Eksamen MA223 våren 2022

Av

Kandidat nr. 285

i

MA223

Statistikk

Forelest av Svein Olav Nyberg

Fakultet for teknologi og realfag

Universitetet i Agder

Grimstad, Mai 2022

Innholdsfortegnelse

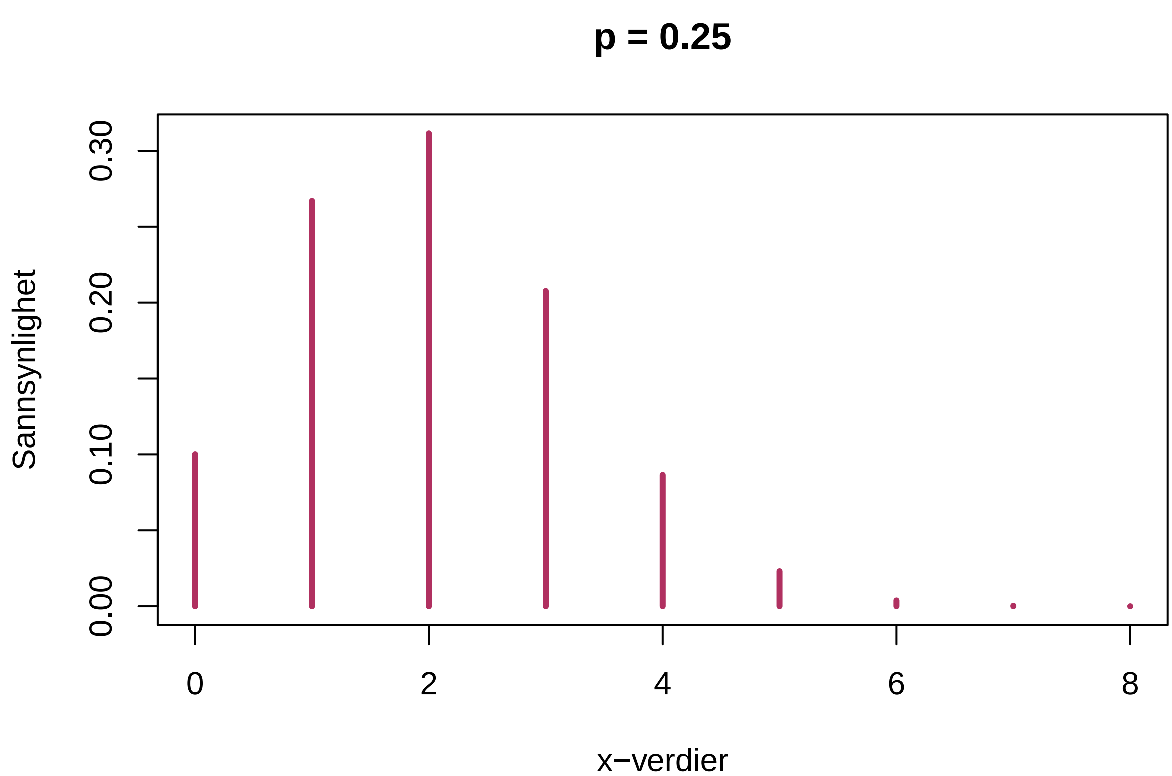
1. Prosess

a)

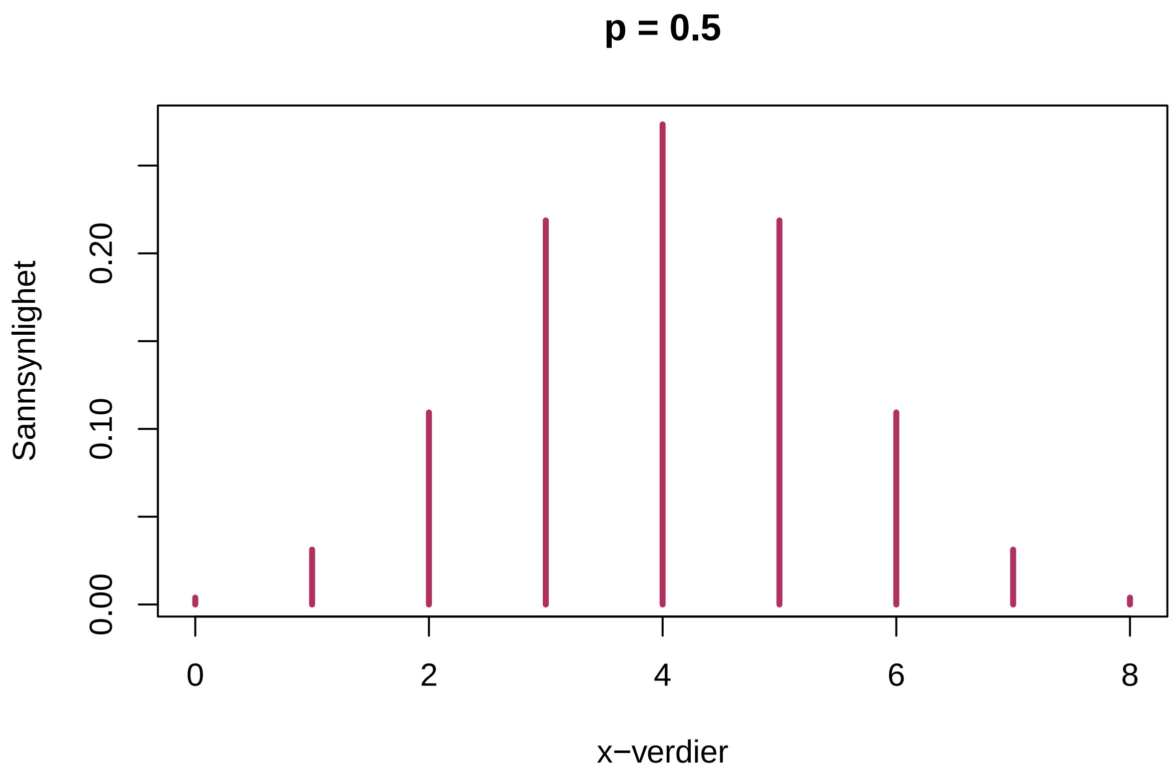
i.

