

8 Analýza kategoriálních dat

Některá data mají spojity charakter s danou jednotkou měření, jako jsou metry, kilogramy apod. Údaje tohoto typu analyzujeme metodami, které jsme popsali v předchozích kapitolách. Při výzkumu se však setkáváme i s kategoriálními údaji, jako je typ zaměstnání, typ automobilu, pohlaví jedince nebo typ stížnosti zákazníka. Ptáme se např.:

- Pomáhá pravidelné užívání malé dávky aspirinu při prevenci srdečního infarktu? Byla provedena znáhodněná dvojitě zaslepená studie, v níž 11037 lékařů užívalo aspirin a 11034 lékařů užívalo placebo. Po pěti letech 104 lékařů z první skupiny zemřelo na srdeční infarkt, ve druhé skupině byl počet takových úmrtí 189. Je tento rozdíl dostatečně veliký, abyhom byli přesvědčeni o preventivním působení aspirinu?
- Jaké důvody uvádějí studenti pro svou účast ve sportovních soutěžích? Liší se důvody pro účast mezi studenty a studentkami?
- Liší se procentuální zastoupení návyku kouření mezi skupinami dospělých s různým sociálním statusem? Abyhom to zjistili, výzkumník zařadí několik stovek záznamů o jedincích podle sociálního statusu (nižší, střední, vyšší) a také podle toho, zda jedinec kouří (kuřák, zanechal kouření, nikdy nekouřil). Statistické hodnocení má zodpovědět, zda existuje závislost mezi proměnnou „sociální status“ a proměnnou „návyk kouření“.

Údaje se mohou týkat nejen jedinců, ale i jiných sledovaných jednotek (školních tříd, institucí, měst, zemí, událostí). Získaná kategoriální data zachycujeme pomocí jedno-, dvou- nebo vícerozměrných tabulek četností nebo relativních četností, procent. Každý rozměr (dimenze) tabulky odpovídá klasifikaci do kategorií podle určité proměnné. Některé proměnné mají podle úlohy charakter závisle proměnné (cílové proměnné), jiné považujeme za nezávislé. Proměnné jsou často nominálního, resp. kvalitativního typu. Také však mohou mít nějaké přirozené řazení (např. vedlejší reakce na lék mohou být žádné, mírné nebo silné) – jsou pak ordinálního typu. Četnostní tabulky vznikají i zařazením jinak spojitych metrických údajů do kategorií, které byly navrženy jako intervaly pokryvající rozsah hodnot sledované proměnné.

Při zkoumání četnostních dat stojíme před podobnými úkoly jako v případě dat metrických. Porovnáváme náhodné chování proměnné s pravděpodobnostním rozdelením, jež je předem přesně specifikované, nebo srovnáváme rozdelení sledované proměnné ve dvou nebo více populacích, aniž bychom předem specifikovali tvar jejich rozdelení. Také nás zajímá síla asociace jednotlivých proměnných mezi sebou.

V této kapitole popisované procedury statistického usuzování vycházejí z toho, že při větších rozsazích výběru můžeme v důsledku působení centrálního limitního teorému přibližně popsat náhodné chování použitých statistik normálním rozdelením nebo χ^2 -rozdelením. Počítací a speciální programy pomáhají realizovat přesné testy při malém počtu pozorování.

8.1 Jednoduché hodnocení četností

Často chceme zodpovědět otázky o velikosti relativní četnosti určité vlastnosti prvků populace. V této kapitole nám proto půjde o zkoumání hypotézy o pravděpodobnosti specifikovaného náhodného jevu a o její intervalový odhad. Jina běžná situace nastává, když chceme porovnat relativní četnosti v několika populacích. Budeme se tedy zabývat i hypotézami, týkajícimi se rozdílu dvou pravděpodobností. Přitom využijeme model binomického rozdelení (kap. 4.5.1). V některých případech je vhodnější popsat chování četností pomocí Poissonova rozdelení. Také pro tento případ uvedeme základní metodu porovnání. Jestliže chceme ověřit, že napozorované relativní četnosti n_1, n_2, \dots, n_k kategoriální proměnné odpovídají teoretickému rozdelení, které je zadáné pravděpodobnostmi p_1, p_2, \dots, p_k , použijeme χ^2 -test dobré shody.

PŘÍKLAD 8.1

Hodnocení četnosti při analýze kategoriálních dat

S analýzou kategoriálních dat a porovnáváním relativních četností se setkáváme běžně při výzkumu veřejného mínění. Moore (1997) uvádí některá data z amerického výzkumu náboru 924 respondentů na roli vojenské cenzury v průběhu války proti Iráku v roce 1991 a z doplňujících výzkumů.

Jedna z otásek v dotazníku zněla: *Myslite si, že vojenské orgány mají více ovlivňovat, jak tiskové agentury informují o válce, nebo si myslíte, že hlavní rozhodnutí o tom, co se uveřejní, mají provést samy tyto agentury?* Názor, že vojenské úřady měly více kontrolovat obsah zpráv, zastávalo 57 % respondentů. V dodatečném výzkumu v univerzitním prostředí dostalo 173 náhodně vybraných studentů formulačně lehce pozměněnou, ale obsahově stejnou otázkou. K větší kontrole zpráv se přiklonilo 55 % nich.

V jiném průzkumu dostalo 198 náhodně vybraných studentů otázku: „*Dominujete se, že míra kontroly tiskových agentur ze strany vojenské administrativy je dostatečná silná?*“ Pouze 16 % studentů vyžadovalo větší cenzuru.

Všechny tři výzkum vycházejí z náhodného výběru. Zjištěné procentuální údaje představují odhadu očekávaných podílů odpovědí v příslušných populacích. Při statistické analýze těchto údajů nás např. zajímá:

- Jačí je 95% interval spolehlivosti pro skutečnou hodnotu podílu populace v prvním výzkumu, která se přikláňá k silnější kontrole zpráv?
- Liší se podíly odpovědí podporujících větší kontrolu zpráv ve druhém a třetím výzkumu tak, že zjištěné rozdíly nelze přičíst působení náhody?

V následujících odstavcích poznáme metody, které nám pomohou tyto otázky zodpovědět.

8.1.1 Porovnání relativní četnosti s teoretickou hodnotou

Posuzujeme relativní četnost přítomnosti určité vlastnosti v populaci pomocí náhodného výběru o rozsahu n . Popíšeme test specifikované hodnoty teoretické relativní četnosti a její intervalový odhad. Test se velmi podobá testu průměrné hodnoty v jednom výběru. Relativní četnost \hat{p} se spočte jako poměr x/n , přičemž x je četnost vlastnosti ve výběru o rozsahu n . Tato hodnota je bodovým odhadem pravděpodobnosti p .

Předpokládejme hodnotu relativní četnosti výskytu sledované vlastnosti p_0 . Testujeme nulovou hypotézu $H_0: p = p_0$ proti alternativní hypotéze $H_1: p \neq p_0$ nebo $H_1: p > p_0$ (nebo $H_1: p < p_0$).

Na základě výsledků z kapitoly 4.5.5 použijeme pro test statistiku z , která měří odchylku empirické relativní četnosti \hat{p} od hypotetické hodnoty p_0 . V tabulce 8.1 uvádíme stručně potřebné výpočty a návod k rozhodování. Například hypotézu H_0 při dvoustranné alternativě nezamítáme, jestliže testovací statistika z leží uvnitř intervalu $\pm z_{\alpha/2}$, kde $z_{\alpha/2}$ je kritická hodnota standardizovaného normálního rozdelení. Jinak vyjádřeno – pokud absolutní hodnota testovací statistiky je větší než $z_{\alpha/2}$, máme evidenci pro zamítnutí hypotézy H_0 .

Dvoustranný interval spolehlivosti pro hladinu spolehlivosti $1 - \alpha$ má tvar:

$$p \in (\hat{p} - z_{\alpha/2} \sqrt{\frac{\hat{p}(1 - \hat{p})}{n}}, \hat{p} + z_{\alpha/2} \sqrt{\frac{\hat{p}(1 - \hat{p})}{n}})$$

Tento interval spolehlivosti lze použít také pro test popsané hypotézy.

Tab. 8.1 Schéma porovnání relativní četnosti s teoretickou hodnotou

Jednostranný test	Dvoustranný test
$H_0: p = p_0$	$H_0: p = p_0$
$H_1: p > p_0$ (nebo $H_1: p < p_0$)	$H_1: p \neq p_0$
Testovací statistika: $z = \frac{\hat{p} - p_0}{\sqrt{p_0 q_0 / n}}, \text{kde } q_0 = 1 - p_0$	Testovací statistika: $z < -z_{\alpha/2} \text{ a } z > z_{\alpha/2}$
Oblast zamítnutí: $z > z_\alpha$ (nebo $z < -z_\alpha$)	Oblast zamítnutí: $z < -z_{\alpha/2} \text{ a } z > z_{\alpha/2}$

PŘÍKLAD 8.2

Konstrukce intervalu spolehlivosti pro četnost kategorálních dat

Zajímá nás odpověď na první otázku z příkladu 8.1 – chceme znát 95% interval spolehlivosti pro skutečnou hodnotu podílu populace v prvním výzkumu, která se přiklání k silnější kontrole zpráv o válce v Iráku. Vypočítáme dvoustranný 95% interval spolehlivosti pro $n = 924$ a bodový odhad $\hat{p} = 0,57$. Zvolíme $z_{\alpha/2} = 1,96$. Dosadíme do vzorce pro intervalový odhad a dostaneme

$$(0,57 - 1,96 \sqrt{\frac{0,57(1-0,57)}{924}}, 0,57 + 1,96 \sqrt{\frac{0,57(1-0,57)}{924}}) = \\ = (0,57 - 0,031; 0,57 + 0,031) = (0,537; 0,601).$$

Hledaný interval spolehlivosti má tedy tvar (53,7%; 60,1%).

