Eksamen STK1100

Kandidat 15267

8. juni 2020

Oppgave 1

a)

Vi velger en tilfeldig valgt person. Dersom vi trekker en person med antistoffer i seg, kjenner vi til sensitiviteten som 95.5% og dersom vi trekker en person som ikke har antistoffer i seg, kjenner vi spesifisiteten som 98.0%. Vi kaller begivenhet A for at personen faktisk er smittet og begivenhet B for at testen gir positivt utslag. Det er en av hundre som faktisk har antistoffene i seg $P(A) = 0.01$, og vi får at:

$$P_{T1}(\mathrm{B}) = P(A) \cdot P(B|A) + P(A') \cdot P(B|A') \tag{1}$$

$$= 0.01 \cdot 0.955 + 0.99 \cdot 0.02 = 0.0294 = \underline{\underline{2.94\ \%}} \tag{2}$$

b)

Vi ønsker å finne ut om personen faktisk har antistoffene i seg gitt at personen tester positivt, altså finne $P(A|B)$. Vi husker definisjonen på Bayes setning:

$$P_{\mathrm{T1}}(A|B) = \frac{P(B|A) \cdot P(A)}{P(A) \cdot P(B|A) + P(A') \cdot P(B|A')} \tag{3}$$

$$= \frac{0.955 \cdot 0.01}{0.01 \cdot 0.955 + 0.99 \cdot 0.02} = 0.3254 = \underline{\underline{32.54\ \%}} \tag{4}$$

Vi ser at selv om en tilfeldig person tester positivt så er sannsynligheten relativt liten for at hen faktisk inneholder antistoffene. Karakteristikkene på test en kan tolkes som svært dårlige siden vi vil helst være sikre på resultatene vi får.

c)

Vi definerer begivenhetene på samme måte men bruker nå sensitiviteten og spesifiteten fra test to:

$$P_{\mathrm{T2}}(A|B) = \frac{P(B|A) \cdot P(A)}{P(A) \cdot P(B|A) + P(A') \cdot P(B|A')} \tag{5}$$

$$= \frac{0.942 \cdot 0.01}{0.01 \cdot 0.942 + 0.99 \cdot 0.002} = 0.8263 = \underline{\underline{82.63\ \%}} \tag{6}$$

Vi ser at vi får en større verdi for at personen som tester positivt faktisk har antistoffene i seg. Siden spesifisiteten er veldig nær å være perfekt i test to, betyr det at denne testen sjeldent tester feil dersom den gir et negativt utslag. Dersom spesifisiteten hadde vært 100 %, ville vi vært sikre på at personen faktisk ikke har antistoffene. Ser vi på uttrykket over så ser vi at vi også ville vært 100 % sikre på at personen som testet positivt faktisk inneholder antistoffene. Vi kan dermed konkludere med at test to i prinsippet er en bedre test.

d)

Vi ønsker å finne $P_{T1}(A'|B)$ og $P_{T2}(A'|B)$. Siden det er opplagt at dersom en person tester positivt, så har hen enten faktisk antistoffene i seg eller ikke. Vi finner sannsynlighetene på følgende måte:

$$P_{T1}(A'|B) = 1 - P_{T1}(A|B) = 1 - 0.3254 = 0.6746 = \underline{\underline{67.46\ \%}} \tag{7}$$

$$P_{T2}(A'|B) = 1 - P_{T2}(A|B) = 1 - 0.8263 = 0.1737 = \underline{\underline{17.37\ \%}} \tag{8}$$

Dersom du tester positivt på test en så er det en relativt stor sannsynlighet for at du faktisk ikke har antistoffene i deg. Dette gjelder ikke for test to derimot. Med denne testen kan man anta å faktisk ha antistoffene i seg selv om ca. 1/5 av testene viser feil.

e)

Vi fortsetter å se på Bayes setning for test to i uttrykket vi hadde i d) og setter inn at $P_{T2}(A'|B) = 0.05$:

$$0.05 = 1 - P_{T2}(A|B) = 1 - \frac{P(B|A) \cdot P(A)}{P(A) \cdot P(B|A) + P(A') \cdot P(B|A')} \tag{9}$$

