NTNU

Kunnskap for ei betre verd

Institutt for matematiske fag

Eksamensoppgåve i **Løsningsskisse TMA4245 Statistikk**

Fagleg kontakt under eksamen: Gunnar Taraldsen[a], Torstein Fjeldstad[b]

Tlf: [a]464 32 506, [b]962 09 710

Eksamensdato: 23. mai 2018

Eksamenstid (frå–til): 09:00–13:00

Hjelpemiddelkode/Tillatne hjelpemiddel: B: Alle trykte og håndskrevne hjelpemidler tillatt. Bestemt, enkel kalkulator tillatt.

Annan informasjon:

Målform/språk: nynorsk

Sidetal: 9

Sidetal vedlegg: 0

Kontrollert av:

Informasjon om trykking av eksamensoppgåve	
Originalen er:	
1-sidig ☐	**2-sidig** ☒
svart/kvit ☒	**fargar** ☐
skal ha fleirvalskjema ☐	

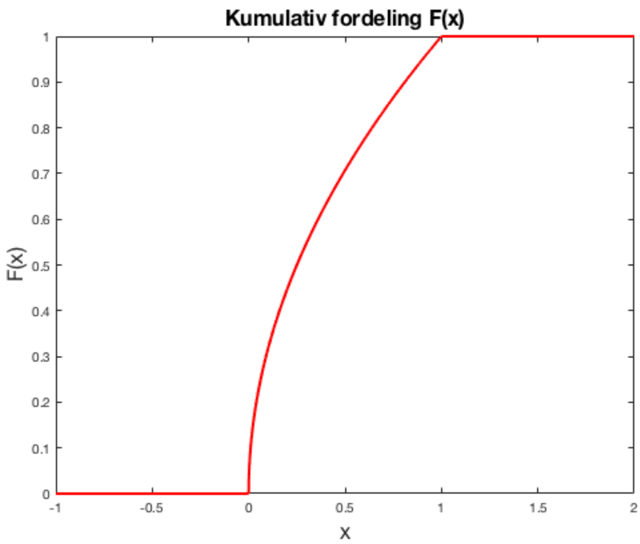
Dato Sign

Oppgåve 1

a) Me finn den kumulative fordelinga til X ved å integrere sannsynstettleiken til X:

$$
\begin{aligned}
F(x) &= \int_{-\infty}^{x} f(x)\mathrm{d}x \\
&= \begin{cases} \int_{-\infty}^{x} 0\mathrm{d}x & = 0 \qquad x \le 0 \\ \int_{-\infty}^{0} 0\mathrm{d}x + \int_0^x \frac{1}{2\sqrt{x}}\mathrm{d}x & = \sqrt{x} \qquad 0 < x < 1 \\ \int_{-\infty}^{0} 0\mathrm{d}x + \int_0^1 \frac{1}{2\sqrt{x}}\mathrm{d}x + \int_1^x 0\mathrm{d}x & = 1 \qquad x \ge 1 \end{cases} .
\end{aligned}
$$

Den kumulative fordelinga er gjeve i Figur 1.



Figur 1: Kumulativ fordeling $F(x)$.

Ved å nytte komplementærsetninga har me

$$
\begin{aligned}
P(X \ge 0.5) &= 1 - P(X \le 0.5) \\
&= 1 - \sqrt{0.5} \\
&\approx 0.293
\end{aligned} .
$$

Frå definisjonen på betinga sannsyn har me

$$\begin{aligned} P(X \leq 0.7 | X \geq 0.5) &= \frac{P(0.5 \leq X \leq 0.7)}{P(X \geq 0.5)} \\ &= \frac{P(X \leq 0.7) - P(X \leq 0.5)}{P(X \geq 0.5)}. \\ &= \frac{\sqrt{0.7} - \sqrt{0.5}}{1 - \sqrt{0.5}} \\ &\approx 0.442 \end{aligned}$$

b) Me har

$$\begin{aligned} P(Y \leq y) &= P(-\ln X \leq y) \\ &= P(\ln X \geq -y) \\ &= P(X \geq e^{-y}) \\ &= 1 - P(X \leq e^{-y}) \\ &= 1 - \sqrt{e^{-y}} \\ &= 1 - e^{-y/2} \end{aligned}$$

for $y > 0$. Me kjenner att dette som den kumulative fordelinga til ein eksponentialfordelt variabel, det vil seie at Y er eksponentialfordelt med sannsynstettleik

$$g(y) = \frac{1}{2} e^{-y/2} \qquad y > 0$$

og forventningsverdi 2.

