Løsningsforslag til oppgaver brukt i STA100

Oppgave 1

a) Populasjonen er alle studenter ved Universitetet i Stavanger, og utvalget er de (ca 100) studentene hun velger ut i undersøkelsen sin.

b) Fordelen med metode 1 er at den er lett og rask å utføre, ulempen er at den gir et svært lite representativt utvalg - kun studenter som har møtt på en bestemt forelesing i ett bestemt kurs blir spurte, disse trenger slett ikke være representative for hva studentene generelt på Universitetet mener om de aktuelle spørsmålene.
Metode 2 er ikke så raskt utført som metode 1, men er klart raskere og lettere å utføre enn metode 3. Metode 2 vil gi et mer korrekt resultat enn metode 1 da et bredere spekter av studenter vil bli spurt, dvs utvalget blir mer representativ. Utvalget vil imidlertid fremdeles ikke være helt representativt da studenter som sjelden eller aldri er på Universitetsområdet har liten sannsynlighet for å bli spurt (og det godt kan tenkes at disse har en annen mening om de aktuelle spørsmålene enn de som ofte er på Universitetet).
Metode 3 er den klart mest tidkrevenede å utføre, men fordelen er at man her velger ut studentene som blir spurt helt tilfeldig blant alle studenter ved Universitetet (tilfeldig utvalg) og vi vil dermed få en undersøkelse basert på et representativt utvalg av studentene. Metode 3 er derfor den beste.
(Grunnen til at metode 3 er den beste er at man her ønsker å finne ut noe om hva alle studentene ved UiS mener, da må man utføre undersøkelsen slik at alle studenter har like stor sannsynlighet for å bli spurte. Dersom man f.eks. kun er interesserte i hva de studentene som er aktive brukere av Universitetsområdet mener kan metode 2 være en god fremgangsmåte.)

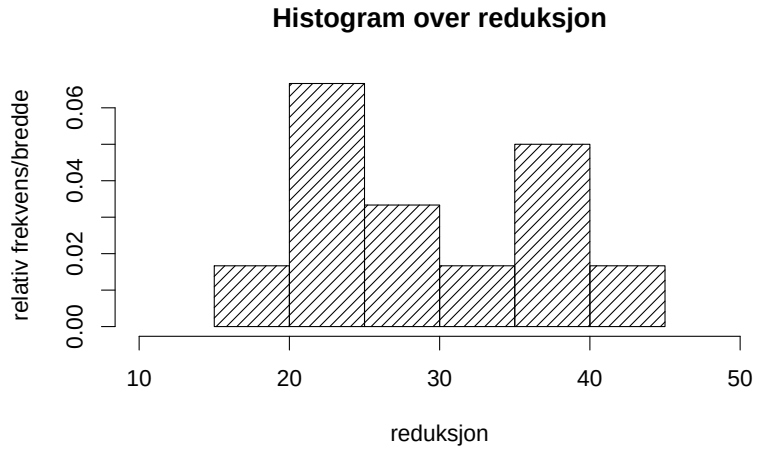
Oppgave 2

a) $\bar{x} = (17 + 22 + 37 + 36 + 22 + 29 + 42 + 23 + 22 + 28 + 36 + 33)/12 = \underline{\underline{28.9}}$
Dataene i sortert rekkefølge:

17 22 22 22 23 28 29 33 36 36 37 42

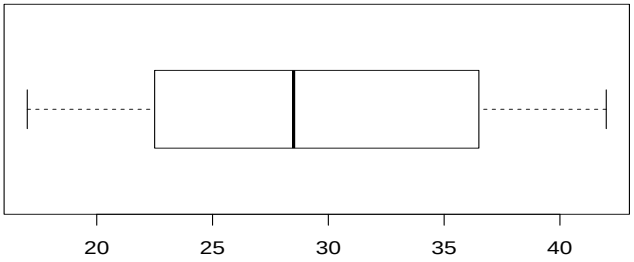
Siden vi har et partall observasjoner må vi ta gjennomsnittet av de to som er mest i midten - dvs medianen blir $\tilde{x} = (28 + 29)/2 = \underline{28.5}$.
Modus (=den verdien som opptrer oftest) ser vi er $\underline{\underline{22}}$.
b) Utvalgsvarians: $s^s = \frac{1}{12-1}\sum_{i=1}^{12}(x_i - \bar{x})^2 = \underline{\underline{61.4}}$.
Utvalgsstandardavvik: $s = \sqrt{61.4} = \underline{\underline{7.8}}$

c) Hvordan histogrammet blir seende ut avhenger av hvordan man velger intervallgrensene - i eksemplet under er intervallgrensene [15,20), [20,25), [25,30), [30,35), [35,40), [40,45) brukt.



d) Nedre kvartil er verdien som har 25% av målingene mindre enn eller lik seg. Her utgjør 3 målinger 25% av dataene så nedre kvartil ligger mellom måling nr 3 og 4 i sortert rekkefølge. Begge disse målingene har verdi 22, nedre kvartil er derfor $\underline{\underline{22}}$. Tilsvarende blir øvre kvartil $\underline{\underline{36}}$. Kvartilbredden blir da 36-22=$\underline{\underline{14}}$. Variasjonsbredden blir 42-17=$\underline{\underline{25}}$.

e) Med median og kvartiler som regnet ut over og største og minste verdi som gitt i datasettet blir boksplottet som vist under (se også figur 2.14 i boka).



Oppgave 3

Alle fire datasettene har $\bar{x} = 4.5$.

i)

$$\begin{aligned} s^s &= \frac{1}{8-1}\sum_{i=1}^{8}(x_i - 4.5)^2 \\ &= \frac{1}{7}[(1-4.5)^2 + (2-4.5)^2 + (3-4.5)^2 + (4-4.5)^2 \\ &\quad +(5-4.5)^2 + (6-4.5)^2 + (7-4.5)^2 + (8-4.5)^2] = \underline{6} \end{aligned}$$

ii) $s^2 = \underline{\underline{14}}$

iii) $s^2 = \underline{\underline{7.14}}$

iv) $s^2 = \underline{54}$

Dvs rekkefølgen blir i), iii), ii), iv).

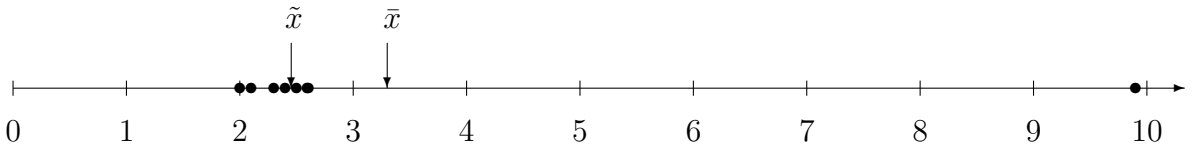
Oppgave 4

a) $\bar{x} = \frac{1}{8}(2.6 + 2.1 + 2.3 + 2.6 + 2.5 + 2.0 + 2.4 + 9.9) = \underline{3.3}$

Dataene i sortert rekkefølge:

2.0 2.1 2.3 2.4 2.5 2.6 2.6 9.9

Siden vi har et partall observasjoner må vi ta gjennomsnittet av de to som er mest i midten - dvs medianen blir $\tilde{x} = \frac{1}{2}(2.4 + 2.5) = \underline{2.45}$.



Vi ser at gjennomsnittet, $\bar{x}$, blir mye påvirket av den ene målingen som er mye større enn de andre. Medianen, $\tilde{x}$, påvirkes ikke i det hele tatt av en slik "ekstremmåling" og ligger midt inne i området hvor hoveddelene av dataene befinner seg.
Dersom vi har ekstremmålinger som skyldes feilregistreringer e.l. blir altså ikke medianen påvirket av dette, mens gjennomsnittet påvirkes. Vi sier derfor gjerne at medianen er et mer robust mål på beliggenhet enn gjennomsnittet.

b) $\bar{x} = \frac{1}{7}(2.6 + 2.1 + 2.3 + 2.6 + 2.5 + 2.0 + 2.4) = \underline{2.36}$

Dataene i sortert rekkefølge:

2.0 2.1 2.3 2.4 2.5 2.6 2.6

$\tilde{x} = \underline{2.4}$. Merk at medianen og gjennomsnittet nå er nesten like!

$s^s = \underline{0.0562}$ og $s = \sqrt{0.0562} = \underline{0.24}$

Tommelfingerregelen fra forelesing om at omtrent 95% av målingene vil ligge i intervallet $[\bar{x} - 2s, \bar{x} + 2s]$ gir oss anslaget at omtrent 95% av målingene vil ligge i intervallet $[2.36 - 2 \cdot 0.24, 2.36 + 2 \cdot 0.24] = \underline{[1.88,\ 2.84]}$.

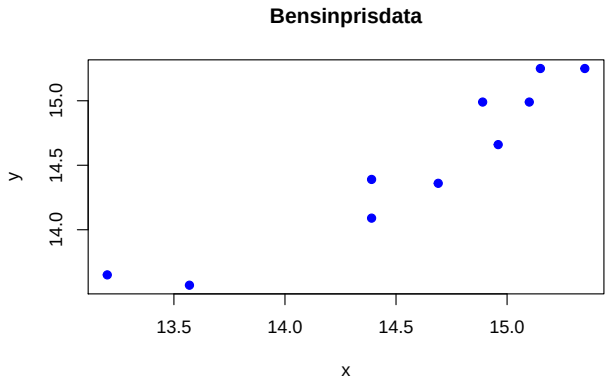
Oppgave 5

a)

$$\begin{aligned}
\overline{x} &= \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{1}{10}145.69 = \underline{14.57}, \qquad \overline{y} = \frac{1}{n}\sum_{i=1}^{n} y_i = \frac{1}{10}145.20 = \underline{14.52} \\
s_x &= \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2} = \sqrt{\frac{1}{10-1}4.436} = \underline{0.70} \\
s_y &= \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(y_i - \overline{y})^2} = \sqrt{\frac{1}{10-1}3.414} = \underline{0.62}
\end{aligned}$$

Både gjennomsnitt og standardavvik er veldig like for de to datasettene. Dvs prisene ved de to bensinstasjonene ser ut til å tendere til å ligge på samme nivå og til å ha omtrent lik variasjon (marginalt lavere variasjon i prisene ved stasjon nr 2/y-dataene).

b)



Vi ser en sterk sammenheng, bensinprisene følger hverandre i stor grad (går opp og ned i takt).

Oppgave 6

La O_1 = finne olje i reservoar 1, og O_2 = finne olje i reservoar 2. Vi har da: $P(O_1) = 0.6$, $P(O_2) = 0.7$, $P(\text{finne olje i begge reservoarene}) = P(O_1 \cap O_2) = 0.45$.

a) Finne olje i minst ett av reservoarene $= O_1 \cup O_2$:
$P(O_1 \cup O_2) = P(O_1) + P(O_2) - P(O_1 \cap O_2) = 0.6 + 0.7 - 0.45 = \underline{\underline{0.85}}$

b) Ikke finne olje i noen av de to $= \overline{O_1 \cup O_2}$:
$P(\overline{O_1 \cup O_2}) = 1 - P(O_1 \cup O_2) = 1 - 0.85 = \underline{\underline{0.15}}$

c) Finne olje i kun reservoar 1 $= O_1 \cap \overline{O_2}$:
$P(O_1 \cap \overline{O_2}) = P(O_1) - P(O_1 \cap O_2) = 0.6 - 0.45 = \underline{\underline{0.15}}$ — bruk Venndiagram!

d) Finne olje i kun reservoar 1 eller kun reservoar 2 $= (O_1 \cap \overline{O_2}) \cup (\overline{O_1} \cap O_2)$:
$P\{(O_1 \cap \overline{O_2}) \cup (\overline{O_1} \cap O_2)\} = P(O_1) + P(O_2) - 2P(O_1 \cap O_2) = 0.6 + 0.7 - 2 \cdot 0.45 = \underline{\underline{0.4}}$ — bruk Venndiagram!

Oppgave 7

a)

$$\begin{aligned} P(B_1 \cap B_2) &= P(B_2|B_1) \cdot P(B_1) = 0.65 \cdot 0.54 = \underline{\underline{0.351}} \\ P(B_1 \cup B_2) &= P(B_1) + P(B_2) - P(B_1 \cap B_2) = 0.54 + 0.54 - 0.351 = \underline{\underline{0.729}} \end{aligned}$$

b)

$$\begin{aligned} B_1 \cap B_2 &= \text{indeksen stiger både dag 1 og dag 2} \\ B_1 \cup B_2 &= \text{indeksen stiger minst en av de to dagene} \\ P(B_1 \cap \overline{B}_2) &= P(\overline{B}_2|B_1) \cdot P(B_1) = (1 - P(B_2|B_1)) \cdot P(B_1) \\ &= (1 - 0.65) \cdot 0.54 = \underline{\underline{0.189}} \end{aligned}$$

c) Merk at snittet mellom to hendelse er en ny hendelse slik at f.eks. $B_1 \cap B_2$ er en hendelse (kan evt innføre $A = B_1 \cap B_2$) og dermed har vi fra multiplikasjonssetningen:

$$P(B_1 \cap B_2 \cap B_3) = P(B_3|B_1 \cap B_2) \cdot P(B_1 \cap B_2) = 0.73 \cdot 0.351 = \underline{\underline{0.256}}$$

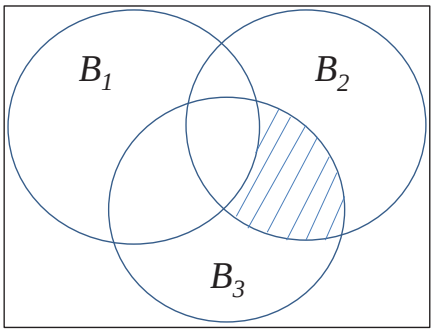
Siste spørsmålet kan løses på flere måter. En måte er å ta utgangspunkt i loven om total sannsynlighet som gir (merk igjen at et snitt som $B_2 \cap B_3$ er en hendelse):

$$\begin{aligned} P(B_2 \cap B_3) &= P(B_1 \cap B_2 \cap B_3) + P(\overline{B}_1 \cap B_2 \cap B_3) \\ \Rightarrow P(\overline{B}_1 \cap B_2 \cap B_3) &= P(B_2 \cap B_3) - P(B_1 \cap B_2 \cap B_3) \\ &= P(B_3|B_2) \cdot P(B_2) - P(B_1 \cap B_2 \cap B_3) \\ &= 0.65 \cdot 0.54 - 0.25623 = \underline{\underline{0.095}} \end{aligned}$$

En alternativ løsning på siste spørsmålet er å starte med multiplikasjonssetningen:

$$\begin{aligned} P(\overline{B}_1 \cap B_2 \cap B_3) &= P(\overline{B}_1|B_2 \cap B_3) \cdot P(B_2 \cap B_3) = (1 - P(B_1|B_2 \cap B_3)) \cdot P(B_2 \cap B_3) \\ &= \left(1 - \frac{P(B_1 \cap B_2 \cap B_3)}{P(B_2 \cap B_3)}\right) \cdot P(B_2 \cap B_3) \\ &= P(B_2 \cap B_3) - P(B_1 \cap B_2 \cap B_3) = P(B_3|B_2) \cdot P(B_2) - P(B_1 \cap B_2 \cap B_3) \\ &= 0.65 \cdot 0.54 - 0.25623 = \underline{\underline{0.095}} \end{aligned}$$

En tredje løsning er å bruke Venn-diagrammet under (der $\overline{B}_1 \cap B_2 \cap B_3$ er skravert) til å forklare at $P(\overline{B}_1 \cap B_2 \cap B_3) = P(B_2 \cap B_3) - P(B_1 \cap B_2 \cap B_3)$ og så fortsette som over.



