Matematisk Modellering 1

Preben Blæsild

4. forelæsning

To normalfordelte observationsrækker
Forskellig varians
Likelihood metoden

Example 3.2

Fagligt problem

I Garrison Bay i staten Washington i USA er det en yndet fritidsinteresse at grave muslinger. For at få viden om, hvordan det påvirker populationen af muslingen *Protothaca stamina*, iværksattes en større indsamling af data.

Data

Man inddelte et undersøgelsesområde i strata. Hvert stratum er 100 m langs kysten og 20 ft bredt. I alt valgte man 5 strata: to over og tre under tidevandslinien. I disse strata placeredes stikprøvekvadrater med kantlængde 0.375 m tilfældigt, og alle muslinger af arten *Protothaca stamina* inden for kvadratet blev indsamlet. I Table 3.2 er kun angivet længden i tiendedele millimeter af muslinger i hvert af to strata, ét lige under og ét lige over tidevandslinien.

Example 3.2

Model og modelkontrol

Lad x_{ij} være den $j$te observation 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 stratum 1 og $i = 2$ til stratum 2 , er $n_1 = 17$ og $n_2 = 11$.

Fraktildiagrammet i Figure 3.2 strider ikke mod denne model og antyder tillige, at variansen i de to grupper ikke er ens.

Example 3.2

Estimation i

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

$$\mu_1 \leftarrow \bar{x}_{1\cdot} = 477.4706 \sim\sim N(\mu_1, \sigma_1^2/n_1)$$
$$\mu_2 \leftarrow \bar{x}_{2\cdot} = 416.5455 \sim\sim N(\mu_2, \sigma_2^2/n_2)$$
$$\sigma_1^2 \leftarrow s_{(1)}^2 = 763.1397 \sim\sim \sigma_1^2\chi^2(f_{(1)})/f_{(1)}$$
$$\sigma_2^2 \leftarrow s_{(2)}^2 = 4034.873 \sim\sim \sigma_2^2\chi^2(f_{(2)})/f_{(2)}$$

De 4 tilsvarende stokastiske variable er uafhængige.

Example 3.2: test af $H_{0\sigma^2}: \sigma_1^2 = \sigma_2^2$ $(87_8–88^3)$, $89_4–90_1$

Example 3.2

Idet

$$s^2_{\text{tæller}} = 4034.873 \quad \text{og} \quad f_{\text{tæller}} = 10$$
$$s^2_{\text{nævner}} = 763.1397 \quad \text{og} \quad f_{\text{nævner}} = 16$$

fås

$$F = \frac{4034.873}{763.1397} = 5.29 \sim\sim F(10,16)$$

og

$$p_{\text{obs}} = 2[1 - F_{F(10,16)}(5.29)] = 0.0033,$$

så hypotesen om ens varianser forkastes.

Faglig konklusion

Det kan ikke antages, at variansen for længden af muslingerne i de to strata er ens.

Test af $H_{0\mu}: \mu_1 = \mu_2$ i M_0 88^4–88_1

I
$$M_0 : X_{ij} \sim N(\mu_i, \sigma_i^2), \quad j = 1, \ldots, n_i, i = 1, 2,$$
testes
$$H_{0\mu} : \mu_1 = \mu_2$$
ved hjælp af teststørrelsen
$$t(\mathbf{x}) = \frac{\bar{x}_{1\cdot} - \bar{x}_{2\cdot}}{\sqrt{s_{(1)}^2/n_1 + s_{(2)}^2/n_2}} \sim\approx t(\bar{f}),$$
hvor
$$\bar{f} = \frac{\left(\frac{s_{(1)}^2}{n_1} + \frac{s_{(2)}^2}{n_2}\right)^2}{\frac{\left(\frac{s_{(1)}^2}{n_1}\right)^2}{f_{(1)}} + \frac{\left(\frac{s_{(2)}^2}{n_2}\right)^2}{f_{(2)}}}$$
Testsandsynligheden er
$$p_{\text{obs}}(\mathbf{x}) = 2(1 - F_{t(\bar{f})}(|\ t(\mathbf{x})\ |)).$$

Konfidensinterval for $\mu_1 - \mu_2$ i M_0

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

$$\mathrm{StdError}(\bar{x}_{1\cdot} - \bar{x}_{2\cdot}) = \sqrt{s_{(1)}^2/n_1 + s_{(2)}^2/n_2},$$

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

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

Example 3.2: test af $H_{0\mu} : \mu_1 = \mu_2$

Example 3.2

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

$$t(\mathbf{x}) = \frac{477.4706 - 416.5455}{\sqrt{763.1397/17 + 4034.873/11}} = \frac{60.9251}{20.2903} = 3.003.$$

$\bar{f} = 12.48$ rundes ned til 12, så

$$p_{\text{obs}}(\mathbf{x}) = 2\left[1 - F_{t(12)}(3.003)\right] = 0.01,$$

og hypotesen $H_{0\mu} : \mu_1 = \mu_2$ forkastes.