Vi legger merke til at $P(B|A') = 1 - P(B'|A')$ der sistnevte sannsynlighet står for spesifisiteten. Videre får vi da at:

$$0.95 = \frac{P(B|A) \cdot P(A)}{P(A) \cdot P(B|A) + P(A') \cdot (1 - P(B'|A'))} \tag{10}$$

$$\Rightarrow P(A) \cdot P(B|A) + P(A') \cdot (1 - P(B'|A')) = \frac{P(B|A) \cdot P(A)}{0.95} \tag{11}$$

$$\Rightarrow 1 - P(B'|A') = \frac{\frac{P(B|A) \cdot P(A)}{0.95} - P(A) \cdot P(B|A)}{P(A')} \tag{12}$$

$$P(B'|A') = 1 - \frac{\frac{P(B|A) \cdot P(A)}{0.95} - P(A) \cdot P(B|A)}{P(A')} \tag{13}$$

Vi setter inn verdier og får:

$$P(B'|A') = 1 - \frac{\frac{0.942 \cdot 0.01}{0.95} - 0.01 \cdot 0.942}{0.99} = 0.9995 = \underline{\underline{99.95\ \%}} \tag{14}$$

som tilsvarer en økning i spesifisiteten på 0.15 %.

f)

Sannsynligheten for at en person som tester negativt faktisk ikke har antistoffene i seg er gitt ved Bayes setning:

$$P_{T1}(A'|B') = \frac{P(B'|A') \cdot P(A')}{P(A) \cdot P(B'|A) + P(A') \cdot P(B'|A')} \tag{15}$$

Vi kjenner sannsynlighetene og putter dem inn:

$$P_{T1}(A|B') = \frac{0.980 \cdot 0.99}{0.01 \cdot 0.045 + 0.99 \cdot 0.980} = 0.9995 = \underline{99.95\%} \tag{16}$$

Vi finner sannsyngligheten for at de 10 personene som tester negativt faktisk ikke har antistoffene i seg på følgende måte:

$$P_{T1}(\text{10 negative tester uten antistoffer}) = P_{T1}(A|B')^{10} = 0.9995^{10} = 0.9950 = \underline{99.5\%} \tag{17}$$

Sannsynligheten for at minst en person faktisk har antistoffene i seg er da gitt på følgende måte:

$$P_{T1}(\text{10 negative tester med minst en med antistoffer}) \tag{18}$$

$$= 1 - P_{T1}(\text{10 negative tester uten antistoffer}) \tag{19}$$

$$= 1 - 0.9950 = 0.0050 = \underline{0.5\ \%} \tag{20}$$

Det er altså 0.5 % av de som testet negativt som faktisk har antistoffene i seg.

Oppgave 2

a)

Vi har fått oppgitt en simultan sannsynlighetstetthet $f(x, y)$ for to kontinuerlige stokastiske variable. I tillegg til dette har vi fått oppgitt grensene dette gjelder for. Vi vet at summen av alle mulige utfall skal bli lik 1, altså må vi integrere over alle mulige utfall på følgende måte:

$$\int_0^2 \int_0^2 kx(x+y)dydx = 1 \tag{21}$$

$$\Rightarrow \int_0^2 \left[kx(xy + \frac{1}{2}y^2) \right]_{y=0}^{y=2} dx = 1 \tag{22}$$

$$\Rightarrow \int_0^2 2kx(x+1)dx = 1 \tag{23}$$

$$\Rightarrow \left[2kx^2 \left(\frac{1}{3}x + \frac{1}{2} \right) \right]_{x=0}^{x=2} = 1 \tag{24}$$

$$\Rightarrow 8k \left(\frac{2}{3} + \frac{1}{2} \right) = 1 \tag{25}$$

$$\Rightarrow k \left(\frac{4}{6} + \frac{3}{6} \right) = \frac{1}{8} \tag{26}$$

$$\Rightarrow k = \frac{1}{8} \cdot \frac{6}{7} = \frac{6}{56} = \underline{\underline{\frac{3}{28}}} \tag{27}$$

b)

Nå går grensene tilhørende Y fra x til 2. Vi benytter samme integral fra a) med nye grenser og setter inn for k:

$$P(Y \geq X) = \frac{3}{28} \int_0^2 \int_x^2 x(x+y)dydx \tag{28}$$

$$= \frac{3}{28} \int_0^2 \left[x(xy + \frac{1}{2}y^2) \right]_{y=x}^{y=2} dx \tag{29}$$

$$= \frac{3}{28} \int_0^2 \left(2x(x+1) - x(x^2 + \frac{1}{2}x^2) \right) dx \tag{30}$$

$$= \frac{3}{28} \int_0^2 \left(2x + 2x^2 - \frac{3}{2}x^3 \right) dx \tag{31}$$

$$= \frac{3}{28} \left[x^2 + \frac{2}{3}x^3 - \frac{3}{8}x^4 \right]_{x=0}^{x=2} \tag{32}$$

$$= \frac{3}{28} \left(4 + \frac{16}{3} - \frac{48}{8} \right) \tag{33}$$

$$= \frac{3}{28} \left(\frac{96 + 128 - 144}{24} \right) = \frac{3}{28} \cdot \frac{80}{24} = \underline{\underline{\frac{5}{14}}} \tag{34}$$

c)

