STK1000: Løsningsforslag Uke 39

2022

Oppgave 4.42

a) La T være hendelsen at en person bruker Twitter (eller lignende). Vi har da utfallsrommet $\{T, T^C\}$.

b) Utfallrommet er nå alle mulige kombinasjoner av de tre individene, altså

$$\{TTT, TTT^C, TT^CT, T^CTT, TT^CT^C, T^CTT^C, T^CT^CT, T^CT^CT^C\}.$$

c) Utfallsrommet for antallet (av de tre) som bruker Twitter $\{0, 1, 2, 3\}$.

d) Utfallsrommet i c) forteller bare antallet mens utfallsrommet i b) forteller også hvem av de tre individene som bruker Twitter. Om denne informasjonen er viktig kommer helt an på hva vi ønsker å bruke den til.

Oppgave 4.44

Fra Oppgave 4.42 har vi at $P(T) = 0.19$ som gir oss $P(T^C) = 0.81$. Siden hendelsene er uavhengige har vi at $P(TT) = P(T) \cdot P(T)$ osv. Dette gir oss

Utfall	TT	TT^C	T^CT	T^CT^C
Antall	2	1	1	0
Sannsynlighet	0.0361	0.1539	0.1539	0.6561

Oppgave 4.46

a) Tid er kontinuerlig.

b) Her snakker vi om et antall, som er diskret.

c) Inntekt har en minste enhet (kroner eller øre) og er derfor diskret. Men når vi har veldig fin oppløsning på en diskret variabel er det ofte praktisk å anse den som kontinuerlig (tenk hvor mange mulige inntekter det finnes mellom 0 og 1 million! 0, 0.01, 0.02, ..., 999999.98, 999999.99, 1000000).

Oppgave 4.51

a) $P(X \geq 0.40) = 1 - 0.40 = 0.60$.

b) $P(X = 0.40) = 0$.

c) $P(0.40 < X < 1.40) = P(X < 1.40) - P(X < 0.40) = P(X < 1) - P(x < 0.40) = 1 - 0.40 = 0.60$.

d) $P(0.22 < X < 0.25 \cup 0.42 \leq X \leq 0.45) = P(0.22 < X < 0.25) + P(0.42 \leq X \leq 0.45) = 0.03 + 0.03 = 0.06$.

e) $P(X < 0.5 \cup X > 0.8) = P(X < 0.5) + P(X > 0.8) = 0.5 + 0.2 = 0.7$. Alternativt kan man bruke løsningsmetoden $P(X < 0.5 \cup X > 0.8) = 1 - P(0.5 \leq X \leq 0.8) = 1 - (P(X \leq 0.8) - P(X \leq 0.5)) = 0.7$.

Oppgave 4.54

For å regne ut $P(7.1 \leq \bar{x} \leq 8.1)$, må vi først standardisere (slik vi gjorde noen øvinger tidligere). $P(7.1 \leq \bar{x} \leq 8.1) = P\left(\frac{7.1-8}{0.1342} \leq \frac{\bar{x}-8}{0.1342} \leq \frac{8.1-8}{0.1342}\right) = P(-6.7 \leq Z \leq 0.77) \approx 0.77$.

Ut fra teksten i oppgaven (den delen om ± 0.1 fra μ), skulle vi nok egentlig regne ut $P(7.9 \leq \bar{x} \leq 8.1)$: $P(7.9 \leq \bar{x} \leq 8.1) = P\left(\frac{7.9-8}{0.1342} \leq \frac{\bar{x}-8}{0.1342} \leq \frac{8.1-8}{0.1342}\right) = P(-0.77 \leq Z \leq 0.77) \approx 0.54$. Ettersom sannsynligheten er litt over 0.5 vil vi oftere enn ikke, estimere et gjennomsnitt $\bar{x}$ som er innenfor ± 0.1 fra μ.

Oppgave 4.58

Fra formelen side 237 har vi at $\mu = 0 \cdot 0.4 + 1 \cdot 0.1 + 2 \cdot 0.1 + 3 \cdot 0.2 + 4 \cdot 0.1 + 5 \cdot 0.1 = 1.8$.