Oppgave 8

a) Opplysningene gitt i oppgaven kan skrives $\underline{\underline{P(A) = 0.18}}$, $\underline{\underline{P(B) = 0.12}}$ og $\underline{\underline{P(A \cap B) = 0.10}}$. $P(A|B)$ er sannsynligheten for at en som har lest et tidsskrift om naturvitenskap/teknologi også har sett et TV-program om temaet (eller andel av de som har lest tidsskrift som også har sett TV-program).

$$P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{0.10}{0.12} = \underline{\underline{0.83}}$$

Siden $P(A|B) = 0.83$ er forskjellig fra $P(A) = 0.18$ er hendelsene $\underline{\underline{\text{ikke uavhengige}}}$.

b)

$$P(\overline{A}|B) = 1 - P(A|B) = 1 - 0.83 = \underline{\underline{0.17}}$$

På siste spørsmålet kan vi bruke Bayes regel:

$$P(B|\overline{A}) = \frac{P(\overline{A}|B)P(B)}{P(\overline{A})} = \frac{0.17 \cdot 0.12}{1 - 0.18} = \underline{\underline{0.025}}$$

Dersom vi tar med en ekstra desimal i de to første spørsmålene slik at vi der får hhv 0.833 og 0.167 vil svaret her bli $\underline{\underline{0.024}}$ som er enda litt mer presist. Generelt er det ingen absolutte regler for hvor mange desimaler som skal tas med ved angivelse av sannynligheter, men det er rimelig å ta med flere desimaler jo nærmere 0 eller 1 svaret er. For sannsynligheter mellom 0.1-0.9 brukes vanligvis 2 eller 3 desimaler.

Oppgave 9

La C_1 = komponent 1 virker, og C_2 = komponent 2 virker. Da er: $P(C_1) = 1 - 0.1 = 0.9$, og $P(C_2) = 1 - 0.2 = 0.8$. C_1 og C_2 er uavhengige begivenheter.

a) Vi har at: system virker = begge komponentene virker = $C_1 \cap C_2$, og pga uavhengighet er $P(C_1 \cap C_2) = P(C_1)P(C_2) = 0.9 \cdot 0.8 = \underline{\underline{0.72}}$.

b) Vi har at: system virker = komponent 1 eller komponent 2 virker = $C_1 \cup C_2$, og $P(C_1 \cup C_2) = P(C_1) + P(C_2) - P(C_1 \cap C_2) = 0.9 + 0.8 - 0.72 = \underline{\underline{0.98}}$.

Oppgave 10

a) Siden en tilfeldig 10% av husstandene mottar vareprøven vil sannsynligheten for at Per blir trukket ut være $\underline{\underline{0.1}}$. Eventuelt med et gunstige på mulige resonnement (der antall mulige er antall måter å fordele 1 prøve til Per og de 9 øvrige til de 99 andre og antall mulige er antall måter å fordele 10 prøver blant 100) :

$$\frac{\text{antall gunstige}}{\text{antall mulige}} = \frac{\binom{1}{1}\binom{99}{9}}{\binom{100}{10}} = \underline{\underline{0.1}}$$

b) Denne kan også løses på (minst) to måter. La A="Per mottar vareprøve" og B="Kari mottar vareprøve". Dersom vi vet at Kari mottar vareprøven er sannsynligheten for at Per kommer til å motta den 9/99=0.091 (siden det da er 9 prøver igjen å fordele på 99 husstander). Dvs $P(A|B) = 0.091$ og dermed blir

$$P(A \cap B) = P(A|B)P(B) = 0.091 \cdot 0.1 = \underline{\underline{0.0091}}$$

(Merk at svaret blir *ikke* $0.1 \cdot 0.1 = 0.010$ fordi dersom den ene blir trekt ut reduserer det sjansen for at den andre blir trekt ut)
Eventuelt med et gunstige på mulige resonnement: Antall gunstige blir nå alle måter å fordele vareprøvene på slik at Per og Kari får en hver og de 8 gjenværende prøvene fordeles blant de 98 andre i området:

$$\frac{\text{antall gunstige}}{\text{antall mulige}} = \frac{\binom{2}{2}\binom{98}{8}}{\binom{100}{10}} = \underline{\underline{0.0091}}$$

c)

$$P(A \cup B) = P(A) + P(B) - P(A \cap B) = 0.1 + 0.1 - 0.0091 = \underline{\underline{0.191}}$$

Eventuelt kan man begynner med å regne ut sannsynligheten for at ingen av naboene blir trekt ut: $\binom{98}{10}/\binom{100}{10} = 0.809$
Sannsynligheten for at minst en blir trekt ut blir da 1-0.809=$\underline{\underline{0.191}}$.

Oppgave 11

a)

$$\frac{\text{antall gunstige}}{\text{antall mulige}} = \frac{5}{\binom{34}{7}} = \underline{\underline{0.00000093 = 9.3 \cdot 10^{-7}}}$$

b) Ved et system på 8 tall er antall gunstige lik antall måter å trekke 7 vinnertall blant de 8 tallene (uordnet uten tilbakelegging):

$$\frac{\text{antall gunstige}}{\text{antall mulige}} = \frac{\binom{8}{7}}{\binom{34}{7}} = \frac{8}{5379616} = \underline{\underline{0.0000015 = 1.5 \cdot 10^{-6}}}$$

Ved system på 9 tall:

$$\frac{\text{antall gunstige}}{\text{antall mulige}} = \frac{\binom{9}{7}}{\binom{34}{7}} = \frac{36}{5379616} = \underline{\underline{0.0000067 = 6.7 \cdot 10^{-6}}}$$

(Utregningene over viser hvorfor man må betale 8 ganger prisen av en enkeltrekke for et system på 8 tall og 36 ganger prisen av en enkeltrekke for et system på 9 tall.)

c) (Dette er en vanskelig oppgave - ikke være bekymret om du ikke får den til.)
En måte å tenke på er å se på antall måter å velge ut de 8 tallene man tipper på. Man velger ut 8 tall fra 34, der 7 er vinnertall, 1 er tilleggstall og 26 er øvrige tall. Antall mulige måter å gjøre det på er antall måter å trekke 8 blant 34. Antall gunstige måter er antall måter å trekke 6 blant de 7 vinnertallene, 1 blant det ene tilleggstallet og 1 blant de 26 øvrige tallene:

$$\frac{\text{antall gunstige}}{\text{antall mulige}} = \frac{\binom{7}{6}\binom{1}{1}\binom{26}{1}}{\binom{34}{8}} = 0.000010 = 1.0 \cdot 10^{-5}$$

Alternativt kan man tenke på følgende måte: Antall mulig måter å trekke 7 vinnertall og ett tilleggstall er antall måter å trekke 7 vinnertall fra 34 ganger antall måter å trekke ett tilleggstall fra de gjenværende 27 tallene. Antall gunstige er antall måter å trekke 6 vinnertall blant de 8 og 1 vinnertall blant de 26 øvrige multiplisert med antall måter å trekke ett tilleggstall blant de 2 gjenværende:

$$\frac{\text{antall gunstige}}{\text{antall mulige}} = \frac{\binom{8}{6}\binom{26}{1}\binom{2}{1}}{\binom{34}{7}\binom{27}{1}} = 0.000010 = 1.0 \cdot 10^{-5}$$

Oppgave 12

a) B og E er uavhengige hvis $P(B \cap E) = P(B) \cdot P(E)$. Har her at $P(B) \cdot P(E) = 0.5 \cdot 0.6 = 0.3 \neq P(B \cap E) = 0.4$. Dvs B og E er $\underline{\underline{\text{ikke uavhengige}}}$.

b)

$$\begin{aligned} P(\text{minst ett av oppdragene}) &= P(B \cup E) = P(B) + P(E) - P(B \cap E) \\ &= 0.5 + 0.6 - 0.4 = \underline{\underline{0.7}} \\ P(\text{ingen av oppdragene}) &= 1 - P(\text{minst ett av oppdragene}) = 1 - 0.7 = \underline{\underline{0.3}} \end{aligned}$$

c)

$$\begin{aligned} P(B|E) &= \frac{P(B \cap E)}{P(E)} = \frac{0.4}{0.6} = \underline{\underline{\frac{2}{3} = 0.67}} \\ P(B|\overline{E}) &= \frac{P(\overline{E}|B)P(B)}{P(\overline{E})} = \frac{(1 - P(E|B))P(B)}{1 - P(E)} = \frac{(1 - P(E \cap B)/P(B))P(B)}{1 - P(E)} \\ &= \frac{P(B) - P(E \cap B)}{1 - P(E)} = \frac{0.5 - 0.4}{1 - 0.6} = \underline{\underline{0.25}} \end{aligned}$$

Evt har vi fra $P(B) = P(B \cap E) + P(B \cap \overline{E})$ at $P(B \cap \overline{E}) = P(B) - P(B \cap E)$ og kan alternativt løse siste spørsmålet slik:

$$P(B|\overline{E}) = \frac{P(B \cap \overline{E})}{P(\overline{E})} = \frac{P(B) - P(E \cap B)}{1 - P(E)} = \frac{0.5 - 0.4}{1 - 0.6} = \underline{\underline{0.25}}$$

d)

$$\begin{aligned} P(O = 0) &= P(\text{ingen oppdrag}) = \underline{\underline{0.3}} \\ P(O = 0.25) &= P(B \cap \overline{E}) = P(\overline{E}|B)P(B) = (1 - P(E|B))P(B) = (1 - \frac{P(E \cap B)}{P(B)})P(B) \\ &= P(B) - P(E \cap B) = 0.5 - 0.4 = \underline{\underline{0.1}} \end{aligned}$$

Eller fra c):

$$\begin{aligned} P(O = 0.25) &= P(B \cap \overline{E}) = P(B|\overline{E})P(\overline{E}) = 0.25 \cdot (1 - 0.6) = \underline{\underline{0.1}} \\ P(O = 0.3) &= P(\overline{B} \cap E) = P(\overline{B}|E)P(E) = (1 - P(B|E))P(E) = (1 - \frac{P(B \cap E)}{P(E)})P(E) \\ &= P(E) - P(B \cap E) = 0.6 - 0.4 = \underline{\underline{0.2}} \\ P(O = 0.55) &= P(B \cap E) = \underline{\underline{0.4}} \end{aligned}$$

e)

$$\begin{aligned} \mathrm{E}(O) &= \sum_o oP(O = o) = 0 \cdot 0.3 + 0.25 \cdot 0.1 + 0.3 \cdot 0.2 + 0.55 \cdot 0.4 = \underline{\underline{0.305}} \\ \mathrm{E}(O^2) &= \sum_o o^2 P(O = o) = 0^2 \cdot 0.3 + 0.25^2 \cdot 0.1 + 0.3^2 \cdot 0.2 + 0.55^2 \cdot 0.4 = 0.14525 \\ \mathrm{Var}(O) &= \mathrm{E}(O^2) - (\mathrm{E}(O))^2 = 0.14525 - 0.305^2 = 0.052225 \\ \mathrm{SD}(O) &= \sqrt{\mathrm{Var}(O)} = \sqrt{0.05225} = \underline{\underline{0.229}} \end{aligned}$$

Oppgave 13

A_i=panel nr i fungerer, $i = 1, 2, 3$; $P(A_1) = P(A_2) = P(A_3) = 0.98$.
Strøm = alle tre panelene fungerer = $A_1 \cap A_2 \cap A_3$, og fordi A_i'ene er uavhengige, får vi:
$P(A_1 \cap A_2 \cap A_3) = P(A_1)P(A_2)P(A_3) = 0.98^3 = \underline{\underline{0.9412}}$.

Oppgave 14

a) Fra de oppgitt sannsynlighetsfordelingene får vi

$$\begin{aligned} P(X = 3) &= \underline{\underline{0.20}} \\ P(X < 3) &= 0.15 + 0.20 + 0.25 = \underline{\underline{0.60}} \\ P(X \leq 3) &= 0.15 + 0.20 + 0.25 + 0.20 = \underline{\underline{0.80}} \\ P(Y \geq 2) &= 0.35 + 0.10 = \underline{\underline{0.45}} \\ P(Y > 2) &= \underline{\underline{0.10}} \end{aligned}$$

b)

$$\begin{aligned} \mathrm{E}(X) &= \sum_x xP(X = x) = 0 \cdot 0.15 + 1 \cdot 0.20 + 2 \cdot 0.25 + 3 \cdot 0.20 + 4 \cdot 0.20 = \underline{\underline{2.1}} \\ \mathrm{E}(X^2) &= \sum_x x^2P(X = x) = 0^2 \cdot 0.15 + 1^2 \cdot 0.20 + 2^2 \cdot 0.25 + 3^2 \cdot 0.20 + 4^2 \cdot 0.20 = 6.2 \\ \mathrm{Var}(X) &= \mathrm{E}(X^2) - (\mathrm{E}(X))^2 = 6.2 - 2.1^2 = \underline{\underline{1.79}} \\ \mathrm{E}(Y) &= \sum_y yP(Y = y) = 0 \cdot 0.15 + 1 \cdot 0.40 + 2 \cdot 0.35 + 3 \cdot 0.10 = \underline{\underline{1.4}} \\ \mathrm{E}(Y^2) &= \sum_y y^2P(Y = y) = 0^2 \cdot 0.15 + 1^2 \cdot 0.40 + 2^2 \cdot 0.35 + 3^2 \cdot 0.10 = 2.7 \\ \mathrm{Var}(Y) &= \mathrm{E}(Y^2) - (\mathrm{E}(Y))^2 = 2.7 - 1.4^2 = \underline{\underline{0.74}} \end{aligned}$$

$\mathrm{E}(X)$ er gjennomsnittlig antall biler leid ut per dag i det lange løp for den første filialen og $\mathrm{E}(Y)$ er tilsvarende for den andre filialen.
c) Vi har at $F_X = 750X - 1500$ og $F_Y = 750Y - 1100$. Vi har regnet ut forventning og varians til X og Y i b), og ved bruk av regneregler for forvetning og varians får vi:

$$\begin{aligned} \mathrm{E}(F_X) &= \mathrm{E}(750X - 1500) = 750\mathrm{E}(X) - 1500 = 750 \cdot 2.1 - 1500 = \underline{\underline{75}} \\ \mathrm{E}(F_Y) &= \mathrm{E}(750Y - 1100) = 750\mathrm{E}(Y) - 1100 = 750 \cdot 1.4 - 1100 = \underline{\underline{-50}} \\ \mathrm{Var}(F_X) &= \mathrm{Var}(750X - 1500) = 750^2\mathrm{Var}(X) = 750^2 \cdot 1.79 = \underline{\underline{1006875}} \\ \mathrm{Var}(F_Y) &= \mathrm{Var}(750Y - 1100) = 750^2\mathrm{Var}(Y) = 750^2 \cdot 0.74 = \underline{\underline{416250}} \end{aligned}$$

Forventet fortjeneste er gjennomsnittlig fortjeneste per dag i det lange løp. Variansene sier oss noe om variasjonen i fortjenesten fra dag til dag (og dermed om risiko/usikkerhet - jo større varians jo større risiko).
Her ser vi at filial 1 går best. Denne har positiv forventet fortjeneste. Filial 2 ser vi at i det lange løp faktisk vil gå med underskudd. Fra variansene ser vi at det vil være større variasjon i fortjenesten fra dag til dag i filial 1.

