OBLIGATORISK INNLEVERING 3, STA-1001 2024

16-04-2024

Innholdsfortegnelse


```
rm(list = ls())
```

OBLIGATORISK INNLEVERING 3, STA-1001 2024

INNLEVERINGSFRIST TIRSDAG 16/4 KL 23:59

Oppgavene er fra kapittel 8 - 9. Øvinga leveres som pdf i Canvas.

1

I denne oppgava skal vi gjøre et simuleringsforsøk for å studere fordelinga til et gjennomsnitt. Vi tenker oss at levetida, X, til en type transistorer følger gammafordeling med $\alpha = 2$ og $\beta = 1.5$:

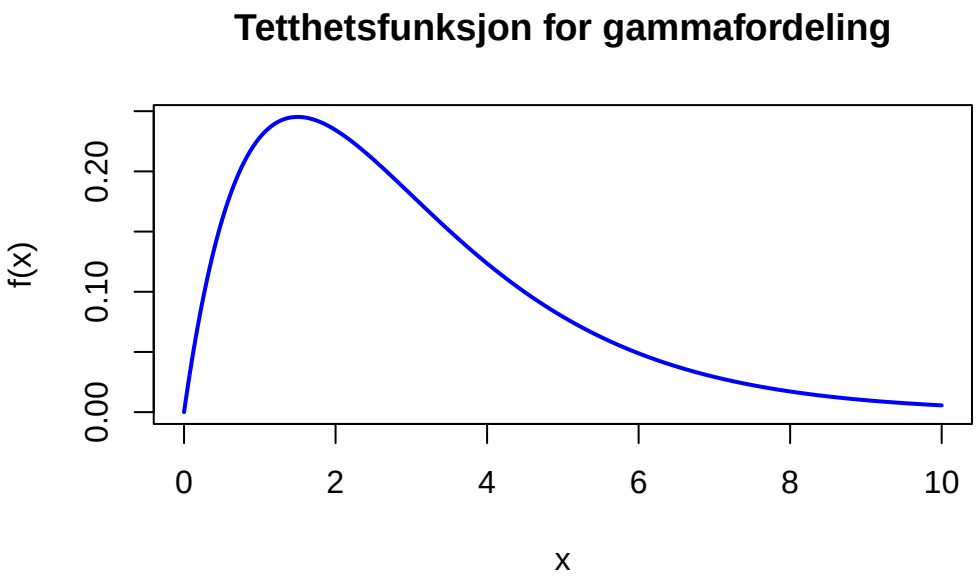
$$f(x) = \frac{1}{1.5^2} x e^{\frac{-x}{1.5}}, x > 0 \qquad E(X) = \alpha\beta = 3, \qquad SD(X) = \sqrt{\alpha\beta^2} = \sqrt{9/2}.$$

OBS: Før du gjør denne oppgava vil jeg at du setter en startverdi ("seed") for generatoren av (pseudo-)tilfeldige tall i R, for eksempel ved å bruke fødselsdatoen din (her 1/1/2001):

```r
set.seed(19970308)
```

a) Illustrer tetthetsfunksjonen for den aktuelle gammafordelinga.

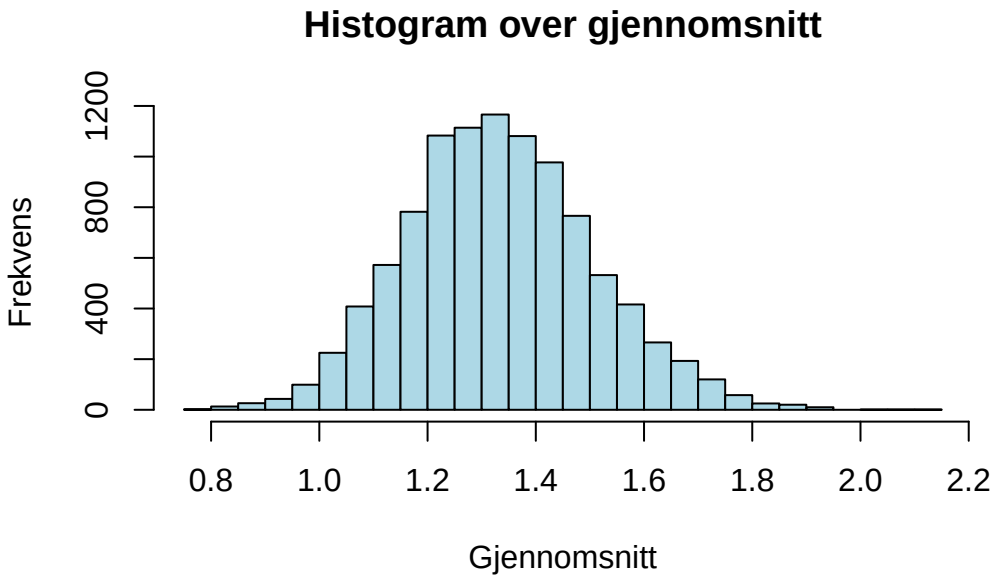
```r
x <- seq(0, 10, 0.01)

f <- function(x) {
  return(x * exp(-x/1.5) / 1.5^2)
}

plot(x, f(x), type = "l", col = "blue", lwd = 2, xlab = "x", ylab = "f(x)", main = "Tett
```



