Eksamen STK-1100

Våren 2020

9. juni 2020

Oppgave 1:

a.)

Finn sannsynligheten for at en tilfeldig valgt testperson får et positivt resultat av test T1.

Løsning:

Vi har at total sannsynlighet for en positiv test er gitt ved

$$P(T_1) = \sum_{i=1}^{k} P(T_1|A_i)P(A_i)$$

som i vårt tilfelle gir

$$P(T_1) = P(T_1|A)P(A) + P(T_1|\overline{A})P(\overline{A})$$

der $P(T_1|A)P(A) = 0.955 \cdot 0.01$ og $P(T_1|\overline{A})P(\overline{A}) = 0.02 \cdot 0.99$. Da har vi totalt

$$\begin{aligned} P(T_1) &= 0.955 \cdot 0.01 + 0.02 \cdot 0.99 \\ &= \underline{\underline{0.02935 \approx 2.935\%}} \end{aligned}$$

b.)

Beregn sannsynligheten for at en tilfeldig valgt testperson faktisk har antistoffer i blodet, gitt en positiv test (T1). Kommenter resultatet.

Løsning:

Vi har at sannsynligheten for faktisk å ha antistoffer gitt en positiv test (T1) er.

$$
\begin{aligned}
P(A|T_1) &= \frac{P(T_1|A)P(A)}{\sum_{i=1}^{k} P(T_1|A_i)P(A_i)} \\
&= \frac{0.955 \cdot 0.01}{P(T_1|A)P(A) + P(T_1|\overline{A})P(\overline{A})} \\
&= \frac{0.955 \cdot 0.01}{0.02935} \\
&\approx \underline{\underline{0.325 = 32.5\%}}
\end{aligned}
$$

Hvilket betyr at det kun er 32.5% sjanse for at man faktisk har antistoffer i blodet når testen er positiv, som er overraskende usikkert med tanke på at sensitivteten og spesifisiten er meget presis og lite feilbar.

c.)

Test T2 har noe høyere spesifisitet enn T1, men lavere sensitivitet. Under samme forutsetninger som tidligere, beregn sannsynligheten for at en tilfeldig testperson faktisk har antistoffer i blodet, gitt en positiv test T2. Sammenlign med T1 og kommenter.

Løsning:

Bruker vi samme ressonement som i oppgave b.) har vi at sannsynligheten er

$$
\begin{aligned}
P(A|T_2) &= \frac{P(T_2|A)P(A)}{\sum_{i=1}^{k} P(T_2|A_i)P(A_i)} \\
&= \frac{0.942 \cdot 0.01}{0.942 \cdot 0.01 + 0.99 \cdot 0.002} \\
&\approx \underline{\underline{0.826 = 82.6\%}}
\end{aligned}
$$

som i motsetning til T1 er langt mer sikkert. Dette betyr kort at å øke spesifisiteten og minke sensitiviteten med få prosent gir et stort utslag i sikkerhet av en slik testing.

d.)

Det kan skape problemer om individer som har testet positivt, egentlig ikke har antistoffer i blodet, da smittefaren for dem vil være stor. Hva er sannsynligheten for å ikke ha antistoffer i blodet, gitt at man har testet positivt? Finn denne sannsynligheten for begge testene og kommenter.

Løsning:

Sannsynligheten for komplementet til løsningen i b.) og c.) er gitt ved

$$\begin{aligned} P(\overline{A}|T_1) &= \frac{P(T_1|\overline{A})P(\overline{A})}{\sum_{i=0}^{k} P(T_1|\overline{A_i})P(\overline{A_i})} \\ &= \frac{0.02 \cdot 0.99}{0.02 \cdot 0.99 + 0.955 \cdot 0.01} \\ &= \underline{\underline{0.6746 \approx 67.5\%}} \end{aligned}$$

og

$$\begin{aligned} P(\overline{A}|T_2) &= \frac{P(T_2|\overline{A})P(\overline{A})}{\sum_{i=1}^{k} P(T_2|\overline{A_i})P(\overline{A_i})} \\ &= \frac{0.002 \cdot 0.99}{0.002 \cdot 0.99 + 0.942 \cdot 0.01} \\ &= \underline{\underline{0.1736 \approx 17.4\%}} \end{aligned}$$

som betyr at sannsynligheten for feil i testing av om en har antistoffer eller ei i blodet er desidert bedre ved test (T2) enn T1. Hvilket igjen styrker argumentet om at økt spesifistet gir mer sikkert resultat under begge omstendigheter.

e.)