Vi finner den marginale sannynlighetstettheten til X ved å integrere den simultane sannsynlighetstettheten med hensyn på y fra $-\infty$ til . Siden funksjonen er definert som null utenfor $y \in [0, 2]$, integrerer vi kun mellom disse to grensene:

$$f_X(x) = \frac{3}{28} \int_0^2 x(x+y)dy \tag{35}$$

$$= \frac{3}{28} \left[x \left(xy + \frac{1}{2}y^2 \right) \right]_{y=0}^{y=2} \tag{36}$$

$$= \underline{\frac{6}{28} x \left(x+1 \right)} \tag{37}$$

Altså har vi at:

$$f_X(x) = \underline{\underline{\begin{cases} \frac{3}{14}x(x+1) & \text{når } 0 \leq x \leq 2 \\ 0 & \text{ellers} \end{cases}}} \tag{38}$$

For at X og Y skal være uavhengige, må kriteriet $f_X(x) \cdot f_Y(y) = f(x, y)$ være oppfylt. Vi regner først ut $f_Y(y)$ på samme måte:

$$f_Y(y) = \frac{3}{28} \int_0^2 x(x+y)dx \tag{39}$$

$$= \frac{3}{28} \left[x^2 \left(\frac{1}{3}x + \frac{1}{2}y \right) \right]_{x=0}^{x=2} \tag{40}$$

$$= \frac{12}{28} \left(\frac{2}{3} + \frac{1}{2}y \right) \tag{41}$$

$$= \frac{3}{7} \left(\frac{4+3y}{6} \right) \tag{42}$$

$$= \underline{\frac{3}{42} \left(4+3y \right)} \tag{43}$$

som gir oss:

$$f_Y(y) = \underline{\begin{cases} \frac{1}{14}(4+3y) & \text{når } 0 \leq y \leq 2 \\ 0 & \text{ellers} \end{cases}} \tag{44}$$

Vi får videre at:

$$f_X(x) \cdot f_Y(y) = \frac{3}{14} x \left(x+1 \right) \cdot \frac{1}{14} \left(4+3y \right) \tag{45}$$

Vi ser med en gang at vi får et ledd som inneholder x^2y som gir oss at:

$$\underline{\underline{f_X(x) \cdot f_Y(y) \neq f(x, y)}} \tag{46}$$

Siden kriteriet ikke er oppfylt, har vi at X og Y er avhengige stokastiske variable.

d)

Vi har nå fått to nye kontinuerlige stokastiske variable $U = X + Y$ og $V = X$. De nye grensene blir da $v \leq u \leq v + 2$ og $0 \leq v \leq 2$. Den simultane tettheten blir da:

$$f(u,v) = \begin{cases} \frac{3}{28}vu & \text{når } v \leq u \leq v+2 \text{ og } 0 \leq v \leq 2 \\ 0 & \text{ellers} \end{cases} \tag{47}$$

Så finner vi den marginale sannsynlighetstettheten til U på samme måte som over:

$$f_U(u) = \frac{3}{28}\int_0^2 vudv \tag{48}$$

$$= \frac{3}{56}\left[v^2u\right]_{v=0}^{v=2} \tag{49}$$

$$= \frac{3}{56} \cdot 4u = \underline{\frac{3}{14}u} \tag{50}$$

som gir oss:

$$f_U(u) = \underline{\underline{\begin{cases} \frac{3}{14}u & \text{når } v \leq u \leq v+2 \\ 0 & \text{ellers} \end{cases}}} \tag{51}$$

Oppgave 3

a)

For å finne medianen til en kontinuerlig stokatisk variabel, må vi finne 50-persentilen, altså $p = F[\eta(p)] = 0.50 = F[\eta(0.50)]$. En lognormalfordeling har den kumulative fordelingen gitt som:

$$F(x;\mu,\sigma) = \Phi\left(\frac{\ln(x)-\mu}{\sigma}\right) \tag{52}$$

Verdier til Φ finnes i tabell A.3 bak i pensumboka som vi kan slå opp. Vi vil ha uttrykket over til å være lik 0.5 og ser fra tabellen at argumentet må være lik 0. Vi får dermed at:

$$\frac{\ln(x)-\mu}{\sigma} = 0 \quad \Rightarrow \quad x = \underline{\underline{e^{\mu}}} \tag{53}$$