Tegn sannsynlighetsfordelingene (pdf) for $X \sim bin_{(8,p)}(x)$ for p = 0.25,p = 0.5,p = 0.75, i hvert sitt diagram.

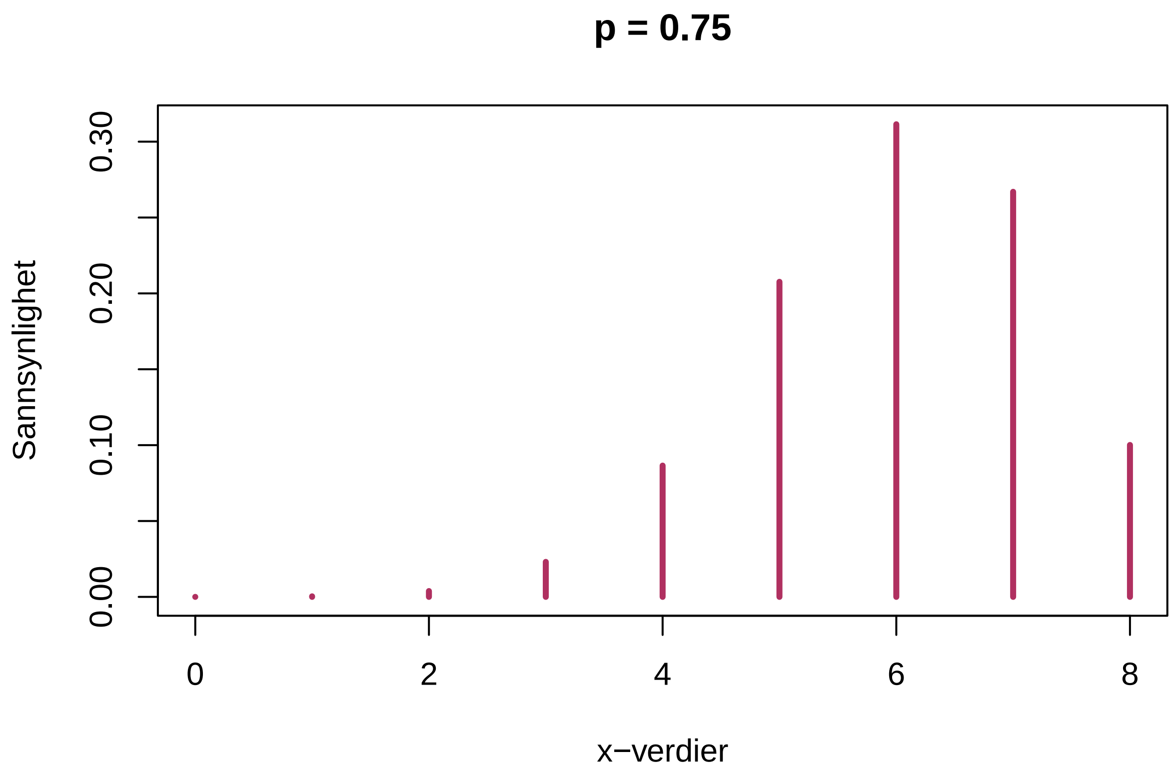
```
# Setter opp variabler for de forskjellige verdiene for p
p_1_a = 0.25
p_2_a = 0.5
p_3_a = 0.75

# Lager x-verdier som går fra 0 til 8, siden n = 8, og har step på 1 siden den skal være diskret
xVerdier = seq(0, 8, 1)

plot(xVerdier, dbinom(xVerdier, 8, p_1_a), ylab="Sannsynlighet", xlab="x-verdier",
     type="h", col="maroon", lwd=3, main="p = 0.25")
```




```
plot(xVerdier, dbinom(xVerdier, 8, p_2_a), ylab="Sannsynlighet", xlab="x-verdier",
     type="h", col="maroon", lwd=3, main="p = 0.5")
```




```
plot(xVerdier, dbinom(xVerdier, 8, p_3_a), ylab="Sannsynlighet", xlab="x-verdier",
     type="h", col="maroon", lwd=3, main="p = 0.75")
```



ii. **Hva slags effekt har en høyere verdi for p på grafen, sammenlignet med en lavere verdi?**

Da vil frekvensen ha en trend mot høyre på grafen, altså at det er flere forekomster av de høyere verdiene hvis p er stor.

iii.

Du skal slå en mynt 8 gonger. Du vinner dersom det blir presis 4 eller presis 5 mynt, men taper elles. Du kan velge mellom tre forskjellige mynter, med pn = P(mynt) respektivt p1 = 0.25, p2 = 0.5, og p3 = 0.75. Hvilken av de tre myntene gir deg størst sannsynlighet for å vinne?

Siden p'en vil gjøre at trenden går mot den prosenten i grafen (altså når $n = 8$ så vil 0.5 ha en trend på rundt 4) gir det mening å velge $p2 = 0.5$ siden den mynten vil ha størt sannsynlighet for å få en verdi rundt 4. Det kan vi se på den forrige oppgaven der en $p = 0.5$ vil ha størst sannsynlighet å ha 4 suksesser, i dette tilfelle vil én suksess være å få mynt.

b)

Du kjører på skogsveiene i Åmli, og gjennomsnittlig antall hull i veienper kilometer er lik kandidatnummeret ditt på denne eksamenen.

