Matematisk Modellering 1

Preben Blæsild

2. forelæsning

Én normalfordelt observationsrække med ukendt varians

χ^2-fordelingen

F-fordelingen

Én normalfordelt observationsrække

I forbindelse med gennemgangen af én normalfordelt observationsrække med ukendt varians ser vi på:

1. Model og modelkontrol
2. Estimater for parametre og deres fordeling
3. Hypoteser, test og testsandsynligheder
4. Konfidensintervaller

Teorien illustreres ved hjælp af data i Example 3.1.

Example 3.1

Fagligt problem

Trykknapper til telefoner sprøjtestøbes af plastic. En kritisk størrelse er diameteren på tasterne, idet tasterne skal passe i en tastaturbund, som også er støbt af plast. Ved produktionen tilstræber man en diameter på 5200 µm.

Data

For at kontrollere produktionen udtager man samme formiddag 40 taster og måler deres diameter i μm. For at lette beregningerne fratrækkes 5160 µm, hvorved observationerne i Table 3.1, side 54 i BG, fremkommer. For disse observationer er den ideelle værdi 40 µm.

Ukendt varians

Model

Som model vil vi anvende *én normalfordelt observationsrække*. Erfaringen viser nemlig, at diametrene af tasterne er normalfordelte. De $n = 40$ observationer

$$x_1, \ldots, x_n$$

antages at være realisationer af uafhængige identisk normalfordelte stokastiske variable

$$X_1, \ldots, X_n$$

med ukendt middelværdi μ og ukendt varians σ^2. Vi ser nu bort fra, at erfaringen med produktionen har vist, at spredningen er 10 µm, dvs vi betragter variansen som ukendt. Vi skriver kort

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

Modellen har to parametre μ og σ^2.

Ukendt varians

Modelkontrol

Modellen M kontrolleres ved hjælp af fraktildiagrammet i Figure 3.1 på side 55. Da punkterne i diagrammet ikke afviger systematisk fra en ret linje, giver det ikke anledning til at tvivle på modellen.

Estimation

Som tidligere estimeres μ ved gennensnittet af observationerne

$$\mu \leftarrow \bar{x}_. = \frac{1}{n}\sum_{i=1}^{n} x_i = 41.65 \sim\sim N(\mu, \frac{\sigma^2}{n})$$

mens σ^2 estimeres ved den empiriske varians, dvs.

$$\sigma^2 \leftarrow s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x}_.)^2 = 53.31 \sim\sim \sigma^2\chi^2(n-1)/(n-1).$$

De tilsvarende stokastiske variable $\bar{X}_.$ og $s^2(\mathbf{X})$ er uafhængige, se BG side 164.

Ukendt varians

Den empiriske varians s^2
er udfald af en stokastiske variabel

$$s^2(\mathbf{X}) = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X}.)^2$$

- med middelværdi

$$\mathrm{E}\, s^2(\mathbf{X}) = \sigma^2$$

- og varians

$$\mathrm{Var}\, s^2(\mathbf{X}) = \frac{2\sigma^4}{(n-1)},$$

- Bemærk, at ved at dividere kvadratsummen med $n-1$ fås en estimator, som har en middelværdi σ^2, der netop den er parameter, der estimeres.

Ukendt varians

Beregninger

Formlerne for $\bar{x}.$ og s^2 er

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

idet S betegner summen af observationerne, og

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

hvor betegnelsen SSD står for Sum of Squares of Deviations. Idet

$$SSD = USS - \frac{S^2}{n},$$

hvor USS er summen af observationernes kvadrat Uncorrected Sum of Squares, fås

$$s^2 = \frac{1}{n-1}(USS - \frac{S^2}{n}).$$

Ukendt varians

Advarsel

I beregningsformlen for den empiriske varians s^2

$$s^2 = \frac{1}{n-1}(USS - \frac{S^2}{n}).$$

er det vigtigt at skelne mellem s^2, den empiriske varians, og S^2, kvadratet på observationernes sum.