Uvedli jsme asymptotické vzorce, které vycházejí z normální approximace rozdělení relativní četnosti a jsou vhodné pouze pro větší rozsahy n : $np > 10$ a $n(1-p) > 10$. Populace má být nejméně desetkrát větší než výběr.

Pro malé výběry n nemá approximace normálním rozdělením oprávnění. Přesný test hypotézy $H_0: p = p_0$ vychází z předpokladu, že pozorovaná četnost x má binomické rozdělení $B(n; p_0)$. Při provedení přesného testu počítáme kumulativní pravděpodobnost realizace četnosti x nebo extrémnejší hodnoty za platnosti nulové hypotézy. Jestliže výsledná pravděpodobnost je menší než hladina významnosti, znamená to evidenci proti platnosti nulové hypotézy. Při výpočtech můžeme použít tabulku kumulativních pravděpodobností binomického rozdělení (tabulka VIII z přílohy B). Například pro výběr $n = 12$ jsme získali četnost $x = 2$ a zkoumáme, zda to odpovídá $p_0 = 0,4$. Pomocí tabulky zjistíme pro jednostranný test kumulativní pravděpodobnost $p(X = 0) + p(X = 1) + p(X = 2) = 0,08344$. Tato hodnota udává dosaženou hladinu významnosti \hat{p} .

Tab. 8.2 Schéma porovnání dvou relativních četností

Jednostranný test	Dvoustranný test
$H_0: (p_1 - p_2) = \Delta$	$H_0: (p_1 - p_2) = \Delta$
$H_1: (p_1 - p_2) > \Delta$ (nebo $H_1: (p_1 - p_2) < \Delta$)	$H_1: (p_1 - p_2) \neq \Delta$
Testovací statistika: $z = \frac{(\hat{p}_1 - \hat{p}_2) - \Delta}{\sqrt{s_{(\hat{p}_1 - \hat{p}_2)}}}$	Testovací statistika: $z > z_\alpha \text{ (nebo } z < -z_\alpha)$
Oblast zamítnutí: $z > z_\alpha$ (nebo $z < -z_\alpha$)	Oblast zamítnutí: $z < -z_{\alpha/2} \text{ a } z > z_{\alpha/2}$

8.1.2 Porovnání dvou relativních četností

Zajímáme se o porovnání dvou pravděpodobností p_1 a p_2 výskytu nějaké vlastnosti ve dvou populacích. Naším cílem je testovat a odhadovat velikost jejich rozdílu $p_1 - p_2$. Popisujeme metodu testování rozdílu $p_1 - p_2$ pomocí asymptoticky platné procedury. Označíme rozdíl pravděpodobností symbolem $\Delta p = p_1 - p_2$. Testová statistika se opírá o standardizovanou odchylku rozdílu empirických četností $\hat{p}_1 - \hat{p}_2$ od předpokládané hodnoty Δ .

Předpokládáme prosté náhodné výběry z obou populací nebo randomizované přiřazení do skupin. Počet prvků se sledovanou vlastností ve výběrových skupinách o rozsahu n_1 a n_2 byl x_1 a x_2 . Teoretické hodnoty p_i odhadujeme pomocí relativních četností $\hat{p}_i = x_i/n_i$. V tabulce 8.2 uvádíme základní schéma rozhodování.

Výpočet odhadu směrodatné odchylky $s_{(\hat{p}_1 - \hat{p}_2)}$ závisí na hodnotě Δ . Označme $\hat{q}_i = 1 - \hat{p}_i$. Jestliže $\Delta \neq 0$, pak

$$s_{(\hat{p}_1 - \hat{p}_2)} = \sqrt{\hat{p}_1 \hat{q}_1 / n_1 + \hat{p}_2 \hat{q}_2 / n_2}.$$

V případě, že $\Delta = 0$, má $s_{(\hat{p}_1 - \hat{p}_2)}$ hodnotu

$$s_{(\hat{p}_1 - \hat{p}_2)} = \sqrt{\hat{p} \hat{q} \left(\frac{1}{n_1} + \frac{1}{n_2} \right)},$$

kde $\hat{p} = (x_1 + x_2)/(n_1 + n_2)$ je spojený odhad teoretické relativní četnosti a $\hat{q} = 1 - \hat{p}$. Kritické hodnoty z_α odpovídají kvantilům standardizovaného normálního rozdělení $N(0; 1)$ s hladinou $1 - \alpha$.

Oba rozsahy n_1 a n_2 musí být dostatečně veliké ($n_1 > 30$ i $n_2 > 30$), abychom mohli pro výběrové rozdělení rozdílu hodnot \hat{p}_1 a \hat{p}_2 uplatnit centrální limitní teorém. Populace mají být nejméně desetkrát větší než výběry.

Dvoustranný interval spolehlivosti pro hladinu spolehlivosti $1 - \alpha$ má tvar:

$$(p_1 - p_2) \in (\hat{p}_1 - \hat{p}_2 - z_{\alpha/2} s_{(\hat{p}_1 - \hat{p}_2)}, \hat{p}_1 - \hat{p}_2 + z_{\alpha/2} s_{(\hat{p}_1 - \hat{p}_2)}),$$

kde $s_{(\hat{p}_1 - \hat{p}_2)} = \sqrt{\hat{p}_1 \hat{q}_1 / n_1 + \hat{p}_2 \hat{q}_2 / n_2}$.

Jestliže podmínka o rozsazích výběru není splněna, ale počty jsou větší než 20, uplatníme arcussinovou transformaci na druhou odmocninu odhadů pravděpodobnosti:

$$\varphi(p) = \arcsin \sqrt{p}$$

Hypotézu rovnosti pravděpodobností pak testujeme pomocí z -statistiky

$$z = \frac{\varphi(p_1) - \varphi(p_2)}{28,648 \sqrt{1/n_1 + 1/n_2}}.$$

Rozptyl empirické hodnoty $\varphi(p)$ již nezávisí na hodnotě p .

Přesný test porovnání dvou relativních četností lze provést Fisherovým testem homogeneity v tabulce četností 2×2 (viz s. 314).

PŘÍKLAD 8.3

Testování hypotézy o rovnosti četností u kategorálních dat

Chceme odpovědět na druhou otázkou z příkladu 8.1 a prozkoumat, zda se liší podíly odpovědí podporujících větší kontrolu zpráv o válce v Iráku ve druhém a třetím výzkumu tak, že zjištěné rozdíly nelze přičíst působení náhody. Hypotézu o rovnosti podílu odpovědi v příslušných populacích otestujeme pomocí approximativního 95% intervalu spolehlivosti. Výpočty vycházejí z těchto údajů:

$$\begin{aligned} n_1 &= 198 & \hat{p}_1 &= 0,16 & \hat{q}_1 &= 1 - 0,16 = 0,84 \\ n_2 &= 173 & \hat{p}_2 &= 0,55 & \hat{q}_2 &= 1 - 0,55 = 0,45 \end{aligned}$$

Testovací situace je

$$\begin{array}{ll} H_0: p_1 = p_2 & H_0: \Delta = 0 \\ \text{nebo jinak vyjádřeno} & H_1: \Delta \neq 0. \end{array}$$

Platí, že $\hat{p}_1 - \hat{p}_2 = 0,16 - 0,55 = -0,39$. Pro 95% interval spolehlivosti na 5% hladině je kritická hodnota $z_{\alpha/2} = z_{0,025} = 1,96$. Základem výpočtu intervalu spolehlivosti je odhad směrodatné chyby pro rozdíl relativních četností:

$$s_{(\hat{p}_1 - \hat{p}_2)} = \sqrt{\frac{\hat{p}_1 \hat{q}_1}{n_1} + \frac{\hat{p}_2 \hat{q}_2}{n_2}} = \sqrt{\frac{0,16 \times 0,84}{198} + \frac{0,55 \times 0,45}{173}} = 0,0459$$

Interval spolehlivosti pro rozdíl pravděpodobností je tedy dán hodnotami

$$(p_1 - p_2) \in (-0,39 - 1,96 \times 0,0459; -0,39 + 1,96 \times 0,0459) = (-0,479; -0,344).$$

Protože interval spolehlivosti nepokrývá 0, můžeme na hladině významnosti 0,05 zamítnout nulovou hypotézu, že v obou skupinách je stejný podíl těch studentů, kteří souhlasí s přísnější kontrolou informačních agentur v době války.

Chceme testovat hypotézu, že podíl jedinců souhlasících se zvýšením kontroly zpravodajství v druhé studentské populaci není větší o více než 10% ve srovnání se zastoupením tohoto názoru v první populaci studentů (předpokládáme, že bylo ověřeno, že význam otázek považují studenti za stejný). Použijeme jednostranný test na 5% hladině významnosti. Kritická mez z -testu je 2,56. Testovací z -statistika má hodnotu

$$z = \frac{(\hat{p}_2 - \hat{p}_1) - \Delta}{s_{(\hat{p}_2 - \hat{p}_1)}} = \frac{(0,55 - 0,16) - 0,10}{0,0459} = 6,318.$$

Vypočtená hodnota z -statistiky svědčí ve prospěch alternativní hypotézy.

8.1.3 Porovnání četností majících Poissonovo rozdělení

Nyní se budeme zabývat situací posouzení četností, které vzniknou sledováním náhodné proměnné s Poissonovým rozdělením (kap. 4.5.2). Hypotéza, že zjištěná četnost x výskytu specifikovaného jevu není v rozporu s předpokladem o Poissonovém rozdělení s parametrem λ o specifikované hodnotě, se zkoumá stejně jako v příkladu na s. 138. Uvedeme (1) test shody dvou parametrů $H_0: \lambda_1 = \lambda_2$ a (2) test shody parametru λ_i ve více než dvou populacích. Máme k dispozici zjištěnou četnost $\{x_i\}$ z nezávislých pozorování.