Alternativt kan ein nytte transformasjonsformelen med $y = u(x) = -\ln(x)$ og $x = w(y) = e^{-y}$:

$$\begin{aligned} g(y) &= \frac{1}{2\sqrt{e^{-y}}} \cdot \left| -e^{-y} \right| \\ &= \frac{1}{2} e^{y/2} e^{-y} \\ &= \frac{1}{2} e^{-y/2} \end{aligned}$$

for $y > 0$.

Anta at me har trekt $Y_1, Y_2, \ldots, Y_n$ frå sannsynstettleiken til Y. La $\bar{Y}_n = \frac{1}{n} \sum_{i=1}^{n} Y_i$ vere gjennomsnittsverdien til Y. Når $n \to \infty$ vil $\bar{Y}_n = \frac{1}{n} \sum_{i=1}^{n} Y_i \to \mathrm{E}(Y)$. Me kan altså tenke på $\mathrm{E}(Y)$ som gjennomsnittsverdien til Y når me gjentar forsøket uendeleg mange gonger.

Oppgåve 2

a) Me definerer følgande hendingar frå oppgåva:

K : ein tilfeldig tilsett er ei kvinne; $P(K) = 0.67$
M : ein tilfeldig tilsett er ein mann; $P(M) = 0.33$
N : ein tilfeldig tilsett har lasta ned minst ein film

Frå oppgåveteksten veit me at $P(N|K) = 0.17$ og $P(N|M) = 0.20$. Frå lova om totalt sannsyn får me:

$$\begin{aligned} P(N) &= P(K)P(N|K) + P(M)P(N|M) \\ &= 0.67 \cdot 0.17 + 0.33 \cdot 0.20 \\ &\approx 0.18 \end{aligned}.$$

Me nyttar Bayes sitt teorem og får

$$\begin{aligned} P(K|N) &= \frac{P(K)P(N|K)}{P(K)P(N|K) + P(M)P(N|M)} \\ &= \frac{0.67 \cdot 0.17}{0.67 \cdot 0.17 + 0.33 \cdot 0.20} \\ &\approx 0.633 \end{aligned}.$$

b) Me nyttar tabell for kumulative sannsyn for poissonfordelinga (for $\mu = 18$) i formelsamlinga. I Tabell 1 (utdrag frå formelsamlinga) ser me at det største heiltalet som oppfyller $P(X \leq c|H_0 : \mu = 18) \leq 0.10$ er $c = 12$. Det vil seie at den kritiske verdien er $c = 12$.

c	10	11	12	13	14
$P(X \leq c)$	0.0304	0.0549	0.0917	0.1426	0.2018

Tabell 1: Kumulative sannsyn $P(X \leq c)$ for poissonfordelinga med forventning 18.

Me forkastar nullhypotesen dersom $X \leq 12$.

Sidan me har observert $x = 13$ vil me ikkje forkaste nullhypotesen.

c) Sannsynet for type-II feil er definert som

$$\begin{aligned} P\left(\text{ikkje forkast } H_0 \text{ når } H_1 \text{ er sann}\right) &= P(X > c|H_1 : \mu = \mu_1) \\ &= 1 - P(X \leq c|H_1 : \mu = \mu_1) \\ &= 1 - \sum_{x=0}^{c} \frac{\mu_1^x}{x!} \exp\left(-\mu_1\right) \end{aligned}$$

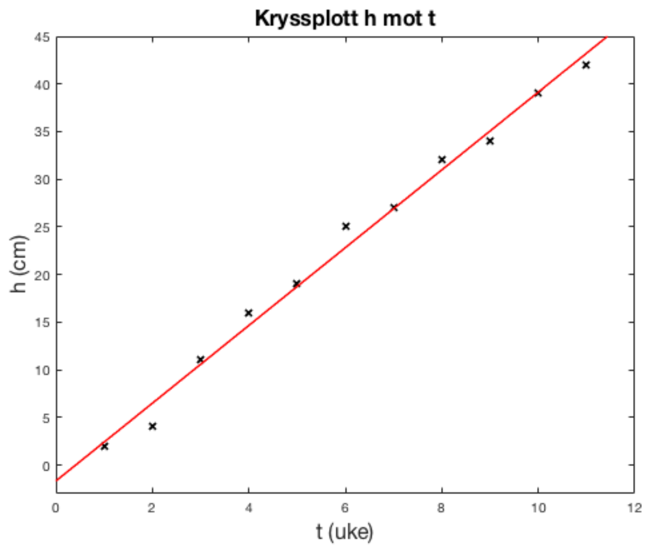
Innsatt $\mu_1 = 14$ får me frå tabell

$$\begin{aligned} P(X > 12 | \mu_1 = 14) &= 1 - \sum_{x=0}^{12} \frac{14^x}{x!} \exp(-14) \\ &= 1 - 0.3585 \\ &= 0.6415 \end{aligned}$$

det vil seie at sannsynet for å gjere ein type-II feil er høgt dersom den sanne verdien til μ er 14.

