Testování statistických hypotéz a analýza kategoriálních dat

Prof. RNDr. Jana Zvárová, DrSc.

Záznam epidemiologických dat

Populace	Rizikový faktor: Přítomen	Rizikový faktor: Nepřítomen	Celkem
Nemocní	a	b	$a+b$
Kontroly	c	d	$c+d$
Celkem	$a+c$	$b+d$	n
	Exponovaní	Neexponovaní	

a, b, c, d ... pozorované (absolutní) četnosti v jednotlivých skupinách

$$n = a + b + c + d$$

Epidemiologické ukazatele

Populace	Rizikový faktor: Přítomen	Rizikový faktor: Nepřítomen	Celkem
Nemocní	a	b	$a+b$
Kontroly	c	d	$c+d$
Celkem	$a+c$	$b+d$	n
	Exponovaní	Neexponovaní	

Incidence exponovaných
$$I_E = \frac{a}{a+c}$$

Incidence neexponovaných
$$I_N = \frac{b}{b+d}$$

Incidenci většinou přepočítáváme na 1 000, 10 000 nebo 100 000 osob.

Epidemiologické ukazatele

Relativní riziko
$$RR = \frac{I_E}{I_N}$$

- odhaduje *sílu asociace mezi rizikovým faktorem a nemocí*
- vyjadřuje, **kolikrát častěji** se může **nemoc vyvinout** v populaci **exponovaných** ve srovnání s populací neexponovaných

Epidemiologické ukazatele

Atributivní riziko

$$AR = I_E - I_N$$

- část incidence exponované populace, která může být vysvětlena *pouze přítomností rizikového faktoru*
- umožňuje odhalit stupeň maximálního poklesu výskytu onemocnění u exponované populace v případě, že umíme odstranit vliv rizikového faktoru

Příklad

Chceme ověřit, zda progresivní polyartritida (PAP) souvisí s výskytem antigenu HLA-DR4.

Domníváme se, že ano (to je naše medicínská hypotéza).

Sestavíme tedy nulovou a alternativní hypotézu (nezapomeňte, že nulovou hypotézu volíme opačně, než je dokazované tvrzení). Tedy:

H_0: PAP nesouvisí s výskytem HLA-DR4
H_1: PAP souvisí s výskytem HLA-DR4

Volba hladiny významnosti

- *hladina významnosti* souvisí s chybami, kterých se při rozhodnutí můžeme dopustit:

Rozhodnutí	Skutečnost	
	H_0 platí	H_1 platí
Nezamítneme H_0 (nevýznamný výsledek)	Správné rozhodnutí	Chyba II. druhu (β)
Zamítneme H_0 (významný výsledek)	Chyba I. druhu (α)	Správné rozhodnutí

- **hladina významnosti (** α **)** je předepsaná hodnota, kterou pravděpodobnost chyby I. druhu nesmí překročit

Obvykle α **= 0,05** (zamítáme na 5% hladině - **významný výsledek**) nebo α **= 0,01** (zamítáme na hladině 1% - **vysoce významný výsledek**)

Sběr dat

- tato fáze je velmi důležitá a měla by být konzultována se statistikem
- sebraný vzorek dat musí být objektivní, reprezentativní a dostatečně velký

Př. (pokračování): Nasbíraná data – pozorované četnosti ve čtyřpolní tabulce

Výskyt PAP	Antigen HLA-DR4 Ano	Ne	Celkem
Ano	46	28	74
Ne	50	184	234
Celkem	96	212	308

Volba vhodného testu

Rozhodnutí o platnosti nebo neplatnosti hypotézy činíme na základě aplikace vhodného **statistického testu**.

Každý statistický test je charakterizován *testovou statistikou* - funkcí, která ze sesbíraných dat "vytvoří" jedno číslo.

Př.:

$$c^2 = \sum \frac{(\text{pozorovaná četnost - očekávaná četnost})^2}{\text{očekávaná četnost}} \sim c^2(df)$$

Krok 5: Výpočet hodnoty testové statistiky

Sesbíraná data je třeba zpracovat a dosadit do předpisu **testové statistiky**.

Příklad (pokračování)

Výpočet očekávaných hodnot:

Výskyt PAP	Antigen HLA-DR4 Ano	Ne	Celkem
Ano	46	28	74
Ne	50	184	234
Celkem	96	212	308

Výskyt antigenu je rozdělen v poměru 96:212. V případě platnosti hypotézy nezávislosti obou znaků očekáváme, že ve stejném poměru budou rozděleny i skupiny s PAP a bez PAP. Tedy pro skupinu **(PAP-Ano, HLA-DR4-Ano)**:

Očekávaný počet = 96/308 . 74 = 23

Naměřené a očekávané hodnoty

Výskyt PAP	Antigen HLA-DR4 Ano	Ne	Celkem
Ano	46 23	28 51	74
Ne	50 73	184 161	234
Celkem	96	212	308

Červeně jsou vyznačeny četnosti očekávané v případě, že platí hypotéza nezávislosti.

Po dosazení: $c^2 = 43{,}61$

Krok 5: Určení kritické hodnoty

Po dosazení naměřených hodnot do testové statistiky **zamítáme hypotézu**, pokud výsledná hodnota přesáhne jistou mez, nazývanou *kritická hodnota*.

? Jak tuto hodnotu určit?

Kritickou hodnotou testu je takové číslo, které testová statistika překročí v případě, že nulová hypotéza je pravdivá, s pravděpodobností nejvýše α.

$$\mathrm{P}_{H_0}\left(T \geq k_a\right) \leq a$$

Kritické hodnoty jsou tabelovány.

