TMA4245 Statistikk
Eksamen mai 2016

Norges teknisk-naturvitenskapelige universitet
Institutt for matematiske fag

Løsningsskisse

Oppgave 1

a) Lar X være kvadratprisen. Har da at $X \sim N(\mu, \sigma^2)$, med $\mu = 30$ og $\sigma^2 = 2,5^2$.

$$P(X < 30) = P(X < \mu) = \underline{\underline{0.5}}$$

$$P(X > 25) = 1 - P(X < 25) = 1 - P\left(\frac{X-30}{2.5} < \frac{25-30}{2.5}\right)$$
$$= 1 - \Phi(-2) = 1 - 0.0228 = \underline{\underline{0.9772}}$$

$$P(X > 25 | X < 30) = \frac{P(X > 25 \cap X < 30)}{P(X < 30)}$$
$$= \frac{P(X < 30) - P(X < 25)}{P(X < 30)} = \frac{0.5 - 0.0228}{0.5} = \underline{\underline{0.9544}}$$

b)
- Gustav ser på 40 kvm med kvadratpris X_G: Pris $\underline{\underline{Y_G = 40X_G}}$.
- Margrethe ser på 50 kvm med kvadratpris X_M: Pris $\underline{\underline{Y_M = 50X_M}}$.
- Prisforskjellen $\underline{\underline{D = Y_M - Y_G = 40X_G - 50X_M}}$.

Da D er en lineær kombinasjon av uavhengige normalfordelte stokastiske variable, så er D normalfordelt med

$$E(D) = E(40X_G - 50X_M) = 40E(X_G) - 50E(X_M) = 10 \cdot 30 = 300$$

$$\mathrm{Var}(D) = \mathrm{Var}(40X_G - 50X_M) = 40^2\mathrm{Var}(X_G) + 50^2\mathrm{Var}(X_M) = 4100 \cdot 2.5^2$$

Altså er

$$P(D < 0) = P\left(\frac{D - 300}{\sqrt{4000}2.5} < \frac{0 - 300}{\sqrt{4100}2.5}\right) = \Phi(-1.87) = \underline{\underline{0.0307}}$$

Sannsynligheten for at leiligheten Margrethe vurderer er billigast er på 3.07 %.

c) Ønsker å finne et 95% konfidensintervall for μ basert på data $x_1, x_2, \ldots, x_n$, der $n = 15$. Har at X_i er uavhengige og normalfordelte med ukjent μ og σ_2. Kan altså bruke T-observator:

$$T = \frac{\bar{X} - \mu}{S/\sqrt{n}} \sim t_{n-1},$$

T er student-t fordelt med $n - 1 = 14$ frihetsgrader. Har dermed

$$P(-t_{\alpha/2,n-1} < T < t_{\alpha/2,n-1}) = 1 - \alpha$$

Setter inn for T, løser hver ulikhet med hensyn på μ og setter så sammen igjen,

$$P\left(-t_{\alpha/2,n-1} < \frac{\bar{X} - \mu}{S/\sqrt{n}} < t_{\alpha/2,n-1}\right) = 1 - \alpha$$

$$P\left(\bar{X} - t_{\alpha/2,n-1}S/\sqrt{n} < \mu < \bar{X} + t_{\alpha/2,n-1}S/\sqrt{n}\right) = 1 - \alpha$$

Konfidensintervallet blir dermed

$$\underline{\underline{\left(\bar{X} - t_{\alpha/2,n-1}S/\sqrt{n}, \bar{X} + t_{\alpha/2,n-1}S/\sqrt{n}\right)}}.$$

Med $n = 15$, $t_{0.025,14} = 2.145$, $\bar{x} = 32$ og $s^2 = 1/(n-1)\sum_1^{15}(x_i - \bar{x})^2 = 1/14 \cdot 74.1$ blir 95 % konfidensintervallet $\underline{\underline{(30.7,\ 33.3)}}$.

Oppgave 2

a)
- Fra histogrammet i figur 1 ser vi at fordelingen til X (sann gjennomstrømming) ser ut til å være symmetrisk om 5, og har dermed forventningsverdi på 5. Videre ser X ut til å kunne være normalfordelt (eller nær normalfordelt). Da skal ca 95% av obserbasjonen ligge innenfor forventningsverdien +/- 2 (1.96) standardavvik. Dermed ser standardavviket ut til å være 2.
- Dersom vi i figur 2 ser på sensormålte verdier (y) for sann gjennomstrømming lik (omlag) x=6, ser vi at de er symmetrisk om y=6, altså er forventningsverdien $E(Y|X = 6) = 6$. Videre så ser mesteparten av y-ene er mellom 4 og 8, og det betinga standardavviket ser dermed ut til å være 1.
- Det er en positiv samvariasjon mellom x og y (når x øker så øker y), og vi har dermed en positiv korrelasjon (og kovarians).

