Oppgave 1

Vi antar at vi har stokastiske variable $X_1, X_2, \ldots, X_n$, 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

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

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 tolerans $\epsilon > 0$ 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, ... 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

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

innsetting gir

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

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:

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

slik at vi har

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

Standardfeilen til den modifiserte estimatoren blir

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

der variansen er

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

Variansen til maksimums likelihood estimatoren er igjen

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

der

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

og

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

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:

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

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.

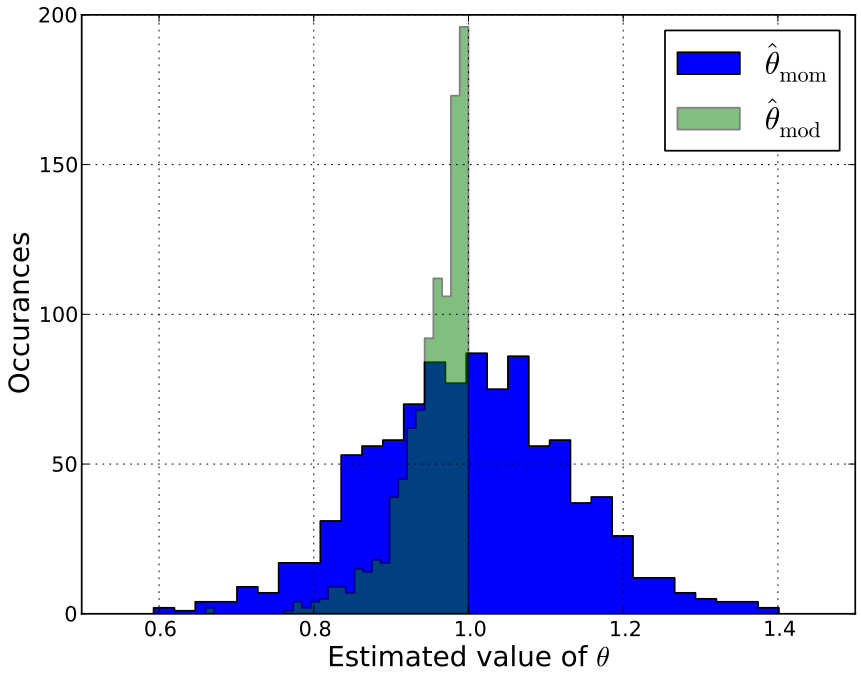

Figure 1: Resultatene av $N = 1000$ forsøk med $n = 20$ samples hver presentert i et histogram.

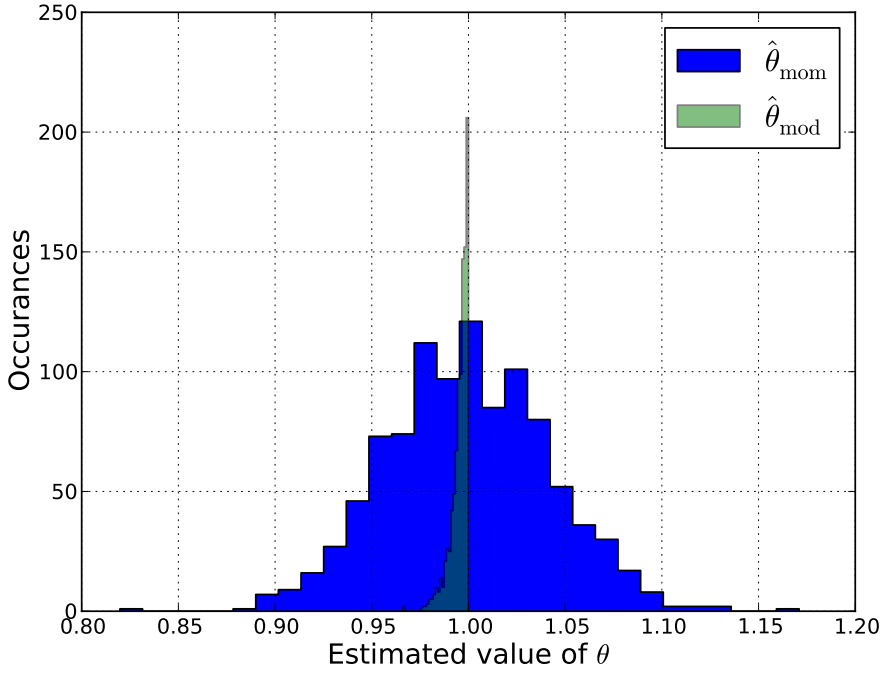

Figure 2: Resultatene av $N = 1000$ forsøk med $n = 200$ samples hver presentert i et histogram.

