Matematisk Modellering 1

Preben Blæsild

1. forelæsning

Praktiske oplysninger
Én normalfordelt observationsrække med kendt varians
t-fordelingen

Hjemmeside

Kursets hjemmeside findes i Blackboard på adressen

https://bb.au.dk

hvor I kan finde ugesedler og meget andet.

Materiale

1. Preben Blæsild and Jørgen Granfeldt (2003):
   Statistics with Applications in Biology and Geology.
   Chapman & Hall.
   På meddelelser mm. henvises der til bogen som BG.
   Bogen har sin egen hjemmeside

   http://imf.au.dk/biogeostatistics

   hvor der foruden SAS programmer til eksempler og opgaver findes en trykfejlsliste
2. Preben Blæsild: Statistical Tables
   Kan købes i Statbogladen, med kan også downloades fra kursets hjemmeside
3. Lommeregner og/eller PC

Omfang

1. Forlæsninger
   - Tirsdag kl. 9–11 i Aud. E
   - Fredag kl. 10–12 i Aud. E
2. Øvelser
   - 3 timer per uge
     (tidspunkter fremgår af Ugeseddel 1)
3. StatLab
   - 2 timer per uge
     (tidspunkter fremgår af Ugeseddel 1)
4. Obligatorisk program
   - 2 afleveringsopgaver
   - 1 obligatorisk opgave, der involverer brugen af SAS
5. Eksamen
   - 3 timers skriftlig prøve med evaluering efter 7-trinsskalaen

SAS

I kurset benyttes programpakken SAS, som Aarhus Universitet har en site-licens til. Alle studerende frit kan derfor hente og installere SAS på deres egen PC, sålænge SAS kun bruges til studiemæssige formål.

1. SAS findes til Windows og Linux samt til Macintosh vha af virtuel Linux. Version 9.4 af SAS anbefales
2. For at få adgang til SAS, skal du bruge dit AU-selvbetjenings–UserID og –password
3. Vejledninger til download og installation findes på

   `http://math.medarbejdere.au.dk/en/ithelp/sas`

   (Da filerne er store, bør download **ikke** foregå over trådløst netværk)
4. Har du ikke egen PC, kan du bruge IMF's computersystem, hvor SAS er installeret. Adgang til systemet kræver et brugernavn.

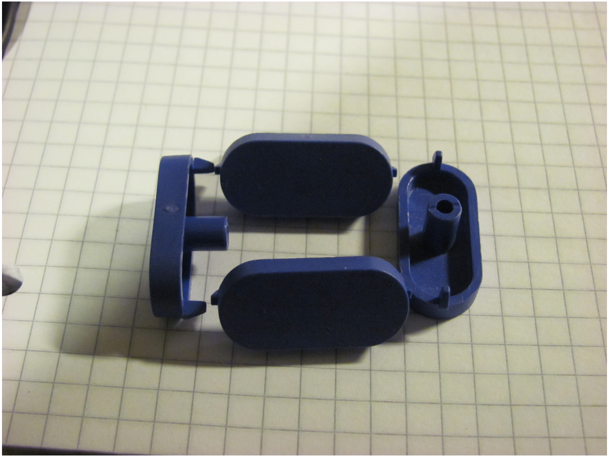
Én normalfordelt observationsrække 53_{18}–79_{1}

I forbindelse med gennemgangen af én normalfordelt observationsrække indføres centrale statistiske begreber såsom:

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

Begreberne introduceres i forbindelse med Example 3.1.

Beocom 1000
(på markedet 1985–2003)



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.

Kendt varians 55_{10}–57^7

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 kendt varians σ_0^2. Erfaringen med produktionen har nemlig vist, at spredningen σ_0 er 10, så det vil vi regne med indledningsvis. Vi skriver kort

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

og omtaler den ukendte middelværdi som en parameter.

Kendt varians 55_{10}–57^7

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

Først ser vi på, hvad man kan sige om middelværdien μ ud fra observationerne. Vi skønner over μ, eller estimerer μ. Traditionelt benytter man gennensnittet af observationerne

$$\bar{x}. = \frac{1}{n}\sum_{i=1}^{n} x_i = 41.65.$$

Det er vigtigt at skelne mellem den teoretiske, men ukendte middelværdi μ og skønnet $\bar{x}.$ for μ. Vi benytter notationen

$$\bar{x}. \to \mu \text{ eller } \mu \leftarrow \bar{x}.,$$

som læses »$\bar{x}.$ estimerer μ« eller »μ estimeres af $\bar{x}.$«.