Oppgave 15

a) $\mathrm{E}(X-Y) = \mathrm{E}(X) - \mathrm{E}(Y) = 0 - (-1) = \underline{\underline{1}}$,
$\mathrm{E}(X+Y) = \mathrm{E}(X) + \mathrm{E}(Y) = 0 + (-1) = \underline{\underline{-1}}$

b) $\mathrm{Var}(X-Y) = \mathrm{Var}(X) + (-1)^2\mathrm{Var}(Y) = 2^2 + 4^2 = \underline{\underline{20}}$,
$\mathrm{Var}(X+Y) = \mathrm{Var}(X) + \mathrm{Var}(Y) = 2^2 + 4^2 = \underline{\underline{20}}$

c) $\mathrm{E}(0.5X + 0.5Y) = 0.5\mathrm{E}(X) + 0.5\mathrm{E}(Y) = 0.5 \cdot 0 + 0.5 \cdot (-1) = \underline{\underline{-0.5}}$
$\mathrm{Var}(0.5X + 0.5Y) = 0.5^2\mathrm{Var}(X) + 0.5^2\mathrm{Var}(Y) = 0.25(2^2 + 4^2) = \underline{\underline{5}}$

d) $\mathrm{Var}(X-Y) = \mathrm{Var}(X) + (-1)^2\mathrm{Var}(Y) + 2 \cdot (-1) \cdot \mathrm{Cov}(X,Y) = 20 - 2 = \underline{\underline{18}}$
$\mathrm{Var}(X+Y) = \mathrm{Var}(X) + \mathrm{Var}(Y) + 2\mathrm{Cov}(X,Y) = 20 + 2 = \underline{\underline{22}}$

Oppgave 16

a)

$$\begin{aligned}
\mathrm{E}(X) &= \sum_x xP(X=x) = 0 \cdot 0.65 + 1 \cdot 0.25 + 2 \cdot 0.1 = \underline{\underline{0.45}} \\
\mathrm{E}(X^2) &= \sum_x x^2P(X=x) = 0^2 \cdot 0.65 + 1^2 \cdot 0.25 + 2^2 \cdot 0.1 = 0.65 \\
\mathrm{Var}(X) &= \mathrm{E}(X^2) - (\mathrm{E}(X))^2 = 0.65 - 0.45^2 = 0.4475 \\
\mathrm{SD}(X) &= \sqrt{\mathrm{Var}(X)} = \sqrt{0.4475} = \underline{\underline{0.67}}
\end{aligned}$$

b)

$$\begin{aligned}
P(Y=0) &= P(X_1 = 0 \cap X_2 = 0) \overset{uavh.}{=} P(X_1=0) \cdot P(X_2=0) = 0.65^2 = \underline{\underline{0.42}} \\
P(Y=1) &= P(X_1 = 1 \cap X_2 = 0) + P(X_1 = 0 \cap X_2 = 1) \overset{uavh.}{=} 0.25 \cdot 0.65 + 0.65 \cdot 0.25 = \underline{\underline{0.33}} \\
P(Y=2) &= P(X_1 = 2 \cap X_2 = 0) + P(X_1 = 0 \cap X_2 = 2) + P(X_1 = 1 \cap X_2 = 1) \\
&\overset{uavh.}{=} 0.1 \cdot 0.65 + 0.65 \cdot 0.1 + 0.25 \cdot 0.25 = \underline{\underline{0.19}} \\
P(Y=3) &= P(X_1 = 1 \cap X_2 = 2) + P(X_1 = 2 \cap X_2 = 1) \overset{uavh.}{=} 0.25 \cdot 0.1 + 0.1 \cdot 0.25 = \underline{\underline{0.05}} \\
P(Y=4) &= P(X_1 = 2 \cap X_2 = 2) \overset{uavh.}{=} 0.1 \cdot 0.1 = \underline{\underline{0.01}}
\end{aligned}$$

Dvs sannsynlighetsfordelingen blir:

y	0	1	2	3	4
$P(Y=y)$	0.42	0.33	0.19	0.05	0.01

Sjekk: $\sum_y P(Y=y) = 0.42 + 0.33 + 0.19 + 0.05 + 0.01 = 1$, dvs ok!

$$\begin{aligned}
\mathrm{E}(Y) &= \sum_y yP(Y=y) = 0 \cdot 0.42 + 1 \cdot 0.33 + 2 \cdot 0.19 + 3 \cdot 0.05 + 4 \cdot 0.01 = \underline{\underline{0.90}} \\
\text{Eller:}\quad \mathrm{E}(Y) &= \mathrm{E}(X_1 + X_2) = \mathrm{E}(X_1) + \mathrm{E}(X_2) = 0.45 + 0.45 = \underline{\underline{0.90}} \\
P(Y>2) &= P(Y=3) + P(Y=4) = 0.05 + 0.01 = \underline{\underline{0.06}}
\end{aligned}$$

c)

$$\begin{aligned}
P(Z=0) &= P(X_1 = 0 \cap X_2 = 0 \cap X_3 = 0 \cap X_4 = 0 \cap X_5 = 0) \\
&\overset{uavh.}{=} P(X_1=0) \cdot P(X_2=0) \cdot P(X_3=0) \cdot P(X_4=0) \cdot P(X_5=0) \\
&= 0.65 \cdot 0.65 \cdot 0.65 \cdot 0.65 \cdot 0.65 = \underline{\underline{0.12}} \\
P(Z=1) &= P(X_1 = 1 \cap X_2 = 0 \cap X_3 = 0 \cap X_4 = 0 \cap X_5 = 0) \\
&\quad + P(X_1 = 0 \cap X_2 = 1 \cap X_3 = 0 \cap X_4 = 0 \cap X_5 = 0) \\
&\quad + \cdots + P(X_1 = 0 \cap X_2 = 0 \cap X_3 = 0 \cap X_4 = 0 \cap X_5 = 1) \\
&\overset{uavh.}{=} 5 \cdot 0.25 \cdot 0.65 \cdot 0.65 \cdot 0.65 \cdot 0.65 = \underline{\underline{0.22}}
\end{aligned}$$

Oppgave 17

a) Vi har en situasjon karakterisert ved:

- Gjentatte delforsøk som gir "suksess"/ikke "suksess" - flere ventiler som enten består eller ikke består trykktesten.
- Lik sannsynlighet p i alle delforsøk - samme sannsynlighet $p = 0.07$ for ikke å bestå trykktesten for alle ventiler.
- Uavhengige delforsøk - uavhengig fra ventil til ventil om den består trykktesten.
- Et bestemt antall, n, delforsøk - $n = 26$ ventiler som skal testes.

Dvs, alle betingelsene for binmoisk fordeling er oppfylte og vi har dermed at X = antall ventiler som ikke består testen er binomisk fordelt, $X \sim Bin(26, 0.07)$.

b)

$$
\begin{aligned}
P(X = 2) &= \binom{26}{2} 0.07^2 (1 - 0.07)^{26-2} = \underline{\underline{0.28}} \\
P(X < 2) &= P(X = 0) + P(X = 1) = \binom{26}{0} 0.07^0 (1 - 0.07)^{26-0} + \binom{26}{1} 0.07^1 (1 - 0.07)^{26-1} \\
&= 0.1516 + 0.2966 = \underline{\underline{0.45}} \\
P(X \geq 1) &= 1 - P(X = 0) = 1 - 0.152 = \underline{\underline{0.85}} \\
P(1 \leq X \leq 4) &= P(X = 1) + P(X = 2) + P(X = 3) + P(X = 4) \\
&\quad 0.2966 + 0.2790 + \binom{26}{3} 0.07^3 (1 - 0.07)^{26-3} + \binom{26}{4} 0.07^4 (1 - 0.07)^{26-4} \\
&= 0.2966 + 0.2790 + 0.1680 + 0.0727 = \underline{\underline{0.82}} \\
\mathrm{E}(X) &= np = 26 \cdot 0.07 = \underline{\underline{1.8}} \\
\mathrm{E}(Y) &= np = 26 \cdot 0.93 = \underline{\underline{24.2}} \\
P(Y = 26) &= P(X = 0) = \underline{\underline{0.15}}
\end{aligned}
$$

I de to siste utregningen har vi innført Y = antall ventiler som består testen, og av samme grunner som i punkt a) har vi $Y \sim \mathrm{Bin}(26, 0.93)$.

Oppgave 18

a) La X være antallet av 10 tilfeldig valgte studenter som er enige med avisen. Vi sjekker her "suksess"/ikke "suksess" (enig/ikke enig), vi gjør et bestemt antall forsøk og vi har (i alle fall tilnærmet) lik sannsynlighet i hvert forsøk og uavhengighet mellom hvert av forsøkene, dvs X er binomisk fordelt: $X \sim Bin(n, p)$, og avisens påstand svarer til at $p = 0.8$.

Dersom avisens påstand er riktig, får vi at $X \sim B(10, 0.8)$, og da blir:

$$
\begin{aligned}
P(X \leq 4) &= P(X = 0) + P(X = 1) + P(X = 2) + P(X = 3) + P(X = 4) \\
&= \binom{10}{0} 0.8^0 (1 - 0.8)^{10-0} + \binom{10}{1} 0.8^1 (1 - 0.8)^{10-1} + \binom{10}{2} 0.8^2 (1 - 0.8)^{10-2} \\
&\quad + \binom{10}{3} 0.8^3 (1 - 0.8)^{10-3} + \binom{10}{4} 0.8^4 (1 - 0.8)^{10-4} \\
&= 0.0000001 + 0.000004 + 0.00007 + 0.0008 + 0.0055 = \underline{\underline{0.006}}
\end{aligned}
$$

Dvs det observerte resultatet ($X = 4$), eller noe enda mer ekstremt, har svært liten sannsynlighet for å inntreffe dersom avisens påstand er riktig, og derfor er det god grunn til å tvile på om avisens påstand er riktig i virkeligheten !
Merk at vi i dette tilfellet også kunne ha funnet sannsynligheten $P(X \leq 4)$ direkte fra tabellen over binomisk fordeling bak i boka. Men denne tabellen over binomisk fordeling dekker kun noen (få) kombinasjoner av verdier for n og p og vi må derfor også kunne regne ut slike sannsynligheter uten tabell.

Oppgave 19

La T være levetiden. Siden vi har for eksponensialfordeling at $\mathrm{E}(T) = 1/\lambda$ får vi her at $1/\lambda = 10000$ dvs $\lambda = 0.0001$. Sannsynlighetstettheten blir da $f(t) = 0.0001e^{-0.0001t}$ for $t > 0$ og $f(t) = 0$ ellers.
a)

$$\begin{aligned}
P(T < 5000) &= \int_0^{5000} 0.0001e^{-0.0001t} = [-e^{-0.0001t}]_0^{5000} \\
&= -e^{-0.0001\cdot 5000} - (-e^{-0.0001\cdot 0}) = \underline{\underline{0.393}} \\
P(T > 20000) &= \int_{20000}^{\infty} 0.0001e^{-0.0001t} \\
&= [-e^{-0.0001t}]_{20000}^{\infty} = 0 - (-e^{-0.0001\cdot 20000}) = \underline{\underline{0.135}} \\
P(5000 < T < 10000) &= \int_{5000}^{10000} 0.0001e^{-0.0001t} = [-e^{-0.0001t}]_{5000}^{10000} \\
&= -e^{-0.0001\cdot 10000} - (-e^{-0.0001\cdot 5000} = \underline{\underline{0.239}}
\end{aligned}$$

b) Den kumulative fordelingsfunksjonen for eksponensialfordeling er
$P(T < t) = F(t) = 1 - e^{-\lambda t} = 1 - e^{-0.0001t}$ og vi får da:

$$\begin{aligned}
P(T < 5000) &= F(5000) = 1 - e^{-0.0001\cdot 5000} = \underline{\underline{0.393}} \\
P(T > 20000) &= 1 - P(T \leq 20000) = 1 - F(20000) = 1 - (1 - e^{-0.0001\cdot 20000}) = \underline{\underline{0.135}} \\
P(5000 < T < 10000) &= P(T < 10000) - P(T < 5000) = F(10000) - F(5000) \\
&= 1 - e^{-0.0001\cdot 10000} - (1 - e^{-0.0001\cdot 5000}) = \underline{\underline{0.239}}
\end{aligned}$$

c) La D_i være hendelsen at diode i lyser i mer enn 20000 timer. Vi har da regnet ut over at $P(D_i) = P(T > 20000) = 0.135$. Siden diodene fungerer uavhengig av hverandre er $P(D_1 \cap D_2) = P(D_1)P(D_2)$, og vi får:

$$P(\text{lys}) = P(D_1 \cup D_2) = P(D_1) + P(D_2) - P(D_1 \cap D_2) = 0.135 + 0.135 - 0.135 \cdot 0.135 = \underline{\underline{0.252}}$$

Oppgave 20

a) La T være levetiden. Siden vi har for eksponensialfordeling at $\mathrm{E}(T) = 1/\lambda$ får vi har at $1/\lambda = 5$ dvs $\lambda = 0.2$. Sannsynlighetstettheten blir da $f(t) = 0.2e^{-0.2t}$ for $t > 0$ og $f(t) = 0$ ellers.
a)

$$P(T > 2) = \int_2^{\infty} 0.2e^{-0.2t}dt = [-e^{-0.2t}]_2^{\infty} = 0 - (-e^{-0.2\cdot 2}) = \underline{\underline{0.67}}$$

b) La X være antall av de 10 enhetene som fungerer etter 2 år. Vi har da en situasjon der vi for hver enhet (hvert forsøk) sjekker om den fungerer eller ikke fungerer etter 2 år ("suksess"/ikke "suksess"). Sannsynligheten for at enheten fungerer er den samme for alle enheter, $p = 0.67$. Enhetene fungerer eller ikke uavhengig av hverandre, og vi følger et bestemt antall, $n = 10$, enheter. Dvs $X \sim \text{Bin}(10, 0.67)$.

$$\begin{aligned} P(X \geq 8) &= P(X = 8) + P(X = 9) + P(X = 10) \\ &= \binom{10}{8}(0.67)^8(1-0.67)^2 + \binom{10}{9}(0.67)^9(1-0.67)^1 + \binom{10}{10}(0.67)^{10}(1-0.67)^0 = \underline{\underline{0.31}} \end{aligned}$$