b)

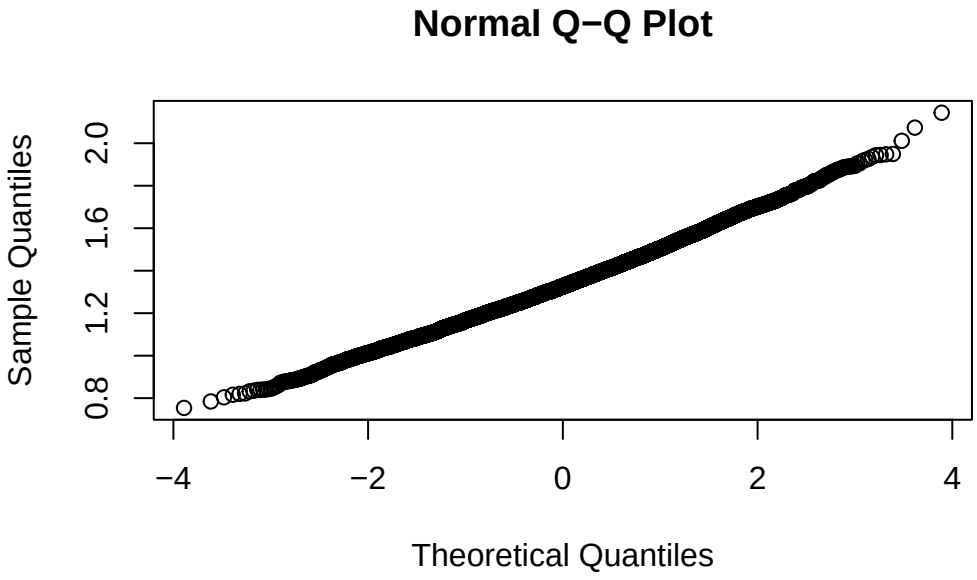
Generer 10000 tilfeldige utvalg av lengde $n = 30$ fra denne fordelinga. (Hint: funksjonene `rgamma` og `replicate`.) Fra hvert utvalg regn ut gjennomsnittet av de 30 datapunkta. Lag et histogram over gjennomsnitta. Bruk QQ-plott til å sjekke om gjennomsnitta er tilnærma normalfordelt. Hva blir konklusjonen din?

```r
n <- 30

X <- replicate(10000, mean(rgamma(n, 2, 1.5)))

hist(X, breaks = 30, col = "lightblue", main = "Histogram over gjennomsnitt", xlab = "G
```



Histogrammet ser ut til å være tilnærma normalfordelt. Vi kan også sjekke dette ved å lage et QQ-plott:

```
qqnorm(X)
```



QQ-plottet viser at gjennomsnitta er tilnærma normalfordelt. Konklusjonen er da at gjennomsnitta er tilnærma normalfordelt.

Vi ønsker å bruke simuleringene til å sjekke tilnærminga i forrige punkt. ### c) Bruk sentralgrenseteoremet til å regne ut #### 1: Tilnærma verdi for sannsynligheten for at gjennomsnittlig levetid av et tilfeldig utvalg på 30 transistorer skal gi en verdi over 3.5, altså $P(X > 3.5)$.

