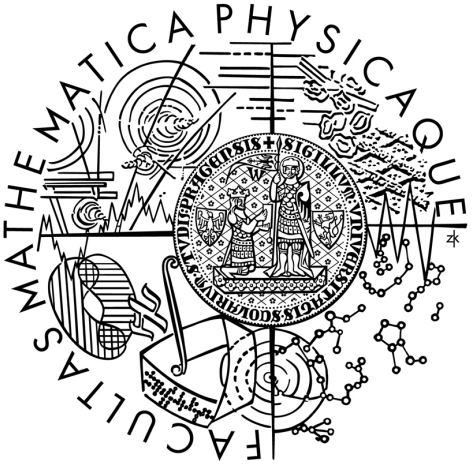
Univerzita Karlova v Praze
Matematicko-fyzikální fakulta



BAKALÁŘSKÁ PRÁCE

Martina Štěrbová

Analýza epidemiologických studií

Katedra pravděpodobnosti a matematické statistiky
Obecná matematika

Vedoucí bakalářské práce: Mgr. Michal Kulich, Ph.D.
Studijní program: Matematika

2008

Na tomto místě bych ráda poděkovala vedoucímu mojí bakalářské práce panu Mgr. Michalu Kulichovi, Ph.D. za cenné rady a čas, který mi věnoval.

Prohlašuji, že jsem svou bakalářskou práci napsala samostatně a výhradně s použitím citovaných pramenů. Souhlasím se zapůjčováním práce a jejím zveřejňováním.

V Praze dne 4. září 2008 Martina Štěrbová

Název práce: Analýza epidemiologických studií
Autor: Martina Štěrbová
Katedra: Katedra pravděpodobnosti a matematické statistiky
Vedoucí bakalářské práce: Mgr. Michal Kulich, Ph.D.
E-mail vedoucího: Michal.Kulich@mff.cuni.cz

Abstrakt: V předložené práci budeme studovat sběr a analýzu dat, které vypovídají o vztahu mezi výskytem nemoci a působením sledovaných faktorů (expozicemi) v populaci. Budeme diskutovat o tom jak určit, zda pozorovaný vztah není pouze náhodný a zda má daná expozice na jedince vliv škodlivý nebo ochranný. Ukážeme různé metody sběru dat. Uvedeme také různé metody měření vztahu mezi expozicí a nemocí. Dále se budeme zabývat eliminací zkreslení vztahu mezi expozicí a nemocí způsobeného přítomností rušivých faktorů.

Práce nabízí i příklady k některým tématům.

Klíčová slova: studie plánů, relativní riziko, zavádějící efekt

Title: Analysis of epidemiological studies
Author: Martina Štěrbová
Department: Department of Probability and Mathematical Statistics
Supervisor: Mgr. Michal Kulich, Ph.D.
Supervisor's e-mail address: Michal.Kulich@mff.cuni.cz

Abstract: In this work we study the collection and analysis of data, which reveal the relationship between the incidence of diseases and effects of controlled factors (exposures) in the population. We will discuss how to determine whether the observed relationship is merely random and whether the exposure has either harmful or protective influence on the individuals. We present different methods of the data collection. Further we focus on the different methods of measuring the relationship between the exposure and the disease. In addition, we deal with eliminating the distortion of the relationship between exposure and illness, caused by the presence of interfering factors.

This work offers examples to some sections.

Keywords: study design, relative risk, confounding effect

Obsah

1 Úvod

Epidemiologie je vědní obor, který se zabývá studiem faktorů ovlivňujících zdraví a událostí spjatých se zdravotním stavem lidské populace a aplikací těchto poznatků při řešení zdravotních problémů. Zásadním předpokladem epidemiologického výzkumu je respektování základních principů medicíny, etiky a sociální spravedlnosti.

Práce epidemiologů zahrnuje zkoumání vzniku nemoci, výběr vhodné studie, sběr a analýzu dat s ohledem na vývoj statistických modelů, sestavení hypotézy a sepsání závěrů. Epidemiologické studie je možno klasifikovat podle několika hledisek. Máme dvě základní skupiny:

Observační studie je studie, při které nezasahujeme do chodu událostí, pouze zaznamenáváme, klasifikujeme, počítáme a statisticky analyzujeme pozorovaná zjištění.

Experimentální studie je studie, která je pod naší kontrolou, určujeme jakému režimu expozice bude kdo podroben. Podmínky realizace lze nastavit podle potřeb výzkumu, proto tyto studie poskytují nejpřesnější informace. Realizace těchto studií je limitována etickými a právními omezeními, protože nelze někoho úmyslně vystavit např. virům nebo podávat horší léky.

V této práci bude popsán sběr a analýza dat vypovídajících o vztahu mezi výskytem nemoci a charakteristickými vlastnostmi jedinců v populaci. K měření velikosti vztahu se v epidemiologii užívají hlavně tzv. ukazatele asociace, mezi které patří zejména relativní riziko, poměr šancí a atributivní riziko, s nimiž se dále seznámíme. Pomocí statistických postupů lze testovat, zda je asociace jevů statisticky významná, ale nelze rozhodnout, zda je jeden jev příčinou jiného. K tomu je potřeba umět zkombinovat mnoho přímých i nepřímých informací a vydedukovat z nich závěry, které co nejlépe odpovídají epidemiologickým i statistickým poznatkům.

2 Měření asociace mezi nemocí a expozicí

Základním cílem epidemiologických studií je zpravidla zjistit, zda existuje asociace (vztah) mezi onemocněním a působením určité látky (expozice) podezřelé jako původce nemoci a zda je tento vztah příčinný. K měření velikosti asociace se v epidemiologii užívají hlavně tzv. ukazatele asociace.

2.1 Relativní riziko

Relativní riziko (relative risk) pro výstup D asociované s binárním rizikovým faktorem E, značíme RR a definujeme následovně:

$$RR = \frac{\mathrm{P}(D|E)}{\mathrm{P}(D|E^c)}. \tag{1}$$

Vlastnosti:

- je zřejmé, že RR musí být nezáporné
- pokud $RR = 1 \Rightarrow D$ a E jsou nezávislé
- pokud $RR > 1 \Rightarrow$ je větší riziko D za přítomnosti E
- pokud $RR < 1 \Rightarrow$ naopak

Relativní riziko není symetrické:

$$\frac{\mathrm{P}(D|E)}{\mathrm{P}(D|E^c)} \neq \frac{\mathrm{P}(E|D)}{\mathrm{P}(E|D^c)}.$$

Pozn.: Pokud vyjde $RR = 10$, je riziko onemocnění za přítomnosti rizikového faktoru 10-krát větší, než kdyby tam nebyl.

2.2 Poměr šancí

Pojem *šance* výskytu jevu, značíme O, je definován jako poměr pravděpodobnosti výskytu jevu A a pravděpodobnosti jevu opačného k jevu A, což formálně zapisujeme jako

$$O(A) = \frac{\mathrm{P}(A)}{\mathrm{P}(A^c)} = \frac{\mathrm{P}(A)}{1 - \mathrm{P}(A)}.$$

Proto šance výskytu onemocnění D v exponované populaci E je dána poměrem podmíněných pravděpodobností $\mathrm{P}(D|E)$ a $\mathrm{P}(D^c|E)$, tedy

$$O(D|E) = \frac{\mathrm{P}(D|E)}{\mathrm{P}(D^c|E)}$$

a podobně šance výskytu onemocnění v neexponované populaci je

$$O(D|E^c) = \frac{\mathrm{P}(D|E^c)}{\mathrm{P}(D^c|E^c)}.$$

Poměr šancí (odds ratio) je definován následovně

$$OR = \frac{O(D|E)}{O(D|E^c)} = \frac{\mathrm{P}(D|E)}{\mathrm{P}(D^c|E)} \Big/ \frac{\mathrm{P}(D|E^c)}{\mathrm{P}(D^c|E^c)} \tag{2}$$

a udává, kolikrát je vyšší šance výskytu nemoci u exponované populace ve srovnání s neexponovanou populací.

