ANALÝZA EXPERIMENTŮ S KVALITATIVNÍ ODPOVĚDÍ

Josef Machek, MFF UK Praha

1. Úvod

Ve všech oborech výzkumné činnosti jsou velmi časté experimenty, při kterých se na jednotlivých experimentálních jednotkách neměří hodnota náhodné veličiny se spojitým rozdělením, nýbrž zjišťuje se přítomnost či nepřítomnost určitého kvalitativního znaku nebo se jednotky třídí do několika skupin /kategorií/ podle jednoho nebo několika znaků. Např. při srovnání několika terapeutických prostředků v klinickém pokusu je u každého ošetřeného jedince zaznamenáno, zda došlo k úplnému uzdravení /kategorie A_1 znaku A/, částečnému zlepšení /kategorie A_2 znaku A/, či nenastalo žádné zlepšení /kategorie A_3/. Výsledky takových pokusů bývají zaznamenány v tabulce následujícího tvaru /tzv. kontingenční tabulce/:

Kategorie znaku A	Kategorie znaku B					
	B_1	B_2	...	B_j	... B_c	
A_1	n_{11}	n_{12}	...	n_{1j}	... n_{1c}	n_{1o}
A_2	n_{21}	n_{22}	...	n_{2j}	... n_{2c}	n_{2o}
....		.....				
A_i	n_{i1}	n_{i2}		n_{ij}	n_{ic}	n_{io}
....		.....				
A_r	n_{r1}	n_{r2}	...	n_{rj}	... n_{rc}	n_{ro}
	n_{o1}	n_{o2}	...	n_{oj}	... n_{oc}	n

Taková kontingenční tabulka vyjadřuje nejčastěji jednu ze dvou následujících experimentálních situací:

a/ Kategorie znaku B /tj. sloupce tabulky/ odpovídají různým experimentálním podmínkám, různým "ošetřením" a kategorie znaku A různým možným kategoriím výsledku. Pak značí n_{oj} počet jednotek, na kterých bylo zkoušeno ošetření B_j, n_{ij} počet jednotek s ošetřením B_j, u kterých byl zjištěn výsledek A_i a n_{io} celkový počet výskytů výsledku A_i při všech ošetřeních. Cílem analýzy je zjistit, zda pravděpodobnosti různých výsledků A_i při různých ošetřeních, B_j, řekněme P_{ij}, jsou či nejsou závislé na j, tj. ověřit hypotézu $P_{ij} = P_i$ pro všechna j = 1,2, ..., c a pro všechna i = 1,2..., r. Test této hypotézy je znám jako "test homogenity". Je-li hypotéza splněna, pak ošetření se neliší ve svých účincích. Jde o úlohu obdobnou analýze rozptylu při jednoduchém třídění.

b/ A a B jsou dva různé znaky, n značí celkový počet pozorovaných jednotek, n_{ij} počet jednotek, u kterých se vyskytla kategorie A_i znaku A současně s kategorií B_j znaku B, n_{io} celkový počet jednotek se znakem A_i a n_{oj} celkový počet jednotek se znakem B_j. Cílem analýzy je ověření, zda znaky A a B jsou navzájem nezávislé či nikoliv, tj. ověření hypotézy, že pravděpodobnost současného výskytu kategorií A_i a B_j, P_{ij}, je rovna součinu pravděpodobnosti P_{io} varianty znaku A pravděpodobnosti P_{oj} varianty B_j znaku B; test hypotézy $P_{ij} = P_{io}.P_{oj}$ pro všechny dvojice i,j je znám pod názvem test homogenní nezávislosti.