Příklad (dokončení)

Testová statistika: $c^2 = 43{,}61$

Testové kriterium: $c^2 \geq c^2_{1-a}(1) = 3{,}84$

Rozhodnutí:

$$c^2 = 43{,}61 \geq 3{,}84 = c^2_{1-0{,}05}(1)$$

H_0 **zamítáme** na hladině 5%

Zjistili jsme významnou souvislost mezi výskytem antigenu HLA-DR4 a PAP na 5% hladině.

Statistická a klinická významnost

Statistická významnost

Je-li statistický test zamítnut (významný) na předepsané hladině α (**hladina významnosti**).

Klinická významnost

Je-li efekt významný z hlediska *klinické praxe* (např. překročení **prahové hodnoty**).

G

Pojmy statistické a klinické významnosti bývají často ztotožňovány. Toto ztotožnění je však třeba provádět opatrně, neboť bývá nepřesné.

Kontingenční tabulky

- *kontingenční tabulky* slouží ke studování vztahů mezi dvěma znaky

Kontingenční tabulka r x s:

Znak 1	Znak 2: Kategorie 1	...	Kategorie *s*	
Kategorie 1	n_{11}	...	n_{1s}	$n_{1\bullet}$
...	...	...	...	
Kategorie *r*	n_{r1}	...	n_{rs}	$n_{r\bullet}$
	$n_{\bullet 1}$	...	$n_{\bullet s}$	n

- **kontingenční tabulka** typu 2x2 se nazývá *čtyřpolní tabulka*

Test hypotézy o shodnosti struktur

- test shodnosti pravděpodobnostní struktury nějakého znaku za různých podmínek

Př.: Stejná věková struktura pacientů ve dvou nemocnicích.

- tzv. χ^2-test dobré shody použitý na kontingenční tabulku

Příklad

Studie percentuálních zastoupení krevních skupin ve třech krajích severního Skotska.
Je ve všech krajích stejné percentuelní zastoupení krevních skupin?

Oblast	A	B	0	AB	Celkem
Eskdale	33	6	56	5	100
Annadale	54	14	52	5	125
Nithdale	98	35	115	5	253
Celkem	185	55	223	15	478

Příklad (dokončení)

H_0: Pravděpodobnosti skupin jsou v jednotlivých krajích stejné.
H_1: Nulová hypotéza neplatí.

Testová statistika:

$$c^2 = n\sum_{i=1}^{3}\sum_{j=1}^{4}\frac{n_{ij}^2}{n_{i.}n_{.j}} - n \sim c^2((3-1).(4-1)) = c^2(6)$$

Testové kriterium: $c^2 \geq c^2_{1-0,05}(6)$

Rozhodnutí: $c^2 = 10,45 < 12,59 = c^2_{1-0,05}(6)$

H_0 tedy nelze zamítnout na 5% hladině.

Dosažená hladina významnosti

Alternativní postup při rozhodnutí o platnosti či neplatnosti hypotézy:

Určíme pravděpodobnost **p**, s jakou bychom mohli obdržet pozorovaná data nebo data stejně nebo více odporující nulové hypotéze za předpokladu, že je nulová hypotéza **pravdivá**, tato hodnota se nazývá ***dosažená hladina významnosti***.

! $\boxed{\text{Čím menší p, tím méně důvěryhodné je } H_0.}$

Pro účely statistické analýzy volíme hladinu významnosti α a **zamítneme H_0**, je-li:

$$p < a$$

McNemarův test

- Máme náhodný výběr 18 pacientů, kteří byli léčeni dvěma různými antihypertenzivy A, B. Každý pacient dostával po dobu jednoho měsíce léka A a po odeznění jeho případných účinků po dobu jednoho měsíce lék B. Výsledek byl klasifikován jako úspěch nebo neúspěch.

Obecný postup při testování hypotéz

- Formulujeme **nulovou hypotézu** $\mathbf{H_0}$ a **alternativu** $\mathbf{H_1}$.
- Zvolíme **hladinu významnosti α.**
- Získáme **data.**
- Vybereme vhodný **statistický test.**
- Spočteme hodnotu **testového kritéria.**
- Najdeme v tabulkách příslušnou **kritickou hodnotu.**
- Provedeme **statistické rozhodování** následujícím způsobem:

  Je-li hodnota testového kritéria větší než kritická hodnota, zamítneme nulovou hypotézu $\mathbf{H_0}$ ve prospěch alternativy $\mathbf{H_1}$ na hladině významnosti **α.**

Pravidla statistického rozhodování

- **hladina testu α :** pravděpodobnost chyby 1. druhu, tj.. zamítnutí platné nulové hypotézy
- **kritický obor** : výsledky pokusu, při nichž se zamítá nulová hypotéza
- **síla testu** (1-β): pravděpodobnost zamítnutí nulové hypotézy, jestliže nulová hypotéza neplatí
- kritický obor i hladina testu se volí před pokusem, nezávisle na jeho výsledku

Dosažená hladina testu

- Hladinu testu **α** volíme předem (nesmí záviset na datech)
- **Dosažená hladina** (p **value**) je nejmenší hladina, na které bychom při daných datech nulovou hypotézu zamítli
- **Dosažená hladina** (p **value**) je pravděpodobnost našeho výsledku a všech výsledků ještě méně podporujících nulovou hypotézu
- Jednoduché pravidlo:

  $$\textbf{p value} < \alpha \ \boxed{\Longrightarrow} \ \mathbf{H_0} \textbf{ zamítáme}$$

Obecné schéma statistického rozhodování

Rozhodnutí	Skutečnost: H_0 platí	Skutečnost: H_0 neplatí
H_0 zamítnout	chyba 1. druhu a	**správně**
H_0 nezamítnout	**správně**	chyba 2. druhu b

Chyba 1.a 2. druhu (α, β)