Hvor høy måtte spesifisiteten ha vært for test T2 for at sannsynligheten i punkt d skal bli mindre enn 5%

Løsning:

Vi ser at $P(T_2|\overline{A})$ er komplementet til spesifisiteten i T2 som betyr at vi kan løse oppgaven med hensyn på komplimentet, og bruke $1 - P(T_2|\overline{A}) = P(\overline{T_2}|\overline{A})$. Det gir

$$P(\overline{A}|T_2) < 0.05$$

$$\frac{xP(\overline{A})}{xP(\overline{A}) + P(T_2|A)P(A)} < 0.05$$

$$0.99x < 0.05(0.99x + 0.00942)$$

$$0.99x - 0.0495x < 0.000471$$

$$x < 0.0005007$$

og spesifisiteten må dermed være $\underline{\underline{1 - 0.0005007 = 0.9994 = 99.94\%}}$.

f.)

10 uavhengige, tilfeldig valgte personer testes med testen T1, og alle tester *negativt. Hva er sannsynligheten for at minst en av disse 10 egentlig har vært smittet av koronaviruset og derfor har antistoffer i blodet?*

Løsning:

Når vi har uavhengige forsøk tenker vi med engang produktsetningen. Det gir

$$P(\overline{T}_1) = (0.980)^n = (0.980)^{10}.$$

Så må vi huske at sannsynligheten for at minst en tester positivt er komplementet til at alle tester negativt (altså $P(T_1 \geq 1) = 1 - P(\overline{T}_1)$). Da har vi sluttvis

$$\begin{aligned} P(T \geq 1) &= 1 - P(\overline{T}_1) \\ &= 1 - (0.980)^{10} = 0.1829 = \underline{\underline{18.3\%}} \end{aligned}$$

Oppgave 2:

De to kontinuerlige stokastiske variablene X og Y har simultan sannsynlighetstetthet

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

a.)

Vis at konstanten $k = \frac{3}{28}$

Løsning:

Det er oppgitt at x, y er begrenset innenfor intervallet $[0, 2]$, og sannsynligheten for at alle x, y i D_f forekommer er dermed

$$
\begin{aligned}
P(X,Y) &= \iint_{x,y\in D_f} f(x,y)\ dxdy \\
&= \int_0^2 \int_0^2 f(x,y)\ dxdy \\
&= \int_0^2 \int_0^2 kx(x+y)\ dxdy \\
&= \int_0^2 \left[\frac{x^3}{3}k + \frac{x^2}{2}yk\right]_0^2\ dy \\
&= \int_0^2 \frac{8}{3}k + 2yk\ dy \\
&= \left[\frac{8}{3}ky + y^2k\right]_0^2 \\
&= \frac{16}{3}k + 4k = 1 \\
&= k = \frac{3}{28}
\end{aligned}
$$

som var det vi skulle vise.

b.)

Beregn sannsynligheten for at Y er større eller lik X, dvs. $P(Y \geq X)$

Løsning:

Når $Y \geq X$ er det naturlig at X må være begrenset av 0 nedenifra og Y ovenifra ettersom at Y varierer mellom verdier i $[0, 2]$. Da vil $P(Y \geq X)$ se følgelig ut

$$
\begin{aligned}
P(Y \geq X) &= \frac{3}{28}\int_0^2 \int_0^y x(x+y)\ dxdy \\
&= \frac{3}{28}\int_0^2 \left[\frac{x^3}{3} + \frac{x^2}{2}y\right]_0^y\ dy \\
&= \frac{3}{28}\int_0^2 \frac{y^3}{3} + \frac{y^3}{2}\ dy \\
&= \frac{3}{28}\left[\frac{y^4}{12} + \frac{y^4}{8}\right]_0^2 \\
&= \frac{3}{28}\left[\frac{4}{3} + \frac{6}{3}\right] \\
&= \underline{\underline{\frac{5}{14}}}
\end{aligned}
$$

c.)