For å finne forventningsverdien til X bruker vi først den momentgenererende funksjonen til Y som vi kjenner som følgende:

$$M_Y(t) = E(e^{tY}) = e^{\mu t + \sigma^2 t^2/2} \tag{54}$$

Med dette kan vi enkelt finne forventningsverdien til X ved å sette $t = 1$ i den momentgenererende funksjonen på følgende vis:

$$E(X) = E(e^Y) = M_Y(1) = e^{\mu+\sigma^2/2} = e^{\mu} \cdot e^{\sigma^2/2} = \underline{\underline{\eta \cdot e^{\sigma^2/2}}} \tag{55}$$

b)

Dersom $E(\hat{\mu}) = \mu$ for alle verdier av μ så har vi at $\hat{\mu}$ er en forventningsrett estimator for μ. Vi finner $E(\hat{\mu})$:

$$E(\hat{\mu}) = E(\overline{Y}) = E\left(\frac{1}{n}\sum_{i=1}^{n} Y_i\right) \tag{56}$$

$$= \frac{1}{n}\sum_{i=1}^{n} E(Y_i) \tag{57}$$

Siden Y_i er symmetriske om μ_i, vil også gjennomsnittet være lik medianen slik at vi videre får:

$$= \frac{1}{n}\sum_{i=1}^{n} \mu = \frac{1}{\cancel{n}}\cancel{n}\mu = \underline{\underline{\mu}} \tag{58}$$

som betyr at $\hat{\mu}$ er en forventningsrett estimator for μ. Vi har at en estimator kan skrives på følgende måte:

$$\hat{\mu} = \mu + \text{estimeringsfeil} \tag{59}$$

Vi bruker MSE for å finne denne estimeringsfeilen og siden vi har at $E(\hat{\mu}) = \mu$, får vi at $\text{MSE}(\hat{\mu}) = V(\hat{\mu}) + [E(\hat{\mu}) - \mu]^2 = V(\hat{\mu}) - [\mu - \mu]^2 = V(\hat{\mu})$, altså kun variansen. Variansen til n identiske og uavhengige fordelinger er $V(\hat{\mu}) = \sigma^2/n$ slik at vi da får:

$$\hat{\mu} = \mu + \underline{\underline{\frac{\sigma^2}{n}}} \tag{60}$$

c)

$\eta^* = e^{\hat{\mu}}$ er ikke en forventningsrett estimator for medianen fordi vi må ta hensyn til medianen til alle observable, altså for alle medianer for X_i.

d)

Vi prøver å sjekke forventingen til $\hat{\eta} = e^{\hat{\mu}-\sigma^2/2n}$ og ser om dette er lik η. Vi starter med å ta ln på begge sider av likhetstegnet:

$$\ln(E(\hat{\eta})) = \ln\Big(E\left(e^{\hat{\mu}-\frac{\sigma^2}{2n}}\right)\Big) \tag{61}$$

$$= E\left(\ln\left(e^{\hat{\mu}-\frac{\sigma^2}{2n}}\right)\right) \tag{62}$$

$$= E\left(\hat{\mu} - \frac{\sigma^2}{2n}\right) \tag{63}$$

$$= E\left(\frac{1}{n}\sum_{i=1}^{n}\ln(X_i) - \frac{\sigma^2}{2n}\right) \tag{64}$$

$$= \frac{1}{n}\sum_{i=1}^{n}\ln(E(X_i)) - \frac{\sigma^2}{2n} \tag{65}$$

Vi har at $E(X_i) = \eta e^{\sigma^2/2n}$ siden vi har n uavhengige fordelinger som vi setter inn og får videre:

$$= \frac{1}{n}\sum_{i=1}^{n}\ln\left(\eta e^{\sigma^2/2n}\right) - \frac{\sigma^2}{2n} \tag{66}$$

$$= \frac{1}{n}\sum_{i=1}^{n}\left(\ln(\eta) + \ln\left(e^{\sigma^2/2n}\right)\right) - \frac{\sigma^2}{2n} \tag{67}$$

$$= \frac{1}{n}\sum_{i=1}^{n}\left(\ln(\eta) + \frac{\sigma^2}{2n}\right) - \frac{\sigma^2}{2n} \tag{68}$$

$$= \frac{1}{\cancel{n}}\cancel{n}\left(\ln(\eta) + \frac{\cancel{\sigma^2}}{\cancel{2n}}\right) - \frac{\cancel{\sigma^2}}{\cancel{2n}} \tag{69}$$

$$= \ln(\eta) \tag{70}$$

Så tar vi eksponenten på begge sider slik at vi da får:

$$E(\hat{\eta}) = \underline{\underline{\eta}} \tag{71}$$

som betyr at $\hat{\eta}$ er en forventningsrett estimator for medianen.