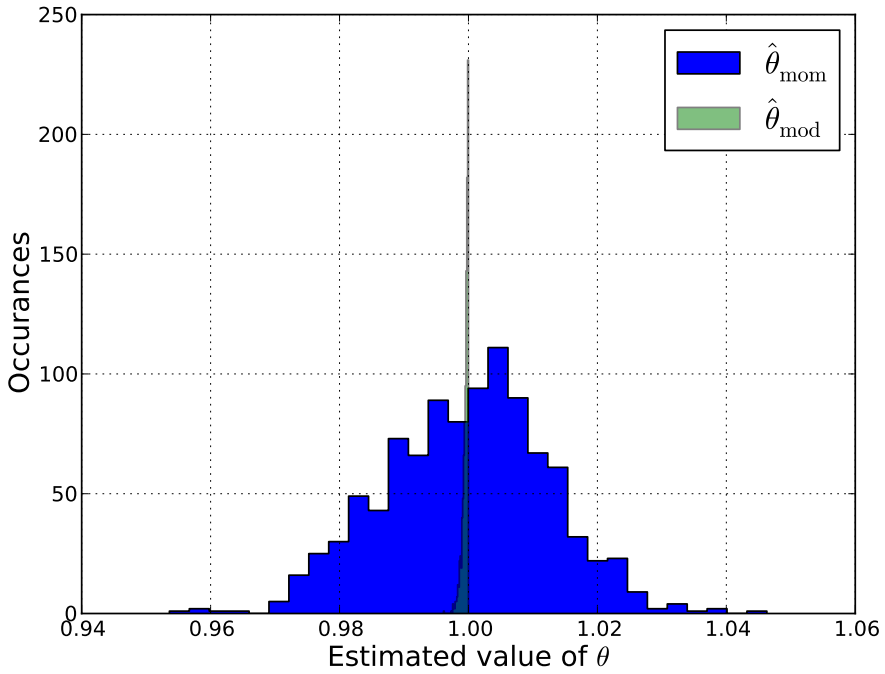

Figure 3: Resultatene av $N = 1000$ forsøk med $n = 2000$ samples hver presentert i et histogram.

Problem 2

a)

Ettersom at vi ikke kjenner det faktiske standardavviket σ, bruker vi istedet det empiriske 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

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

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

$$(13.76, 14.97).$$

b)

Vi skal nå gjennomføre et numerisk eksperiment hvor vi genererer $N = 1000$ datasett, hvert av størrelse $n = 14$ observasjoner. Vi lar de 14 stokastiske variable være uavhengige og normalfordelt med parametere $\mu = 14.5$ og $\sigma = 1$. For hvert av datasettene regner vi ut et 95 % konfidensintervall, merk at vi i dette tilfelle kjenner standardavviket σ, så vi slipper å gå veien om det empirsike standardavviket S. Etter å ha generert de 1000 datasettene og tilhørende konfidensintervall, teller vi opp antallet av intervallene som inneholder den faktiske verdien av μ. Koden er lagt ved som vedlegg 2 på slutten av oppgaven.

Vi kjører programmet vårt 100 ganger, og lager et histogram av resultatene, ser figur 4. Vi ser fra resultatene at μ er inneholdt i konfidensintervallet ca. 95 % av datasettene. Vi legger samtidig merke til at dette langt ifra alltid er nøyaktig 95 %, og at hvor mange ganger μ ligger innenfor konfidensintervallet er en stokastisk variabel.

