NTNU

Kunnskap for ei betre verd

Institutt for matematiske fag

Eksamensoppgåve i **TMA4245 Statistikk**

Fagleg kontakt under eksamen:

Tlf:

Eksamensdato:

Eksamenstid (frå–til):

Hjelpemiddelkode/Tillatne hjelpemiddel:

Annan informasjon:

Målform/språk: nynorsk

Sidetal: 10

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 Elektrisk sparkesykkel

a) Sidan X er binomisk fordelt har me punktsannsyn

$$P(X = 4) = \binom{17}{4} 0.2^4 (1 - 0.2)^{17-4} = 0.209.$$

Nyttar komplementærsetninga

$$\begin{aligned} P(X \geq 4) &= 1 - P(X \leq 3) \\ &= 1 - \sum_{x=0}^{3} \binom{17}{x} 0.2^x (1 - 0.2)^{17-x} \\ &= 1 - (0.023 + 0.096 + 0.191 + 0.239) \\ &= 0.449. \end{aligned}$$

Nyttar definisjonen på vilkårsbunde (bokmål: betinget) sannsyn

$$\begin{aligned} P(X \geq 6 \mid X \geq 4) &= \frac{P(X \geq 6 \cap X \geq 4)}{P(X \geq 4)} \\ &= \frac{P(X \geq 6)}{P(X \geq 4)} \\ &= \frac{1 - P(X \leq 5)}{P(X \geq 4)} \\ &= \frac{1 - (0.023 + 0.096 + 0.191 + 0.239 + 0.209 + 0.136)}{0.449} \\ &= 0.232. \end{aligned}$$

b) Sentralgrenseteoremet seier at dersom $X_1, X_2, \ldots, X_n$ er uavhengige og identisk fordelte stokastiske variablar med med forventningsverdi $\mathrm{E}(X_i) = \mu$ og $\mathrm{Var}(X_i) = \sigma^2$ vil sannsynsfordelinga til

$$\frac{\bar{X} - \mu}{\sqrt{\frac{\sigma^2}{n}}}$$

gå mot ei standard normalfordeling når $n \to \infty$.

La $X_1, X_2, \ldots X_{215}$ vere uavhengige stokastiske variablar der $X_i = 1$ dersom pasient nummer i hadde ein bruddskade etter ei sparkesykkelulukke og 0 elles. Då er X_i Bernoullifordelt med suksessannsyn p for $i = 1, 2, \ldots, 215$. Me har vidare at $\mu = \mathrm{E}(X_i) = p$ og $\sigma^2 = \mathrm{Var}(X_i) = p(1 - p)$. Det er kjend at $\widehat{p} = \frac{X}{n}$ er sannsynsmaksimeringsestimatoren til p, der $X = \sum_{i=1}^{215} X_i$.

Frå sentralgrenseteoremet har me at

$$Z = \frac{\bar{X} - \mu}{\sqrt{\frac{\sigma^2}{n}}} = \frac{\widehat{p} - p}{\sqrt{\frac{p(1-p)}{n}}}$$

går mot eit standard normalfordeling. Me bytter p i nemnaren med $\widehat{p}$ og får

$$Z = \frac{\widehat{p} - p}{\sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}}} \approx n(z; 0, 1).$$

Me tar utgangspunkt i

$$P\left(-z_{0.25} \leq \frac{\widehat{p} - p}{\sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}}} \leq z_{0.025}\right) = 0.95$$

og løyser ulikskapane med hensyn på p:

$$P\left(\widehat{p} - z_{0.025}\sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}} \leq p \leq \widehat{p} + z_{0.025}\sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}}\right) = 0.95.$$

Eit 95 % konfindensintervall for p er difor

$$\left[\widehat{p} - z_{0.025}\sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}}, \widehat{p} + z_{0.025}\sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}}\right].$$

Innsatt tala får me $[0.193, 0.309]$.

Oppgåve 2 Lading av elbil

a) Me ynskjer å utføre følgjande hypotesetest

$$\mathrm{H}_0 : \mu = 3000 \qquad \text{mot} \qquad \mathrm{H}_1 : \mu > 3000.$$

Sidan både forventningsverdien og variansen er ukjende har me under nullhypotesen at

$$T = \frac{\bar{X} - \mu}{\sqrt{\frac{S^2}{17}}} \sim t_{16}$$

Me forkastar nullhypotesen dersom $T^{\text{obs}} \geq t_{16,0.05} = 1.745$. Innsatt tala våre har me

$$T^{\text{obs}} = \frac{3200 - 3000}{\sqrt{\frac{300^2}{17}}} = 2.749.$$

Me forkastar altså nullhypotesen.