Mezi uživateli statistických metod je asi nejrozšířenější - pro oba typy hypotéz - test založený na následující známé skutečnosti: jestliže marginální četnosti n_{io}, n_{oj} jsou dost velké /zpravidla se žádá, aby $n_{io}.n_{oj}/n \geq 5$ pro všechny dvojice i, j; podrobnější diskusi lze najít v [1] a [2]/, pak statistika

$$\chi^2 = \sum_{i=1}^{r} \cdot \sum_{j=1}^{c} \frac{(n_{ij} - n_{io}.n_{oj}/n)^2}{n_{io}.n_{oj}/n}$$

má v případě a/ za platnost hypotézy $P_{ij} = P$ /pro všechny dvojice i, j/ a v případě b/ za platnosti hypotézy $P_{ij} = P_{io}.P_{oj}$ rozdělení χ^2 s (r-1)(c-1) stupni volnosti. Překročí-li statistika χ^2 vypočtená z konkrétních dat 100(1 - α)%-ní kvantil rozdělení χ^2 s příslušným počtem stupňů volnosti, považuje se hypotéza homogenity, resp. hypotéza nezávislosti za pochybnou, zamítá se.

Právě popsaný test nemá příliš velkou účinnost, protože není zaměřen proti specifickým alternativám, je příliš universální. Závěry, které lze po jeho uskutečnění udělat, jsou také příliš všeobecné. Předmětem tohoto referátu jsou proto některé další způsoby analýzy kontingenčních tabulek, kterými lze v určitých zvláštních případech doplnit nebo nahradit tradiční test χ^2.

2. Odpověď s ordinálními kategoriemi

Říkáme, že kategorie A_1, A_2, ... , A_r znaku A jsou ordinální /nebo že znak A je ordinální/, jestliže je lze uspořádat do nějakého přirozeného pořadí, jako kdyby klasifikace byla založena na nějakém měřitelném znaku; např. pokusná osoba splnila daný úkol v psychotechnickém testu bez nesnází a správně /kategorie A_1/, s potížemi, ale správně /kategorie A_2/, s chybami /kategorie A_3/, vůbec nesplnila /kategorie A_4/. Pro analýzu kontingenčních tabulek s ordinálními kategoriemi lze modifikovat některé pořadové testy, jak uvádí E. L. Lehmann [4].

2. 1 Srovnání dvou ošetření při odpovědi s ordinálními kategoriemi

Nechť n_{ij}, resp. n_{i2}, $i = 1,2, \ldots ,r$, značí počet výsledků z kategorie A_i při prvním, resp. druhém ošetření n_{o1}, n_{o2} celkový počet jednotek, na nichž bylo zkoušeno první, resp. druhé ošetření $n_{io} = n_{i1} + n_{i2}$ celkový počet výsledků z kategorie A_i a konečně P_{i1} a P_{i2} pravděpodobnost výsledku A_i při prvním a druhém ošetření. Test hypotézy $P_{i1} = P_{i2}$, $i = 1,2, \ldots ,r$, tj. test hypotézy, že není rozdílu mezi ošetřeními, proti alternativnímu typu "$P_{i1} - P_{i2}$ je rostoucí funkcí i" lze odvodit ze známého Wilcoxonova testu. Na jednotky z téže kategorie A_i se pohlíží jako na jednotky, u kterých došlo ke shodě pozorovaných hodnot v obvyklém Wilcoxonově testu, tzn., že se jim přiřadí stejné "pořadí" rovné aritmetickému průměru pořadových čísel, která by tyto jednotky dostaly, kdyby klasifikace byla založena na veličině se spojitým rozdělením a sestrojí se statistika obdobná Wilcoxonově statistice W = součet pořadových čísel jednotek s ošetřením 1. Testová statistika má tedy tvar

$$W = n_{11}\frac{n_{1o}+1}{2} + n_{21}\left(n_{1o} + \frac{n_{2o}+1}{2}\right) + \ldots$$
$$+ \ldots + n_{r1}\left(n_{1o} + n_{2o} + \ldots n_{r-1,o}\frac{n_{ro}+1}{2}\right).$$

Za platnost hypotézy nulového rozdílu mezi ošetřeními má statistika W asymptoticky normální rozdělení se střední hodnotou

$$E(W) = n_{o1}(n+1)/2$$

a s rozptylem

$$Var(W) = \left\{(n_{o1}.n_{o2})/12\right\} \cdot \left\{n + 1 - \sum(N_{io}^3 - n_{io})/n(n-1)\right\}.$$