e)

For en $N(\mu, \sigma^2)$ fordeling av uavhengige stokastiske variabler Y_i med observasjonene y_i der $i = 1, 2, ..., n$, har vi at $\overline{Y} \sim N(\mu, \sigma^2/n)$. Med dette følger at $Z = (\overline{Y} - \mu)/(\sigma/\sqrt{n})$ og med observasjonene y_i har vi også at $Z = (\overline{y} - \mu)/(\sigma/\sqrt{n})$. Vi kan derfor finne 95 % konfidensintervallet til medianen slik:

$$P\left(-1.96 \leq Z \leq 1.96\right) = 0.95 \tag{72}$$

$$\Rightarrow P\left(-1.96 \leq \frac{\overline{y} - \mu}{\sigma/\sqrt{n}} \leq 1.96\right) = 0.95 \tag{73}$$

$$\Rightarrow P\left(-1.96\frac{\sigma}{\sqrt{n}} \leq \overline{y} - \mu \leq 1.96\frac{\sigma}{\sqrt{n}}\right) = 0.95 \tag{74}$$

$$\Rightarrow P\left(-\overline{y} - 1.96\frac{\sigma}{\sqrt{n}} \leq -\mu \leq -\overline{y} + 1.96\frac{\sigma}{\sqrt{n}}\right) = 0.95 \tag{75}$$

$$\Rightarrow P\left(\overline{y} + 1.96\frac{\sigma}{\sqrt{n}} \geq \mu \geq \overline{y} - 1.96\frac{\sigma}{\sqrt{n}}\right) = 0.95 \tag{76}$$

$$\Rightarrow P\left(e^{\overline{y}-1.96\frac{\sigma}{\sqrt{n}}} \leq e^{\mu} \leq e^{\overline{y}+1.96\frac{\sigma}{\sqrt{n}}}\right) = 0.95 \tag{77}$$

$$\Rightarrow P\left(e^{\overline{y}-1.96\frac{\sigma}{\sqrt{n}}} \leq \eta \leq e^{\overline{y}+1.96\frac{\sigma}{\sqrt{n}}}\right) = 0.95 \tag{78}$$

Dette gir oss at 95%-konfidensintervallet for medianen er gitt ved:

$$\underline{\underline{\left[e^{\overline{y}-1.96\frac{\sigma}{\sqrt{n}}}, e^{\overline{y}+1.96\frac{\sigma}{\sqrt{n}}}\right]}} \tag{79}$$

f)

Vi får at:

$$E\left(\hat{\eta}e^{\frac{\sigma^2}{2}}\right) = e^{\frac{\sigma^2}{2}} \cdot E\left(\hat{\eta}\right) = e^{\frac{\sigma^2}{2}} \cdot \eta = \underline{\underline{E(X)}} \tag{80}$$

som beviser at $\hat{\eta}e^{\frac{\sigma^2}{2}}$ er en forventningsrett estimator for $E(X)$. Vi ser på Equation 78 igjen og ganger alle ledd inni fordelingen med $e^{\sigma^2/2}$ og får følgende:

$$P\left(e^{\overline{y}-1.96\frac{\sigma}{\sqrt{n}}} \cdot e^{\frac{\sigma^2}{2}} \leq \eta \cdot e^{\frac{\sigma^2}{2}} \leq e^{\overline{y}+1.96\frac{\sigma}{\sqrt{n}}} \cdot e^{\frac{\sigma^2}{2}}\right) = 0.95 \tag{81}$$

$$P\left(e^{\overline{y}-1.96\frac{\sigma}{\sqrt{n}}+\frac{\sigma^2}{2}} \leq E(X) \leq e^{\overline{y}+1.96\frac{\sigma}{\sqrt{n}}+\frac{\sigma^2}{2}}\right) = 0.95 \tag{82}$$

som viser til at

$$\left[e^{\overline{y}-1.96\frac{\sigma}{\sqrt{n}}+\frac{\sigma^2}{2}}, e^{\overline{y}+1.96\frac{\sigma}{\sqrt{n}}+\frac{\sigma^2}{2}}\right] \tag{83}$$

er et 95%-konfidensialintervall for $E(X)$.

g)

Når vi kjører scriptet får vi følgende verdier; $\eta^*_{n=10} = 2.160$, $\eta^*_{n=30} = 2.051$, $\tilde{\eta}_{n=10} = 2.012$ og $\tilde{\eta}_{n=30} = 2.002$. Vi ser at de genererte estimatene har en del større presisjon for $\tilde{\eta}$ ovenfor η^*. Dette gjelder for både $n = 10$ og $n = 30$ da begge disse har en relativ feil på under 1 %. Vi har tidligere funnet at $\hat{\eta}$ er en forventningsrett estimator for η og siden vi estimerer en verdi for σ og bruker denne for å finne $\tilde{\eta}$, så ser vi at vi ligger veldig nære den faktiske verdien til η. Feilen er riktignok ikke veldig stor for η^* heller men allikevel noe større.