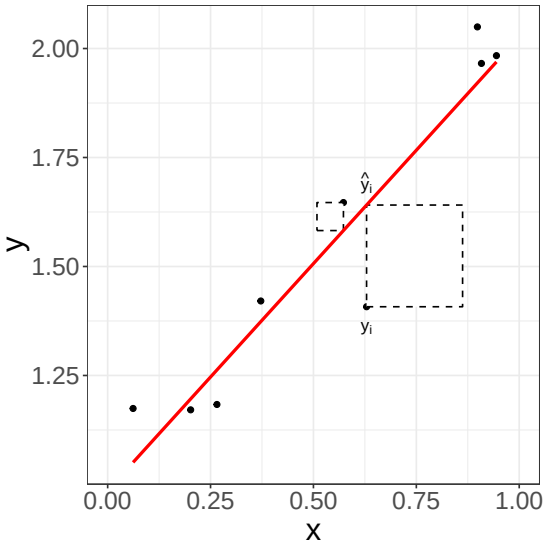
Ja, det grunnlag til å tru at straumforbruket er høgare enn 3000 kilowattimar.

Oppgåve 3 Avrenning

a) Me kan finne minste kvadraters metode estimatorar b_0 og b_1 til β_0 og β_1 ved å minimere summen av kvadratfeila for den estimerte modellen:

$$\text{SSE} = \sum_{i=1}^{25}(y_i - \widehat{y}_i)^2 = \sum_{i=1}^{25}(y_i - b_0 - b_1 x_i)^2.$$

Dette gjeres ved å setje dei deriverte $\partial\text{SSE}/\partial b_0$ og $\partial\text{SSE}/\partial b_1$ lik 0 og løyse det lineære likningssystemet for b_0 og b_1.



Ved lineær regresjon har me følgjande modellantakingar

- lineær samanheng for $\text{E}(Y|x)$: i Figur 1(a) er det ein klar lineær samanheng mellom x og y.
- varians uavhengig av x, det vil seie $\text{Var}(Y|x) = \sigma^2$: det er ikkje noko tydeleg trend i spredning av predikert avrenning i Figure 1(b) mot residuala.
- støyledda ϵ_i er normalfordelt og uavhengige: observasjonane verkar å vere jamnt fordelt omkring den tilpassa linja i Figur 1(a) utan noko tydeleg trend i spredning. Residuala verkar og å vere sentrert rundt 0. Ein kunne og ha sett på eit normalsannsynplott av residuala for å avgjere om dei er normalfordelte.

b) Estimert forventa avrenning når $x = 2000$ er $-1364 + 1.08 \cdot 2000 = 796$.

Me har at

$$\begin{aligned}
\mathrm{E}\left(\widehat{Y}_0 - Y_0\right) &= \mathrm{E}(\widehat{\beta}_0 + \widehat{\beta_1}x_0 - \beta_0 - \beta_1 x_0 - \epsilon) \\
&= \mathrm{E}(\widehat{\beta}_0) + x_0\mathrm{E}(\widehat{\beta_1}) - \beta_0 - \beta_1 x_0 - \mathrm{E}(\epsilon) \\
&= \beta_0 + \beta_1 x_0 - \beta_0 - \beta_1 x_0 \\
&= 0
\end{aligned}$$

sidan $\widehat{\beta}_0$ og $\widehat{\beta}_1$ er forventningsrette estimatorar. Vidare er

$$\begin{aligned}
\mathrm{Var}\left(\widehat{Y}_0 - Y_0\right) &= \mathrm{Var}\left(\widehat{\beta}_0 + \widehat{\beta}_1 x_0 - \beta_0 - \beta_1 x_0 - \epsilon\right) \\
&= \mathrm{Var}\left(\bar{Y} - \widehat{\beta}_1\bar{x} + \widehat{\beta}_1 x_0 - \epsilon\right) \\
&= \mathrm{Var}(\bar{Y}) + (x_0 - \bar{x})^2\mathrm{Var}(\widehat{\beta}_1) + \mathrm{Var}(\epsilon)
\end{aligned}$$