Test 1. Asymptoticky platný test hypotézy $H_0: \lambda_1 = \lambda_2$ provedeme pomocí statistiky

$$z = \frac{|x_1 - x_2| - 1}{\sqrt{x_1 + x_2}},$$

která má za platnosti nulové hypotézy a pro $x_1 + x_2 > 5$ přibližně rozdělení $N(0; 1)$.

PŘÍKLAD 8.4

Ověření rovnosti četností u jevu řídího se Poissonovým rozdělením

Zjistili jsme četnosti $x_1 = 13$ a $x_2 = 3$. Hypotézu H_0 rovnosti parametru Poissonova rozdělení v obou situacích testujeme pomocí statistiky $z = (13 - 3 - 1)/\sqrt{16} = 2,25$, která vychází větší než příslušná kritická mez 1,96. Máme evidenci pro zamítnutí H_0 na 5% hladině významnosti.

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 , $i = 1, 2, \dots, k$ přibližně χ^2 -rozdělení s $k - 1$ stupni volnosti. Hodnota \bar{x} je odhadem společného parametru λ .

PŘÍKLAD 8.5

Ověření rovnosti četnosti pro více případů jevu řidícího se Poissonovým rozdělením

Pro čtyři záznamy četnosti 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 kritickámez χ^2 -rozdělení s 3 stupni volnosti na hladině významnosti 0,05, můžeme zamítнуть 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 kategorii. 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 kritickoumez, 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 zamítнуть? 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 zamítнуть.

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 popisovali v minulé kapitole, nebo o podobnosti s analýzou rozptylu, již popiseme 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 závisle 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 tvář 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. Stejně 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*“

Tab. 8.4 Data o cestujících při ztroskotání Titaniku

Status	Muži		Ženy	
	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

Status	Muži		Ženy			
	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

jícího“ pravděpodobnost přežití cestujícího. Tuto otázku pomohou zodpovědět metody, které nyní popíšeme. Poznamenejme, že nás 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.

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 (kritérii) 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 $n_{i\cdot}$ počet prvků z výběru, které patří do kategorie A_i (bez ohledu na hodnotu proměnné B), a podobně $n_{\cdot j}$ počet prvků patřících do kategorie B_j . Platí následující vztahy:

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

Čísla $n_{i\cdot}$, resp. $n_{\cdot j}$ 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	\dots	B_s	Součty řádkové
A_1	n_{11}	n_{12}	\dots	n_{1s}	n_1
A_2	n_{21}	n_{22}	\dots	n_{2s}	n_2
\vdots	\vdots	\vdots	\ddots	\vdots	\vdots
A_r	n_{r1}	n_{r2}	\dots	n_{rs}	n_r
Součty sloupcové	n_1	n_2	\dots	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, začnáme 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 odchylinky 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, \dots, 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, \dots, W_r a každý prvek z W_i je klasifikován do jedné z kate-

gorií proměnné B . Nechť P_{ij} je relativní četnost prvků subpopulace W_i , jež jsou v j -té kategorii proměnné B . Potom se hypotéza homogenity se může vyjádřit následující rovnicí $P_{1j} = P_{2j} = \dots = P_{rj}$ pro všechna $j = 1, 2, \dots, s$, 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 spolu s udáním 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		Rádkové 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 podmíněně rozdelení 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 kategorialních proměnných A a B . Zkoumáme, zda hodnoty proměnné A neovlivňují rozdelení 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

$$p_{ij} = p_i \cdot p_j, \quad p_i = \sum_{j=1}^s p_{ij}, \quad p_j = \sum_{i=1}^r p_{ij},$$

které platí pro všechna i a j ($i = 1, 2, \dots, r$; $j = 1, 2, \dots, s$). 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, jež 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				Rádkové součty
	hry	atletika	gymn.	plování	
hry	133	6	2	4	145
atletika	15	10	4	3	32
gymn.	4	1	25	0	30
plování	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 pakují 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 lze najde v předešlým kapitolem o odhadu a testování hypotéz o relativní četnosti.

Tabulka typu $r \times s$

Při 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 $n_{i\cdot}/n$ a $n_{\cdot j}/n$. Protože očekávaná hodnota četnosti v políčku $m_{ij} = p_{ij}n$ a odhad p_{ij} je $n_{i\cdot}n_{\cdot j}/n^2$, tak

$$m_{ij} = \frac{n_i.n_j}{n}$$

pro $i = 1, 2, \dots, r$; a $j = 1, 2, \dots, s$. 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

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

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 mely být větší než jedna. Jestliže se v některých polohách vyskytnou nulové hodnoty, přejdeme k analýze odvozené tabulky vzniklé sloučením málo obsazených kategorií.

Jestliže zamítneme hypotézu nezávislosti nebo homogeneity, 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 ta bulky (opět typu $r \times s$). Příčinu nehomogeneity můžeme zjistit tak, že zopakujeme χ^2 -test pro tabulku, jež je zredukována o sloupce nebo řádky, které představují kandidáty nehomogeneity. 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 řadce 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).

Formuli pro vyjádření očekávaných četností $m_{ij} = n_i n_j / n$ lze po zlogaritmování zapsat ve formě $\ln(m_{ij}) = \ln(n_i n_j / n) = k + \ln(n_i) + \ln(n_j)$ kde $k = -\ln(n)$. To znamená, že za podmínky nezávislosti nebo homogenity pro každé políčko platí, že logaritmus očekávané hodnoty četnosti je lineární funkcí logaritmů řádkových a sloupcových marginálních součtů. To je analogický zápis k aditivnímu modelu analýzy rozptylu dvojného třídění, jejž poznáme v kapitole 9. Tato analogie vede k použití tzv. los-lineárních modelů pro analýzu četnostních tabulek (viz kap. 13.9).

Koeficienty závislosti pro kontingenční tabulku

Pro měření síly vztahu v kontingenční tabulce bylo navrženo několik koeficientů, které fungují podobně jako korelační koeficient. Interpretovat jejich číselné hodnoty je však dosti obtížné vzhledem ke všem možným kombinacím vztahů mezi kvalitativními údaji. Pro korelační tabulkou můžeme vypočítat příslušný Pearsonův korelační koeficient r obou původně spojitéch proměnných pomocí speciálního výpočetního schématu. Příkladem koeficientů vazby v kontingenční tabulce jsou **korigovaný koeficient kontingence podle Pearsona**

$$C_{\text{kor}} = \frac{C}{C_{\text{max}}},$$

kde

$$C = \sqrt{\frac{\chi^2}{\chi^2 + n}}, \quad C_{\max} = \sqrt{(m - 1)m};$$

a) Cramerův koeficient

$$V = \sqrt{\frac{\chi^2}{n(m-1)}},$$

příčemž χ^2 je popsaná testovací statistika a m je větší z obou čísel r a s . Platí, že C_{kor} i V jsou z intervalu $(0, 1)$; při nulové hodnotě není v tabulce žádný vztah; jestliže koeficienty mají hodnotu 1, je v ní vztah úplný.

Pokračování příkladu 8.9 (s. 311)

Vrátime se k tabulce 8.10, která popisuje na základě získaných četností vztah mezi oblíbeností sledování určitého sportu v televizi a oblíbeností aktivního provozování sportu. Již vizuální prozkoumání tabulky 8.11 odhalí patrnou závislost. Při platnosti hypotézy nezávislosti by si totiž měla být rádková procenta velmi podobná. Výpočtem testovací statistiky dostaneme hodnotu $\chi^2 = 273,3$, jež je větší než kritickámez $M_{krit} = 16,9$ pro hladinu významnosti 0,05 a počet stupňů volnosti $9 = (4 - 1) \times (4 - 1)$. Existuje tedy statisticky potvrzený vztah mezi oblíbeností sportu, který žáci provozují, a sportem, jenž rádi sledují v televizi. Korigovaný koeficient podle Pearsona C_{kor} má pro tuto tabulku hodnotu 0,82, což indikuje silný vztah. Cramerovo V má hodnotu 0,62.

Tab. 8.11 Tabulka řádkových relativních četností pro data z tabulky 8.10

Proměnná A oblíbenost při sledování televize (%)	Proměnná B oblíbenost při sportování [%]				Řádkové součty [%]
	hry	atletika	gymn.	plavání	
hry	91,7	4,1	1,4	2,8	100
atletika	46,9	31,3	12,5	9,3	100
gymn.	13,3	3,3	83,4	0	100
plavání	33,3	0	3,7	63	100
Sloupcové součty	68,8	7,2	13,6	10,4	100

Tabulka 2 × 2

Uvažujeme dvě náhodné proměnné X a Y , které nabývají jenom dvě hodnoty: 0 a 1. Po klasifikaci n prvků výběru získáme četnostní tabulku typu 2×2 s četnostmi a, b, c a d (tab. 8.12). Pro výpočet statistiky χ^2 můžeme použít zjednodušený vzorec

$$\chi^2 = \frac{n(ad - bc)^2}{(a + b)(c + d)(a + c)(b + d)}.$$

Kritické hodnoty jsou uvedeny v tabulce χ^2 -rozdělení o jednom stupni volnosti. Zvláštností tabulky typu 2×2 je, že v ní lze uvažovat směr poruchy nulové hypotézy, a proto se musíme rozhodnout, zda použijeme jednostranný, nebo dvoustranný test. Příslušný výtah z tabulky kritických hodnot χ^2 uvádíme v tabulce 8.13. Přitom předpokládáme, že pravděpodobnost sledovaného jevu v první, resp. druhé subpopulaci je p_1 , resp. p_2 .

Tabulka musí být dostatečně obsazena, aby χ^2 test platil. Za kritérium použijeme dva vztahy $a + b \approx c + d > 5$ nebo $a + b > 5, c + d > (a + c)/3$. Tabulkou s menším obsazením políček testujeme Fisherovým testem.