Bruk sentralgrenseteoremet til å regne ut

```
n <- 30
alpha <- 2
beta <- 1.5

E <- alpha * beta
SD <- sqrt(alpha * beta^2)

z <- (3.5 - E) / (SD / sqrt(n))

tabellverdi <- pnorm(z)


print(1-tabellverdi)
```

```
[1] 0.0983528
```

```
print(1- 0.9015)
```

```
[1] 0.0985
```

2:

Et tilnærma 95% prediksjonsintervall for gjennomsnittlig levetid, X, av av et tilfeldig utvalg på 30 transistorer.

```
conf <- 0.95
alpha <- 1 - conf

z <- qnorm(1 - alpha / 2)

E <- alpha * beta
SD <- sqrt(alpha * beta^2)

print(E + c(-1, 1) * z * (SD / sqrt(n)))
```

```
[1] -0.04502279  0.19502279
```

```
print(3+(E + c(-1, 1) * z * (SD / sqrt(n))))
```

```
[1] 2.954977 3.195023
```

Lyspæren vil ha en levetid på mellom 2.954977 og 3.195023 med 95% sannsynlighet.

d)

Bruk de 10000 gjennomsnitta til å gi et estimat av sannsynligheten $P(X > 3.5)$

```
mean(X > 3.5)
```

```
[1] 0
```

Er det god overenstemmelse med svaret i punkt c)?

Ja det virker sånn. Gir en 0% sannsynlighet for at $X > 3.5$.

e)

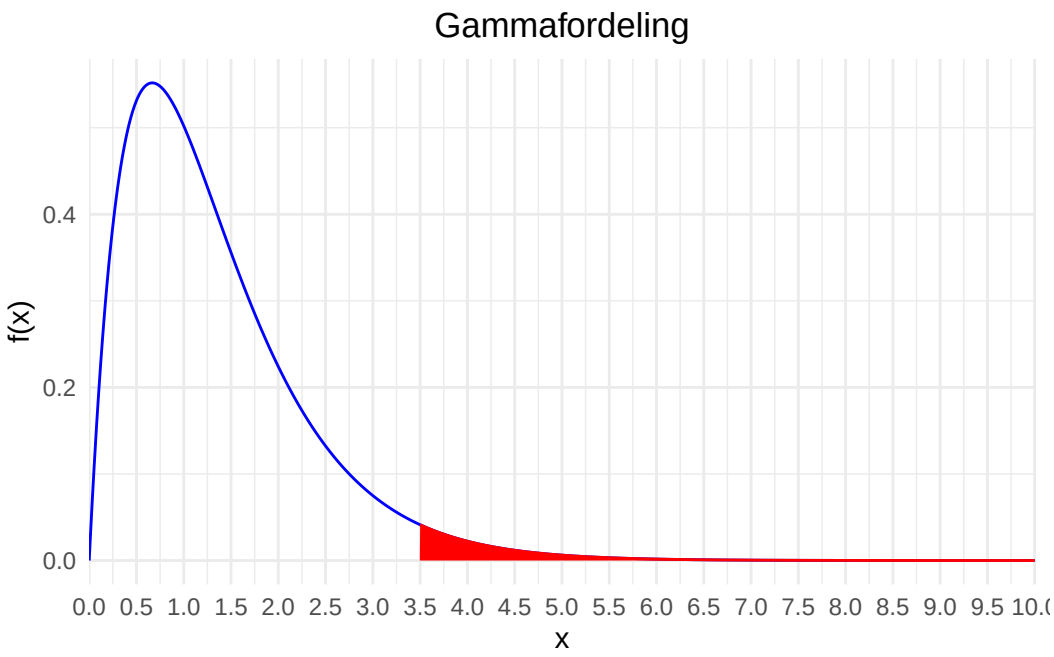
Frivillig utfordring: Bruk det du veit om fordelinga av summen av gammafordelte stokastiske variabler (og funksjonen `pgamma` i R) til å finne eksakt verdi for sannsynligheten $P(X > 3.5)$

```
1 - pgamma(3.5, 2, 1.5)
```

```
[1] 0.03279699
```

Det vi kan se er at sannsynligheten er 3.28% for at vi får verdier over 3.5. Som vist i graf i oppgave a) så kan vi se at dette ser ut til å stemme

```
library(tidyverse)


x <- seq(0, 10, by = 0.01)


df <- data.frame(x = x, y = dgamma(x, shape = 2, rate = 1.5))


df_fill <- df[df$x >= 3.5,]



df %>%
  ggplot(aes(x = x, y = y)) +
  geom_line(color = "blue") +
  geom_area(data = df_fill, fill = "red", color = "red") +
  theme_minimal() +
  labs(title = "Gammafordeling", x = "x", y = "f(x)") +
  theme(plot.title = element_text(hjust = 0.5)) +
  scale_x_continuous(n.breaks = 20, expand=c(0,0))
```