Oppgave 4.62

Fra formelen side 237 har vi at $\mu = 0 \cdot 0.8507 + 1 \cdot 0.1448 + 2 \cdot 0.0045 = 0.1538$.

Oppgave 4.63

Fra side 237 i boka har vi at $\mu = -2 \cdot 0.1 - 1 \cdot 0.2 + 0 \cdot 0.4 + 1 \cdot 0.3 =$ -0.1. Fra side 245 i boka har vi at variansen er $\sigma^2 = (-2-(-0.1))^2 \cdot 0.1 + (-1-(-0.1))^2 \cdot 0.2 + (0-(-0.1))^2 \cdot 0.4 + (1-(-0.1))^2 \cdot 0.3 = 0.89$. Dette gir oss et standardavvik på $\sigma = 0.94$.

Oppgave 4.65

Null korrelasjon mellom X og Y (uavhengige, dvs korrelasjonskoeffisient $\rho = 0$)

a) $\sigma_Z^2 = 8^2\sigma_X^2 = 64 \cdot 3^2 = 576$, og $\sigma_Z = 24$.

b) $\sigma_Z^2 = 11^2\sigma_X^2 = 1089$, og $\sigma_Z = 33$.

c) $\sigma_Z^2 = \sigma_X^2 + \sigma_Y^2 = 3^2 + 2^2 = 13$, og $\sigma_Z \approx 3.606$.

d) $\sigma_Z^2 = \sigma_X^2 + \sigma_Y^2 = 3^2 + 2^2 = 13$, og $\sigma_Z \approx 3.606$.

e) $\sigma_Z^2 = 2^2\sigma_X^2 + 2^2\sigma_Y^2 = 52$, og $\sigma_Z \approx 7.211$.

Oppgave 4.67

Samvariasjon mellom X og Y: korrelasjonskoeffisient $\rho = 0.4$

a) $\sigma_Z^2 = 8^2\sigma_X^2 = 64 \cdot 3^2 = 576$, og $\sigma_Z = 24$.

b) $\sigma_Z^2 = 11^2\sigma_X^2 = 1089$, og $\sigma_Z = 33$.

c) $\sigma_Z^2 = \sigma_X^2 + \sigma_Y^2 + 2\rho\sigma_X\sigma_Y = 3^2 + 2^2 + 2 \cdot 0.4 \cdot 3 \cdot 2 = 17.8$, og $\sigma_Z \approx 4.22$.

d) $\sigma_Z^2 = \sigma_X^2 + \sigma_Y^2 - 2\rho\sigma_X\sigma_Y = 3^2 + 2^2 - 2 \cdot 0.4 \cdot 3 \cdot 2 = 8.2$, og $\sigma_Z \approx 2.86$.

e) $\sigma_Z^2 = 2^2\sigma_X^2 + 2^2\sigma_Y^2 - 2\rho\sigma_{2X}\sigma_{2Y} = 2^2 \cdot 3^2 + 2^2 \cdot 2^2 - 2 \cdot 0.4 \cdot (2 \cdot 3) \cdot (2 \cdot 2) = 32.8$, og $\sigma_Z \approx 5.73$.

Oppgave 4.73

Vi har altså $\rho = 1$. Variansen til $X + Y$ er $\sigma_{X+Y}^2 = \sigma_X^2 + \sigma_Y^2 + 2\rho\sigma_X\sigma_Y = \sigma_X^2 + 2\sigma_X\sigma_Y + \sigma_Y^2 = (\sigma_X + \sigma_Y)^2$. Dette gir standardavviket til summen av X og Y verdi $\sigma_{X+Y} = \sigma_X + \sigma_Y$.

Oppgave 4.76

Hvis man selger kun 5 forsikringer vil man mest sannsynlig ikke få noen krav, men hvis man er riktig uheldig og får selv ett enkelt krav vil det bli veldig dyrt for selskapet (mye dyrere enn det man har fått inn gjennom slaget av 5 forsikringer).

Hvis man i stedet selger tusenvis av forsikringer vil det gjennomsnittlige kravet fra kundene være veldig nær det sanne gjennomsnittet på \$300. Man har derfor kontroll på risikoen til selskapet, og kan jevnt tjene penger til å dekke øvrige driftsutgifter etc ettersom prisen for forsikringene vil (mest sannsynlig) dekke forsikringskravene.