Oppgave 4

a)

Vi finner hvor mange som drikker minst 10 liter per år ved å se på den kumulative fordelingen til X som er gitt ved følgende:

$$F(X;\mu,\sigma) = \Phi\left(\frac{\ln(X)-\mu}{\sigma}\right) \tag{84}$$

der Φ er definert som:

$$\Phi\left(z\right) = P\left(Z \leq z\right) = \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}} e^{-\frac{y^2}{2}}\, dy \tag{85}$$

hvor verdier til $\Phi(z)$ er gitt i tabeller i intervallet $z \in [-3.49, 3.49]$. Først regner vi på argumentet slik at vi kan slå opp hva den kumulative fordelingen blir. Vi får:

$$\frac{\ln(x)-\mu}{\sigma} = \frac{\ln(10)-0.7}{1.2} = \underline{1.34} \tag{86}$$

Vi slår opp i tabell A.3 i pensumboka og finner at den kumulative verdien blir:

$$\Phi(1.34) = \underline{0.9099} \tag{87}$$

Siden vi ønsker å finne ut av hvor mange norske kvinner som drikker minst 10 liter ren alkohol i året, får vi altså følgende:

$$P(X \geq 10) = 1 - P(X \leq 10) = 1 - 0.9099 = 0.0901 = \underline{\underline{9.01\%}} \tag{88}$$

b)

Vi husker at medianen $\tilde{\mu}$ er definert som 50-persentilen slik at vi har at den kumulative fordelingen blir:

$$F(\eta(0.5)) = \Phi\left(\frac{\ln(\tilde{\mu})-\mu}{\sigma}\right) = 0.5 \tag{89}$$

Ser vi på tabellen igjen så ser vi at argumentet til Φ må være lik null. Dette gir oss dermed:

$$\frac{\ln(\tilde{\mu})-\mu}{\sigma} = 0 \quad \Rightarrow \quad \tilde{\mu} = e^{\mu} = e^{0.7} = \underline{\underline{2.01}} \tag{90}$$

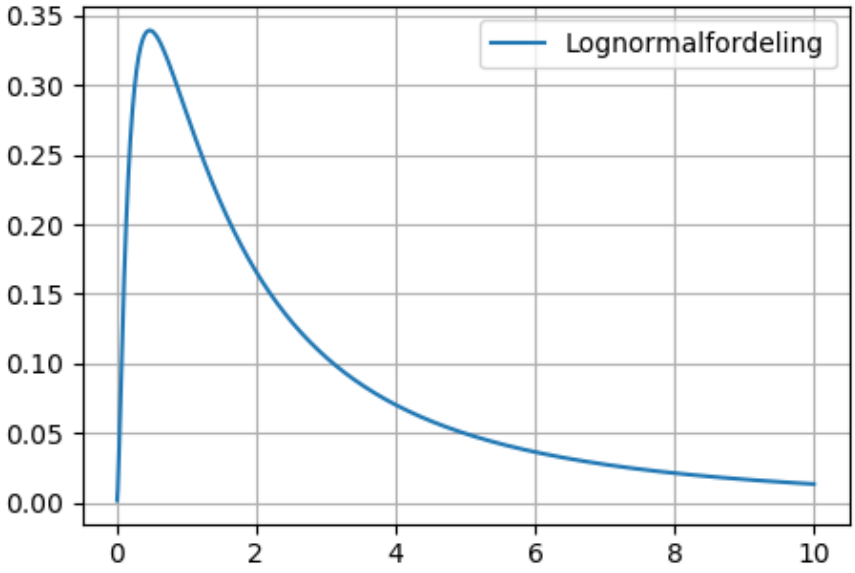
Forventningen til en lognormalfordeling X er gitt ved:

$$E(X) = e^{\mu+\frac{\sigma^2}{2}} \tag{91}$$

slik at vi da kan finne forventningsverdien med gitte verdier for μ og σ:

$$E(X) = e^{0.7+\frac{1.2^2}{2}} = \underline{\underline{4.14}} \tag{92}$$

Vi ser altså at forvetningsverdien er større enn medianen. Hvis vi ser på peaken på Figure 1, så ser vi at denne ligger omtrent på $x = 0.5$, altså alkoholinntaket per år med størst sannsynlighet. Med dette kan det virke som at medianen egner seg best i forbindelse med alkoholforbruket i løpet av et år for voksne norske kvinner med størst sannsynlighet.



