Løsningsforslag til hjemmeeksamen STK1100 våren 2020

Oppgave 1

Vi har to tester T1 og T2 med følgende karakteristika:

	Sensitivitet	Spesifisitet
T1	95.5%	98.0%
T2	94.2%	99.8%

Sensitivitet er sannsynligheten for at testen gir et positivt resultat (testen indikerer antistoffer i blodet), gitt at man faktisk har antistoffer i blodet. *Spesifisitet* er sannsynligheten for at testen gir et negativt resultat (testen indikerer ikke antistoffer i blodet), gitt at man ikke har antistoffer i blodet. Vi vil i det følgende se på en populasjon der 1% har slike antistoffer.

Hvis vi definerer hendelsene

B= positivt resultat

A= har antistoffer i blodet

kan vi bruke opplysningene i oppgaveteksten til å finne følgende betingede sannsynligheter:

For test T1:

$P(B|A)$= sensitivitet $= 0.955$

$P(B'|A')$= spesifisitet $= 0.980$

For test T2:

$P(B|A)$= sensitivitet $= 0.942$

$P(B'|A')$= spesifisitet $= 0.998$

Vi har også at $P(A) = 0.01$ og dermed at $P(A') = 1 - P(A) = 0.99$.

a) Vi skal finne sannsynligheten for at en tilfeldig valgt testperson får et positivt resultat av test T1, dvs. $P(B)$. Fra setningen om total sannsynlighet har vi at

$$P(B) = P(B|A) \cdot P(A) + P(B|A') \cdot P(A') = 0.955 \cdot 0.01 + 0.02 \cdot 0.99 = 0.0294,$$

der vi har brukt at $P(B|A') = 1 - (B'|A') = 1 - 0.980 = 0.02$.

b) For å beregne sannsynligheten for at en tilfeldig valgt testperson faktisk har antistoffer i blodet, gitt en positiv test (T1), bruker vi Bayes setning og setter inn:

$$P(A|B) = \frac{P(B|A) \cdot P(A)}{P(B)} = \frac{0.955 \cdot 0.01}{0.0294} = 0.325.$$

Det betyr at det bare er litt over 30% sannsynlighet for å ha antistoffer i blodet, gitt en positiv test, noe som er svært lavt. T1 er ikke en spesielt trygg test.

c) For å beregn sannsynligheten for at en tilfeldig testperson faktisk har antistoffer i blodet, gitt en positiv test T2, må vi gjøre som i a og b med tallene for T2. Vi finner $P(B) = 0.942 \cdot 0.01 + 0.002 \cdot 0.99 = 0.0114$ og

$$P(A|B) = \frac{P(B|A) \cdot P(A)}{P(B)} = \frac{0.942 \cdot 0.01}{0.0114} = 0.826.$$

Nå har vi altså over 80% sannsynlighet for å ha antistoffer i blodet, gitt en positiv test. Dette er betraktelig høyere enn for T1, og skyldes den økte spesifisiteten til T2.

d) Sannsynligheten for å ikke ha antistoffer i blodet, gitt at man har testet positivt, finner man enkelt som

$$P(A'|B) = 1 - P(A|B).$$

For T1 blir sannsynligheten $1-0.325 = 0.675$, mens for T2 blir den $1-0.826 = 0.174$. Det er klart at begge disse er høye, men T2 er svært mye bedre enn T1, som kommentert ovenfor.

e) For at sannsynligheten i punkt d skal bli mindre enn 5% for test T2, må vi øke spesifisiteten $P(B'|A')$ ytterligere. Vi tar utgangspunkt i at vi ønsker

$$\begin{aligned}
1 - P(A|B) &\leq 0.05 \\
P(A|B) &\geq 0.95 \\
\frac{P(B|A) \cdot P(A)}{P(B)} &\geq 0.95 \\
\frac{P(B|A) \cdot P(A)}{P(B|A) \cdot P(A) + P(B|A') \cdot P(A')} &\geq 0.95 \\
\frac{P(B|A) \cdot P(A)}{P(B|A) \cdot P(A) + (1 - P(B'|A')) \cdot P(A')} &\geq 0.95
\end{aligned}$$

