $\pi_0 = \frac{1}{2}$.

χ^2 test dobré shody (alias Raoův skórový test)

Stačí si uvědomit, že vektor $(Y, n-Y)$ má multinomické rozdělení $M(n, (\pi, 1-\pi))$ a použít χ^2-test dobré shody ze sekce 1.1. Testová statistika (6) zde má tvar

$$\mathsf{X}^2 = \frac{(Y - n\pi_0)^2}{n\pi_0} + \frac{(n - Y - n(1-\pi_0))^2}{n(1-\pi_0)} = \frac{(Y - n\pi_0)^2}{n\pi_0} + \frac{(Y - n\pi_0)^2}{n(1-\pi_0)} = \frac{(Y - n\pi_0)^2}{n\pi_0(1-\pi_0)}$$

a má asymptoticky za H_0 rozdělení χ^2_1. Na hladině α tedy zamítneme nulovou hypotézu (10), pokud

$$\mathsf{X}^2 \geq q\chi^2_1(1-\alpha).$$

V kontextu binomického rozdělení je tento test znám jako Raoův[2] skórový test.

[2] Calyampudi Radhakrishna Rao (1920 - nyní), indický matematik a statistik

3.2 Porovnání dvou binomických rozdělení

Může též nastat situace, že máme dvě náhodné veličiny s binomickým rozdělením:

$$Y_1 \sim Bi(n_1, \pi_1)$$
$$Y_2 \sim Bi(n_2, \pi_2)$$

a budeme chtít otestovat hypotézu

$$H_0 : \pi_1 = \pi_2 \qquad (11)$$
$$\text{proti } H_1 : \pi_1 \neq \pi_2.$$

Pravděpodobnosti π_1 a π_2 si odhadneme pomocí relativní četnosti

$$\hat{\pi}_1 = \frac{Y_1}{n_1} \qquad \hat{\pi}_2 = \frac{Y_2}{n_2}$$

a tyto odhady porovnáme pomocí testové statistiky

$$X^2 = \frac{(\hat{\pi}_1 - \hat{\pi}_2)^2}{\tilde{\pi}(1-\tilde{\pi})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}, \qquad (12)$$

kde $\tilde{\pi} = \frac{Y_1+Y_2}{n_1+n_2}$ je jakýsi „sdružený“ odhad pravděpodobnosti π. Statistika X^2 má za platnosti H_0 asymptoticky rozdělení χ^2_1, tedy na hladině α zamítáme nulovou hypotézu (11), pokud

$$X^2 \geq q\chi^2_1(1-\alpha).$$

Tento test je také nazýván Raoův skórový.

4. „Čtyřpolní“ kontingenční tabulky

Pokud X i Y nabývají pouze 2 kategorií, tak se kontingenční tabulka nazývá „čtyřpolní“, neboť má uvnitř jen čtyři políčka:

		X		
		1	2	
Y	1	n_{11}	n_{12}	$n_{1\cdot}$
	2	n_{21}	n_{22}	$n_{2\cdot}$
		$n_{\cdot 1}$	$n_{\cdot 2}$	n

		X		
		1	2	
Y	1	π_{11}	π_{12}	$\pi_{1\cdot}$
	2	π_{21}	π_{22}	$\pi_{2\cdot}$
		$\pi_{\cdot 1}$	$\pi_{\cdot 2}$	1

Tabulka 3: Čtyřpolní tabulka napozorovaných četností (vlevo) a teoretických pravděpodobností (vpravo).

4.1 Test symetrie - McNemarův

V případě čtyřpolní tabulky se Bowkerův test nazývá McNemarův a jeho nulová hypotéza má tvar

$$H_0 : \pi_{12} = \pi_{21}$$
$$\text{proti: } H_1 : \pi_{12} \neq \pi_{21}.$$

Testová statistika (9) se redukuje na tvar

$$Q = \frac{(n_{12} - n_{21})^2}{n_{12} + n_{21}} \qquad (13)$$

a pokud platí H_0, tak má tato statistika asymptoticky (tj. pro velké napozorované četnosti) rozdělení χ^2_1.

