Test 2. Jestliže četnosti odpovídají pozorováním v k populacích, pak použijeme testovací statistiku

$$\chi^2 = \frac{\sum (x_i - \bar{x})^2}{\bar{x}},$$

která má za platnosti nulové hypotézy rovnosti parametrů λ_i, $i = 1, 2, \ldots, k$ přibližně χ^2-rozdělení s $k - 1$ stupni volnosti. Hodnota $\bar{x}$ je odhadem společného parametru λ.

PŘÍKLAD 8.5

Ověření rovnosti četností pro více případů jevu řídícího se Poissonovým rozdělením

Pro čtyři záznamy četností určitého jevu jsme získali hodnoty 5, 12, 8 a 19. Odhad společného parametru je $\bar{x} = 11$, počet stupňů volnosti je $4 - 1 = 3$. Testovací statistika $\chi^2 = [(5-11)^2 + (12-11)^2 + (8-11)^2 + (19-11)^2]/11 = 10$. Protože $\chi^2 = 10 > 7{,}81$, kde 7,81 je kritická mez χ^2-rozdělení s 3 stupni volnosti na hladině významnosti 0,05, můžeme zamítnout hypotézu homogenity parametrů λ_i.

8.2 χ^2-test dobré shody

Přezkušujeme, zda tvar pravděpodobnostního rozdělení kategoriální proměnné X má specifikovanou podobu. Při pozorování proměnné X se zjistily četnosti $\{n_i\}$ jednotlivých kategorií. Předpokládáme, že pravděpodobnostní rozdělení proměnné je určené pravděpodobnostmi $\{p_i\}$. Označíme ho symbolem $F_0(x)$. Symbolem $F(x)$ označíme rozdělení, jež náhodná proměnná skutečně má.

Test dobré shody testuje hypotézu:

$$H_0: F(x) = F_0(x) \quad \text{proti alternativě} \quad H_1: F(x) \neq F_0(x)$$

Rozdíl mezi pozorovanými a očekávanými četnostmi zachycuje testovací statistika, která má tvar:

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

kde k = počet možných hodnot kategoriální proměnné,

n_i = pozorovaná četnost v kategorii i,

np_i = teoretická (očekávaná) četnost v kategorii i vypočítaná za předpokladu platnosti H_0, přičemž n označuje rozsah výběru a p_i teoretickou pravděpodobnost kategorie i.

Statistika χ^2 má za platnosti nulové hypotézy asymptoticky χ^2-rozdělení. Při hledání kritické hodnoty použijeme $k - 1$ stupňů volnosti. Jestliže hodnota statistiky χ^2 překročí kritickou mez, signalizuje to špatnou shodu dat s teoretickým rozdělením.

PŘÍKLAD 8.6

Test dobré shody

V n nezávislých náhodných pokusech očekáváme, že četnosti náhodných jevů A_1, A_2, A_3, které v pokusu vůbec mohou nastat, jsou v poměru 1 : 2 : 1. V 80 pokusech jsme získali jejich četnosti 14, 50 a 16. Máme naši hypotézu zamítnout? Pro vypočtení testovací statistiky χ^2 vytvoříme tabulku 8.3. V tomto případě použijeme 2 stupně volnosti. Pro 5% hladinu významnosti je kritická hodnota pro χ^2 rozdělení 5,99. Protože 5,10 < 5,99, nemůžeme naši hypotézu zamítnout.