Løser denne med hensyn på $P(B'|A')$ og setter inn alle de kjente sannsynlighetene, og får

$$P(B'|A') \geq 1 - \frac{P(B|A) \cdot P(A)}{P(A')} \cdot \frac{0.05}{0.95} = 1 - \frac{0.942 \cdot 0.01}{0.99} \cdot \frac{0.05}{0.95} = 0.9995.$$

Spesifisiteten til T2 må altså øke til minst 99.95% for at sannsynligheten for at en tilfeldig testperson får falsk positiv test skal bli mindre enn 5%.

f) $n = 10$ uavhengige, tilfeldig valgte personer testes med testen T1, og alle tester negativt. Vi skal finne sannsynligheten for at minst én av disse er en falsk negativ.

$$
\begin{aligned}
& P(\text{minst en av de 10 negative er falsk negativ}) \\
= \; & 1 - P(\text{ingen av de 10 negative er falsk negativ}) \\
= \; & 1 - P(\text{alle de 10 negative testene er sanne negative}) \\
= \; & 1 - (P(\text{negativ test er sann negativ}))^{10} \\
= \; & 1 - (P(A'|B'))^{10} = 1 - 0.9996^{10} = 0.004,
\end{aligned}
$$

der vi har funnet $P(A'|B')$ ved

$$P(A'|B') = P(B'|A')P(A')/P(B') = 0.98 \cdot 0.99/(1 - 0.0294) = 0.9996.$$

Kommentar: Denne oppgaven er laget i midten av mai. Den er basert på realistiske tall og opplysninger, men mye kan være endret på få uker.

Oppgave 2

De to kontinuerlige stokastiske variablene X og Y har simultan sannsynlighetstetthet

$$f(x,y) = \begin{cases} kx(x+y) & \text{når } \; 0 \le x \le 2 \; \text{ og } \; 0 \le y \le 2 \\ 0 & \text{ellers} \end{cases}$$

a) Vi finner konstanten k ved å bruke at $\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(x,y)dxdy = 1$:

$$
\begin{aligned}
k\int_0^2\int_0^2 x(x+y)dxdy &= k\int_0^2\int_0^2 (x^2+xy)dydx \\
&= k\int_0^2 \left[x^2y + \frac{1}{2}xy^2\right]_{y=0}^{2} dx \\
&= k\int_0^2 (2x^2+2x)dx \\
&= k\left[\frac{2}{3}x^3 + x^2\right]_{x=0}^{2} \\
&= k(\frac{16}{3} + \frac{12}{3}) = \frac{k\cdot 28}{3}
\end{aligned}
$$

som blir $= 1$ hvis konstanten $k = \frac{3}{28}$.

b) Vi finner at

$$\begin{aligned} P(Y \geq X) &= \frac{3}{28}\int_0^2 \int_0^y (x^2+xy)dxdy \\ &= \frac{3}{28}\int_0^2 \left[\frac{1}{3}x^3+\frac{1}{2}x^2y\right]_{x=0}^{y} dy \\ &= \frac{3}{28}\int_0^2 (\frac{1}{3}y^3+\frac{1}{2}y^3)dy \\ &= \frac{3}{28}\int_0^2 \frac{5}{6}y^3dy = \frac{3}{28}\cdot\frac{5}{6}\left[\frac{1}{4}y^4\right]_0^2 = \frac{5}{14} \end{aligned} \tag{1}$$

c) Når $0 \leq y \leq 2$ har vi at

$$f_Y(y) = \int_{-\infty}^{\infty} f(x,y)dx = \frac{3}{28}\int_0^2 (x^2+xy)dx = \frac{3}{28}\left[\frac{1}{3}x^3+\frac{1}{2}x^2y\right]_{x=0}^{2} = \frac{3}{28}(\frac{8}{3}+2y),$$

slik at vi får

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

Når $0 \leq x \leq 2$ har vi at

$$f_X(x) = \int_{-\infty}^{\infty} f(x,y)dy = \frac{3}{28}\int_0^2 (x^2+xy)dy = \frac{3}{28}\left[x^2y+\frac{1}{2}xy^2\right]_{y=0}^{2} = \frac{3}{28}(2x^2+2x),$$

slik at vi får

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

X og Y kan ikke være uavhengige, siden vi enkelt ser fra formlene at $f(x,y) \neq f_X(x)\cdot f_Y(y)$.