Oppgave 21

a) $P(X < 107) = P(\frac{X-100}{8} < \frac{107-100}{8}) = P(Z < 0.88) = \underline{\underline{0.81}}$, der $Z \sim N(0,1)$ (dvs Z er standard normalfordelt) og vi finner $P(Z < 0.88)$ ved å slå opp i normalfordelingstabellen.

b) $P(X < 97) = P(\frac{X-100}{8} < \frac{97-100}{8}) = P(Z < -0.38) = \underline{\underline{0.35}}$

c) $P(X > 110) = 1 - P(X < 110) = 1 - P(Z < 1.25) = 1 - 0.8944 = \underline{\underline{0.1056}}$

d) $P(95 < X < 106) = P(X < 106) - P(X < 95) = P(Z < \frac{106-100}{8}) - P(Z < \frac{95-100}{8})$
$= P(Z < 0.75) - P(Z < -0.63) = 0.7734 - 0.2643 = \underline{\underline{0.51}}$

e) $P(60 < X < 108) = P(X < 108) - P(X < 60) = P(Z < 1) - P(Z < -5)$
$= 0.8413 - 0 = \underline{\underline{0.84}}$
(Minste verdi som tabellen dekker er -3.7, for mindre verdier er sannsynligheten for alle praktiske formål lik 0.)

f) Her må vi først transformere til standard normalfordeling og deretter bruke omvendt tabelloppslag:

$$\begin{aligned} P(X > b) &= P(Z > (b-100)/8) = 1 - P(Z \leq (b-100)/8) = 0.8708 \\ &\Leftrightarrow \quad P(Z \leq (b-100)/8) = 1 - 0.8708 = 0.1292 \\ &\Leftrightarrow \quad (b-100)/8 = -1.13, \quad \text{som medfører at } b = \underline{\underline{90.96}} \end{aligned}$$

Oppgave 22

La X være mengde øl på en ølboks.

a) $$P(X < 0.49) = P(\frac{X-0.5}{0.015} < \frac{0.49-0.5}{0.015}) = P(Z < -0.67) = 0.2514 \approx \underline{\underline{0.25}}$$

b)
$$\begin{aligned} P(0.47 \leq X \leq 0.53) &= P(X \leq 0.53) - P(X < 0.47) \\ &= P(\frac{X-0.5}{0.015} \leq \frac{0.53-0.5}{0.015}) - P(\frac{X-0.5}{0.015} < \frac{0.47-0.5}{0.015}) \\ &= P(Z \leq 2) - P(Z < -2) = 0.9772 - 0.0228 = \underline{\underline{0.9544}} \end{aligned}$$

c)
$$\begin{aligned} P(X > k) &= 1 - P(X \leq k) = 0.99 \\ P(X \leq k) &= P(Z \leq \frac{k-0.50}{0.015}) = 0.01 \\ &\Rightarrow \quad \frac{k-0.50}{0.015} = -2.33 \quad \Rightarrow \quad k = -2.33 \cdot 0.015 + 0.50 = \underline{\underline{0.465}} \end{aligned}$$

d) Vi har en situasjon karakterisert ved:

- Flere enkeltforsøk som hvert resulterer i "suksess" eller ikke "suksess" (flere ølbokser som enten inneholder mindre enn 0.49 liter øl eller ikke)
- Sannsynligheten for "suksess" er den samme i alle enkeltforsøk, $p = P(X < 0.49) = 0.25$ (samme sannsynlighet for at en ølboks inneholder mindre enn 0.49 liter øl for hver boks).
- Enkeltforsøkene er uavhengige (uavhengige mengder øl på hver boks).
- Et bestemt antall, $n = 6$ enkeltforsøk.

Dermed har vi at V="antall av 6 bokser som inneholder mindre enn 0.49" er binomisk fordelt med parametre $n = 6$ og $p = 0.25$.

$$\begin{aligned} P(V \geq 2) &= 1 - P(V \leq 1) = 1 - P(V = 0) - P(V = 1) \\ &= 1 - \binom{6}{0}(0.25)^0(0.75)^{(6)} - \binom{6}{1}(0.25)^1(0.75)^{(5)} \\ &= 1 - 0.1780 - 0.3560 = \underline{\underline{0.466}} \end{aligned}$$

e) Dersom vi lar V=antall av 60 bokser som inneholder mindre enn 0.49 liter øl har vi tilsvarende som i forrige punkt at $V \sim \text{Bin}(60, 0.25)$. Vi er da i en situasjon med $np(1-p) = 60 \cdot 0.25 \cdot 0.75 = 11.25 > 5$, dvs vi kan bruke tilnærmingen til normalfordeling.

$$\begin{aligned} P(V \geq 20) &= 1 - P(V \leq 19) \approx 1 - P(Z \leq \frac{19 + 0.5 - \text{E}(V)}{\sqrt{\text{Var}(V)}}) \\ &= 1 - P(Z \leq \frac{19 + 0.5 - np}{\sqrt{np(1-p)}}) \\ &= 1 - P(Z \leq \frac{19 + 0.5 - 60 \cdot 0.25}{\sqrt{60 \cdot 0.25 \cdot 0.75}}) \\ &= 1 - P(Z \leq 1.34) = 1 - 0.9099 = \underline{\underline{0.09}} \end{aligned}$$

(Det er ikke farlig om heltallskorreksjonen, +0.5, droppes. Svaret blir da enten 0.07 eller 0.12 alt etter om man tar $P(V \geq 20) = 1 - P(V < 20)$ eller $P(V \geq 20) = 1 - P(V \leq 19)$ som utgangspunkt.)

f) Husk at $\text{E}(\bar{X}) = \mu$ og $\text{Var}(\bar{X}) = \frac{\sigma^2}{n}$.

$$P(\bar{X} < 0.49) = P(\frac{\bar{X} - \text{E}(\bar{X})}{\sqrt{\text{Var}(\bar{X})}} < \frac{0.49 - \text{E}(\bar{X})}{\sqrt{\text{Var}(\bar{X})}}) = P(Z < \frac{0.49 - 0.50)}{\sqrt{\frac{0.015^2}{6}}}) = P(Z < -1.63) = \underline{\underline{0.0516}}$$

g)

$$P(\bar{X} < 0.49) = P(\frac{\bar{X} - \text{E}(\bar{X})}{\sqrt{\text{Var}(\bar{X})}} < \frac{0.49 - \text{E}(\bar{X})}{\sqrt{\text{Var}(\bar{X})}}) = P(Z < \frac{0.49 - 0.50)}{\sqrt{\frac{0.015^2}{60}}}) = P(Z < -5.16) \approx \underline{\underline{0}}$$

(Minste verdi som tabellen dekker er -3.7, for mindre verdier er sannsynligheten for alle praktiske formål lik 0.)

h) Variansen til gjennomsnittet blir mindre og mindre jo flere observasjoner vi tar gjennomsnitt av, og sannsynligheten for å få en verdi et stykke unna forventningsverdien blir da også mindre og mindre.

Oppgave 23

a)

$$\begin{aligned}
\mathrm{E}(X) &= \sum_x xP(X=x) = 0\cdot 0.93 + 1\cdot 0.04 + 2\cdot 0.02 + 3\cdot 0.01 = \underline{\underline{0.11}} \\
\mathrm{E}(X^2) &= \sum_x x^2P(X=x) = 0^2\cdot 0.93 + 1^2\cdot 0.04 + 2^2\cdot 0.02 + 3^2\cdot 0.01 = 0.21 \\
\mathrm{Var}(X) &= \mathrm{E}(X^2) - (\mathrm{E}(X))^2 = 0.21 - 0.11^2 = \underline{\underline{0.198}}
\end{aligned}$$

b) La $Y = \sum_{i=1}^{100} = X_1 + X_2 + \cdots + X_{100}$ være totalt antall feil på 100 pumper. Vi har da at $\mathrm{E}(Y) = \sum_{i=1}^{100} \mathrm{E}(X_i) = 100 \cdot 0.11 = \underline{\underline{11}}$ og $\mathrm{Var}(Y) \overset{uavh.}{=} \sum_{i=1}^{100} \mathrm{Var}(X_i) = 100 \cdot 0.198 = 19.8$. Siden Y er en sum av mange uavhengige variable gir sentralgrenseteoremet (SGT) at den er tilnærmet normalfordelt, og vi får:

$$P(Y<10) = P(Y \leq 9) \overset{SGT}{\approx} P(Z \leq \frac{9+0.5-\mathrm{E}(Y)}{\sqrt{\mathrm{Var}(Y)}}) = P(Z \leq \frac{9+0.5-11}{\sqrt{19.8}}) = P(Z \leq -0.34) = \underline{\underline{0.37}}$$

(Det er ikke farlig om heltallskorreksjonen, +0.5, droppes. Svaret blir da 0.33 eller 0.41.)

Oppgave 24

a) La X antall bakterier i en volumenhet. Fra opplysningene i oppgaven vil da X være Poisson-fordelt med $\lambda = 2.5$ og $t = 1$.

$$\begin{aligned}
P(X=2) &= \frac{2.5^2}{2!}e^{-2.5} = \underline{\underline{0.257}} \\
P(X<3) &= P(X=0) + P(X=1) + P(X=2) = \frac{2.5^0}{0!}e^{-2.5} + \frac{2.5^1}{1!}e^{-2.5} + 0.257 = \underline{\underline{0.544}} \\
P(X\geq 4) &= 1 - P(X<4) = 1 - (P(X<3) + P(X=3)) = 1 - (0.544 + \frac{2.5^3}{3!}e^{-2.5}) = \underline{\underline{0.242}}
\end{aligned}$$

Evt kan man her spare litt tid på å bruke tabell.

b) La Y antall bakterier i tre volumenheter. Siden vi nå ser på tre volumenheter er $t = 3$, mens $\lambda = 2.5$ som før.

$$\begin{aligned}
P(Y=2) &= \frac{7.5^2}{2!}e^{-7.5} = \underline{\underline{0.0156}} \\
P(Y<3) &= P(X=0) + P(X=1) + P(X=2) = \frac{7.5^0}{0!}e^{-7.5} + \frac{7.5^1}{1!}e^{-7.5} + 0.0156 = \underline{\underline{0.020}} \\
P(X\geq 4) &= 1 - P(X<4) = 1 - (P(X<3) + P(X=3)) = 1 - (0.020 + \frac{7.5^3}{3!}e^{-7.5}) = \underline{\underline{0.941}}
\end{aligned}$$

c) La X antall bakterier i 10 volumenheter. Vi har da $t = 10$ og med $\lambda = 2.5$ har vi at X er Poissonfordelt med $\lambda t = 2.5 \cdot 10 = 25$. Siden $\lambda t > 10$ kan vi bruke tilnærming til normalfordeling. Husk at i Poissonfordeling er $\mathrm{E}(X) = \mathrm{Var}(X) = \lambda t$.

$$P(X > 20) = 1 - P(X \leq 20) = 1 - P(Z \leq \frac{20 + 0.5 - 25}{\sqrt{25}} = 1 - P(Z \leq -0.9) = 1 - 0.1841 = \underline{\underline{0.82}}$$

(Det er ikke farlig om heltallskorreksjonen, +0.5, droppes. Svaret blir da enten 0.84 eller 0.79 alt etter om man tar $P(X > 20) = 1 - P(X \leq 20)$ eller $P(X > 20) = 1 - P(X < 21)$ som utgangspunkt.)

Oppgave 25

X = antall skudd som gir mål (av de 120 forsøkene). Har at $X \sim \mathrm{Bin}(120, 0.8)$. Vi ser at $\mathrm{Var}(X) = 120 \cdot 0.8 \cdot (1 - 0.8) = 19.2 > 10$ dvs vi kan bruke tilnærming til normalfordeling. Har også $\mathrm{E}(X) = 120 \cdot 0.8 = 96$.

a) $$P(X < 90) = P(X \leq 89) \approx P(Z \leq \frac{89 + 0.5 - 96}{\sqrt{19.2}}) = P(Z < -1.48) = \underline{\underline{0.069}}$$

(Det er ikke farlig om heltallskorreksjonen, +0.5, droppes. Svaret blir da enten 0.055 eller 0.087.)

b) $P(X \geq 105) = 1 - P(X \leq 104) \approx 1 - P(Z \leq 1.94) = 1 - 0.9738 = \underline{\underline{0.026}}$. (Uten heltallskorreksjon blir svaret enten 0.020 eller 0.034.)

c) $P(X = 100) = \begin{pmatrix} 120 \\ 100 \end{pmatrix} 0.8^{100} 0.2^{20} = \underline{\underline{0.063}}$. (Her er det enkelt å beregne den eksakte sannsynligheten vha. uttrykket for binomiske sannsynligheter siden det kun er ett ledd som skal regnes ut. Det er også mulig å bruke normaltilnærming.)

Oppgave 26

La X_i være stokastisk variabel for utfallet av måling nr i, $i = 1, \ldots, 80$. Vi har for alle i at $\mathrm{E}(X_i) = \mu$ = virkelig smeltepunkt til ny legering, og $\mathrm{SD}(X_i) = 7$.
Sentralgrenseteoremet gir at $\overline{X} = (\sum_{i=1}^{80} X_i)/80$ er tilnærmet normalfordelt med forventning μ og standardavvik $7/\sqrt{80} = 0.78$.

$$\begin{aligned} & P(\text{gj.sn. avviker mindre enn 1.54 grader fra virkelig smeltepunkt}) \\ = \; & P(\mid \overline{X} - \mu \mid < 1.54) \\ = \; & P(-1.54 < \overline{X} - \mu < 1.54) \\ = \; & P(\overline{X} - \mu < 1.54) - P(\overline{X} - \mu < -1.54) \\ \approx \; & P(Z < 1.97) - P(Z < -1.97) = 0.9756 - 0.0244 = \underline{\underline{0.95}} \end{aligned}$$

Oppgave 27

Husk fra pensum at lineærkombinasjoner av normalfordelte variable er normalfordelt og husk regnereglene for forventning og varians.

a) La $D = X - Y$.

$$\begin{aligned} P(X < Y) &= P(X - Y < 0) = P(D < 0) \\ \mathrm{E}(D) &= \mathrm{E}(X) - \mathrm{E}(Y) = 180 - 167 = 13 \\ \mathrm{Var}(D) &= \mathrm{Var}(X) + (-1)^2\mathrm{Var}(Y) = 36 + 36 = 72 \\ P(D < 0) &= P(Z < \frac{0 - 13}{\sqrt{72}}) = P(Z < -1.53) = \underline{\underline{0.063}} \end{aligned}$$

Betyr at sannsynligheten for at en tilfeldig valgt (med hensyn på høyde) mann er lavere enn en tilfeldig valgt kvinne er 6.3%. (F.eks. dersom man kan anta at par finner hverandre uavhengig av høyde vil kvinnen være høyere enn mannen i 6.3% av parene)

b) La $S = X + Y$.

$$\begin{aligned} \mathrm{E}(S) &= \mathrm{E}(X) + \mathrm{E}(Y) = 180 + 167 = 347 \\ \mathrm{Var}(S) &= \mathrm{Var}(X) + \mathrm{Var}(Y) = 36 + 36 = 72 \\ P(S > 350) &= 1 - P(S \leq 350) = 1 - P(Z < \frac{350 - 347}{\sqrt{72}}) \\ &= 1 - P(Z < 0.35) = 1 - 0.6368 = \underline{\underline{0.36}} \end{aligned}$$

Betyr at sannsynligheten for at summen av høyden til en tilfeldig valgt mann og en tilfeldig valgt kvinne er mer enn 350 cm er 36%. (Dersom par finner hverandre uavhengig av høyde har 36% av parene en samlet høyde på mer enn 350cm.)

c) La $G = (X + Y)/2 = \frac{1}{2}(X + Y)$.

$$\begin{aligned} \mathrm{E}(G) &= \frac{1}{2}(\mathrm{E}(X) + \mathrm{E}(Y)) = \frac{1}{2}(180 + 167) = 173.5 \\ \mathrm{Var}(G) &= \frac{1}{2^2}(\mathrm{Var}(X) + \mathrm{Var}(Y)) = \frac{1}{4}(36 + 36) = 18 \\ P(G < 170) &= P(Z < \frac{170 - 173.5}{\sqrt{18}}) = P(Z < -0.82) = \underline{\underline{0.21}} \end{aligned}$$

Betyr at sannsynligheten for at gjennomsnittet av høyden til en tilfeldig valgt mann og en tilfeldig valgt kvinne er mindre enn 170 cm er 21%. (Dersom par finner hverandre uavhengig av høyde har 21% av parene en gjennomsnittshøyde på mindre enn 170cm.)