Finn de marginale sannsynlighetstetthetene for X og Y. Er X og Y uavhengige?

Løsning:

Vi husker at marginaltettheten til X er gitt ved

$$
\begin{aligned}
f_X(x) &= \int_{-\infty}^{\infty} f(x,y)\ dy \\
&= \int_{-\infty}^{0} f(x,y)\ dy + \int_0^x f(x,y)\ dy
\end{aligned}
$$

hvor $\int_{-\infty}^{0} f(x,y)\ dy = 0$ ettersom at x er definert i $[0,2]$. Da har vi

$$
\begin{aligned}
f_X(x) &= \int_0^x f(x,y)\ dy \\
&= \frac{3}{28}\int_0^x x^2 + xy\ dy \\
&= \frac{3}{28}\left[x^2y + \frac{y^2}{2}x\right]_0^x \\
&= \frac{9x^3}{56}.
\end{aligned}
$$

Som betyr at $f_X(x)$ kan skrives slik.

$$f_X(x) = \begin{cases} \frac{9x^3}{56} & \text{for} \quad 0 \leq x \leq 2 \\ 0 & \text{ellers} \end{cases}$$

For marginaltettheten til Y følger vi samme ressonement som gir

$$\begin{aligned} f_Y(y) &= \int_{-\infty}^{\infty} f(x,y)\ dy \\ &= \int_0^y f(x,y)\ dx \\ &= \frac{3}{28} \int_0^y x^2 + xy\ dx \\ &= \frac{3}{28} \left[\frac{x^3}{3} + \frac{x^2}{2} y \right]_0^y \\ &= \frac{3}{28} \left(\frac{y^3}{3} + \frac{y^3}{2} \right) \\ &= \frac{15y^3}{168} = \frac{5y^3}{56}. \end{aligned}$$

Hvilket betyr at $f_Y(y)$ kan skrives slik.

$$f_Y(y) = \begin{cases} \frac{5y^3}{56} & \text{for} \quad 0 \leq y \leq 2 \\ 0 & \text{ellers} \end{cases}$$

Skal X, Y være uavhengige må de oppfylle kravet

$$f(x,y) = f_X(x) \cdot f_Y(y)$$

samt vil ulikheten

$$f(a \leq X \leq b, c \leq Y \leq d) = f_X(a \leq X \leq b) \cdot f_Y(c \leq Y \leq d)$$

holde. Sjekker vi for vårt tilfelle har vi

$$\begin{aligned} f_X(x) \cdot f_Y(y) &= \frac{9x^3}{56} \cdot \frac{5y^3}{56} \\ &= \frac{45x^3y^3}{3136} \neq f(x,y) \end{aligned}$$

som betyr at X, Y er avhengige stokastiske variable.

d.)

Finn sannsynlighetstettheten til $U = X + Y$. (Vink : Finn først den simultane tettheten til U og V, der $U = X + Y$ og $V = X$.)

Løsning:

Vi har $U = X + Y$ og $V = X$. La så $g(u, v)$ være simultantettheten til $f_{XY}(x, y)$, slik at $g(u, v) > 0$ for $0 < u < 2 \wedge 0 < v < 2$. Finner vi så den inverse transformasjonen til U, V har vi $\boxed{Y = U - X = U - V}$ og $\boxed{X = V}$. Det gir

$$g(u, v) = f_{XY}(v, u - v) \begin{vmatrix} \frac{\partial x}{\partial u} & \frac{\partial x}{\partial v} \\ \frac{\partial y}{\partial u} & \frac{\partial y}{\partial v} \end{vmatrix}$$

hvor

$$\begin{vmatrix} \frac{\partial x}{\partial u} & \frac{\partial x}{\partial v} \\ \frac{\partial y}{\partial u} & \frac{\partial y}{\partial v} \end{vmatrix} = \begin{vmatrix} 0 & 1 \\ 1 & -1 \end{vmatrix} = |-1|$$