d) Vi har to nye variabler $U = X+Y$ og $V = X$. Skal finne den marginale tettheten $g_U(u)$ til U, via simultantettheten $g(u,v)$ til U og V. Først skriver vi de opprinnelige variablene X og Y uttrykt ved U og V:

$$X = V \quad \text{slik at} \quad v_1(u,v) = v$$

$$Y = U - V \quad \text{slik at} \quad v_2(u,v) = u - v$$

Vi finner Jacobi-matrisen

$$M = \begin{bmatrix} \frac{\partial v_1}{\partial u} & \frac{\partial v_1}{\partial v} \\ \frac{\partial v_2}{\partial u} & \frac{\partial v_2}{\partial v} \end{bmatrix} = \begin{bmatrix} 0 & 1 \\ 1 & -1 \end{bmatrix}$$

og får derfor $\det(M) = -1$ og $|\det(M)| = 1$. Fra områdene $0 \leq x \leq 2$ og $0 \leq y \leq 2$ finner vi områdene $0 \leq v \leq 2$ og $v \leq u \leq v+2$, og får derfor

$$g(u,v) = f(v, u-v) \cdot 1 = \begin{cases} \frac{3}{28}uv & \text{når } 0 \leq v \leq 2 \text{ og } v \leq u \leq v+2 \\ 0 & \text{ellers.} \end{cases}$$

Vi finner den marginale sannsynlighetstettheten til U ved

$$g_U(u) = \int_{-\infty}^{\infty} g(u,v)dv.$$

Det kan lønne seg å lage en figur for å finne området å integrere over. For $0 \leq u \leq 2$ har vi

$$g_U(u) = \int_0^u \frac{3}{28}uvdv = \left[\frac{3}{28}u\frac{1}{2}v^2\right]_{v=0}^{u} = \frac{3}{28} \cdot \frac{1}{2}u^3.$$

For $2 < u \leq 4$ har vi

$$g_U(u) = \int_{u-2}^{2} \frac{3}{28}uvdv = \left[\frac{3}{28}u\frac{1}{2}v^2\right]_{v=u-2}^{2} = \frac{3}{28} \cdot \frac{1}{2}(4u^2 - u^3).$$

Vi får derfor at sannsynlighetstettheten til variabelen $U = X + Y$ blir

$$g_U(u) = \begin{cases} \frac{3}{56}u^3 & \text{når } 0 \leq u \leq 2 \\ \frac{3}{56}u^2(4-u) & \text{når } 2 < u \leq 4 \\ 0 & \text{ellers.} \end{cases}$$

Oppgave 3

a) Vi har at $X = e^Y$, der $Y \sim N(\mu, \sigma^2)$. Medianen η er gitt ved

$$P(X \leq \eta) = 0.50.$$

Vi har nå at

$$\begin{aligned} P(X \leq \eta) &= P(e^Y \leq \eta) \\ &= P(Y \leq \ln \eta) \\ &= P\left(\frac{Y-\mu}{\sigma} \leq \frac{\ln\eta - \mu}{\sigma}\right) \\ &= P\left(Z \leq \frac{\ln\eta - \mu}{\sigma}\right), \end{aligned}$$

der $Z = (Y-\mu)/\sigma$ er standardnormalfordelt. Vi vet at $P(Z \leq 0) = 0.50$. Vi får dermed at $P(X \leq \eta) = 0.50$ hvis vi bestemmer η av ligningen

$$\frac{\ln\eta - \mu}{\sigma} = 0.$$

Det gir $\ln \eta = \mu$, og det følger at medianen er gitt ved $\eta = e^{\mu}$.

