Jednovýběrový vážený t-test pro pozorování s různými rozptyly

Michal Kulich

Katedra pravděpodobnosti a matematické statistiky
Matematicko-fysikální fakulta Univerzity Karlovy

Robust 2012, Němčičky

Obsah

Motivace: Projekt ACCEPT

- Skupinově randomizovaná studie na prevenci šíření viru HIV
- Probíhá v Jižní Africe, Tanzanii, Zimbabwe a Thajsku od r. 2004
- Aktivní intervence [CBVCT]: mobilní testování, propagace, osvěta, společné aktivity
- Standardní intervence [SVCT]: testování na klinice
- Cíl: odhadnout a otestovat vliv intervence na incidenci HIV (riziková funkce)

Motivace: Projekt ACCEPT

- 48 geografických oblastí bylo spárováno do 24 párů
- *Randomizace:* v každém páru byla jedné oblasti přiřazena aktivní intervence a druhé standardní intervence
- intervence probíhaly po 3 roky, skončily v r. 2010
- 2010 – 2011 *postintervenční šetření*: náhodný výběr domácností, z každé vybrané domácnosti všichni členové ve věku 18–32 let
- Úspěšnost postintervenčního šetření: $> 85\%$, nezávisí na intervenci
- Součástí pointervenčního šetření byl sběr krevních vzorků a jejich testování na HIV
- Celkem nasbíráno přes 54 000 vzorků krve

Motivace: Projekt ACCEPT

Počet vzorků podle HIV statutu a místa

Centrum	Oblastí	n HIV+	% HIV+
Thajsko	7 párů	79	1.0
Zimbabwe	4 páry	1531	12.9
Tanzánie	5 párů	536	5.9
Vulindlela (SA)	4 páry	3642	30.8
Soweto (SA)	4 páry	1652	14.1

Thajské páry nelze použít pro vyhodnocení vlivu intervence na riziko HIV infekce (málo dat). Zbýbá 17 párů ve 4 centrech.

Motivace: Projekt ACCEPT

Odhad rizika HIV infekce je založen na počtu nedávno nakažených případů zjištěných k tomu určeným laboratorním testem (MAA).

Označme

- $i = 1, \dots, n$ index páru, $n = 17$ počet párů
- N_i^1 a N_i^0 počty HIV-negativních vzorků v oblastech s aktivní a standardní intervencí i-tého páru
- R_i^1 a R_i^0 počty nedávno nakažených vzorků v oblastech s aktivní a standardní intervencí i-tého páru

Motivace: Projekt ACCEPT

Odhady incidence (rizika) v oblastech s aktivní a standardní intervencí i-tého páru jsou

$$\tilde{\lambda}_i^1 = \frac{R_i^1}{(N_i^1 + R_i^1/2)\tau_R} \quad \text{a} \quad \tilde{\lambda}_i^0 = \frac{R_i^0}{(N_i^0 + R_i^0/2)\tau_R},$$

kde $\tau_R = 0.71$ roku je tzv. časové okno testu MAA.

Efekt intervence v i-tém páru odhadneme jako logaritmus podílu pozorovaných rizik, tj.

$$\tilde{\theta}_i = \log \tilde{\lambda}_i^1 - \log \tilde{\lambda}_i^0 = \log \frac{R_i^1(N_i^0 + R_i^0/2)}{R_i^0(N_i^1 + R_i^1/2)}.$$

Tento odhad nezávisí na časovém okně.

Motivace: Projekt ACCEPT

Chceme

1. odhadnout efekt aktivní intervence na riziko nakažení $\theta_0 \equiv \mathrm{E}\,\tilde{\theta}_i = \mathrm{E}\,\log(\tilde{\lambda}_i^1/\tilde{\lambda}_i^0)$ a relativní riziko $\varrho_0 \equiv \exp\{\theta_0\}$;
2. testovat hypotézu $H_0 : \theta_0 = 0$ proti alternativě $H_1 : \theta_0 \neq 0$;
3. zkonstruovat 95%-ní interval spolehlivosti pro ϱ_0.