Mitt kandidat nummber er 285, så da vil $\lambda = 285$.

i.

Hva slags prosess må du bruke for å drive statistikk om vei 'hullene i Åmlis skogsveier, og hva er verdien på parameteren(e) for prosessen?

Her vil jeg velge å bruke en poisson-prosess ettersom vi forventer λ antall hull på den gitte avstanden, altså 285 hull per 1000 meter.

ii.

Hva er sannsynlighetsfordelingen for antall hull de neste 100 meterne? For å da finne sannsynlighetsfordelingen bruker jeg formelen for kjent λ etter θ antall sukksesser. Enheten som er i denne oppgaven er per 1000 meter, derfor blir θ her da $\frac{100m}{1000m} = 0.1$.

$$N_{+\theta} \sim pois_{\lambda\theta}(x)$$

$$N_{+\theta} \sim pois_{(285\,\cdot\,0.1)}(x)$$

iii. **Hva er sannsynligheten for at du finner mer enn 30 hull de neste 100 meterne?**

Bruker fordelingen fra forrige steg for å finne hva sannsynligheten er for $x = 30$.

```
lambda = 285
unit = 0.1 # km

sanns_30_hull = 1 - ppois(30, 285 * 0.1)
```

Da får jeg at det er en $0.3440776 \approx 0.3441 = 34.41\%$ sannsynlighet for å finne mer enn 30 hull.

c)

i.

Hvilken "pdf" og hvilken "CDF" tilhører ikke en sannsynlighetsfordeling?

Jeg vil si at pdf-fordeling c. og cdf-fordeling A. ikke tilfører en sannsynlighetsfordeling ettersom det ikke er en gjevn spredning, men det ser mer ut som ren plotting av data.

ii.

Hvilken pdf hører sammen med hvilken CDF?

- a hører med B
- b hører med D
- c hører til A
- d hører til C

iii.

Hvilke sannsynlighetsfordelinger er diskrete?

Ettersom at alle grafene har x-verdier som ikke er heltall, er ikke noen av grafene diskret siden diskret verdier må være heltall.

iv.

Hvilke sannsynlighetsfordelinger kan være sannsynlighetsfordelinger for andeler og sannsynligheter? Hvorfor?

Kontinuerlige fordelinger kan være for både sannsynligheter og for andeler.

2. Inferens

a)

Tabellversjonen av Bayes teorem: Du hører på statistikk-podcasten til to grupper. La oss kalle dem gruppe Kul og gruppe Flink

i.

Prior: La prior sannsynlighet være proporsjonal med antall podcasts hvergruppe har laget. Kul har laget 7 podcasts, Flink har laget 4. Hva er de respektive prior sannsynlighetene?

```
# Prior

# Antall kul = 7
# Antall flink = 4
# totalt = 11
ant_kul = 7
ant_flink = 4
tot_podcast = 11

# _k = kul --- _f = flink
A_k = ant_kul / tot_podcast
A_k
```

```
## [1] 0.6363636
```

```
A_f = ant_flink / tot_podcast
A_f
```

```
## [1] 0.3636364
```

ii.

begge de to gruppene trekker de lodd om hvem i gruppa som skal innledesendingen. Kul har 4 gutter og 2 jenter, mens Flink har 2 gutter og 4 jenter. Den sendingen du hører på blir innledet av en jente. Oppdatér sannsynlighetene for hvilken av gruppene du lytter til nå

```
# Likelihood

# Setter opp A som innledet av gutt, og B som innledet av jente
tot_gruppemedlemmer = 2 + 4
B_gitt_kul = 2 / tot_gruppemedlemmer
B_gitt_kul
```

```
## [1] 0.3333333
```

```
B_gitt_flink = 4 / tot_gruppemedlemmer
B_gitt_flink
```

```
## [1] 0.6666667
```

```
# Joint probability

samsannsynlighet_kul = A_k * B_gitt_kul
samsannsynlighet_kul
```

```
## [1] 0.2121212
```

```
samsannsynlighet_flink = A_f * B_gitt_flink
samsannsynlighet_flink
```

```
## [1] 0.2424242
```

```
# Total probability

tot_samsannsynlighet = samsannsynlighet_kul + samsannsynlighet_flink
tot_samsannsynlighet
```

```
## [1] 0.4545455
```

```
# Posterior

post_kul = samsannsynlighet_kul / tot_samsannsynlighet
post_kul
```

```
## [1] 0.4666667
```

```
post_flink = samsannsynlighet_flink / tot_samsannsynlighet
post_flink
```

```
## [1] 0.5333333
```

iii.