Tab. 8.3 Příklad výpočtu testovací statistiky pro test dobré shody

n_i	np_i	$n_i - np_i$	$(n_i - np_i)^2$	$(n_i - np_i)^2/np_i$
14	20	−6	36	1,80
50	40	10	100	2,50
16	20	−4	16	0,80
80	80	$\chi^2 = 5{,}10$		

8.3 Závislost kategoriálních proměnných

V tomto odstavci se budeme zabývat statistickou analýzou četnostních tabulek, které vznikají, když popisujeme a analyzujeme vztah kategoriálních proměnných. Jedná se o analogii korelační analýzy spojitých proměnných, kterou jsme probrali v minulé kapitole, nebo o podobnost s analýzou rozptylu, již popíšeme v kapitole následující. Rozdíl mezi oběma metodami spočívá v tom, že v případě analýzy četnostních tabulek obě kategoriální proměnné považujeme za náhodné, zatímco v analýze rozptylu posuzujeme vliv faktoru s určitým počtem hladin jako nezávisle proměnné na chování náhodné závisle proměnné, jež má kategoriální charakter.

PŘÍKLAD 8.7

Analýza závislosti kategoriálních proměnných

V roce 1912 se na své první plavbě srazil luxusní zámořský parník Titanic s plovoucí ledovou krou a potopil se. Někteří cestující se dostali na záchranné čluny, ostatní zemřeli. Představme si, že zkáza Titaniku je experimentem, jak se lidé chovají tváří v tvář smrti, když jenom někteří mohou uniknout. Předpokládáme, že, pasažéři jsou nestranným vzorkem z populace stratifikované podle majetkových poměrů. V tabulce 8.4 uvádíme data zvlášť pro muže a ženy (Lord, 1998 – nejsou zachyceni cestující, u nichž není znám jejich sociální status). Při popisné analýze takovýchto dat se doporučuje uvést údaje v tabulkách jako procenta z řádkových nebo sloupcových součtů. Tím se lépe prezentují rozdílnosti rozdělení v jednotlivých kategoriích. V tabulce 8.5 uvádíme řádková procenta. Procenta nebo absolutní četnosti také zobrazujeme pomocí sloupkových grafů.

Pro jednoduchou inferenční analýzu lze použít metody pro srovnání procent z předchozích odstavců. Snadno lze spočítat, že celkově zemřelo 680 mužů a 168 se jich zachránilo. Žen zemřelo 126, uniknout smrti se podařilo 317. Existuje evidence, že muži v této situaci více umírají? Jaké jsou pro to důvody? Můžeme se však také zeptat, zda existují statisticky významné rozdíly v procentuálních podílech zemřelých žen mezi jednotlivými třídami. Nechceme však srovnávat páry tříd, ale vyhodnotit globální hypotézu, zda vůbec existuje nějaký rozdíl. Stejné vyhodnocení můžeme provést pro muže. Zajímáme se, zda existuje stochastický vztah mezi proměnnou *třída cestujícího* a proměnnou, která popisuje status přežití cestujícího (ANO, NE). Jinak řečeno, ptáme se, zda ovlivňuje proměnná „třída cestujícího" pravděpodobnost přežití cestujícího. Tuto otázku pomohou zodpovědět metody, které nyní popíšeme. Poznamenejme, že náš příklad pracuje dohromady se třemi proměnnými: pohlaví, třída cestujícího a status přežití. Analýzou kontingenčních tabulek vznikajících tříděním podle tří a více kategoriálních proměnných se zabýváme v kapitole 13.9.

Tab. 8.4 Data o cestujících při ztroskotání Titaniku

	Muži		Ženy	
Status	**zemřeli**	**přežili**	**zemřely**	**přežily**
I. třída	111	61	6	126
II. třída	150	22	13	90
III. třída	419	85	107	101

Tab. 8.5 Data o cestujících přepočtená na procenta řádkových součtů

	Muži			Ženy		
Status	**zemřeli**	**přežili**	**počet celkem**	**zemřely**	**přežily**	**počet celkem**
I. třída	64,5 %	35,5 %	172	4,4 %	95,6 %	135
II. třída	84,7 %	15,3 %	177	12,6 %	87,4 %	103
III. třída	83,1 %	16,9 %	504	51,4 %	48,6 %	208

Omezíme se na tabulky dvoudimenzionální, což jsou tabulky vzniklé tříděním podle dvou proměnných. Ve statistice takové tabulky nazýváme **kontingenční tabulky**.