Nabízí se odhad

$$\hat{\theta} = \frac{1}{n} \sum_{i=1}^{n} \tilde{\theta}_i$$

a párový t-test založený na pozorováních $\tilde{\theta}_1, \dots, \tilde{\theta}_n$.

Motivace: Projekt ACCEPT

- Náhodné veličiny $\tilde{\theta}_1, \ldots, \tilde{\theta}_n$ jsou odhady spočítané z jednotlivých párů. Jsou založeny na výrazně se lišících počtech pozorování. Proto mají různé rozptyly. Jejich rozdělení není normální.
- Přestože odhad $\hat{\theta} = \frac{1}{n} \sum_{i=1}^{n} \tilde{\theta}_i$ zůstává nestranný a konsistentní, a párový t-test dosahuje stanovené hladiny aspoň asymptoticky, dostali bychom přesnější odhad a lepší sílu testu, kdybychom použili *vhodně vážený průměr* jednotlivých párů.

Obsah

Předpoklady a značení

- Nechť $\tilde{\theta}_1, \ldots, \tilde{\theta}_n$ jsou nezávislé náhodné veličiny se stejnou střední hodnotou $\mathrm{E}\,\tilde{\theta}_i = \theta_0$ a obecně různými neznámými rozptyly $\operatorname{var}\tilde{\theta}_i = \sigma_i^2$.
- Nechť $w_1, \ldots, w_n$ je posloupnost nezáporných konstant (alespoň jedna je kladná). Označme

$$a_i = \frac{w_i}{\sum_{j=1}^n w_j} \quad \text{(normalizované váhy)},$$

$$q_i = \frac{w_i^2}{\sum_{j=1}^n w_j^2} = \frac{a_i^2}{\sum_{j=1}^n a_j^2} \quad \text{(normalizované kvadratické váhy)}.$$

- Definujme vážený odhad θ_0

$$\hat{\theta}_W = \frac{\sum_{i=1}^n w_i \tilde{\theta}_i}{\sum_{i=1}^n w_i} = \sum_{i=1}^n a_i \tilde{\theta}_i .$$

Vlastnosti váženého odhadu

Je zřejmé, že

$$\mathrm{E}\,\hat{\theta}_W = \theta_0 \quad \text{a} \quad \operatorname{var}\hat{\theta}_W = \sum_{i=1}^{n} a_i^2 \sigma_i^2 .$$

Potřebujeme odhadnout $\operatorname{var}\hat{\theta}_W$, aniž bychom znali nebo uměli odhadnout jednotlivá σ_i^2. Vyjdeme z rovnosti

$$\mathrm{E}\sum_{i=1}^{n} a_i^2 \tilde{\theta}_i^2 = \sum_{i=1}^{n} a_i^2 (\sigma_i^2 + \theta_0^2) = \operatorname{var}\hat{\theta}_W + \theta_0^2 \sum_{i=1}^{n} a_i^2 .$$

Nahradíme-li θ_0 odhadem $\hat{\theta}_W$, dostaneme odhad $\operatorname{var}\hat{\theta}_W$ ve tvaru

$$S_W^2 = \sum_{i=1}^{n} a_i^2 \tilde{\theta}_i^2 - \hat{\theta}_W^2 \left(\sum_{i=1}^{n} a_i^2 \right).$$

Vážený t-test

Uvažujme test hypotézy $H_0 : \theta_0 = \zeta$ proti alternativě $H_1 : \theta_0 \neq \zeta$.

Testová statistika bude mít tvar

$$T_W = \frac{\hat{\theta}_W - \zeta}{S_W} = \sqrt{\frac{1}{\sum_{i=1}^n a_i^2}} \frac{\hat{\theta}_W - \zeta}{\sqrt{\sum_{i=1}^n q_i \tilde{\theta}_i^2 - \hat{\theta}_W^2}},$$

kde q_i jsou normované kvadratické váhy.

Limitní rozdělení testové statistiky

Tvrzení

Nechť posloupnosti $\frac{1}{n}\sum_{i=1}^{n} w_i$ a $\frac{1}{n}\sum_{i=1}^{n} w_i^2$ konvergují ke kladným a konečným limitám, nechť $\max(a_1, \ldots, a_n) \to 0$ pro $n \to \infty$.