Tab. 8.12 Četnostní tabulka pro dvě dvouhodnotové náhodné proměnné

Proměnná Y	Proměnná X		Řádkové součty
	0	1	
0	a	b	a + b
1	c	d	c + d
Sloupcové součty	a + c	b + d	n

Tab. 8.13 Kritické hodnoty pro test nezávislosti v čtyřpolní tabulce

Hladina významnosti	0,05	0,01	0,001
Dvoustranný test ($H_0: p_1 = p_2; H_1: p_1 \neq p_2$)	3,84	6,63	10,82
Jednostranný test ($H_0: p_1 = p_2; H_1: p_1 > p_2$)	2,7	5,41	9,55

„Fisherův test“ nezávislosti v kontingenční tabulce patří k přesným testům nezávislosti náhodných proměnných a používá se při malých rozsazích výběru. Předpokládejme, že společné rozdělení obou proměnných je pro výběr popsáno čtyřpolní tabulkou 8.12. Zjišťujeme, jak pravděpodobná je konfigurace četností získaná nebo ještě extrémnější za platnosti nulové hypotézy. Při potřebných výpočtech pravděpodobností jednotlivých konfigurací četností v tabulce fixujeme sloupcové a řádkové součty. Tento postup je oprávněný, protože sloupcové a řádkové součty neobsahují žádnou informaci o tom, zda je splněna podmínka nezávislosti. V tabulce pak ze všech hodnot (a, b, c, d) může mít pouze jedna z nich určitou volnost. Ostatní jsou jí automaticky určeny. Budeme za ni považovat hodnotu a a k ní příslušnou náhodnou proměnnou označíme Z . Jestliže platí hypotéza nezávislosti (nebo homogeneity), pak podmíněná pravděpodobnost této hodnoty je dána výrazem

$$P^*(Z = a) = \frac{(a + b)!(c + d)!(a + c)!(b + d)!}{n! a! b! c! d!}.$$

Pomocí pravděpodobnosti P^* můžeme určit kritickou oblast pro hodnotu náhodné proměnné Z tak, že spočítáme kumulativní pravděpodobnosti extrémních hodnot proměnné Z . Ty by neměly být nižší než zvolená hladina významnosti α , pokud platí hypotéza nezávislosti.

PŘÍKLAD 8.11**Testování hypotézy homogeneity v kontingenční tabulce**

Pro ilustraci postupu budeme hodnotit hypotézu homogeneity v tabulce 8.14. Spočítáme pravděpodobnosti jednotlivých hodnot pro náhodnou proměnnou Z (tabulka 8.15). Jestliže chceme provést jednostranný test hypotézy, že pravděpodobnost výskytu hodnoty $Y = 0$ je menší ve skupině s hodnotou $X = 0$ než ve skupině s hodnotou $X = 1$, musíme sečítat pravděpodobnosti pro dvě extrémní hodnoty $Z = 0$ a $Z = 1$. Dostaneme pravděpodobnost $p = 0,0029 + 0,0355 = 0,0384$. Tuto hodnotu porovnáme se zvolenou hladinou významnosti, např. 0,05. Protože $p < 0,05$, výsledek indikuje, že konfigurace dat popsaná uvedenou tabulkou četností je dosti výjimečná za platnosti nulové hypotézy.

Tab. 8.14 Příklad četnostní tabulky pro dvě dvouhodnotové proměnné

		Proměnná X	
		0	1
Proměnná Y	0	1	6
	1	13	8

Tab. 8.15 Tabulka pravděpodobnosti různých hodnot testovací statistiky Z

Z	0	1	2	3	4	5	6	7
P(Z)	0,0029	0,0355	0,1539	0,3077	0,3077	0,1539	0,0355	0,0029

Z provedené analýzy plyne, že náhodná proměnná Z má za platnosti hypotézy nezávislosti asymptoticky normální rozdělení s průměrem $E(Z) = (a+c)(a+d)/n$ a rozptylem $Var(Z) = (a+b)(a+d)(c+d)(b+d)/[(n-1)n^2]$. Pomocí tohoto poznatku lze provést asymptoticky platný test nezávislosti v čtyřpolní tabulce také pomocí z-statistiky

$$z = \frac{Z - E(Z)}{\sqrt{Var(Z)}}.$$

Koefficienty závislosti pro tabulku 2×2

Pro měření síly vztahu dvou dichotomických proměnných v tabulce 2×2 bylo navrženo mnoho koefficientů. Tyto koefficienty nabývají hodnot 0, když obě proměnné jsou na sobě nezávislé, a záporné nebo kladné hodnoty, jestliže proměnné jsou záporně nebo kladně asociované. Podle definice dvě dichotomické proměnné A a B jsou spolu kladně asociované, když platí jedna ze dvou ekvivalentních podmínek:

1. $p(A = 1, \text{ za podmínky } B = 1) > p(A = 1, \text{ za podmínky } B = 2)$
2. $p_{11} > p_1 p_1$, kde $p_{11} = p_{11} + p_{12}$ a $p_1 = p_{11} + p_{21}$

První podmínka vyjadřuje, že relativní četnost hodnoty A = 1 je jiná v subpojácích definovaných hodnotami B = 1 a B = 2. Druhá podmínka zdůrazňuje, že jevy A = 1 a B = 1 nejsou nezávislé, a popisuje směr odchylky od nezávislosti. Jestliže nerovnosti jsou obrácené, mluvíme o záporné asociaci. Bylo ukázáno,

vhodná míra asociace by měla být funkci tak zvaného poměru (podílu) šancí OR (odds ratio – viz též kap. 4.1.3):

$$OR = p_{11}p_{22}/p_{12}p_{21} = (p_1/(1-p_1))/(p_2/(1-p_2))$$

Jestliže vezmeme jeho logaritmus, dostaneme koefficient, jenž nabývá hodnoty 0, když proměnné jsou nezávislé, a hodnoty záporné, resp. kladné, když proměnné jsou spolu kladně, resp. záporně asociované. Test nezávislosti v tabulce 2×2 je testem hypotézy, že koefficient OR je roven jedničce.

Koefficient závislosti, který vychází z OR, je Yuleovo Q

$$Q = \frac{p_{11}p_{22} - p_{12}p_{21}}{p_{11}p_{22} + p_{12}p_{21}} = \frac{OR - 1}{OR + 1}.$$

Je-li OR = 1, což odpovídá nezávislosti obou náhodných proměnných, pak Q = 0. Jestliže OR roste, resp. klesá, Q se blíží k hodnotě 1, resp. k hodnotě -1. Hodnotu koefficientu Q odhadujeme dosazením četností z příslušných políček tabulky:

$$Q = \frac{ac - bd}{ac + bd}$$

PŘÍKLAD 8.12

Testování nezávislosti v kontingenční tabulce

testujeme hypotézu nezávislosti v tabulce 2×2 s údaji o zájmech studentů o sport (příklad 8.8):

$$\chi^2 = \frac{140(30 \times 63 - 11 \times 36)^2}{41 \times 99 \times 66 \times 74} = 15,8$$

Přestože kritická hodnota χ^2 rozdělení na hladině 0,05 s jedním stupněm volnosti je 3,84, můžeme zamítout hypotézu nezávislosti v této tabulce. Uzavíráme, že dívky a chlapci se liší v intenzitě zájmu o účast ve sportovním oddíle. Zájem chlapců je vyšší než zájem dívek. Cílem je také zhodnotit věcnou významnost tohoto výsledku. Odhad části zájemu dívek $30/66 = 0,45$ u chlapců a u dívek $11/74 = 0,14$. Tato čísla říkají, že chlapci mají zhruba dva až třikrát tak velký zájem jako dívky, což je jistě věcně významný rozdíl. Hodnotu koefficientu závislosti Q odhadujeme číslem

$$Q = \frac{30 \times 63 - 11 \times 36}{30 \times 63 + 11 \times 36} = 0,65.$$

8.3.2 Analýza párových dichotomických proměnných

Často na zkoumaných osobách (prvcích souboru) sledujeme dichotomickou proměnnou (hodnoty + nebo -) dvakrát, před pokusem a po něm; máme zjistit, zda došlo ke statisticky významné změně v rozdelení této proměnné. McNemar navrhl pro tento případ test, který je speciálním případem znaménkového testu pro dvě závislé skupiny.

Poměr četností v obou kategorických dichotomických proměnných se bude mezi oběma měřenými více nebo méně měnit. Velikost této změny posuzujeme McNemarovým testem tak, že uvažujeme, kolik osob při prvním měření přejde při druhém měření do jiné kategorie uvažované proměnné. Vztah výsledků obou měření zobrazujeme četnostní tabulkou typu 2×2 , v každém řádku klasifikují výsledky v prvním a druhém sloupci výsledky z druhého měření (tab. 8.16). Například číslo a je četnost osob, jež jak v prvním, tak ve druhém měření měly hodnotu sledované proměnné +. Zajímá nás, zda čísla b, c se od sebe liší pouze v rámci náhodného kolísání. Jenom tyto dvě četnosti rozhodují o tom, zda je charakteristika ve druhém měření jinak rozdělená než při prvním měření. Tuto hypotézu testujeme statistikou

$$\chi^2 = \frac{(b - c)^2}{b + c},$$

Tab. 8.16 Četnostní tabulka pro analýzu párových dichotomických proměnných

		II. měření		Součet
		+	-	
I. měření	+	a	b	
	-	c	d	

Tab. 8.17 Tabulka teoretických pravděpodobností pro analýzu párových dichotomických proměnných

		II. měření		Součet
		+	-	
I. měření	+	p_{11}	p_{12}	p_1
	-	p_{21}	p_{22}	p_2
Součet		p_1	p_2	

kterou srovnáváme s kritickou hodnotou χ^2 rozdělení o jednom stupni volnosti (vhodné pro počty údajů $b + c > 8$). Jestliže nahradíme četnosti teoretickými pravděpodobnostmi p_{ij} , abychom popsali společné rozdělení výskytu jednotlivých výsledků, dostaneme tabulku 8.17. Tabulku jsme také doplnili pravděpodobnostmi p_i a p_j , které určují marginální rozdělení dichotomické proměnné ve druhém a prvním měření. McNemarův test testuje hypotézu $H_0: p_1 = p_2$.

