Oppgave 1

Vi antar at vi har stokastiske variable X_1, X_2, ..., X_n, som er er uavhengige og uniformt fordelt på intervallet $[0, \theta]$, dvs at de har tetthet

$$f(x;\theta) = \begin{cases} 1/\theta & \text{for } 0 \leq x \leq \theta, \\ 0 & \text{ellers.} \end{cases}$$

Parameteren θ er ukjent, og skal estimeres.

a)

Vi starter med å finne forventingen av X_i, som per definisjon gitt ved

$$E(X_i) = \int_{-\infty}^{\infty} x f(x;\theta)\ \mathrm{d}x,$$

setter inn uttrykket for sannsynlighetsfordelingen og ser at bare $x \in (0, \theta]$ gir bidrag til integralet:

$$E(X_i) = \int_0^\theta \frac{x}{\theta}\ \mathrm{d}x = \frac{\theta^2}{2\theta} = \frac{\theta}{2}.$$

Vi kan finne variansen til X_i fra uttrykket

$$V(X_i) = E(X_i^2) - E(X_i)^2,$$

vi må da først finne $E(X_i^2)$, som gjøres likt som tidligere:

$$E(X_i^2) = \int_{-\infty}^{\infty} x^2 f(x;\theta)\ \mathrm{d}x = \int_0^\theta \frac{x^2}{\theta}\ \mathrm{d}x = \frac{\theta^2}{3}.$$

Variansen er da

$$V(X_i) = E(X_i^2) - E(X_i)^2 = \frac{\theta^2}{12}.$$

Vi har altså vist følgende

$$E(X_i) = \frac{\theta}{2}, \qquad V(X_i) = \frac{\theta^2}{12}.$$

b)

Vi finner momentestimatoren for θ ved å sette det første distribusjonsmomentet, $E(X_i)$, lik det første samplingsmomentet, $\overline{X}$

$$E(X_i) = \frac{1}{n}\sum_{i=1}^{n} X_i,$$

vi setter inn for forventningen og løser for estimatoren

$$\frac{\hat{\theta}_{\mathrm{mom}}}{2} = \overline{X} \quad \Rightarrow \quad \hat{\theta}_{\mathrm{mom}} = 2\overline{X}.$$

Forventningen av momentestimatoren blir

$$E(\hat{\theta}_{\text{mom}}) = E(2\overline{X}),$$

setter inn for $\overline{X}$ og bruker linearitet av forventningen

$$E(\hat{\theta}_{\text{mom}}) = E\left(\frac{2}{n}\sum_{i=1}^{n} X_i\right) = \frac{2}{n}\sum_{i=1}^{n} E(X_i) = \frac{2}{n}\frac{n\theta}{2} = \theta.$$

Ettersom at

$$E(\hat{\theta}_{\text{mom}}) - \theta = 0,$$

ser vi at momentestimatoren er en forventningsrett estimator.

c)

Standardfeilen til momentestimatoren er gitt ved

$$\sigma_{\hat{\theta}_{\text{mom}}} = \sqrt{V(\hat{\theta}_{\text{mom}})},$$

vi finner derfor først variansen til estimatoren, bruker da at vi generelt for variansen har

$$V\left(a + \sum_i b_i X_i\right) = \sum_i b_i^2 V(X_i).$$

Finner at

$$V(\hat{\theta}_{\text{mom}}) = V\left(\sum_{i=1}^{n} \frac{2}{n} X_i\right) = \frac{4}{n^2}\sum_{i=1}^{n} V(X_i) = \frac{4}{n^2}\frac{n\theta^2}{12} = \frac{\theta^2}{3n}.$$

Slik at standardfeilen blir

$$\sigma_{\hat{\theta}_{\text{mom}}} = \frac{\theta}{\sqrt{3n}}.$$

Vi kan nå vise at estimatoren er konsistent ved å bruke Chebychevs ulikhet, som sier at for enhver stokastisk variabel X, med forventning μ og varians σ^2, så vil

$$P(|X - \mu| > t) \leq \frac{\sigma^2}{t^2},$$

gjelde for alle $t > 0$. Momentestimatoren $\hat{\theta}_{\text{mom}}$ er en stokastisk variabel med forventning $\mu = \theta$ og varians $\sigma^2 = \theta^2/3n$, vi setter dette inn i ulikheten og får

$$P(|\hat{\theta}_{\text{mom}} - \theta| > t) \leq \frac{\theta^2}{3nt^2}.$$

Vi ser at for en hvilken $t > 0$ vi velger, kan vi gjøre uttrykket på høyre side mindre enn en hvilken som helst toleranse $\epsilon > 0$ ved å velge $n > N$ for en eller annen N. Det vil si at momentestimatoren konvergerer mot θ i sannsynlighet når n vokser.