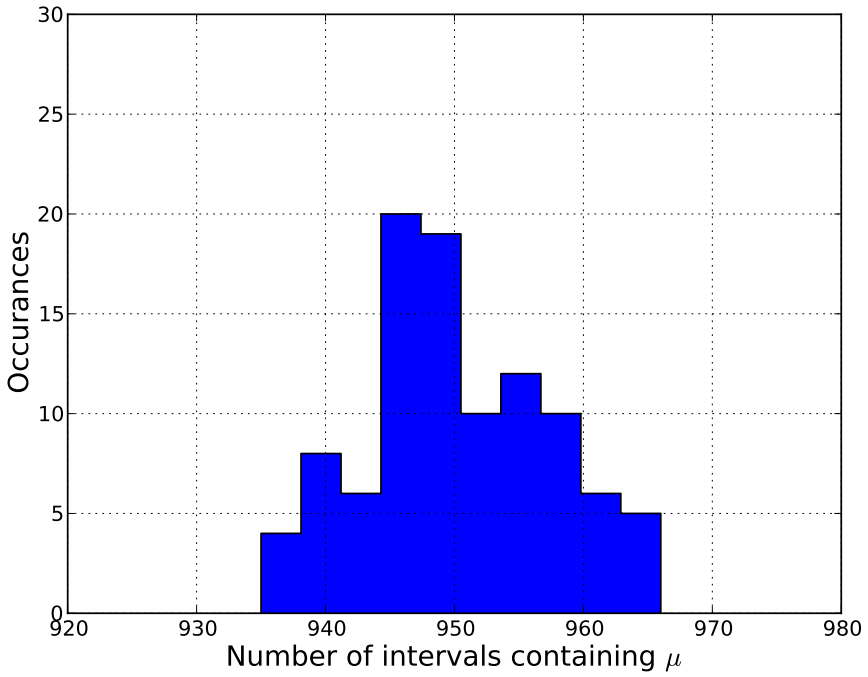

Figure 4: Resultatene av 100 forsøk med 1000 datasett hver.

c)

Vi gjennomfører nå akkurat det samme eksperimentet, men bruker nå det empiriske standardavviket S istedetfor den faktiske variansen σ. Vi gjennomfører igjen 100 forsøk og plotter resultatene i figur 5.

Vi ser igjen at antall intervaller som inneholder μ er en stokastisk variabel. I dette tilfellet ser derimot ikke lenger histogrammet ut til å være sentrert rundt 950, eller 95 %, forekomster, men mer mot 93 %. Denne lille forskjellen skyldes at vi ikke bruker vår kunnskap om standardavviket σ, men må isteden bruke det empiriske standardavviket, som da innfører litt mer usikkerthet.

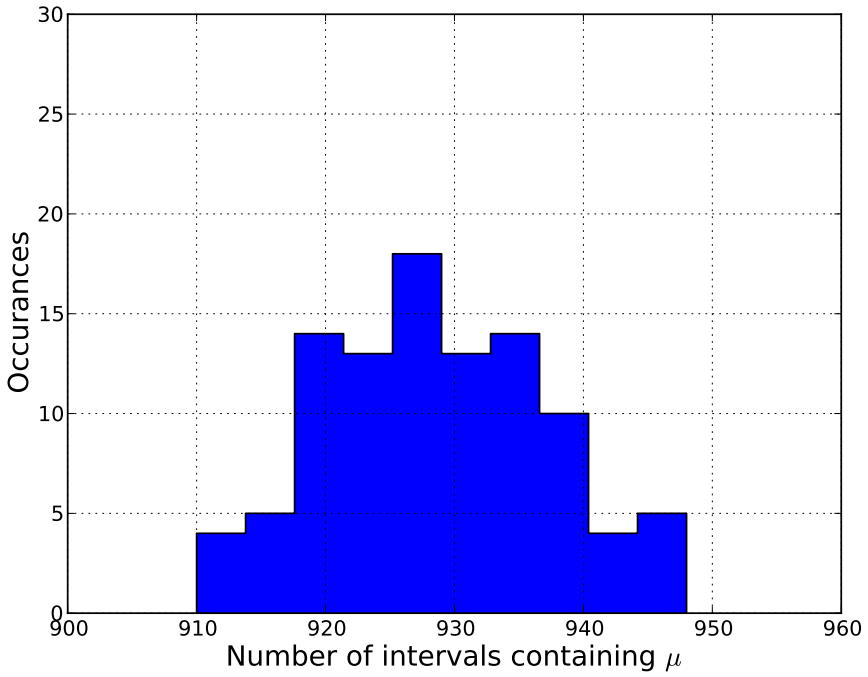

Figure 5: Resultatene av 100 forsøk med 1000 datasett hver.