Figur 1: Logormalfordeling som viser sannsynligheten for alkoholinntak for kvinner i løpet av et år for $x \in (0, 10]$ med $\mu = 0.7$ og $\sigma = 1.2$.

c)

I Python-scriptet får vi at den estimerte medianen blir $\hat{\mu} = 4.40$ og konfidensintervallet til medianen er $[3.07, 6.56]$. Den estimerte forventningen blir $\dot{\eta} = 7.74$ der $\dot{\eta} = \hat{\eta} e^{s^2/2}$ og konfidensintervallet til forventningen er $[5.39, 11.55]$.

d)

Etter å ha brukt parametrisk bootstrapping til å bestemme standardfeilen til både den estimerte medianen og den estimerte forventningen, kommer vi fram til at standardfeilene er $s_{\hat{\mu}} = 0.89$ og $s_{\dot{\eta}} = 1.99$. Merk at siden vi genererer nye bootstrap-utvalg $y_1^*, y_2^*, ..., y_B^*$, vil standardfeilen variere fra hver gang vi kjører koden.

e)

Resultatene vi fant i oppgave b) er verdier som baserer seg på gitte størrelser der vi har satt $\mu = 0.7$ og $\sigma = 1.2$. Vi har ikke gjort et forsøk på flere mennesker men heller gjort et grovt estimat for disse verdiene noe som gjør at vi ikke oppnår systematiske feil. I tillegg ser vi at både den estimerte medianen og den estimerte forventningen er større for de utvalgte folkegruppene slik vi ser c). Disse estimatene baserer seg på et forsøk av et utvalg på 30 personer som da fører til feil eller usikkerheter. Vi har funnet 95%-konfidensiallintervallet til estimatene som sier noe om området de bør ligge på med 95% for at det stemmer med virkeligheten. I tillegg fant vi standardfeil som forteller oss noe presisjonen i estimeringene vi har gjort. Med disse verdiene er det vanskelig å si nøyaktig hva årsinntaket på rent alkohol for norske kvinnelige studenter faktisk er. Vi kan derimot si med stor sikkerhet at inntaket for norske kvinnelig studenter er større enn for norske kvinner generelt i landet. Med de estimatene vi har gjort, får vi den minste estimerte medianen til å være nedre konfidensintervall minus standardfeilen, altså $\hat{\mu}_{\min} = 3.07 - 0.89 = 2.18$. Vi gjør det samme med den estimerte forventningen, $\dot{\eta}_{\min} = 5.39 - 1.99 = 3.4$. Vi ser at den minste estimerte medianen er større enn medianen vi fant i b). Den minste estimerte forventningen er mindre enn forventningen i b).

```
import numpy as np


############################
######## oppgave 3g) ########
############################
print("###### Oppgave 3g) ######")

mu = 0.7
sigma = 1.2
B = 10000

def y(n, M, S):
    y = np.zeros(n)
    for i in range(n):
        y[i] = np.random.normal(loc=M,scale=S)
    return y

def mu_hat(y_list):
    n = len(y_list)
    return (1/n) * np.sum(y_list)

def s_sqrd(y_list):
    n = len(y_list)
    y_mean = mu_hat(y_list)
    s_sqrd = np.sum((y_list-y_mean)**2) / (n-1)
    return s_sqrd

def eta_star(y_list):
    return np.exp(mu_hat(y_list))

def eta_tilde(y_list):
    n = len(y_list)
    return np.exp(mu_hat(y_list) - s_sqrd(y_list)/(2*n))

n1 = 10
n2 = 30
```

```
eta_star_1 = np.zeros(B)
eta_star_2 = np.zeros(B)
eta_tilde_1 = np.zeros(B)
eta_tilde_2 = np.zeros(B)

for i in range(B):
    y1 = y(n1, mu, sigma)
    y2 = y(n2, mu, sigma)
    eta_star_1[i] = eta_star(y1)
    eta_star_2[i] = eta_star(y2)
    eta_tilde_1[i] = eta_tilde(y1)
    eta_tilde_2[i] = eta_tilde(y2)

eta_star_1_mean = np.sum(eta_star_1) / B
eta_star_2_mean = np.sum(eta_star_2) / B
eta_tilde_1_mean = np.sum(eta_tilde_1) / B
eta_tilde_2_mean = np.sum(eta_tilde_2) / B

print("eta = e^mu =                  %.3f" % np.exp(mu))