2

En kjemiker skal estimere koffeininnhold, μ, i en brustype. Dette gjøres ved å ta prøver som analyseres i en maskin. Prøveresultata fra maskinen er uavhengige og normalfordelte med forventning μ og standardavvik σ, der μ representerer korrekt koffeininnhold og σ representerer unøyaktigheten til maskinen.

Vi antar at både koffeininnholdet i brusen og unøyaktigheten til maskinen er ukjente. For å finne estimater for disse tar kjemikeren $n = 10$ prøver som analyseres. Resultatet:

```
koffein <- c(27.38, 33.71, 27.32, 29.37, 30.70, 27.68, 30.19, 26.54, 26.30, 32.05)
```

a)

Hvilke estimatorer vil du bruke til å finne estimater for koffeinnholdet μ og maskinvariansen σ^2? Hva veit du om egenskapene til disse estimatorene?

Jeg vil bruke gjennomsnittet (sample mean) som estimator for μ og den empiriske variansen (sample variance) som estimator for σ^2 hvor variansen bruker Bessel's korreksjon, for å gjøre S^2 til en unbiased estimator av σ^2. konsistent estimator for μ.

```
var(koffein)
```

```
[1] 6.256293
```

```
mean(koffein)
```

```
[1] 29.124
```

```
qt(0.975, 9)
```

```
[1] 2.262157
```

b)

Utled et 95%-konfidensintervall for koffeininnholdet μ til brustypen. Rekn ut intervallestimatet med dataene ovenfor. Hva slags konklusjon kan du trekke fra intervallet?

```
mean(koffein) - qt(0.975, 9) * (sd(koffein) / sqrt(10))
```

```
[1] 27.33471
```

```
mean(koffein) + qt(0.975, 9) * (sd(koffein) / sqrt(10))
```

```
[1] 30.91329
```

```
mean(koffein)
```

```
[1] 29.124
```

Lower er da 27.335 og upper bound er 30.91 med vårt sample mean på 29.124. Dette betyr at vi med 95% sannsynlighet kan si at koffeininnholdet i brustypen er mellom 27.335 og 30.91. $27.335 > \mu < 30.913$

c)

Dersom kjemikeren skulle ta ei prøve til, kan du finne et intervall som med 90% sannsynlighet vil inneholde prøveresultatet?

```
mean(koffein) - qt(0.9, 9) * (sd(koffein) / sqrt(10))
```

```
[1] 28.03007
```

```
mean(koffein) + qt(0.9, 9) * (sd(koffein) / sqrt(10))
```

```
[1] 30.21793
```

Han vil finne at prøveresultatet vil være mellom 28.03007 og 30.21793 med 90% sannsynlighet.

Å analysere koffeininnholdet med denne maskinen (maskin X) tar lang tid. Så for å redusere tida får kjemikeren tak i en annen og raskere maskin (maskin Y) som har noe mindre nøyaktighet. På maskin Y får han gjort $m = 20$ prøver på samme tid som maskin X bruker på $n = 10$ prøver men unøyaktigheten er 3 ganger større. La $X_1, \dots, X_n$ være resultata fra maskin X og $Y_1, \dots, Y_m$ være resultata fra maskin Y.

Vi har uavhengige tilfeldige utvalg:

$$X_i \sim N(\mu, \sigma), \quad i = 1, \dots, n \quad \text{og} \quad Y_j \sim N(\mu, 3\sigma), \quad j = 1, \dots, m$$

Han lurer på hvordan han skal kombinere resultata fra de to maskinene, og kommer fram til to mulige estimatorer:

$$\hat{\mu}_1 = \frac{\sum_{i=1}^{n} X_i + \sum_{j=1}^{m} Y_j}{n+m} \quad \text{og} \quad \hat{\mu}_2 = \frac{9\sum_{i=1}^{n} X_i + \sum_{j=1}^{m} Y_j}{9n+m}$$

d)

Vis at begge estimatorene er forventningsrette estimatorer for μ.