Předpokládáme přitom, že každý jednotlivec, resp. experimentální jednotka populace W může být klasifikována podle dvou proměnných (kritérií) A a B. Proměnná A má r kategorií (úrovní) a proměnná B má s kategorií (úrovní). Označme n_{ij} počet prvků z výběru o rozsahu n, které podle proměnné A patří do kategorie A_i a podle proměnné B do kategorie B_j. Dále označme $n_{i.}$ počet prvků z výběru, které patří do kategorie A_i (bez ohledu na hodnotu proměnné B), a podobně $n_{.j}$ počet prvků patřících do kategorie B_j. Platí následující vztahy:

$$\sum_{i=1}^{r} n_{ij} = n_{.j}, \quad \sum_{j=1}^{s} n_{.j} = n, \quad \sum_{j=1}^{s} n_{ij} = n_{i.}, \quad \sum_{i=1}^{r} n_{i.} = n.$$

Čísla $n_{i.}$, resp. $n_{.j}$ někdy nazýváme marginální řádkové, resp. sloupcové součty kontingenční tabulky. Čísla n_{ij} jsou pozorováním získané četnosti v políčku $[i, j]$ a sestavují se do kontingenční tabulky (tab. 8.6), o níž říkáme, že je typu $r \times s$. Tabulku četností někdy doplňujeme tabulkami řádkových, resp. sloupcových procent, které vztahují v procentech četnosti n_{ij} v políčkách k marginálním řádkovým, resp. sloupcovým součtům (viz tab. 8.7). Také můžeme četnosti n_{ij} vyjádřit v procentech vzhledem k rozsahu výběru n. Všechny tyto tabulky nám usnadňují analýzu původní tabulky. Poznamenejme, že pro kvantitativní proměnné můžeme jejich vhodnou transformací vytvořit kategorie, podle nichž pak třídíme prvky

Tab. 8.6 Konstrukce kontingenční tabulky

Úrovně	B_1	B_2	$\cdots$	B_s	**Součty řádkové**
A_1	n_{11}	n_{12}	$\cdots$	n_{1s}	$n_{1.}$
A_2	n_{21}	n_{22}	$\cdots$	n_{2s}	$n_{2.}$
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
A_r	n_{r1}	n_{r2}	$\cdots$	n_{rs}	$n_{r.}$
Součty sloupcové	$n_{.1}$	$n_{.2}$	$\cdots$	$n_{.s}$	n

Tab. 8.7 Konstrukce tabulky s řádkovými procenty

Úrovně	B_1	B_2	⋯	B_s	Součty řádkové
A_1	$100n_{11}/n_{1.}$	$100n_{12}/n_{1.}$	⋯	$100n_{1s}/n_{1.}$	100
A_2	$100n_{21}/n_{2.}$	$100n_{22}/n_{2.}$	⋯	$100n_{2s}/n_{2.}$	100
⋮	⋮	⋮	⋱	⋮	⋮
A_r	$100n_{r1}/n_{r.}$	$100n_{r2}/n_{r.}$	⋯	$100n_{rs}/n_{r.}$	100
Procenta sloupcová	$100n_{.1}/n$	$100n_{.2}/n$	⋯	$100n_{.s}/n$	

výběru. Tím převedeme analýzu kvantitativních údajů (např. pomocí korelačního koeficientu) do oblasti analýzy kontingenčních tabulek.

Když jsme vytvořili tabulku, začínáme zkoumat vzájemný vztah obou proměnných A a B – nejdříve pomocí vhodného zobrazení, později lze testovat různé hypotézy. Hypotézy pro kontingenční tabulky se obvykle definují v pojmech stochastické nezávislosti, a to pomocí určitých podmínek V kontextu stochastické nezávislosti proměnných A a B tyto podmínky indukují, že čísla $n_{ij}/n_{i.}$, resp. $n_{ij}/n_{.j}$ (řádkové, resp. sloupcové relativní četnosti) jsou pro všechna čísla i, resp. j až na náhodné odchylky konstantní.

