STK1100
Oblig 1

Egil Furnes
Studentnummer: 693784

Oppgave 1

a)

Det er til sammen 11 etasjer å velge mellom for første person, deretter 10, 9, og 8 etasjer. Sannsynligheten for at de 4 personene går av i hver sin etasje blir følgende muligheter delt på antall mulige etasjer, i dette tilfellet 54%.

$$\frac{11}{11} \cdot \frac{10}{11} \cdot \frac{9}{11} \cdot \frac{8}{11} = \frac{11 * 10 * 9 * 8}{11^4} = \frac{7920}{14641} \approx 0.540 = 54\%$$

b)

Her bruker jeg bare komplimentet til sannsynligheten, hvor to eller flere personer går av i samme etasje bare er $1 - 54\% = 46\%$.

$$P^C = 1 - P = 1 - 0.54 = 0.46 = 46\%$$

c)

Benytter meg av en binomisk fordeling med grupper av 3 personer blant 4 mulige personer, og finner at det er totalt 4 slike grupper.

$$\binom{4}{3} = \frac{4!}{3!(4-3)!} = \frac{4 \cdot 3 \cdot 2 \cdot 1}{3 \cdot 2 \cdot 1 \cdot 1} = 4$$

d)

Sannsynligheten for at parallellkoblingen $(1, 2)$ eller $(4, 5)$ fungerer er $1 - 0.05^2 = 0.9975$ og koblingen (3) er fortsatt 0.95. Dermed er sannsynligheten for at systemet som helhet fungerer følgende:

$$0.9975 \cdot 0.9975 \cdot 0.95 = 0.9452 \approx 94.5\%$$

Oppgave 2

a)

Til tross for at anti-jukse-programmet avslører men tekst med 90% sannsynlighet, finner vi at bare 40% av tilfellene hvor det flagges for juks faktisk er det, grunnet den eksepsjonelt lave andelen av juksere i den totale befolkningen.

$$\frac{0.05 \cdot 0.90}{0.05 \cdot 0.90 + 0.95 \cdot 0.07} = 0.4035 \approx 40.4\%$$

b)

Bruker i dette tilfellet resultatet fra forrige deloppgave og setter opp en ligning, hvor vi finner at sannsynligheten for at programmet feilaktig fastlår at teksten er AI-generert må være så liten som 0.5% og ikke 10% som tidligere.

$$\frac{0.05 \cdot 0.90}{0.05 \cdot 0.90 + 0.95 \cdot x} = 0.90$$

$$x = \frac{0.05 * 0.90 - 0.05 * 0.90^2}{0.90 * 0.95} = \frac{1}{190} = 0.00526 \approx 0.5\%$$

Sannsynligheten for at programmet feilaktig anslår at teksten er AI-generert må være under 0.5% for at sannsynligheten i **a)** skal være over 90%.

Oppgave 3

```
# laster inn pakker
library(tidyverse)
library(readr)

# leser inn data lokalt fra datamaskinen
data <- read_table("doedelighet.txt") %>%
  mutate(menn = menn/1000,
         kvinner = kvinner/1000)

# definerer en funksjon for at beregne gjenverande levealder
life <- function(qx){
  n = length(qx)
  lx = numeric(n); Lx <- numeric(n); Tx <- numeric(n); ex <- numeric
     (n)
  lx[1] = 1000000
  for (i in 1:(n-1)) {
    dx = lx[i]*qx[i]
    lx[i+1] = lx[i]-dx
    Lx[i] = lx[i]-0.5*dx
  }
  Lx[n] <- lx[n]*0.5
  Tx <- rev(cumsum(rev(Lx)))
  ex <- Tx/Lx
  return(list(lx=lx, Lx=Lx, Tx=Tx, ex=ex))
}

men <- life(data$menn) %>% as_tibble()
women <- life(data$kvinner) %>% as_tibble()
aldre = c(0, 25, 50, 85)
```

a)

```
# a)
e_men = men$ex[aldre+1]
names(e_men) = paste("Age", aldre)
print("men:"); print(round(e_men,2))
```

```
[1] "men:"
 Age 0 Age 25 Age 50 Age 85
 81.32  56.79  32.79   6.22
```