da vil sannsynlighetstettheten til U være

$$\begin{aligned} g_U(u) &= \int_0^2 f_{XY}(v, u - v)|-1|\ dv \\ &= \frac{3}{28} \int_0^2 v^2 + v(u - v^2)\ dv \\ &= \frac{3}{28} \int_0^2 vu\ dv \\ &= \frac{3}{28} \left[\frac{v^2}{2} u \right]_0^2 \\ &= \frac{3}{28} \left(\frac{4u}{2} \right) \\ &= \frac{3}{14} u. \end{aligned}$$

slik at

$$g_U(u) = \begin{cases} \frac{3}{14} u & \text{for} \quad 0 \leq u \leq 2 \\ 0 & \text{ellers} \end{cases}$$

Oppgave 3:

Anta at $Y \sim N(\mu, \sigma)$ og la $X = e^Y$. Da er X lognormalfordelt med parametere μ og σ.

a.)

Vis at medianen i den lognormale fordelingen er $\eta = e^{\mu}$. *Vis også at* $E(X) = E(e^Y) = \eta e^{\frac{\sigma^2}{2}}$ *(Vink: For å bestemme forventningen, kan du bruke at den momentgenererende funksjonen til* Y *er* $M_Y(t) = E(e^{tY}) = e^{\mu t + \frac{\sigma^2 t^2}{2}}$*)*

Løsning:

Vi har at $\frac{Y-\mu}{\sigma} \sim N(0,1)$ slik at den kumulative fordelingen til X er gitt ved

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

videre vet vi at for en standardnormalfordelt variabel er μ, og medianen lik null ettersom fullstendig symmetri om midtpunktet i fordelingen. Dermed har vi

$$\begin{aligned} \Phi\left(\frac{\ln(\eta)-\mu}{\sigma}\right) &= 0.5 \\ \left(\frac{\ln(\eta)-\mu}{\sigma}\right) &= \Phi^{-1}(0.5) \\ \ln(\eta) - \mu &= 0 \\ \eta &= e^{\mu} \end{aligned}$$

som var det vi skulle vise.

Vi velger å benytte oss av den momentgenererende funksjonen til Y for å finne forventningen til X. Da har vi

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

hvor vi observerer at forventningen til X er den momentgenererende funksjonen til Y i punktet $t = 1$.

$$M_Y(1) = E(e^Y) = E(X) = e^{\mu + \frac{\sigma^2}{2}}$$

som var det vi skulle vise.

b.)

Vis at

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

er en forventningsrett estimator for μ og bestem fordelingen til $\hat{\mu}$.

Løsning:

Vi har at

$$\begin{aligned} E[\hat{\mu}] &= \frac{1}{n}\sum_{i=1}^{n} E[Y_i] \\ &= \frac{1}{n}\sum_{i=1}^{n} M'_Y(0) \\ &= \frac{1}{n}\sum_{i=1}^{n} (\mu + 2t\sigma^2)e^{\mu t + \frac{\sigma^2 t^2}{2}} \\ &= \frac{1}{n}\mu n \\ &= \mu \end{aligned}$$

der vi brukte den momentgenererende funksjonen til Y i punktet $t = 0$ for å finne forventningen til Y og vise at $\hat{\mu}$ er forventningsrett. $\hat{\mu}$ vil dessuten være normalfordelt.

c.)

En mulig estimator for medianen η er $\eta^* = e^{\hat{\mu}}$. Forklar hvorfor denne estimatorene ikke er forventningsrett.

Løsning:

Vi har at

$$E(\eta^*) = E(e^{\hat{\mu}})$$

så observerer vi at $e^{\hat{\mu}} = e^{\ln(X_n)} = X_n$ hvilket gir

$$\begin{aligned} E(\eta^*) &= E(X_n) \\ &= e^{\hat{\mu} + \frac{\sigma^2}{2n}} \\ &= \eta e^{\frac{\sigma^2}{2n}} \end{aligned}$$

som er forskjellig fra medianen η, og beviser at η^* ikke er forventningsrett.

d.)