I Example 3.1 er $n = 40$, $S = 1666$ og $USS = 71468$, så

$$\bar{x}. = \frac{1666}{40} = 41.65,$$

og

$$s^2 = \frac{1}{39}(71468 - \frac{1666^2}{40}) = 53.31.$$

Ukendt varians 57^8–59_9

Hypotesen $H_0 : \mu = \mu_0$
Hvis variansen var kendt og lig med σ^2 betragtede vi u-teststørrelsen

$$u(\mathbf{X}) = u(X_1, \ldots, X_n) = \frac{\bar{X}. - \mu_0}{\sqrt{\sigma^2/n}}.$$

I denne kender vi ikke σ^2, så derfor benyttes istedet for den empiriske varians s^2, dvs. vi får t-teststørrelsen

$$t(\mathbf{X}) = t(X_1, \ldots, X_n) = \frac{\bar{X}. - \mu_0}{\sqrt{s^2(\mathbf{X})/n}},$$

ved at erstatte den ukendte varians i u-teststørrelsen med et estimat.

Ukendt varians

Testsandsynlighed i t-testet

Som ved u-testet er numerisk store værdier kritiske for H_0. Da

$$t(\mathbf{X}) \sim t(n-1)$$

bliver testsandsynligheden

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

I Example 3.1 fås

$$t(\mathbf{x}) = t(x_1, \ldots, x_n) = \frac{\bar{x}. - \mu_0}{\sqrt{s^2/n}} = \frac{41.65 - 40}{\sqrt{53.31/40}} = 1.429,$$

og

$$p_{\text{obs}}(\mathbf{x}) = 2(1 - F_{t(n-1)}(1.429)) = 0.161.$$

Ukendt varians

Faglig konklusion Da $p_{\text{obs}}(\mathbf{x}) = 0.161 > 0.05$ forkastes hypotesen $H_0 : \mu = 40$ ikke.

Vi har således ikke i den her betragtede kontrol fundet afvigelser fra den tilstræbte ideelle produktion.

Ukendt varians

Konfidensinterval for μ

Hypotesen $H_0 : \mu = \mu_0$ forkastes ikke ved et test på niveau α hvis og kun hvis

$$
\begin{aligned}
& p_{\text{obs}}(\mathbf{x}) = 2(1 - F_{t(n-1)}(|\ t(\mathbf{x})\ |) \geq \alpha \\
\iff\ & F_{t(n-1)}(|\ t(\mathbf{x})\ |) \leq 1 - \alpha/2 \\
\iff\ & |\ t(\mathbf{x})\ | \leq t_{1-\alpha/2}(n-1) \\
\iff\ & \left|\ \frac{\bar{x}. - \mu_0}{\sqrt{s^2/n}}\ \right| \leq t_{1-\alpha/2}(n-1).
\end{aligned}
$$

Altså $H_0 : \mu = \mu_0$ forkastes ikke hvis og kun hvis μ_0 tilhører intervallet med grænser

$$\bar{x}. \mp t_{1-\alpha/2}(n-1)\sqrt{s^2/n}$$

Ukendt varians 63^1–63_{13}

$1-\alpha$ konfidensintervallet for μ
består af de μ_0 for hvilke hypotesen $H_0 : \mu = \mu_0$ forkastes ikke ved et test på niveau α, dvs intervallet med grænser

$$\bar{x}. \mp t_{1-\alpha/2}(n-1)\sqrt{s^2/n}.$$

Specielt har 95 % konfidensintervallet for μ grænser

$$\bar{x}. \mp t_{0.975}(n-1)\sqrt{s^2/n}.$$

For data i Example 3.1 beregnes intervallet på side 63 i BG.

Ukendt varians

Std Error

Nævneren $\sqrt{s^2/n}$ i t-teststørrelsen

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

der er den estimerede spredning på på estimatet $\bar{x}.$ for μ, betegnes StdError , "standard error". Med denne betegnelse bliver t-teststørrelsen

$$t(\mathbf{x}) = \frac{\text{estimat} - \mu_0}{\text{StdError}},$$

og 95 % konfidensintervallet for μ

$$\text{estimat} \mp t_{0.975}(f)\,\text{StdError},$$

hvor f er antallet af frihedsgrader for variansskønnet; formler der kan være nyttige i forbindelse med udskrifter fra statistikpakker.