Vlastnosti:

- je zřejmé, že OR musí být nezáporné
- pokud $OR = 1 \Rightarrow D$ a E jsou nezávislé
- pokud $OR > 1 \Rightarrow$ je větší riziko D za přítomnosti E
- pokud $OR < 1 \Rightarrow$ naopak

Poměr šancí je symetrický:

$$\frac{\mathrm{P}(D|E)}{\mathrm{P}(D^c|E)} \Big/ \frac{\mathrm{P}(D|E^c)}{\mathrm{P}(D^c|E^c)} = \frac{\mathrm{P}(E|D)}{\mathrm{P}(E^c|D)} \Big/ \frac{\mathrm{P}(E|D^c)}{\mathrm{P}(E^c|D^c)} \,.$$

2.3 Atributivní riziko

Ne v každém případě je nemoc důsledkem expozice, proto nás zajímá, kolik nemocí v populaci můžeme vysvětlit přítomností rizikového faktoru E. Atributivní riziko (attributable risk) je další měření vztahu mezi D a E vyjadřující podíl nemoci ve studované populaci, jehož vznik lze vysvětlit vlivem expozice. Značíme AR a definujeme následujícím způsobem:

$$AR = \frac{\mathrm{P}(D) - \mathrm{P}(D|E^c)}{\mathrm{P}(D)}. \tag{3}$$

Dále můžeme dosadit $\mathrm{P}(D) = \mathrm{P}(D|E)\,\mathrm{P}(E) + \mathrm{P}(D|E^c)\,\mathrm{P}(E^c)$ podle věty o celkové pravděpodobnosti a získat

$$\begin{aligned} AR =& \frac{\mathrm{P}(D|E)\,\mathrm{P}(E) + \mathrm{P}(D|E^c)\,\mathrm{P}(E^c) - \mathrm{P}(D|E^c)}{\mathrm{P}(D)} \\ &\vdots \\ =& \frac{\mathrm{P}(E) \cdot (RR - 1)}{1 + \mathrm{P}(E) \cdot (RR - 1)}. \end{aligned} \tag{4}$$

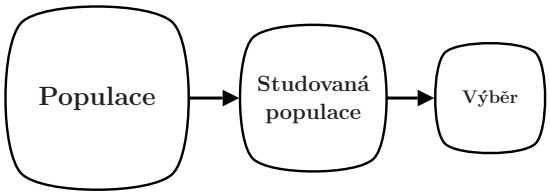
Vlastnosti:

- může nabývat i záporných hodnot, nikdy nepřekročí hodnotu 1
- pokud $AR = 0 \Rightarrow D$ a E jsou nezávislé
- pokud $AR < 0 \Rightarrow$ expozice E má ochranný efekt
- pokud $0 < AR \leq 1 \Rightarrow$ s roustoucí E roste riziko onemocnění

3 Plány studie

Pojem plán studie (study design) zahrnujeme všechny aspekty spojené se zařazením jedinců do studie a sběrem dat i aspekty spojené s vyhodnocením studie. Jde tedy zejména o vymezení sledované populace a znaků (proměnných), které budou zjišťovány, o metody zjišťování a sběru dat (a tím i o typ studie) a o volbu vhodných postupů technického zpracování a statistické analýzy dat. Tyto aspekty je třeba uvažovat v kontextu proveditelnosti studie z hlediska finančních nákladů a technických možností.

Je důležité definovat následující pojmy. *Cílová populace* je množina jedinců, ve které můžeme použít naše odhady a závěry týkající se vztahu mezi nemocí a expozicí. *Studovaná populace* je dostupná populace z cílové populace, kterou jsme schopni testovat. *Studovaný výběr* zahrnuje testované jedince ze studované populace, pro které sbíráme data o nemoci, expozici a dalších faktorech.



Obrázek 1: Schematické znázornění vztahu mezi cílovou populací, studovanou populací a výběrem.

Jak získáme studované náhodné výběry? Máme tři základní formy, které se nejvíce používají v epidemiologických studiích. V každé se soustředíme na vztah mezi D a E. Data získaná z výběru zapisujeme do tabulky 1, tzv. tabulka 2×2.

3.1 Populační studie

Hlavními kroky této studie jsou:

1. Vezmeme prostý náhodný výběr o velikosti n ze studované populace.

Tabulka 1: Tabulka 2×2.

		nemoc	
		D	D^c
vliv	E	·	·
	E^c	·	·

2. Zjistíme presenci a absenci D a E u všech testovaných jedinců.

V bodě dva není zahrnuto pořadí v jakém jsou jevy D a E měřeny. Proto je nutné provést další klasifikaci na:

Prospektivní studie (prospective study) je epidemiologická studie, ve které se k měření expozice přistupuje dříve než dojde k výskytu sledovaného zdravotního jevu, např. onemocnění, úmrtí.

Retrospetivní studie (retrospective study) je studie, ve které se k měření expozice přistupuje až po zjištění sledovaného jevu.

Tato klasifikace může mít značný vliv na kvalitu a platnost měření vystavení vlivu nemoci. Například při odhadu vlivu v retrospektivní studii musíme ohodnotit stupeň expozice v místě před onemocněním a zjistit, zda měření nejsou ovlivněna stavem nemocného jednince. Ještě poznamenejme, že prospektivní měření jevu D může vyžadovat deset nebo dvacet let po výběru.

Z dat získaných pomocí studie založené na populaci můžeme získat ze studovaného výběru následující pravděpodobnosti:

- $\mathrm{P}(D \wedge E)$, $\mathrm{P}(D \wedge E^c)$, $\mathrm{P}(D^c \wedge E)$, $\mathrm{P}(D^c \wedge E^c)$
- $\mathrm{P}(D)$, $\mathrm{P}(E)$, $\mathrm{P}(D^c)$, $\mathrm{P}(E^c)$
- $\mathrm{P}(D|E)$, $\mathrm{P}(D|E^c)$, $\mathrm{P}(E|D)$, $\mathrm{P}(E|D^c)$

3.2 Kohortové studie

Termín *kohorta* původně označoval oddíl římského vojska, desetinu legie, a mnohem později se začal přeneseně používat pro skupiny jedinců v epidemiologickém výzkumu. Často jsou ve výzkumu definovány právě dvě kohorty, exponovaná a neexponovaná. Hlavními kroky studie jsou (viz obr. 2):

1. Vytvoříme dvě podskupiny populace na základě presence a absence jevu E.
2. Vezmeme prostý náhodný výběr z obou skupin z prvního kroku o velikostech n_E a n_{E^c}.
3. Postupně zjišťujeme presenci a absenci jevu D u jednotlivců v obou náhodných výběrech.

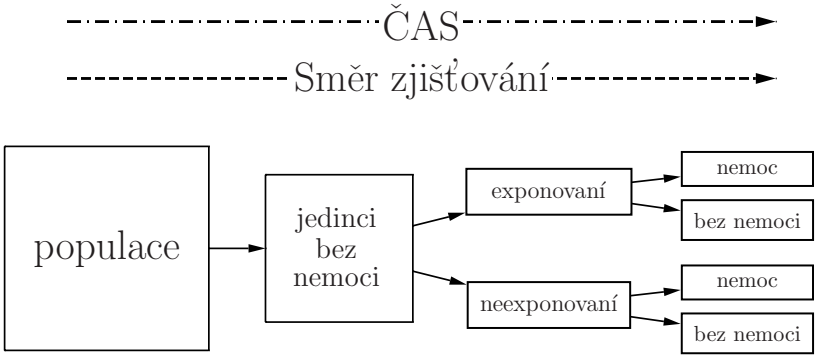
Z dat získaných touto studií nemůžeme zjistit následující pravděpodobnosti:

- $\mathrm{P}(D \wedge E)$, $\mathrm{P}(D \wedge E^c)$, $\mathrm{P}(D^c \wedge E)$, $\mathrm{P}(D^c \wedge E^c)$
- $\mathrm{P}(D)$, $\mathrm{P}(E)$, $\mathrm{P}(D^c)$, $\mathrm{P}(E^c)$

Pouze následující podmíněné pravděpodobnosti jsou nám k dispozici:

- $\mathrm{P}(D|E)$, $\mathrm{P}(D|E^c)$, $\mathrm{P}(D^c|E)$, $\mathrm{P}(D^c|E^c)$

Tedy můžeme spočítat RR a OR, ale AR nelze, protože nemáme k dispozici $P(E)$. Ještě poznamenejme, že se jedná o formu dlouhodobobé studie.



Obrázek 2: Základní princip kohortové studie, tj. sledování skupin od expozice k následku.

3.3 Studie kontrol a případů

Studie tohoto typu začíná identifikací případů, tj. jedinců se sledovanou nemocí. Tito jedinci jsou pak porovnáváni s jinou skupinou jedinců, kteří nevykazují danou nemoc, tzv. *kontrolami*. V obou takto sestavených skupinách porovnáváme, jaká byla v minulosti expozice. Pokud byla expozice vyšší mezi případy, potenciální rizikový faktor může skutečně být rizikovým faktorem. Hlavními kroky studie jsou (viz obr. 3):

1. Vytvoříme dvě podskupiny populace na základě presence a absence jevu D.
2. Vezmeme prostý náhodný výběr z obou skupin z předchozího kroku o velikostech n_D a n_{D^c}.
3. Postupně zjišťujeme presenci a absenci jevu E u jednotlivců v obou náhodných výběrech.