Vis at $\hat{\eta} = e^{\hat{\mu} - \frac{\sigma^2}{2n}}$ *er en forvetningsrett estimator for medianen.*

Løsning:

Bruke samme ressonement som i forrige i oppgave

$$\begin{aligned}
E(\hat{\eta}) &= E(e^{\hat{\mu} - \frac{\sigma^2}{2n}}) \\
&= E(e^{\hat{\mu}} e^{-\frac{\sigma^2}{2n}}) \\
&= e^{-\frac{\sigma^2}{2n}} E(e^{\hat{\mu}}) \\
&= e^{\mu + \frac{\sigma^2}{2n} - \frac{\sigma^2}{2n}} = \eta
\end{aligned}$$

hvilket beviser at $\hat{\eta}$ er forventningsrett.

e.)

Vis at

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

er et 95% *konfidensintervall for medianen.*

Løsning:

Vi antar at vi har et utvalg som er tilstrekkelig stort slik at sentralgrensesetning gjelder. Da har vi

$$Z = \frac{\overline{y} - \mu}{\frac{\sigma}{\sqrt{n}}}$$

og derfor er

$$P\left(-z_{\frac{\alpha}{2}} \le \frac{\overline{y} - \mu}{\frac{\sigma}{\sqrt{n}}} \le z_{\frac{\alpha}{2}} \right) \approx 1 - \alpha$$

$$P\left(-\frac{z_{\frac{\alpha}{2}} \sigma}{\sqrt{n}} \le \overline{y} - \mu \le \frac{z_{\frac{\alpha}{2}} \sigma}{\sqrt{n}} \right) \approx 1 - \alpha$$

$$P\left(\overline{y} - \frac{z_{\frac{\alpha}{2}} \sigma}{\sqrt{n}} \le \mu \le \overline{y} + \frac{z_{\frac{\alpha}{2}} \sigma}{\sqrt{n}} \right) \approx 1 - \alpha$$

$$P\left(e^{\overline{y} - \frac{z_{\frac{\alpha}{2}} \sigma}{\sqrt{n}}} \le e^{\mu} \le e^{\overline{y} + \frac{z_{\frac{\alpha}{2}} \sigma}{\sqrt{n}}} \right) \approx 1 - \alpha$$

hvilket gir et konfidensintervall for medianen

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

som var det vi skulle vise.

f.)

Vis at $\hat{\eta}e^{\frac{\sigma^2}{2}}$ *er en forventningsrett estimator for* $E(X)$*. Vis også at vi får et 95 % konfidensintervall for* $E(X)$ *ved å gange nedre og øvre grense av konfidensintervallet i punkt e med* $e^{\frac{\sigma^2}{2}}$*.*

Løsning:

Vi har at

$$\begin{aligned} E(\hat{\eta}e^{\frac{\sigma^2}{2}}) &= e^{\frac{\sigma^2}{2}}E(\hat{\eta}) \\ &= e^{\frac{\sigma^2}{2}}\eta \\ &= E(X) \end{aligned}$$

som var det vi skulle vise. Videre ganger vi øvre og nedre grense i konfidensintervallet med $e^{\frac{\sigma^2}{2}}$.

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

$$\left[e^{\overline{y}-\frac{1.96\sigma}{\sqrt{n}}}e^{\frac{\sigma^2}{2}},E(X),\ e^{\overline{y}+\frac{1.96\sigma}{\sqrt{n}}}e^{\frac{\sigma^2}{2}}\right]$$

og dermed har vi vist at vi får et 95 % konfidensintervall for $E(X)$.

g.)

Gjør beregningen beskrevet ovenfor for $n = 10$ *og* $n = 30$ *når* $\mu = 0.7$ *og* $\sigma = 1.2$*. Bruk* $B = 10\ 000$*. Diskuter hva resultatene sier deg om (mulig) skjevhet og standardfeil for estimatorene* η^* *og* $\tilde{\eta}$*.*

Løsning:

For å løse oppgaven lager vi et python-script. Det kan se slik ut

```
import numpy as np
import matplotlib.pyplot as plt
import scipy.stats as stats

sigma = 1.2
my = 0.7

def f3(B, n):