d)

Som vi har påpekt tidligere, er antallet av konfidensintervallene som faktisk inneholder μ en stokastisk variabel. Denne stokastiske variabelen vil ha en binomisk fordeling, ettersom at det for hvert intervall er en sannsynlighet p for at μ er inneholdt, og en sannsynlighet $1-p$ for at μ ikke er inneholdt. Ettersom at p er lik for alle datasettene vil resultatet være binomisk fordelt.

e)

Vi skal nå vise at $13 \cdot S^2/\sigma^2$ er χ^2-fordelt med 13 frihetsgrader. Vi starter med å bevise et lemma som viser hvordan differansen av to χ^2-fordelte variabler er fordelt.

Lemma: Hvis $X_3 = X_1 + X_2$, der $X_1 \sim \chi^2_{\nu_1}$, $X_3 \sim \chi^2_{\nu_3}$, der $\nu_3 > \nu_1$, og X_1 og X_2 er uavhengige, da er $X_2 \sim \chi^2_{\nu_3 - \nu_1}$.

Bevis: Ettersom at X_1 og X_2 er uavhengige, så vil den momentgenererende funksjonen til X_3 være gitt ved produktet av de momentgenererende funksjonene til X_1 og X_2, slik at

$$M_{X_3}(t) = M_{X_1+X_2}(t) = M_{X_1}(t)M_{X_2}(t).$$

Vi vet samtidig at den momentgenererende funksjonen for en χ^2-fordelt variabel med ν frihetsgrader er $(1-2t)^{-\nu/2}$. slik at

$$(1-2t)^{-\nu_3/2} = (1-2t)^{-\nu_1/2}M_{X_2}(t).$$

Vi løser nå for $M_{X_2}(t)$:

$$M_{X_2}(t) = (1-2t)^{-(\nu_3-\nu_1)/2},$$

og ettersom at $\nu_3 > \nu_1$ ser vi at dette svarer til den momentgenererende funksjonen til en χ^2-fordelt variabel med $\nu_3 - \nu_1$ frihetsgrader. Siden momentgenererende funksjoner er entydige, må altså $X_2 \sim \chi^2_{\nu_3-\nu_1}$.

Vi starter nå med å definere en stokastisk variabel X_3 som følger:

$$X_3 = \sum_{i=1}^{14} \left(\frac{X_i - \mu}{\sigma} \right)^2.$$

Vi vet nå at $X_3 \sim \chi^2_{14}$, fordi X_3 er summen av 14 kvadrerte standardnormalfordelte variable. Vi utvider nå innsiden av summen ved å legge til og trekke fra $\overline{X}$

$$X_3 = \sum_{i=1}^{14} \left(\frac{X_i - \overline{X}}{\sigma} + \frac{\overline{X} - \mu}{\sigma} \right)^2,$$

vi ganger ut og finner

$$X_3 = \sum_{i=1}^{14} \left(\frac{X_i - \overline{X}}{\sigma} \right)^2 + \sum_{i=1}^{14} 2\left(\frac{X_i - \overline{X}}{\sigma} \right)\left(\frac{\overline{X} - \mu}{\sigma} \right) + \sum_{i=1}^{14} \left(\frac{\overline{X} - \mu}{\sigma} \right)^2$$

Den midterste summen blir null, fordi

$$\sum_{i=1}^{14} 2\left(\frac{X_i - \overline{X}}{\sigma} \right)\left(\frac{\overline{X} - \mu}{\sigma} \right) = 2\left(\frac{\overline{X} - \mu}{\sigma^2} \right) \sum_{i=1}^{14} (X_i - \overline{X}) = 0.$$

Mens den siste summen avhenger ikke av i, slik at vi har

$$X_3 = \sum_{i=1}^{14} \left(\frac{X_i - \overline{X}}{\sigma} \right)^2 + 14 \left(\frac{\overline{X} - \mu}{\sigma} \right)^2.$$

som kan skrives om til

$$X_3 = \sum_{i=1}^{14} \left(\frac{X_i - \overline{X}}{\sigma} \right)^2 + \left(\frac{\overline{X} - \mu}{\sigma/\sqrt{14}} \right)^2.$$

