# STK1110 – Oblig 1

Jonas Gahr Sturtzel Lunde (`jonassl`)

October 3, 2018

## Oppgave 1

### a)

Vi velger oss estimatoren $\hat{\mu} = \overline{X} \approx 276.89$. 90% konfidensintervallet til $\mu$ med $\overline{X}$ estimatoren er utledet i oppgave 2a til å være

$$P\left(\overline{X} - t_{0.05,\,n-1} \cdot \frac{S}{\sqrt{n}} < \sigma < \overline{X} + t_{0.95,n-1} \cdot \frac{S}{\sqrt{n}}\right) = 0.9 \tag{1}$$

Setter vi inn for alle verdier, får vi (se vedlagt kode) et 90% konfidensintervall på $[266.4, 287.4]$.

### b)

Vi velger oss estimatoren $\hat{\sigma} = S \approx 26.36$. 90% konfidensintervallet til $\sigma$ med $S$ estimatoren er utladet i oppgave 2b til å være

$$P\left(\sqrt{\frac{(n-1)}{\chi_{0.05,\,n-1}}}S < \sigma < \sqrt{\frac{(n-1)}{\chi_{0.95,\,n-1}}}S\right) = 0.9 \tag{2}$$

Setter vi inn for alle verdier, får vi (se vedlagt kode) et 90% konfidensintervall på $[20.8, 36.5]$.

### c)

Konfidensintervallene ble utledet under antagelsen om at fordelingen er normalfordelt. Vi er avhengige av å vite fordelingen for å kunne utlede et konfidensintervall, og målingene virker å være normalfordelte. I forventningsverdiens tilfelle kan vi også lene oss på sentralgrenseteoremet, men det er få datapunkter, og dette gjelder uansett ikke for standardavviket.

### d)

Ikke-parametrisk bootstrap gir et 90% konfidensintervall på (se vedlagt kode) $[267.2, 287.2]$ for $\mu$ og $[18.3, 32.4]$ for $\sigma$. Vi sitter altså med

| | $\mu$ | $\sigma$ |
|---|---|---|
| Analytisk | $[266.4, 287.4]$ | $[20.8, 36.5]$ |
| Bootstrap | $[267.2, 287.2]$ | $[18.3, 32.4]$ |

Som vi ser stemmer intervallet for forventningsverdien $\mu$ meget godt overens, mens det er vesentlige forskjeller i intervallet for standardavviket $\sigma$.

### e)

Vi har en nullhypotese $H_0 : \mu = 265$, og en alternativ hypotese $H_a = \mu < 265$. Vi har allerede antatt at dataen er normalfordelt, som betyr at $\overline{X}$ skal være t-fordelt. Dersom nullhypotesen holder, skal $\overline{X}$ være t-fordelt rundt $\mu_0$ som

$$t = \frac{\overline{X} - \mu_0}{S/\sqrt{n}} \tag{3}$$

der vi bruker estimatoren $S$ som erstatning for standardavviket $\sigma$. Vi forkaster nullhypotesen dersom denne verdien er mindre enn $-t_\alpha = -1.734$. Vi har at

$$t = \frac{276.89 - 265}{26.36/\sqrt{19}} = 1.966 \tag{4}$$

Dette er åpenbart ikke mindre enn $-1.734$, og nullhypotesen holder. Dett er å forvente, ettersom den alternative hypotesen var at $\mu$ skulle være mindre enn en verdi som allerede er en del mindre enn det observerte gjennomsnittet.

Dersom vi gjør en "two-tailed" test, forkaster vi nullhypotesen dersom $t > t_{0.025} = 2.101$ eller $t < -t_{0.025} = -2.101$. Ingen av disse holder, selv om det nå åpenbart er betydelig nærmere, ettersom det oberserverte gjennomsnittet var på den øvre delen av nullhypotesen.

## Oppgave 2

### a)

Ettersom ligning (1) fra oppgaven er en t-fordeling med $n-1$ frihetsgrader, vet vi at den følger

$$P\left(t_{\alpha/2,\,n-1} < \frac{\overline{X} - \mu}{S/\sqrt{n}} < t_{1-\alpha/2,\,n-1}\right) = 1 - \alpha \tag{5}$$

der $t_{\alpha/2,\,n-1}$ og $t_{1-\alpha/2,\,n-1}$ er $\alpha/2$ og $1-\alpha/2$ persentilene til en t-fordeling med $n-1$ frihetsgrader.