der me har brukt at $\bar{Y}$ og $\widehat{\beta}_1$ er uavhengige stokastiske variablar. Me treng $\mathrm{Var}(\widehat{\beta}_1)$:

$$\begin{aligned}
\mathrm{Var}(\widehat{\beta}_1) &= \frac{1}{\left(\sum_{i=1}^{25}(x_i - \bar{x})^2\right)^2}\mathrm{Var}\left(\sum_{i=1}^{25}(x_i - \bar{x})Y_i\right) \\
&= \frac{1}{\left(\sum_{i=1}^{25}(x_i - \bar{x})^2\right)^2}\sum_{i=1}^{25}(x_i - \bar{x})^2\mathrm{Var}(Y_i) \\
&= \frac{\sigma^2\sum_{i=1}^{25}(x_i - \bar{x})^2}{\left(\sum_{i=1}^{25}(x_i - \bar{x})^2\right)^2} \\
&= \frac{\sigma^2}{\sum_{i=1}^{25}(x_i - \bar{x})^2}.
\end{aligned}$$

Ved å setje dette inn får me

$$\begin{aligned}
\mathrm{Var}\left(\widehat{Y}_0 - Y_0\right) &= \mathrm{Var}(\bar{Y}) + (x_0 - \bar{x})^2\frac{\sigma^2}{\sum_{i=1}^{25}(x_i - \bar{x})^2} + \mathrm{Var}(\epsilon) \\
&= \frac{\sigma^2}{25} + \frac{\sigma^2(x_0 - \bar{x})^2}{\sum_{i=1}^{25}(x_i - \bar{x})^2} + \sigma^2 \\
&= \sigma^2\left(1 + \frac{1}{25} + \frac{(x_0 - \bar{x})^2}{\sum_{i=1}^{25}(x_i - \bar{x})^2}\right).
\end{aligned}$$

Me finn eit 95 % prediksjonsintervall for Y_0 ved å sjå på

$$\frac{\widehat{Y}_0 - Y_0}{\sqrt{\sigma^2\left(1 + \frac{1}{25} + \frac{(x_0 - \bar{x})^2}{\sum_{i=1}^{25}(x_i - \bar{x})^2}\right)}}$$

som er standard normalfordelt. Sidan σ^2 er ukjend får me

$$\frac{\widehat{Y}_0 - Y_0}{\sqrt{s^2\left(1+\frac{1}{25}+\frac{(x_0-\bar{x})^2}{\sum_{i=1}^{25}(x_i-\bar{x})^2}\right)}} \sim t_{23}.$$

Ta utgangspunkt i

$$P\left(-t_{23,0.025} \leq \frac{\widehat{Y}_0 - Y_0}{\sqrt{s^2\left(1+\frac{1}{25}+\frac{(x_0-\bar{x})^2}{\sum_{i=1}^{25}(x_i-\bar{x})^2}\right)}} \leq t_{23,0.05}\right) = 0.95.$$

Eit 95 % prediksjonsintervall for Y_0 er gjeve som

$$\left[\widehat{Y}_0 - t_{23,0.025}\sqrt{s^2\left(1+\frac{1}{25}+\frac{(x_0-\bar{x})^2}{\sum_{i=1}^{25}(x_i-\bar{x})^2}\right)}, \widehat{Y}_0 + t_{23,0.025}\sqrt{s^2\left(1+\frac{1}{25}+\frac{(x_0-\bar{x})^2}{\sum_{i=1}^{25}(x_i-\bar{x})^2}\right)}\right].$$

Oppgåve 4 Vekt gullbarre

a) Ein god estimator $\mu^\star$ for den ukjende parameteren μ er forventningsrett, det vil seie $\mathrm{E}(\mu^\star) = \mu$. Av fleire forventningsrette estimatorar $\mu_1^\star, \ldots \mu_K^\star$ vel me den estimatoren med lågast varians.

