Matematisk Modellering 1

Preben Blæsild

3. forelæsning

To normalfordelte observationsrækker
Ens varians

Modellen for k *normalfordelte observationsrækker* er, at data

$$x_{11}, \dots, x_{1j}, \dots, x_{1n_1}$$

$$x_{i1}, \dots, x_{ij}, \dots, x_{in_i}$$

$$x_{k1}, \dots, x_{kj}, \dots, x_{kn_k}$$

er realisationer af uafhængige, normalfordelte stokastiske variable

$$X_{ij} \sim N(\mu_i, \sigma_i^2), \quad j = 1, \dots, n_i, i = 1, \dots, k.$$

Notation vedrørende modellen

$$X_{ij} \sim N(\mu_i, \sigma_i^2), \quad j = 1, \ldots, n_i, i = 1, \ldots, k.$$

I den $i$te række er:

- antallet af observationer: n_i
- den empiriske middelværdi eller gennemsnittet:

$$\bar{x}_{i\cdot} = \frac{1}{n_i} \sum_{j=1}^{n_i} x_{ij}$$

- den empiriske varians

$$s_{(i)}^2 = \frac{1}{n_i - 1} \sum_{j=1}^{n_i} (x_{ij} - \bar{x}_{i\cdot})^2$$

- antallet af frihedsgrader: $f_{(i)} = n_i - 1$

Bemærk beregningskemaet side 100 i BG.

Example 2.5

37^{20}–39_1

Data

i dette eksempel stammer fra to eksperimenter, hvor man målte fluers reaktion, efter de var blevet udsat for en nervegas. Målingen for den enkelte flue består i den tid - reaktionstiden - der går fra fluen bringes i kontakt med giften og indtil den ikke længere kan stå på benene.

Fagligt problem

Afhænger reaktionstiden af kontakttiden?

Transformation

Fraktildiagrammerne i Figure 2.16 og 2.17 på side 39 viser, at vi er nødt til at logaritmetransformere reaktionstiderne for at opnå normalitet.

Example 2.5

Model og modelkontrol

Lad x_{ij} være den naturlige logaritme til $j$te observation af reaktionstiden i den $i$te gruppe. Vi betragter da modellen

$$M_0 : X_{ij} \sim N(\mu_i, \sigma_i^2), \quad j = 1, \ldots, n_i, i = 1, 2.$$

Idet $i = 1$ svarer til en kontakttid på 30 s og $i = 2$ til 60 s, er $n_1 = 16$ og $n_2 = 15$.

Fraktildiagrammet i Figure 2.17 strider ikke mod denne model og antyder tillige, at kontakttiden ikke har nogen betydning for reaktionstiden.

Example 2.5 69 og 81–82

Estimation i

$$M_0 : X_{ij} \sim N(\mu_i, \sigma_i^2), \quad j = 1, \ldots, n_i, i = 1, 2.$$

$$\begin{aligned}
\mu_1 &\leftarrow \bar{x}_{1\cdot} = 2.8072 \sim\sim N(\mu_1, \sigma_1^2/n_1) \\
\mu_2 &\leftarrow \bar{x}_{2\cdot} = 2.7154 \sim\sim N(\mu_2, \sigma_2^2/n_2) \\
\sigma_1^2 &\leftarrow s_{(1)}^2 = 1.102779 \sim\sim \sigma_1^2 \chi^2(f_{(1)})/f_{(1)} \\
\sigma_2^2 &\leftarrow s_{(2)}^2 = 1.023778 \sim\sim \sigma_2^2 \chi^2(f_{(2)})/f_{(2)}
\end{aligned}$$

De 4 tilsvarende stokastiske variable er uafhængige.

Hypoteser og test

I modellen for to normalfordelte observationsrækker

$$M_0 : X_{ij} \sim N(\mu_i, \sigma_i^2), \quad j = 1, \dots, n_i, i = 1, 2,$$

er det sædvanligvis hypotesen ens middelværdier,

$$H_{0\mu} : \mu_1 = \mu_2,$$

der er af primær interesse.
Testet for $H_{0\mu}$ afhænger af, om det kan antages, at de to rækker har ens varians. Vi tester derfor først hypotesen

$$H_{0\sigma^2} : \sigma_1^2 = \sigma_2^2.$$

Test af $H_{0\sigma^2} : \sigma_1^2 = \sigma_2^2$ $\quad 81^{10}$–81_4, 89_4–90_1