```

```python
    eta_star = np.zeros(B)
    eta_tilde = np.zeros(B)

    for i in range(B):
        s = np.random.normal(my, sigma**2, n)
        uhat = 1/n * np.sum(s)
        S = 1/(n-1) * np.sum(np.power((s - uhat), 2))
        eta_star[i] = np.exp(uhat)
        eta_tilde[i] = np.exp(uhat - S/(2*n))

    eta_star_avg = np.mean(eta_star)
    eta_tilde_avg = np.mean(eta_tilde)

    std1 = np.std(eta_star, ddof = 1)
    std2 = np.std(eta_tilde, ddof = 1)

    return eta_star_avg, eta_tilde_avg, std1, std2

eta_s_1, eta_t_1, std11, std12 = f3(10000, 10)
eta_s_2, eta_t_2, std21, std22 = f3(10000, 30)

print(f""" Eta stjerne gjennomsnitt (n = 10): {eta_s_1: .2f}  Eta tilde gjennomsnitt (n = 10): {eta_t_1: .2f}
Avvik eta stjerne (n = 10): {std11: .2f} Avvik eta tilde (n = 10): {std12: .2f}
""")

print(f""" Eta stjerne gjennomsnitt (n = 30){eta_s_2: .2f}  Eta tilde gjennomsnitt (n = 30): {eta_t_2: .2f}
Avvik eta stjerne (n = 30): {std21: .2f} Avvik eta tilde (n = 30): {std22: .2f}
""")
```

som ved gjennomkjøring gir følgende output

```
Eta stjerne gjennomsnitt (n = 10):  2.23
Avvik eta stjerne (n = 10):  1.08

Eta tilde gjennomsnitt (n = 10):  2.01
Avvik eta tilde (n = 10):  0.98

Eta stjerne gjennomsnitt (n = 30) 2.09
Avvik eta stjerne (n = 30):  0.56

 Eta tilde gjennomsnitt (n = 30):  2.02
Avvik eta tilde (n = 30):  0.54
```

Hvor vi observerer at $\tilde{\eta}$ gir et delvis bedre estimat enn η^* både ved å se på differansen i gjennomsnitt fra $n = 10$ til $n = 30$ og på størrelsen i avvik som er noe mindre i begge tilfellene. Skjevheten er det noe vanskelig å kommentere, men det er opplagt at for store avvik følger det stor skjevhet, og avvikene for tilstrekkelig store utvalg blir mindre og mindre. I dette tilfellet er avvikene noe store for mindre utvalg, men minker betraktelig for fler som betyr at skjevheten kan være liten.

Oppgave 4:

a.)

Hvor stor andel av (ikke-avholdende) norske kvinner drikker minst 10 liter ren alkohol per år?

Løsning

Vi husker at den kumulative fordelingsfunksjonen til X er gitt ved

$$\Phi\left(\frac{\ln(X)-\mu}{\sigma}\right)$$

hvor $\mu = 0.7$ og $\sigma = 1.2$. Da har vi at andelen kvinner som drikker minst 10 liter ren alkohol per år er

$$\begin{aligned} P(X \geq 10) &= 1 - \Phi\left(\frac{\ln(10)-\mu}{\sigma}\right) \\ &= 1 - \Phi\left(\frac{\ln(10)-0.7}{1.2}\right) \\ &= 1 - 0.90906 = \underline{\underline{0.09094 \approx 9.1\%}} \end{aligned}$$

hvor vi har slått opp i standardnormalfordelingstabellen og funnet at $\Phi(1.335) = 0.9090956$.

b.)

Bestem medianen til det årlige alkoholforbruket og forventet årlig alkoholforbruk for (ikke-avholdende) voksne norske kvinner. Hvilken av de to verdiene egner seg etter din mening best til å beskrive alkoholforbruket for voksne norske kvinner?

Løsning:

Vi husker at medianen er gitt ved $\eta = e^{\mu}$ og forventingen $E(X) = \eta e^{\frac{\sigma^2}{2}}$. Setter vi inn verdiene får vi

$$\begin{aligned} \eta &= e^{\mu} \\ &= \underline{\underline{2.013}} \text{ L} \\ E(X) &= \eta e^{\frac{\sigma^2}{2}} \\ &= 2.013(e^{\frac{1.2^2}{2}}) \\ &= \underline{\underline{4.137}} \text{ L.} \end{aligned}$$

hvor det i vårt tilfelle er medianen som beskriver det årlige alkoholforbruket mest presist, grunnet at standaravviket er > 1 og gir skjevhet mot høyre hvilket medianen ikke påvirkes like mye av.

c.)