Me har

$$\mathrm{E}(\widehat{\mu}) = \mathrm{E}(Y) = \mu$$

og

$$\mathrm{E}(\tilde{\mu}) = \mathrm{E}\left(\frac{1}{2}(X+Y)\right) = \frac{1}{2}\mathrm{E}(X) + \frac{1}{2}\mathrm{E}(Y) = \frac{\mu}{2} + \frac{\mu}{2} = \mu.$$

Sidan begge estimatorane er forventningsrette må me avgjere kva for ein av dei som har minst varians:

$$\mathrm{Var}(\widehat{\mu}) = \mathrm{Var}(Y) = 0.5^2 = 0.25$$

og

$$\begin{aligned}\mathrm{Var}(\tilde{\mu}) &= \mathrm{Var}\left(\frac{1}{2}(X+Y)\right)\\ &= \frac{1}{2}^2\mathrm{Var}(X) + \frac{1}{2}^2\mathrm{Var}(Y) + 2\cdot\frac{1}{2}\cdot\frac{1}{2}\mathrm{Cov}(X,Y)\\ &= \frac{1}{4}\cdot 1^2 + \frac{1}{4}\cdot 0.5^2 + \frac{1}{2}\cdot(-0.2)\\ &= 0.2125.\end{aligned}$$

Sidan $\mathrm{Var}(\tilde{\mu}) < \mathrm{Var}(\widehat{\mu})$ vil me føretrekke $\tilde{\mu}$ som estimator for μ.

Oppgåve 5 Momentgenererande funksjon

a) To stokastiske variablar V og W har same sannsynsfordeling dersom dei momentgenererande funksjonane deira er like for alle t. Me må altså vise at den momentgenererande funksjonen til Y er $M_Y(t) = \frac{1}{(1-t\beta)^{n\alpha}}$.

Merk: det opphavelege eksamenssettet inneholdt ein typografisk feil i definisjonen for den momentgenererande funksjonen då det stod $M_X(t) = (1 - t/\beta)^{-\alpha}$. Utrekninga (og konklusjonen) ville ha vore på same måte som vist under. Denne feilen blir tatt omsyn til ved sensur.

Sidan $X_1, X_2, \ldots, X_n$ er uavhengige stokastiske variablar har me at den momentgenererande funksjonen til Y er gjeve ved

$$\begin{aligned} M_Y(t) &= M_{X_1+X_2+\cdots+X_n}(t) \\ &= M_{X_1}(t)M_{X_2}(t)\cdots M_{X_n}(t) \\ &= \frac{1}{(1-\beta t)^{\alpha}}\frac{1}{(1-\beta t)^{\alpha}}\cdots\frac{1}{(1-\beta t)^{\alpha}} \\ &= \frac{1}{(1-\beta t)^{n\alpha}} \end{aligned}$$

som er den momentgenererande funksjonen til ein gammafordelt stokastisk variabel med parametrar $n\alpha$ og β.

Rimelegheitsfunksjonen basert på det tilfeldige utvalet $Y = y$ er

$$L(\beta; y) = \frac{1}{\beta^{n\alpha}\Gamma(n\alpha)} y^{n\alpha-1} e^{-y/\beta}$$

for $y > 0$. Log-rimelegheitsfunksjonen er

$$l(\beta; y) = -n\alpha \log \beta - \log \Gamma(n\alpha) - (n\alpha - 1)y - \frac{y}{\beta}.$$

Me deriverer log-rimelegheitsfunksjonen, set uttrykket lik 0 og løyser likninga for β:

$$\frac{dl(\beta; y)}{\beta} = -\frac{n\alpha}{\beta} + \frac{y}{\beta^2} = 0.$$

Me får då

$$\beta n\alpha = y \Rightarrow \beta = \frac{y}{n\alpha}.$$

Det vil seie at sannsynsmaksimeringsestimatoren for β basert på y er $\widehat{\beta} = \frac{Y}{n\alpha}$.