Testy nezávislosti ve čtyřpolní tabulce

Předpokládejme, že u našich dvou veličin X a Y chceme vyšetřit nezávislost:

$$\begin{aligned} H_0 &: X \text{ a } Y \text{ jsou nezávislé} \\ \text{proti } H_1 &: X \text{ a } Y \text{ jsou závislé.} \end{aligned} \tag{14}$$

4.2 Test nezávislosti pomocí χ^2-testu dobré shody

χ^2 statistiku (8) lze pro čtyřpolní tabulku přepsat na tvar

$$\mathsf{X}^2 = \frac{n(n_{11}n_{22} - n_{12}n_{21})^2}{n_{1.}n_{2.}n_{.1}n_{.2}}. \tag{15}$$

Tato statistika má samozřejmě při platnosti H_0 opět asymptoticky rozdělení χ^2_1 a aby tato aproximace byla dobrá, je potřeba aby všechny očekávané četnosti byly ≥ 5. Na hladině α zamítáme nulovou hypotézu (14), pokud

$$\mathsf{X}^2 \geq q\chi^2_1(1-\alpha).$$

Je-li některá očekávaná četnost < 5, navrhl statistik Frank Yates[3] tzv. korekci kontinuity (opravu na spojitost), která zlepšuje aproximaci skutečného rozdělení statistiky X^2 rozdělením χ^2_1. Testová statistika s „Yatesovou korekcí“ má tvar

$$\mathsf{X}^2_{corr} = \frac{n(|n_{11}n_{22} - n_{12}n_{21}| - \frac{n}{2})^2}{n_{1.}n_{2.}n_{.1}n_{.2}}$$

a při rozhodování o nulové hypotéze (14) ji opět porovnáváme s kvantilem $q\chi^2_1(1-\alpha)$.

4.3 Test nezávislosti - Fisherův faktoriálový

Tento test se někdy též nazývá „přesný“ nebo „exaktní“. Netradiční na něm je to, že nepočítá hodnotu testové statistiky, ale počítá přímo p-hodnotu, a to následujícím způsobem. Pravděpodobnost, že se realizuje tabulka s marginálními četnostmi $n_{1.}, n_{2.}, n_{.1}, n_{.2}$ je rovna

$$P(n_{1.}, n_{2.}, n_{.1}, n_{.2}) = \frac{n_{1.}!\, n_{2.}!\, n_{.1}!\, n_{.2}!}{n!\, n_{11}!\, n_{22}!\, n_{12}!\, n_{21}!},$$

kde vykřičník značí faktoriál daného čísla. Právě pro výskyt faktoriálů v tomto vzorci se testu říká „Fisherův faktoriálový“. P-hodnota je pak rovna součtu těchto pravděpodobností od všech tabulek, které mají stejné marginální četnosti jako naše tabulka, ale jejichž četnosti $n_{11}, n_{22}, n_{12}, n_{21}$ ještě více svědčí[4] proti H_0 než ty v naší tabulce.

Během odvozování vzorce pro pravděpodobnost výše se ve výpočtu objevuje výraz pro teoretický poměr šancí. Podívejme se nyní na jeho význam. **Šance** (angl. odds) na výskyt jevu A je

$$\mathcal{O}(A) = \frac{P(A)}{1 - P(A)}.$$

Pokud máme onemocnění, u něhož je pravděpodobnost nákazy 1/3, pak šance na nakažení je $\frac{1/3}{2/3} = \frac{1}{2}$, tedy 1:2.

Nyní se podívejme na první řádek Tabulky 3 (vpravo):

		X		
		1	2	
Y	1	π_{11}	π_{12}	$\pi_{1.}$

[3] 1902-1994, anglický statistik

[4] to „svědčí proti H_0“ se kvantifikuje pomocí poměru šancí, o kterém si povíme v zápětí

Vidíme, že $\pi_{1.}$ je pravděpodobnost jevu $[Y = 1]$ a π_{11} je pravděpodobnost jevu $[X = 1, Y = 1]$. Z definice podmíněné pravděpodobnosti tedy máme

$$P(X = 1|Y = 1) = \frac{P(X = 1, Y = 1)}{P(Y = 1)} = \frac{\pi_{11}}{\pi_{1.}},$$

a tudíž šance na $[X = 1]$ za podmínky $[Y = 1]$ je

$$\mathcal{O}(X = 1|Y = 1) = \frac{P(X = 1|Y = 1)}{1 - P(X = 1|Y = 1)} = \frac{\frac{\pi_{11}}{\pi_{1.}}}{1 - \frac{\pi_{11}}{\pi_{1.}}} = \frac{\pi_{11}}{\pi_{1.} - \pi_{11}} = \frac{\pi_{11}}{\pi_{12}}.$$