Oppgave 4.77

For forsikringskrav $X_1, X_2, \ldots, X_5$ er vi interessert i gjennomsnittet $\bar{x} = \frac{1}{5}\sum_{i=1}^{5} X_i$. Fra regnereglene om forventningsverdi vet vi at $\mu_{\bar{x}} = \frac{1}{5}\sum_{i=1}^{5} \mu = \mu = \$300$. Fra regnereglene om varians har vi $\sigma^2_{\bar{x}} = \frac{1}{5^2}\sum_{i=1}^{5} \sigma^2 = \frac{1}{5}\sigma^2 = \frac{400^2}{5} = 32000$. Vi har derfor $\sigma_{\bar{x}} \approx 178.9$.

Vi gjentar samme beregninger med 20 i stedet for 5 forsikringer og får $\mu_{\bar{x}} = 300$ og $\sigma_{\bar{x}} \approx 89.4$. Merk at $\mu_{\bar{x}}$ ikke påvirkes av antall forsikringer mens $\sigma_{\bar{x}}$ blir mindre jo flere forsikringer vi har. Forsikre deg om at du forstår hvordan dette er relatert til konklusjonen i oppgave 4.76.

Oppgave 4.90

La A betegne nok søvn og B nok trening. Vi har da $P(A) = 0.46$, $P(B) = 0.40$ og $P(A \cap B) = 0.27$.

a) $P(A \cap B^C) = P(A) - P(A \cap B) = 0.46 - 0.27 = 0.19$.

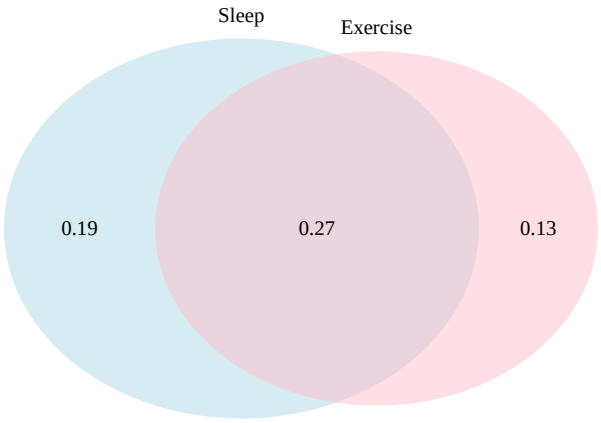
b) $P(A^C \cap B) = P(B) - P(A \cap B) = 0.40 - 0.27 = 0.13$.

c) $P(A^C \cap B^C) = 1 - P(A \cup B) = 1 - [P(A) + P(B) - P(A \cap B)] = 1 - (0.46 + 0.40 - 0.27) = 0.41$.

d) Her vil svarene variere etter hvordan man har løst oppgavene. I a) har vi brukt en variasjon av komplimentregelen (*compliment rule*) og regelen for total sannsynlighet, $P(A) = P(A \cap B) + P(A \cap B^C)$. Denne refereres ofte til som **loven om total sannsynlighet** (*law of total probability*), og kan være nyttig å lære seg. I b) har vi brukt det samme regel som i a) (bare bytt om A og B). I c) har vi brukt addisjonsregelen og komplimentregelen ved at $P(A^C \cap B^C) = P((A \cup B)^C) = P(C^C) = 1 - P(C) = 1 - P(A \cup B)$. Dette er lettest å se med et venn-diagram (se neste oppgave)!

Oppgave 4.91

Følgende figur viser et Venn-diagram av sannsynlighetene. En skisse kan med fordel også inkludere området rundt $(A^C \cap B^C)$ med areal 0.41.



Oppgave 4.94

a) Her bruker vi Bayes for å fylle inn tabellen. For eksempel har vi at

$$P(\text{Mann} \cap \text{Fire år}) = P(\text{Fire år}) \cdot P(\text{Mann} \mid \text{Fire år}) = 0.59 \cdot 0.44 = 0.2596.$$

Videre vet vi at $P(\text{To år}) = 1 - P(\text{Fire år}) = 0.41$, $P(\text{Mann} \mid \text{To år}) = 0.40$, og $P(\text{Kvinne} \mid \text{x år}) = P(\text{Mann}^C \mid \text{x år}) = 1 - P(\text{Mann} \mid \text{x år})$. Vi har da alle verktøyene til regne ut de resterende feltene.