Løser vi ulikheten inni parantesen for $\mu$ får vi at

$$\overline{X} - t_{\alpha/2,\,n-1} \cdot \frac{S}{\sqrt{n}} < \mu < \overline{X} + t_{1-\alpha/2,n-1} \cdot \frac{S}{\sqrt{n}}$$

som er $100(1-\alpha)\%$ konfidensintervallet til $\mu$.

### b)

Ettersom ligning (1) fra oppgaven er kjikvadrat-fordelt med $n-1$ frihetsgrader, vet vi at den tilfredsstiller

$$P\left(\chi_{\alpha/2,\,n-1} < \frac{(n-1)}{\sigma^2} S^2 < \chi_{\alpha/2,\,n-1}\right) = 1 - \alpha \tag{6}$$

der $\chi_{\alpha/2,\,n-1}$ og $\chi_{1-\alpha/2,\,n-1}$ er $\alpha/2$ og $1-\alpha/2$ persentilene til en kjikvadrat-fordeling med $n-1$ frihetsgrader.

Løser vi ulikheten inni parantesen for $\sigma$ får vi at

$$\sqrt{\frac{(n-1)}{\chi_{\alpha/2,\,n-1}}} S < \sigma < \sqrt{\frac{(n-1)}{\chi_{1-\alpha/2,\,n-1}}} S \tag{7}$$

som er $100(1-\alpha)\%$ confidensintervallet til $\sigma$.

### c) & d)

Kode vedlagt i appendiks. Resultatet presenteres i neste oppgave.

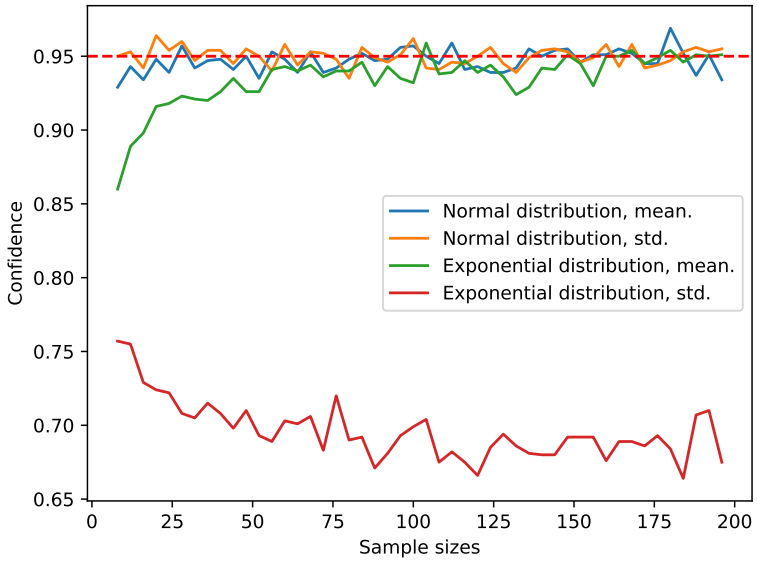


Figure 1: Observerte hit-rates av konfidensintervaller for forventnignsverdi og standardavvik til eksponensialfordeling og normalfordeling mot forventet konfidens.

### e) & f)

I figur 1 er det plottet andelen av de genererte 95% konfidensintervallene som treffer de faktiske $\mu$ og $\sigma$ verdiene til normal- og eksponensialfordelingene. Vi ser at andelen stemmer godt overens med den forventede verdien på 0.95 for både standardavvik og forventningsverdi for normalfordelingen (med noe støy). Dette er forventet, ettersom konfidensintervallet er utledet med hensyn på en normalfordeling.

De målte konfidensintervallene til forventningsverdien til eksponensialfordelingen stemmer ikke overens med forventet verdi for lave sample sizes, men begynner å nærme seg forventede verdier ved $n \approx 60$. Her begynner sentralgrenseteoremet å gjelde, som sier at den målte forventningsverdien til *enhver* fordeling vil være normalfordelt ved store samples.