Ze získaných dat nemůžeme zjistit následující pravděpodobnosti:

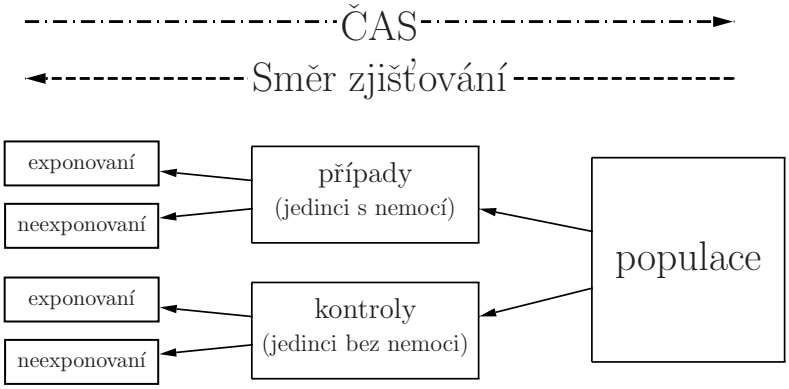
- $\mathrm{P}(D \wedge E)$, $\mathrm{P}(D \wedge E^c)$, $\mathrm{P}(D^c \wedge E)$, $\mathrm{P}(D^c \wedge E^c)$
- $\mathrm{P}(D)$, $\mathrm{P}(E)$, $\mathrm{P}(D^c)$, $\mathrm{P}(E^c)$

Pouze podmíněné pravděpodobnosti za podmínky D můžeme získat:

- $\mathrm{P}(E|D)$, $\mathrm{P}(E|D^c)$, $\mathrm{P}(E^c|D)$, $\mathrm{P}(E^c|D^c)$

Tedy můžeme spočítat pouze OR pro E za podmínky D (shoduje se s OR pro D za podmínky E). V situaci, kdy se D objevuje vzácně v populaci exponované i neexponované, se OR blíží k RR, takže OR získané díky této metodě můžeme použít jako odhad RR.

Na první pohled se zdá, že nám tento plán neumožňuje odhadnout AR. Nicméně v případě vzácně se vyskytujících chorob, můžeme získat aproximaci



Obrázek 3: Schéma studie případů a kontrol.

a to následujícím způsobem. Máme

$$\begin{aligned}\mathrm{P}(D|E^c) &= \mathrm{P}(D|E^c)\cdot(\mathrm{P}(E^c)+\mathrm{P}(E)) \\ &= \mathrm{P}(E^c)\,\mathrm{P}(D|E^c)+\frac{\mathrm{P}(D|E)\,\mathrm{P}(E)}{RR}.\end{aligned}$$

Dosazením do (3) a použitím Bayesova vzorce, dostáváme

$$\begin{aligned}AR &= 1-\frac{\mathrm{P}(E^c)\,\mathrm{P}(D|E^c)}{\mathrm{P}(D)}-\frac{\mathrm{P}(D|E)\,\mathrm{P}(E)}{RR\cdot\mathrm{P}(D)} \\ &= 1-\mathrm{P}(E^c|D)-\frac{\mathrm{P}(E|D)}{RR} \\ &= \mathrm{P}(E|D)\left(1-\frac{1}{RR}\right). \qquad (5)\end{aligned}$$

4 Význam tabulky 2×2

Již máme ponětí o tom, jak shromažďovat informace o nemoci expozici. V této kapitole budeme zjišťovat, jestli D a E mají nějakou spojitost, zda jsou závislé. V řeči testování hypotéz

$$\mathrm{H}_0 \ : \ D \text{ a } E \ \text{ jsou nezávislé } \ (\text{tj. } RR = 1, \ OR = 1). \tag{6}$$

4.1 Plán populační studie

Nezávislost D a E je ekvivalentní s $\mathrm{P}(D \wedge E) = \mathrm{P}(D)\,\mathrm{P}(E)$. Všechny tyto pravděpodobnosti můžeme získat z tabulky 2×2, která třídí testované jedince. Z tabulky 2 dostáváme:

- $\mathrm{P}(D \wedge E)$ můžeme odhadnout pomocí $\frac{a}{n}$
- $\mathrm{P}(D)$ můžeme odhadnout pomocí $\frac{a+c}{n}$
- $\mathrm{P}(E)$ můžeme odhadnout pomocí $\frac{a+b}{n}$

Pokud čísla $a,\ b,\ c,\ d$ odpovídají dokonale hypotéze H_0, má statistika

$$\frac{a}{n} - \frac{(a+c)}{n} \cdot \frac{(a+b)}{n} = \frac{(ad-bc)}{n^2} \tag{7}$$

hodnotu 0. Čím více se bude statistika vzdalovat od 0, tím spíše danou hypotézu H_0 zamítáme, kladná hodnota značí pozitivní závislost mezi jevy E a D a záporná naopak. Abychom mohli sestrojit přesný kritický obor testu na hladině α, potřebujeme znát rozdělení naší testové statistiky (7) za platnosti H_0.

Víme, že náhodná veličina $(ad-bc)$ má pro n jdoucí do nekonečna asymtoticky normální rozdělení. Za platnosti H_0 má náhodná veličina $(ad-bc)$ nulovou střední hodnotou a její rozptyl odhadneme (viz [1, str. 60])

$$\widehat{\mathrm{Var}}(ad-bc) = \frac{(a+b)(a+c)(b+d)(c+d)}{n}.$$

Tabulka 2: Vzorová tabulka.

		nemoc		
		D	D^c	
vliv	E	a	b	$a+b$
	E^c	c	d	$c+d$
		$a+c$	$b+d$	$n=a+b+c+d$

Tedy pro n jdoucí do nekonečna má veličina $U = (ad - bc)/\sqrt{\widehat{\mathrm{Var}}(ad - bc)}$ asymptoticky normální rozdělení $\mathrm{N}(0, 1)$. Proto víme, že rozdělení U^2 se blíží rozdělení χ^2 s jedním stupněm volnosti, tj. χ^2_1. Nulovou hypotézu na hladině významnosti α zamítáme v případě, že

$$U^2 = \frac{(ad - bc)^2}{\widehat{\mathrm{Var}}(ad - bc)} = \frac{n(ad - bc)^2}{(a + b)(c + d)(a + c)(b + d)} \geq \chi^2_1(\alpha), \tag{8}$$

kde $\chi^2_1(\alpha)$ je kritická hodnota χ^2_1 na hladině α.

Tímto testem se zbavujeme znaménka, proto neříká nic o tom, zda je závislost pozitivní nebo negativní v tom smyslu, zda má expozice efekt škodlivý nebo ochraný, tedy šlo o test s oboustrannou alternativou. Pokud bychom se touto problematikou chtěli zabývat, testujeme hypotézu následujícími způsoby. Když platí

$$U \geq u(\alpha), \tag{9}$$

kde $u(\alpha)$ je kritická hodnota normálního rozdělení $\mathrm{N}(0, 1)$, zamítáme H_0 na hladině α ve prospěch alternativní hypotézy, že expozice má vliv škodlivý. Když platí

$$U \leq -u(\alpha), \tag{10}$$

kde $u(\alpha)$ je kritická hodnota normálního rozdělení $\mathrm{N}(0, 1)$, zamítáme H_0 na hladině α ve prospěch alternativní hypotézy, že expozice má vliv ochranný. Takto se tedy provedou jednostranné testy hypotézy.

4.2 Plán kohortové studie

Jedná se o observační plán studie, kde je pacient vybrán na základě expozice, tj. exponovaný a neexponovaný. Nezávislost D a E je ekvivalentní s

$$\mathrm{P}(D|E) = \mathrm{P}(D|E^c).$$

Tedy testujeme nulovou hypotézu

$$\mathrm{H}_0 : \mathrm{P}(D|E) = \mathrm{P}(D|E^c).$$

Pro zjednodušení označme $p_1 = \mathrm{P}(D|E)$ a $p_2 = \mathrm{P}(D|E^c)$. Použitím tabulky 2 jsme získali odhady

$$\hat{p}_1 = \frac{a}{a + b} = \frac{a}{n_1}$$
$$\hat{p}_2 = \frac{c}{c + d} = \frac{c}{n_2},$$

kde n_1 a n_2 jsou velikosti výběrů E a E^c. Odhady $\hat{p}_1$ a $\hat{p}_2$ jsou náhodné veličiny s asymptoticky normáním rozdělením, střední hodnota a rozptyl náhodné

veličiny $\hat{p}_1$ jsou p_1 a $p_1(1-p_1)/n_1$, analogicky i pro $\hat{p}_2$. Rozdíl mezi dvěma odhady $\hat{p}_1$ a $\hat{p}_2$ má asymptoticky normální rozdělení se střední hodnotou $p_1 - p_2$ a rozptylem $p_1(1-p_1)/n_1 + p_2(1-p_2)/n_2$, protože je zřejmé, že $\hat{p}_1$ a $\hat{p}_2$ jsou nezávislé (exponované a neexponované výběry jsou nezávislé).