PŘÍKLAD 8.13

Analýza párových dichotomických proměnných McNemarovým testem

Prezkušuje se, zda výuka o pozitivním působení sportu na zdraví vede ke změně postojů žáků ke sportování. Hypotézy:

H_0 : Počet žáků, kteří změní svůj postoj pozitivním směrem, je pouze náhodně odlišný od počtu žáků, kteří změní svůj postoj negativním směrem.

H_1 : Po výuce je počet žáků s pozitivní změnou větší než počet žáků se změnou v negativním směru (jednostranná hypotéza).

Tab. 8.18 Příklad dat pro analýzu párových dichotomických proměnných – změna postojů žáků ke sportu

		Postoj po výuce		Součet
		+	-	
Postoj před výukou	+	5	3	8
	-	16	2	18
Součet		21	5	26

Zkoumání přineslo výsledky uvedené v tabulce 8.18. Za platnosti nulové hypotézy se změnění názory v obou směrech (pozitivním a negativním) u přibližně stejného počtu žáků (až na náhodné kolísání). Testovací statistika má hodnotu

$$\chi^2 = \frac{(3 - 16)^2}{3 + 16} = \frac{169}{19} = 8,89.$$

Protože $8,89 > 3,84$ (jednostranná kritická mez χ^2 o jednom stupni volnosti pro hladinu významnosti 0,05), lze uzavřít, že zvolená výuka má pozitivní vliv na postoj žáků vzhledem k provozování sportu s cílem podpory zdraví.

Interval spolehlivosti pro rozdíl marginálních pravděpodobností $\Delta = p_{.1} - p_{1.}$ má tvar

$$((b-c)/n - z_{\alpha/2}SE; (b-c)/n + z_{\alpha/2}SE),$$

kde $n = a + b + c + d$ a SE je směrodatná chyba poměru $(b-c)/n$, daná vztahem

$$SE = \frac{1}{n} \sqrt{b+c - \frac{(b-c)^2}{n}}.$$

Je patrné, že hypotéza $H_0: p_{.1} = p_{1.}$ je ekvivalentní hypotéze $H_0: p_{12} = p_{21}.$ Z toho plyne, že pokud za referenční počet pozorování vezmeme číslo $b + c$, pak McNemarův test je ekvivalentní testu, že četnost a má binomické rozdělení $B(b+c; 0,5).$

PŘÍKLAD 8.14

Analýza párových dichotomických proměnných McNemarovým testem

Členové horolezeckého klubu diskutují o obtížnosti jednotlivých stěn. Zvláště je zajímají stěny A a B. Aby je porovnali, provedou analýzu 108 údajů klubového deníku o tom, zda jednotliví členové při prvním pokusu stěnu zdolali, nebo ne (tabulka 8.19).

Úvaha vede k tomu, že k pokusům, které u obou stěn vedly k úspěchu nebo naopak k neúspěchu, není potřeba při posuzování obtížnosti stěny přihlížet. Jedinou informaci přináší údaje o těch členech klubu, kteří jednu stěnu zdolali a druhou nikoliv. Z tabulky je vidět, že 9 členů zdolalo stěnu A, ale ne stěnu B. Číslo 9 považujeme za počet kladných znamének pro stěnu A. Naopak 14 členů zdolalo stěnu B, ale ne stěnu A. Toto číslo můžeme zase považovat za počet záporných znamének pro stěnu A. Vypočítáme testovací statistiku podle McNemara:

$$\chi^2 = \frac{(9 - 14)^2}{9 + 14} = \frac{25}{23} = 1,08$$

Z hodnoty testovací statistiky plyne, že není dostatek evidence, jež by svědčila pro větší obtížnost jedné ze stěn. V dalším odstavci řešenou úlohu zobecníme.

Tab. 8.19 Příklad dat pro test podle McNemara – porovnání obtížnosti dvou stěn

Stěna B	Stěna A	
	úspěch	neúspěch
úspěch	73	14
neúspěch	9	12

8.3.3 Cochranův test a test podle Bowkera

McNemarovým testem, který jsme popsali v předchozím odstavci, se prověřuje homogenita rozdělení alternativních dat dvou závislých výběrů. Hypotéza homogenity ve více závislých výběrech alternativních dat se prověřuje Q -testem podle Cochranova. Testuje se hypotéza H_0 , že všechny výběry pocházejí ze stejné základní populace. Jeho provedení ukážeme na situaci, jež je podobná předchozímu příkladu o horolezcích. Nyní se na rozdíl od příkladu 8.14 srovnává obtížnost tří výstupů. Data mohou být popsána tabulkou 8.20. Jednička znamená úspěšný pokus o zdolání stěny. Hodnoty T_i , B_i a N , které potřebujeme pro výpočet Cochranovy testovací statistiky, jsou počty jedniček v řádcích, sloupcích a v celé tabulce.

Cochran navrhl metodu pro testování hypotézy typu $H_0: Všechny stěny mají stejnou obtížnost$ proti alternativní hypotéze $H_1: Aspoň jedna stěna má jinou obtížnost než ostatní.$ Jednotlivé stěny považujeme za „ošetření“ a horolezci jsou bloky. Jestliže máme t ošetření provedených v b blocích a alternativní odpověď na ošetření (např. 0, 1), vhodnou statistikou pro test je

$$Q = \frac{t(t-1) \sum_i T_i^2 - (t-1)N^2}{tN - \sum_j B_j^2},$$

kde T_i označuje součet „jedniček“ pro ošetření i , B_j je součet „jedniček“ v bloku j a N je počet všech jedniček. Statistika Q má za platnosti nulové hypotézy asymptoticky χ^2 rozdělení o $t-1$ stupních volnosti. Pro dvě ošetření je navržená statistika přesně rovna testovací statistice podle McNemara.

Jiným zobecněním McNemarova testu je test symetrie v tabulce typu $N \times N$, který navrhl Bowker. Jeho test se může např. použít při hodnocení vedlejších účinků nového léku. Jestliže máme pacienty, kteří jsou ošetřeni starým a poté novým lékem, můžeme zaznamenat jejich vedlejší účinky více než dvěma kategoriemi. Tabulka 8.21 ukazuje možné výsledky takového experimentu.

Tab. 8.20 Příklad dat pro ověření homogenity Q -testem podle Cochranova

	Horolezci					T_i
	Adam	Bára	Cyril	Dana	Emil	
Stěna 1	1	1	0	0	1	3
Stěna 2	1	0	0	1	0	2
Stěna 3	0	1	1	1	1	4
B_i	2	2	1	2	2	$N = 9$

Tab. 8.21 Příklad dat pro test symetrie podle Bowkera – vedlejší účinky léků

Vedlejší účinek původního léku	Vedlejší účinek nového léku			Celkem
	žádný	lehký	vážný	
žádný	83	4	3	90
lehký	17	22	5	44
vážný	4	9	11	24
Celkem	104	35	19	158

Bowker navrhl test, jenž testuje, zda se alespoň jeden páru pravděpodobnosti symetricky položených políček v tabulce $N \times N$ nacházejících se mimo diagonálu od sebe liší. Jedná se o zobecnění McNemarova testu a testovací statistika má tvar

$$\chi^2 = \sum (n_{ij} - n_{ji})^2 / (n_{ij} + n_{ji}),$$

kde se sčítá přes všechna i od 1 do $n - 1$ a $j > i$. Za platnosti nulové hypotézy symetrie má tato statistika asymptoticky χ^2 -rozdělení s $0,5(n-1)n$ stupni volnosti. Pro náš příklad dosazením dostaneme

$$\chi^2 = (4 - 17)^2 / 21 + (3 - 4)^2 / 7 + (5 - 9)^2 / 14 = 9,33$$

se 3 stupni volnosti. Lze zjistit, že $p = P(\chi^2 > 9,33) = 0,026$. Z toho plyne, že existuje evidence pro rozdílnou incidenci vedlejších účinků u obou léků.

8.3.4 Kappa koeficient shody

Četnostní tabulky vznikají různým způsobem a také je lze hodnotit z rozličných pohledů. Předpokládejme, že četnostní tabulka zachycuje četnosti shod a rozdílností dvou posuzovatelů, kteří klasifikují n objektů do K kategorií při posuzování specifikované vlastnosti objektů. Protože oba posuzovateli zařazují objekty do stejněho počtu kategorií, má vzniklá tabulka četností stejný počet řádků a sloupců – je typu $K \times K$. Posuzovateli jsou pro posuzování vyškoleni, ale přesto mezi nimi může docházet k rozdílům. První pozorovatel zařadí objekt do kategorie i a druhý pozorovatel do kategorie j . Zřejmě nemusí vždy platit, že $i = j$. To znamená, že součet četností na diagonále tabulky, tj. shod, se nemusí rovnat počtu posuzovaných objektů.

Podobně jako Pearsonův korelační koeficient r při porovnávání dvou kvantitativních metod měření stejné veličiny posuzuje pouze sílu vztahu výsledků

získaných oběma metodami a ne jejich shodu, tak také Pearsonův koeficient C_{kor} a Cramerův koeficient V měří v popsané tabulce pouze sílu vztahu a ne shodu posuzovatelů. Tu zachycujeme pomocí kappa koeficientu.