Gruppe Kul skåler for statistikken innen 5 minutt etter introen på 70% av podcastene sine. Gruppe Flink skåler ikke på sine podcasts. Hva er sannsynligheten for at de kommer til å skåle innen 5 minutter på podcasten du nå hører på?

```
# Går igjennom tabellen enda en gang, men nå bruker jeg posterior fra forrige gang som prior

# Likelihood
B_gitt_kul_2 = 0.7 # fordi de skåler 70% av tiden
B_gitt_flink_2 = 0 # fordi de aldri skåler

# Joint probability

samsannsynlighet_kul_2 = B_gitt_kul_2 * post_kul
samsannsynlighet_flink_2 = B_gitt_flink_2 * post_flink

totalsannsynlighet_skaal = samsannsynlighet_kul_2 + samsannsynlighet_flink_2
```

Da blir den totale sannsynligheten for at vi hører på en podcast det skåles innen 5 minutter $0.3266667 \approx 32.67\%$.

b)

Bernoulli-prosess: Du vil anslå π, andelen som identifiserer seg som Jedi fremfor Sith. For å få til dette, gjør du et eksperiment med Jon og Laurits. Jon og Laurits er på Outland med deg 4. mai. "May the 4th Be With You". Jon deler ut Sith-drops, mens Laurits deler ut Jedi-drops. Kundene velger selv hvilket drops de vil ta. Du teller hvor mange hver av dem får delt ut. Antallene finner du i tabellen under, Jedi i kolonne 2, og Sith i kolonne 3. Du finner dine tall i raden med ditt kandidatnummer (rad1, lengst til venstre) Eksempel: Er du kandidat 547, er Jedi=43 og Sith=20.

Mitt kandidat nr er 285, derfor blir Jedi = 41 og Sith = 22.

i.

Bruk Jeffreys' prior hyperparametre for π. Du finner observasjonene i tabellen under. Finn posterior sannsynlighetsfordeling forπ, og tegn både pdf (og cdf) for sannsynlighetsfordelingen.

```
# Sette inn Jeffreys prior som a_0 og b_0
a_0_bern = 0.5
b_0_bern = 0.5

k_bern = 41
l_bern = 22

a_1_bern = a_0_bern + k_bern
b_1_bern = b_0_bern + l_bern
a_1_bern
```

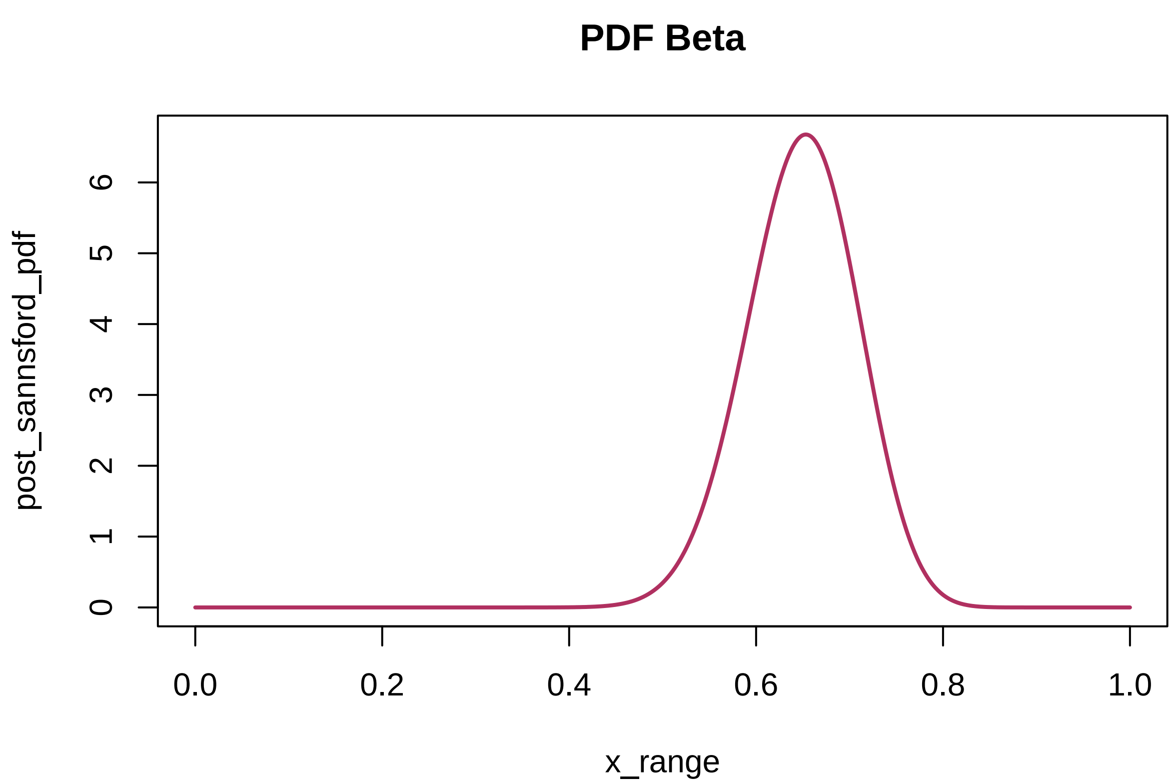
```
## [1] 41.5
```

```
b_1_bern
```

```
## [1] 22.5
```

Sannsynlighetsfordelingen for π: $\beta_{(41.5,22.5)}(t)$

```
x_range = seq(0, 1, 0.001)

post_sannsford_pdf = dbeta(x_range, a_1_bern, b_1_bern)

plot(x_range, post_sannsford_pdf, type="l", col="maroon", lwd=2, main="PDF Beta")
```



ii.