Za platnosti H_0 (tj. $p_1 = p_2$) je $E(\hat{p}_1 - \hat{p}_2) = 0$ a $\mathrm{Var}(\hat{p}_1 - \hat{p}_2) = p(1-p)(\frac{1}{n_1} + \frac{1}{n_2})$, kde $p_1 = p_2 = p$. Dále $\widehat{\mathrm{Var}}(\hat{p}_1 - \hat{p}_2) = \hat{p}(1-\hat{p})(\frac{1}{n_1} + \frac{1}{n_2})$, kde $\hat{p} = \frac{a+c}{n}$ je odhad hodnoty p založený na sjednocené populaci (exponovaní a neexponovaní). Tedy

$$(\hat{p}_1 - \hat{p}_2)/\sqrt{\widehat{\mathrm{Var}}(\hat{p}_1 - \hat{p}_2)} \sim_{as} \mathrm{N}(0,1), \tag{11}$$

za platnosti H_0, tj. $(\hat{p}_1 - \hat{p}_2)^2/\widehat{\mathrm{Var}}(\hat{p}_1 - \hat{p}_2)$ se blíží χ^2_1. Pokud zjistíme, že

$$U^2 = \frac{(\hat{p}_1 - \hat{p}_2)^2}{\widehat{\mathrm{Var}}(\hat{p}_1 - \hat{p}_2)} = \frac{n(ad-bc)^2}{(a+b)(c+d)(a+c)(b+d)} \geq \chi^2_1(\alpha), \tag{12}$$

kde $\chi^2_1(\alpha)$ je kritická hodnota χ^2_1 na hladině α, pak H_0 zamítáme na hladině α. Jedná se o χ^2 test nezávislosti pro čtyřpolní tabulky [2].

Tímto testem se opět zbavujeme znaménka, proto neříká nic o tom, zda je závislost pozitivní nebo negativní v tom smyslu, zda má expozice efekt škodlivý nebo ochraný. Pokud bychom se touto problematikou znovu chtěli zabývat testujeme hypotézu stejnými způsoby jako v předchozím případě způsoby. Když platí

$$U \geq u(\alpha), \tag{13}$$

kde $u(\alpha)$ je kritická hodnota normálního rozdělení $\mathrm{N}(0,1)$, zamítáme H_0 na hladině α ve prospěch alternativní hypotézy, že expozice má vliv škodlivý. Když platí

$$U \leq -u(\alpha), \tag{14}$$

kde $u(\alpha)$ je kritická hodnota normálního rozdělení $\mathrm{N}(0,1)$, zamítáme H_0 na hladině α ve prospěch alternativní hypotézy, že expozice má vliv ochranný. Takto se tedy provede jednostranný test hypotézy.

4.3 Plán studie případů a kontrol

Jedná se o stejný případ jako v předchozí kapitole, jen zaměníme role E a D. Testujeme hypotézu $H_0 : P(E|D) = P(E|D^c)$ a z tabulky 2 získáváme odhady

$$\widehat{p_1} = \frac{a}{a+c} = \frac{a}{n_1}$$

$$\widehat{p_2} = \frac{b}{b+d} = \frac{b}{n_2},$$

kde $p_1 = P(E|D)$ a $p_2 = P(E|D^c)$. Dále postupujeme naprosto analogicky. Za platnosti H_0 je $E(\hat{p}_1 - \hat{p}_2) = 0$ a $\mathrm{Var}(\hat{p}_1 - \hat{p}_2) = p(1-p)(\frac{1}{n_1} + \frac{1}{n_2})$, kde $p_1 =$

Tabulka 3: Konzumace ryb a onemocnění srdce.

	CHD	**bez CHD**	$\sum$
jedli ryby	34	227	261
nejedli ryby	42	163	205
$\sum$	79	390	466

$p_2 = p$. Dále $\widehat{\mathrm{Var}}(\hat{p}_1 - \hat{p}_2) = \hat{p}(1-\hat{p})(\frac{1}{n_1} + \frac{1}{n_2})$, kde $\hat{p} = \frac{a+b}{n}$ je odhad hodnoty p založené na sjednocené populaci (zdraví a nemocní). Dojdeme k totožné testové statistice (12), akorát došlo k záměně b a c.

4.4 Příklad

V tomto příkladě nás bude zajímat, jestli konzumace ryb ovlivňuje kardiovaskulární onemocnění. Budeme testovat hypotézu, že konzumace ryb nemá vliv na kardiovaskulární onemocnění. K výpočtu použijeme tabulku 3. Nejdříve použijeme test χ^2 pro čtyřpolní tabulky (11)

$$\begin{aligned} U^2 &= \frac{n(ad-bc)^2}{(a+b)(c+d)(a+c)(b+d)} \\ &= \frac{466(34 \cdot 163 - 227 \cdot 42)^2}{261 \cdot 205 \cdot 79 \cdot 390} \\ &= 4,68. \end{aligned}$$

Zjistili jsme tedy, že $U^2 = 4,68$, protože $U^2 \geq \chi^2_1(0,05) = 3,84$, zamítneme hypotézu, že konzumace ryb nemá vliv na kardiovaskulární onemocnění. V tomto případě si můžeme položit otázku, zda konzumace ryb má škodlivý efekt.

Pro kontrolu provedeme ještě jednostranný test této hypotézy

$$\begin{aligned} U &= \frac{\frac{a}{n_1} - \frac{c}{n_2}}{\sqrt{\frac{a+c}{n}\left(1 - \frac{a+c}{n}\right)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \\ &= \frac{\frac{34}{261} - \frac{42}{205}}{\sqrt{\frac{79}{466}\left(1 - \frac{79}{466}\right)\left(\frac{1}{261} + \frac{1}{205}\right)}} \\ &= -2,13. \end{aligned}$$

Zjistili jsme tedy, že $U = -2,13$, protože $U = -2,13 \leq -u(0,05) = -1,64$, zamítáme hypotézu, že konzumace ryb nemá vliv na vznik kardiovaskulární onemocnění ve prospěch alternativní hypotézy, že konzumace ryb má ochranný efekt na vznik tohoto onemocnění!

5 Odhady a odvození míry asociace

Nyní se zaměříme na odhad míry vztahu mezi nemocí a expozicí, navazujeme na druhou kapitolu. Chceme tyto vztahy vyjádřit pomocí intervalů spolehlivosti.

5.1 Poměr šancí

Odhad OR můžeme získat pro plán populační studie a kohortové studie, potřebujeme pouze odhad $P(D|E)$ a $P(D|E^c)$, které snadno získáme z tabulky 2. Odhady podmíněných pravděpodobností, tedy dosadíme do vzorce OR a získáme

$$\widehat{OR} = \left[\frac{a/(a+b)}{b/(a+b)}\right] \Big/ \left[\frac{c/(c+d)}{d/(c+d)}\right] = \frac{ad}{bc}\,.$$

OR je symetrické, proto pro dvojice $P(D|E)$, $P(D|E^c)$ a $P(E|D)$, $P(E|D^c)$ vychází odhad stejně, tedy $\widehat{OR}$ nezávisí na volbě plánu.

Pro n_1 i n_2 jdoucí do nekonečna má $\widehat{OR}$ asymptoticky normální rozdělení. Ovšem prro konečná n_1 a n_2 je rozdělení $\widehat{OR}$ asymetrické (zprava zešikmené), proto pro rychlejší konvergenci k normálnímu rozdělení použijeme logaritmickou transformaci.

Interval spolehlivosti pro OR

To, že $\ln\widehat{OR}$ má opravdu asymptoticky normální rozdělení, můžeme dokázat pomocí následující věty, která nám navíc umožní získat odhad rozptylu $\ln\widehat{OR}$.

Věta (Δ - metoda)
Nechť $\{X_n\}_{n=1}^{\infty} = \boldsymbol{X}$ je náhodný vektor, který splňuje $\sqrt{n}(\boldsymbol{X} - \boldsymbol{\mu}) \xrightarrow{d} \mathrm{N}_k(\mathbf{0}, \boldsymbol{\sigma^2})$ pro nějaké $\boldsymbol{\mu} \in \mathbb{R}^k$, $\boldsymbol{\sigma^2}$ (pozitivně semidefinitní matice). Nechť g je spojitá diferencovatelná funkce z $\mathbb{R}^k$ do $\mathbb{R}^p$. Označme $D(\mathbf{x}) = \frac{\partial g(\mathbf{x})}{\partial \mathbf{x}}$ (spojitá, $p \times k$ matice funkcí). Pak

$$\sqrt{n}(g(\boldsymbol{X}) - g(\boldsymbol{\mu})) \xrightarrow{d} \mathrm{N}_p(0, D(\boldsymbol{\mu})\,\boldsymbol{\sigma^2}\,D(\boldsymbol{\mu})^T).$$

Důkaz: Viz Serfling (2002). □

Tvrzení
Náhodná veličina $\ln\widehat{OR}$ má asymptoticky normální rozdělení a její rozptyl lze odhadnout

$$\widehat{\mathrm{Var}}\left(\ln\widehat{OR}\right) = \frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d}.$$