Jsou-li veličiny X a Y nezávislé, měla by šance jevu $[X = 1]$ být stejná pro $Y = 1$ i pro $Y = 2$, tj. mělo by platit $\mathcal{O}(X = 1|Y = 1) = \mathcal{O}(X = 1|Y = 2)$. A tedy příslušný **poměr šancí** (odds ratio)

$$\beta = \frac{\mathcal{O}(X = 1|Y = 1)}{\mathcal{O}(X = 1|Y = 2)} = \frac{\frac{\pi_{11}}{\pi_{12}}}{\frac{\pi_{21}}{\pi_{22}}} = \frac{\pi_{11}\pi_{22}}{\pi_{12}\pi_{21}} \tag{16}$$

by měl být roven 1. Hypotézy (14) lze tedy také zformulovat jako

$$H_0 : \beta = 1$$
$$H_1 : \beta \neq 1.$$

Odhadem teoretického poměru šancí (16) je empirický poměr šancí

$$b = \frac{\frac{n_{11}}{n}\frac{n_{22}}{n}}{\frac{n_{12}}{n}\frac{n_{21}}{n}} = \frac{n_{11}n_{22}}{n_{12}n_{21}}, \tag{17}$$

kde jsou jednotlivé pravděpodobnosti odhadnuty pomocí relativních četností. Empirický poměr šancí je také součástí výstupu funkce `fisher.test` v Rku.

Poznámka 7 *Pokud bychom si poměr šancí definovali jako*

$$\beta = \frac{\mathcal{O}(Y = 1|X = 1)}{\mathcal{O}(Y = 1|X = 2)},$$

tj. prohodili bychom roli X a Y, nakonec bychom došli k úplně stejnému výsledku, že totiž $\beta = \frac{\pi_{11}\pi_{22}}{\pi_{12}\pi_{21}}$.

4.4 Test nezávislosti pomocí porovnání dvou binomických rozdělení

Třetí způsob jak testovat nezávislost ve čtyřpolní kontingenční tabulce je podívat se na ni z pohledu binomických rozdělení. Podívejme se opět na první řádek levé Tabulky 3:

		X		
		1	2	
Y	1	n_{11}	n_{12}	$n_{1.}$

Definujeme-li si jako „úspěch“ to, že $X = 1$, pak četnost n_{11} je počet úspěchů z $n_{1.}$ pokusů, kde pravděpodobnost úspěchu je $P(X = 1|Y = 1) = \pi_{11}$. Tedy n_{11} je realizací veličiny N_{11}, která má binomické rozdělení $Bi(n_{1.}, \pi_{11})$. Podíváme-li se nyní na druhý řádek tabulky, tj. díváme se na počet úspěchů za podmínky, že $Y = 2$, analogicky zjistíme, že n_{21} je realizací veličiny N_{21} s rozdělením $Bi(n_{2.}, \pi_{21})$. Pokud jsou veličiny X a Y nezávislé, tak by pravděpodobnost „úspěchu“ měla být stejná ať už je $Y = 1$, nebo 2, tj. obě ta binomická rozdělení by měla mít stejnou pravděpodobnost. Hypotézy (8) tak lze v tomto případě přeformulovat jako

$$H_0 : \pi_{11} = \pi_{21}$$
$$\text{proti } H_1 : \pi_{11} \neq \pi_{21}.$$

Hypotézu o rovnosti parametrů dvou binomických rozdělení ale testovat umíme, a to pomocí Raova skórového testu ze sekce 3.2. Použije se opět testová statistika (12), která má nyní tvar

$$X^2 = \frac{\left(\frac{N_{11}}{n_{1\cdot}} - \frac{N_{21}}{n_{2\cdot}}\right)^2}{\tilde{\pi}(1-\tilde{\pi})\left(\frac{1}{n_{1\cdot}} + \frac{1}{n_{2\cdot}}\right)}, \tag{18}$$

kde $\tilde{\pi} = \frac{N_{11}+N_{21}}{n_{1\cdot}+n_{2\cdot}}$. Nulovou hypotézu zamítáme, pokud $X^2 \geq q\chi_1^2(1-\alpha)$.

Pokud bychom ve svých úvahách prohodili role X a Y a místo na řádky se dívali na sloupce tabulky, dostali bychom stejnou hodnotu statistiky X^2. Oba postupy jsou tedy ekvivalentní.