Da $\sigma_1^2 \leftarrow s_{(1)}^2$ og $\sigma_2^2 \leftarrow s_{(2)}^2$, må man under $H_{0\sigma^2}$ forvente, at

$$\frac{s_{(1)}^2}{s_{(2)}^2} \approx 1.$$

Under $H_{0\sigma^2}$ er $s_{(1)}^2 / s_{(2)}^2 \sim\sim F(f_{(1)}, f_{(2)})$-fordelt. Da denne fordeling ikke er symmetrisk, beregnes testsandsynligheden således

$$p_{\text{obs}}(\mathbf{x}) = \begin{cases} 2F_{F(f_{(1)},f_{(2)})}(\frac{s_{(1)}^2}{s_{(2)}^2}), & \text{hvis } \frac{s_{(1)}^2}{s_{(2)}^2} < 1, \\ 2[1 - F_{F(f_{(1)},f_{(2)})}(\frac{s_{(1)}^2}{s_{(2)}^2})], & \text{hvis } \frac{s_{(1)}^2}{s_{(2)}^2} \geq 1, \end{cases}$$

hvor $F_{F(f_{(1)},f_{(2)})}$ er fordelingsfunktionen for $F(f_{(1)}, f_{(2)})$-fordelingen.

Test af $H_{0\sigma^2} : \sigma_1^2 = \sigma_2^2$ og Statistical Tables 89_4–90_1

Statistical Tables har kun p-fraktiler i $F(f_1, f_2)$-fordelingen for $p \geq 0.5$, men testsandsynligheden for testet af $H_{0\sigma^2} : \sigma_1^2 = \sigma_2^2$ kan altid beregnes på følgende måde: Sæt

$s^2_{\text{tæller}} = \max\{s^2_{(1)}, s^2_{(2)}\}$ med frihedsgrader $f_{\text{tæller}}$,

$s^2_{\text{nævner}} = \min\{s^2_{(1)}, s^2_{(2)}\}$ med frihedsgrader $f_{\text{nævner}}$,.

F-teststørrelsen er

$$F = \frac{s^2_{\text{tæller}}}{s^2_{\text{nævner}}}.$$

Testsandsynligheden beregnes som

$$p_{\text{obs}}(\mathbf{x}) = 2\left[1 - F_{F(f_{\text{tæller}}, f_{\text{nævner}})}(F)\right],$$

hvor $F(f_{\text{tæller}}, f_{\text{nævner}})$ betegner F-fordelingen med $f_{\text{tæller}}$ frihedsgrader i tælleren og $f_{\text{nævner}}$ frihedsgrader i nævneren.

Example 2.5: test af $H_{0\sigma^2} : \sigma_1^2 = \sigma_2^2$ 81^{10}–81_4

Example 2.5

Da forholdet mellem variansskønnene

$$F = \frac{s_{(1)}^2}{s_{(2)}^2} = \frac{1.102779}{1.023778} = 1.077,$$

fås

$$p_{\text{obs}}(\mathbf{x}) = 2\left[1 - F_{F(15,14)}(1.077)\right] = 2 \times 0.477 = 0.894,$$

så $H_{0\sigma^2} : \sigma_1^2 = \sigma_2^2$ forkastes ikke.

Faglig konklusion

Data strider ikke mod at antage, at de to observationsrækker har ens varians.

Ens varians

Modellen M_1

Forkastes hypotesen $H_{0\sigma^2} : \sigma_1^2 = \sigma_2^2$ ikke, reduceres modellen M_0 til

$$M_1 : X_{ij} \sim N(\mu_i, \sigma^2), \quad j = 1, \ldots, n_i, i = 1, 2,$$

som har tre ukendte parametre, middelværdierne μ_1 og μ_2 i de to rækker og den fælles varians σ^2.

Estimation i M_1

$$\mu_1 \leftarrow \bar{x}_{1\cdot} \sim\sim N(\mu_1, \sigma^2/n_1),$$
$$\mu_2 \leftarrow \bar{x}_{2\cdot} \sim\sim N(\mu_2, \sigma^2/n_2),$$

$$\sigma^2 \leftarrow s_1^2 = \frac{f_{(1)}s_{(1)}^2 + f_{(2)}s_{(2)}^2}{f_{(1)} + f_{(2)}} = \frac{SSD_{(1)} + SSD_{(2)}}{f_{(1)} + f_{(2)}} = \frac{SSD_1}{f_1} \sim\sim \sigma^2\chi^2(f_1)/f_1,$$

hvor $f_1 = n_1 + n_2 - 2$, antallet af observationer minus antallet af ukendte parametre i middelværdien.

Example 2.5: estimation i M_1 82_6–82_2

Example 2.5

Estimaterne for parametrene i

$$M_1 : X_{ij} \sim N(\mu_i, \sigma^2), \quad j = 1, \ldots, n_i, i = 1, 2,$$

er

$$\begin{aligned} \mu_1 &\leftarrow \bar{x}_{1\cdot} = 2.8072 \sim\sim N(\mu_1, \sigma^2/16), \\ \mu_2 &\leftarrow \bar{x}_{2\cdot} = 2.7154 \sim\sim N(\mu_2, \sigma^2/15), \\ \sigma^2 &\leftarrow s_1^2 = 1.064638 \sim\sim \sigma^2\chi^2(29)/29. \end{aligned}$$

De 3 tilsvarende stokastiske variable er uafhængige.