Důkaz:
Víme, že

$$\sqrt{n_1}\,(\hat{p}_1 - p_1) \xrightarrow{d} \mathrm{N}\,(0, p_1(1-p_1)) \;\text{ pro }\; n_1,\, n_2 \to \infty$$

použitím Δ - metody dostáváme

$$\sqrt{n_1}\left(g(\hat{p}_1) - g(p_1)\right) \xrightarrow{d} \mathrm{N}\left(0, p_1(1-p_1)\cdot[g'(p_1)]^2\right) \text{ pro } n_1,\, n_2 \to \infty,$$

kde $g(\hat{p}_1) = \ln\frac{\hat{p}_1}{1-\hat{p}_1}$, $g(p_1) = \ln\frac{p_1}{1-p_1}$ a $g'(p_1) = \frac{1}{p_1(1-p_1)}$, tedy

$$\sqrt{n_1}\left(\ln\frac{\hat{p}_1}{1-\hat{p}_1} - \ln\frac{p_1}{1-p_1}\right) \xrightarrow{d} \mathrm{N}\left(0, \frac{1}{p_1(1-p_1)}\right) \text{ pro } n_1,\, n_2 \to \infty.$$

Náhodná veličina $\ln\frac{\hat{p}_1}{1-\hat{p}_1}$ má tedy asymptoticky normální rozdělení s rozptylem $\frac{1}{n_1 p_1(1-p_1)}$. Analogicky zjistíme, že $\ln\frac{\hat{p}_2}{1-\hat{p}_2}$ má taktéž asymptoticky normální rozdělení s rozptylem $\frac{1}{n_2 p_2(1-p_2)}$. Protože $\ln\frac{\hat{p}_1}{1-\hat{p}_1}$ a $\ln\frac{\hat{p}_2}{1-\hat{p}_2}$ jsou nezávislé (exponované a neexponované výběry jsou nezávislé), pak podle [2, Lemma 4.7] má jejich rozdíl opět asymptoticky normální rozdělení a rozptyl je roven součtu jejich rozptylů, tj.

$$\mathrm{Var}\left(\ln\frac{\hat{p}_1}{1-\hat{p}_1} - \ln\frac{\hat{p}_2}{1-\hat{p}_2}\right) = \frac{1}{n_1 p_1(1-p_1)} + \frac{1}{n_2 p_2(1-p_2)}.$$

Odhad rozptylu dostaneme tak, že za p_1 dosadíme $\hat{p}_1$ a za p_2 dosadíme $\hat{p}_2$, tedy

$$\begin{aligned}\widehat{\mathrm{Var}}\left(\ln\widehat{OR}\right) &= \widehat{\mathrm{Var}}\left(\ln\frac{\hat{p}_1}{1-\hat{p}_1} - \ln\frac{\hat{p}_2}{1-\hat{p}_2}\right)\\ &= \frac{1}{n_1\hat{p}_1(1-\hat{p}_1)} + \frac{1}{n_2\hat{p}_2(1-\hat{p}_2)}\\ &= \frac{1}{a}+\frac{1}{b}+\frac{1}{c}+\frac{1}{d}. \qquad \square\end{aligned}$$

Tento postup můžeme aplikovat na všechny tři plány. Dále máme, že

$$\left(\ln\widehat{OR} - \ln OR\right) \Big/ \sqrt{\widehat{\mathrm{Var}}(\ln\widehat{OR})} \sim_{as} \mathrm{N}(0,1),$$

proto $100(1-\alpha)\%$ interval spolehlivosti pro $\ln OR$ lze vyjádřit ve tvaru

$$\left(\ln\widehat{OR} - z_{1-\frac{\alpha}{2}}\sqrt{\widehat{\mathrm{Var}}(\ln\widehat{OR})}, \ln\widehat{OR} + z_{1-\frac{\alpha}{2}}\sqrt{\widehat{\mathrm{Var}}(\ln\widehat{OR})}\right),$$

kde $z_{1-\frac{\alpha}{2}}$ je $(1-\frac{\alpha}{2})$-kvantil standardizovaného normálního rozdělení $\mathrm{N}(0,1)$, např. pro hladinu významnosti $\alpha = 0,05$ je tato hodnota rovna 1,96 [6]. Na základě mezí pro $\ln OR$ vypočteme $100(1-\alpha)\%$ interval spolehlivosti pro OR

$$\left(\widehat{OR}\cdot\exp\left\{-z_{1-\frac{\alpha}{2}}\sqrt{\widehat{\mathrm{Var}}(\ln\widehat{OR})}\right\}, \widehat{OR}\cdot\exp\left\{z_{1-\frac{\alpha}{2}}\sqrt{\widehat{\mathrm{Var}}(\ln\widehat{OR})}\right\}\right).$$

5.2 Relativní riziko

Pro výpočet RR potřebujeme $\mathrm{P}(D|E)$ a $\mathrm{P}(D|E^c)$. Obojí můžeme získat ze studie založené na populaci nebo kohortové studie. Postupujeme stejně jako v předchozí kapitole, zavedeme substituci z tabulky 2 a dostáváme

$$\widehat{RR} = \frac{a/(a+b)}{c/(c+d)}.$$

Rozdělení $\widehat{RR}$ je také zešikmené, opět použijeme transformaci pomocí logaritmu, abychom se přiblížili normálnímu rozdělení. Odhadneme $\ln RR$ jako $\ln \widehat{RR} = \ln(a/(a+b))/(c/(c+d))$. Použijeme stejné značení jako v kapitole 5.1 a získáme $\ln \widehat{RR} = \ln \hat{p}_1 - \ln \hat{p}_2$. Stejným postupem jako v předchozím tvrzení dostáváme rozptyl

$$\widehat{\mathrm{Var}}(\ln \widehat{RR}) = \frac{b}{a(a+b)} + \frac{d}{c(c+d)}.$$

Dále máme, že

$$\left(\ln \widehat{RR} - \ln RR\right) \Big/ \sqrt{\widehat{\mathrm{Var}}(\ln \widehat{RR})} \sim_{as} \mathrm{N}(0,1),$$

proto $100(1-\alpha)\%$ interval spolehlivosti pro $\ln RR$ lze vyjádřit ve tvaru

$$\left(\ln \widehat{RR} - z_{1-\frac{\alpha}{2}} \cdot \sqrt{\widehat{\mathrm{Var}}(\ln \widehat{RR})}, \ln \widehat{RR} + z_{1-\frac{\alpha}{2}} \cdot \sqrt{\widehat{\mathrm{Var}}(\ln \widehat{RR})}\right),$$

kde $z_{1-\frac{\alpha}{2}}$ je $(1-\frac{\alpha}{2})$-kvantil standardizovaného normálního rozdělení $\mathrm{N}(0,1)$. Na základě mezí pro $\ln OR$ vypočteme $100(1-\alpha)\%$ interval spolehlivosti pro RR jako u OR.

5.3 Atributivní riziko

Pro populační studii

Pro výpočet $\widehat{AR}$ potřebujeme odhady $\mathrm{P}(E)$ a RR nebo odhady $\mathrm{P}(D)$ a $\mathrm{P}(D|E^c)$. V prvním případě použijeme vzorec 4. V druhém případě dosadíme následující odhady $\hat{\mathrm{P}}(D) = (a+c)/n$ a $\hat{\mathrm{P}}(D|E^c) = c/(c+d)$ do (3) a získáme

$$\widehat{AR} = \frac{\frac{a+c}{n} + \frac{c}{c+d}}{\frac{a+c}{c}} = \frac{ad-bc}{(a+c)(c+d)}.$$

Rozdělení $\widehat{AR}$ je zešikmeno opačným směrem (zleva), doporučuje se použít transformaci $\ln(1-\widehat{AR})$. Atributivní riziko má smysl pouze tehdy, je-li zkoumaný faktor skutečně rizikový, pak hodnota atributivního rizika leží mezi 0 a 1. Rozptyl $\ln(1-\widehat{AR})$ odhadneme následovně [1, str. 85]

$$\widehat{\mathrm{Var}}(\ln(1-\widehat{AR})) = \frac{b + \widehat{AR}(a+d)}{nc}.$$

Tabulka 4: Cestující na palubě Titanicu.

	Muži		**Ženy**		
	nepřežili	**přežili**	**nepřežili**	**přežili**	$\sum$
1. třída	118	61	5	139	323
2. třída	146	25	12	94	277
3. třída	418	75	110	106	709
$\sum$	682	161	127	339	1309

Interval spolehlivosti pro $\ln(1 - AR)$ získáme standardním způsobem jako v předchozích případech a interval pro AR dostaneme pomocí exponenciely, pak odečteme 1 a nakonec násobíme -1.

Pro kohortovou studii

Není možné odhadnout AR, protože nemůžeme odhadnout $\mathrm{P}(E)$.