d)

Siden de stokastiske variable X_1, X_2, $\ldots$ X_n er uavhengige, så vil likelihooden være produktet av de individuelle sannsynlighetsfordelingene

$$f(x_1, x_2, \ldots, x_n; \theta) = \prod_{i=1}^{n} f(x_i; \theta).$$

Ved å sette inn $f(x_i; \theta)$ ser vi at produktet blir

$$f(x_1, x_2, \ldots, x_n; \theta) = \begin{cases} (1/\theta)^n & \text{for } 0 \leq x_1, \ldots, x_n \leq \theta \\ 0 & \text{ellers.} \end{cases}.$$

e)

Vi skal nå finne maksimum likelihood estimatoren $\hat{\theta}_{\max}$, det vil si den verdien av den ukjente parameteren θ slik at likelihooden er maksimert:

$$f(x_1, x_2, \ldots, x_n; \hat{\theta}_{\max}) \geq f(x_1, x_2, \ldots, x_n; \theta).$$

Vi er altså ute etter å finne et globalt maksimum i likelihood funksjon

$$f(x_1, x_2, \ldots, x_n; \theta) = \begin{cases} (1/\theta)^n & \text{for } 0 \leq x_1, \ldots, x_n \leq \theta \\ 0 & \text{ellers.} \end{cases}.$$

Vi ser med en gang at en mindre θ betyr en større sannsynlighet, så lenge ikke $\theta < x_1, x_2, \ldots, x_n$, vi lar derfor

$$\hat{\theta}_{\max} = \max_{1 \leq i \leq n} X_i.$$

Merk at vi ikke kan finne dette maksimumet ved å derivere likelihood-funksjonen fordi likelihood funksjonen har en diskontinuitet akkurat i dette punktet.

f)

Sannsynlighetstettheten til den største sampelen fra en stokastisk variabel med sannsynlighetstetthet $f(x)$ og kumulativ sannsynlighet $F(x)$ er gitt ved[1]

$$g_n(y) = n\big[F(y)\big]^{n-1} \cdot f(y).$$

Vi vet at for X_i har vi tettheten

$$f(x; \theta) = \begin{cases} 1/\theta & \text{for } 0 \leq x \leq \theta, \\ 0 & \text{ellers,} \end{cases}$$

slik at den kumulative sannsynligheten blir

$$F(x; \theta) = \begin{cases} 0 & \text{for } x < 0 \\ \int_0^x \frac{1}{\theta} \, \mathrm{d}x' = \frac{x}{\theta} & \text{for } 0 \leq x \leq \theta \\ 1 & \text{for } x > \theta, \end{cases}$$

[1] Se *Devore & Berk* avsnitt 5.5, side 268-269

$$\int_0^x \frac{1}{\theta}\ \mathrm{d}x' = \frac{x}{\theta}.$$

Innsetting gir da at sannsynlighetstettheten til maksimum likelihood estimatoren er

$$g_n(y) = n\left(\frac{y}{\theta}\right)^{n-1}\frac{1}{\theta} = \frac{ny^{n-1}}{\theta^n}.$$

g)

Ettersom at vi nå kjenner sannsynlighetsfordelingen til maksimum likelihood estimatoren, kan vi finne forventningen til estimatoren direkte

$$E(\hat{\theta}_{\mathrm{max}}) = \int_{-\infty}^{\infty} y \cdot g_n(y)\ \mathrm{d}y,$$

innsetting gir

$$E(\hat{\theta}_{\mathrm{max}}) = \int_0^{\theta} \frac{nx^n}{\theta^n}\ \mathrm{d}y = \frac{n}{n+1}\theta.$$

h)