Ukendt varians

Hypotesen $H_0 : \sigma^2 = \sigma_0^2$
Da $s^2 \to \sigma^2$, må man under hypotesen forvente, at

$$\frac{s^2}{\sigma_0^2} \approx 1 \quad \text{eller} \quad \frac{fs^2}{\sigma_0^2} \approx f,$$

hvor $f = n - 1$. Under H_0 er $fs^2/\sigma_0^2 \sim\sim \chi^2(f)$-fordelt. Da denne fordeling ikke er symmetrisk, beregnes testsandsynligheden således

$$p_{\text{obs}}(\mathbf{x}) = \begin{cases} 2F_{\chi^2(f)}(\frac{fs^2}{\sigma_0^2}), & \text{hvis } \frac{fs^2}{\sigma_0^2} \leq f, \\ 2(1 - F_{\chi^2(f)}(\frac{fs^2}{\sigma_0^2})), & \text{hvis } \frac{fs^2}{\sigma_0^2} \geq f, \end{cases}$$

hvor $F_{\chi^2(f)}$ er fordelingsfunktionen for $\chi^2(f)$-fordelingen.

$\chi^2(f)$-fordelingen: fraktiler og tabel 162_{12}–164_{12}

Hvis

$$F_{\chi^2(f)}$$

er fordelingsfunktionen for χ^2-fordelingen med f frihedsgrader, er p-fraktilen for χ^2-fordelingen

$$\chi^2_p(f) = F^{-1}_{\chi^2(f)}(p), \quad p \in]0,1[.$$

En tabel over p-fraktilerne for χ^2-fordelingen findes på siderne 6–9 i Statistical Tables.

$\chi^2(f)$-fordelingen: tabelopslag 162_{12}–164_{12}

- I rækken med $f = 3$ ses i søjlen 0.95 at

$$F_{\chi^2(3)}(7.81) = 0.95$$

- I rækken med $f = 8$ ses i søjlen 0.60 at

$$\chi^2_{0.60}(8) = 8.35$$

- Beregn $F_{\chi^2(19)}(14.0)$. I rækken med $f = 19$ ses

$$\chi^2_{0.20}(19) = 13.7 \quad \text{og} \quad \chi^2_{0.30}(19) = 15.4$$

  så

$$F_{\chi^2(19)}(14.0) \in (0.20, 0.30).$$

$\chi^2(f)/f$-fordelingen: tabel 162_{12}–164_{12}

Tabellen over fraktiler i $\chi^2(f)/f$-fordelingen på siderne 10–13 i Statistical Tables er opbygget som tabellen over fraktiler i $\chi^2(f)$-fordelingen dog med den forskel, at der er færre værdier af p og flere værdier af f.

Bemærk, at tabellen kan bruges til af finde fraktiler i $\chi^2(f)$-fordelingen for værdier af f, som ikke findes i tabellen på side 6–9, idet

$$\chi^2_p(f) = f \times \chi^2_p(f)/f.$$

For eksempel

$$\begin{aligned}\chi^2_{0.95}(200) =& 200 \times \chi^2_{0.95}(200)/200\\ =& 200 \times 1.1700\\ =& 234.00.\end{aligned}$$

Ukendt varians

Example 3.1 Hypotesen $H_0 : \sigma^2 = 10^2 = 100$

Da

$$\frac{39 \times 53.31}{100} = 20.79 \leq 39,$$

bliver

$$p_{\text{obs}}(\mathbf{x}) = 2F_{\chi^2(39)}(20.79) = 0.01486.$$

Faglig konklusion

Da $p_{\text{obs}}(\mathbf{x}) < 0.05$ forkastes hypotesen H_0.

Variansen (spredningen) på målingerne her er mindre end den varians (spredning) man har erfaring for.

Ukendt varians

$1-\alpha$ konfidensinterval for σ^2

$$\left[\frac{(n-1)s^2}{\chi^2_{1-\alpha/2}(n-1)}\,,\ \frac{(n-1)s^2}{\chi^2_{\alpha/2}(n-1)}\right]$$

I Example 3.1 bliver 95 % konfidensintervallet for σ^2

$$= \left[\frac{39 \times 53.31}{58.1201}\,,\ \frac{39 \times 53.31}{23.6543}\right]$$

$$= [35.77\ ,\ 87.89]\,.$$

og 95 % konfidensintervallet for σ

$$[5.981\ ,\ 9.375]\,.$$

Ukendt varians

Hovedpunkter: side 78–79

SAS: side 75-77
Standardberegninger for én observationsrække kan udføres ved hjælp af PROC MEANS