	Men	Kvinner	Totalt
Fire år	0.2596	0.3304	0.59
To år	0.1640	0.2460	0.41
Totalt	0.4236	0.5764	1

b) Vi bruker Bayes for å gjøre regnestykket om til å bare inneholde tall vi kan lese av fra tabellen vi har laget i a).

$$P(\text{Fire år} \mid \text{Kvinne}) = \frac{P(\text{Fire år} \cap \text{Kvinne})}{P(\text{Kvinne})} = \frac{0.3304}{0.5764} \approx 0.57.$$

Oppgave 4.97

Vi må skrive om uttrykket så det bare inneholder sannsynligheter vi kjenner. Med addisjonsregelen har vi at

$$P(A \cup B) = P(A) + P(B) - P(A \cap B) = 0.138 + 0.261 - 0.082 = 0.317.$$

Oppgave 4.98

Vi kan bruke Bayes her

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)} = \frac{0.082}{0.261} = 0.3142.$$

Vi vet at $P(A) = 0.138 \neq P(A \mid B)$. Da kan ikke A og B være uavhengige.

Oppgave 5.2

a) En parameter beskriver en befolkning, ikke et utvalg.

b) Variabilitet handler om spredning mens bias handler om spredningen er rundt riktig verdi (om snittet er lik parameteren). Hvis man bruker spillet dart som en analogi er variabilitet hvor stor spredning det er på kastene dine, mens bias er hvor nært det gjennomsnittlig kastet er midten av skiven. Dette er illustrert i figur 5.4 i boken.

c) Generelt ja, men utsagnet er litt upresist. Hvis det mindre og det større utvalget er trykket på samme måte vil det større utvalget gi mindre variabilitet i estimatene dine. Hvis det er trukket på en annen måte kan vi ikke si hvilket som er best. Nedsider med store utvalg er at de kan være tunge å regne på og at det kan være mye arbeid og/eller dyrt å få tak i dem.

d) En utvalgsfordeling ('sampling distribution') er en teoretisk fordeling. Den oppsummerer ikke bare hvilke verdier en observator ('statistic') kan ta for alle mulige utvalg av samme størrelse fra populasjonen, men beskriver hvordan observatoren varierer i (det muligens hypotetiske) scenariet av repeterte forsøk. Merk også at utvalgsfordelingen naturlig nok også avhenger at størrelsen på utvalget, og denne detaljen manglet i påstanden d).

Oppgave 5.6

a) Høy bias og høy variabilitet.

b) Lav bias og lav variabilitet.

c) Lav bias og høy variabilitet.

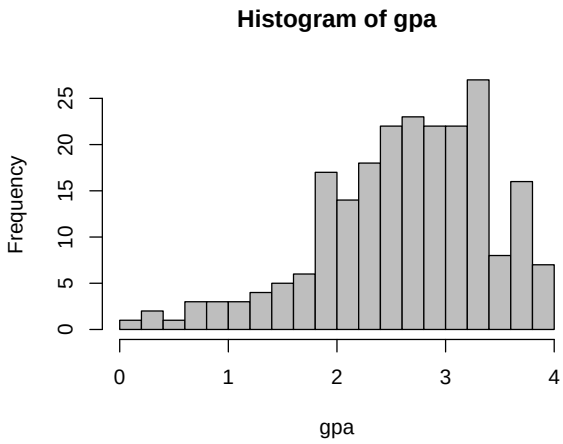
d) Høy bias og lav variabilitet.

