Experimental Design and Statistics - AGA46E

M. Maciak (Czech University of Life Sciences, Prague)

Lab Session 7 - Summer Term 2015

1 Dvouvyberove problemy - teoreticke priklady

- Predpokladajme, ze vyska muzu v populaci ma normalne rozdeleni $N(\mu_1, 40)$ a vyska zen v populaci ma normalne rozdeleni $N(\mu_2, 40)$ (predpokladame tedy homoskedasticitny pripad). Na zaklade nahodnehych vyberu $X_1, \ldots, X_{25} \sim N(\mu_1, 40)$ a $Y_1, \ldots, Y_{30} \sim N(\mu_2, 40)$ jsme spocetli vyberove prumery $\overline{X}_{25} = 182\ cm$ a $\overline{Y}_{30} = 176.8\ cm$. Sestrojte konfidencny interval pro rozdil $\mu_1 - \mu_2$ na hladine spolehlivosti $95\ \%$. Na stejne hladine take testujte nulovou hypotezu $H_0 : \mu_1 = \mu_2$ oproti alternative $H_A : \mu_1 \neq \mu_2$.
- Uvazujme stejny pripad s vyskou muzu a zen v populaci, avsak parametr rozptylu $\sigma^2 > 0$ je neznamy (nadale vsak predpokladame homoskedasticitu). Vyberove rozptyly jsou $s^2_{25} = 42$ a $s^2_{30} = 38$. Sestrojte $95\ \%$ konfidencny interval za techto predpokladu a take testujte nulovou hypotezu $H_0 : \mu_1 = \mu_2$ oproti alternative $H_A : \mu_1 \neq \mu_2$.
- Jak se reseni zmeni, jestlize budeme predpokladat heteroskedasticitny pripad, tudiz nestejne rozptyly, avsak vyberove rozptyly s^2_{25} a s^2_{30} budu stejne, jako v predchozim pripadu?
- Otestujte nulovou hypotezu o homoskedasticite v predchozim pripade ($H_0 : \sigma^2_1 = \sigma^2_2$). Uvazujte hladinu spolehlivosti $90\ \%$.

2 Dvouvyberove problemy v Rku

Pouzite datovy soubor passengerData3.RData a nactete data do softwaru R pomoci prikazu `read.csv()`; Vyuzite nektere zakladne popisne statistiky v Rku (summary statistics) aby ste ziskali zakladny prehled o celkovych datech. Pak udelejte nasledujici:]

- Pouzite help v Rku z zjistete, jak funguje funkce (prikaz) `t.test()`; Zjistete, jake dodatecne parametre jsou potrebne a jakym zpusobem se specifikuji jednotlive pripady, ktere pro statisticky dvouvyberovy t-test rozlisujeme.
- Najdete 95% interval spolehlivosti pro stredni ocekavanu vysku muzskych a zenskych pasazeru; Zaroven otestujte nulovu hypotezu, ze mezi ocekavanou vyskou muzu a zen neni zadny rozdil (pouzijte $\alpha = 0.05$ a funkci `t.test()` ktera je v Rku). Jak by vypadal prislusny 95% interval spolehlivosti pro rozdil v ocekavanej vysce muzu a zen?
- Uvazujte prumernu dobu letu a prumernu cekaci dobu spoctenu za obdobi jedneho mesice (12 mesicu dohromady). Lze rict, ze ocekavana doba cekani je kratsi, nez ocekavana doma letu? Pouzite hladinu vyznamnosti $\alpha = 0.05$.
- Uvazujte cekaci doby samostatne pro muzske a zenske pasazery a testujte nulovou hypotezu, ze ocekavana doba cekani je stejna pro muze i pro zeny. Zvolte rozumnou hladinu vyznamnosti $\alpha \in (0, \frac{1}{2})$); Jak vypada prislusny konfidencny interval spolehlivosti pro rozdil v ocekavanej dobe cekani muzu a zen
- Stejne priklady se pokuste spocitat manualne, pomoci vzorecku, ktere byli na prednasce. Pro prislusne kvantilove hodnoty pouzite tabulky prislusnych rozdeleni, nebo statististicky software R.

Strucny prehled prednasky

- pro nahodne vybery $X_1, \ldots, X_n \sim N(\mu_1, \sigma_1^2)$ a $Y_1, \ldots, Y_m \sim N(\mu_2, \sigma_2^2)$ plati:

$$\overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i \qquad \overline{Y}_m = \frac{1}{m}\sum_{i=1}^{m} Y_i$$

$$s_X^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X}_n)^2 \qquad s_Y^2 = \frac{1}{m-1}\sum_{i=1}^{m}(Y_i - \overline{Y}_m)^2$$

- kdyz jsou nahodne vybery zavisle (paired samples), pak nutne plati, ze $n = m$ a zaroven

$$Z_i = X_i - Y_i \sim N(\mu_Z, \sigma_Z^2), \quad \text{for } \mu_Z = \mu_1 - \mu_2 \text{ and } \sigma_Z^2 = \sigma_1^2 + \sigma_2^2.$$

  Prislusny pocet stupnu volnosti je $df = n - 1$.

- kdyz jsou oba nahodne vybery vzajemne nezavisle a s nerovnymi rozptyly $\sigma_1^2 \neq \sigma_2^2$, pak plati

$$\overline{X}_n - \overline{Y}_m \sim N\left(\mu_1 - \mu_2, \frac{\sigma_X^2}{n} + \frac{\sigma_Y^2}{m}\right), \quad \text{so stupnemi volnosti } df = \frac{\left(\frac{\sigma_X^2}{n} + \frac{\sigma_Y^2}{m}\right)^2}{\frac{\sigma_X^4}{n^2(n-1)} + \frac{\sigma_Y^4}{m^2(m-1)}}$$

- kdyz jsou nahodne vybery nezavisle se stejnymi rozptyly $\sigma_1^2 = \sigma_2^2$ (pouze predpoklad, zjiskane odhady rozptylu muzu byt obecne ruzne), pak plati

$$\overline{X}_n - \overline{Y}_m \sim N\left(\mu_1 - \mu_2, \sigma_{XY}^2\left(\frac{1}{n} + \frac{1}{m}\right)\right), \quad \text{for } \sigma_{XY}^2 = \frac{(n-1)\sigma_X^2 + (m-1)\sigma_Y^2}{m+n-2},$$

  co se take nazyva **pooled estimate** s prislusnymi stupnemi volnosti $df = n + m - 2$.