Da $t_{0.975}(12) = 2.1788$, bliver 95 % konfidensintervallet for $\mu_1 - \mu_2$

$$60.9251 \mp 2.1788 \times 20.2903 = [16.72, 105.13].$$

Faglig konklusion

Det kan ikke antages, at middelværdien af længden af muslingerne er den samme i de to strata.

To normalfordelte observationsrækker

Hovedpunkter: side 94–95

SAS: side 91–93

Beregningerne i modellerne for to normalfordelte observationsrækker laves ved hjælp af PROC TTEST.

Example 2.5 (Continued) side 91–92

Hvis `react` er et SAS-datasæt med de variable `group`, der angiver de to grupper, og `lnreact`, der indeholder log-reaktionstiderne, beregnes de relevante størrelser således ved hjælp af PROC TTEST:

```
PROC TTEST DATA=react;
CLASS group;
VAR lnreact;
RUN;
```

- I `CLASS` sætningen angives den variabel, der inddeler observationerne i de to rækker, her `group`
- I `VAR` sætningen angives den variable med observationerne, her `lnreact`

Example 2.5 (Continued) side 91–92

Udskriften fra programmet har 3 tabeller, `Statistics`, `T-Tests` og `Equality of Variances`, som skal læses i omvendt række følge:

1) I tabellen `Equality of Variances` angives F-testet for $H_{0\sigma^2} : \sigma_1^2 = \sigma_2^2$, frihedsgraderne for F-fordelingen, samt testsandsynligheden $p_{\text{obs}}(\mathbf{x})$
2) I tabellen `T-Tests` ses t-testene for $H_{0\mu} : \mu_1 = \mu_2$. Hvis man har forkastet hypotesen $H_{0\sigma^2}$, skal man bruge sidste linje, hvis ikke, skal man bruge første linje.
3) I tabellen `Statistics` findes en række beregnede størrelser såsom estimat og 95 % konfidensinterval for en middelværdi og estimat og 95 % konfidensinterval for en spredning.

fortsættes på næste slide

3) Tabellen `Statistics` fortsat

Indholdet af tabellen afhænger af, hvad der står i søjlen med `CLASS` variablens navn, her `group`

a) Rækkerne, hvor der står 1 og 2, skal benyttes, hvis $H_{0\sigma^2}$ og/eller $H_{0\mu}$ forkastes. Her står estimaterne $\bar{x}_{1\cdot}$, $\bar{x}_{2\cdot}$, $s_{(1)}$ og $s_{(2)}$ for henholdsvis μ_1, μ_2, σ_1 og σ_2, samt under navnet `Std Error` $s_{(1)}/\sqrt{n_1}$ og $s_{(2)}/\sqrt{n_2}$. Desuden er grænserne for 95 % konfidensintervallerne for μ_1, μ_2, σ_1 og σ_2 angivet.

b) Rækken, hvor der står `Diff (1-2)`, skal benyttes, hvis $H_{0\sigma^2}$ ikke forkastes. Under `Mean` ses $\bar{x}_{1\cdot} - \bar{x}_{2\cdot}$, som er estimat for $\mu_1 - \mu_2$, omgivet af grænserne for det tilsvarende 95 % konfidensinterval. Under `Std Dev` findes s_1, kvadratroden af estimatet s_1^2 for den fælles varians σ^2. Estimatet for standardafvigelsen σ omgives af grænserne for det tilsvarende 95 % konfidensinterval.

Likelihood metoden

Likelihood metoden er en generel metode til at beregne estimater og test på.

Metoden giver estimater, maksimum likelihood estimater, (mle), og test, likelihood ratio test, (lrt), der stort set svarer til de størrelser, vi har betragtet indtil nu.

Undtagelser:

- estimation af variansen
  - ml-estimatet for σ^2 i én normalfordelt observationsrække fås ved at dividere kvadratsumsafvigelsen med antallet af observationer, n
  - det middelværdirette estimat fås ved at dividere med

$$f = \#\text{antal observationer} - \#\text{antal parametre}$$

- Testene for $\sigma^2 = \sigma_0^2$ og $\sigma_1^2 = \sigma_2^2$ er lidt anderledes end lr-testene.

Likelihood metoden: eksempler

En generel omtale af likelihood metoden findes i Chapter 11 i BG.