Den momentgenererende funksjonen til Y er $M_Y(t) = E(e^{tY}) = e^{\mu t + \sigma^2 t^2/2}$. Dermed har vi at

$$E(X) = E(e^Y) = M_Y(1) = e^{\mu + \sigma^2/2} = e^{\mu} e^{\sigma^2/2} = \eta\, e^{\sigma^2/2}.$$

b) Vi antar nå at $X_1, \ldots, X_n$ er uavhengige og lognormalt fordelte, der μ er en ukjent parameter mens σ er kjent. Videre setter vi $Y_i = \ln(X_i)$ for $i = 1, \ldots, n$. Da er Y_i-ene uavhengige og $N(\mu, \sigma^2)$-fordelte.

Vi vil først estimere μ og ser på estimatoren

$$\widehat{\mu} = \overline{Y} = \frac{1}{n} \sum_{i=1}^{n} Y_i = \frac{1}{n} \sum_{i=1}^{n} \ln(X_i).$$

Vi har at $E(Y_i) = \mu$ og $V(Y_i) = \sigma^2$. Det følger at

$$E(\widehat{\mu}) = E\left(\frac{1}{n} \sum_{i=1}^{n} Y_i \right) = \frac{1}{n} \sum_{i=1}^{n} E(Y_i) = \frac{1}{n} \cdot n\mu = \mu,$$

så $\widehat{\mu}$ er forventningsrett. Videre er

$$V(\widehat{\mu}) = V\left(\frac{1}{n} \sum_{i=1}^{n} Y_i \right) = \left(\frac{1}{n} \right)^2 \sum_{i=1}^{n} V(Y_i) = \frac{1}{n^2} \cdot n\sigma^2 = \frac{\sigma^2}{n}.$$

Vi vet at en lineærkombinasjon av uavhengige og normalfordelte variabler selv er normalfordelt (jf. side 309 i læreboka). Det følger at $\widehat{\mu} \sim N(\mu, \sigma^2/n)$.

Kommentar: Det er kjent at hvis $Y_1, \ldots, Y_n$ er uavhengige og alle er $N(\mu, \sigma^2)$-fordelte, så er $\overline{Y} \sim N(\mu, \sigma^2/n)$; jf. side 297 i læreboka. Det er i orden om en student bare henviser til dette resultatet uten å bestemme forventning og varians slik vi har gjort over.

c) Vi ser så på estimering av medianen $\eta = e^{\mu}$ i den lognormale fordelingen. En mulig estimator er $\eta^* = e^{\widehat{\mu}}$. Vi har at $\eta^* = h(\widehat{\mu})$, der $h(u) = e^u$. For en slik ikke-lineær funksjon har vi at

$$E(\eta^*) = E\left[h(\widehat{\mu})\right] \neq h\left[E(\widehat{\mu})\right] = h(\mu) = e^{\mu} = \eta.$$

Derfor er ikke η^* en forventningsrett estimator for η.

Kommentar: Vi har at $e^u > 1 + u$ når $u \neq 0$. Av det følger det at

$$E(e^{\widehat{\mu}}) = e^{\mu}\, E(e^{\widehat{\mu} - \mu}) > e^{\mu} E\left[1 + (\widehat{\mu} - \mu)\right] = e^{\mu}\left[1 + E(\widehat{\mu}) - \mu\right] = e^{\mu}.$$

Det forventes ikke at studentene har med et slikt ressonement.

d) Fra punkt b har vi at $\widehat{\mu} \sim N(\mu, \sigma^2/n)$. Den momentgenererende funksjonen til $\widehat{\mu}$ er derfor gitt ved

$$M_{\widehat{\mu}}(t) = E(e^{t\widehat{\mu}}) = e^{\mu t + (\sigma^2/n)\, t^2/2}.$$