Pro studii kontrol a případů

V tomto případě použijeme ekvivalentní formulaci atributivního rizika [1, str. 84]

$$AR = \mathrm{P}(E|D)\left(1 - \frac{1}{RR}\right).$$

S těmito daty můžeme odhadnout $\mathrm{P}(E|D)$ a OR, které lze použít pro odhad RR. Takže použitím $\hat{\mathrm{P}}(E|D) = a/(a+c)$ a $\widehat{OR} = ad/bc$ dostáváme

$$\widehat{AR} = \frac{a}{a+c}\left(1 - \frac{bc}{ad}\right) = \frac{ad - bc}{d(a+c)}.$$

Odhad rozptylu $\ln(1 - \widehat{AR})$ je dán

$$\widehat{\mathrm{Var}}(\ln(1 - \widehat{AR})) = \frac{a}{c(a+c)} + \frac{b}{d(b+d)}.$$

Stejným způsobem jako v předchozím případě dostaneme intervaly spolehlivosti pro $\ln(1 - AR)$ a AR.

5.4 Příklad

Na stránkách *http://www.crcpress.com/e_products/downloads/* jsem našla data o cestujících lodi Titanic. Obsahují informace o tom kolik umřelo žen a mužů,

Tabulka 5: Muži a ženy na palubě Titanicu.

	nepřežili	**přežili**	$\sum$
muži	682	161	843
ženy	127	339	466
$\sum$	809	500	1309

Tabulka 6: 1. a 2. třída versus 3. třída.

muži i ženy	**nepřežili**	**přežili**	$\sum$
3. třída	528	181	709
1. a 2. třída	281	319	600
$\sum$	809	500	1309

v jakých třídách cestovali. Hledala jsem závislosti mezi pohlavím a úmrtím, mezi úmrtím a třídou, kterou cestovali (viz tabulka 4).

Pomocí tabulky 5 spočteme odhad relativního rizika $\widehat{RR} = 2,969$, který udává, kolikrát častěji se vyskytne úmrtí u mužů oproti ženám na palubě, tedy tento výsledek ukazuje, že se umrtí u mužů vyskytovalo 2,969-krát častěji než u žen. Metodou z kapitoly 5.2 dostáváme 95% interval spolehlivosti pro $\ln RR$ jako (0,936; 1,24) a následně 95% interval spolehlivosti pro RR jako (2,55; 3,46). I z intervalu spolehlivosti je patrné, že pohlaví ovlivnilo smrt pasažérů.

Pomocí tabulky 6 spočteme odhad relativního rizika $\widehat{RR} = 1,59$, který udává, kolikrát častěji se vyskytne úmrtí u pasažérů třetí třídy oproti pasažérům 1. a 2. třídy, tedy tento výsledek ukazuje, že se umrtí u cestujících 3. třídy vyskytovalo 1,59-krát častěji než u cestujících 1. a 2. třídy. Metodou z kapitoly 5.2 dostáváme 95% interval spolehlivosti pro $\ln RR$ jako (0,368; 0,56) a následně 95% interval spolehlivosti pro RR jako (1,445; 1,75). I z intervalu spolehlivosti je patrné, že cestování ve 3. třídě bylo mnohem nebezpečnější.

6 Zavádějící faktor

V epidemiologii prakticky nelze zaručit, aby se srovnávané skupiny lišily pouze ve sledovaném faktoru, např. expozici. Právě působení dalších faktorů, které souvisejí s expozicí i s nemocí, stojí u kořenů jevu nazývaného *zavádějící efekt* (confounding effect).

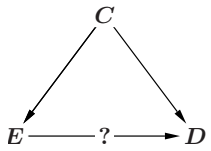
6.1 Kauzální závěr

Zapomeňme na chvíli, že většinu epidemiologických dat získáváme pozorováním lidí. Představme si nejlepší možný svět pro experimentování, kdybychom chtěli vědět zda změna v expozici způsobuje změny při vzniku nemoci za stejných podmínek. V laboratoři bychom nechali běžet experimenty, jeden s expozicí a druhý bez. Rozdíl mezi výsledky bychom přisoudili změně v přítomnosti expozice. Kdybychom tento postup opakovali na dalších skupinách s expozicí a bez expozice, pak by nám průměrný rozdíl ve výsledcích testů poskytl míru kauzálního účinku expozice. Tato strategie vyžaduje určení velikosti expozice před výskytem nemoci.

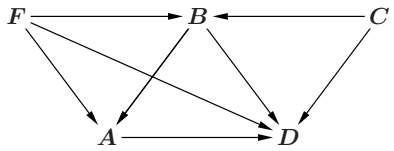
Zavádějící proměnná

Nyní pro vysvětlení pojmu zavádějící proměnné uvedeme příklad o vlivu pití kávy na rakovinu slinivky. Nechť E je vliv pití kávy, D rakovina slinivky a C označuje pohlaví jedince. Poznamenejme, že riziko rakoviny slinivky je u mužů vyšší než u žen. Speciálně pro tento příklad předpokládejme, že muži pijí kávu s větší pravděpodobností než ženy. Je vidět, že rizikovost onemocnění závisí i na pohlaví. Tedy můžeme prohlásit, že pohlaví jedince je zavádějící proměnná. Na obr. 4 můžeme vidět schématický popis zavádějící proměnné, kde šipky reprezentují kauzální účinek. V této jednoduché situaci musí mít C dvě vlastnosti

- musí být kauzálně spojená s D,
- musí být kauzálně spojená s E.



Obrázek 4: Schématický popis zavádějící proměnné.



Obrázek 5: Orientovaný acyklický graf.

Kontrola zavádějící proměnné pomocí stratifikace

Nejjednodušší postup, jak kontrolovat zavádějící proměnnou, tj. jak eliminovat zavádějící účinky, je tzv. stratifikace. Celou populaci rozdělíme do skupin tedy strat, které sdílí stejné hodnoty C, v našem případě jedna strata pouze ženy a druhá pouze muži. Měření vztahu mezi D a E se v jednotlivých stratech může lišit. Pokud zjistíme, že vztah mezi D a E se mění v jednotlivých stratech C, můžeme prohlásit, že C modifikuje účinek E na D.

6.2 Kauzální grafy

Velký počet zavádějících proměnných situaci komplikuje, proto je nutné zavést složitější grafy než v předchozím případě.

Základní terminologie Pojmem *orientovaný graf* se v teorii grafů označuje takový graf, jehož hrany jsou upořádané dvojice. *Orientovaná cesta* je posloupnost uzlů propojených hranami se stejnou orientací, ve které se žádná hrana neopakuje. *Acyklický graf* je orientovaný graf bez cyklů, tj. neexistuje cesta, která by měla začátek a konec v jednom uzlu. Nechť $A \rightarrow B \rightarrow C$, tedy B je *syn* A, C je *syn* B, B má *předka* A a C má dva *předky* A a B.

Nově definujeme pojem *zadní cesta* (backdoor path), je to cesta z uzlu X do uzlu Y, která začíná opuštěním X přes hranu, která směřuje do X (tj. proti proudu), a potom pokračuje do Y bez ohledu na orientaci hran. Na obr. 5 vidíme, že například $A{-}F{-}B{-}D$ je zadní cesta z A do D. Uzel X na vybrané cestě se nazývá *kolidátor* (collider), jestliže hrany této cesty do uzlu pouze směřují. Cesta je *blokovaná* (blocked), pokud obsahuje nejméně jeden kolidátor. Na obr. 5 máme acyklický graf, kde D je kolidátor v cestě $C{-}D{-}A{-}F{-}B$, která je zásluhou D blokovaná.

Vztahy v kauzálních grafech

Na obr. 5 vidíme, že

- F způsobuje A, B a D
- F způsobuje D přes A i B
- F nezpůsobuje C

Presence a absence cesty nám udává vztah mezi uzly. Závislost mezi dvěma proměnnými X a Y implikuje existenci nejméně jedné neblokované cesty mezi X a Y, ekvivalentně, všechny blokované cesty mezi X a Y implikují jejich nezávislost. Obráceně, existence neblokovné cesty mezi X a Y má skoro vždy za následek, že X a Y jsou závislé.

Pokud jsou všechny cesty mezi dvěma uzly blokované, pak jsou nezávislé, v našem případě (obr. 5) F a C jsou nezávilé. Přítomnost neblokované cesty nám dává závislost mezi uzly.