b)
- Den inntegna linja $y = x$ ser ut til å være den beste linære tilpassinga, og estimatene blir (ca) $a = 0$ og $b = 1$.
- Tilpassa modell blir dermed $\hat{y}|x = x$, og for $x = 4$ blir $\hat{y} = 4$
- Vi har to hovedantakelser: 1) Forventningen til Y er lineær i x, og 2) Støyleddene er uavhengig identisk normalfordelt. Anatakelse 1) er OK (observasjonene for y ser ut til å være sentert om en rett linje), ang 2) så ser støyleddene ut til å kunne være både normalfordelt (symmetriske om linja) og uavhenge, men variansen (σ_ϵ^2) ser ut til å øke med x.

Oppgave 3

a) X_1 er poissonfordelt med forventning $\mu = \lambda t_1 = 10$. Ved å sette inn i oppgitt formel for punktsannsynlighet får man da at

$$P(X_1 = 8) = \frac{10^8}{8!} e^{-10} = \underline{\underline{0.1126}}.$$

Dessuten får vi at

$$P(X_1 \geq 8) = 1 - P(X_1 < 8) = 1 - P(X_1 \leq 7) = 1 - 0.2202 = \underline{\underline{0.7798}},$$

der vi i den siste overgangen har benyttet tabell over poissonfordeling med $\mu = 10$ i formelsamlinga. Ved igjen å benytte tabell over poissonfordeling får vi

$$P(8 \leq X_1 \leq 12) = P(X_1 \leq 12) - P(X_1 < 8) = P(X_1 \leq 12) - P(X_1 \leq 7)$$

$$= 0.7916 - 0.2202 = \underline{\underline{0.5714}}.$$

b) For å bestemme hvilken av de tre foreslåtte estimatorene som er å foretrekke må vi først finne ut hvilke som er forventingsrette. Fra oppgitt formel for $E[\widehat{\lambda}]$ vet vi allerede at $\widehat{\lambda}$ er forveningsrett. Ved å benytte regneregler for forventingsverdi og at $E[X_i] = \lambda t_i$ følger det at

$$E[\widetilde{\lambda}] = E\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right] = \frac{1}{n} E\left[\sum_{i=1}^{n} X_i\right] = \frac{1}{n}\sum_{i=1}^{n} E\left[X_i\right] = \frac{1}{n}\sum_{i=1}^{n} \lambda t_i$$

$$= \frac{\lambda}{n}\sum_{i=1}^{n} t_i = \frac{\lambda}{5}(1+2+5+1+5) = 2.8 \cdot \lambda \neq \lambda$$

og

$$E\left[\widehat{\widehat{\lambda}}\right] = E\left[\frac{1}{n}\sum_{i=1}^{n} \frac{X_i}{t_i}\right] = \frac{1}{n} E\left[\sum_{i=1}^{n} \frac{X_i}{t_i}\right] = \frac{1}{n}\sum_{i=1}^{n} E\left[\frac{X_i}{t_i}\right] = \frac{1}{n}\sum_{i=1}^{n} \frac{E[X_i]}{t_i}$$

$$= \frac{1}{n}\sum_{i=1}^{n} \frac{\lambda t_i}{t_i} = \frac{1}{n}\sum_{i=1}^{n} \lambda = \frac{1}{n} \cdot n\lambda = \lambda.$$

Vi ser følgelig at $\widetilde{\lambda}$ er forventningsskjev, mens $\widehat{\lambda}$ og $\widehat{\widehat{\lambda}}$ er forventingsrette. Siden vi foretrekker at en estimator er forventingsrett vet vi dermed at vi foretrekker $\widehat{\lambda}$ eller $\widehat{\widehat{\lambda}}$. Vi foretrekker den av disse to som har minst varians. Ved å benytte oppgitt formel for variansen til $\widehat{\lambda}$ får vi

$$\mathrm{Var}[\widehat{\lambda}] = \frac{\lambda}{\sum_{i=1}^{n} t_i} = \frac{\lambda}{1+2+5+1+5} = 0.0714 \cdot \lambda,$$

og ved å benytte regneregler for varians for uavhengige stokastiske variabler og at $\mathrm{Var}[X_i] = \lambda t_i$ får vi

$$\mathrm{Var}\left[\widehat{\widehat{\lambda}}\right] = \mathrm{Var}\left[\frac{1}{n}\sum_{i=1}^{n} \frac{X_i}{t_i}\right] = \frac{1}{n^2}\mathrm{Var}\left[\sum_{i=1}^{n} \frac{X_i}{t_i}\right] = \frac{1}{n^2}\sum_{i=1}^{n} \mathrm{Var}\left[\frac{X_i}{t_i}\right]$$

$$= \frac{1}{n^2}\sum_{i=1}^{n}\frac{\text{Var}[X_i]}{t_i^2} = \frac{1}{n^2}\sum_{i=1}^{n}\frac{\lambda t_i}{t_i^2} = \frac{\lambda}{n^2}\sum_{i=1}^{n}\frac{1}{t_i}$$
$$= \frac{\lambda}{5^2}\left(\frac{1}{1}+\frac{1}{2}+\frac{1}{5}+\frac{1}{1}+\frac{1}{5}\right) = 0.116\cdot\lambda.$$