Test af $H_{0\mu} : \mu_1 = \mu_2$ i M_1

Testet er baseret på en sammenligning af estimaterne for middelværdierne. Da

$$\bar{x}_{1\cdot} \sim\sim N(\mu_1, \frac{\sigma^2}{n_1}) \quad \text{og} \quad \bar{x}_{2\cdot} \sim\sim N(\mu_2, \frac{\sigma^2}{n_2}),$$

og fordelingerne er uafhængige, gælder der under $H_{0\mu}$, at

$$\bar{x}_{1\cdot} - \bar{x}_{2\cdot} \sim\sim N\left(0, \sigma^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)\right).$$

Derfor er

$$\frac{\bar{x}_{1\cdot} - \bar{x}_{2\cdot}}{\sqrt{\sigma^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \sim\sim N(0, 1).$$

Da σ^2 er ukendt, indsættes i stedet skønnet s_1^2.

Test af $H_{0\mu}: \mu_1 = \mu_2$ i M_1 (fortsat)

Dette giver teststørrelsen $t(\mathbf{x})$, hvor

$$t(\mathbf{x}) = \frac{\bar{x}_{1\cdot} - \bar{x}_{2\cdot}}{\sqrt{s_1^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \sim\sim t(f_1).$$

Da numerisk store værdier af $t(\mathbf{x})$ er kritiske, bliver testsandsynligheden for t-testet

$$p_{\mathrm{obs}}(\mathbf{x}) = 2\left[1 - F_{t(f_1)}(|\ t(\mathbf{x})\ |)\right].$$

Konfidensintervaller i M_1

Da den estimerede spredning på $\bar{x}_{1\cdot} - \bar{x}_{2\cdot}$ er

$$\mathrm{StdError}(\bar{x}_{1\cdot} - \bar{x}_{2\cdot}) = \sqrt{s_1^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)},$$

bliver 95 % konfidensintervallet for $\mu_1 - \mu_2$

$$\bar{x}_{1\cdot} - \bar{x}_{2\cdot} \mp t_{0.975}(f_1)\,\mathrm{StdError}(\bar{x}_{1\cdot} - \bar{x}_{2\cdot})$$

95 % konfidensintervallet for σ^2 er

$$\left\{\frac{f_1 s_1^2}{\chi^2_{0.975}(f_1)}, \frac{f_1 s_1^2}{\chi^2_{0.025}(f_1)}\right\}.$$

Example 2.5: test af $H_{0\mu}: \mu_1 = \mu_2$ 83_{17}–83_{11}

Example 2.5

Den observerede værdi af teststørrelsen $t(\mathbf{x})$ er

$$t(\mathbf{x}) = \frac{2.8072 - 2.7154}{\sqrt{1.064638\left(\frac{1}{16} + \frac{1}{15}\right)}} = \frac{0.0918}{0.3708} = 0.2475,$$

så hypotesen $H_{0\mu}: \mu_1 = \mu_2$ forkastes ikke, da

$$p_{\text{obs}}(\mathbf{x}) = 2\left[1 - F_{t(29)}(0.2475)\right] = 0.81.$$

Da $t_{0.975}(29) = 2.0452$, bliver 95 % konfidensintervallet for $\mu_1 - \mu_2$

$$0.0918 \mp 2.0452 \times 0.3708 = [-0.6666, 0.8502].$$

Faglig konklusion

I forsøget er det ikke påvist, at kontakttiderne har nogen indflydelse på reaktionstiderne.

Ens varians og ens middelværdi

Modellen M_2

Forkastes hypotesen $H_{0\mu} : \mu_1 = \mu_2$ ikke ved test i modellen M_1, reduceres M_1 til

$$M_2 : X_{ij} \sim N(\mu, \sigma^2), \quad j = 1, \ldots, n_i, i = 1, 2,$$

dvs alle observationerne har samme fordeling og kan betragtes som én observationsrække.

Estimater i M_2

$$\begin{aligned}
\mu \leftarrow \bar{x}_{..} &= \frac{S.}{n.} \sim\sim N(\mu, \frac{\sigma^2}{n.}), \\
\sigma^2 \leftarrow s^2 &= \frac{1}{n. - 1}\left[USS. - \frac{S.^2}{n.}\right] \sim\sim \sigma^2\chi^2(n. - 1)/(n. - 1),
\end{aligned}$$

hvor $n. = n_1 + n_2$, $S. = S_1 + S_2$ og $USS. = USS_1 + USS_2$.