Fra ark 3.93(R)

a)

Vi kan lese inn dataene fra url'en. Merk at vi ikke her har en komma-separert csv fil, men vi har "tab" som separerer kolonnene. Videre bruker file komma i stedet for punktum for å gi desimaler. Vi må derfor legge til noen ekstra argument til `read.csv` for at filen skal leses riktig.

```
url = 'https://www.uio.no/studier/emner/matnat/math/STK1000/h18/csdata.txt'
data = read.csv(url, sep = '\t', dec = ',')
gpa = data$gpa
hist(gpa, col = 'gray', breaks = 20)
```




```
summary(gpa)
```

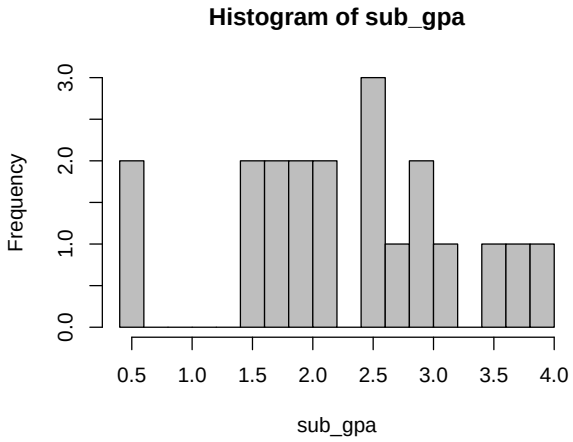
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
```

```
##   0.120   2.167   2.740   2.635   3.212   4.000
```

Vi ser at dataene er litt skjevfordelte med et gjennomsnitt 2.635.

b)

```
sub_gpa = sample(gpa, 20)
hist(sub_gpa, col = 'gray', breaks = 20)
```



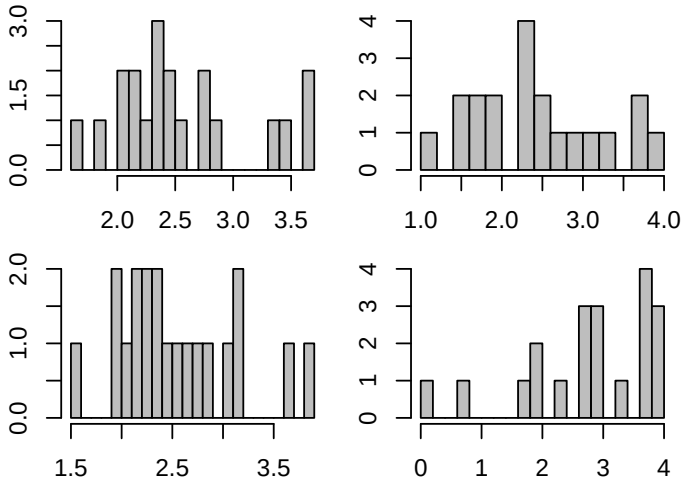

```
summary(sub_gpa)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.400   1.690   2.350   2.290   2.875   3.860
```

Vi ser at vi får et litt annet gjennomsnitt enn for det fulle data settet. Vi ser også at fordelingen ikke er helt like.

c) I koden under gjenntar vi opperasjonene i b) fire ganger. `par` kommandoen gjør at vi kan plott alle fire ved siden av hverandre, men det er ikke så viktig å kunne bruke.

```
par(mfrow=c(2, 2), oma = c(5,4,0,0) + 1, mar = c(0,0,1,1) + 1)
sub_gpa = sample(gpa, 20)
hist(sub_gpa, col = 'gray', breaks = 20, main = '')
sub_gpa = sample(gpa, 20)
hist(sub_gpa, col = 'gray', breaks = 20, main = '')
sub_gpa = sample(gpa, 20)
hist(sub_gpa, col = 'gray', breaks = 20, main = '')
sub_gpa = sample(gpa, 20)
hist(sub_gpa, col = 'gray', breaks = 20, main = '')
```



Vi ser at vi får ganske forskjellige fordelinger. Altså ser ikke vårt lille utvalg ut til å være veldig representativt for populasjonen.

Fra ark 3.94(R)

a)

Når vi skal gjenta en operasjon mange ganger har vi noen verktøy vi kan bruke i R. En for-løkke trolig den enkleste måten å gjøre dette på.