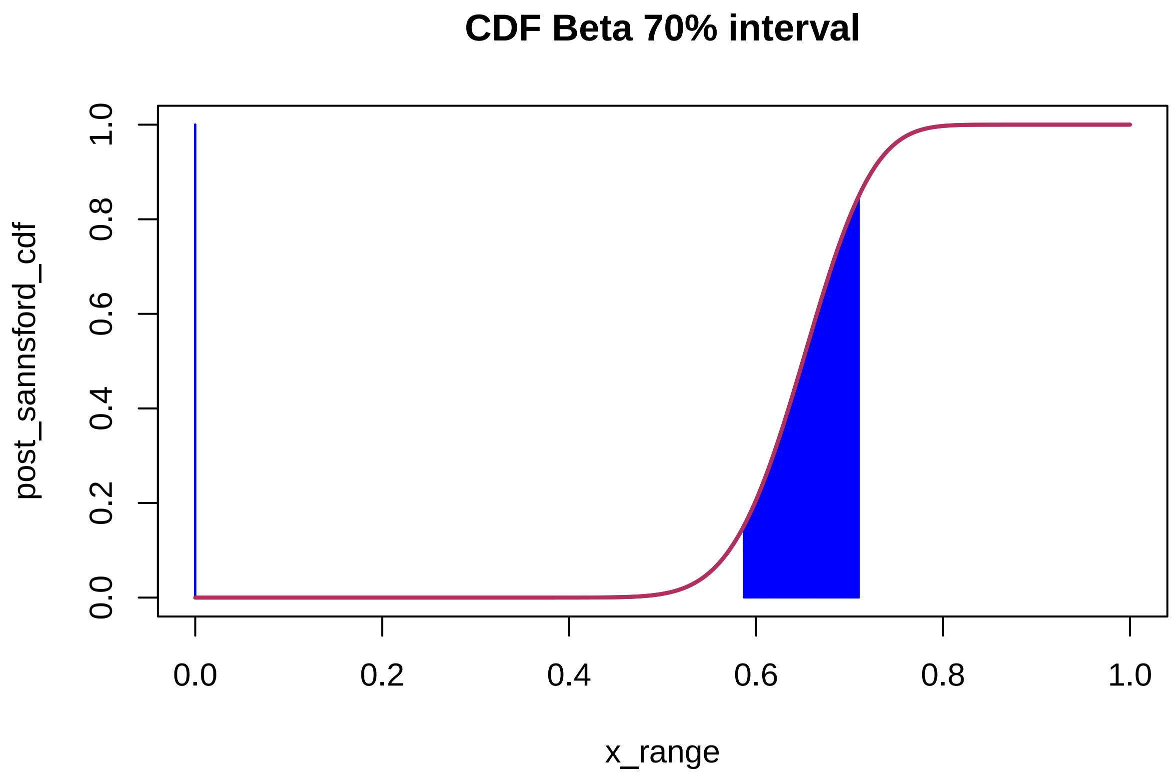
Regn ut et 70% intervallestimat ("kredibilitetsintervall") for π, tegn CDF for sannsynlighets-fordelingen for π og markér intervallestimatet på dennekurven.

```
nedre_prosent = (1 - 0.7)/2
ovre_prosent = 1 - nedre_prosent

end_interval = qbeta(ovre_prosent, a_1_bern, b_1_bern)
end_interval
```

```
## [1] 0.7100813
```

```
begin_interval = qbeta(nedre_prosent, a_1_bern, b_1_bern)
begin_interval
```

```
## [1] 0.5865321
```



iii.

Tegn en konfidenskurve for π , og markér 70% intervallestimatet for π på denne kurven.

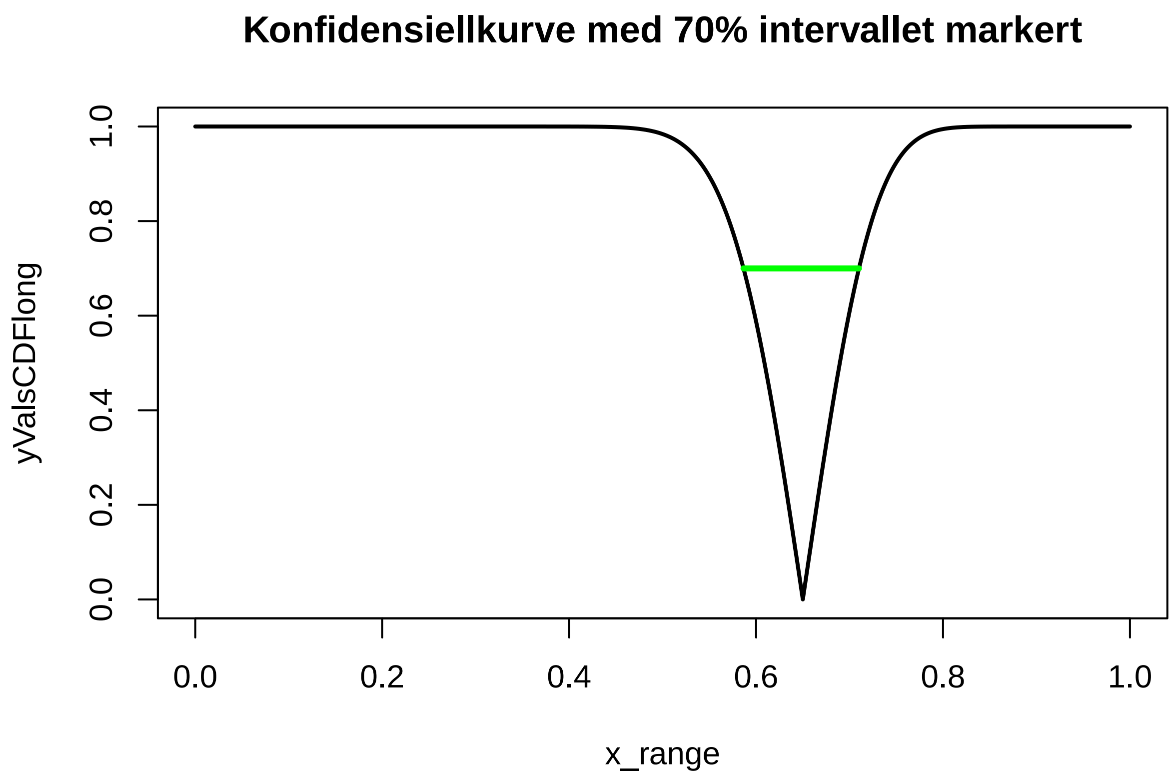
```
a_1_bern
```

```
## [1] 41.5
```

```
b_1_bern
```

```
## [1] 22.5
```

```r
yValsCDFlong=abs(2*pbeta(x_range, a_1_bern, b_1_bern) - 1)

plot(x_range, yValsCDFlong, type="l", lwd=2, main="Konfidensiellkurve med 70% intervallet markert")
segments(begin_interval, 0.7, end_interval, 0.7, col="green", lwd=3)
```



c)