Jestliže jednu z proměnných kontrolujeme během výběru – třeba proměnnou A, nazýváme ji faktor. Tato proměnná vlastně určuje r disjunktních subpopulací $W_1, W_2, \ldots, W_r$ z populace W. V tomto případě se může hypotéza nezávislosti popsat jako hypotéza homogenity chování proměnné B vzhledem k faktoru A. Pro oba případy jsou statistické výpočty v podstatě stejné.

Dále si uvedeme podrobnější popisy a příklady pro obě zmíněné výběrové situace.

Hypotéza homogenity

Stručně řečeno tato hypotéza předpokládá, že pravděpodobnostní rozdělení kategoriální proměnné B je stejné v různých populacích, které jsou identifikovány faktorem A. Příslušné statistické testy nazýváme někdy testy dobré shody. Řeší se podobný problém jako v analýze rozptylu, kde porovnáváme shodu průměrů metrických proměnných. Zde však jde o shodu rozdělení kategoriální proměnné. Úrovně faktoru A stratifikují v tomto případě celou populaci W do r disjunktních subpopulací $W_1, W_2, \ldots, W_r$ a každý prvek z W_i je klasifikován do jedné

z kategorií proměnné B. Nechť P_{ij} je relativní četnost prvků subpopulace W_i, jež jsou v j-té kategorii proměnné B. Hypotéza homogenity se potom může vyjádřit následující rovnicí $P_{1j} = P_{2j} = \ldots = P_{rj}$ pro všechna $j = 1, 2, \ldots, s$, což znamená, že pro každou kategorii má být relativní četnost prvků v dané subpopulaci stejná pro všechny subpopulace. Poznamenejme, že úrovně faktoru A určující subpopulace představují hodnoty kvalitativní proměnné, ale proměnná odpovídající proměnné B mohla být původně metrická a teprve nějakou transformací se převedla na proměnnou diskrétní.

Hypotézu homogenity můžeme testovat dále uvedenými metodami, jestliže máme k dispozici prostý náhodný výběr z každé subpopulace určené faktorem A nebo jsme provedli přiřazení objektů do jednotlivých skupin pomocí randomizace.

PŘÍKLAD 8.8

Hypotéza homogenity v kontingenční tabulce ($r = s = 2$)

Populace W studentů je stratifikovaná podle pohlaví a proměnná B je určena tím, zda student má zájem o účast ve školním sportovní oddíle. Je zřejmé, že proměnná B je kategoriální. Dotazování se provádí tak, že zvlášť se provede náhodný výběr 66 chlapců a 74 dívek. Z chlapců, resp. dívek mělo zájem 30, resp. 11 jedinců. Zařazením osob podle zájmu dostaneme tabulku typu 2 × 2, jejíž obecný tvar ukazuje tabulka 8.8.

Jestliže P_{11} je relativní část chlapců se zájmem o sport a P_{21} je relativní část dívek se zájmem o sport v celé škole, pak hypotéza homogenity má tvar: $P_{11} = P_{21}$ (z toho plyne také: $P_{12} = P_{22}$). V pojmech nezávislosti nulová hypotéza vyjadřuje, že relativní četnost jedinců zajímajících se o účast ve sportovním oddíle je nezávislá na pohlaví. Celý problém samozřejmě můžeme převést do procentuálního vyjádření. Výsledky pro náhodný výběr 66 chlapců a 74 dívek obsahuje tabulka 8.9.