Vi gjenkjenner det første leddet på venstre side som $13 \cdot S^2/\sigma^2$, og det siste leddet som en standardnormalfordeltvariabel kvadrert, slik at

$$X_3 = 13 \cdot S^2/\sigma^2 + X_1, \text{ der } X_1 \equiv \left(\frac{\overline{X} - \mu}{\sigma/\sqrt{14}} \right)^2.$$

Vi vet nå at $X_3 \sim \chi^2_{14}$ og $X_1 \sim \chi^2_1$, men før vi kan si med sikkerhet at $13 \cdot S^2/\sigma^2 \sim \chi^2_{14-1}$ må vi vise at $13 \cdot S^2/\sigma^2$ og X_1 er uavhengige. Vi ser at $13 \cdot S^2/\sigma^2$ bare avhenger av den stokastiske variabelen S^2, mens X_1 avhenger av $\overline{X}$, og vi vet fra før at S^2 og $\overline{X}$ er uavhengige. Så da har vi vist at

$$X_2 = 13 \cdot S^2/\sigma^2 \sim \chi^2_{13}.$$

f)

Vi vil nå beregne et 99% konfidensintervall for σ^2. Vi har vist at $13 \cdot S^2/\sigma^2 \sim \chi^2_{13}$, slik at vi vet[2]

$$P(3.565 \leq 13\frac{S^2}{\sigma^2} \leq 29.817) = 0.99.$$

Vi løser ulikheten

$$3.565 \leq 13\frac{S^2}{\sigma^2} \leq 29.817,$$

for σ^2 ved å ta resiproken, og finner

$$\frac{13S^2}{3.565} \geq \sigma^2 \geq \frac{13S^2}{29.817},$$

setter vi inn for $S^2 = 1.336$ (se oppg 2a), finner vi at et 99% konfidensintervall for σ^2 er

$$\sigma^2 \in [0.436, 4.872].$$

[2]Se *Devore & Berk* tabell A.7, side 791.

Oppgave 3

Vi lar nå X være antall tilfeller av en sjelden medfødt sykdom i året, og antar at X er Poisson-fordelt med parameter λ, slik at

$$P(X = k) = \frac{\lambda^k}{k!} e^{-\lambda}, \text{ der } k = 0, 1, 2, \ldots$$

a)

Parameteren λ har vært 1, slik at $H_0 : \lambda = 1$, vi skal nå undersøke hvor mange tilfeller man må observere et gitt år for å forkaste H_0 til fordel for $H_a : \lambda > 1$ på et nivå $\alpha = 0.05$.

Størrelsen α beskriver sannsynligheten for å gjøre en type I feil, det vil si at vi forkaster H_0 selv om den stemmer. For å finne denne sannsynligheten tenker vi oss altså at vi forkaster H_0 hvis det er k eller flere tilfeller av sykdommen et gitt år. Ettersom at parameteren er $\lambda = 1$, kan vi lett regne oss frem til sannsynligheten for en type I feil for hver mulig verdi for k:

$$\begin{aligned}
\alpha &= P(\text{type I error}) \\
&= P(X \geq k) \\
&= 1 - \sum_{i=0}^{k-1} P(X = i) \\
&= 1 - \sum_{i=0}^{k-1} \frac{1}{ek!}.
\end{aligned}$$

Vi har da at

k	α
0	1.00
1	0.63
2	0.26
3	0.08
4	0.02

Slik at om man vil ha en $\alpha \leq 0.05$, må man velge å forkaste H_0 om man observerer minst $k = 4$ tilfeller av sykdommen et gitt år.

b)

Vi fant i forrige oppgave at vi skulle observere minst $k = 4$ tilfeller før man forkastet H_0, vi antar nå at parameteren faktisk har forandret seg, slik at $\lambda = 2$, og ønsker å finne β, som er sannsynligheten for en type II feil, det vil si at vi holder på H_0 når den faktisk er falsk.

$$\begin{aligned}
\beta &= P(\text{type II error}) \\
&= P(X \leq 3; \lambda = 2) \\
&= \sum_{i=0}^{3} P(X = i; \lambda = 2) \\
&= \frac{2^0}{0!}e^{-2} + \frac{2^1}{1!}e^{-2} + \frac{2^2}{2!}e^{-2} + \frac{2^3}{3!}e^{-2} \\
&= 0.812.
\end{aligned}$$

Så vi ser at sannsynligheten for en type II feil er mye større en en type I feil i dette tilfellet.

c)

Lar nå p være sannsynligheten for at et barn blir født med sykdommen. Vi observerer n_i fødte barn og X_i antall tilfeller for hvert år i, der $i = 1, \ldots, m$. Antar at antall tilfeller hvert år er uavhengige og Poisson-fordelt $X_i \sim \text{Poisson}(n_i p), i = 1, \ldots, m$ og skal finn sannsynlighetsmaksimering-estimatoren for sannsynligheten p.