Četnosti n_{ij} ($i = 1, 2, \dots, K$) označují počet objektů zařazených posuzovatelem A do i -té kategorie a posuzovatelem B do j -té kategorie. Jestliže sečteme četnosti n_{ii} na diagonále, dostaneme počet shod obou pozorovatelů. Vydelením tohoto součtu celkovým počtem objektů získáme odhad pravděpodobnosti p_0 , že se při klasifikaci objektu oba posuzovateli shodou. Součet všech ostatních četností v tabulce vydelený celkovým počtem objektů je odhadem pravděpodobnosti p_e , že posuzovatelé objekt klasifikují rozdílně. Koeficient kappa se vypočte jednoduše:

$$\kappa = \frac{p_0 - p_e}{1 - p_e}$$

Koeficient κ nabývá hodnoty jedna při úplném souhlasu obou pozorovatelů a hodnoty 0, jestliže počet shod odpovídá náhodné shodě obou pozorovatelů. Negativní hodnoty se objevují, jestliže shoda je slabší než shoda očekávaná při zcela náhodné shodě obou posuzovatelů. To samozřejmě nastává zřídka. Hypotéza nulové hodnoty koeficientu κ se zkoumá málokdy.

Kappa koeficient většinou slouží pro popisné účely shody posuzovatelů. Je navržen pro nominální klasifikace. Pro větší hodnoty dimenze K je význam koeficientu κ omezený. Jestliže kategorie jsou ordinální, pak závažnost neshody posuzovatelů závisí na velikosti diference kategorií, do nichž posuzovateli objekt zařadili. V takovém případě použijeme vážený κ koeficient. Při jeho výpočtu se velikost souhlasu klasifikace mezi kategoriemi i a j oceňuje vahou w_{ij} , která vyjadřuje míru souhlasu kategorií. Vážený koeficient κ s vahami w_{ij} se vypočte podle vzorce:

$$\kappa = \frac{\sum_{i=1}^K \sum_{j=1}^K w_{ij} p_{ij} - \sum_{i=1}^K \sum_{j=1}^K w_{ij} p_i p_j}{1 - \sum_{i=1}^K \sum_{j=1}^K w_{ij} p_i p_j}$$

Pro volbu vah $\{w_{ij} = 1 - (i - j)^2 / (K - 1)^2\}$ je váha shody větší, pokud posuzovaná četnost leží blíže k diagonále tabulky. Tento upravený koeficient se používá ve studiích o spolehlivosti měření nebo shody pozorovatelů, když se využívají např. jednotlivé položky na Likertově škále.

PŘÍKLAD 8.15

Posuzování shody dvou pozorovatelů kappa koeficientem shody

Dva lékaři se vyjádřili nezávisle k problému, zda je vhodné vybraným pacientům doporučit operaci, nebo ne. Po zpracování případů několika desítek pacientů výzkumník vypočítal koeficient κ jako míru souhlasu rozhodnutí obou lékařů. Vycházel z odhadu pravděpodobnosti p_{ij} jednotlivých kombinací rozhodnutí obou lékařů (tab. 8.22). Tabulka obsahuje i pomocné výpočty pro zjištění hodnoty p_e . Z tabulky plyne, že $p_e = 0,485$ a $p_0 = 0,70$. Koeficient κ má tedy hodnotu $(0,7 - 0,485)/(1 - 0,485) = 0,417$. Tato hodnota naznačuje relativně malý souhlas obou lékařů při rozhodování o terapii.

Tab. 8.22 Příklad dat pro výpočet koeficientu shody dvou posuzovatelů

Lékař B	Lékař A		Součet
	operace potřebná	operace nepotřebná	
operace potřebná	0,40	0,05	0,45
operace nepotřebná	0,25	0,30	0,55
Součet	0,65	0,35	1,00
p_i, p_j	$0,65 \times 0,45 = 0,2925$	$0,35 \times 0,55 = 0,1925$	$0,2925 + 0,1925 = 0,485$

8.4 Ordinální kategoriální data

Test nezávislosti v kontingenční tabulce lze použít i pro kategoriální ordinální data. Jestliže však uvažujeme o trendu v datech v důsledku ordinality obou proměnných, je vhodnější aplikovat metodu, která reaguje silněji na takto vzniklé porušení hypotézy nezávislosti. V kapitole o testech středních hodnot (kap. 6.4) jsme popsali Wilcoxonův, resp. Mannův-Whitneyův test a jeho aplikaci na data kategoriálního ordinálního typu při srovnání dvou skupin. Tento test je vhodný, pokud srovnáváme pouze dvě skupiny nebo jedna z proměnných je binárního typu. Oblíbeným testem zjištění trendu v datech je test Cochranova a Armitageho, kterým se testuje rovnost pravděpodobností p_i určité vlastnosti v k populacích (viz Zvára, 2000, s. 190) proti alternativní hypotéze $p_1 \leq p_2 \leq \dots \leq p_k$, přičemž aspoň jedna nerovnost je ostrá.

V zásadě musíme odlišovat případy, kdy mají obě proměnné ordinální charakter, od případů, kdy jej má pouze jedna z nich. Jestliže nastal druhý případ,

Tab. 8.23 Příklad čtyřpolní tabulky se dvěma ordinálními proměnnými

Proměnná Y	Proměnná X	
	0	1
0	4	2
1	2	3

Tab. 8.24 Úprava dat z tabulky 8.23

X	0	0	0	0	0	0	1	1	1	1	1
Y	0	0	0	0	1	1	0	0	1	1	1

pak lze použít při zkoumání hypotézy nezávislosti nebo homogenity Kruskalův-Wallisův test s opravou na stejné hodnoty (ties), jenž je pro dvě skupiny totožný s testem Manna a Whitneye.

Pokud jsou obě proměnné ordinálního typu, vyřešíme úlohu aplikací nějakého neparametrického koeficientu korelace. Je tomu tak proto, že data v tabulce lze přepsat jako dvě řady indexů, pro které vypočítáme vhodný korelační koeficient. Například čtyřpolní tabulku 8.23 lze rozepsat jako 11 dvojic údajů (x_i, y_i) tak, jak to ukazuje tabulka 8.24. Nejčastěji se používá při zkoumání závislosti kategoriálních ordinálních proměnných X a Y Kendallův korelační koeficient upravený na shodné hodnoty nebo některé jeho varianty. Při analýze trendu v ordinálních datech lze použít test podle Jonckheere-Terpstra (viz kap. 9.1.5). Jeho uplatnění je ekvivalentní testu významnosti Kendallova korelačního koeficientu.

Jednou z variant Kendallova koeficientu t_k je Goodmanův-Kruskalův koeficient γ , který také vychází z počtu konkordancí P a diskordancí Q . Obě tyto veličiny jsme popsali v kapitole 7.2.7. Tento koeficient se hodí pro zachycení asociace v kontingenční tabulce, jež vznikla tříděním objektů podle ordinálních proměnných. Koeficient γ se spočte podle vzorce

$$\gamma = \frac{P - Q}{P + Q}.$$

Uvedeme postup výpočtu hodnot P a Q na příkladu. Máme tabulku 8.25, která vznikla tříděním objektů podle ordinálních kritérií X a Y . Každé z kritérií má tři kategorie. Počet konkordancí a diskordancí spočteme podle tabulky 8.26.

Tab. 8.25 Data o objektech třídených podle dvou ordinálních proměnných

	x_1	x_2	x_3
y_1	a	d	c
y_2	d	e	f
y_3	g	h	i

Tab. 8.26 Výpočet konkordancí a diskordancí pro data z tabulky 8.25

Typ	Počet párů	Označení
konkordance	$a(e + f + h + i) + b(f + i) + d(h + i) + e(i)$	P
diskonkordance	$c(d + e + g + h) + b(d + g) + f(g + h) + e(g)$	Q

Testování hypotézy nezávislosti se opírá o výraz

$$z = \frac{P - Q}{\sqrt{\text{Var}(P - Q)}},$$

kde rozptyl $\text{Var}(P - Q)$ se spočte pomocí formule

$$\text{Var}(P - Q) = \frac{(1 - \sum p_i^3)(1 - \sum p_j^3)n^3}{9}.$$

PŘÍKLAD 8.16

Posuzování nezávislosti kategorialních ordinálních proměnných na základě Goodmanova-Kruskaľova koeficientu

Metody pro testování hypotézy nezávislosti dvou ordinálních proměnných demonstrujeme na příkladu tabulky, jež vznikla při sledování závislosti aktivity ve sportu na úrovni výdělků u mužů. Jak proměnná popisující intenzitu sportovní aktivity, tak příjemová skupina mají charakter kategorialních ordinálních proměnných. Získané údaje jsou zobrazeny v tabulkách 8.27 a 8.28 spolu s dopočítanými hodnotami pro výpočet testovací statistiky. Vypočítáme koeficient závislosti γ :

$$\gamma = \frac{291\,800 - 200\,100}{291\,800 + 200\,100} = 0,186$$

Vypočítáme také rozptyl hodnoty $P - Q$ a testovací statistiku z :

$$z = \frac{P - Q}{\sqrt{\text{Var}(P - Q)}} = \frac{291\,800 - 200\,100}{17\,015,34} = 5,38$$

Tab. 8.27 Příklad kontingenční tabulky pro analýzu dat popsaných dvěma kategorialními ordinálními proměnnými

Sportovní aktivita	Příjemová skupina			Součet	$p_j = n_j/n$
	spodní	střední	vyšší		
často sportuje	170	160	90	420	0,28
občas sportuje	290	220	120	630	0,42
nesportuje	140	120	190	450	0,30
Součet	600	500	400	1500	
$p_j = n_j/n$	0,40	0,33	0,27		

Tab. 8.28 Výpočet konkordancí a diskordancí pro data z tabulky 8.27

Typ	Počet párů	Označení
konkordance	$170(220 + 120 + 120 + 190) + 160(120 + 190) + 290(120 + 190) + 220(190)$	$P = 291\,800$
diskonkordance	$90(290 + 220 + 140 + 120) + 160(290 + 140) + 120(140 + 120) + 220(140)$	$Q = 200\,100$

Asymptoticky platným testem jsme prokázali na hladině významnosti 0,05, že koeficient γ se liší od nuly, protože hodnota 5,38 je větší než 1,95 (kritická mez).