Tab. 8.8 Uspořádání dat při testování hypotézy homogenity

	Zájem o sport		Řádkové součty
	ano	ne	
Chlapci	a	b	$a + b$
Dívky	c	d	$c + d$
Sloupcové součty	$a + c$	$b + d$	N

Tab. 8.9 Příklad dat při testování hypotézy homogenity

	Zájem o sport		Řádkové součty
	ano	ne	
Chlapci	30	36	66
Dívky	11	63	74
Sloupcové součty	41	99	140

Hypotéza nezávislosti

V hypotéze nezávislosti se považují obě proměnné A a B za náhodné proměnné, přičemž předpokládáme jejich úplnou nezávislost. To znamená, že hodnota proměnné A neovlivňuje podmíněné rozdělení proměnné B a naopak. Situace je analogická posuzování velikosti korelačního koeficientu dvou metrických proměnných. Uvažujeme populaci W, přičemž každý prvek této populace je klasifikován podle dvou kategoriálních proměnných A a B. Zkoumáme, zda hodnoty proměnné A neovlivňují rozdělení proměnné B a naopak. Nulová hypotéza zní, že obě proměnné jsou na sobě stochasticky nezávislé. Tuto hypotézu lze vyjádřit podmínkami pro pravděpodobnosti p_{ij}, což jsou pravděpodobnosti, že na osobě zjistíme hodnotu proměnné A v kategorii i a hodnotu proměnné B v kategorii j. Nechť $p_{i.}$, resp. $p_{.j}$ je pravděpodobnost v populaci W, že proměnná A nabude hodnoty i, resp. proměnná B nabude hodnoty j. Pak hypotézu nezávislosti obou proměnných můžeme vyjádřit rovnicemi

$$p_{ij} = p_{i.}p_{.j}, \qquad p_{i.} = \sum_{j=1}^{s} p_{ij}, \qquad p_{.j} = \sum_{i=1}^{r} p_{ij},$$

které platí pro všechna i a j ($i = 1, 2, \ldots, r$; $j = 1, 2, \ldots, s$). Uvedené vyjádření vyplývá ze vzorce pro výpočet pravděpodobnosti současného výskytu dvou nezávislých jevů. Tímto vztahem jsme se zabývali v kapitole 7.2.2. Poznamenejme, že jak A, tak B můžeme měřit nejdřív jako kvantitativní proměnné, které poté vhodnou transformací převedeme do diskrétní podoby. Hypotézu nezávislosti můžeme testovat dále uvedenými metodami, jestliže máme k dispozici prostý náhodný výběr z uvažované populace.

PŘÍKLAD 8.9

Hypotéza nezávislosti v kontingenční tabulce

Populace W sestává z žáků středních škol, kteří uvedli nejoblíbenější sport, jejž rádi provozují, a rovněž sport, na nějž se rádi dívají v televizi. Po provedení výběru o rozsahu 234 a zjištění hodnot obou proměnných byla vytvořena kontingenční tabulka pro zkoumání závislosti vztahu obou proměnných (tab. 8.10). Zajímá nás hypotéza H_0: Oblíbenost sledování jednotlivých sportů v televizi nezávisí na oblíbenosti při vlastním sportování.

Tab. 8.10 Příklad kontingenční tabulky zachycující společné rozdělení dvou proměnných pro ověření hypotézy o jejich nezávislosti

Proměnná A oblíbenost při sledování televize	Proměnná B oblíbenost při sportování				Řádkové součty
	hry	atletika	gymn.	plavání	
hry	133	6	2	4	145
atletika	15	10	4	3	32
gymn.	4	1	25	0	30
plavání	9	0	1	17	27
Sloupcové součty	161	17	32	24	234

8.3.1 Posuzování závislosti v kontingenčních tabulkách

Budeme se zabývat obecnou tabulkou typu $r \times s$ a zvlášť čtyřpolní tabulkou typu 2×2, pro kterou jsou výpočty podstatně jednodušší a v podstatě v jiné formě opakují řešení pro porovnání dvou relativních četností. Při analýze kontingenčních tabulek typu $r \times s$ se častěji provádějí testy než odhady. Problém odhadu relativních četností má význam hlavně v tabulce čtyřpolní. Příslušné výpočty čtenář najde v předcházející kapitole o odhadu a testování hypotéz o relativní četnosti.