Forventet gjenværende levealder for menn ved fødsel er med dette 81.32 år og følgende 56.79, 32.79 og 6.22 år for henholdsvis aldrene 25, 50 og 85.

b)

```r
# b)
e_women = women$ex[aldre+1]
names(e_women) <- paste("Age", aldre)
print("women:"); print(round(e_women, 2))
print(tibble(Age=aldre, Men=round(e_men,2), Women=round(e_women,2),
             Difference = round(e_women-e_men,2)))
```

```
[1] "women:"
 Age 0 Age 25 Age 50 Age 85
 84.69  59.97  35.59   7.28
```

Forventet gjenværende levealder for kvinner ved fødsel er med dette 84.69 år og følgende 59.97, 35.59 og 7.28 år for henholdsvis aldrene 25, 50 og 85 – altså noe lengre enn for menn!

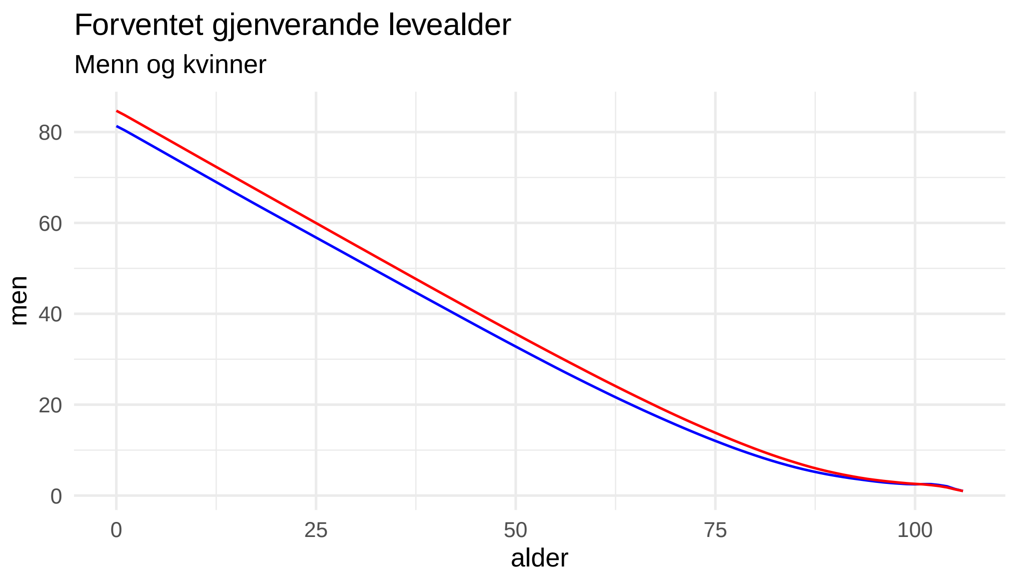
```
# A tibble: 4 × 4
    Age   Men Women Difference
  <dbl> <dbl> <dbl>      <dbl>
1     0 81.3  84.7        3.37
2    25 56.8  60.0        3.19
3    50 32.8  35.6        2.8
4    85  6.22  7.28       1.05
```

Her er utskrift av forskjellene, hvor forventet levealder for kvinner er noe høyere enn for menn.

c)

```r
# c)
tibble(alder = min(data$alder):max(data$alder),
       men = men$ex,
       women = women$ex) %>%
  ggplot(aes(x = alder)) +
  geom_line(aes(y = men), col = "blue") +
  geom_line(aes(y = women), col = "red") +
  ggtitle(label = "Forventet gjenverande levealder",
          subtitle = "Menn og kvinner") +
  theme_minimal()
```

Fra plottet ser vi at kvinner og menn har en relativt forskjellig forventet levealder fra fødselen (hvor x-aksen alder er 0), deretter ser vi at disse sannsynlighetene konvergerer mot maksalderen som ligger på litt over 100 år.