Potom za platnosti hypotézy $H_0 : \theta_0 = \zeta$

$$T_W \xrightarrow{d} \mathrm{N}(0, 1).$$

pro $n \to \infty$.

Aplikace pro malé rozsahy výběru

- Je-li počet pozorování n malý (nebo máme mnoho takřka nulových vah), nemůžeme se spolehnout na asymptotickou normalitu.
- Chtěli bychom analogii k jednovýběrovému t-testu, kde používáme referenční rozdělení t_{n-1}.
- Počet stupňů volnosti musí záviset na vahách.
- Povšimneme-li si tvaru testové statistiky

$$T_W = \sqrt{\frac{1}{\sum_{i=1}^n a_i^2}} \frac{\hat{\theta}_W - \zeta}{\sqrt{\sum_{i=1}^n q_i \tilde{\theta}_i^2 - \hat{\theta}_W^2}},$$

a porovnáme-li ji s klasickým t-testem, vidíme, že roli rozsahu výběru n zde přebírá výraz $\frac{1}{\sum_{i=1}^n a_i^2}$.

Aplikace pro malé rozsahy výběru

- Z důvodu lepší aproximace pro malé rozsahy výběru tedy doporučujeme používat referenční rozdělení t_f, kde počet stupňů volnosti je $f = 1/\sum_{i=1}^{n} a_i^2 - 1$.
- Nastavíme-li váhy rovnoměrně, tedy $a_i = 1/n$, vyjde $f = n - 1$ a dostáváme klasický t-test jako speciální případ.
- Nastavíme-li váhy tak, že $n - m$ vah je nulových a zbylých m vah má hodnoty $1/m$, vyjde $f = m - 1$ a dostáváme klasický t-test se sníženým počtem pozorování.

Oboustranný interval spolehlivosti pro θ_0 s přibližnou pravděpodobností pokrytí $1 - \alpha$ má krajní body

$$\hat{\theta}_W \pm S_W t_f(1 - \alpha/2).$$

Volba vah

- Rozptyl váženého odhadu $\hat{\theta}_W$ je minimální, zvolíme-li $w_i = 1/\sigma_i^2$. Rozptyly jednotlivých pozorování však neznáme.
- Místo toho se můžeme pokusit získat co nejlepší odhady σ_i^2, třeba i za cenu přidání dalších předpokladů o rozdělení $\tilde{\theta}_i$.
- Používáme-li však ve vahách odhady rozptylů, potřebujeme zobecnění teoretických výsledků na náhodné váhy.

Obsah

Předpoklady a definice

- Nechť $W_1, \ldots, W_n$ je posloupnost nezáporných nezávislých a stejně rozdělených náhodných veličin s nenulovou střední hodnotou a konečnými rozptyly. Nechť $\mathrm{E}\left(\tilde{\theta}_i \,\middle|\, W_i\right) = \theta_0$.
- Definujme

$$
\begin{aligned}
A_i &= \frac{W_i}{\sum_{i=1}^n W_i} \quad \text{normalizované váhy,} \\
Q_i &= \frac{W_i^2}{\sum_{i=1}^n W_i^2} \quad \text{normalizované kvadratické váhy,} \\
\hat{\theta}_W &= \frac{\sum_{i=1}^n W_i \tilde{\theta}_i}{\sum_{i=1}^n W_i} = \sum_{i=1}^n A_i \tilde{\theta}_i \quad \text{vážený odhad,} \\
T_W &= \sqrt{\frac{1}{\sum_{i=1}^n A_i^2}} \frac{\hat{\theta}_W - \zeta}{\sqrt{\sum_{i=1}^n Q_i \tilde{\theta}_i^2 - \hat{\theta}_W^2}} \quad \text{testovou statistiku.}
\end{aligned}
$$

Výsledky

Věta

Za stanovených předpokladů platí

1. $\hat{\theta}_W \xrightarrow{p} \theta_0$,
2. $T_W \xrightarrow{d} \mathrm{N}(0,1)$ za platnosti hypotézy $H_0 : \theta_0 = \zeta$.

