MATEMATICKO-FYZIKÁLNÍ FAKULTA
Univerzita Karlova

Matěj Lebeda

DVOUVÝBĚROVÝ WELCHŮV *t*-TEST

19. října 2021

Oborový seminář

Obsah

Reálný problém

- Reálný problém: 2 umělá hnojiva A a B, sledujeme produkce obilí po použití obou hnojiv.
- CÍL: otestovat, zda se shodují střední hodnoty produkcí.

Značení

- $F_X \ldots$ distribuční fce rozdělení n. v. X_i z náh. výběru $\mathbf{X} = (X_1, \ldots, X_n)^T$
- $F_Y \ldots$ distribuční fce rozdělení n. v. Y_i z náh. výběru $\mathbf{Y} = (Y_1, \ldots, Y_m)^T$
- $\mathcal{L}_+^2 \ldots$ třída distr. fcí rozdělení s konečným kladným rozptylem
- $\mu_X := \mathrm{E}X_1, \ \mu_Y := \mathrm{E}Y_1$
- $S_{X,Y}^2 := \frac{(n-1)S_X^2 + (m-1)S_Y^2}{n+m-2} \ldots$ vážený průměr výběrových rozptylů
- $u_\alpha \ldots \alpha$-kvantil rozdělení $\mathcal{N}(0,1)$
- $t_n(\alpha) \ldots \alpha$-kvantil t-rozdělení o n stupních volnosti

Asymptotický t-test

- *Model:* $F_X \in \mathcal{L}^2_+,\ F_Y \in \mathcal{L}^2_+$, přičemž $\mathrm{var}(X_1) = \mathrm{var}(Y_1) =: \sigma^2,\ X, Y$ nezávislé.
- $H_0 : \mu_X = \mu_Y, \quad H_1 : \mu_X \neq \mu_Y$
- *Testová statistika:*

$$T_{n,m} = \frac{\overline{X_n} - \overline{Y_m}}{\sqrt{S^2_{X,Y}(\frac{1}{n} + \frac{1}{m})}}$$

- *Rozdělení test. statistiky za H_0:* $T_{n,m} \overset{\text{as.}}{\sim} \mathcal{N}(0,1)$
- *Zamítací pravidlo:* Zamítáme $H_0 \Leftrightarrow |T_{n,m}| \geq u_{1-\alpha/2}$
- *Alternativně* můžeme zamítat tehdy, když $|T_{n,m}| \geq t_{n+m-2}(1-\alpha/2)$
- *p-hodnota:* $p = 2(1 - F(|t|))$

Konstrukce testové statistiky

- Čitatel: $\overline{X_n} - \overline{Y_m} \ldots$ nestranný a konzistentní odhad par. $\mu_X - \mu_Y$.
- Jmenovatel: obvykle odmocnina z odhadu rozptylu čitatele:
- $\mathrm{var}(\overline{X_n} - \overline{Y_m}) \overset{X \perp Y}{=} \mathrm{var}(\overline{X_n}) + \mathrm{var}(\overline{Y_m}) = \frac{\sigma^2}{n} + \frac{\sigma^2}{m} = \sigma^2(\frac{1}{n} + \frac{1}{m})$
- $S^2_{X,Y} = \frac{(n-1)S^2_X + (m-1)S^2_Y}{n+m-2}$ je nestranný odhad σ^2.

Přesný t-test

- *Model:* $F_X \sim \mathcal{N}(\mu_X, \sigma^2)$, $F_Y \sim \mathcal{N}(\mu_Y, \sigma^2)$, X, Y nezávislé.
- $H_0 : \mu_X = \mu_Y, \quad H_1 : \mu_X \neq \mu_Y$
- *Testová statistika:*

$$T_{n,m} = \frac{\overline{X_n} - \overline{Y_m}}{\sqrt{S^2_{X,Y}(\frac{1}{n} + \frac{1}{m})}}$$

- *Rozdělení test. statistiky za H_0:* $T_{n,m} \sim t_{n+m-2}$
- *Zamítací pravidlo:* Zamítáme $\Leftrightarrow |T_{n,m}| \geq t_{n+m-2}(1 - \alpha/2)$.
- *p-hodnota:* $p = 2(1 - F(|t|))$

Welchův *t*-test

- Modifikace standardního *t*-testu pro předpoklad neshodných rozptylů.
- *Model:* $X_1, \ldots, X_n \sim \mathcal{N}(\mu_X, \sigma_1^2),\ Y_1, \ldots, Y_m \sim \mathcal{N}(\mu_Y, \sigma_2^2)$ náhodné výběry nezávislé na sobě.
- *Nulová a alternativní hypotéza:* $H_0 : \mu_X = \mu_Y, \quad H_1 : \mu_X \neq \mu_Y$

Konstrukce testové statistiky

- Statistiku přirozeně založíme na $\overline{X_n} - \overline{Y_m}$.
- Normalita a nezávislost dat $\Rightarrow \overline{X_n} - \overline{Y_m} \overset{H_0}{\sim} \mathcal{N}(0, \frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{m})$, tj. $\frac{\overline{X_n} - \overline{Y_m}}{\sqrt{\frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{m}}} \overset{H_0}{\sim} \mathcal{N}(0, 1)$.
- Nestranný a konzistentní odhad $\frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{m}$ je $S^2 := \frac{S_X^2}{n} + \frac{S_Y^2}{m}$.
- Odtud *testová statistika*:

$$T_{n,m}^* = \frac{\overline{X_n} - \overline{Y_m}}{\sqrt{S^2}}$$

- *Problém:* Za H_0 neznáme přesné rozdělení $T_{n,m}^*$! (z CS věty získáme pouze as.).