Oppgave 28

a) μ er forventningsverdien, dvs gjennomsnitt i det lange løp/over hele populasjonen. $\bar{x}$ er gjennomsnittet av et utvalg. σ^2 er variansen som er et mål på variasjon i populasjonen/variasjon i det lange løp. s^2 er utvalgsvariansen som er et mål på variasjonen i et utvalg.
μ er en parameter som sier noe om hele populasjonen (gjennomsnitt i det lange løp), mens $\bar{x}$ bare regnes ut fra et utvalg.
$\bar{x}$ er et estimat (anslag) på verdien til μ.
En estimator er forventningsrett dersom forventningsverdien til estimatoren er lik den ukjente parameteren estimatoren skal brukes til å anslå. Betyr at dersom forsøket gjentas mange ganger vil estimatoren i gjennomsnitt i det lange løp gi rett verdi. (En estimator som ikke er forventningsrett vil ved gjentatte forsøk i gjennomsnitt i det lange løp gi gal verdi - dvs vi gjør en systematisk feil ved å bruke en estimator som ikke er forventningsrett.)

Oppgave 29

a) Forventningen estimeres med gjenomsnittet, og derfor har vi:
April, estimat av forventet nedbørsmengde: $\hat{\mu} = \bar{x} = \underline{\underline{52.7}}$ mm
August, estimat av forventet nedbørsmengde: $\hat{\mu} = \bar{x} = \underline{\underline{119.1}}$ mm

Standardavvik estimeres med utvalgssstandardavvik, og derfor har vi:
April, estimat av standardavvik (=utv.stand.avvik): $\hat{\sigma} = s = \underline{\underline{26.4}}$
August, estimat av standardavvik (=utv.stand.avvik): $\hat{\sigma} = s = \underline{\underline{57.4}}$

b) Ja, det det ser rimelig ut å bruke normalantakelse for disse dataene - histogrammene har tendens til normalfordelingsform. Konfidensintervallet for forventningsverdien μ krever normalfordeling dersom man har få målinger, men her har vi så mange målinger ($n = 41$) at vi pga sentralgrenseteoremet kan regne ut et tilnærmet konfidensintervall for μ også uten å anta normalfordeling. Konfidensintervallet for standardavvik krever normalfordeling.

c) Siden det virker rimelig å anta at målingene er normalfordelte, men vi ikke kjenner verdiene til μ eller σ, er vi her i en situasjon med normalfordeling med ukjent μ og ukjent σ. Et $(1-\alpha)100\%$ konfidensintervall for μ er da gitt ved

$$[\bar{X} - t_{\alpha/2,n-1}\frac{S}{\sqrt{n}}, \bar{X} + t_{\alpha/2,n-1}\frac{S}{\sqrt{n}}]$$

Siden vi skal ha et 95% konfidensintervall må vi sette $\alpha = 0.05$ og vi finner i tabell E.5 i læreboka at $t_{\alpha/2,n-1} = t_{0.025,40} = 2.021$. Videre setter vi for april inn $\bar{x} = 52.7$, $n = 41$ og $s = 26.4$, og vi får da 95% konfidensintervall for μ for april:

$$[52.7 - 2.021\frac{26.4}{\sqrt{41}},\ 52.7 + 2.021\frac{26.4}{\sqrt{41}}] = \underline{\underline{[44.4,\ 61.0]}}$$

For august har vi $\bar{x} = 119.1$, $n = 41$ og $s = 57.4$, og vi får da 95% konfidensintervall for μ:

$$[119.1 - 2.021\frac{57.4}{\sqrt{41}},\ 119.1 + 2.021\frac{57.4}{\sqrt{41}}] = \underline{\underline{[101.0,\ 137.2]}}$$

Vi ser at konfidensintervallet for forventet aprilnedbør dekker klart lavere verdier enn intervallet for augustnedbør. Dvs, selv når vi tar høyde for usikkerhet i estimatene virker det helt klart at forventet aprilnedbør er en god del lavere enn forventet augustnedbør.

d) Ved å fremdeles anta at målingene er normalfordelte kan vi finne et konfidensintervall for σ ved å ta utgangspunkt i resultatet at $(n-1)\frac{S^2}{\sigma^2} \sim \chi(n-1)$ og regne videre som forklart på forelesning får vi at et $(1-\alpha)100\%$ konfidensintervall for σ er gitt ved

$$[\sqrt{\frac{(n-1)S^2}{\chi_{\alpha/2,n-1}}},\ \sqrt{\frac{(n-1)S^2}{\chi_{1-\alpha/2,n-1}}}]$$

Med $\alpha = 0.05$ og $n = 41$ finner vi i tabell E.6 i boka at $\chi_{\alpha/2,n-1} = \chi_{0.025,40} = 59.34$ og $\chi_{1-\alpha/2,n-1} = \chi_{0.975,40} = 24.43$. For april med $s = 26.4$ får vi da

$$[\sqrt{\frac{(41-1)26.4^2}{59.34}},\ \sqrt{\frac{(41-1)26.4^2}{24.43}}] = \underline{\underline{[21.7,\ 33.8]}}$$

For august med $s = 57.4$ får vi

$$[\sqrt{\frac{(41-1)57.4^2}{59.34}},\ \sqrt{\frac{(41-1)57.4^2}{24.43}}] = \underline{\underline{[47.1,\ 73.4]}}$$

Dvs, det virker helt klart at også variasjonen i nedbør i august er høyere enn i april (siden alle verdier i konfidensintervallet for σ for august er høyere enn verdiene i intervallet for april).

Oppgave 30

a) X er egentlig strengt tatt hypergeometrisk fordelt med N = totalt antall personer i hele befolkningen, M = antall personer i befolkningen som er for saken og $n = 1200$ som er utvalgsstørrelse. Siden n er betydelig mindre enn antall personer i befolkningen, N, (dvs $N > 10n$), er X tilnærmet binomisk fordelt: $X \sim \text{Bin}(n, p)$, der $p = M/N$ er andelen i befolkningen som er for.

b) Undersøkelsen indikerer at det ikke er flertall dersom færre enn halvparten av de spurte er for, dvs dersom $X < 600$. Siden vi er i en situasjon med $np(1-p) = 1200 \cdot 0.52 \cdot (1-0.52) = 299.5 > 5$ kan vi bruke tilnærmingen til normalfordeling.

$$\begin{aligned} P(X < 600) &= P(X \leq 599) \approx P(Z \leq \frac{599 + 0.5 - \text{E}(X)}{\sqrt{\text{Var}(X)}}) = P(Z \leq \frac{599 + 0.5 - np}{\sqrt{np(1-p)}}) \\ &= P(Z \leq \frac{599 + 0.5 - 1200 \cdot 0.52}{\sqrt{1200 \cdot 0.52 \cdot 0.48}}) = P(Z \leq -1.42) = \underline{\underline{0.08}} \end{aligned}$$

(Det er ikke farlig om heltallskorreksjonen, +0.5, droppes. Svaret blir da enten 0.07 eller 0.08 alt etter om man tar $P(X \leq 599)$ eller $P(X < 600)$ som utgangspunkt.)

c) Estimator for suksess-sannynligheten, p, i binomisk modell: $\widehat{p} = X/n$.
Estimat av p: $\hat{p} = 740/1200 = \underline{\underline{0.617}}$.
Et (tilnærmet) $(1-\alpha)100\%$ konfidensintervall for p er gitt ved:

$$[\hat{p} - z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\ ,\ \ \hat{p} + z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}]$$

Innsatt $\hat{p} = 0.617$, $n = 1200$ og $z_{\alpha/2} = z_{0.05} = 1.645$ gir dette tilnærmet 90% konfidensintervall for p:

$$[0.617 - 1.645\sqrt{\frac{0.617(1-0.617)}{1200}},\ \ 0.617 + 1.645\sqrt{\frac{0.617(1-0.617)}{1200}}] = \underline{\underline{[0.594\ ,\ \ 0.640]}}$$

d) Det er 10 % sannsynlighet for at et 90 % konfidensintervall ikke vil komme til å dekke den virkelige verdien. Derfor vil i det lange løp 10 % av intervallene ikke dekke den virkelige verdien.

Oppgave 31

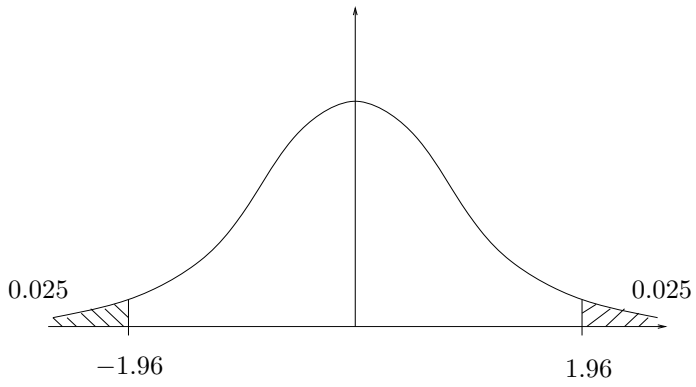
a)

$$\hat{\mu} = \overline{X}$$

$$Z = \frac{\overline{X} - \mathrm{E}(\overline{X})}{\sqrt{\mathrm{Var}(\overline{X})}} = \frac{\overline{X} - \mu}{\sqrt{\sigma^2/n}} = \frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \sim \mathrm{N}(0,1)$$

$$\begin{aligned} P(-1.96 \leq \frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \leq 1.96) &= 0.95 \\ P(-1.96\frac{\sigma}{\sqrt{n}} \leq \overline{X} - \mu \leq 1.96\frac{\sigma}{\sqrt{n}}) &= 0.95 \\ P(\overline{X} - 1.96\frac{\sigma}{\sqrt{n}} \leq \mu \leq \overline{X} + 1.96\frac{\sigma}{\sqrt{n}}) &= 0.95 \end{aligned}$$

Dvs et 95% konfidensintervall for μ er gitt ved:



$$\underline{\underline{[\overline{X} - 1.96\frac{\sigma}{\sqrt{n}}, \overline{X} + 1.96\frac{\sigma}{\sqrt{n}}]}}$$

Innsatt $\overline{x} = 0.497$, $n = 24$ og $\sigma = 0.015$ gir dette 95% konfidensintervall for μ:

$$[0.497 - 1.96\frac{0.015}{\sqrt{24}}, 0.497 + 1.96\frac{0.015}{\sqrt{24}}] = \underline{\underline{[0.491, 0.503]}}$$

b) Husk først at med $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ der $X_1, \ldots X_n$ er uavhengige med $\mathrm{E}(X) = \mu$ og $\mathrm{Var}(X) = \sigma^2$ så vil $\mathrm{E}(\bar{X}) = \frac{1}{n}\sum_{i=1}^{n}\mathrm{E}(X_i) = \frac{1}{n}\sum_{i=1}^{n}\mu = \mu$ og $\mathrm{Var}(\bar{X}) = \frac{1}{n^2}\sum_{i=1}^{n}\mathrm{Var}(X_i) = \frac{1}{n^2}\sum_{i=1}^{n}\sigma^2 = \sigma^2/n$. Det samme gjelder for $\bar{Y}$. Vi får da:

$$\begin{aligned} \mathrm{E}(\hat{\mu}_1) &= \frac{1}{2}\mathrm{E}(\bar{X}) + \frac{1}{2}\mathrm{E}(\bar{Y}) = \frac{1}{2}\mu + \frac{1}{2}\mu = \underline{\underline{\mu}} \\ \mathrm{E}(\hat{\mu}_2) &= \frac{2}{3}\mathrm{E}(\bar{X}) + \frac{1}{3}\mathrm{E}(\bar{Y}) = \frac{2}{3}\mu + \frac{1}{3}\mu = \underline{\underline{\mu}} \\ \mathrm{Var}(\hat{\mu}_1) &= \mathrm{Var}(\frac{1}{2}\bar{X} + \frac{1}{2}\bar{Y}) \overset{uavh}{=} \frac{1}{2^2}\mathrm{Var}(\bar{X}) + \frac{1}{2^2}\mathrm{Var}(\bar{Y}) \end{aligned}$$

$$
\begin{aligned}
&= \frac{1}{2^2}\cdot\frac{0.015^2}{24} + \frac{1}{2^2}\cdot\frac{0.015^2}{12} = \underline{\underline{0.000007}} \\
\mathrm{Var}(\hat{\mu}_2) &= \mathrm{Var}(\frac{2}{3}\bar{X} + \frac{1}{3}\bar{Y}) \overset{uavh}{=} \frac{2^2}{3^2}\mathrm{Var}(\bar{X}) + \frac{1}{3^2}\mathrm{Var}(\bar{Y}) \\
&= \frac{2^2}{3^2}\cdot\frac{0.015^2}{24} + \frac{1}{3^2}\cdot\frac{0.015^2}{12} = \underline{\underline{0.000006}}
\end{aligned}
$$

Estimatoren $\underline{\underline{\hat{\mu}_2 \text{ er best}}}$ siden den har minst varians og begge estimatorene er forventningsrette.

c) Estimat for μ: $\hat{\mu}_2 = \frac{2}{3}\cdot 0.497 + \frac{1}{3}\cdot 0.494 = \underline{\underline{0.496}}$.
Siden $\hat{\mu}_2 = \frac{2}{3}\bar{X} + \frac{1}{3}\bar{Y}$ er en lineærkombinasjon av normalfordelte variable er $\hat{\mu}_2$ normalfordelt. Siden $\hat{\mu}_2$ også er forventningsrett har vi (se forelesningsnotatene/boka regel 6.8) at et $(1-\alpha)100\%$ konfidensintervall for μ er gitt ved

$$[\hat{\mu}_2 - z_{\alpha/2}\mathrm{SD}(\hat{\mu}_2), \hat{\mu}_2 + z_{\alpha/2}\mathrm{SD}(\hat{\mu}_2)]$$