Oppgave 4

a)

Med q markører for 4 bokstaver har man 4^q typer profiler.

Sannsynligheten for at en vilkårlig person har en spesifikk DNA profil $\mathcal{S}$.

For $q = 5$

$$\frac{1}{4^5} = \frac{1}{1024} \approx 0.097\%$$

For $q = 10$

$$\frac{1}{4^{10}} = \frac{1}{1048576} \approx 0.000095\%$$

For $q = 20$

$$\frac{1}{4^{20}} = \frac{1}{1.09 \cdot 10^{12}} \approx 9.09 \cdot 10^{-11}\%$$

Med en populasjon på $N = 5500000$ individer, blir forventet antall personer som har $\mathcal{S}$ for de ulike verdiene av q.

For $q = 5$

$$5500000 \cdot \frac{1}{4^5} = 5371.09$$

For $q = 10$

$$5500000 \cdot \frac{1}{4^{10}} = 5.24$$

For $q = 20$

$$5500000 \cdot \frac{1}{4^{20}} = 5.00 * 10^{-6}$$

b)

A =Mistenkt er bidragsyter
B =Mistenkt har DNA-profil $\mathcal{S}$.

$P(B|A) = 1$ betyr med ord at mistenkt er bidragsyter gitt at mistenkt har DNA-profil $\mathcal{S}$. Vi har sett at selv med $q = 5$ utvalgte posisjoner er sannsynligheten for at en enkeltperson sin DNA-profil er $\mathcal{S}$ lik bare 0.097%. I forskning bruker man typisk $p = 5\%$ som statistisk signifikant, og selv med en relativt god margin i en kriminalsak vil 0.097% være statistisk signifikant.

For $P(B|A')$ vet vi fra tidligere at sannsynligheten for at vilkårlig person har en spesifikk DNA profil $\mathcal{S}$ er lik

$$P(B|A') = \frac{1}{4^q}$$

c)

Skal finne $P(A|B)$ gitt $N = 5500000$ og $q = 5, 10$ og 20.

Kjenner til Bayes' formel

$$P(A|B) \cdot P(B) = P(B|A) \cdot P(A)$$

Vet også følgende

$$P(A) = \frac{1}{N} = \frac{1}{5500000}, \quad P(B|A) = 1$$

$$P(A') = 1 - P(A) = \frac{N-1}{N}$$

$$P(B) = \frac{1}{N} + \frac{1}{4^q} \cdot \frac{N-1}{N} = \frac{1}{N} + \frac{N-1}{N \cdot 4^q}$$

Kan da sette inn i Bayes' formel

$$P(A|B) = \frac{\frac{1}{N}}{\frac{1}{N} + \frac{N-1}{N \cdot 4^1}} | \cdot N \cdot 4^q = \frac{4^q}{4^q + (N-1)}$$

For $q = 5$

$$P(A|B) = \frac{4^5}{4^5 + (5500000 - 1)} \approx 1.86 \cdot 10^{-4}$$

For $q = 10$

$$P(A|B) = \frac{4^{10}}{4^{10} + (5500000 - 1)} \approx 0.1601$$

For $q = 20$

$$P(A|B) = \frac{4^{20}}{4^{20} + (5500000 - 1)} \approx 0.9999$$

For q markører ser vi at sannsynligheten for at den mistenkte personen er en bidragsyter øker betraktelig. Fra nesten helt usannsynlig for $q = 5$ til nesten helt garantert for $q = 20$.

d)

Definerer $C =$ bidragsyter er et av individene i databasen. Videre er X antall personer med spor $\mathcal{S}$ innen databasen.

Hvis C *ikke* har inntruffet vet vi at ingen personer i databasen har spor $\mathcal{S}$, og at den personen er en av de andre personene i N men ikke i n.

e)

f)

g)