Kendallův koeficient τ_{au-c} je další modifikací Kendallova koeficientu korelace t . Vypočítá se podle vzorce

$$t_c = \frac{2m(P - Q)}{n^2(m - 1)},$$

kde m je menší z obou dimenzí r, s kontingenční tabulky, pro niž se t_c počítá. Hodí se pro výpočet korelace v tabulce s libovolnými hodnotami r a s . Z rozdílu $P - Q$ vycházejí také dva výpočty korelačních koeficientů podle Somerse, které se používají pro hodnocení schopnosti řádkové, resp. sloupcové proměnné predikovat hodnoty sloupcové, resp. řádkové proměnné. Oba koeficienty mají symetrický charakter.

Pro analýzu kontingenční tabulky s oběma proměnnými ordinálního typu také používá test nezávislosti, který vychází ze skóru u_i a v_j , přiřazených jednotlivým kategoriím obou ordinálních proměnných před provedením testu. Při analýze tabulek je obvykle musíme sami navrhnut. Přitom přihlížíme k podstatě soumaného problému. Obě řady skóru musí tvorit vzestupnou posloupnost.

V uvedené čtyřpolní tabulce jsou skóry 0 a 1. Testovací statistiku lze vytvořit pomocí hodnoty S

$$S = \sum_{i,j} u_i v_j n_{ij}.$$

Za platnosti hypotézy nezávislosti obou proměnných má průměr a rozptyl statistiky S hodnoty:

$$E(S) = \frac{\sum_i u_i n_i \cdot \sum_j v_j n_{j.}}{n}$$

$$Var(S) = \frac{\left[\sum_i u_i^2 n_i - \left(\sum_i u_i n_i \right)^2 \right] \left[\sum_j v_j^2 n_{.j} - \left(\sum_j v_j n_{.j} \right)^2 \right]}{n^2(n-1)}$$

Test lze provést na základě poznatku, že testovací z -statistika

$$z = \frac{S - E(S)}{\sqrt{Var(S)}}$$

má za platnosti nulové hypotézy asymptoticky standardizované normální rozdělení. Tento test představuje zobecnění testu Cochranova a Armitageho.

PŘÍKLAD 8.17

Posuzování nezávislosti kategoriálních ordinálních proměnných

Následující příklad popsal Sprent (2001, 378). Zkoumáme vedlejší účinky nového léku při různých dávkách. Získali jsme tabulku 8.29. Je patrné, že obě proměnné můžeme považovat za kategoriální a ordinální (rozdíl mezi dávkami chápeme kvalitativně). Chceme otestovat

Tab. 8.29 Příklad dat, u nichž testujeme hypotézu nezávislosti, resp. existenci trendu

Dávka	Vedlejší efekt				Součet
	žádný (1)	slabý (2)	střední (3)	velký (4)	
100mg (1)	50	0	1	0	51
200mg (2)	60	1	0	0	61
300mg (3)	40	1	1	0	42
400mg (4)	30	1	1	2	34
Součet	180	3	3	2	188

hypotézu nezávislosti a ověřit existenci trendu. Skóry jsme zvolili $u_i = i$, $v_j = j$. Jednotlivé marginální četnosti mají hodnoty, které jsou také uvedeny v tabulce. Výpočty dostaneme $S = 484$, $E(S) = 469,7$ a $Var(S) = 35,797$, tudíž testovací statistika $z = 2,39$. Tato hodnota indikuje přítomnost trendu v datech, jestliže test provádíme na 5% hladině významnosti. Poznamenejme, že výsledky testu ovlivňuje volba skóru.

8.5 Problém třetí proměnné a Simpsonův paradox

Doporučuje se zkoumat efekty různých proměnných na vztah v kontingenční tabulce, abychom mu lépe porozuměli. Řešíme takzvaný problém třetí proměnné. Působení třetí proměnné může ovlivnit naši interpretaci vztahu v kontingenční tabulce. Obecně jsme se touto otázkou zabývali v úvodní kapitole (s. 41) a specificky již v souvislosti s korelační analýzou (kap. 7.2.4). Pokud jsme všechny proměnné měřili, postupujeme těmito kroky: 1. Popíšeme vztah mezi dvěma proměnnými kontingenční tabulkou. 2. Rozdělíme data podle hodnot třetí proměnné a vytvoříme tak podskupiny. 3. Sestrojíme pro každou podskupinu kontingenční tabulku vztahu původních dvou proměnných. 4. Porovnáváme vztah nalezený v jednotlivých podskupinách se vztahem v původní tabulce.

Třetí proměnnou nazýváme někdy kontrolní proměnnou. Postupem odhalujeme rušivé působení kontrolní proměnné. Vztah v původní tabulce nazýváme někdy **vztahem nultého rádu**. Vztah, který získáme pro určitou hodnotu kontrolní proměnné, nazýváme **parciálním vztahem**. Možnosti, jež mohou nastat, mají čtyři podoby.

- Parciální vztah je stejný jako vztah nultého rádu. Třetí proměnná nemá na vztah žádný efekt.
- Parciální vztah je redukován na nulu. V tomto případě záleží na tom, jaký je logický vztah kontrolní proměnné k oběma proměnným. Jestliže ji pojíme jako společnou příčinu předcházející oběma proměnným, které příčinně ovlivňuje, pak je vztah nultého rádu zdánlivý.
- Různý efekt na jednotlivé parciální vztahy. Například některé parciální vztahy jsou nulové, jiné zůstávají beze změny. To poukazuje na moderující působení kontrolní proměnné. Tento jev se nazývá interakce.
- Parciální vztahy jsou podstatně silnější než vztah nultého rádu. Vztah nultého rádu obvykle zkoumáme pouze tehdy, když je významně veliký. Někdy je však slabý vztah zesílen při zavedení třetí proměnné. Třetí proměnná se

nazývá v tomto případě supresorní proměnná. Někdy se dokonce vztah zcela obrátí. Pak tento jev nazýváme Simpsonův paradox. Popíšeme ho podrobněji v dalším odstavci.

Interpretace dat, již jsme popsali, je komplexnější než zkoumání bez uvážení třetí proměnné. Tento přístup můžeme dále rozšířit na kombinace více kontrolních proměnných. Je pravděpodobné, že v konkrétním výzkumu se setkáváme s konfiguracemi dat, které nejsou tak jasně rozlišeny. V praxi je však třeba počítat s tím, že vícenásobné příčinné působení je spíše pravidlem než výjimkou. Proto je zapotřebí uvažovat složitější modely dat. V kapitole 13.9 popíšeme jeden z takových modelů, jenž se nazývá log-lineární model pro četnostní data. Používáme ho při analýze vícerozměrných kontingenčních tabulek.

Simpsonův paradox

Jako Simpsonův paradox se označuje obrácení závislosti nebo směru působení v kontingenční tabulce působením třetí proměnné při srovnání se vztahem nultého řádu. Je to jev natolik překvapivý, že se mu věnujeme podrobněji. Může se objevit, kdykoli slučujeme data. Jestliže se hodnoty v tabulce sloučí např. tím, že se zruší určitá klasifikační dimenze (členění podle poblaví, kvalifikace apod.), nová tabulka nemusí reprezentovat skutečné vztahy mezi proměnnými. Tento jev je pojmenován podle Edwarda Simpsona (Simpson, 1951), který ho poprvé popsal.

Podstatu Simpsonova paradoxu demonstруjeme pomocí jednoduchého příkladu. Tabulky 8.30 a 8.31 obsahují data o výsledcích léčení těžce nemocných pacientů za jeden rok ve dvou nemocnicích A a B. Případy byly rozdeleny podle závažnosti stavu, v němž byli pacienti přijati. Pro obě tabulky jsme doložili mortalitu. Hodnoty ukazují, že mortalita v nemocnici A je menší než v nemocnici B pro oba typy pacientů. Nás logický závěr by měl být, že data indikují lepší zdravotnickou péči v nemocnici A. Sloučením obou tabulek dostáváme novou tabulku 8.32. Výpočet mortality ukazuje, že nemocnice A dosahuje horších výsledků, což vede k závěru, že lepší zdravotnickou péči má naopak pacient v nemocnici B.

Nepochybňě ve statistice platí, že čím větší je množství dat, tím dosahujeme spolehlivější výsledky. Simpsonův paradox jako by toto pravidlo zpochybňoval. Poukazuje na to, že musíme být velice opatrní, když slučujeme několik malých skupin dat do větší množiny. Může se stát, že závěr z větší množiny může být pravým opakem závěrů z menších množin.