Konfidensintervallet til standardavviket til forventningsverdien er feil hele veien, ettersom sentralgrenseteoremet bare gjelder forventningsverdier, og konfidensintervallet vårt ikke er tilpasset en eksponensialfordeling.

## Oppgave 3

### a)

$$F(x) = \int_{-\infty}^{x} f(x)\,\mathrm{d}x = \int_{\kappa}^{x} \theta\kappa^{\theta}x^{-\theta-1}\,\mathrm{d}x = \left[\theta\kappa^{\theta}\frac{x^{-\theta}}{-\theta}\right]_{\kappa}^{\theta} = 1 - \left(\frac{\kappa}{x}\right)^{\theta}$$

$$F(x) = 1 - \left(\frac{\kappa}{x}\right)^{\theta} = \frac{1}{2} \;\Rightarrow\; \frac{\kappa}{x} = \frac{1}{2}$$

### b)

$$E(X) = \int_{-\infty}^{\infty} x \cdot f(x)\,\mathrm{d}x = \int_{\kappa}^{\infty} \theta\kappa^{\theta}x^{-\theta}\,\mathrm{d}x = \left[\theta\kappa^{\theta}\frac{x^{-\theta+1}}{-\theta+1}\right]_{\kappa}^{\infty} = 0 - \theta\kappa^{\theta}\frac{\kappa^{-\theta+1}}{-\theta+1} = \frac{\theta\kappa}{\theta-1}$$

## c)

Vi omskriver den gitte definisjonen til å gi $X$ som definisjon av $Y$:

$$\begin{aligned} Y = 2\theta[\ln(X) - \ln(\kappa)] &= 2\theta \ln(X/\kappa) \\ e^{Y/2\theta} &= X/\kappa \\ X &= \kappa e^{Y/2\theta} \end{aligned}$$

Vi setter dette inn i uttrykket vårt for den kummulative fordelingsfunksjonen:

$$F(Y) = 1 - \left(\frac{\kappa}{\kappa e^{y/2\theta}}\right)^{\theta} = 1 - e^{-y/2}$$

$$f(y) = F'(y) = \frac{1}{2}e^{-y/2} = \frac{1}{2^{2/2}\Gamma(2/2)}x^{2/2-1}e^{-x/2}$$

som vi ser er en kjikvadratfordeling med 2 frihetsgrader.

## d)

$$E(X) = \frac{\theta\kappa}{\theta - 1} = \overline{X} \;\Rightarrow\; \theta\kappa = \theta\overline{X} - \overline{X} \;\Rightarrow\; \theta = \frac{\overline{X}}{\overline{X} - \kappa}$$

Momentestimatoren til $\theta$ er altså $\hat{\theta} = \frac{\overline{X}}{\overline{X}-\kappa}$.

## e)

Sannsynligheten for at kombinasjonen av tilfeldige variable $X_1, X_2, ..., X_n$ i $n$ uavhengige forsøk blir $x_1, x_2, ..., x_n$ vil være produktet av de individuelle sannsynlighetene

$$f(x_1, x_2, ..., x_n; \theta) = \prod_{i=1}^{n} \theta\kappa^{\theta}\left(\frac{1}{x_i}\right)^{\theta+1} = \theta^n\kappa^n\theta\left(\prod_{i=1}^{n}(x_i)^{-1}\right)^{\theta+1}$$

Vi skal finne maks-verdien til denne fordelingen. Vi tar først logaritmen av fordelingen, ettersom den deler toppunkt med sin logaritme, og dette er enklere å regne med.

$$\begin{aligned} \ln[f(x_1, x_2, ..., x_n; \theta)] &= \ln(\theta^n) + \ln(\kappa^n\theta) + \ln\left[\left(\prod_{i=1}^{n} x_i^{-1}\right)^{\theta+1}\right] \\ &= n\ln(\theta) + n\theta\ln(\kappa) + (\theta+1)\left[\sum_{i=1}^{n} -\ln(x_i)\right] \end{aligned}$$

Deriverer, setter lik 0, og løser for $\theta$:

$$\begin{aligned} \frac{\mathrm{d}}{\mathrm{d}\theta}\ln[f(x_1, x_2, ..., x_n; \theta)] = \frac{n}{\theta} + n\ln(\kappa) - \sum_{i=1}^{n}\ln(x_i) &= 0 \\ n &= \theta\left[\sum_{i=1}^{n}\ln(x_i) - n\ln(\kappa)\right] \\ \hat{\theta} &= \frac{n}{\sum_{i=1}^{n}\ln(x_i) - n\ln(\kappa)} \end{aligned}$$

som da er maximum likelihood estimatoren for $\theta$.