Poisson-prosess: En studentgruppe på fornybar energi har gjort et bachelor-prosjekt der de blandt annet har observert oppslag om strømpriser i de største nyhetskanalene. Vi skal bruke deres data til å gjøre inferens rundt hyppigheten til disse oppslagene.

i.

Gruppen observerte 13 oppslag i de største nyhetskanalene i løpet av de 5 siste månedene av 2021. Bruk denne observasjonen sammen med nøytrale prior hyperparametre for Poisson-prosess til å finne en posterior sannsynlighetsfordeling for rateparameteren λ, gjennomsnittlig oppslag per måned.

```r
kappa_0_pois = 0
tau_0_pois = 0

n = 13
t = 5

kappa_1 = kappa_0_pois + n
tau_1 = tau_0_pois + t
```

Sannsynlighetsfordelingen blir da $\lambda \sim \gamma_{(13,\ \ 5)}(l)$

ii.

Hva er sannsynligheten for at det blir akkurat 3 slike oppslag neste måned?

Bruker sannsynlighetsfordelingen jeg fant i forrige steg

```
# Setter inn 3 for l siden det er hvor mange oppslag vi skal finne
lambda_pois = dgamma(3, kappa_1, tau_1)
lambda_pois
```

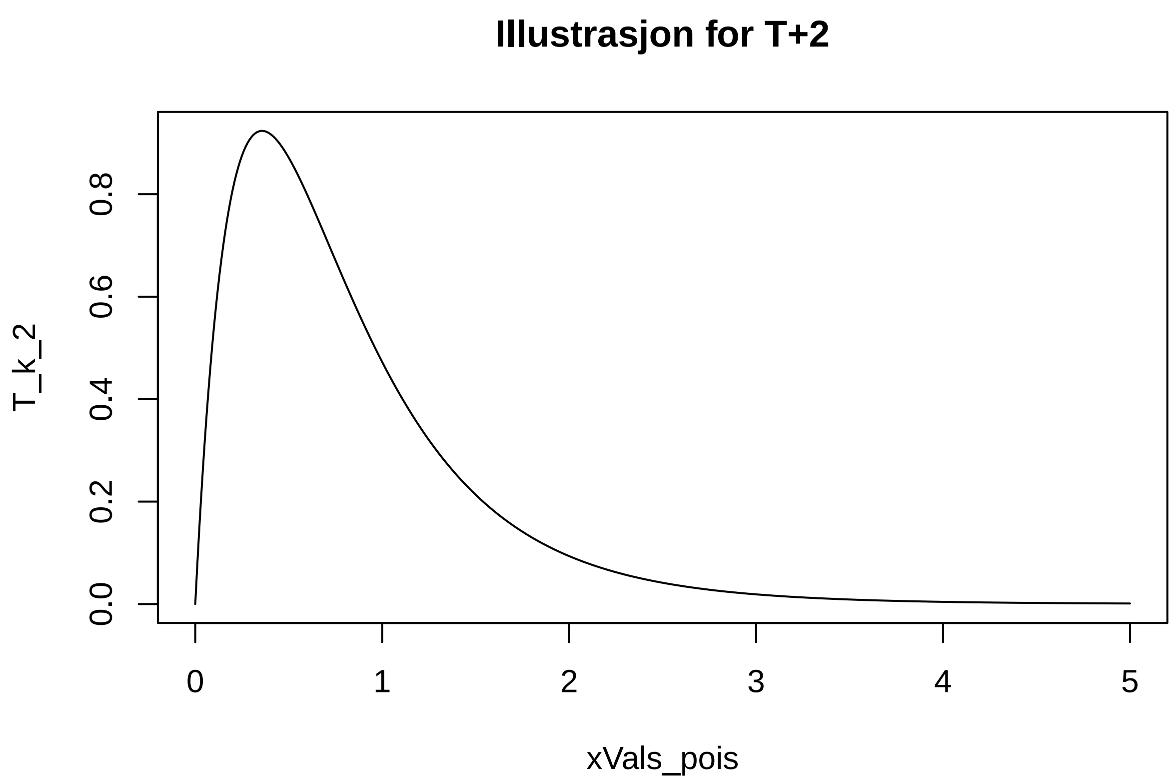
```
## [1] 0.4142962
```

$\lambda \sim \gamma_{(13,\ \ 5)}(3) \approx 0.414$

iii.