Oppgåve 3

a) Eit kryssplott av h mot t er gjeve i Figur 2. Me ser at det er ein lineær samanheng mellom h og t; det vil seie at antakinga om $\mathrm{E}(H_i|t_i) = a + b(t_i - 6)$ er rimeleg. Det er heller ingen openbar trend i variansen, det vil seie at antakinga om konstant varians σ^2 verkar rimeleg. Me kan ikkje seie noko direkte om antakinga om uavhengige H_i-ar frå kryssplottet direkte.



Figur 2: Kryssplott h (cm) mot t (uke) samt den tilpassa lineære modellen funne i oppgåve **b)**.

b) Me vil minimere

$$\mathrm{SSE} = \sum_{i=1}^{11} (h_i - a - b(t_i - 6))^2$$

med omsyn på a og b. Me deriverer med omsyn på a og b, set likningssystema lik null og løyser med omsyn på dei ukjende parametra. Me deriverer først med omsyn på a:

$$\begin{aligned}
\frac{\partial \mathrm{SSE}}{\partial a} &= -2\sum_{i=1}^{11}\left(h_i - a - b(t_i-6)\right) \\
a &= \frac{1}{11}\sum_{i=1}^{11}\left(h_i - b(t_i-6)\right) \qquad . \\
a &= \frac{1}{11}\sum_{i=1}^{11}h_i - \frac{b}{n}\sum_{i=1}^{11}(t_i-6)
\end{aligned}$$

Sidan me har fiksert t_i slik at $\sum_{i=1}^{11}(t_i-6)=0$ får me etter å ha satt $\frac{\partial \mathrm{SSE}}{\partial a}=0$:

$$a = \frac{1}{11}\sum_{i=1}^{11}h_i.$$

Dersom me set uttrykket for a inn i uttrykket for SSE før me deriverer med omsyn på b får me

$$\begin{aligned}
\frac{\partial \mathrm{SSE}}{\partial b} &= \frac{\partial}{\partial b}\sum_{i=1}^{11}\left(h_i - a - b(t_i-6)\right)^2 \\
&= \frac{\partial}{\partial b}\sum_{i=1}^{11}\left(h_i - \frac{1}{11}\sum_{j=1}^{11}h_j - b(t_i-6)\right)^2 \\
&= -2\sum_{i=1}^{11}(t_i-6)\left(h_i - \frac{1}{11}\sum_{j=1}^{11}h_j - b(t_i-6)\right) \qquad . \\
&= -2\sum_{i=1}^{11}(t_i-6)\left(h_i - \frac{1}{11}\sum_{j=1}^{11}h_j\right) + 2b\sum_{i=1}^{11}(t_i-6)^2
\end{aligned}$$

Me set uttrykket lik null og løyser med omsyn på b

$$\begin{aligned}
b &= \frac{\sum_{i=1}^{11}(t_i-6)\left(h_i - \frac{1}{11}\sum_{j=1}^{11}h_j\right)}{\sum_{i=1}^{11}(t_i-6)^2} \\
&= \frac{\sum_{i=1}^{11}(t_i-6)\,h_i}{\sum_{i=1}^{11}(t_i-6)^2}
\end{aligned}$$

der den siste overgangen kjem frå

$$\begin{aligned}
\sum_{i=1}^{11}(t_i-6)\frac{1}{11}\sum_{j=1}^{11}h_j &= \left(\sum_{j=1}^{11}h_j\right)\left(\sum_{i=1}^{11}(t_i-6)\right). \\
&= 0
\end{aligned}$$

Me får difor at

$$\begin{aligned}\hat{b} &= \frac{\sum_{i=1}^{11}(t_i-6)\,H_i}{\sum_{i=1}^{11}(t_i-6)^2} \\ \hat{a} &= \frac{1}{11}\sum_{i=1}^{11} H_i\end{aligned}.$$

Ved å setje inn dei oppgjevne tala får me

$$\begin{aligned}\hat{b} &= \frac{449}{110} &\approx\ 4.082 \\ \hat{a} &= \frac{251}{11} &\approx 22.818\end{aligned}.$$

Den tilpassa linja er skissert inn i Figur 2. Som diskutert i oppgåve **a)** ser me at ein lineær modell verker rimeleg då det er ein lineær samanheng mellom h og t. Me ser og at observasjonane verker å vere tilfeldig spreidd rundt den tilpassa linja, noko som styrker antakinga om normalfordelte feilledd.