## f)

Vi har en ny stokastisk variabel

$$Y = 2n\frac{\theta}{\hat{\theta}} = 2\theta\left[\sum_{i=1}^{n}\ln(x_i) - n\ln(\kappa)\right] = \sum_{i=1}^{n} 2\theta[\ln(x_i) - \ln(\kappa)] = \sum_{i=1}^{n} Z$$

hvor $Z \sim \chi^2_2$, altså $Z$ er kjikvadratfordelt med 2 frihetsgrader.

Fra s.316 i læreboka har vi at summen av kjikvadratfordelinger selv er en kjikvadratfordeling, med ny frihetsgrad lik summen av frihetsgradene, som betyr at $Z$ er en kjikvadratfordeling med $2n$ frihetsgrader:

$$\sum_{i=1}^{n} Z \sim \chi^2_{2n}$$

## g)

$$E[\hat{\theta}] = 2n\theta \cdot E[Y^{-1}]$$

Ettersom $Y$ er en kjikvadratfordelt tilfeldig variabel med $\nu = 2n$ frihetsgrader, setter vi inn for dens forventningsverdi, definert i ligning (5) i oppgaven

$$E[\hat{\theta}] = 2n\theta \cdot \frac{2^{-1}\Gamma\left(\frac{2n}{2} - 1\right)}{\Gamma\left(\frac{2n}{2}\right)} = n\theta\frac{\Gamma(n-1)}{\Gamma(n)} = n\theta\frac{(n-2)!}{(n-1)!} = \theta\frac{n}{n-1}$$

For å finne variansen finner vi først $E[\hat{\theta}^2]$.

$$E[\hat{\theta}^2] = 2^2n^2\theta^2 \cdot \frac{2^{-2}\Gamma\left(\frac{2n}{2} - 2\right)}{\Gamma\left(\frac{2n}{2}\right)} = n^2\theta^2\frac{\Gamma(n-2)}{\Gamma(n)} = n\theta\frac{(n-3)!}{(n-1)!} = \theta^2\frac{n^2}{(n-1)(n-2)}$$

Vi bruker da definisjonen

$$\begin{aligned} V[\hat{\theta}] = E[\hat{\theta}^2] - E[\hat{\theta}]^2 &= \theta^2\frac{n^2}{(n-1)(n-2)} - \theta^2\frac{n^2}{(n-1)^2} \\ &= \theta^2\frac{n^2(n-1) - n^2(n-2)}{(n-1)^2(n-2)} = \theta^2\frac{n^2}{(n-1)^2(n-2)} \end{aligned}$$

## h)

Estimatoren er ikke unbiased, fordi forventningsverdien ikke er $\theta$. Biasen til estimatoren er definert som

$$Bias(\hat{\theta}) = E[\hat{\theta}] - \theta = \frac{1}{n-1}\theta$$

Vi ser at vi kan gjøre estimatoren unbiased ved å gange den med $n - 1/n$:

$$\frac{n-1}{n}\hat{\theta} = \frac{n-1}{\sum_{i=1}^{n}\ln(X_i) - n\ln(\kappa)}$$

# Appendiks

## Kode til oppgave 1

```
import numpy as np
import matplotlib.pyplot as plt
import scipy.stats as st
age = np.array([249, 254, 243, 268, 253, 269, 287, 241, 273, 306,\
                303, 280, 260, 256, 278, 344, 304, 283, 310])
n = len(age)

# __TASK 1A & 1B
X_bar = np.mean(age)
S = np.std(age)

df = n-1
conf_int_mean = [X_bar - st.t.ppf(0.95, df)*S/np.sqrt(n),\
                 X_bar - st.t.ppf(0.05, df)*S/(np.sqrt(n))]
conf_int_std = [np.sqrt((n-1)/(st.chi2.ppf(0.95, df)))*S,\
                np.sqrt((n-1)/(st.chi2.ppf(0.05, df)))*S]
print(conf_int_mean)
print(conf_int_std)

# __TASK 1C__
N = 1000000

sample = np.random.choice(age, size=(N,n), replace=True)

averages = np.sort(np.mean(sample, axis=1))
stds = np.sort(np.std(sample, axis=1))

averages_conf_int = (averages[int(0.05*N-1)], averages[int(0.95*N-1)])
stds_conf_int = (stds[int(0.05*N-1)], stds[int(0.95*N-1)])

print(averages_conf_int)
print(stds_conf_int)
```