Gi et estimat og et tilnærmet 95 % konfidensintervall for medianen til det årlige alkoholforbruket for studentgruppen. Gi også et estimat og et tilnærmet 95 %konfidensintervall for forventet årlig alkoholforbruk.

Løsning:

Oppgaven kan løses ved å lage et python-script som kan se følgende ut

```python
import numpy as np
import matplotlib.pyplot as plt
import scipy.stats as stats

def f2(n):

    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])

    estimat_median = np.median(x)
    estimat_expected1 = np.zeros(len(x))

    for i in range(len(x)):

        uhat = (1/n) * np.sum(np.log(x))
        S = 1/(n-1) * np.sum(np.power((np.log(x) - uhat), 2))
        estimat_expected1[i] = estimat_median*(np.exp(S/2))

    low = np.exp(uhat - ((1.96*S)/(np.sqrt(n))))
    high = np.exp(uhat + ((1.96*S)/(np.sqrt(n))))

    low2 = np.exp(uhat - ((1.96*S)/(np.sqrt(n)))) * np.exp(S/2)
    high2 = np.exp(uhat + ((1.96*S)/(np.sqrt(n)))) * np.exp(S/2)

    return low, high, low2, high2, estimat_expected1[0], estimat_median

```

```
lowM, highM, lowE, highE, estimat_forv, estimat_med = f2(30)

print(f""" Nedre grense median: {lowM: .2f} Ovre grense median: {highM: .2f
    }
""")
print(f""" Nedre grense forventning: {lowE: .2f} Ovre grense forventning: {
    highM: .2f}
""")
print(f""" Estimat median: {estimat_med: .2f} Estimat forventning: {
    estimat_forv: .2f}
""")
```

som gir følgende output

```
Nedre grense median:  2.99

Øvre grense median:  6.72

Nedre grense forventning:  5.27

Øvre grense forventning:  6.72

Estimat median:  4.10

Estimat forventning:  7.22
```

d.)

Bruk parametrisk bootstrap til å bestemme standardfeilen til estimatene i forrige punkt.

Løsning:

Et python-script for parametrisk bootstrap kan se noe slikt ut

```
import numpy as np
import matplotlib.pyplot as plt
import scipy.stats as stats

def f3(n, B):

    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])

    meanstar = np.zeros(B)
    medianstar = np.zeros(B)

    for i in range(B):
        uhat = (1/n) * np.sum(np.log(x))
        S = 1/(n-1) * np.sum(np.power(np.log(x) - uhat, 2))
```

```python
        s = np.random.normal(uhat, S, n)
        meanstar[i] = np.mean(s)
        medianstar[i] = np.median(s)

    standard1 = np.std(meanstar)
    standard2 = np.std(medianstar)

    return standard1, standard2

standardfeil1, standarfeil2 = f3(30, 10000)

print(f"""Standardfeil forventning : {standardfeil1: .2f} Standarfeil median: {standarfeil2: .2f}
""")
```

hvilket gir følgende standardfeil

```
Standardfeil forventning :  0.21

Standarfeil  median:  0.25
```

e.)

Diskuter hva resultatene i punktene c og d sier deg om alkoholforbruket til den aktuelle gruppen av kvinnelige studenter sammenlignet med alle voksne norske kvinner.

Løsning:

Vi observerer at konfidensintervallet til medianen er langt bredere enn forventningen. Dette betyr at medianen kan ligge mellom flere mulige verdier og er derfor mer usikkert. Interessant nok er det verdt å merke seg at esitmert forventing ligger utenfor konfidensintervallet selv etter å ha trukket fra standardfeilen, hvilket kan indikere at estimatet er særdeles upresist. Men sammenlignet med utvalget vårt, samt en gjennomsnittlig norsk kvinne er ca. 7 liter ren alkohol et ganske presist anslag på årlig alkoholforbruk.