Vi ser dermed at av de to forventningsrette estimatorene er det $\widehat{\lambda}$ som har minst varians slik at av de tre foreslåtte estimatorene fortrekker vi $\widehat{\lambda}$.

c) Rimelighetsfunksjonen for λ blir

$$L(\lambda) = f(x_1, x_2, \ldots, x_n; \lambda) = \prod_{i=1}^{n}\left[\frac{(\lambda t_i)^{x_i}}{x_i!}e^{-\lambda t_i}\right]$$

Ved å benytte regneregler for ln får vi dermed at log-rimeligelighetsfunksjonen blir gitt ved

$$l(\lambda) = \ln L(\lambda) = \sum_{i=1}^{n}\left[x_i\ln(\lambda t_i) - \ln(x_i) - \lambda t_i\right]$$
$$= \sum_{i=1}^{n} x_i\ln(\lambda t_i) - \sum_{i=1}^{n}\ln(x_i!) - \lambda\sum_{i=1}^{n} t_i.$$

For å finne for hvilken verdi av λ denne funksjonen har sitt maksimum finner vi den deriverte og krever at denne er lik null,

$$l'(\lambda) = \sum_{i=1}^{n} x_i\frac{1}{\lambda t_i}t_i - 0 - \sum_{i=1}^{n} t_i = \frac{1}{\lambda}\sum_{i=1}^{n} x_i - \sum_{i=1}^{n} t_i = 0 \Leftrightarrow \lambda = \frac{\sum_{i=1}^{n} x_i}{\sum_{i=1}^{n} t_i}$$

Sannsynlighetsmaksimeringsestimatoren for λ blir følgelig

$$\underline{\widehat{\lambda} = \frac{\sum_{i=1}^{n} X_i}{\sum_{i=1}^{n} t_i}.}$$

d) Vi ønsker å bestemme om det er grunnlag for å påstå at besøksintensiteten for større for SeMeg enn for konkurrenten. Følgelig må vi som H_1 velge at $\lambda > \lambda_0$. Vi må dermed teste

$$\underline{H_0 : \lambda = \lambda_0 \quad \text{mot} \quad H_1 : \lambda > \lambda_0.}$$

For å konstruere en testobservator benytter vi, som oppgitt i oppgaveteksten, at $\widehat{\lambda}$ er tilnærmet normalfordelt. Dermed vil

$$\frac{\widehat{\lambda} - \text{E}[\widehat{\lambda}]}{\sqrt{\text{Var}[\widehat{\lambda}]}} = \frac{\widehat{\lambda} - \lambda}{\sqrt{\frac{\lambda}{\sum_{i=1}^{n} t_i}}} \tag{3.1}$$

være tilnærmet standard normalfordelt. Vi får vår testobservator ved å erstatte λ med verdien til λ når H_0 er riktig,

$$\underline{Z = \frac{\widehat{\lambda} - \lambda_0}{\sqrt{\frac{\lambda_0}{\sum_{i=1}^{n} t_i}}}.}$$

som altså er tilnærmet standard normalfordelt når H_0 er riktig.

For å regne ut p-verdien trenger vi først å regne ut observert verdi for testobservatoren og å bestemme et forkastningskriterium. Dersom λ er stor (dvs H_1 riktig) vil $\widehat{\lambda}$ tendere til å blir stor, og dermed vil da også Z tendere til å bli stor. Det er dermed rimelig å forkaste H_0 dersom Z er stor. Forkastningskriteriet blir dermed på formen at vi skal forkaste H_0 dersom

$$Z \geq k.$$

Innsatt observerte verdier får vi at

$$\widehat{\lambda} = \frac{\sum_{i=1}^{n} x_i}{\sum_{i=1}^{n} t_i} = \frac{8+20+48+10+62}{1+2+5+1+5} = 10.5714.$$

Observert verdi av testobservatoren blir dermed

$$z_{\text{obs}} = \frac{10.5714 - 10}{\sqrt{\frac{10}{1+2+5+1+5}}} = 0.6761.$$

Dette gir at p-verdien blir

$$p = P(Z \geq z_{\text{obs}}|H_0) = P(Z \geq 0.6761|H_0) = 1 - P(Z < 0.6761|H_0)$$

$$= 1 - P(Z \leq 0.6761|H_0) = 1 - \Phi(0.6761) = 1 - 0.7517 = \underline{0.2483}.$$