## Kode til oppgave 2

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



N = 1000
alpha = 0.05
mu = 1
sigma = 1
ns = range(8, 200, 4)

#__TASK 2C & 2D___
hit_rates_norm_mean = []
hit_rates_norm_std = []
for n in ns:
    df = n - 1

    t_ppf_lower = st.t.ppf(1 - alpha/2, df)
    t_ppf_upper = st.t.ppf(alpha/2, df)
    chi_pdf_lower = st.chi2.ppf(1 - alpha/2, df)
    chi_pdf_upper = st.chi2.ppf(alpha/2, df)

    data = np.random.normal(loc=mu, scale=sigma, size=(N, n))

    X_bar = np.mean(data, axis=1)
    S = np.std(data, axis=1)

    conf_int_mean = np.zeros((N,2))
    conf_int_std = np.zeros((N,2))
    conf_int_mean[:,0] = X_bar - t_ppf_lower*S/np.sqrt(n)
```

```
    conf_int_mean[:,1] = X_bar - t_ppf_upper*S/np.sqrt(n)
    conf_int_std[:,0] = np.sqrt((n-1)/(chi_pdf_lower))*S
    conf_int_std[:,1] = np.sqrt((n-1)/(chi_pdf_upper))*S

    does_contain_mu = (conf_int_mean[:,0] < mu) * (mu < conf_int_mean[:,1])
    does_contain_sigma = (conf_int_std[:,0] < sigma) * (sigma < conf_int_std[:,1])
    hit_rates_norm_mean.append( np.sum(does_contain_mu)/N )
    hit_rates_norm_std.append( np.sum(does_contain_sigma)/N )


#__TASK 2E & 2F___
hit_rates_exp_mean = []
hit_rates_exp_std = []
for n in ns:
    df = n - 1

    t_ppf_lower = st.t.ppf(1 - alpha/2, df)
    t_ppf_upper = st.t.ppf(alpha/2, df)
    chi_pdf_lower = st.chi2.ppf(1 - alpha/2, df)
    chi_pdf_upper = st.chi2.ppf(alpha/2, df)

    data = np.random.exponential(scale=sigma, size=(N, n))

    X_bar = np.mean(data, axis=1)
    S = np.std(data, axis=1)

    conf_int_mean = np.zeros((N,2))
    conf_int_std = np.zeros((N,2))
    conf_int_mean[:,0] = X_bar - t_ppf_lower*S/np.sqrt(n)
    conf_int_mean[:,1] = X_bar - t_ppf_upper*S/np.sqrt(n)
    conf_int_std[:,0] = np.sqrt((n-1)/(chi_pdf_lower))*S
    conf_int_std[:,1] = np.sqrt((n-1)/(chi_pdf_upper))*S

    does_contain_mu = (conf_int_mean[:,0] < mu) * (mu < conf_int_mean[:,1])
    does_contain_sigma = (conf_int_std[:,0] < sigma) * (sigma < conf_int_std[:,1])
    hit_rates_exp_mean.append( np.sum(does_contain_mu)/N )
    hit_rates_exp_std.append( np.sum(does_contain_sigma)/N )

plt.plot(ns, hit_rates_norm_mean, label="Normal distribution, mean.")
plt.plot(ns, hit_rates_norm_std, label="Normal distribution, std.")
plt.plot(ns, hit_rates_exp_mean, label="Exponential distribution, mean.")
plt.plot(ns, hit_rates_exp_std, label="Exponential distribution, std.")
plt.axhline(y=0.95, ls='--', color='r')
plt.ylabel("Confidence")
plt.xlabel("Sample sizes")
plt.legend()
plt.savefig("opg2.pdf")
```