Estimator $\hat{\mu}_1$:

$$\mu_{\overline{X}} = E(\overline{X}_1) = E\left(\frac{\sum_{i=1}^{n} X_i + \sum_{j=1}^{m} Y_j}{n+m}\right) = \frac{1}{n+m}\left(\sum_{i=1}^{n} E(X_i) + \sum_{j=1}^{m} E(Y_j)\right) = \frac{1}{n+m}(n\mu + m\mu) = \mu$$

Estimator $\hat{\mu}_2$:

$$\hat{\mu}_2 = E(\overline{X}_2) = E\left(\frac{9\sum_{i=1}^{n} X_i + \sum_{j=1}^{m} Y_j}{9n+m}\right) = \frac{1}{9n+m}\left(9\sum_{i=1}^{n} E(X_i) + \sum_{j=1}^{m} E(Y_j)\right) = \frac{1}{9n+m}(9n\mu + m\mu)$$

e)

Finn variansene til estimatorene (uttrykt ved σ). Hvilken estimator ville du ha valgt?

Estimator $\hat{\mu}_1$:

$$\sigma^2(\hat{\mu}_1) = \frac{1}{(n+m)^2}\left(\sum_{i=1}^{n} \mathrm{Var}(X_i) + \sum_{j=1}^{m} \mathrm{Var}(Y_j)\right) = \frac{1}{(n+m)^2}(n\sigma^2 + m(3\sigma)^2) = \frac{1}{(n+m)^2}(n\sigma^2 + 9m\sigma^2) =$$

Estimator $\hat{\mu}_2$:

$$\sigma^2(\hat{\mu}_2) = \frac{1}{(9n+m)^2}\left(9\sum_{i=1}^{n} \mathrm{Var}(X_i) + \sum_{j=1}^{m} \mathrm{Var}(Y_j)\right) = \frac{1}{(9n+m)^2}(9n\sigma^2 + m(3\sigma)^2) = \frac{1}{(9n+m)^2}(9n\sigma^2 + 9$$

Siden variansen i $\hat{\mu}_2$ er mindre enn i $\hat{\mu}_1$ ville jeg valgt $\hat{\mu}_2$.

Kjemikeren vil bruke estimator $\hat{\mu}_2$. Han vil lage et konfidensintervall for koffeininnholdet μ, basert på data fra begge maskinene. Merk at han antar nå at standardavviket er kjent, $\sigma = 7$. Resultat av $m = 20$ prøver fra maskin Y:

```
koffein_Y <- c(38.46, 27.38, 42.86, 19.11, 31.25, 27.88, 25.19, 26.06, 28.23, 32.40, 22.
```

```
sigma <- 7
m = 20


z <- qnorm(1 - alpha / 2)
SE <- sigma / sqrt(m)
```

```
mean(koffein_Y) + z * SE
```

```
[1] 31.35733
```

```
mean(koffein_Y) - z * SE
```

```
[1] 25.22167
```

Gitt at dataen er normalfordelt så vil gjennomsnittet ligge mellom 25.22167 og 31.35733 med 95% sikkerhet.

f)

Bruk oppgitte verdier til å rekne ut et estimat for μ ved bruk av estimator $\hat{\mu}_2$.

Utled et 95% konfidensintervall for μ med utgangspunkt i dette estimatet.

Finn intervallestimatet med de oppgitte verdiene.

```
n <- 10


mean_koffein_hat2 <- (9 * n * mean(koffein) + m * mean(koffein_Y)) / (9 * n + m)

z <- qnorm(1 - alpha / 2)

SE_hat2 <- sigma / sqrt(9 * n + m)
```

```
mean_koffein_hat2
```

```
[1] 28.97227
```

```r
mean_koffein_hat2 - z * SE_hat2
```

```
[1] 27.66415
```

```r
mean_koffein_hat2 + z * SE_hat2
```

```
[1] 30.2804
```

Med 95% sikkerhet så kan vi si at gjennomsnittet ligger mellom 27.66415 og 30.2804

3

Kjemikeren i forrige punkt vil sammenlikne forskjellen i varians av prøver tatt med maskin X og Y i forrige oppgave. Vi antar som før at vi har to uavhengige tilfeldige utvalg:

$$X_i \sim N(\mu, \sigma_1), \quad i = 1, \dots, n \quad \text{og} \quad Y_j \sim N(\mu, \sigma_2), \quad j = 1, \dots, m$$

a)