Vi ser fra forventningen til maksimums likelihood estimatoren at den ikke er forventningsrett. Ettersom at forventningen er lineær, ser vi at vi kan lage en forventningsrett estimator ved å gange inn en faktor:

$$\hat{\theta}_{\mathrm{mod}} = \frac{n+1}{n}\hat{\theta}_{\mathrm{max}},$$

slik at vi har

$$E(\hat{\theta}_{\mathrm{mod}}) = E\left(\frac{n+1}{n}\hat{\theta}_{\mathrm{max}}\right) = \frac{n+1}{n}E(\hat{\theta}_{\mathrm{max}}) = \frac{n+1}{n}\frac{n}{n+1}\theta = \theta.$$

Standardfeilen til den modifiserte estimatoren blir

$$\sigma_{\hat{\theta}_{\mathrm{mod}}} = \sqrt{V(\hat{\theta}_{\mathrm{mod}})},$$

der variansen er

$$V(\hat{\theta}_{\mathrm{mod}}) = V\left(\frac{n+1}{n}\hat{\theta}_{\mathrm{max}}\right) = \frac{(n+1)^2}{n^2}V(\hat{\theta}_{\mathrm{max}}).$$

Variansen til maksimums likelihood estimatoren er igjen

$$V(\hat{\theta}_{\mathrm{max}}) = E(\hat{\theta}^2_{\mathrm{max}}) - E(\hat{\theta}_{\mathrm{max}})^2,$$

der

$$E(\hat{\theta}_{\mathrm{max}}) = \frac{n}{n+1}\theta,$$

og

$$E(\hat{\theta}^2_{\mathrm{max}}) = \int_0^{\theta} \frac{nx^{n+1}}{\theta^n}\ \mathrm{d}x = \frac{n}{n+2}\theta^2.$$

Innsetting gir da

$$V(\hat{\theta}_{\text{mod}}) = \frac{(n+1)^2}{n^2}\left(\frac{n}{n+2}\theta^2 - \frac{n^2}{(n+1)^2}\theta^2\right),$$

som videre gir

$$V(\hat{\theta}_{\text{mod}}) = \frac{n^2+2n+1-n^2-2n}{n(n+2)}\theta = \frac{\theta^2}{n(n+2)}.$$

Slik at standardfeilen er

$$\sigma_{\hat{\theta}_{\text{mod}}} = \frac{\theta}{\sqrt{n(n+2)}}.$$

i)

Vi har nå funnet to forventningsrette estimatorer:

$$\hat{\theta}_{\text{mom}} = 2\overline{X}, \qquad \hat{\theta}_{\text{mod}} = \frac{n+1}{n}\left(\max_{1\le i\le n} X_i\right).$$

Med standardfeilene

$$\sigma_{\hat{\theta}_{\text{mom}}} = \frac{\theta}{\sqrt{3n}}, \qquad \sigma_{\hat{\theta}_{\text{mod}}} = \frac{\theta}{\sqrt{n(n+2)}}.$$

Vi ser nå at selv om begge estimatorene konvergerer mot θ med økende n, så går momentestimatoren som $\mathcal{O}(1/\sqrt{n})$, mens den modifiserte estimatoren går som $\mathcal{O}(1/n)$. Den modifiserte estimatoren har altså en mye bedre asymptotisk oppførsel en momentestimatoren, om vi har mulighet til å ta mange samples er det altså den modifiserte estimatoren som er å foretrekke.

j)

Vi skal nå gjennomføre et numerisk eksperiment for å teste resultatet vi har funnet. Vi trekker $n = 20$ samples fra en uniform fordeling med parameter $\theta = 1$, vi regner så ut hva de to estimatorene estimerer at θ faktisk er, vi gjør dette 1000 ganger på rad og lager et histogram av resultatene. For å undersøke den asymptotiske oppførselen til de to estimatoren gjentar vi forsøket for $n = 200$ og $n = 2000$. Se vedlegg 1 for kildekoden og figur 1, 2 og 3 for resultatene.

Vi ser at den modifiserte estimatoren har en mye skarpere topp og ligger generelt sett nærmere den faktiske verdien av θ enn det momentestimatoren gjør, og vi ser at dette bare blir klarere når n økes, slik som vi har kommet frem til. Vi ser også at momentestimatoren legger seg ganske symmetrisk om θ, mens den modifiserte estimatoren derimot er asymmetrisk og har mer tendens til å legge seg under θ.