```
gpa_means = numeric(25) # Vi lager 20 nullerer, tilsvarende c(0, 0, 0, 0, ... 0).
for (i in 1:25) { # Dette gjenntar opperasjonene under med i=1, i=2, ..., i=25
  sub_gpa = sample(gpa, 20) # Vi trekker et tilfeldig utvalg på 20 individer.
  mean_sub = mean(sub_gpa) # Vi regner ut gjennomsnittet.
  gpa_means[i] = mean_sub # Vi setter in gjennomsnittet på posisjon "i" i listen med tall.
}
gpa_means
```

```
##  [1] 2.2900 2.5555 2.4485 2.5610 2.7850 2.7370 2.4830 2.5750 2.7730 2.9690 2.9485
## [12] 2.3785 2.7500 2.4585 2.6110 2.7080 2.5145 2.6550 2.3880 2.9035 2.5030 2.5970
## [23] 2.5435 2.6550 2.4475
```

Vi plotter resultatene på samme måte som tidligere i oppgave c)

Nå som du har forstått en for-løkke kan du prøve å skrive om 4.94 c) med en tilsvarende for-løkke.

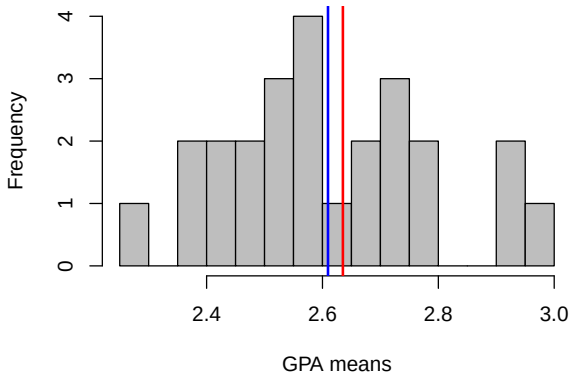
b)

I Oppgave c) lager vi histogrammet over gjennomsnittene i a). Vi merker også gjennomsnittet i populasjonen med en rød linje. Histogrammet ser ut til å være distribuert rundt populasjonsgjennomsnittet uten noe videre bias.

c)

Vi regner ut gjennomsnitt og standardavvik for de 25 gjennomsnittene funnet i a). I figuren under har vi laget histogrammet over gjennomsnittene og med en rød linje som gir populasjonsgjennomsnittet og den blå linjen gir gjennomsnittet av de 25 utvalgene.

```
hist(gpa_means, col = 'gray', breaks = 20, main = '', xlab = 'GPA means')
abline(v = mean(gpa), col = 'red', lwd = 2)
abline(v = mean(gpa_means), col = 'blue', lwd = 2)
```




```
c(mean(gpa), sd(gpa)) # Fra populasjonen
```

```
## [1] 2.6352232 0.7793949
```

```
c(mean(gpa_means), sd(gpa_means)) # Fra de 25 utvalgene
```

```
## [1] 2.6095400 0.1785101
```

Vi ser at gjennomsnittet av de 25 gjennomsnittene er ganske nærme populasjonsgjennomsnittet. Dette er litt tilfeldig, så det kan være at du får litt annerledes resultater enn dette. Vi ser også (som forventet) at standardavviket til hele populasjonen er større en for vår 25 gjennomsnitt av tilfeldige utvalg.

Midtveiseksamen Høsten 2008

5) $P(X < 10) = P(Z < \frac{10-15}{5}) = P(Z < -1) = 0.1587$ som gir c).

Midtveiseksamen Høsten 2011

9) B).

10) C).

15) $P(X \geq 3) = 1 - P(X \leq 2) = 1 - (0.1 + 0.2) = 0.7$. Altså D).

17) $P(X < 1 \cup X > 1.8) = P(X < 1) + P(X > 1.8) = \frac{1+0.2}{2} = 0.6$. Altså D)

Midtveiseksamen Høsten 2012

11) $P(X < 5) = \frac{4}{9} = 0.44$, som gir C)

12) Vi krever at arealet skal være 1, som vil si at den ene halvdelen må være 0.5. Vi vet da at $\frac{0.5 \cdot h}{2} = 0.5$, som gir $h = 2$ (husk at arealet av en rettvinklet trekant er halve arealet ev tilsvarende firkant med sider gitt de to katetene). Dette gir A).

13) C).

14) D).

15) $P(A \cup D) = P(D)$, som gir C).