Ettersom at antall tilfeller hvert år er uavhengig er likelihood-funksjonen:

$$f(x_1, x_2, \ldots, x_m; p) = \prod_{i=1}^{m} \frac{(n_i p)^{x_i}}{x_i!} e^{-n_i p}.$$

Tar logaritmen og får

$$\ln[f] = \sum_{i=1}^{m} \Big(x_i \ln n_i p - \ln x_i! - n_i p \Big).$$

Deriverer med hensyn på p og setter lik 0 for å finne maksimummet

$$\frac{\mathrm{d} \ln[f]}{\mathrm{d}p} = \sum_{i=1}^{m} \frac{\mathrm{d}}{\mathrm{d}p} \Big(x_i \ln n_i p - \ln x_i! - n_i p \Big) = 0,$$

som blir

$$\sum_{i=1}^{m} \left(\frac{x_i}{p} - n_i \right) = 0,$$

Vi løser nå for p,

$$\hat{p} = \frac{\sum_i X_i}{\sum_i n_i},$$

vi kan nå utvide brøken med $1/m$, slik at estimatoren er forholdet av gjennomsnittene

$$\hat{p} = \frac{\overline{X}}{\overline{n}},$$

d)

Vi finner nå forventningen og variansen til estimatoren.

Forventningen blir

$$\begin{aligned} E(\hat{p}) &= E\left(\frac{\overline{X}}{\bar{n}}\right) \\ &= \frac{1}{\sum_i n_i} E\left(\sum_{i=1}^{m} X_i\right) \\ &= \frac{1}{\sum_i n_i} \sum_{i=1}^{m} E(X_i) \\ &= \frac{1}{\sum_i n_i} \sum_{i=1}^{m} n_i p \\ &= \frac{\sum_i n_i}{\sum_i n_i} p \\ &= p. \end{aligned}$$

Så vi ser at estimatoren er forventningsrett.

Variansen blir

$$\begin{aligned} V(\hat{p}) &= V\left(\frac{\sum_i X_i}{\sum_i n_i}\right) \\ &= \left(\frac{1}{\sum_i n_i}\right)^2 \sum_{i=1}^{m} V(X_i) \\ &= \left(\frac{1}{\sum_i n_i}\right)^2 \sum_{i=1}^{m} n_i p \\ &= \frac{p}{\sum_i n_i}. \end{aligned}$$

Så vi ser at variansen avtar når den totale summen av barn født over de m årene øker.

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()
```

Vedlegg 2 - Kildekode til oppgave 2b og 2c

```
from numpy import random, sqrt, zeros

# Number of data set
N = 1000
# Number of samples per set
n = 14
# Number of experiments
m = 100
# Parameters
mu = 14.5
sigma = 1

# Conduct experiment
results = zeros(m)
for i in range(m):
    s = 0
    for j in xrange(N):
        # Draw n random numbers
        x = random.normal(mu, sigma, n)
        # Calculate interval
        x_avrg = sum(x)/n
        S = sqrt(sum((x-x_avrg)**2)/(n-1))
        lower = x_avrg - 1.96*S/sqrt(n)
        upper = x_avrg + 1.96*S/sqrt(n)
        # Check if interval contains mu
        if lower <= mu and upper >= mu:
            s += 1
    results[i] = s

# Plot results as a histogram
import matplotlib.pyplot as plt
plt.hist(results, histtype='stepfilled', color='b')
plt.axis([900,960,0,30])
plt.xlabel(r"Number of intervals containing $\mu$", fontsize=16)
plt.ylabel(r"Occurances", fontsize=16)
plt.grid()
plt.savefig('2c.pdf')
plt.show()
```