Hypotéza $P_{i1} = P_{i2}$, $i = 1,2, \ldots, r$ se tedy zamítá ve prospěch zmíněné alternativní hypotézy, když

$$W > E(W) + u_{1-\alpha}\,(Var(W))^{1/2}.$$

2. 2 Srovnání c > 2 ošetření při odpovědi s ordinálními kategoriemi

Při analýze dvojrozměrné kontingenční tabulky, ve které sloupce odpovídají ošetřením a řádky ordinálním kategoriím výsledku, lze použít analogie Kruskalovy-Wallisovy statistiky pro neparametrickou analýzu

rozptylu při jednoduchém třídění, totiž

$$K = \left\{ \frac{12}{n(n+1)} \sum_{j=1}^{c} \frac{R_j^2}{n_{oj}} - 3(n+1) \right\} / \left\{ 1 + \frac{\Sigma (d_i^3 - d_i)}{n(n+1)} \right\},$$

kde n_j jsou rozsahy výběrů, R_j součet pořadových čísel přiřazených jednotkám z j-tého výběru při uspořádání všech $n = \sum_j n_j$ jednotek podle velikosti a d_i počet jednotek se stejným pořadovým číslem, i-tým podle velikosti. Při aplikaci na kontingenční tabulku s c ošetřeními a r kategoriemi odpovědi dostane - jako při postupu z předcházejícího odstavce - všech n_{io} jednotek s výsledkem z kategorie A_i stejné pořadové číslo rovné

$$n_{1o} + n_{2o} + \ldots + n_{i-1,o} + (n_{io} + 1)/2,$$

takže bude

$$R_j = \sum_{i=1}^{r} n_{ij} \, [n_{1o} + n_{2o} + \ldots + (n_{io} + 1)/2],$$

$$d_i = n_{io}.$$

Za platnosti hypotézy, že mezi ošetřeními není rozdílu, má statistika K asymptoticky rozdělení χ^2 s (c - 1) stupni volnosti. Rozdíly v účincích ošetření se tedy považují za statisticky významné na hladině významnosti α, když statistika K překročí $100(1-\alpha)$%ní kvantil rozdělení χ^2 s (c-1) stupni volnosti.

2. 3 Test nezávislosti dvou ordinálních znaků

V případě, že kontingenční tabulka obsahuje výsledky n nezávislých pozorování tříděných podle dvou ordinálních znaků a žádá se ověření jejich nezávislosti proti alternativní hypotéze typu "vyšší kategorie znaku A se vyskytují častěji ve spojení s vyššími kategoriemi znaku B" /případně naopak/, lze použít testu odvozeného ze Spearmanova koeficientu korelace

$$r_S = 1 - 6D/(n^3 - n),$$

kde

$$D = \sum_{i=1}^{n} (R_i - S_i)^2,$$

R_i a S_i jsou pořadová čísla pozorování X_i a Y_i v náhodném výběru n dvojic (X_i, Y_i).

Při aplikaci na kontingenční tabulku s oběma znaky ordinálními se přiřadí všem jednotkám s kategorií A_i znaku A a zároveň s kategorií B_j znaku B /tj. všem jednotkám v poli i, j kontingenční tabulky/ pořadová čísla R_{ij} a S_{ij} rovná

$$R_{ij} = n_{1o} + n_{2o} + \ldots + n_{i-1,o} + (n_{io} + 1)/2,$$

$$S_{ij} = n_{o1} + n_{o2} + \ldots + n_{o,j-1} + (n_{oj} + 1)/2.$$

Veličina D ve vzorci pro r_S pak je rovna

$$D = \sum_{i=1}^{r} \sum_{j=1}^{c} n_{ij}(R_{ij} - S_{ij})^2.$$

Z asymptotických vlastností Spearmanova koeficientu korelace při výskytu stejných pozorování /tzv. vazeb/ pak plyne, že statistika r_S má za platnosti hypotézy nezávislosti znaků A a B asymptoticky normální rozdělení se střední hodnotou

$$E(r_S) = \frac{1}{2(n^3 - n)} \left\{ \sum_{i=1}^{r} (n_{io}^3 - n_{io}) + \sum_{j=1}^{c} (n_{oj}^3 - n_{oj}) \right\}$$

s rozptylem

$$Var(r_S) = \frac{1}{n - 1} \left\{ \left[1 - \frac{\Sigma (n_{io}^3 - n_{io})}{n^3 - n} \right] . \left[1 - \frac{\Sigma (n_{oj}^3 - n_{oj})}{n^3 - n} \right] \right\}.$$