Tabulka typu $r \times s$

Pro testování hypotéz homogenity i nezávislosti používáme stejný postup. Nejdříve vypočítáme tzv. očekávané frekvence m_{ij} v políčku $[i, j]$ za předpokladu,

že platí nulová hypotéza. Pravděpodobnost p_{ij}, že u objektu zjistíme kombinaci hodnot obou proměnných i a j, musí mít v tomto případě hodnotu $p_{i.}p_{.j}$. Hodnoty obou pravděpodobností odhadneme podíly $n_{i.}/n$ a $n_{j.}/n$. Protože očekávaná hodnota četnosti v políčku $m_{ij} = p_{ij}n$ a odhad p_{ij} je $n_{i.}n_{.j}/n^2$, tak

$$m_{ij} = \frac{n_{i.}n_{.j}}{n}$$

pro $i = 1, 2, \ldots, r$; a $j = 1, 2, \ldots, s$. Tento vztah vyplývá z úvah provedených v kapitole 7.2.2. Čísla n_{ij}/n odhadují pravděpodobnosti p_{ij} stejně jako čísla $n_{i.}n_{.j}/n^2$, ale bez podmínky, že platí nulová hypotéza. Testovací statistiku χ^2 spočteme podle vzorce

$$\chi^2 = \sum \frac{(\textit{pozorované četnosti} - \textit{očekávané četnosti})^2}{\textit{očekávané četnosti}},$$

tedy

$$\chi^2 = \sum_{i=1}^{r} \sum_{j=1}^{s} (n_{ij} - m_{ij})^2 / m_{ij}.$$

Je patrné, že χ^2-statistika měří celkovou nepodobnost čísel n_{ij} a m_{ij}. Čím jsou rozdíly zjištěných a očekávaných četností větší, tím je větší testovací statistika χ^2. Hodnotu χ^2 srovnáme s kritickou hodnotou χ^2-rozdělení o stupních volnosti $(r-1)(s-1)$ na zvolené hladině významnosti. Jestliže hodnota χ^2 je větší než tabulková hodnota, hypotézu zamítáme. Jestliže program dopočítá také p-hodnotu, tak ji srovnáváme se zvolenou hladinou významnosti. Tento test je platný asymptoticky, proto ho můžeme použít pouze při dostatečném počtu pozorování. Všechny očekávané hodnoty by měly být větší než jedna. Jestliže se v některých políčkách vyskytnou nulové hodnoty, přejdeme k analýze odvozené tabulky vzniklé sloučením málo obsazených kategorií.

Jestliže zamítneme hypotézu nezávislosti nebo homogenity, lze tabulku dále analyzovat a hledat důvody, proč je nulová hypotéza porušena. K tomu nám slouží tzv. normalizované reziduální hodnoty $(n_{ij} - m_{ij})/\sqrt{m_{ij}}$, které vyneseme do tabulky (opět typu $r \times s$). Příčinu nehomogenity můžeme zjistit tak, že zopakujeme χ^2-test pro tabulku, jež je zredukována o sloupce nebo řádky, které představují kandidáty nehomogenity. Jestliže tento χ^2-test již nesignalizuje závislost (χ^2-statistika nepřekročí kritickou mez), je podezření potvrzeno. Nebo vybereme čtyři symetricky od sebe položená políčka, jež vždy po dvou leží v jedné řádce nebo sloupci, a vzniklou tabulku 2×2 opět testujeme. Významnost výsledku testu indikuje zdroj poruchy modelu nezávislosti. Poznamenejme, že – podobně jako v korelační analýze – nedokazuje prokázaná závislost kauzální vztah proměnných. Příčiny zdánlivých asociací zjišťujeme analýzou vícerozměrných tabulek, kterým se věnujeme v kapitole 13.9 (srov. též Simpsonův paradox, kap. 8.5).