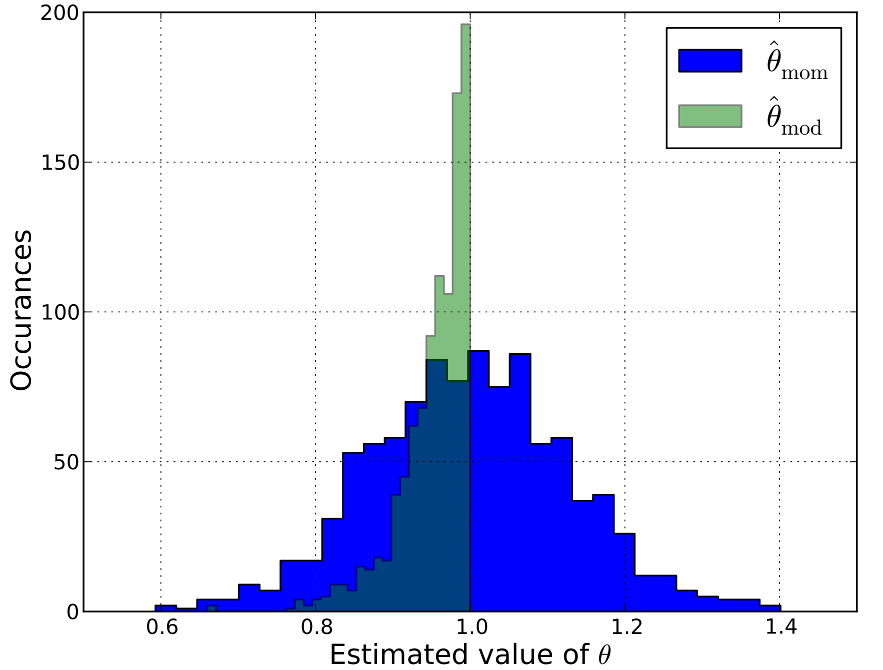

Figure 1: Resultatene av $N = 1000$ forsøk med $n = 20$ samples hver presentert i et histogram.

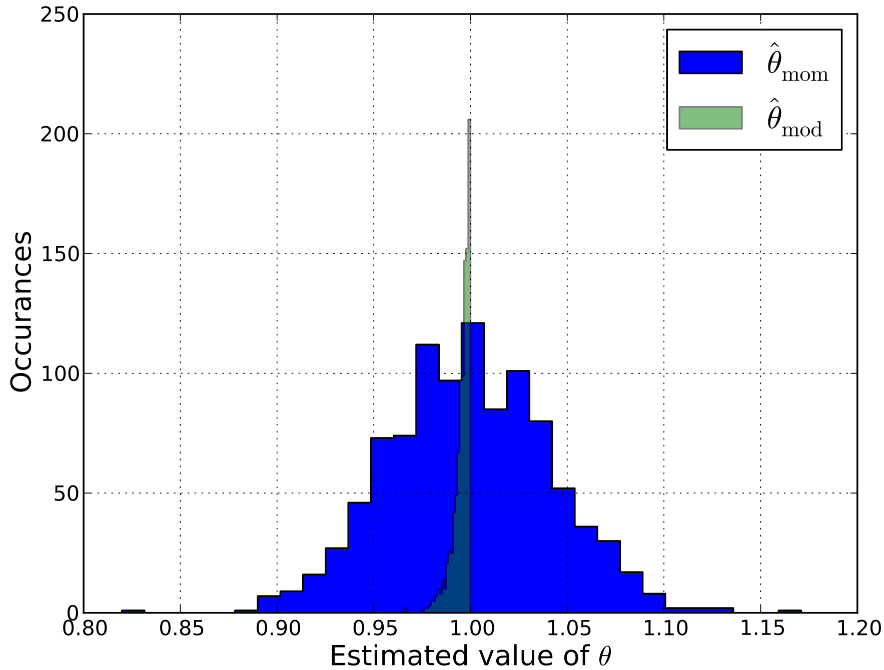

Figure 2: Resultatene av $N = 1000$ forsøk med $n = 200$ samples hver presentert i et histogram.