3. Analýza odchylek pozorovaných četností od hypotetických

Jak jsme poznamenali už v úvodu, univerzální test χ^2 použitelný k testování homogenity i k testování nezávislosti dvou znaků, poskytuje jen velmi obecný závěr typu "ošetření mají různé účinky", případně "znaky A a B nejsou vzájemně nezávislé". K podrobnějšímu rozboru povahy rozdílů mezi ošetřeními nebo charakteru závislosti mezi znaky doporučuje S. J. Haberman [5] analyzovat jednotlivé sčítance skládající statistiku χ^2, tzv. normované residuální odchylky

$$e_{ij} = (n_{ij} - n_{io}.n_{oj}/n)/ (n_{io}.n_{oj}/n)^{1/2}.$$

Normované residuální odchylky e_{ij} mají - za platnosti hypotézy homogenity, případně nezávislosti - asymptoticky normální rozdělení s nulovou střední hodnotou a s rozptylem V_{ij}, jehož maximálně věrohodný odhad je

$$\hat{V}_{ij} = (1 - n_{io}/n)(1 - n_{oj}/n).$$

Tzv. modifikovaná rezidua $d_{ij} = e_{ij}/(\hat{V}_{ij})^{1/2}$ mají tedy - za platnosti testované hypotézy - při velkých rozsazích výběrů přibližně normální rozdělení s nulovou střední hodnotou a s jednotkovým rozptylem. S. J. Haberman doporučuje doplnit test χ^2 jednoduchou grafickou analýzou těchto modifikovaných residuí, např. seřazením hodnot d_{ij} podle velikosti v posloupnost

$$d_{(1)}, d_{(2)}, \ldots, d_{(rc)}$$

a studiem grafu bodů

$$(u_{(3k-1)/(3rc+1)}, d_{(k)}), \quad k = 1,2, \ldots, rc,$$

kde u_p značí 100p%ní kvantil rozdělení N(0,1). /Místo úseček $u_{(3k-1)/(3rc+1)}$ lze také použít tabelovaných středních hodnot pořádkových statistik výběru rozsahu rc z N(0, 1)/. Za platnosti testované hypotézy budou body seskupeny - až na náhodné odchylky - kolem přímky procházející počátkem se směrnicí 1. Residua, jejichž obrazy se od této přímky výrazně odchylují, odpovídají třídám, ve kterých je hypotéza nezávislosti, eventuálně homogenity, porušena.

4. Mnohonásobná srovnávání několika ošetření

Uvažujme dvojrozměrnou kontingenční tabulku jako v oddílu 1, ve které $B_1, B_2, \ldots, B_c$ /tj. sloupce/ odpovídají různým "ošetřením" a $A_1, A_2, \ldots, A_r$ /tj. řádky/ různým kategoriím odpovědi. Někdy je jako závěr analýzy takových dat přijatelný rozklad skupiny c ošetření na takové podskupiny, že ošetření z jedné podskupiny se mezi sebou neliší, zatímco ošetření z různých podskupin ano. Pro řešení takových úloh navrhl K. R. Gabriel [6] simultánní test odvozený z testu poměrem věrohodnosti /související také se statistikami odvozenými z teorie informace [7]/.

Nechť P je daná skupina p ošetření. Označme $n_i^P = \sum_{j \in P} n_{ij}$,

$n_o^P = \sum_{i=1}^{r} n_i^P$. K ověření hypotézy "$P_{ij} = P_i$ pro všechna $j \in P$" se užívá statistiky

$$I(P) = \sum_{j \in P} \sum_{i=1}^{r} n_{ij} . \ln n_{ij} - \sum_{j \in P} n_{oj} \ln n_{oj} - \sum_{i=1}^{r} n_i^P . \ln n_i^P + n_o^P \ln n_o^P .$$

Za platnosti testované hypotézy má veličina $2I(P)$ asymptoticky rozdělení χ^2 s $(p - 1)(r - 1)$ stupni volnosti.