Av dette får vi at ‘

$$E(e^{\widehat{\mu}}) = M_{\widehat{\mu}}(1) = e^{\mu+\sigma^2/(2n)}.$$

Det følger at

$$E(\widehat{\eta}) = E\left(e^{\widehat{\mu}-\sigma^2/(2n)}\right) = e^{-\sigma^2/(2n)}E(e^{\widehat{\mu}}) = e^{-\sigma^2/(2n)}e^{\mu+\sigma^2/(2n)} = e^{\mu} = \eta,$$

så $\widehat{\eta}$ er en forventningsrett estimator for medianen η.

e) Vi har at $\overline{Y} \sim N(\mu, \sigma^2/n)$. Ved å standardisere får vi at

$$Z = \frac{\overline{Y} - E(\overline{Y})}{\sqrt{V(\overline{Y})}} = \frac{\overline{Y} - \mu}{\sigma/\sqrt{n}} \sim N(0,1).$$

Vi har at $P(-1.96 \leq Z \leq 1.96) = 0.95$. Derfor er

$$P\left(-1.96 \leq \frac{\overline{Y} - \mu}{\sigma/\sqrt{n}} \leq 1.96\right) = 0.95.$$

Ved å løse ulikhetene gir dette

$$P\left(\overline{Y} - 1.96\,\sigma/\sqrt{n} \leq \mu \leq \overline{Y} + 1.96\,\sigma/\sqrt{n}\right) = 0.95.$$

Funksjonen $h(u) = e^u$ er strengt voksende. Derfor har vi at

$$P\left(e^{\overline{Y}-1.96\,\sigma/\sqrt{n}} \leq e^{\mu} \leq e^{\overline{Y}+1.96\,\sigma/\sqrt{n}}\right) = 0.95. \tag{2}$$

Det viser at

$$\left[e^{\overline{y}-1.96\,\sigma/\sqrt{n}},\, e^{\overline{y}+1.96\,\sigma/\sqrt{n}}\right]$$

er et 95% konfidensintervall for medianen $\eta = e^{\mu}$.

f) Vi ser på estimatoren $\widehat{\eta}\, e^{\sigma^2/2}$ for forventningen $E(X) = \eta\, e^{\sigma^2/2}$. Vi har at

$$E(\widehat{\eta}\, e^{\sigma^2/2}) = e^{\sigma^2/2}E(\widehat{\eta}) = e^{\sigma^2/2}\eta = E(X),$$

så estimatoren er forventningsrett.

Ved å gange alle leddene i ulikehetene i (2) med $e^{\sigma^2/2}$ finner vi videre at

$$P\left(e^{\sigma^2/2}e^{\overline{Y}-1.96\,\sigma/\sqrt{n}} \leq e^{\sigma^2/2}e^{\mu} \leq e^{\sigma^2/2}e^{\overline{Y}+1.96\,\sigma/\sqrt{n}}\right) = 0.95.$$

Dermed har vi at

$$P\left(e^{\sigma^2/2}e^{\overline{Y}-1.96\,\sigma/\sqrt{n}} \leq E(X) \leq e^{\sigma^2/2}e^{\overline{Y}+1.96\,\sigma/\sqrt{n}}\right) = 0.95,$$

så et 95% konfidensintevall for for forventningen $E(X)$ er

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

g) Vi vil til slutt se på situasjonen der σ er en ukjent parameter. Vi vil estimere medianen $\eta = e^{\mu}$, og vi har de to estimatorene $\eta^* = e^{\widehat{\mu}}$ og $\widetilde{\eta} = e^{\widehat{\mu} - S^2/(2n)}$, der $\widehat{\mu} = \overline{Y}$ og $S^2 = \sum_{i=1}^{n}(Y_i - \overline{Y})^2/(n-1)$. Ved stokastisk simulering vil vi sammenligne de to estimatorene for situasjonen der $\mu = 0.7$ og $\sigma = 1.2$. Da er $\eta = e^{\mu} = e^{0.7} = 2.01$.