Tab. 8.30 Příklad Simpsonova paradoxu – údaje o pacientech přijatých v relativně dobrém stavu

	Přežili	Zemřeli	Celkem	Mortalita
Nemocnice A	590	10	600	0,016
Nemocnice B	870	30	900	0,033

Tab. 8.31 Příklad Simpsonova paradoxu – údaje o pacientech v kritickém stavu

	Přežili	Zemřeli	Celkem	Mortalita
Nemocnice A	210	190	400	0,47
Nemocnice B	30	70	100	0,70

Tab. 8.32 Příklad Simpsonova paradoxu – sloučená data zahrnující oba typy pacientů

	Přežili	Zemřeli	Celkem	Mortalita
Nemocnice A	800	200	1000	0,2
Nemocnice B	900	100	1000	0,1

PŘÍKLAD 8.18

Příklad Simpsonova paradoxu: přesvědčení o „horké ruce“ košikářů

Ukážeme reálná data z oblasti sportovní statistiky, která demonstrují Simpsonův paradox. Mnoho lidí zajímajících se o basketbal věří, že u košikářů se často setkáváme se sériemi spalných nebo dobrých hodů na koš. Při zkoumání této teze o „horké ruce“ se pracovalo různými statistikami úspěšných a neúspěšných hodů na koš z americké profesionální ligy (Wardorf, 1985). Výzkumná otázka zněla, zda existuje rozdíl mezi dokumentovaným přesvědčením fanoušků o existenci tohoto jevu, a skutečnými poměry ve statistikách. Kromě toho byla analyzována data o trestných hodech. V tabulkách 8.33 a 8.34 uvádíme příslušné údaje v sezónách 1980–1982 pro Larryho Birda a Ricka Robeye z mužstva Boston Celtics. Hodnotila se vždy dvojice trestných hodů. Statistické výpočty vedou k podmíněném výděláním dobrého druhého pokusu za předpokladu, že první pokus byl úspěšný (+|+), resp. neúspěšný p(+|-).

$$\text{Bird: } p(+|+) = 251/285 = 0,881, \quad p(+|-) = 48/53 = 0,906$$

$$\text{Robey: } p(+|+) = 54/91 = 0,593, \quad p(+|-) = 49/80 = 0,612$$

Oba hráči data svědčí proti hypotéze „horké ruky“, protože $p(+|+) < p(+|-)$.

Tab. 8.33 Příklad Simpsonova paradoxu – výsledky trestních hodů L. Birda

První pokus	Druhý pokus		Celkem
	podařený (+)	mimo (-)	
podařený (+)	251	34	285
mimo (-)	48	5	53
Celkem	299	39	338

Tab. 8.34 Příklad Simpsonova paradoxu – výsledky trestních hodů R. Robeye

První pokus	Druhý pokus		Celkem
	podařený (+)	mimo (-)	
podařený (+)	54	37	91
mimo (-)	49	31	80
Celkem	103	68	171

Tab. 8.35 Příklad Simpsonova paradoxu – souhrnné výsledky trestních bodů pro oba hráče

První pokus	Druhý pokus		Celkem
	podařený (+)	mimo (-)	
podařený (+)	305	71	376
mimo (-)	97	36	133
Celkem	402	107	509

Souhrnná data pro oba hráče uvádí tabulka 8.35. Výpočty podmíněných pravděpodobností vedou pro sdružená data k hodnotám:

$$\text{Bird + Robey: } p(+|+) = 305/376 = 0,811, \quad p(+|-) = 97/133 = 0,738$$

Data ve sdružené tabulce četností podporují hypotézu, že teze „horké ruky“ má oprávnění, protože $p(+|+) = 0,811 > p(+|-) = 0,738$. Jelikož však $p(+|+) < p(+|-)$ pro Birda i Robeya a zároveň $p(+|+) > p(+|-)$ pro sdružená data, nastal v tomto případě Simpsonův paradox, takže o platnosti hypotézy „horké ruky“ musíme dálé pochybovat.

Tab. 8.36 Kontingenční tabulka pro ověření hypotézy o „horké ruce“

První pokus	Druhý pokus		Řádkové součty
	podařený (+)	mimo (-)	
podařený (+)	a	b	$a + b$
mimo (-)	c	d	$c + d$
Sloupcové součty		$a + c$	$b + d$
		n	

Tab. 8.37 Data úspěšnosti trestních hodů hráčů basketbalu

Hráč	$p(+ +)$	$p(+ -)$	a	$E(a)$	$Var(a)$	z
Kevin McHale	0,73	0,59	93	88,23	7,633	1,73
Cedric Maxwell	0,81	0,76	245	240,20	14,667	1,25
Robert Parish	0,77	0,72	164	160,75	13,061	0,90
Nate Archibald	0,83	0,82	203	202,26	8,380	0,26
Rick Robey	0,59	0,61	545	4,81	10,257	-0,25
Gerald Henderson	0,76	0,78	77	77,58	4,858	-0,26
Larry Bird	0,88	0,91	251	252,12	4,575	-0,52
Chris Ford	0,71	0,77	36	37,03	3,100	-0,58
M. L. Carr	0,68	0,81	39	41,20	3,620	-1,16

Náhodné chování údajů o trestních hodech Larryho Birda přezkoušíme pomocí statistického testu hypotézy $H_0: p(+|+) = p(+|-)$. Tato hypotéza je ekvivalentní hypotéze nezávislosti v čtyřpolní kontingenční tabulce 8.36. Testovací z -statistiky má tvar

$$z = [a - E(a)]/Var(a) = (251 - 252,12)/\sqrt{4,675} = -0,52$$

(ento test jsme popsali na s. 316).

Wardorf zpracoval údaje o dalších 7 hráčích ze stejného oddílu, aby reprezentativněji zkoumal hypotézu „horké ruky“. Ty jsou spolu s testovací statistikou z pro test nezávislosti obou pokusů uvedeny v tabulce 8.37. Testovací statistiky mají hodnoty od $-1,16$ do $1,73$. Četnosti kladných a záporných testovacích hodnot celkově nesvědčí ve prospěch teze „horké ruky“. Podmíněné pravděpodobnosti druhého úspěšného pokusu, jestliže byl první pokus úspěšný, resp. neúspěšný, pro sloučenou tabulkou četností všech devíti hráčů popisuje tabulka 8.38. Pro sloučenou tabulkou platí, že pravděpodobnost úspěšného pokusu po druhém prvním pokusu je o $4,6\% = 78,9\% - 74,3\%$ větší než po neúspěšném prvním pokusu. Hodnota testovací statistiky pro test nezávislosti $z = 2,24$ svědčí ve prospěch teze „horké ruky“.

Tab. 8.38 Podmíněné pravděpodobnosti úspěšnosti druhého trestného hodu vypočtené ze sloučených dat devíti hráčů

První pokus	Druhý pokus		Řádkové součty
	podařený (+)	mimo (-)	
podařený (+)	0,789	0,211	1,00
mimo (-)	0,743	0,257	1,00

Souhrnně lze říci, že oddelené analýzy dat jednotlivých hráčů indikují, že čtyři hráči mají lepší bilanci druhých pokusů po podařeném prvním pokusu a pět hráčů má lepší bilanci druhých pokusu po nepodařeném prvním pokusu. V kontrastu s těmito výsledky analýza tabulky sloučených dat ukazuje statisticky významný výsledek ve prospěch teze „horké ruky“.

Ve vztahu k přesvědčení fanoušků o platnosti teze „horké ruky“ lze získané výsledky interpretovat tak, že jde zřejmě o kognitivní nedostatek, kdy fanoušek není schopen uchovávat v paměti stovky kontingenčních tabulek o výkonech jednotlivých hráčů. Waldorf (1995) usuzuje, že fanoušek vychází pravděpodobně ze souhrnné tabulky o všech hráčích, která, jak jsme ukázali, v případě jednoho klubu svědčí pro tezu „horké ruky“. Mentální ekvivalent Simpsonova paradoxu vedl ke statistické iluzi, že výsledky odpovídají tezi o „horké ruce“.

Waldorf prokazuje v další analýze dat hráčů pomocí McNemarova testu symetrie pro každého hráče a spojením získaných výsledků, že trestné hody nemají stacionární charakter. Pravděpodobnosti vstřelení koše v prvním a druhém pokusu jsou rozdílné. Ve druhém pokusu mají hráči v průměru větší pravděpodobnost, že vstřelí koš.

Souhrn

Analýza kategoriálních dat se zabývá zkoumáním rozdělení četnostních dat získaných měřením kategoriálních proměnných, jež mohou být kvalitatívного (nominalního) nebo ordinálního typu. Tyto proměnné nás zajímají odděleně nebo studujeme jejich vztahy.

Pro dichotomické proměnné použijeme testy a intervaly spolehlivosti, které se týkají relativních četností. Sílu vztahu kategoriálních proměnných lze posuzovat různými koeficienty. Jejich interpretace je však obtížná, jestliže proměnné mají více kategorií. Nejznámější ze statistických testů pro kvalitatívní proměnné je asymptoticky platný χ^2 test nezávislosti v kontingenční tabulce. Přesný test se opírá o Fisherův přístup pomocí výpočtu kumulativních pravděpodobností extrémních četnostních konfigurací. Vztah dvou ordinálních proměnných kategoriálního typu hodnotíme testem Kendallová koeficientu korelace (kap. 7.2.7), testem podle Jonckheere a Terpstra (kap. 9.1.5) nebo z-testovací statistikou trendu vypočítanou pomocí vhodných skóru, přiřazených kategoriím obou proměnných. Porovnání chování kategoriální ordinální proměnné ve dvou populacích provádíme testem podle Wilcooxona (kap. 6.4.6) a ve více populacích testem Kruskala a Wallise (kap. 9.1.4).

Také v případě kategoriálních dat se setkáváme se spárovanými daty. V případě dat binárního typu je hodnotíme testem podle McNemara (kap. 8.3.2) nebo jeho zobecněním podle Bowkera (kap. 8.3.3). Pro vícekategoriální data ordinálního typu použijeme při posuzování opakování měření na skupině jedinců testy podle Friedmana (kap. 9.3.1). Jestliže chceme hodnotit závislost binární proměnné na několika nezávislých proměnných, použijeme k tomu logistickou mnohonásobnou regresi (kap. 13.2).

Naše závěry o závislosti proměnných mohou být ovlivněny působením třetí proměnné. Uvedli jsme příklad takového působení, který je znám pod názvem Simpsonův paradox. Pokud třetí proměnné zahrnujeme do analýzy, musíme použít vícerozměrné metody (kap. 13.9).

Monografii o analýze kategoriálních proměnných napsal Agresti (1990).