Nyní nechť P_1, P_2, P_3, ... jsou různé skupiny ošetření, které lze utvořit z celkového počtu c srovnávaných ošetření. Skupinu P_k nazvěme homogenní, jestliže pro všechna $j \in P_k$ a $t \in P_k$ platí $P_{ij} = P_{it}$, $i = 1,2, \ldots ,r$. Jestliže pro některou dvojici $j \in P_k$ a $t \in P_k$ a některé i je $P_{ij} \neq P_{it}$, řekneme, že skupina P_k je heterogenní. Simultánní test spočívající v prohlášení skupiny P_k za heterogenní, jakmile statistika $2I(P_k)$ překročí $100(1 - \alpha)$%ní kvantil rozdělení s $(r - 1)(c - 1)$ stupni volnosti, má tyto vlastnosti:

1/ Jestliže některá skupina P je testem uznána homogenní, pak i všechny její podskupiny jsou uznány za homogenní.

2/ Pravděpodobnost, že některá ze skupin ve skutečnosti homogenních bude na základě testu prohlášena za heterogenní, je nejvýše rovna α.

Princip, na kterém je tento postup založen, je podobný principům simultánních testů pro větší počet kontrastů v analýze rozptylu a má také stejné výhody i nedostatky. K. R. Gabriel v citované práci uvažuje i možnost různě kombinovat též kategorie odpovědi A_i.

5. Faktoriální pokusy s alternativní odpovědí

Uvažujme pokus typu faktoriálního experimentu, ve kterém se zjišťuje vliv k faktorů; i-tý faktor má r_i tzv. úrovní, takže celkem jde o $r_1 . r_2 \ldots r_k$ ošetření. Výsledek každého ošetření se klasifikuje jen do dvou kategorií, řekněme "úspěch" a "neúspěch". Předpokládejme, že ošetření $(i_1, i_2, \ldots ,i_k)$, tj. kombinace i_1-té úrovně prvního faktoru, i_2-té úrovně druhého faktoru atd., bylo aplikováno na $n(i_1, i_2, \ldots, i_k)$ experimentálních jednotek, z nichž u $x(i_1, i_2, \ldots ,i_k)$ byl pozorován jako výsledek "úspěch". Rozbor dat je tu zjednodušen tím, že stačí analyzovat četnosti "úspěchů". Cílem rozboru je zjištění, jak dalece jednotlivé faktory ovlivňují pravděpodobnost "úspěchu" a jak dalece se případně jejich účinky vzájemně kombinují.

Označme $P(i_1,i_2, \ldots ,i_k)$ pravděpodobnost "úspěchu" při ošetření $(i_1,i_2, \ldots ,i_k)$. Odhadem této pravděpodobnosti je přirozeně relativní četnost úspěchů, tj.

$$p(i_1,i_2, \ldots ,i_k) = x(i_1,i_2, \ldots ,i_k) / n(i_1,i_2, \ldots ,i_k)$$

K vyšetření vlivu jednotlivých faktorů nebo jejich kombinací na pravděpodobnosti "úspěchu" se užívá několika postupů.

Např. Grizzle, Starmer a Koch [8] /také Johnson a Koch [9]/ připouštějí pro analýzu takových dat užití metody nejmenších čtverců na lineární model

$$P(i_1,i_2, \ldots) = m + m(i_1) + m(i_2) + \ldots + m(i_1,i_2) + m(i_1,i_3) + \ldots ,$$

kde m(...) jsou parametry obdobné hlavním efektům a interakcím ve faktoriálním experimentu s kvantitativní odpovědí. Ukazují, že tento postup dává dobré výsledky při vysokých hodnotách $n(i_1,i_2,\ldots)$ a při pravděpodobnostech úspěchu z intervalu 0,15 až 0,85. Pokud by poslední podmínka nebyla splněna, mohl by postup vést k nepřijatelným hodnotám pro pravděpodobnosti $P(i_1,i_2, \ldots)$.