print("eta* snitt for n=10:          %.3f" % eta_star_1_mean, \
                            "      Relativ feil:   %.2f %%" % \
                            (100*abs(eta_star_1_mean-np.exp(mu))/np.exp(mu)))
print("eta~ snitt for n=10:          %.3f" % eta_tilde_1_mean, \
                            "      Relativ feil:   %.2f %%" % \
                            (100*abs(eta_tilde_1_mean-np.exp(mu))/np.exp(mu)))
print("eta* snitt for n=30:          %.3f" % eta_star_2_mean, \
                            "      Relativ feil:   %.2f %%" % \
                            (100*abs(eta_star_2_mean-np.exp(mu))/np.exp(mu)))
print("eta~ snitt for n=30:          %.3f" % eta_tilde_2_mean, \
                            "      Relativ feil:   %.2f %%" % \
                            (100*abs(eta_tilde_2_mean-np.exp(mu))/np.exp(mu)))




##############################
######## oppgave 4c) ########
##############################
print("")
print("###### Oppgave 4c) ######")

forbruk = np.array([1.0, 3.4, 5.0, 14.4, 11.5, 8.2, 0.6, 2.7, 26.8, 3.0,
                    1.3, 20.2, 4.0, 14.0, 3.3, 1.8, 1.7, 4.6, 7.4, 7.1,
                    5.2, 23.6, 1.6, 1.1, 15.5, 3.0, 1.9, 4.2, 27.4, 1.5])
ln_forbruk = np.log(forbruk)
N = len(forbruk)

median_estimat = eta_tilde(ln_forbruk)
print("Estimert median:                        %.2f" % median_estimat)

y_snitt = mu_hat(ln_forbruk)
s = np.sqrt(s_sqrd(ln_forbruk))

nedre_ki_median = np.exp(y_snitt - 1.96*s/np.sqrt(N))
ovre_ki_median = np.exp(y_snitt + 1.96*s/np.sqrt(N))
print("Konfidensintervall for medianen:        [%.2f, %.2f]" % \
                                    (nedre_ki_median, ovre_ki_median))


forv_estimat = median_estimat * np.exp(s**2/2)
print("Estimert forventning:                   %.2f" % forv_estimat)

nedre_ki_forv = np.exp(y_snitt - 1.96*s/np.sqrt(N) + s**2/2)
ovre_ki_forv = np.exp(y_snitt + 1.96*s/np.sqrt(N) + s**2/2)
print("Konfidensintervall for forventningen:   [%.2f, %.2f]" % \
                                    (nedre_ki_forv, ovre_ki_forv))
```

```
##############################
######## oppgave 4d) ########
##############################
print("")
print("###### Oppgave 4d) ######")

median_boot = np.zeros(B)
forv_boot = np.zeros(B)

for i in range(B):
    y_boot = y(N, y_snitt, s)
    s_boot = np.sqrt(s_sqrd(y_boot))
    median_boot[i] = eta_tilde(y_boot)
    forv_boot[i] = median_boot[i] * np.exp(s_boot**2/2)

median_boot_mean = mu_hat(median_boot)
median_std = np.sqrt(1/(B-1) * np.sum((median_boot - median_boot_mean)**2))
forv_boot_mean = mu_hat(forv_boot)
forv_std = np.sqrt(1/(B-1) * np.sum((forv_boot - forv_boot_mean)**2))

print("Standardfeil til median:                %.2f" % median_std)
print("Standardfeil til forventning:           %.2f" % forv_std)



"""
Resultater:
###### Oppgave 3g) ######
eta = e^mu =                2.014
eta* snitt for n=10:        2.160       Relativ feil:   7.28 %
eta~ snitt for n=10:        2.012       Relativ feil:   0.09 %
eta* snitt for n=30:        2.051       Relativ feil:   1.86 %
eta~ snitt for n=30:        2.002       Relativ feil:   0.56 %

###### Oppgave 4c) ######
Estimert median:                        4.40
Konfidensintervall for medianen:        [3.07, 6.56]
Estimert forventning:                   7.74
Konfidensintervall for forventningen:   [5.39, 11.55]

###### Oppgave 4d) ######
Standardfeil til median:                0.88
Standardfeil til forventning:           1.97
"""
```