Vi ser først på situasjonen der $n = 10$. Vi genererer $B = 10\,000$ verdier av estimatorene og beregner gjennomsnittsverdi og empirisk standardavvik av de B verdiene ved Python kommandoene:

```
import numpy as np
import scipy.stats as stats
n=10
B=10000
my=0.7
sigma=1.2
eta_star = []
eta_tilde = []
for _ in range(B):
    y=stats.norm.rvs(my,sigma,size=n)
    y_mean=np.mean(y)
    y_s2=np.var(y,ddof=1)
    eta_star.append(np.exp(y_mean))
    eta_tilde.append(np.exp(y_mean-y_s2/(2*n)))
print(np.mean(eta_star),np.std(eta_star,ddof=1))
print(np.mean(eta_tilde),np.std(eta_tilde,ddof=1))
```

Vi finner at gjennomsnittsverdien og standardavviket av verdiene for η^* er[1]

$$\overline{\eta^*} = 2.17 \qquad \text{og} \qquad s_{\eta^*} = 0.87,$$

mens vi for $\widetilde{\eta}$ får

$$\overline{\widetilde{\eta}} = 2.02 \qquad \text{og} \qquad s_{\widetilde{\eta}} = 0.81.$$

Vi gjentar beregningene for $n = 30$ og finner da

$$\overline{\eta^*} = 2.06, \qquad s_{\eta^*} = 0.46, \qquad \overline{\widetilde{\eta}} = 2.02, \qquad s_{\widetilde{\eta}} = 0.45$$

Vi ser at for $n = 10$ er gjennomsnittet av η^*-verdiene lik 2.17 som er en del større enn medianen $\eta = 2.01$. Men gjennomsnittet av verdiene av $\widetilde{\eta}$ er 2.02 som er veldig nær medianen. Vi ser også at $s_{\widetilde{\eta}}$ er mindre enn s_{η^*}, så $\widetilde{\eta}$ er den beste estimatoren av de to. Også for $n = 30$ er $\widetilde{\eta}$ den beste estimatoren, men nå er det ikke så stor forskjell på de to estimatorene.

[1] Merk at resultatene kan variere litt fra en simulering til en annen, så dine resultater kan avvike litt fra de som er gitt her.

Oppgave 4

a) Andelen (ikke-avholdende) norske kvinner som drikker minst 10 liter ren alkohol per år, svarer til sannsynligheten $P(X \geq 10)$. Som i oppgave 3 setter vi $Y = \ln(X)$. Vi har at $Y \sim N(\mu, \sigma^2)$ der $\mu = 0.7$ og $\sigma = 1.2$. Ved å bruke at $Z = (Y - \mu)/\sigma \sim N(0, 1)$ og tabell over standardnormalfordelingen, finner vi at

$$\begin{aligned} P(X \geq 10) &= P(e^Y \geq 10) \\ &= P(Y \geq \ln 10) \\ &= P\left(\frac{Y-\mu}{\sigma} \geq \frac{\ln 10 - \mu}{\sigma}\right) \\ &= P\left(Z \geq \frac{\ln 10 - 0.7}{1.2}\right) \\ &= 1 - P\left(Z \leq 1.34\right) \\ &= 1 - 0.91 = 0.09 \end{aligned}$$

Det betyr at 9% av voksne norske kvinner drikker minst 10 liter ren alkohol per år.

Kommentar: Vi kan også bestemme sannsynligheten direkte ved Python kommandoen `1-stats.norm.cdf(np.log(10),0.7,1.2)`.

b)Ved resultatet i oppgave 3a har vi at median alkoholforbruk er

$$\eta = e^{\mu} = e^{0.7} = 2.0,$$

og at forventet alkoholforbruk er

$$E(X) = \eta e^{\sigma^2/2} = e^{\mu+\sigma^2/2} = e^{0.7+1.2^2/2} = 4.1.$$

At forventningsverdien er 4.1, betyr at det gjennomsnittlige forbruket av alkohol blant ikke-avholdende norske kvinner er 4.1 liter ren alkohol per år. At medianen er 2.0, betyr at halvparten av ikke-avholdende norske kvinner drikker mindre enn 2.0 liter ren alkohol per år, og halvparten drikker mer enn det. Siden gjennomsnittsverdien blir trukket opp av de som drikker mye, er medianen best egnet til å beskrive alkoholforbruket for en "typisk" ikke-avholdende norsk kvinne.