Proto je účelné aplikovat teorii lineárních modelů na transformovaná data

$$y(i_1,i_2, \ldots) = h(p(i_1,i_2, \ldots)),$$

kde funkce h(.) je zvolena tak, aby proměnná y mohla nabývat jakýchkoliv reálných hodnot a aby parametry získaného lineárního modelu měly přijatelnou interpretaci. Nejlépe se osvědčuje tzv. logistická /či logitová/ transformace

$$h(p) = \ln [p/(1-p)].$$

Zapíše-li se pravděpodobnost "úspěchu" při dané kombinaci úrovní faktorů ve tvaru

$$P(i_1,i_2, \ldots) = \frac{\exp \{\beta_o + \beta_{i_1} + \beta_{i_2} + \ldots + \beta_{i_1 i_2} + \beta_{i_1 i_3} + \ldots\}}{1 - \exp \{\beta_o + \beta_{i_1} + \beta_{i_2} + \ldots + \beta_{i_1 i_2} + \beta_{i_1 i_3} + \ldots\}}$$

vyjde

$$h[P(i_1,i_2, \ldots)] = \beta_o + \beta_{i_1} + \beta_{i_2} + \ldots + \beta_{i_1 i_2} + \beta_{i_1 i_3} + \ldots ;$$

takto transformovanou pravděpodobnost $P(i_1,i_2, \ldots)$ nazveme logitem a označíme $\lambda(i_1,i_2, \ldots)$. Logit $\lambda(i_1,i_2, \ldots)$ má velmi jasný praktický význam: je to logaritmus tzv. "šance na úspěch" - je-li např. roven 2, znamená to, že v průměru připadá e^2 úspěchů na jeden neúspěch, apod. Podobně rozdíly mezi logity mají jasnou interpretaci: např. $\lambda(i_1,i_2, \ldots) - \lambda(i_1',i_2 \ldots)$ je logaritmus podílu "šance na úspěch" při ošetření $(i_1, \ldots)$ a při ošetření $(i_1',i_2, \ldots)$, čili ukazuje, kolikrát vzroste či klesne "šance na úspěch" změnou ošetření. Analýzu dat s alternativní odpovědí založenou na logitové transformaci podrobně vykládá D.R. Cox v [10]. Různé metody odhadu parametrů β_o, β_{i_1}, β_{i_2}, ... $\beta_{i_1 i_2}$, ... a testy hypotéz o nich porovnává Ch. L. Odoroff [11].

Podíly "šancí na úspěch" nazývají mnozí autoři interakcemi a jejich logaritmy - čili rozdíly logitů - logaritmickými interakcemi. O použití logaritmických interakcí při analýze složitějších kontingenčních tabulek než jsou ty, kterými se zabýváme v tomto referátu, pojednává J. Anděl [12], [13].

Literatura

[1] H. Cramér: Metody matematyczne w statystyce, Warszawa 1958, /překlad z angličtiny/

[2] W. G. Cochran: Some methods of strengthening the common test; Biometrics 10, /1954/, 417-451

[3] M. G. Kendall, A. Stuart: The Advanced Theory of Statistics, Vol. 2, London 1961

[4] E. L. Lehmann: Nonparametrics: Statistical Methods Based on Ranks, McGraw-Hill

[5] S. J. Habermann: The analysis of residuals, Biometrics 29, /1973/, 205-220

[6] K. R. Gabriel: Simultaneous test procedures for multiple comparisons on categorical data. Journal of the American Statistical Association 61 /1966/

[7] S. Kullback: Information Theory and Statistics, J. Wiley and Sons, New York, 1959

[8] J. E. Grizzle, C. F. Starmer, C. G. Koch: Analysis of categorical data by linear models, Biometrics 25, /1969/, 489-504

[9] W. D. Johnson, G. G. Koch: A note on the weighted least squares analysis of the Ries-Smith contingency table data, Technometrics 13, /1971/, 438-442

[10] D. R. Cox: Analysis of Binary Data, London, 1969

[11] Ch. L. Odoroff: A comparison of minimum logit chi-square estimation and maximum likelihood estimation in 2x2x2 and 3x2x2 contingency tables, Journal of the American Statistical Association 65, /1970/, 167-184

[12] J. Anděl: Kontingenční tabulky, Praha, 1972, skripta pro postgraduální kurs MFF UK

[13] J. Anděl: On interactions in contingency tables. Aplikace matematiky 18, 1973, 99-109