Oppgåve 6 Sannsyn

a) Frå komplementærsetninga har me

$$\begin{aligned} P(2X > 3) &= 1 - P\left(X \leq 3/2\right) \\ &= 1 - \Phi\left(Z \leq \frac{3/2 - 1}{1}\right) \\ &= 1 - \Phi(0.5) \\ &= 1 - 0.691 \\ &= 0.309. \end{aligned}$$

Sidan X og Y er uavhengige har me

$$P(2X > 3 \mid Y > 0) = P(2X > 3) = 0.309.$$

Sidan $X - Y$ er ein lineærkombinasjon av uavhengige normalfordelte variablar er den og normalfordelt med forventningsverdi

$$\mathrm{E}(X - Y) = \mathrm{E}(X) - \mathrm{E}(Y) = 1 - 0 = 1$$

og varians

$$\mathrm{Var}(X - Y) = \mathrm{Var}(X) + \mathrm{Var}(Y) = 1 + 1 = 2.$$

Me har då

$$\begin{aligned} P(-1 \leq X - Y \leq 1) &= P\left(\frac{-1 - 1}{\sqrt{2}} \leq Z \leq \frac{1 - 1}{\sqrt{2}}\right) \\ &= \Phi(0) - \Phi(-\sqrt{2}) \\ &= 0.5 - 0.079 \\ &= 0.421. \end{aligned}$$

b) Kvar X_i kan anten vere mindre eller lik x, eller større enn x. Sannsynet for at ein tilfeldig vald X_i er mindre enn eller lik x er $p = P(X_i \leq x) = P(X \leq x)$ for $i = 1, 2, \ldots, 10$. Sidan kvar av hendingane $\{X_1 \leq x\}, \ldots, \{X_{10} \leq x\}$ er uavhengige av kvarandre har me eit binomisk forsøk med konstant suksessannsyn p og $n = 10$ forsøk. Me har derfor

$$\begin{aligned} P(\text{høgst 5 av } X_i - \text{ane er større enn } x) &= \sum_{k=0}^{5} \binom{10}{k} p^k (1-p)^{10-k} \\ &= \sum_{k=0}^{5} \binom{10}{k} \left[P(X \leq x)\right]^k \left(1 - P(X \leq x)\right)^{10-k} \\ &= \sum_{k=0}^{5} \binom{10}{k} \left[\Phi(x-1)\right]^k \left(1 - \Phi(x-1)\right)^{10-k} \end{aligned}$$

der $\Phi(x)$ er den kumulative fordelingsfunksjonen til standard normalfordelinga.

Sidan X_i og Y_j er parvis uavhengig for alle par av i og j har me

$$\begin{aligned} P(\max\{X_1,\ldots,X_{10},Y_1,\ldots,Y_{15}\} \leq z) &= P(X_1 \leq z,\ldots,X_{10} \leq z, Y_1 \leq z,\ldots,Y_{15} \leq z) \\ &= \prod_{i=1}^{10} P(X_i \leq z) \prod_{j=1}^{15} P(Y_j \leq z) \\ &= [P(X \leq z)]^{10} [P(Y \leq z)]^{15} \\ &= [\Phi(z-1)]^{10} [\Phi(z)]^{15} \end{aligned}$$

der $\Phi(z)$ er den kumulative fordelingsfunksjonen til standard normalfordelinga.

Oppgåve 7 Hypotesetest uniformfordelinga

a) Under nullhyposen er sannsynstettleiken til X_1 gjeve ved

$$f(x;0) = \begin{cases} 1 & x \in [0,1] \\ 0 & \text{elles.} \end{cases}$$

Sannsynet for type I-feil for forkastningsregel 1

$$P(\text{forkast } \mathrm{H}_0 \text{ når } \mathrm{H}_0 \text{ er sann}) = P(X_1 \geq 0.95) = 0.05.$$

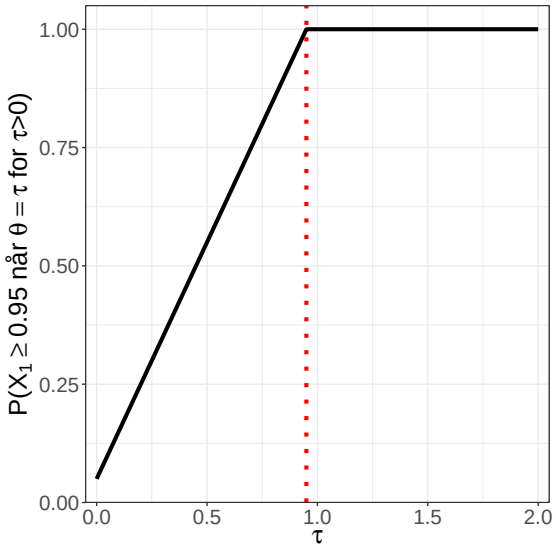
Me ser på to tilfeller: for $0 < \tau < 0.95$

$$\begin{aligned} P(X_1 \geq 0.95 \text{ når } \theta = \tau \text{ for } \tau > 0) &= \int_{0.95}^{1+\tau} 1 \mathrm{d}x \\ &= 1 + \tau - 0.95 \\ &= \tau + 0.05. \end{aligned}$$

og for $\tau \geq 0.95$ vil alltid $X_1 \geq 0.95$. Det vil seie at me vil at me alltid forkastar H_0 når $\tau \geq 0.95$ ved forkastningsregel 1.