Her $41.65 \to \mu$.

Kendt varians 55_{10}–57^{7}

Fordeling af estimator

Den stokastiske variabel, som estimatet er et udfald af, omtales som en estimator.

Af formel (3.82) i side 161 i BG fås, at der for estimatoren for middelværdien μ gælder, at

$$\bar{X}. = \frac{1}{n}\sum_{i=1}^{n} X_i \sim N(\mu, \frac{\sigma_0^2}{n}).$$

Bemærk følgende egenskaber ved estimatoren:

1. $\mathrm{E}(\bar{X}.) = \mu$, så middelværdien af estimatoren er netop den parameter, vi estimerer
2. Variansen af estimatoren går mod 0, når n, antallet af observationer vokser, dvs. for store n ligger estimatet $\bar{x}.$ tæt ved den parameter μ, vi estimerer.

Kendt varians 55_{10}–57^{7}

Hypotese

Om produktionen rammer den ideelle værdi 5200 µm, kan i modellen M formuleres som spørgsmålet, om middelværdien μ er lig med 40. Vi betragter derfor nulhypotesen

$$H_0 : \mu = \mu_0 = 40.$$

Test af hypotese

Da $\bar{x}. \to \mu$, forekommer det rimeligt

- at forkaste H_0, hvis $\bar{x}.$ ligger langt fra μ_0
- ikke at forkaste H_0, hvis $\bar{x}.$ ligger tæt på μ_0

Spørgsmålet er altså om differensen $\bar{x}. - \mu_0 = 41.65 - 40 = 1.65$ »*er tæt*« på 0. Svaret må afhænge af variansen på $\bar{X}.$.

Kendt varians 55_{10}–57^{7}

Teststørrelse

Vi beregner derfor teststørrelsen

$$u(\mathbf{x}) = u(x_1, \ldots, x_n) = \frac{\bar{x}. - \mu_0}{\sqrt{\sigma_0^2/n}} = \frac{41.65 - 40}{\sqrt{100/40}} = 1.044,$$

som er differensen normeret med spredningen på gennemsnittet. Teststørrelsen $u(\mathbf{x})$ er en realisation af den stokastiske variabel

$$u(\mathbf{X}) = u(X_1, \ldots, X_n) = \frac{\bar{X}. - \mu_0}{\sqrt{\sigma_0^2/n}} = \frac{\bar{X}. - 40}{\sqrt{100/40}},$$

som ifølge formel (3.82) og (3.83) i BG er $N(0,1)$-fordelt under nulhypotesen.

Kendt varians 55_{10}–57^{7}

Testsandsynlighed

Værdier af teststørrelsen $u(\mathbf{x})$, som ligger tæt på 0, er ikke kritiske for hypotesen H_0. Værdier af u, som numerisk er større end eller lig den observerede værdi 1.044, siges at være mindst lige så kritiske for hypotesen som den observerede.

Man beregner derfor testsandsynligheden $p_{\text{obs}}(\mathbf{x})$, som er sandsynligheden under nulhypotesen for en mere kritisk værdi af teststørrelsen end den observerede. Det vil sige

$$\begin{aligned} p_{\text{obs}}(\mathbf{x}) &= \Phi(-1.044) + (1 - \Phi(1.044)) \\ &= 2(1 - \Phi(1.044)) \\ &= 0.296. \end{aligned}$$

Kendt varians 55_{10}–57^7

Forkastes H_0?

Der er altså en sandsynlighed på ca. 30 % for at få udfald af u, der er mere kritiske end det observerede 1.044.

Hvad nu, hvis vi havde fundet at $u(\mathbf{x}) = 3.39$ og dermed en testsandsynlighed på $p_{\text{obs}}(\mathbf{x}) = 0.06\,\%$?

- Med en testsandsynlighed på 30 % kan hypotesen ikke afvises.
- Med et testsandsynlighed på 0.06 % er det ikke særlig sandsynligt, at H_0 er sand.

Vi forkaster derfor H_0, hvis $p_{\text{obs}}(\mathbf{x}) = 0.06\,\%$, og vi forkaster ikke, hvis $p_{\text{obs}}(\mathbf{x})$ er ca. 30 %.

Hvor går grænsen mellem at forkaste og ikke forkaste?