Volba vah v projektu ACCEPT

- V Projektu ACCEPT máme

$$\tilde{\theta}_i = \log \tilde{\lambda}_i^1 - \log \tilde{\lambda}_i^0 = \log \frac{R_i^1(N_i^0 + R_i^0/2)}{R_i^0(N_i^1 + R_i^1/2)}.$$

- Kdybychom předpokládali, že R_i^0 a R_i^1 mají Poissonovo rozdělení s dostatečně velkými parametry, dostali bychom aproximací Poissonova rozdělení normálním a aplikací delta metody

$$\frac{\tilde{\theta}_i - \theta_0}{\sqrt{\frac{1}{R_i^1} + \frac{1}{R_i^0}}} \dot{\sim} \mathrm{N}(0,1).$$

- V i-tém páru použijeme jako váhu harmonický průměr R_i^0 a R_i^1

$$W_i = \frac{1}{\frac{1}{R_i^1} + \frac{1}{R_i^0}} = \frac{R_i^1 R_i^0}{R_i^1 + R_i^0}.$$

Volba vah v projektu ACCEPT

- Náhodné váhy
$$W_i = \frac{R_i^1 R_i^0}{R_i^1 + R_i^0}$$
jsou za uvedených předpokladů na rozdělení R_i^0 a R_i^1 dobré odhady optimálních vah $1/\sigma_i^2$. Porušení předpokladů způsobí horší přesnost odhadu θ_0 a menší sílu testu, ale asymptotické rozdělení testové statistiky nebude narušeno.
- Dodefinujeme $W_i = 0$ pokud $R_i^0 = 0$ a $R_i^1 = 0$.
- Pokud v některé oblasti nepozorujeme ani jednu nedávnou infekci, daný pár je vyloučen z odhadování efektu intervence.

Obsah

Metodika

Budeme simulovat použití klasického a váženého t-testu v Projektu ACCEPT.

- Počet párů, rozsahy výběrů v oblastech, počty nakažených byly nastaveny podle skutečnosti.
- Laboratorní data byla získána výběrem s vracením z externí databáze 4000 laboratorních vzorků.
- Počty nedávno nakažených byly určeny aplikací laboratorního testu MAA na generovaná laboratorní data.
- Simulace byly opakovány 1000-krát.

Simulace za hypotézy

Metoda	Prům. odh. efekt	Emp. SE pro $\hat{\theta}$	Prům. odh. SE pro $\hat{\theta}$	Hladina	Pokrytí
Vážený t-test	0.982	0.087	0.090	0.050	0.959
Klasický t-test	0.981	0.135	0.141	0.037	0.969

Normalizované váhy:

Zimbabwe	0.04 – 0.06
Tanzánie	0.005 – 0.02
Vulindlela	0.10 – 0.15
Soweto	0.03 – 0.08

Stupně volnosti váženého testu:	10.1
Stupně volnosti klasického testu:	16

Simulace za alternativy

Metoda	Prům. odh. efekt	Emp. SE pro $\hat{\theta}$	Prům. odh. SE pro $\hat{\theta}$	Síla	Pokrytí
Vážený t-test	0.685	0.095	0.091	0.952	0.925
Klasický t-test	0.661	0.150	0.148	0.753	0.954

Závěr:

Projekt ACCEPT bude analyzován váženým t-testem.

Obsah

Závěrečné poznámky

- *Vážený t-test pro data s neznámými nestejnými rozptyly* může být užitečný v řadě případů, kdy data jsou sama výsledky statistické analýzy experimentálních jednotek nebo jde o měření s rozdílnou přesností, o niž máme aspoň hrubou představu.
- Analýza Projektu ACCEPT váženým t-testem byla dokončena, výsledky ještě nejsou veřejné.
- Aproximaci počtu stupňů volnosti lze nejspíš zlepšit (viz Welchův, Satterthwaiteův test).

Závěrečné poznámky

- Šlo by použít bootstrap?
- Šlo by zavést vážený permutační test?
- Co když neplatí $\mathrm{E}\,\tilde{\theta}_i = \theta_0 \;\forall i$?
- Podmínka $\mathrm{E}\left(\tilde{\theta}_i \,\middle|\, W_i\right) = \theta_0$ není triviální. V řadě případů ji nelze beze zbytku splnit.

A to je KONEC...