c) Rimeligheitsfunksjonen er gjeve som

$$\begin{aligned}L\left(a,b,\sigma^2\right) &= \prod_{i=1}^{11}\frac{1}{\sqrt{2\pi}}\cdot\frac{1}{\sqrt{\sigma^2}}\cdot\exp\left(-\frac{1}{2\sigma^2}\left(h_i-(a+b(t_i-6))\right)^2\right) \\ &= \left(\frac{1}{2\pi}\right)^{11/2}\cdot\left(\frac{1}{\sigma^2}\right)^{11/2}\cdot\exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{11}\left(h_i-(a+b(t_i-6))\right)^2\right)\end{aligned}.$$

For å finne sannsynsmaksimeringsestimatorane for a, b og σ^2 må me derivere med omsyn på desse, setje uttrykka lik null og løyse likningssystemet.

Log-rimelighetsfunksjonen er gjeve som

$$l\left(a,b,\sigma^2\right) = -\frac{11}{2}\ln(2\pi) - \frac{11}{2}\ln\left(\sigma^2\right) - \frac{1}{2\sigma^2}\sum_{i=1}^{11}\left(h_i-(a+b(t_i-6))\right)^2.$$

Me maksimerer log-rimeligheitsfunksjonen ved å derivere med omsyn på a og b

$$\begin{aligned}\frac{\partial l(a,b,\sigma^2)}{\partial a} &= \frac{1}{\sigma^2}\sum_{i=1}^{11}\left(h_i-(a+b(t_i-6))\right) &= 0 \\ \frac{\partial l(a,b,\sigma^2)}{\partial b} &= \frac{1}{\sigma^2}\sum_{i=1}^{11}(t_i-6)\left(h_i-(a+b(t_i-6))\right) &= 0\end{aligned}$$

som svarar til uttrykka for minste kvadraters metode gjeve i oppgåve **b)**.

d) Me nyttar vanlege reknereglar for forventningsverdi

$$
\begin{aligned}
\mathrm{E}\left(\hat{b}\right) &= \mathrm{E}\left(\frac{\sum_{i=1}^{11}(t_i-6)\,H_i}{\sum_{i=1}^{11}(t_i-6)^2}\right) \\
&= \frac{\sum_{i=1}^{11}(t_i-6)\,\mathrm{E}\left(H_i\right)}{\sum_{i=1}^{11}(t_i-6)^2} \\
&= \frac{\sum_{i=1}^{11}(t_i-6)\left(a+b(t_i-6)\right)}{\sum_{i=1}^{11}(t_i-6)^2} \\
&= \frac{a\sum_{i=1}^{11}(t_i-6)+b\sum_{i=1}^{11}(t_i-6)^2}{\sum_{i=1}^{11}(t_i-6)^2} \\
&= b
\end{aligned}
$$

Nyttar kjende reknereglar for varians (hugs at $H_1, H_2, \ldots, H_{11}$ er uavhengige)

$$
\begin{aligned}
\mathrm{Var}\left(\hat{b}\right) &= \mathrm{Var}\left(\frac{\sum_{i=1}^{11}(t_i-6)\,H_i}{\sum_{i=1}^{11}(t_i-6)^2}\right) \\
&= \frac{\sum_{i=1}^{11}(t_i-6)^2\,\mathrm{Var}\left(H_i\right)}{\left(\sum_{i=1}^{11}(t_i-6)^2\right)^2} \\
&= \frac{\sigma^2\sum_{i=1}^{11}(t_i-6)^2}{\left(\sum_{i=1}^{11}(t_i-6)^2\right)^2} \\
&= \frac{\sigma^2}{\sum_{i=1}^{11}(t_i-6)^2}
\end{aligned}
$$

Sidan $\hat{b}$ er ein lineærkombinasjon av uavhengige, normalfordelte stokastiske variblar er $\hat{b}$ normalfordelt.