Finn sannsynlighetsfordelingen for T+2 , ventetiden på de neste 2 forekoms-tene, og regn ut et 90% intervallestimat (”prediktivt intervall”) for T+2.

```
xVals_pois = seq(0, 5, 0.01)
T_k_2 = dbetapr(xVals_pois, 2, kappa_1, tau_1)
plot(xVals_pois, T_k_2, type="l", main="Illustrasjon for T+2")
```



Fordelingen for T_{+2} blir da $\underline{\underline{T_{+2} \sim g\gamma_{(2,13,5)}(t)}}$

```
# Finner øvre og nedre verdier for grafen

nedre_prosent_pois = (1 - 0.9)/2
```

```r
ovre_prosent_pois = 1 - nedre_prosent_pois

intervall_90_estimat_pois = qbetapr(c(nedre_prosent_pois, ovre_prosent_pois), 2, kappa_1, tau_1)
intervall_90_estimat_pois
```

```
## [1] 0.1334667 2.1096878
```

Nedre estimat er 0.1334667 og øvre er da 2.1096878.

d)

Gaussisk prosess: En bachelorprosjektgruppe våren 2022 kaller seg "Gærnin-gene på Labben" (GL). De har testet forskjellige betongtyper, og vi har fått låne dataene til denne eksamenen. Vi skal se på trykkfastheten til A = Leca 300 vs. B = Leca 300 med mer sement. Vi antar at X^A , trykkfastheten for en tilfeldig prøve betong av type A, følger sannsynlighetsfordelingen $X^A \sim \phi_{(\mu,A,\sigma,A)}$, og tilsvarende for B at $X^B \sim \phi_{(\mu,B,\sigma,B)}$.

i.

De første målingene for trykkfasthet for betongtype A er: {x 1 = 20.0, x 2 = 21.5, x 3 = 20.0, x 4 = 20.2, x 5 = 18.4} (N/mm 2). Bruk nøytral prior og finn posterior fordelinger for μ A og τ A , og prediktiv fordeling for X + A .

```r
# Utregning av alle verdier når mu og sigma er ukjent
A_data_gaus = c(20.0, 21.5, 20.0, 20.2, 18.4)

n = length(A_data_gaus)

# Nøytrale prior
K_0 = 0
Sigma_0 = 0
nu_0 = -1
C_0 = 0


# Posterior hyperparametre
Sigma_X = sum(A_data_gaus)

Sigma_XX = sum(A_data_gaus^2)

SSx = Sigma_XX - n * mean(A_data_gaus)^2

K_1 = K_0 +  n
Sigma_1 = Sigma_0 + Sigma_X

m_1 = (Sigma_1 / K_1)

nu_1 = nu_0 + n
C_1 = C_0 + Sigma_XX

SS_1 = C_1 - K_1 * m_1^2

s_1_2 = SS_1 / nu_1

s_1 = sqrt(s_1_2)
```

```r
# Alle verdiene for A
n
```

```
## [1] 5
```

```r
Sigma_X
```

```
## [1] 100.1
```

```r
Sigma_XX
```

```
## [1] 2008.85
```

```r
SSx
```

```
## [1] 4.848
```

```r
K_1
```

```
## [1] 5
```

```r
Sigma_1
```

```
## [1] 100.1
```

```r
m_1
```

```
## [1] 20.02
```

```r
nu_1
```

```
## [1] 4
```

```r
C_1
```

```
## [1] 2008.85
```

```r
SS_1
```

```
## [1] 4.848
```

```r
s_1_2
```

```
## [1] 1.212
```

```r
s_1
```

```
## [1] 1.100909
```

$\tau_A \sim \gamma_{(\frac{\nu_1}{2},\ \frac{SS_1}{2})}(t)$

$\underline{\underline{\tau_A \sim \gamma_{(2,\ 2.424)}(t)}}$

$\mu_A = t_{(m_1, s_1 \cdot \sqrt{\frac{1}{K_1}}, \nu_1)}(x)$

$\underline{\underline{\mu_A = t_{(20.02, 0.4923413, 4)}(x)}}$

$X_+^A = t_{(m_1, s_1 \cdot \sqrt{1+\frac{1}{K_1}}, \nu_1)}(x)$

$\underline{\underline{X_+^A = t_{(20.02, 1.2059851, 4)}(x)}}$

ii.

```r
# Konfidenskurve med 80% intervall
nedre_prosent_gaus = (1 - 0.8)/2
ovre_prosent_gaus = 1 - nedre_prosent_gaus

nu_1/2
```

```
## [1] 2
```