Užití kauzálních grafů k odhalení prezence nebo absence zavádění

Výhodou těchto grafů je, že nám umožní zjistit, zda vztah mezi E a D je ovlivněn zavádějící proměnnou. Na obrázku 6(a) vidíme klasický kauzální graf popisující C jako zavádějící proměnnou a na obrázku 6(b) C není zavádějící, protože je sama ovlivňována proměnnou E, navíc v tomto případě cesty $E{-}D$ a $E{-}C{-}D$ popisují dva odlišné kauzální účinky. Určit zda existuje zavádění

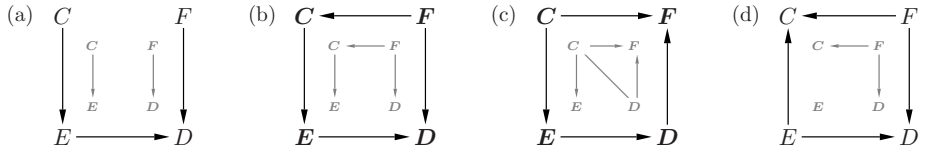
Obrázek 6: Kauzální grafy

můžeme pomocí dvou kroků. Mějme tedy acyklický orientovaný graf, postupujeme následovně:

- Vymažeme všechny cesty z E ukazující na jiné uzly.
- V tomto redukovaném grafu určíme, zda existuje neblokovaná zadní cesta $E{\rightarrow}D$, pokud ano, pak vztah $E{-}D$ je ovlivňován zavádějícími proměnnými.

6.3 Řízení zavádění v kauzálních grafech

Úspěšná strategie jak zvládnout zavádění (viz kapitola 6.1) je stratifikace populace do skupin se stejnou hodnotou zavádějící proměnné C. Stratifikace C je ekvivaletní s jejím odstraněním z grafu. Uvažujeme tedy kauzální grafy pro každou stratu zvlášť bez C, ve kterých opět zjišťujeme, zda se tam vyskytuje zavádění pomocí metody z kapitoly 6.2. Pokud se v grafu objevuje kolidátor není dobré graf rozvrstvit podle něj, protože se začnou tvořit nové cesty mezi proměnnými, které tam doposud nebyly.



Obrázek 7: Pouze v grafu (b) jsme našli zavádění, které po stratifikaci na C mizí, byla použita metoda z kapitoly 6.3, jejíž výsledek je ilustrován uvnitř původních kauzálních grafů.

Pravidla nalezení zavádějící proměnné v kauzálním grafu

Definujeme množinu stratifikačnch faktorů $S = \{C_1, \cdots, C_s\}$. Dále nás zajímá, jestli po stratifikaci na faktory v S zůstává v grafu stále zavádění. To můžeme zjistit pomocí následujících tří kroků:

1. Vymazat šipky vedoucí z E do jiných uzlů.
2. Přidat neorientovanou hranu mezi uzly, které mají stejného potomka v množině stratifikovaných faktorů S.
3. V novém grafu ověřit existenci neblokované zadní cesty z E do D, která neprochází skrze uzly v S. Pokud neexistuje, máme pod kontrolou všechny zavádějící proměnné.

Na obr. 7 jsou uvedeny čtyři příklady, na nichž je uvedený postup aplikován.

7 Kontrola vnějších faktorů

V předchozí kapitole jsme se zabývali stratifikací, nyní se budeme zabývat stratifikovanou analýzou, která je založená na měření síly asociace v jednotlivých, dobře definovaných, homogenních stratech, určených hodnotou zavádějící proměnné. Je-li pohlaví zavádějícím faktorem, pak se síla asociace mezi expozicí a odezvou vypočte nejprve pro každé pohlaví (pro stratum mužů a stratum žen) zvlášť a poté se sumarizuje v celkovou asociaci. Tato jednoduchá statistická technika umožňuje i identifikaci potenciálních statistických interakcí mezi expozicí a zavádějícími faktory. Ta se projeví výraznou odchylkou síly asociace ve stratech se specifickou úrovní kontrolovaných faktorů.

7.1 Sumární test závislosti pro tabulky 2×2

Chceme zevšeobecnit χ^2 test pro vztah z kapitoly 4. Předpokládejme, že data jsou stratifikována do I strat, takže pokud stratifikujeme podle pohlaví a věku, kde věk dělíme do tří kategorií (25-35 let, 35-45 let, 45-60 let), dostáváme $I = 6$. Pro každou stratu vytvoříme zvlášť tabulku 2×2, viz tabulka 7.1. Po stratifikaci dat potřebujeme sumární (celkový) test závislosti D a E, který používá stratifikovaná data, dále chceme zjistit i intervaly spolehlivosti pro OR a RR.

Cochranův-Mantelův-Haenszelův test

Stanovíme nulovou hypotézu a to H_0 : D a E jsou nezávislé podmíněně, je-li dáno C. Přesná formulace hypotézy

$$\mathrm{H}_0: \quad OR_1 = OR_2 = \cdots = OR_I = 1$$

nebo

$$\mathrm{H}_0: \quad RR_1 = RR_2 = \cdots = RR_I = 1,$$

kde OR_i je poměr šancí pro $E-D$ v i-té stratě a RR_i relativní riziko v i-té stratě.

Chceme sestrojit test hypotézy H_0. Musíme se rozhodnout, na co má být test citlivý (na jaký typ porušení testované hypotézy H_0). Máme dvě možnosti:

Tabulka 7: Vzorová tabulka.

		nemoc		
		D_i	D_i^c	
vliv	E_i	a_i	b_i	$a_i + b_i$
	E_i^c	c_i	d_i	$c_i + d_i$
		$a_i + c_i$	$b_i + d_i$	$n_i = a_i + b_i + c_i + d_i$

První možnost

$$\chi^2 = \sum_{\mathrm{i=1}}^{\mathrm{I}} \frac{n_i(a_i d_i - b_i c_i)^2}{(a_i + b_i)(c_i + d_i)(a_i + c_i)(b_i + d_i)} \tag{15}$$

Při tomto testu nezávislosti vypočteme tedy χ^2 podle (15). Pokud vyjde $\chi^2 \geq \chi^2_I(\alpha)$, kde $\chi^2_I(\alpha)$ je kritická hodnota χ^2_I na hladině α, zamítáme H_0 na hladině α. Tento test zamítne hypotézu, pokud se jednotlivé hodnoty $\widehat{OR}_i$ (resp. $\widehat{RR}_i$) výrazně liší od 1, i když jejich průměr je blízko 1. Tento test bychom použili v případě, že připouštíme jako alternativní hypotézu situaci, kdy se závislost E a D liší mezi jednotlivými straty.

Druhá možnost

$$\chi^2_{CHM} = \left(\sum_{i=1}^{I} (a_i d_i - b_i c_i) \right)^2 \Bigg/ \sum_{i=1}^{I} \frac{(a_i + b_i)(c_i + d_i)(a_i + c_i)(b_i + d_i)}{n_i} \tag{16}$$

Při tomto testu nezávislosti vypočteme tedy χ^2_{CHM} podle (16). Pokud vyjde $\chi^2_{CHM} \geq \chi^2_1(\alpha)$, kde $\chi^2_1(\alpha)$ je kritická hodnota χ^2_1 na hladině α, zamítáme H_0 na hladině α. Tento test není citlivý na kolísání jednotlivých $\widehat{OR}_i$ (resp. $\widehat{RR}_i$), pokud průměr je blízký 1. Testová statistika (16) se nazývá Cochranova-Mantelova-Haenszelova testová statistika. Statistika χ^2_{CHM} je zobecněním χ^2 statistiky pro test nezávislosti ve čtyřpolní tabulce [2, str. 284].

Cochranův-Mantelův-Haenszelův test je tedy založen na předpokladu, že v základní populaci je velikost asociace ve všech stratech stejná, a že tudíž odlišnosti mezi jednotlivými straty ve velikosti asociace (pozorované prostřednictvím dílčích odhadů poměrů šancí nebo relativních rizik) jsou jen projevem výchylek způsobených náhodnými vlivy.

7.2 Sumární odhady a intervaly spolehlivosti pro OR

Nyní popíšeme dvě základní strategie pro odhad poměru šancí a relativního rizika.

Woolfova metoda

Woolfova metoda je založena na přepokladu, že OR_i jsou ve všech stratech stejná (toto bychom mohli nazvat homogenita závislosti), a použití váženého průměru, protože máme-li dva či více souborů s výrazně rozdílným počtem hodnot, ze kterých chceme vypočítat celkový průměr, musíme zohlednit tyto rozdílné počty pomocí vah w_i, které musí být navíc nezáporné (tj. $w_i \geq 0,\ i =$

$0, \cdots, 1)$. Protože rozdělení poměru šancí je zešikmené, použijeme opět logaritmickou transformaci. Tedy vážený průměr je dán

$$\ln \widehat{OR}_W = \frac{\sum_{i=1}^{I} w_i \ln \widehat{OR}_i}{\sum_{i=1}^{I} w_i},$$

kde w_i jsou váhy a $\ln \widehat{OR}_i$ je předchozí odhad $\ln OR$ pro i-tou stratu. Váhy pro dostatečně velké výběry ve všech stratech jsou definovány jako převrácené hodnoty odhadů rozptylů, tj.

$$w_i = \left(\frac{1}{a_i} + \frac{1}{b_i} + \frac{1}{c_i} + \frac{1}{d_i} \right)^{-1}.$$