Me tar utgongspunkt i

$$
\begin{aligned}
Z &= \frac{\hat{b}-\mathrm{E}\left(\hat{b}\right)}{\sqrt{\mathrm{Var}\left(\hat{b}\right)}} \\
&= \frac{\hat{b}-\mathrm{E}\left(\hat{b}\right)}{\sqrt{\frac{\sigma^2}{\sum_{i=1}^{11}(t_i-6)^2}}} \quad \sim n(z;0,1)
\end{aligned}
$$

Sidan σ^2 er ukjend nyttar me ein estimator for σ^2, S^2. Me får då

$$
\begin{aligned}
T &= \frac{\hat{b} - \mathrm{E}\left(\hat{b}\right)}{\sqrt{\frac{S^2}{\sum_{i=1}^{11}(t_i-6)^2}}} \\
&= \frac{\frac{\hat{b}-b}{\sqrt{\frac{\sigma^2}{\sum_{i=1}^{11}(t_i-6)^2}}}}{\sqrt{\frac{\frac{(11-2)\cdot S^2}{\sigma^2}}{11-2}}} \\
&= \frac{Z}{\sqrt{\frac{V}{9}}}
\end{aligned}
$$

der Z er standard normalfordelt og V er kjikvadratfordelt med 9 fridomsgrader og Z og V er uavhengige. Difor er T student-t fordelt med 9 fridomsgrader.

Me får då

$$
\begin{aligned}
0.95 &= P\left(-t_{0.025,9} \leq T \leq t_{0.025,9}\right) \\
&= P\left(-t_{0.025,9} \leq \frac{\hat{b} - \mathrm{E}\left(\hat{b}\right)}{\sqrt{\frac{S^2}{\sum_{i=1}^{11}(t_i-6)^2}}} \leq t_{0.025,9}\right) \\
&= P\left(\hat{b} - t_{0.025,9}\sqrt{\frac{S^2}{\sum_{i=1}^{11}(t_i-6)^2}} \leq b \leq \hat{b} + t_{0.025,9}\sqrt{\frac{S^2}{\sum_{i=1}^{11}(t_i-6)^2}}\right)
\end{aligned}
$$

Det vil seie at eit 95 % konfidensintervall for veksthastigheten er

$$
\left[\hat{b} - t_{0.025,9}\sqrt{\frac{S^2}{\sum_{i=1}^{11}(t_i-6)^2}}, \hat{b} + t_{0.025,9}\sqrt{\frac{S^2}{\sum_{i=1}^{11}(t_i-6)^2}}\right].
$$

Innsatt tal får me

$$
[3.786, 4.378].
$$

Anta at me gjer eit forsøk basert på dei stokastiske variablane $X_1, X_2, \ldots, X_n$. Me er interessert i sannsynet

$$
P\left(\widehat{\theta}_L(X_1, X_2, \ldots, X_n) \leq \theta \leq \widehat{\theta}_U(X_1, X_2, \ldots, X_n)\right) = 1 - \alpha.
$$

For eit spesifikt utfall $x_1, x_2, \ldots, x_n$ vil hendinga at den sanne (ukjende) verdien til θ er innafor intervallet anten vere oppfyllt eller ikkje. Dersom me gjentar dette forsøket uendeleg mange ganger vil me få ny verdiar for

$x_1, x_2, \ldots, x_n$ i kvart forsøk (merk at verdien til θ er lik i alle forsøka) som resulterer i eit nytt konfidensintervall i kvart forsøk der den sanne verdien til θ anten er innanfor eller utanfor intervallet. Dersom me gjentar forsøket uendeleg mange gonger vil ein andel $1-\alpha$ dekke den sanne verdien til θ.

Oppgåve 4 Sidan kvar person kan svare høgst ein gong har me eit forsøk *utan* tilbakeleggjing.

- Me trekk eit utval av $n = 20$ personer utan tilbakeleggjing frå ein populasjon av storleik $N = 100$.
- Kvar av dei $N = 100$ personane har to alternativ: "for" eller "imot" der det er anteke at k personar er "for" og $N - k$ personar er "imot".

Difor er X hypergeometrisk fordelt.

Hypotesen i oppgåva kan formulerast som

$$H_0 : N - k = 100 - k = 50 \qquad \text{mot} \qquad H_1 : N - k = 100 - k > 50$$

alternativt

$$H_0 : k = 50 \qquad \text{mot} \qquad H_1 : k < 50.$$

Under H_0 har me at X er hypergeometrisk fordelt med punktsannsyn

$$h(x; 100, 20, k = 50) = \frac{\binom{50}{x}\binom{50}{20-x}}{\binom{100}{20}} \qquad \text{for} \qquad 0 \leq x \leq 20.$$

P-verdien til testen er gjeve som

$$P(X \leq 3 | k = 50) = \sum_{x=0}^{3} \frac{\binom{50}{x}\binom{50}{20-x}}{\binom{100}{20}}.$$
$$\approx 0.0004$$

Det vil si at dersom H_0 er riktig er sannsynligheten for å få det vi fikk i prøveavstemningen, eller noe mer ekstremt, så liten som 0.0004. Det er dermed rimeleg å konkludere at H_0 er feil, og at fleirtalet imot streik