Med $\alpha = 0.05$ har vi $z_{\alpha/2} = 1.96$. Videre sette vi inn $\hat{\mu}_2 = 0.496$ og $\mathrm{SD}(\hat{\mu}_2) = \sqrt{\mathrm{Var}(\hat{\mu}_2)} = \sqrt{0.000006} = 0.0024$, og vi får da 95% konfidensintervallet for μ:

$$[0.496 - 1.96\cdot 0.0024, 0.496 + 1.96\cdot 0.0024] = \underline{\underline{[0.491, 0.501]}}$$

Oppgave 32

a) Gjennomsnitt = $\underline{\underline{771}}$, utvalgsstandardavvik = $\underline{\underline{10.17}}$ og median = $\underline{\underline{772}}$.

b)

$$
\begin{aligned}
P(X_1 + X_2 < 1540) &= P(Z < \frac{1540 - \mathrm{E}(X_1 + X_2)}{\sqrt{\mathrm{Var}(X_1 + X_2)}}) = P(Z < \frac{1540 - (780 + 780)}{\sqrt{10^2 + 10^2}}) \\
&= P(Z < -1.41) = \underline{\underline{0.079}}
\end{aligned}
$$

$$
\begin{aligned}
P(\frac{X_1 + X_2}{2} < 775) &= P(Z < \frac{775 - \mathrm{E}(\frac{X_1+X_2}{2})}{\sqrt{\mathrm{Var}(\frac{X_1+X_2}{2})}}) = P(Z < \frac{775 - (780 + 780)/2}{\sqrt{(10^2 + 10^2)/4}}) \\
&= P(Z < -0.71) = \underline{\underline{0.239}}
\end{aligned}
$$

Evt:

$$
P(\frac{X_1 + X_2}{2} < 775) = P(X_1 + X_2 < 1550) = \cdots = \underline{\underline{0.239}}
$$

c) Vi har en situasjon med normalfordelte data med ukjent μ og ukjent σ. Vi må da bruke et t-intervall, dvs et 95% konfidensintervall for μ er gitt ved:

$$\underline{\underline{[\overline{X} - t_{0.025,n-1}\frac{S}{\sqrt{n}}, \overline{X} + t_{0.025,n-1}\frac{S}{\sqrt{n}}]}}$$

Innsatt $\overline{x} = 771$, $s = 10.17$, $n = 4$, $t_{0.025,3} = 3.182$ gir dette 95% konfidensintervall for μ:

$$[771 - 3.182\frac{10.17}{\sqrt{4}},\ 771 + 3.182\frac{10.17}{\sqrt{4}}] = \underline{\underline{[754.8,\ 787.2]}}$$

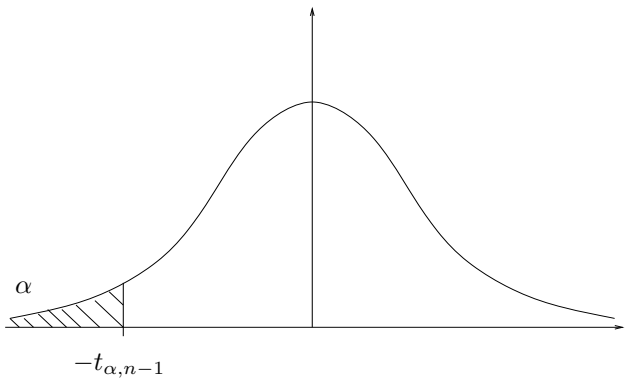
d)

$$H_0: \ \mu \geq 780 \qquad \text{mot} \qquad H_1: \ \mu < 780$$

Siden σ er ukjent baserer vi testen på

$$T = \frac{\bar{X} - \mu_0}{S/\sqrt{n}} = \frac{\bar{X} - 0}{S/\sqrt{n}} \sim t(n-1)$$

Med signifikansnivå 5%, dvs $\alpha = 0.05$, og $n = 4$ målinger forkaster vi H_0 dersom $T \leq -t_{0.05,3} = -2.353$.



Observert: $t = \frac{771-780}{10.17/\sqrt{4}} = -1.77$
Siden $-1.77 > -2.353$ blir konklusjonen at vi ikke forkaster H_0. Dataene gir ikke grunnlag for å konkludere at forventet brødvekt er mindre enn 780 gram.

Oppgave 33

a) Vi må anta at målingene er uavhengige og normalfordelte. Vi må også anta at alle målingene er fra samme normalfordeling (dvs at målingene er identisk fordelte - når vi gjør gjentatte målinger av samme fenomen som her vil dette normalt være oppfylt). Kort sagt: Vi må anta at målingene er uavhengige og identisk normalfordelte.

b) $\mathrm{E}(X_i) = \mu$ er forventet blykonsentrasjon og estimeres med $\widehat{\mu} = \overline{X} = \frac{1}{25}(X_1 + \cdots + X_{25})$.
$\mathrm{SD}(X_i) = \sigma$ estimeres med $S = \sqrt{\frac{1}{24}\sum_{i=1}^{25}(X_i - \overline{X})^2}$.
Fra teksten har vi: Estimat av μ: 0.38. Estimat av σ: 0.06
Et 98% konfidensintervall for μ når σ er ukjent er gitt ved:

$$\left[\overline{X} - t_{0.01,n-1}\frac{S}{\sqrt{n}}, \ \ \overline{X} + t_{0.01,n-1}\frac{S}{\sqrt{n}}\right]$$

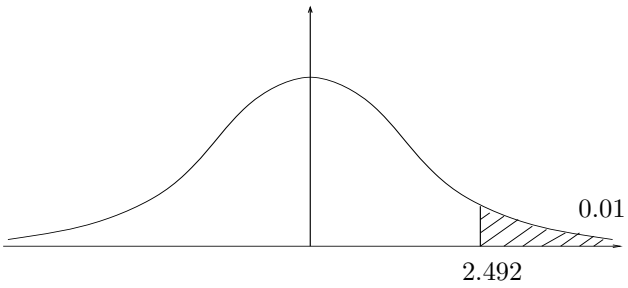
Med data over og $t_{0.01,24} = 2.492$:

$$\left[0.38 - 2.492\frac{0.06}{\sqrt{25}}, \ \ 0.38 + 2.492\frac{0.06}{\sqrt{25}}\right] = [0.35, \ 0.41]$$

c) Vi vil teste

$$H_0 : \mu \leq 0.35 \qquad \text{mot} \qquad H_1 : \mu > 0.35$$

Dersom H_0 er korrekt har vi at $T = \frac{\overline{X}-\mu_0}{\sqrt{S^2/n}} = \frac{\bar{X}-0.35}{S/\sqrt{n}} \sim t_{n-1} = t_{24}$ Med signifikansnivå $\alpha = 0.01$ forkaster vi H_0 dersom $T \geq t_{0.01,24} = 2.492$



Utfall av testobservator:

$$t = \frac{0.38 - 0.35}{0.06/\sqrt{25}} = 2.5 > 2.492.$$

Dvs. utfallet er i forkastningsområdet. Konklusjon: Forkast H_0 - det er grunn til å konkludere at forventet blykonsentrasjon er høyere enn 0.35.

Oppgave 34

a) La $V = X_1 + X_2 + \cdots + X_{300} = \sum_{i=1}^{300} X_i$ være totalt antall solgte biler i løpet av ett år. Vi har da at $\mathrm{E}(V) = \sum_{i=1}^{300} \mathrm{E}(X_i) = 300 \cdot 0.45 = 135$ og $\mathrm{Var}(V) \overset{uavh.}{=} \sum_{i=1}^{300} \mathrm{Var}(X_i) = 300 \cdot 0.4475 = 134.25$. Siden V er en sum av mange uavhengige variable gir sentralgrenseteoremet (SGT) at den er tilnærmet normalfordelt, og vi får:

$$\begin{aligned} P(V \geq 125) &= 1 - P(V \leq 124) \overset{SGT}{\approx} 1 - P(Z \leq \frac{124 + 0.5 - \mathrm{E}(V)}{\sqrt{\mathrm{Var}(V)}}) \\ &= 1 - P(Z \leq \frac{124 + 0.5 - 135}{\sqrt{134.25}}) = 1 - P(Z \leq -0.91) = 1 - 0.1814 = \underline{0.82} \end{aligned}$$

(Det er ikke farlig om heltallskorreksjonen, faktoren +0.5, droppes. Svaret blir da enten 0.81 eller 0.83 alt etter om man tar $P(V \geq 125) = 1 - P(V < 125)$ eller $P(V \geq 125) = 1 - P(V \leq 124)$ som utgangspunkt.)

Oppgave 35

Målingene betraktes som utfall av $n = 100$ uavhengige identisk fordelte stokastiske variable $X_1, \cdots, X_{100}$. $\mathrm{E}(X_i) = \mu$ er forventet størkningstid og estimeres med $\widehat{\mu} = \overline{X} = \frac{1}{100}(X_1 + \cdots + X_{100})$.
$\mathrm{Var}(X_i) = \sigma^2$ estimeres med $S^2 = \frac{1}{99}\sum_{i=1}^{100}(X_i - \overline{X})^2$,
Fra teksten har vi: Estimat av μ: 32. Estimat av σ: 4

a) På grunn av sentralgrenseteoremet vil vi ha at $\overline{X}$ vil være tilnærmet normalfordelt (siden $n = 100 > 30$) selv om $X_1, \cdots, X_{100}$ ikke er normalfordelte. Siden konfidensintervaller og hypotesetester for μ baserer seg på $\overline{X}$ vil intervallene/testene for normalfordelte data fremdeles gjelder som en tilnærming.

b) Et tilnærmet 95% konfidensintervall for μ er gitt ved:

$$\left[\overline{X} - z_{0.025}\frac{S}{\sqrt{n}}, \overline{X} + z_{0.025}\frac{S}{\sqrt{n}}\right]$$

Med data:

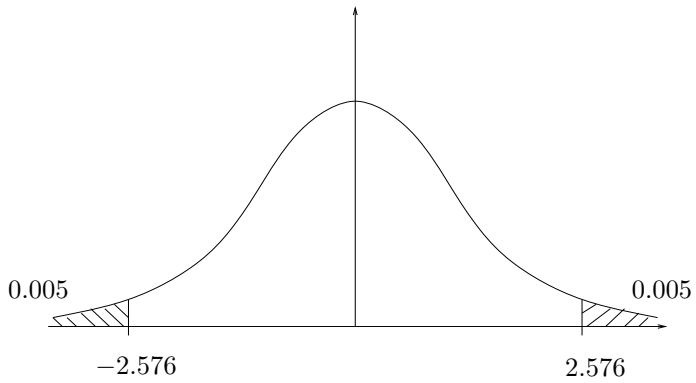
$$\left[32 - 1.96\frac{4}{\sqrt{100}}, 32 + 1.96\frac{4}{\sqrt{100}}\right] = \underline{\underline{[31.2, 32.8]}}$$

c) Vi skal teste

$$H_0 : \mu = 30 \qquad \text{mot} \qquad H_1 : \mu \neq 30$$

Dersom H_0 er korrekt har vi (pga sentralgrenseteoremet) at

$$Z = \frac{\overline{X} - \mathrm{E}(\overline{X})}{\sqrt{\widehat{\mathrm{Var}}(\overline{X})}} = \frac{\overline{X} - \mu_0}{\sqrt{S^2/n}} = \frac{\bar{X} - 0.35}{\sqrt{S^2/n}} \approx \mathrm{N}(0,1)$$



Med (tilnærmet) signifikansnivå $\alpha = 0.01$ forkaster vi H_0 dersom $Z \leq -z_{0.005} = -2.576$ eller dersom $Z \geq z_{0.005} = 2.576$
Utfall av testobservator/teststørrelse:

$$z = \frac{32 - 30}{\sqrt{4^2/100}} = 5.0 > 2.576.$$

Dvs.: Utfallet av teststørrelsen er i forkastningsområdet. Konklusjon: $\underline{\underline{\text{Forkast } H_0}}$ - det er grunn til å hevde at forventet størkningstid er forskjellig fra 30 minutt (den ser ut til å være høyere).

Oppgave 36

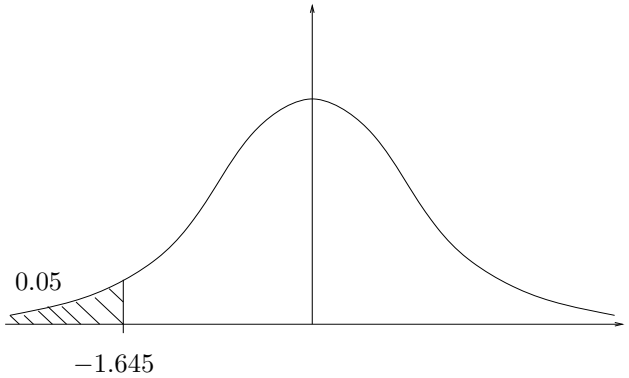
a) La p andelen pumper med feil. Skal da teste

$$H_0: \quad p \geq 0.07 \qquad \text{mot} \qquad H_1: \quad p < 0.07$$

Estimator: $\hat{p} = \frac{X}{n}$. Dersom H_0 er korrekt er

$$Z = \frac{\hat{p} - \mathrm{E}(\hat{p})}{\sqrt{\mathrm{Var}(\hat{p})}} = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} \approx N(0,1)$$

Med signifikansnivå 5%, dvs $\alpha = 0.05$, forkaster vi H_0 dersom $Z \leq -z_{0.05} = -1.645$.



Observert:

$$z_{obs} = \frac{3/100 - 0.07}{\sqrt{\frac{0.07(1-0.07)}{100}}} = -1.57$$

Dvs vi forkaster ikke H_0. Dataene gir ikke grunnlag for å konkludere at andelen pumper med feil er redusert. Dvs vi kan ikke konkludere at $p < 0.07$.
p-verdien blir her:

$$p - \text{verdi} = P(Z \leq z_{obs}) = P(Z < -1.57) = 0.058$$

Dvs vi forkaster ikke testen på 5% nivå(men vi ser at vi ville forkaste testen på 10% nivå siden p-verdien<0.10).