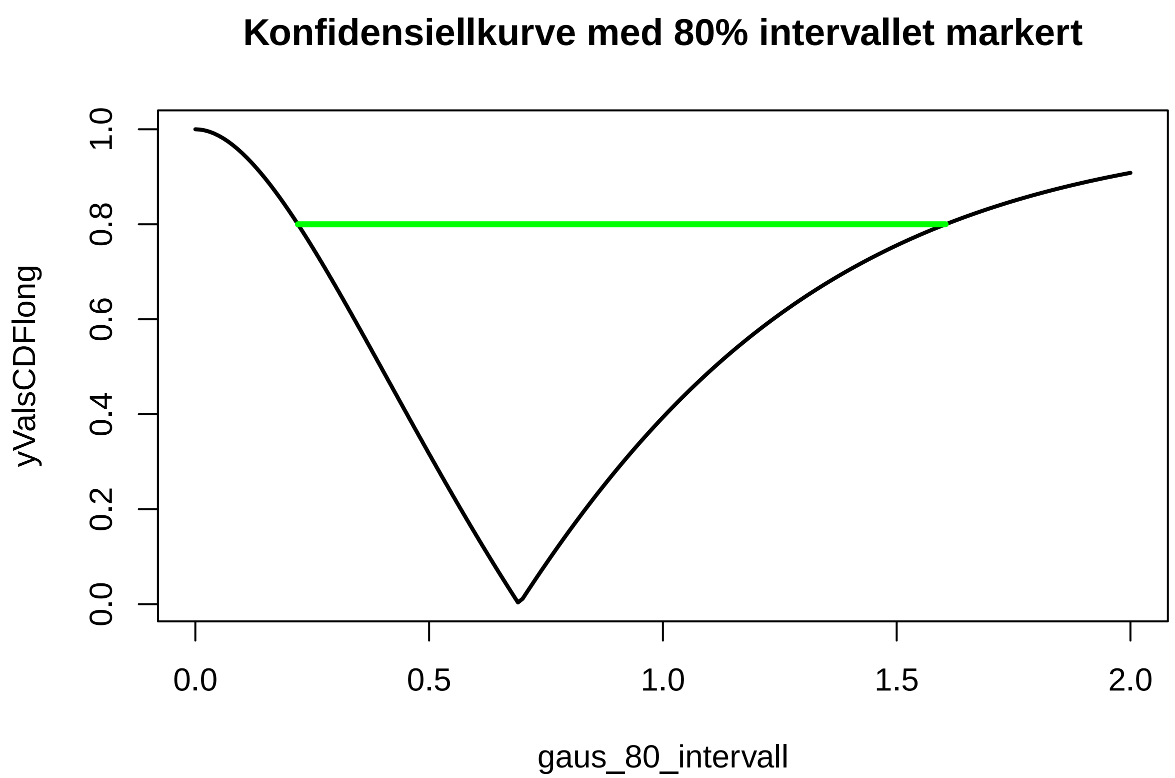
```r
SS_1/2
```

```
## [1] 2.424
```

```r
intervall_80_estimat_gaus = qgamma(c(nedre_prosent_gaus, ovre_prosent_gaus), nu_1/2, SS_1/2)
intervall_80_estimat_gaus
```

```
## [1] 0.2193942 1.6046700
```

```r
gaus_80_intervall = seq(0, 2, 0.01)

yValsCDFlong=abs(2*pgamma(gaus_80_intervall, nu_1/2, SS_1/2) - 1)

plot(gaus_80_intervall, yValsCDFlong, type="l", lwd=2,
     main="Konfidensiellkurve med 80% intervallet markert")
segments(intervall_80_estimat_gaus[1], 0.8,
         intervall_80_estimat_gaus[2], 0.8, col="green", lwd=3)
```



iii.

```
# Legger inn ny data for A
A_data_gaus_2 = c(17.3, 14.9, 19.4)

n_2 = length(A_data_gaus_2)

# Regner ut nye hyperparametre basert på parametrene fra forrige utregning
# Setter da altså <variabel>_1 + <oppdatering> med en
# gang istedet for å gjøre <variabel>_0
Sigma_X = sum(A_data_gaus_2)

Sigma_XX = sum(A_data_gaus_2^2)

SSx = Sigma_XX - n_2 * mean(A_data_gaus_2)^2

K_2 = K_1 +  n_2
Sigma_2 = Sigma_1 + Sigma_X
m_2 = (Sigma_2 / K_2)
nu_2 = nu_1 + n_2
C_2 = C_1 + Sigma_XX
SS_2 = C_2 - K_2 * m_2^2

s_2_2 = SS_2 / nu_2

s_2 = sqrt(s_2_2)
```

Får da ny fordeling for μ_A:

$\mu_A = t_{(m_2, s_2 \cdot \sqrt{\frac{1}{K_2}}, \nu_2)}(x)$

$\underline{\underline{\mu_A = t_{(18.9625, 0.7306889, 7)}(x)}}$

iv.

```
# Gjør akkurat det samme jeg gjorde for A for å finne mu for B

# Utregning av alle verdier når mu og sigma er ukjent
B_data_gaus = c(25.3, 19.7, 26.1, 21.8, 21.8, 20.6)

n = length(B_data_gaus)

# Nøytrale prior
K_0 = 0
Sigma_0 = 0
nu_0 = -1
C_0 = 0


# Posterior hyperparametre
Sigma_X = sum(B_data_gaus)

Sigma_XX = sum(B_data_gaus^2)
```

```r
SSx = Sigma_XX - n * mean(B_data_gaus)^2

K_1 = K_0 +  n
Sigma_1 = Sigma_0 + Sigma_X

m_1 = (Sigma_1 / K_1)

nu_1 = nu_0 + n
C_1 = C_0 + Sigma_XX

SS_1 = C_1 - K_1 * m_1^2

s_1_2 = SS_1 / nu_1

s_1 = sqrt(s_1_2)
```

```r
# Alle verdiene for B
n
```

```
## [1] 6
```

```r
Sigma_X
```

```
## [1] 135.3
```

```r
Sigma_XX
```

```
## [1] 3084.23
```

```r
SSx
```

```
## [1] 33.215
```

```r
K_1
```

```
## [1] 6
```

```r
Sigma_1
```

```
## [1] 135.3
```

```r
m_1
```

```
## [1] 22.55
```

```r
nu_1
```

```
## [1] 5
```

```r
C_1
```

```
## [1] 3084.23
```

```r
SS_1
```

```
## [1] 33.215
```

```r
s_1_2
```

```
## [1] 6.643
```

```r
s_1
```

```
## [1] 2.577402
```

$\mu_B = t_{(m_1, s_1 \cdot \sqrt{\frac{1}{K_1}}, \nu_1)}(x)$

$\underline{\underline{\mu_B = t_{(22.55, 1.0522199, 5)}(x)}}$

v.

```
alpha = 0.05

P_H_0_A = 1 - pt.scaled(0, 18.9625, 0.7306889, 7)
P_H_0_A
```

```
## [1] 0.5410197
```

```
P_H_0_B = 1 - pt.scaled(0, 22.55, 1.0522199, 5)
P_H_0_B
```

```
## [1] 0.5823944
```