Example 3.1 side 76

For de to variable `diameter` og `diam_40` i SAS-datasættet `diam` beregner PROC MEANS ved hjælp af

```
PROC MEANS ALPHA=0.05 DATA=diam VARDEF=DF
     N MEAN STDERR CSS USS VAR SUM CLM T PRT;
VAR diameter diam_40;
RUN;
```

- basale størrelser: antal obs (`N`), sum (`SUM`), gennemsnit (`MEAN`), kvadratsum (`USS`), *SSD* (`CSS`), empirisk varians (`VAR`), standard error (`STDERR`)
- størrelser relateret til t-testet: t-teststørrelsen (`T`), testsandsynligheden (`PRT`), grænserne for 95 % konfidens-intervallet for middelværdien (`CLM`)

F-fordelingen: definition

166^{14}–167_7

F-fordelingen

skal vi benytte i forbindelse med fordelingen af forholdet mellem to uafhængige varianser.

Lad Z_1 og Z_2 være to uafhængige stokastiske variable så $Z_i \sim \chi^2(f_i)/f_i$, $i = 1, 2$. Da er den stokastiske variabel

$$F = \frac{Z_1}{Z_2}$$

F-fordelt med (f_1, f_2) frihedsgrader, eller med f_1 frihedsgrader i tælleren og f_2 frihedgrader i nævneren.

Symbolsk er definitionen

$$F(f_1, f_2) = \frac{\chi^2(f_1)/f_1}{\chi^2(f_2)/f_2},$$

hvor tæller og nævner symboliserer uafhængige stokastiske variable.

F-fordelingen: egenskaber og tabel 166^{14}–167_7

Der gælder, at

$$t \sim t(f) \quad \Rightarrow \quad t^2 \sim F(1, f).$$

Det følger direkte af definitionen, at

$$Y \sim F(f_1, f_2) \quad \Rightarrow \quad \frac{1}{Y} \sim F(f_2, f_1),$$

der bevirker følgende relation mellem p-fraktilen $F_p(f_1, f_2)$ for $F(f_1, f_2)$ fordelingen og $(1-p)$-fraktilen for F fordelingen hvor der er byttet om på frihedsgraderne i tæller og nævner:

$$F_p(f_1, f_2) = \frac{1}{F_{1-p}(f_2, f_1)}.$$

Det er derfor tilstrækkeligt at tabellere fraktilerne for F-fordelingen for værdier af $p \geq 0.5$. Siderne 14–49 i Statistical Tables indeholder p-fraktiler for F-fordelingen for forskellige værdier af p.

F-fordelingen: tabelopslag

Det er nemt at slå fraktiler op i tabellen. Øverst til højre på hver side i tabellen angives p i procent. På side 26 i søjlen med $f_1 = 9$ og rækken med $f_2 = 15$ ses, at

$$F_{0.95}(9, 15) = F^{-1}_{F(9,15)}(0.95) = 2.59.$$

Mere problematisk er det at finde værdier af fordelingsfunktionen ved hjælp af tabellen. Skal vi for eksempel finde $F_{F(13,6)}(7.66)$, ser vi, om vi kan finde en side, hvor der i søjlen $f_1 = 13$ og rækken $f_2 = 6$ står 7.66. Det sker på side 35, hvor $p = 0.99$, så

$$F_{F(13,6)}(7.66) = 0.99.$$

F-fordelingen: tabelopslag

Beregn $F_{F(16,10)}(0.1891)$
Da 0.1891 er mindre end 0.50 fraktilen i $F(16,10)$ fordelingen, bruger vi formlen

$$F_{F(f_1,f_2)}(x) = 1 - F_{F(f_2,f_1)}(x^{-1}).$$

Først finder vi, hvor den reciprokke værdi $1/0.1891 = 5.29$ ligger i $F(10,16)$ fordelingen. Opslag i tabellen side 38 og 42 viser, at

$$F_{0.999}(10,16) = 5.81 \geq 5.29 \geq F_{0.995}(10,16) = 4.27,$$

dvs.

$$0.999 \geq F_{F(10,16)}(5.29) \geq 0.995,$$

så

$$0.001 \leq F_{F(16,10)}(0.1891) \leq 0.005.$$