Til saman har me

$$\begin{aligned} P(\text{forkast } \mathrm{H}_0 \text{ når } \theta = \tau \text{ for } \tau > 0) &= P(X_1 \geq 0.95 \text{ når } \theta = \tau \text{ for } \tau > 0) \\ &= \begin{cases} \tau + 0.05 & 0 < \tau < 0.95 \\ 1 & \tau \geq 0.95 \end{cases} \end{aligned}$$



Figur 1: Testen sin styrke for forkastningsregel 1.

Ei skisse av testen sin styrke er gjeve i Figur 1.

Oppgåva kan løysast på fleire måter. Merk at under nullhypotesen er

$$f(x_1, x_2; \theta = 0) = f(x_1; 0) f(x_2; 0) = \begin{cases} 1 & 0 \leq x_1 \leq 1, 0 \leq x_2 \leq 1 \\ 0 & \text{elles .} \end{cases}$$

Ved å sjå på sannsynet til $X_1 + X_2 > k$, gjeve nullhypotesen, får me

$$\begin{aligned} P(X_1 + X_2 > k \text{ når } \theta = 0) &= \int_{k-1}^{1} \int_{k-x_1}^{1} 1 \mathrm{d}x_2 \mathrm{d}x_1 \\ &= \int_{k-1}^{1} 1 - k + x_1 \mathrm{d}x_1 \\ &= \frac{(2-k)^2}{2}. \end{aligned}$$

Dersom me krev sannsyn for type I-feil som forkastningsregel 1 må me ha $P(X_1 + X_2 > k \text{ når } \theta = 0) = 0.05$. Det vil seie

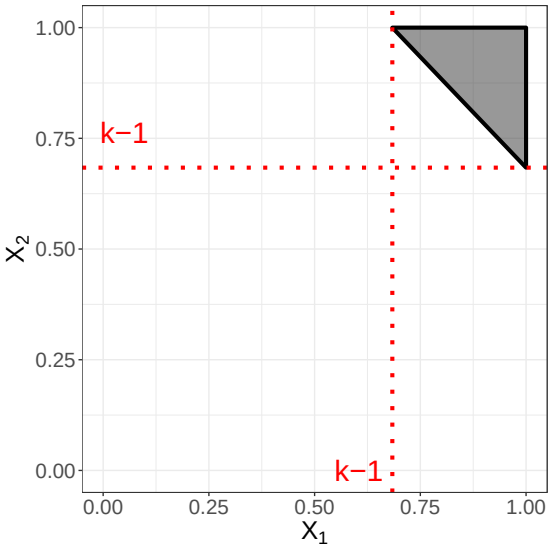
$$\frac{(2-k)^2}{2} = 0.05 \Rightarrow k = 2 - \sqrt{0.1}.$$

Hugs at simultantettleiken er konstant for $(x_1, x_2) \in [0, 1] \times [0, 1]$. Eit geometrisk argument kan sjåast frå Figur 2. Merk at den største verdien me kan ha er $k = 2$ når $X_1 = X_2 = 1$. Me krev at arealet av den skraverte trekanten

må vere 0.05. Dersom me fikserer $X_1 = 1$ er den minste verdien X_2 kan ha $X_2 = k-1$ for at kravet om $X_1+X_2 \geq k$ skal vere oppfylt. Grunna symmetri er den minste verdien X_1 kan ta dersom $X_2 = 1$ lik $X_1 = k-1$. Arealet av trekanent er derfor

$$\frac{1}{2}\left((1-(k-1))(1-(k-1)\right) = \frac{(2-k)^2}{2}$$

som gir den same verdien for k som over.



Figur 2: Skravert område viser der $X_1 + X_2 \geq k$ er slik at arealet er lik 0.05.