Oppgave 37

a) Forventet antall er $\lambda t = 70 \cdot 3 = 210$.
For å regne sannsynligheten bruker vi at en Poisson-fordelt variabel med forventning/varians større enn 10 er tilnærmet normalfordelt. Vi har også at når Y er Poissonfordelt med parameter λt så er $\mathrm{E}(Y) = \mathrm{Var}(Y) = \lambda t = 210$. Dvs:

$$\begin{aligned} P(Y > 200) &= 1 - P(Y \leq 200) \approx 1 - P(Z \leq \frac{200 + 0.5 - 210}{\sqrt{210}}) \\ &= 1 - P(Z \leq -0.66) = 1 - 0.2546 = 0.75 \end{aligned}$$

Angir man svaret med tre desimaler får man over 0.745 mens man uten heltallskorreksjon får 0.755.

b) Totalt antall bakterier man vil finne i de fem prøvene er $Y_1 + Y_2 + Y_3 + Y_4 + Y_5$, og totalt volum for de fem prøvene er $3 \cdot 5 = 15$. Dvs estimatoren er totalt antall bakterier delt på totalt volum, eller med andre ord gjennomsnittlig antall bakterier per volumenhet i prøvene. Dette er derfor en rimelig estimator for λ som er gjennomsnittlig antall bakterier per volumenhet i hele kulturen.
Sjekker forventningsretthet og regner varians:

$$\begin{aligned}
\mathrm{E}(\hat{\lambda}) &= \mathrm{E}((Y_1 + Y_2 + Y_3 + Y_4 + Y_5)/15) = (\mathrm{E}(Y_1) + \mathrm{E}(Y_2) + \mathrm{E}(Y_3) + \mathrm{E}(Y_4) + \mathrm{E}(Y_5))/15 \\
&= (\lambda \cdot 3 + \lambda \cdot 3 + \lambda \cdot 3 + \lambda \cdot 3 + \lambda \cdot 3)/15 = (15\lambda)/15 = \underline{\underline{\lambda}} \\
\mathrm{Var}(\hat{\lambda}) &= \mathrm{Var}((Y_1 + Y_2 + Y_3 + Y_4 + Y_5)/15) = \frac{1}{15^2}\mathrm{Var}(Y_1 + Y_2 + Y_3 + Y_4 + Y_5) \\
&\overset{uavh.}{=} \frac{1}{15^2}(\mathrm{Var}(Y_1) + \mathrm{Var}(Y_2) + \mathrm{Var}(Y_3) + \mathrm{Var}(Y_4) + \mathrm{Var}(Y_5)) \\
&= \frac{1}{15^2}(\lambda \cdot 3 + \lambda \cdot 3 + \lambda \cdot 3 + \lambda \cdot 3 + \lambda \cdot 3) = \underline{\underline{\frac{\lambda}{15}}}
\end{aligned}$$

c) Med $\sum_{i=1}^{5} Y_i = 1084$ blir $\hat{\lambda} = 1084/15 = 72.27$
Siden vi her er i en situasjon der Y_i-ene er tilnærmet normalfordelte (Poisson med forventning >10) vil også $\hat{\lambda} = (Y_1 + Y_2 + Y_3 + Y_4 + Y_5)/15$ være tilnærmet normalfordelt. Vi kan da bruke det generelle resultatet fra pensum (se forelesningsnotatene/regel 6.8 i boka) at når $\hat{\lambda}$ er tilnærmet normalfordelt og forventningsrett så vil et tilnærmet 95% konfindensintervall for λ være gitt ved:

$$[\hat{\lambda} - z_{0.025} \cdot \mathrm{SD}(\hat{\lambda}), \hat{\lambda} + z_{0.025} \cdot \mathrm{SD}(\hat{\lambda})]$$

Her er $\mathrm{SD}(\hat{\lambda}) = \sqrt{\lambda/15}$ ukjent men kan estimeres ved $\widehat{\mathrm{SD}}(\hat{\lambda}) = \sqrt{\hat{\lambda}/15} = \sqrt{72.27/15} = 2.19$.
Tilnærmet 95% konfidensintervall blir dermed:

$$[72.27 - 1.96 \cdot 2.19, 72.27 + 1.96 \cdot 2.19] = \underline{\underline{[68.0,\ 76.6]}}$$

Oppgave 38

a)

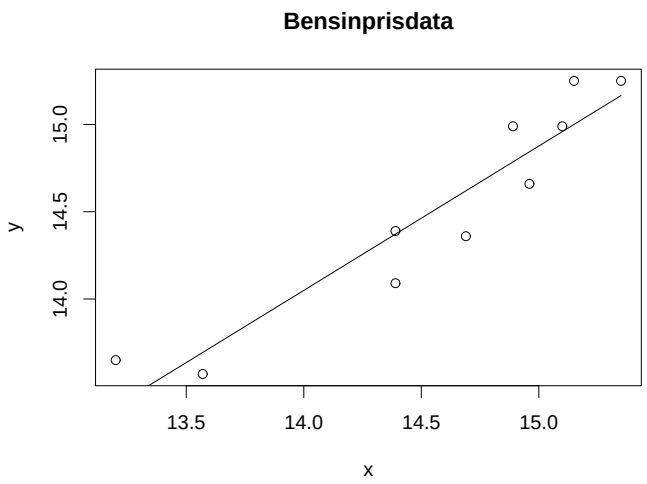
$$r = \frac{s_{xy}}{s_x \cdot s_y} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}} = \frac{3.670}{\sqrt{4.436 \cdot 3.414}} = \underline{\underline{0.94}}$$

Meget sterk positiv korrelasjon, bensinprisene følger hverandre i stor grad (går opp og ned i takt).

b)

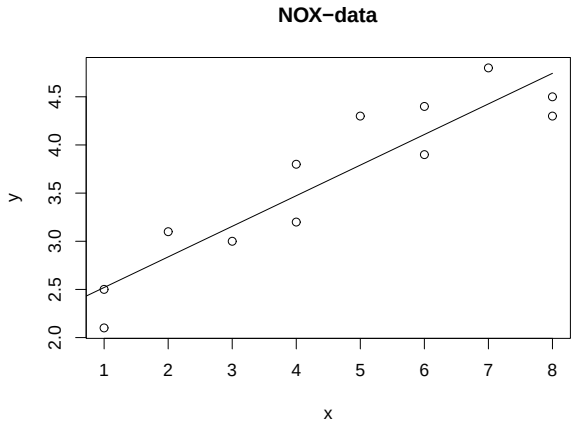
$$\begin{aligned}
\hat{\beta} &= \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{3.670}{4.436} = 0.827 \\
\hat{\alpha} &= \bar{y} - \hat{\beta}\bar{x} = 145.20/10 - 0.827 \cdot (145.69/10) = 2.47
\end{aligned}$$

Dvs estimert regresjonlinje blir $\underline{\underline{\hat{y} = 2.47 + 0.83x}}$.



Oppgave 39

a)



$$r = \frac{s_{xy}}{s_x \cdot s_y} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2 \sum_{i=1}^n (y_i - \bar{y})^2}} = \frac{21.89}{\sqrt{68.9 \cdot 8.19}} = \underline{\underline{0.92}}$$

Meget sterk positiv korrelasjon mellom mengde tilsetningsstoff og reduksjon i NOX-utslipp.

b) Modell: $Y = \alpha + \beta x + e$ der vi antar at $e \sim N(0, \sigma)$ og vi anser x som ikke-stokastisk. Spesielt betyr dette at $\mathrm{E}(Y) = \alpha + \beta x$ og at $\mathrm{Var}(Y) = \sigma^2$. Vi antar videre at feilleddene $e_1, \ldots, e_n$ for ulike målinger er uavhengige.
En praktisk tolkning av stigningstallet β vil her være endring i forventet reduksjon i NOX-utslipp per enhet økning i tilsetningsstoffet.

c)

$$\hat{\beta} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2} = \frac{21.89}{68.9} = 0.318$$
$$\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x} = 43.9/12 - 0.318 \cdot (55/12) = 2.20$$

Dvs estimert regresjonlinje blir $\underline{\underline{\hat{y} = 2.20 + 0.32x}}$.

d) Estimat av varians rundt regresjonslinja: $s^2 = SS_E/(n-2) = 1.235/10 = \underline{\underline{0.124}}$.
Om det er sammenheng tester vi ved testen:

$$H_0: \ \beta = 0 \qquad \text{mot} \qquad H_1: \ \beta \neq 0$$

Vi forkaster H_0 dersom $T = \frac{\hat{\beta}}{SE(\hat{\beta})} = \frac{\hat{\beta}}{\sqrt{s^2/\sum_{i=1}^n (x_i-\bar{x})^2}}$ faller utenfor forkastningsgrensene som for en tosidig test med $\alpha = 0.05$ og $n-2 = 10$ er $t_{\alpha/2,n-2} = t_{0.025,10} = 2.228$ og $-t_{0.025,10} = -2.228$.
Observert:

$$t = \frac{0.318}{\sqrt{0.124/68.9}} = 7.5$$

Dvs vi er langt utenfor forkastningsgrensene og forkaster H_0 og kan konkludere at det er sammenheng mellom mengde tilsetningsstoff og reduksjon i NOX-utslipp.

Oppgave 40

Med den lineæare regresjonsmodellen $Y = \alpha + \beta x + e$ der vi antar at $e \sim N(0,\sigma)$ og vi anser x som ikke-stokastisk har vi at $\mathrm{E}(Y) = \alpha + \beta x$ og at $\mathrm{Var}(Y) = \sigma^2$. Husk også at vi antar at alle målingene er uavhengige.

a) Merk først at: $n\bar{x} = n\frac{1}{n}\sum_{i=1}^n x_i = \sum_{i=1}^n x_i$.

Dette gir videre at: $\sum_{i=1}^n (x_i-\bar{x})\bar{x} = \bar{x}\sum_{i=1}^n (x_i-\bar{x}) = \bar{x}(\sum_{i=1}^n x_i - \sum_{i=1}^n \bar{x}) = \bar{x}(\sum_{i=1}^n x_i - n\bar{x})$
$= \bar{x}(\sum_{i=1}^n x_i - \sum_{i=1}^n x_i) = 0$.

Videre får vi fra dette at:
$\sum_{i=1}^n (x_i-\bar{x})^2 = \sum_{i=1}^n (x_i-\bar{x})(x_i-\bar{x}) = \sum_{i=1}^n (x_i-\bar{x})x_i - \sum_{i=1}^n (x_i-\bar{x})\bar{x} = \sum_{i=1}^n (x_i-\bar{x})x_i$.

Vi får da:

$$\begin{aligned}
\mathrm{E}(\hat{\beta}) &= \mathrm{E}\left(\frac{\sum_{i=1}^n (x_i-\bar{x})Y_i}{\sum_{i=1}^n (x_i-\bar{x})^2}\right) = \frac{\sum_{i=1}^n (x_i-\bar{x})\mathrm{E}(Y_i)}{\sum_{i=1}^n (x_i-\bar{x})^2} = \frac{\sum_{i=1}^n (x_i-\bar{x})(\alpha+\beta x_i)}{\sum_{i=1}^n (x_i-\bar{x})^2} \\
&= \frac{\sum_{i=1}^n (x_i-\bar{x})\alpha + \sum_{i=1}^n (x_i-\bar{x})\beta x_i}{\sum_{i=1}^n (x_i-\bar{x})^2} = \frac{\alpha(\sum_{i=1}^n x_i - \sum_{i=1}^n \bar{x}) + \beta\sum_{i=1}^n (x_i-\bar{x})x_i}{\sum_{i=1}^n (x_i-\bar{x})^2} \\
&= \frac{\alpha(\sum_{i=1}^n x_i - n\bar{x}) + \beta\sum_{i=1}^n (x_i-\bar{x})x_i}{\sum_{i=1}^n (x_i-\bar{x})x_i} = \frac{\alpha(\sum_{i=1}^n x_i - \sum_{i=1}^n x_i) + \beta\sum_{i=1}^n (x_i-\bar{x})x_i}{\sum_{i=1}^n (x_i-\bar{x})x_i} \\
&= \frac{\beta\sum_{i=1}^n (x_i-\bar{x})x_i}{\sum_{i=1}^n (x_i-\bar{x})x_i} = \underline{\beta}
\end{aligned}$$

b)

$$\begin{aligned}
\mathrm{Var}(\hat{\beta}) &= \mathrm{Var}\left(\frac{\sum_{i=1}^n (x_i-\bar{x})Y_i}{\sum_{i=1}^n (x_i-\bar{x})^2}\right) = \frac{1}{(\sum_{i=1}^n (x_i-\bar{x})^2)^2}\mathrm{Var}(\sum_{i=1}^n (x_i-\bar{x})Y_i) \overset{\text{uavh.}}{=} \frac{\sum_{i=1}^n (x_i-\bar{x})^2\mathrm{Var}(Y_i)}{(\sum_{i=1}^n (x_i-\bar{x})^2)^2} \\
&= \frac{\sum_{i=1}^n (x_i-\bar{x})^2\sigma^2}{(\sum_{i=1}^n (x_i-\bar{x})^2)^2} = \underline{\underline{\frac{\sigma^2}{\sum_{i=1}^n (x_i-\bar{x})^2}}}
\end{aligned}$$

Oppgave 41

a) Den generelle lineæare regresjonsmodellen er $Y = \alpha + \beta x + e$ der vi antar at $e \sim N(0, \sigma)$ og vi anser x som ikke-stokastisk. Spesielt betyr dette at $\mathrm{E}(Y) = \alpha + \beta x$ og at $\mathrm{Var}(Y) = \sigma^2$. Vi antar videre at feilleddene $e_1, \ldots, e_n$ for ulike målinger er uavhengige.
For hver leilighet er antall kvadratmeter en fast størrelse som ikke er stokastisk, vi setter derfor denne som x-variabel. Salgsprisen derimot er stokastisk, vi vet ikke hva den blir før salget er gjort, og vi setter derfor salgsprisen som Y-variabel.
En praktisk tolkning av β-parameteren er at den sier oss hvor mye forventet salgspris endrer seg per kvadratmeter endring i boligarealet.
Parameteren α har i denne situasjonen ikke noen direkte praktisk tolkning (pris for leilighet på 0 kvm...?), men vi trenger den for å justere regresjonslinjen til rett nivå.
Fra de oppgitte resultatene under tabellen med dataene får vi at gjennomsnittlig kvadratmeterpris er $\sum_{i=1}^{23} y_i / \sum_{i=1}^{23} x_i = 73420/1865 = \underline{\underline{39.4}}$ (dvs 39.4 tusen kroner).
(En annen måte å gi et gjennomsnittsmål på her er å regne ut pris per kvadratmeter for hver enkelt leilighet og så ta gjennomsnittet av dette for alle leilighetene.)

b) Fra datautskriften ser vi at den estimerte regresjonslinja her blir $\underline{\underline{\hat{y} = -362.95 + 43.84x}}$.
Ved en 10 kvadratmeter økning i areal øker estimert forventet pris med $43.84 \cdot 10 = \underline{\underline{438.4}}$.
Estimert forventet pris ved 75 kvadratmeter er: $\hat{y} = -362.95 + 43.84 \cdot 75 = \underline{\underline{2925}}$.

c) Det er sammenheng mellom x og Y dersom $\beta \neq 0$, og vi leser p-verdien for testen

$$H_0: \ \beta = 0 \qquad \text{mot} \qquad H_1: \ \beta \neq 0$$

rett ut fra datautskriften. Siden p-verdien$=2.7 \cdot 10^{-8} = 0.000000027$ (som er mindre enn alle vanlige signifikansnivå) ser vi at vi forkaster denne testen og kan konkludere med sammenheng.
Et $(1-\alpha)100\%$ konfidensintervall for regresjonslinja er gitt ved:

$$\hat{\alpha} + \hat{\beta}x \pm t_{\alpha/2,n-2} \cdot s \cdot \sqrt{\frac{1}{n} + \frac{(x-\bar{x})^2}{(s/SE(\hat{\beta}))^2}}$$