Kendt varians 55_{10}–57^7

Signifikansniveau

Hvis vi:

- forkaster H_0, hvis $p_{\text{obs}}(\mathbf{x}) < \alpha$

og

- ikke forkaster H_0, hvis $p_{\text{obs}}(\mathbf{x}) \geq \alpha$

kaldes α testets signifikansniveau.

I dette kursus benyttes signifikansniveauet $\alpha = 0.05$ med mindre andet er nævnt.

Kendt varians 55_{10}–57^7

Faglig konklusion

Da $p_{\text{obs}}(\mathbf{x}) = 0.296 > 0.05$ forkastes hypotesen $H_0 : \mu = 40$ ikke.

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

Kendt varians 62_{14}–62_1

Konfidensinterval for μ

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

$$\begin{aligned} & p_{\text{obs}}(\mathbf{x}) = 2(1 - \Phi(|\ u(\mathbf{x})\ |) \geq \alpha \\ \iff & \Phi(|\ u(\mathbf{x})\ |) \leq 1 - \alpha/2 \\ \iff & |\ u(\mathbf{x})\ | \leq u_{1-\alpha/2} \\ \iff & \left|\ \frac{\bar{x}. - \mu_0}{\sqrt{\sigma_0^2/n}}\ \right| \leq u_{1-\alpha/2}. \end{aligned}$$

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

$$\bar{x}. \mp u_{1-\alpha/2}\sqrt{\sigma_0^2/n}$$

Intervallet omtales som $(1 - \alpha)$ konfidensintervallet for μ.

Kendt varians

Hovedpunkter side 78 i BG.

t-fordelingen: definition 164_{11}–166^{13}

t-fordelingen, som vi første gang skal bruge i forbindelse med én observationsrække med ukendt varians, introduceres side 164 i BG: Hvis U og Z er to uafhængige stokastiske variable således at $U \sim N(0,1)$ og $Z \sim \chi^2(f)/f$, er størrelsen

$$t = \frac{U}{\sqrt{Z}}$$

t-fordelt med f frihedsgrader og vi skriver $t \sim t(f)$. Symbolsk kan definitionen af t-fordelingen gengives som

$$t(f) = \frac{N(0,1)}{\sqrt{\chi^2(f)/f}},$$

hvis vi husker på at nævner og tæller symboliserer *uafhængige* stokastiske variable.

Fordelingen kaldes undertiden Student fordelingen eller Student's t-fordeling.

t-fordelingen: egenskaber 164_{11}–166^{13}

Tætheden, som er angivet og illustreret på side 165 i BG,

- er symmetrisk omkring 0
- ligner tæthedsfunktionen for normalfordelingen men er »fladere« omkring 0 og har »tungere haler«
- konvergerer mod tæthedsfunktionen for $N(0, 1)$-fordelingen for $f \to \infty$.

t-fordelingen: fraktiler og tabel 164_{11}–166^{13}

Lad $F_{t(f)}$ betegne fordelingsfunktionen for $t(f)$-fordelingen, og lad $t_p(f) = F^{-1}_{t(f)}(p)$ betegne p-fraktilen for $t(f)$-fordelingen. Da fordelingen er symmetrisk gælder der, at

$$F_{t(f)}(-x) = 1 - F_{t(f)}(x), \quad x \in \mathbb{R},$$

og

$$t_{1-p}(f) = -t_p(f), \quad p \in \left]0, 1\right[.$$

En tabel over p-fraktilerne for $t(f)$-fordelingen for $p \geq 0.5$. findes på side 5 i Statistical Tables.

t-fordelingen: tabelopslag 164_{11}–166^{13}

- I rækken med $f = 6$ ses

$$F_{t(6)}(1.440) = 0.90 \text{ så } F_{t(6)}(-1.440) = 0.10.$$

- I rækken med $f = 17$ ses

$$t_{0.975}(17) = 2.110 \text{ så } t_{0.025}(17) = -2.110.$$

- Beregn $F_{t(39)}(1.429)$. $t(39)$ ikke tabellagt. Bruges $t(40)$ ses,

$$t_{0.90}(40) = 1.303 \text{ og } t_{0.95}(40) = 1.684,$$

$$\text{så } F_{t(39)}(1.429) \in [0.90, 0.95] \,.$$

- Sidste række i tabellen giver

$$P(|t(10)| \geq 1.372) = 0.20.$$