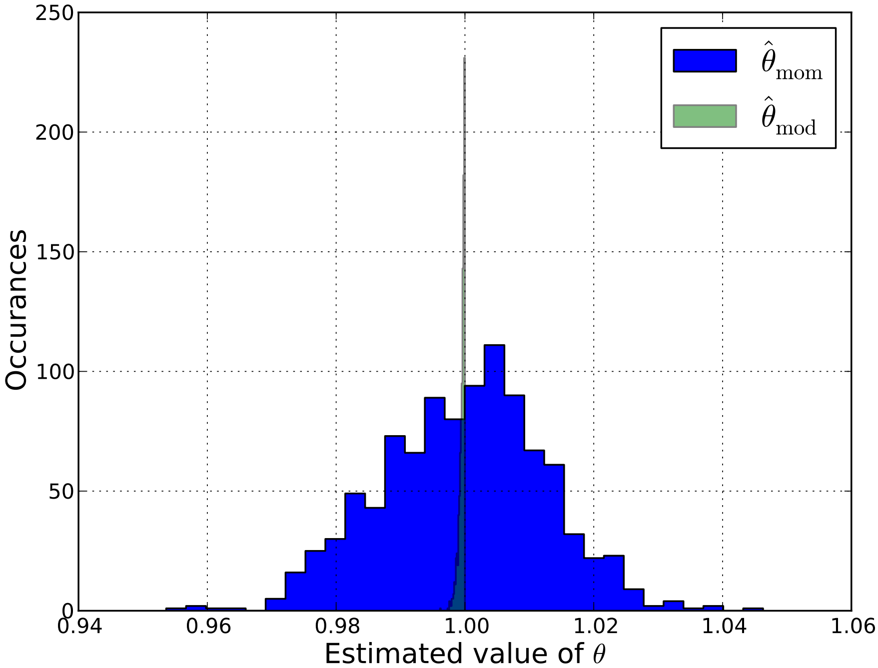

Figure 3: Resultatene av $N = 1000$ forsøk med $n = 2000$ samples hver presentert i et histogram.

Problem 2

Ettersom at vi ikke kjenner det faktiske standardavviket σ, bruker vi istedet sample standardavviket S, vi finner derfor først $\overline{X}$ og S:

$$\overline{X} = 14.36, \quad S = \sqrt{\frac{\sum_i (x_i - \overline{x})^2}{n-1}} = 1.156.$$

Et 95% konfidensintervall for forventningen μ kan da finnes fra

$$P\Big(-1.96 < \frac{\overline{X} - \mu}{S/\sqrt{n}} < 1.96 \Big) \approx 0.95.$$

Vi manipulerer ulikheten og finner at

$$\overline{X} - \frac{1.96S}{\sqrt{n}} < \mu < \overline{X} + \frac{1.96S}{\sqrt{n}}.$$

Slik at konfidensintervallet er

$$\overline{X} \pm \frac{1.96S}{\sqrt{n}}.$$

Ved innsett av verdier finner vi at et 95% konfidensintervall for μ er

$$(13.76, 14.97).$$

Vedlegg 1 - Kildekode til oppgave 1j

```
from numpy.random import random
from numpy import zeros

# Number of samples per trial
n = 20
# Number of trials
N = 1000

results = zeros((N,2))
for i in range(N):
  # Draws n random numbers uniformly from [0,1)
  x = random(n)
  # Calculate estimators
  mom = 2*sum(x)/len(x)
  mod = (n+1)/n * max(x)
  # Store results
  results[i,0] = mom
  results[i,1] = mod

# Plot results as a histogram
import matplotlib.pyplot as plt
plt.hist(results[:,0], bins=30, histtype='stepfilled',
         color='b', label=r'$\hat{\theta}_{\rm mom}$')
plt.hist(results[:,1], bins=30, histtype='stepfilled',
         color='g', alpha=0.5, label=r'$\hat{\theta}_{\rm mod}$')
plt.xlabel(r"Estimated value of $\theta$", fontsize=16)
plt.ylabel(r"Occurances", fontsize=16)
plt.legend(prop={'size':18})
plt.grid()
#plt.savefig('estimators_hist_n%s.pdf' % str(n))
plt.show()
```