Pokud velikosti výběrů v každé stratě budou dostatečně velké, pak každé $\ln \widehat{OR}_i$ bude mít asymptoticky normální rozdělení. Vzhledem k homogenitě závislosti má $\ln \widehat{OR}_W$ asymptoticky normálním rozdělení. Protože střední hodnota každého $\ln \widehat{OR}_i$ je $\ln \widehat{OR}$ za předpokladu nezávislosti, pak středni hodnota $\ln \widehat{OR}_W$ za stejného předpokladu je

$$\mathrm{E} \ln(\widehat{OE}_W) = \frac{\sum_{\mathrm{i}=1}^{\mathrm{I}} w_i \, \mathrm{E}(\ln \widehat{OR}_i)}{\sum_{\mathrm{i}=1}^{\mathrm{I}} w_i} = \frac{\sum_{\mathrm{i}=1}^{\mathrm{I}} w_i \ln OR}{\sum_{\mathrm{i}=1}^{\mathrm{I}} w_i} = \ln OR.$$

Dále odhad rozptylu získáme

$$\begin{aligned} \widehat{\mathrm{Var}}(\ln \widehat{OR}_W) &= \frac{\sum_{\mathrm{i}=1}^{\mathrm{I}} w_i^2 \widehat{\mathrm{Var}}(\ln \widehat{OR}_i)}{(\sum_{\mathrm{i}=1}^{\mathrm{I}} w_i)^2} \\ &= \frac{\sum_{\mathrm{i}=1}^{\mathrm{I}} w_i^2 (w_i)^{-1}}{(\sum_{\mathrm{i}=1}^{\mathrm{I}} w_i)^2} \\ &= \frac{1}{\sum_{\mathrm{i}=1}^{\mathrm{I}} w_i}. \end{aligned}$$

Tedy $100(1-\alpha)\%$ interval spolehlivosti pro $\ln OR$, založený na Woolfově metodě, je dán

$$\ln \widehat{OR}_W \pm z_{1-\frac{\alpha}{2}} \cdot \sqrt{\widehat{\mathrm{Var}}(\ln \widehat{OR}_W)},$$

kde $z_{1-\frac{\alpha}{2}}$ je $(1-\frac{\alpha}{2})$-kvantil standardizovaného normálního rozdělení $\mathrm{N}(0,1)$. Interval spolehlivosti pro OR získáme jako v přechozích případech.

Mantelova-Haenszelova metoda

Jedná se o další metodu průměrování odhadů přes všechny straty za přepokladu, že OR_i jsou ve všech stratech stejná (homogenita závislosti). Metoda funguje velmi dobře i bez ohledu na velikosti výběrů, tj. tuto statistiku lze vypočítat,

i když se v dílčích tabulkách vyskytují nulové pozorované četnosti, ale musíme dávát pozor, aby se nám ve jmenovateli následujícího vzorce neobjevila 0. $\widehat{OR}_{MH}$ můžeme vyjádřit jako vážený průměr poměru šancí $\widehat{OR}_i$ z jednotlivých tabulek

$$\widehat{OR}_{MH} = \frac{\sum_{i=1}^{I} w_i^* \widehat{OR}_i}{\sum_{i=1}^{I} w_i^*},$$

kde váhy jsou dány $w_i^* = b_i c_i / n_i$ [1, str. 130]. Protože víme, že $\widehat{OR}_i = a_i d_i / b_i c_i$ (viz kapitola 5.1), máme k dispozici ekvivalentní vyjádření Mantelova-Haenszelova odhadu pro OR

$$\widehat{OR}_{MH} = \frac{\sum_{i=1}^{I} (a_i d_i / n_i)}{\sum_{i=1}^{I} (b_i c_i / n_i)}.$$

Cochranova-Mantelova-Haenszelova statistika se rovná 0 (nezávislost) právě tehdy, když $\sum_{\mathrm{i=1}}^{\mathrm{I}} (a_i d_i - b_i c_i) = 0$, což je ekvivalentní s $\sum_{\mathrm{i=1}}^{\mathrm{I}} (a_i d_i / n_i) = \sum_{\mathrm{i=1}}^{\mathrm{I}} (b_i c_i / n_i)$=1. Rozptyl odhadneme následovně (Robins a kolektiv 1986):

$$\widehat{\mathrm{Var}}(\ln \widehat{OR}_{MH}) = \frac{\sum_{i=1}^{I} \left(\frac{a_i+d_i}{n_i}\right)\left(\frac{a_i d_i}{n_i}\right)}{2\left(\sum_{i=1}^{I} \frac{a_i d_i}{n_i}\right)^2} + \frac{\sum_{i=1}^{I} \left(\frac{a_i+d_i}{n_i} \cdot \frac{b_i c_i}{n_i} + \frac{b_i+c_i}{n_i} \cdot \frac{a_i d_i}{n_i}\right)}{2\left(\sum_{i=1}^{I} \frac{a_i d_i}{n_i}\right)\left(\sum_{i=1}^{I} \frac{b_i c_i}{n_i}\right)} + \frac{\sum_{i=1}^{I} \left(\frac{b_i+c_i}{n_i}\right)\left(\frac{b_i c_i}{n_i}\right)}{2\left(\sum_{i=1}^{I} \frac{b_i c_i}{n_i}\right)^2}.$$

Tedy $100(1-\alpha)\%$ interval spolehlivosti pro $\ln OR$, založený na Mantelově-Haenszelově metodě, je dán

$$\ln \widehat{OR}_{MH} \pm z_{1-\frac{\alpha}{2}} \cdot \sqrt{\widehat{\mathrm{Var}}(\ln \widehat{OR}_{MH})},$$

kde $z_{1-\frac{\alpha}{2}}$ je $(1-\frac{\alpha}{2})$-kvantil standardizovaného normálního rozdělení $\mathrm{N}(0,1)$. Interval spolehlivosti pro OR získáme jako v přechozích případech.

7.3 Sumární odhady a intervaly spolehlivosti pro RR

Obě metody použité k vytvoření odhadu a intervalu spolehlivosti pro OR z 2×2 tabulek jdou snadno upravit tak, aby zahrnuly i odhad RR pro data z populační studie nebo kohortové studie.

Jako v případě OR pomocí Woolfovy metody dostáváme $100(1-\alpha)\%$ interval spolehlivosti pro $\ln OR$

$$\ln \widehat{RR}_W \pm z_{1-\frac{\alpha}{2}} \cdot \sqrt{\widehat{\mathrm{Var}}(\ln \widehat{RR}_W)},$$

kde $z_{1-\frac{\alpha}{2}}$ je $(1-\frac{\alpha}{2})$-kvantil standardizovaného normálního rozdělení $\mathrm{N}(0,1)$. Interval spolehlivosti pro RR získáme jako v předchozích případech.

Další metodou je opět Mantelův-Haenszelův odhad pro RR, viz kapitola 7.2, tedy

$$\widehat{RR}_{MH} = \frac{\sum_{i=1}^{I} w_i^* \ln \widehat{RR}_i}{\sum_{i=1}^{I} w_i^*},$$

kde $w_i^* = c_i(a_i + b_i)/n_i$ [1, str. 135]. Dále Mantelův-Haenszelův odhad pro RR

$$\widehat{RR}_{MH} = \frac{\sum_{i=1}^{I} \frac{a_i(c_i+d_i)}{n_i}}{\sum_{i=1}^{I} \frac{c_i(a_i+b_i)}{n_i}}.$$

Ke konstrukci příslušného intervalu spolehlivosti je potřeba odhad rozptylu [1, str. 135] pro $\ln \widehat{RR}_{MH}$, tedy

$$\widehat{\mathrm{Var}}(\ln \widehat{RR}_{MH}) = \frac{\sum_{\mathrm{i=1}}^{\mathrm{I}}[(a_i + b_i)(c_i + d_i)(a_i + c_i) - a_i c_i n_i]/n_i^2}{\sum_{\mathrm{i=1}}^{\mathrm{I}}(a_i(c_i + d_i)/n_i) \sum_{\mathrm{i=1}}^{\mathrm{I}}(c_i(a_i + b_i)/n_i)}.$$

Intervaly spolehlivosti pro $\ln RR$ i pro RR získáme analogicky.

Literatura

[1] *Nicholas P Jewell:* **Statistics for epidemiology**
Chapman and Hall/CRC, Boca Raton 2004

[2] *Anděl J.:* **Základy matematické statistiky**
Matfyzpress, Praha 2005

[3] *Anděl J.:* **Statistické metody**
Matfyzpress, Praha 2003

[4] *Zvárová J., Malý M. a kolektiv:* **Biomedicínská statistika III.**
Karolinum, Praha 2003

[5] *Zajíček L.:* **Vybrané úlohy z matematické analýzy pro 1. a 2. ročník**
Matfyzpress, Praha 2006

[6] *Dupač V., Hušková M.:* **Pravděpodobnost a matematická statistika**
Karolinum, Praha 2003

[7] *Serfling R. J.:* **Approximation theorems of mathematical statistics**
Wiley, New York 2002