Aproximace rozdělení testové statistiky

- $T^*_{n,m} = \frac{\frac{\overline{X_n} - \overline{Y_m}}{\sqrt{\frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{m}}}}{\sqrt{K}}$, kde $K = \frac{S^2}{\frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{m}}$. Pak má čitatel $\mathcal{N}(0,1)$ rozdělení.
- *Idea:* aproximovat rozdělení K rozdělením n. v. $\frac{Z_r}{r}$, kde $Z_r \sim \chi^2_r$.
- r určíme tak, aby $\mathrm{var}(K) = \mathrm{var}(\frac{Z_r}{r})$.
- Označme si $\xi := \frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{m}$.

$$\mathrm{RHS} = \frac{1}{r^2}\mathrm{var}(Z_r) = \frac{1}{r^2} \cdot 2r = \frac{2}{r}$$

$$\mathrm{LHS} = \frac{1}{\xi^2}\left(\frac{1}{n^2} \cdot 2\frac{\sigma_1^4}{n-1} + \frac{1}{m^2} \cdot 2\frac{\sigma_2^4}{m-1}\right)$$

Počet stupňů volnosti

- Porovnáním LHS a RHS dostaneme:

$$r = \frac{\left(\frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{m}\right)^2}{\frac{\sigma_1^4}{n^2(n-1)} + \frac{\sigma_2^4}{m^2(m-1)}}$$

- Tedy dle této aproximace $K \approx \frac{\chi_r^2}{r}$, a tedy $T_{n,m}^* \approx t_r$.
- Vzorec pro r obsahuje neznámé parametry - je třeba ho odhadnout.

Welchův odhad

- Uvažujme následující odhad $r = \frac{\left(\frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{m}\right)^2}{\frac{\sigma_1^4}{n^2(n-1)} + \frac{\sigma_2^4}{m^2(m-1)}}$:

$$\hat{r} = \frac{S^4}{\frac{S_X^4}{n^2(n+1)} + \frac{S_Y^4}{m^2(m+1)}} - 2$$

- Přepíšeme-li jej následovně:

$$\hat{r} = \frac{S^4 - 2\left(\frac{S_X^4}{n^2(n+1)} + \frac{S_Y^4}{m^2(m+1)}\right)}{\frac{S_X^4}{n^2(n+1)} + \frac{S_Y^4}{m^2(m+1)}}, \tag{1}$$

pak čitatel (1) je nestranný odhad čitatele r a jmenovatel (1) je nestranný odhad jmenovatele r.

Příklad

- Uvažujme motivační příklad:
- $X = (X_1, \ldots, X_{13})$ - produkce obilí po použití hnojiva A
- $Y = (Y_1, \ldots, Y_{16})$ - produkce obilí po použití hnojiva B

Hnojivo	Produkce obilí
A	452, 874, 554, 447, 356, 754, 558, 574, 664, 682, 547, 435, 245
B	546, 547, 774, 465, 459, 665, 467, 365, 589, 534, 456, 651, 654, 665, 546, 537

- Výsledky spočtené v R:

t	-0.15135
df	19.169
p	0.8813
$\overline{X}$	549.3846
$\overline{Y}$	557.5000
$\sqrt{S_X^2}$	168.7629
$\sqrt{S_Y^2}$	104.6219

Odhad počtu stupňů volnosti v R

- Satterthwaiteův odhad:

$$\tilde{r} = \frac{\left(\frac{S_X^2}{n} + \frac{S_Y^2}{m}\right)^2}{\frac{S_X^4}{n^2(n-1)} + \frac{S_Y^4}{m^2(m-1)}}$$

K-rozměrný test středních hodnot

- K nezávislých náhodných výběrů:

$$
\begin{aligned}
Y_1 &= (Y_{11}, \ldots, Y_{1n_1}), \\
Y_2 &= (Y_{21}, \ldots, Y_{2n_2}), \\
&\vdots \\
Y_K &= (Y_{K1}, \ldots, Y_{Kn_K}).
\end{aligned}
$$

- *Model:* $Y_{k1} \sim \mathcal{L}_2^+$, $\mathrm{E}\, Y_{k1} =: \mu_k$, $k = 1, \ldots, K$
- *Nulová a alt. hypotéza:* $H_0 : \mu_1 = \ldots = \mu_K$, $H_1 : \exists\, i \neq j : \mu_i \neq \mu_j$

Testová statistika

$$F_\omega = \frac{\sum_{k=1}^{K} \omega_k (\overline{Y}_{k+} - \overline{Y}_\omega)^2}{K-1} \cdot \frac{1}{1 + 2\Lambda(K-2)},$$

kde:

- $\omega_k := \frac{n_k}{S_k^2}$ ... váha k-té skupiny,
- $\overline{Y}_{k+} = \frac{1}{n_k} \sum_{i=1}^{n_k} Y_{ki}$ ... průměr v k-té skupině,
- $\overline{Y}_\omega = \frac{\sum_{k=1}^{K} \omega_k \overline{Y}_{k+}}{\sum_{k=1}^{K} \omega_k}$ ... odhad stř. hodnoty za H_0 a
- $\Lambda = \frac{\sum_{k=1}^{K} \frac{1}{n_k - 1} \left(1 - \frac{\omega_k}{\sum_{j=1}^{K} \omega_j}\right)^2}{K^2 - 1}$ korekce
- *Rozdělení F_ω za H_0:* $(K-1)F_\omega \xrightarrow{d} \chi^2_{K-1}$