Dette betyr at dersom H_0 er riktig og man gjentar et slikt forsøk gjentatte ganger, vil man observere det man her har observert eller noe mer ekstremt såpass ofte som i cirka en av fire ganger. Det vi har observert er altså ikke en urimelig verdi selv om H_0 er riktig og det er helt klart ikke grunnlag for å forkaste H_0. Det er dermed ikke grunnlag for å påstå at besøksintensiteten for SeMeg er større enn for konkurrenten.

e) Bestemmer først kritisk verdi k når $\alpha = 0.05$. Det generelle kravet er

$$P(\text{Forkast } H_0|H_0 \text{ riktig}) = \alpha = 0.05,$$

som i situasjonen i denne oppgaven blir at

$$P(Z \geq k|H_0 \text{ riktig}) = 0.05.$$

Ved å benytte at Z er (tilnærmet) standard normalfordelt når H_0 er riktig finner vi fra tabell over standard normalfordelingen at

$$k = z_{0.05} = 1.645.$$

Spørmålet i oppgaven kan matematisk formuleres som at vi ønsker å finne λ slik at

$$P(\text{Forkast } H_0|\lambda) \geq 0.9.$$

Starter med å sette inn forkastningskriteriet i situasjonen vi betrakter og setter inn uttrykket vi har for Z,

$$P(Z \geq 1.645|\lambda) \geq 0.9,$$

$$P\left(\frac{\widehat{\lambda}-\lambda_0}{\sqrt{\frac{\lambda_0}{\sum_{i=1}^n t_i}}} \geq 1.645 \middle| \lambda\right) \geq 0.9.$$

For å kunne benytte tabell over standard normalfordeling omskriver vi denne ligningen slik at den inneholder den standard normalfordelte variabelen gitt i (3.1),

$$P\left(\widehat{\lambda} \geq \lambda_0 + 1.645\sqrt{\frac{\lambda_0}{\sum_{i=1}^n t_i}} \middle| \lambda\right) \geq 0.9,$$

$$P\left(\frac{\widehat{\lambda}-\lambda}{\sqrt{\frac{\lambda}{\sum_{i=1}^n t_i}}} \geq \frac{\lambda_0 + 1.645\sqrt{\frac{\lambda_0}{\sum_{i=1}^n t_i}} - \lambda}{\sqrt{\frac{\lambda}{\sum_{i=1}^n t_i}}} \middle| \lambda\right) \geq 0.9.$$

Ved å illustrere kravet over ved å tegne opp sannsynlighetstettheten til en standard normalfordelt variabel ser vi da at kravet blir

$$\frac{\lambda_0 + 1.645\sqrt{\frac{\lambda_0}{\sum_{i=1}^n t_i}} - \lambda}{\sqrt{\frac{\lambda}{\sum_{i=1}^n t_i}}} \leq z_{0.9} = -z_{0.10} = -1.282. \tag{3.2}$$

Løser denne ulikheten ved først å løse den tilsvarende ligningen

$$\frac{\lambda_0 + 1.645\sqrt{\frac{\lambda_0}{\sum_{i=1}^n t_i}} - \lambda}{\sqrt{\frac{\lambda}{\sum_{i=1}^n t_i}}} = -1.282,$$

$$\lambda_0 + 1.645\sqrt{\frac{\lambda_0}{\sum_{i=1}^n t_i}} - \lambda = -1.282\sqrt{\frac{\lambda}{\sum_{i=1}^n t_i}},$$

Vi setter så inn tall for λ_0 og $\sum_{i=1}^n t_i$, og skriver ligningen som en andregradsligning for $\sqrt{\lambda}$,

$$\left(\sqrt{\lambda}\right)^2 - \frac{1.282}{\sqrt{14}}\sqrt{\lambda} - \left(10 + 1.645\sqrt{\frac{10}{14}}\right) = 0.$$

Denne andregradsligningen har to løsninger, -3.2080 og 3.5506. Siden $\sqrt{\lambda}$ må være positiv er det kun den positive løsningen som er av interesse for oss, så ligningen har løsning

$$\sqrt{\lambda} = 3.5506 \quad \Rightarrow \quad \lambda = 3.5506^2 = 12.6068.$$

Ved å sjekke ulikhetskravet (3.2) for en vilkårlig verdi av λ mindre enn 12.6068 og eventuelt en vilkårlig verdi større enn 12.6068 får man at ulikheten er oppfylt dersom

$$\underline{\underline{\lambda \geq 12.6068}}.$$