Her illustreres metoden i følgende modeller:

- én normalfordelt observationsrække med kendt varians
- én normalfordelt observationsrække med ukendt varians

Likelihood metoden: kendt varians, estimation 70_4–72_5

Model

$$X_i \sim N(\mu, \sigma_0^2), \quad i = 1, \ldots, n.$$

Likelihood funktion

$$L(\mu) = \left(\frac{1}{2\pi\sigma_0^2}\right)^{\frac{n}{2}} e^{-\frac{1}{2\sigma_0^2}\sum_{i=1}^{n}(x_i-\mu)^2}$$

Log likelihood funktion

$$l(\mu) = -\frac{n}{2}\ln(2\pi\sigma_0^2) - \frac{1}{2\sigma_0^2}\sum_{i=1}^{n}(x_i - \mu)^2.$$

Likelihood ligningen

$$0 = \frac{\partial l}{\partial \mu}(\mu) = \frac{1}{\sigma_0^2}\sum_{i=1}^{n}(x_i - \mu).$$

Maksimum likelihood estimat

$$\hat{\mu} = \bar{x}. = \frac{1}{n}\sum_{i=1}^{n} x_i,$$

Likelihood metoden: kendt varians, test

Hypotese

$$H_0 : \mu = \mu_0$$

Likelihood ratio test størrelse

$$Q(\mathbf{x}) = \frac{\max\limits_{\mu \,\in\, H_0} L(\mu)}{\max\limits_{\mu \,\in\, \mathbb{R}} L(\mu)} = \frac{L(\mu_0)}{L(\bar{x}.)}.$$

Log likelihood ratio test størrelse

$$\ln Q = l(\mu_0) - l(\bar{x}.) = -\frac{1}{2\sigma_0^2}[\sum_{i=1}^{n}(x_i - \mu_0)^2 - \sum_{i=1}^{n}(x_i - \bar{x}.)^2]$$

$$= -\frac{n(\bar{x}. - \mu_0)^2}{2\sigma_0^2} = -\frac{1}{2}u^2(\mathbf{x}),$$

hvor

$$u(\mathbf{x}) = \frac{\bar{x}. - \mu_0}{\sqrt{\sigma_0^2/n}}.$$

Likelihood metoden: kendt varians, testss. 70_4–72_5

Da de observationer, som er mere kritiske for H_0 end observationen $\mathbf{x}$ er

$$\begin{aligned}\{\mathbf{y} \mid Q(\mathbf{y}) \leq Q(\mathbf{x})\} &= \{\mathbf{y} \mid -2\ln Q(\mathbf{y}) \geq -2\ln Q(\mathbf{x})\} \\ &= \{\mathbf{y} \mid u^2(\mathbf{y}) \geq u^2(\mathbf{x})\} \\ &= \{\mathbf{y} \mid \; \mid u(\mathbf{y}) \mid \geq \mid u(\mathbf{x}) \mid\},\end{aligned}$$

og

$$U(\mathbf{Y}) \sim N(0,1),$$

bliver testsandsynligheden

$$p_{\text{obs}}(\mathbf{x}) = 2(1 - \Phi(\mid u(\mathbf{x}) \mid).$$

Likelihood ratio testet er altså ækvivalent med u-testet.

Likelihood metoden: ukendt varians, estimation 72_4–74_1

Model

$$X_i \sim N(\mu, \sigma^2), \quad i = 1, \ldots, n.$$

Likelihood funktionen er nu en funktion af begge ukendte parametre

$$L(\mu, \sigma^2) = (\frac{1}{2\pi\sigma^2})^{\frac{n}{2}} e^{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2},$$

og log likelihood funktionen er

$$l(\mu, \sigma^2) = -\frac{n}{2}\ln(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \mu)^2.$$

Maksimum likelihood estimatet for (μ, σ^2) er

$$(\hat{\mu}, \hat{\sigma}^2) = (\bar{x}_., \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x}_.)^2).$$

Likelihood metoden: ukendt varians, estimation 72_4–74_1

Bemærk

Da

$$E(\hat{\sigma}^2) = E(\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x}.)^2) = \frac{(n-1)}{n}\sigma^2,$$

bruges den empiriske varians

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x}.)^2 = \frac{n}{n-1}\hat{\sigma}^2$$

som estimat for σ^2.

Likelihood metoden: ukendt varians, test 72_4–74_1

Hypotese

$$H_0 : \mu = \mu_0$$

Likelihood ratio test størrelse

$$Q(x) = \frac{\max\limits_{\sigma^2 \in \mathbb{R}_+} L(\mu_0, \sigma^2)}{\max\limits_{(\mu,\sigma^2)\in\mathbb{R}\times\mathbb{R}_+} L(\mu, \sigma^2)} = \frac{L(\mu_0, \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu_0)^2)}{L(\bar{x}., \hat{\sigma}^2)}$$

$$= \cdots = \left[1 + \frac{t^2(x)}{n-1}\right]^{-\frac{n}{2}},$$

hvor

$$t(\mathbf{x}) = t(x_1, \ldots, x_n) = \frac{\bar{x}. - \mu_0}{\sqrt{s^2/n}}.$$

Likelihood metoden: ukendt varians, test

Små værdier af $Q(\mathbf{x})$ er kritiske for H_0, og da dette er ækvivalent med store værdier af $t^2(\mathbf{x})$, bliver testsandsynligheden

$$p_{\text{obs}}(\mathbf{x}) = 2(1 - F_{t(f)}(|\ t(\mathbf{x})\ |).$$

Likelihood ratio testet er altså ækvivalent med t-testet.