c) Vi leser inn de 30 observasjonene og finner estimater for median og forventningvverdi ved Python kommandoene:

```
n=30
x=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])
y=np.log(x)
y_mean=np.mean(y)
y_s=np.std(y,ddof=1)
eta_tilde=np.exp(y_mean-y_s**2/(2*n))
```

```
eta_low=np.exp(y_mean-1.96*y_s/np.sqrt(n))
eta_up=np.exp(y_mean+1.96*y_s/np.sqrt(n))
EX_hatt=eta_tilde*np.exp(y_s**2/2)
EX_low=eta_low*np.exp(y_s**2/2)
EX_up=eta_up*np.exp(y_s**2/2)
print(eta_tilde,eta_low,eta_up)
print(EX_hatt,EX_low,EX_up)
```

Vi finner at estimatet for median alkoholforbruk i gruppen av studenter er 4.40 liter per år og at et 95% konfidensintervall for medianen er fra 3.07 liter per år til 6.56 liter per år. For forventningen blir estimater 7.74 liter per år med et 95% konfidensintervall fra 5.39 liter per år til 11.55 liter per år.

Kommentar: For å estimere det empiriske standardavviket til logaritmen av alkoholdforbrukene, har vi over brukt kommandoen `np.std(y,ddof=1)`. Her gir opsjonen `ddof=1` at vi deler på $n-1$ når vi beregner det empiriske standardavviket; jf. formel (1) i oppgaveteksten. Hvis vi ikke har med denne opsjonen får vi estimatet der det deles på n i stedet for $n-1$. Mange studenter vil nok ikke ha med opsjonen `ddof=1` siden foreleserne inntil nylig ikke har vært oppmerksom på denne opsjonen. Estimatet for medianen vil da fortsatt være 4.40, men konfidensintervallet vil gå fra 3.08 til 6.52. For forventningen vil estimatet bli 7.60 med et konfidensintervall fra 5.33 til 11.26. Disse svarene skal også regnes som riktige.

d) Vi bruker så parametrisk bootstrap til å bestemme standardfeilene til estimatene i forrige punkt. Python kode:

```
B=10000
eta_tilde__vec = []
EX_hatt_vec = []
for _ in range(B):
    ystar=stats.norm.rvs(y_mean,y_s,size=n)
    ystar_mean=np.mean(ystar)
    ystar_s=np.std(ystar,ddof=1)
    eta_tilde__vec.append(np.exp(ystar_mean-ystar_s**2/(2*n)))
    EX_hatt_vec.append(np.exp(ystar_mean-ystar_s**2/(2*n))*np.exp(ystar_s**2/2))
print(np.std(eta_tilde__vec,ddof=1),np.std(EX_hatt_vec,ddof=1))
```

Vi finner at estimert standardfeil er 0.87 for estimatet av medianen og at estimert standardfeil er 1.89 for estimatet av forventningsverdien[2].

e) For ikke-avholdende, voksne norsk kvinner fant vi i punkt b at median alkoholforbruk er 2.0 liter per år og forventet alkoholforbruk er 4.1 liter per år. For den aktuelle gruppen av kvinnelige studenter er estimatet for medianen 4.40 liter per år og estimatet for forventningsverdien er 7.74 liter per år. Estimert medianforbruk for studentgruppen er altså mere en dobbelt så stort som forbruket blant voksne norske kvinner, og estimert forventet forbruk er nesten dobbelt så stort. Vi merker

[2]Merk at resultatene kan variere litt fra en bootstraping til en annen, så dine resultater kan avvike litt fra de som er gitt her.

oss videre at medianforbruket og forventet forbruk blant voksne kvinner ikke er inneholdt i de tilsvarende konfidensintervallene for studentgruppen. Vi kan derfor være rimelig sikre på at den aktuelle gruppen av kvinnelige studenter har et høyere alkoholforbruk en tilfellet er for voksne norske kvinner.