Utled et 95% konfidensintervall for forholdet mellom variansene, $\frac{\sigma_1^2}{\sigma_2^2}$.

```r
F_dis <- var(koffein) / var(koffein_Y)
```

```r
F_dis * qf(alpha / 2, n - 1, m - 1)
```

```
[1] 0.025832
```

```r
F_dis * qf(1 - alpha / 2, n - 1, m - 1)
```

```
[1] 0.2740311
```

b)

bruk dataene fra forrige oppgave til å gi et intervallestimat. Vil du konkludere at antakelsen om at $\sigma_2 = 3\sigma_1$ fra forrige oppgave er lite rimelig?

```r
print(sqrt(var(koffein)))
```

```
[1] 2.501258
```

```
print(sqrt(var(koffein_Y)))
```

```
[1] 8.108841
```

```
 var(koffein_Y) / var(koffein)
```

```
[1] 10.50995
```

10.50995 er utenfor intervallet på 0.025832 og 0.274 så nei.

4

I denne oppgava skal vi ta for oss en stokastisk variabel Y som følger ei fordeling med sannsynlighetstetthet

$$f(y) = 10^{-\beta}\beta y^{\beta-1}, \quad 0 < y < 10$$

der parameteren $\beta > 0$.

Vi vil nytte et tilfeldig utvalg $Y_1, \ldots, Y_n$ fra fordelinga til å estimere den ukjente parameteren β.

a)

Vis at likelihoodfunksjonen (sannsynlighetsmaksimeringsfunksjonen) basert på det tilfeldige utvalget blir

$$L(\beta) = \beta^n \cdot 10^{-n\beta} \cdot \left(\prod_{i=1}^{n} y_i\right)^{\beta-1}$$

For å komme frem til det så starter vi med

$$L(\beta) = \prod_{i=1}^{n} f(y_i)$$

hvor ymax er 10

$$= \prod_{i=1}^{n} 10^{-\beta}\beta y_i^{\beta-1}$$

$$= 10^{-n\beta}\beta^n \prod_{i=1}^{n} y_i^{\beta-1}$$

$$= \beta^n \cdot 10^{-n\beta} \cdot \left(\prod_{i=1}^{n} y_i\right)^{\beta-1}$$

b)

Utled uttrykket for sannsynlighetsmaksimeringsestimatoren til β.

For å finne sannsynlighetsmaksimeringsestimatoren så tar vi logaritmen av begge sider av likelihoodfunksjonen og deriverer med hensyn på β og setter lik 0.

$$\ln(L(\beta)) = n\ln(\beta) - n\beta\ln(10) + (\beta - 1)\sum_{i=1}^{n}(\ln(y_i))$$

Deriverer så med hensyn på β og setter lik 0

$$\frac{\partial}{\partial\beta}\ln(L(\beta)) = \frac{n}{\beta} - n\ln(10) + \sum_{i=1}^{n}\ln(y_i) = 0$$

Løser så for beta

$$\frac{n}{\beta} - n\ln(10) + \sum_{i=1}^{n}\ln(y_i) = 0$$

$$\frac{n}{\beta} = n\ln(10) - \sum_{i=1}^{n}\ln(y_i)$$

Flytter over n for å få beta alene

$$\beta = \frac{n}{n\ln(10) - \sum_{i=1}^{n}\ln(y_i)}$$

Et tilfeldig utvalg på $n = 10$ observasjoner er registrert i vektoren data i R.

c)

Rekn ut estimatet for sannsynlighetsestimatoren og skisser likelihoodfunksjonen (kan gjøres i R).

```r
data <- c(6.90, 8.15, 6.19, 6.37, 7.57, 9.33, 9.64, 3.98, 8.83, 5.63)
prod(data)
```

```
[1] 298705699
```

```r
sum(log(data))
```

```
[1] 19.51497
```

```r
length(data) / (length(data) * log(10) - sum(log(data)))
```

```
[1] 2.848287
```

```r
beta <- seq(0.1, 10, 0.01)


L <- beta^length(data) * 10^(-length(data) * beta) * prod(data)^(beta - 1)
plot(beta, L, type = "l")
abline(v = length(data) / (length(data) * log(10) - sum(log(data))), col = "red")
```