Vi har allerede regnet ut at for $x = 75$ er $\hat{\alpha} + \hat{\beta}x = 2925$. Fra datautskriften finner vi videre at: $s = 409.05$, $SE(\hat{\beta}) = 5.11$ og $n = 23$. Vi har også fra t-fordelingstabellen at $t_{\alpha/2,n-2} = t_{0.025,21} = 2.080$. Vi mangler da bare $\bar{x}$ som vi fra den oppgitte informasjonen under dataene får at er: $\bar{x} = 1865/23 = 81.09$. Intervallet blir da:

$$2925 \pm 2.080 \cdot 409.05\sqrt{\frac{1}{23} + \frac{(75-81.09)^2}{(409.05/5.11)^2}} = 2925 \pm 188.8 = \underline{\underline{[2736,\ 3114]}}$$

Et $(1-\alpha)100\%$ prediksjonsintervall for regresjonslinja er gitt ved:

$$\hat{\alpha} + \hat{\beta}x \pm t_{\alpha/2,n-2} \cdot s \cdot \sqrt{1 + \frac{1}{n} + \frac{(x-\bar{x})^2}{(s/SE(\hat{\beta}))^2}}$$

Innsatt samme tall som over får vi intervallet:

$$2925 \pm 2.080 \cdot 409.05\sqrt{1 + \frac{1}{23} + \frac{(75-81.09)^2}{(409.05/5.11)^2}} = 2925 \pm 871.5 = \underline{\underline{[2054,\ 3797]}}$$

Konfidensintervallet forteller oss at med stor sikkerhet er forventet salgspris for en leilighet på 75kvm et sted mellom 2736 og 3114. Dvs gjennomsnittlig salgspris for veldig mange leiligheter på 75kvm ligger med stor sikkerhet i dette intervallet.
Prediksjonsintervallet forteller oss at ca 95% av alle leiligheter på 75kvm vil ha en salgspris i intervallet fra 2032 til 3818.

d) Plottet øverst til venstre er et spredningsplott av dataene med regresjonlinja og prediksjonsintervall tegnet inn. Plottet øverst til høyre er et plott av residualene mot x (kvm=kvadratmeter). Plottet nederst til venstre er et histogram av residualene og plottet nederst til høyre er et normalplott av residualene.
Residualet til observasjon nr i er definert som: $\underline{\underline{\epsilon_i = y_i - \hat{\alpha} - \hat{\beta}x_i = y_i + 362.95 - 43.84x_i}}$. For den første observasjonen blir dette: $\epsilon_1 = 2952 + 362.95 - 43.84 \cdot 75 = \underline{\underline{27}}$.
Når vi tilpasser regresjonsmodellen $Y = \alpha + \beta x + e$ antar vi

1. $\mathrm{E}(Y) = \alpha + \beta x$, dvs lineær sammenheng.
2. $\mathrm{Var}(Y) = \sigma^2$, dvs konstant varians.
3. Feilleddet e normalfordelt.
4. Feilledden for ulike målinger $e_1, \ldots, e_n$ er uavhengige.

De tre første antagelsene kan vi sjekke fra de oppgitt plottene. Lineær sammenheng ser ut for å være ok da plott av residualene mot x (kvm) ikke viser noe bestemt systematisk mønster i Y-retning. Konstant varians derimot virker her å ikke være helt oppfylt - vi ser fra plottet av residualene mot x at det er større variasjon for de største leilighetene. Dette er et brudd på modellantagelsene som kan bidra til at analysene ikke er helt å stole på (alle beregninger hvor estimert standardavvik s inngår kan være beheftet med feil). Det er mulig å justere modellen slik at den tar høyde for økende varians, men det er utenfor pensum i dette kurset. Den tredje antagelsen om normalfordeling kan vi sjekke ved de to siste plottene, og dette ser ut til å være noenlunde ok. Histogrammet viser en noenlunde symetrisk form med topp i midten og normalplottet gir punkter omtrent på en rett linje - begge deler indikerer at normalfordeling er ok. Dog er der en par punkter i normalplottet som faller et stykke fra linja så litt forbehold må vi ta her også. Den fjerde antagelsen om uavhengighet kan vi ikke sjekke fra noen av de gitte plottene. En sjekk av dette kunne vi fått ved å plotte observasjonene i rekkefølge for når salget fant sted og se om det var noe bestemt mønster.

e) Fra datautskriften har vi at andel forklart variasjon er $r^2 = \underline{\underline{0.78}}$ - dvs regresjonslinja fanger opp mye av variasjonen i salgspris. Den gjenværende variasjonen omkring regresjonslinja kan blant annet skyldes faktorer som variasjon i beliggenhet, alder, teknisk standard, hvor flink eiendomsmegleren er, hvordan budrunden utvikler seg, etc.

Oppgave 42

a) Fra datautskriften ser vi at:
$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \hat{\beta}_3 x_3 = \underline{\underline{-5.37 + 2.22x_1 - 1.23x_2 - 0.0033x_3}}$ som spesielt gir $\hat{y} = -5.37 + 2.22 \cdot 20 - 1.23 \cdot 10 - 0.0033 \cdot 100 = \underline{\underline{26.4}}$.
Den estimerte parameterverdien for temperatur, $\hat{\beta}_1 = 2.22$, indikerer at når temperaturen øker med 1 grad så *øker* ozonmengden med 2.22.
Den estimerte parameterverdien for vind, $\hat{\beta}_2 = -1.23$, indikerer at når vinden øker med 1 enhet så *minker* ozonmengden med 1.23.
Den estimerte parameterverdien for sol, $\hat{\beta}_3 = -0.0033$, indikerer at når sol øker med 1 enhet så *minker* ozonmengden med 0.0033.

b)

$$H_0: \quad \beta_1 = \beta_2 = \beta_3 = 0 \qquad \text{mot} \qquad H_1: \quad \text{minst en } \beta_i \neq 0$$

Vi har fra pensum/formelarket at vi baserer denne testen på at under nullhypotesen er

$$F = \frac{SS_R/k}{SS_E/(n-k-1)} = \frac{MSR}{MSE} \sim F(\alpha, k, n-k-1)$$

og vi forkaster nullhypotesen dersom F blir stor. Fra den oppgitte R-utskriften ser vi at p-verdien for testen er $0.0066 < 0.05$, dvs $\underline{\underline{\text{vi forkaster } H_0}}$ $\underline{\underline{\text{på 5\% nivå}}}$. Værvariablene har innflytelse på ozonmengden.
Fra datautskriften ser vi at vi ikke vil forkaste $H_0 : \beta_i = 0$ mot $H_1 : \beta_i \neq 0$ verken for β_2 $(p - verdi = 0.124)$ eller for β_3 $(p - verdi = 0.895)$. Dette tyder på at minst en av variablene vind og sol kan utelates, spesielt tyder den høye $p - verdi$en for sol på at denne variabelen ikke har betydning.

c) Residual: $\underline{\underline{\epsilon_i = y_i - \hat{y}_i}}$ (observert verdi minus estimert regresjonslinje i punktet).
Måling nr 7 har temperatur 16 og ozon 8, residualet blir da: $\epsilon_7 = 8 - (-25.52 + 2.50 \cdot 16) = \underline{\underline{-6.5}}$.
For en enkel lineær regresjonsmodell som her kan antagelsen om lineær sammenheng mellom x og E(Y) og antagelsen om konstant varians undersøkes ved å plott residualene mot tilhørende $x - verdier$. Dette plottet bør ikke vise noe spesielt mønster, og ha lik variasjon i y-retning for ulike x-verdier.
Ved å lage et histogram eller et normalplott av residualene kan antagelsen om at feilleddene $e_1 \ldots e_{24}$ er normalfordelte sjekkes. Et histogram bør ha en form som ligner en normalfordelings symmetrisk om 0. Et normalfordelingsplott bør gi punkter som ligger omtrent på ei rett linje dersom normalfordelingsantagelsen er god.
Ved å plott residualene mot tid/observasjonsnummer kan man sjekke om antagelsen om at $e_1 \ldots e_{24}$ er uavhengige virker rimelig. Et slikt plott bør ikke ha noe bestemt mønster.
Alle residualplottene vist her ser OK ut, den tilpassede modellen ser med andre ord ut til å være en god modell. Vi har ikke noe mønster og konstant variasjon i plottet av residualene mot x-variabelen (tyder på at linære sammenheng og konstant varians er ok). I normalplottet ligger punktene noenlunde på en rett linje (tyder på at normalfordeling er OK). Plottet av residualene mot observasjonsnummer viser ikke noe mønster (uavhengighet OK).

d) $H_0: \ \beta = 2$ mot $H_1: \ \beta \neq 2$

Vi har fra pensum/formelarket at

$$T = \frac{\hat{\beta} - \beta}{SE(\hat{\beta})} \sim t(n-2)$$

Under $H_0: \ \beta = 2$ får vi da at

$$T = \frac{\hat{\beta} - 2}{SE(\hat{\beta})} \sim t(n-2)$$

og vi forkaster da H_0 dersom $T \geq t_{\alpha/2,n-2} = t_{0.025,22} = 2.074$ eller $T \leq -t_{\alpha/2,n-2} = -2.074$.
Observert: $t_{obs} = \frac{2.50-2}{0.683} = 0.73$
Dvs, vi forkaster ikke H_0 på 5% nivå, dataene gir ikke grunn til å konkludere at forholdet mellom endring i temperatur og ozon er anderledes her enn andre steder.

e) Et $(1-\alpha)100\%$ prediksjonsintervall for regresjonslinja er gitt ved:

$$\hat{\alpha} + \hat{\beta}x \pm t_{\alpha/2,n-2} \cdot s \cdot \sqrt{1 + \frac{1}{n} + \frac{(x-\bar{x})^2}{(s/SE(\hat{\beta}))^2}}$$

For $x = 25$ får vi $\hat{\alpha} + \hat{\beta}x = -25.52 + 2.50 \cdot 25 = 36.98$. Fra datautskriften finner vi videre at: $s = 12.3$, $SE(\hat{\beta}) = 0.683$ og $n = 24$. Vi har også fra t-fordelingstabellen at $t_{\alpha/2,n-2} = t_{0.025,22} = 2.074$. Vi mangler da bare $\bar{x}$ som vi fra den oppgitte informasjonen får at er: $\bar{x} = 456/24 = 19$. Intervallet blir da:

$$36.98 \pm 2.074 \cdot 12.3\sqrt{1 + \frac{1}{24} + \frac{(25-19)^2}{(12.3/0.683)^2}} = 36.98 \pm 27.39 = \underline{\underline{[9.6, \ 64.4]}}$$

Oppgave 43

a) Modell: $Y = \alpha + \beta x + e$ der $e \sim N(0, \sigma)$ vi anser x som ikke-stokastisk.
Antagelser:

- $\mathrm{E}(Y) = \alpha + \beta x$, dvs lineære sammenheng mellom x og $\mathrm{E}(Y)$.
- $\mathrm{Var}(Y) = \sigma^2$, dvs konstant varians, uavhengig av x.
- $e \sim N(0, \sigma)$, dvs normalfordelt variasjon om regresjonslinja.
- $e_1, \ldots, e_n$ uavhengige, dvs avvikene fra regresjonslinja ("feilleddene") for ulike målinger er uavhengige.

Residualet til observasjon nr i er definert som: $\underline{\epsilon_i = y_i - \hat{\alpha} - \hat{\beta}x_i = y_i + 0.637 - 0.772x_i}$.
For den første observasjonen blir dette: $\epsilon_1 = 0.95 + 0.637 - 0.772 \cdot 2.11 = \underline{\underline{-0.04}}$.
For den siste observasjonen blir dette: $\epsilon_{27} = 0.51 + 0.637 - 0.772 \cdot 1.74 = \underline{\underline{-0.20}}$.
Fra plottet kan vi sjekke om lineær sammenheng og konstant varians ser ut for å være oppfylt. Konstant varians ser ok ut, men det er et tydlig U-formet mønster i residualene som tyder på at lineær sammeheng ikke er oppfylt. Her virker det rimelig å prøve en modell med ikke-lineær sammeheng.

b) Residualplottet for den kvadratiske modellen ser vesentlig bedre ut, ikke noe mønster lenger, og fremdeles konstant varians. Dvs, kvadratisk sammenheng og konstant varians ser ut for å være rimelige antagelser.
Videre bør det lages et histogram eller normalplott av residualene for å vurdere normalfordeling, og et plott av residualene mot innsamlingsrekkefølgen for å vurdere uavhengighet.

c) Fra datautskriften ser vi at: $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x + \hat{\beta}_2 x^2 = \underline{\underline{3.204 - 3.646x + 1.215x^2}}$ som spesielt gir $\hat{y} = 3.204 - 3.646 \cdot 2.2 + 1.215 \cdot 2.2^2 = \underline{\underline{1.06}}$.
Vi trenger kvadratleddet i modellen dersom $\beta_2 \neq 0$, dvs vi skal teste:

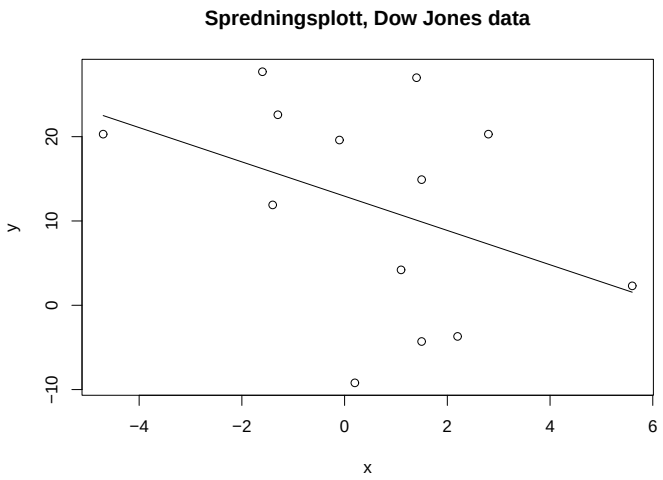
$$H_0: \quad \beta_2 = 0 \qquad \text{mot} \qquad H_1: \quad \beta_2 \neq 0$$

Vi leser p-verdien for testen rett ut fra datautskriften. Vi ser at for kvadratleddet i modellen så er p-verdien$= 7.2 \cdot 10^{-7} = 0.0000007$ (som er mindre enn alle vanlige signifikansnivå) dvs $\underline{\underline{\text{vi forkaster } H_0}}$ og konkluderer med at kvadratleddet trengs.

d) Vi ser at R^2-justert er 0.912 for modellen med andregradspolynom og 0.910 for modellen med tredjegradspolynom. Dvs, R^2-justert er så vidt høyere for andregradsmodellen, og denne modellen har også har færrest parametre så vurdert ut fra R^2-justert er andregradsmodellen å foretrekke.
Dersom vi ser på p-verdiene for de enkelte regresjonsparametrene ser vi at alle er signifikante i andregradsmodellen, mens ingen er signifikante i tredjegradsmodellen. Dette er også et sterkt signal om at tredjegradsmodellen ikke er å anbefale. (Det som skjer i praksis her er nok at de ulike polynomvariablene blir sterkt korrelerte og modelltilpasningen blir da ustabil.)

Oppgave 44

a) Fra datautskriften ser vi at den estimerte regresjonslinja her blir $\underline{\underline{\hat{y} = 12.94 - 2.03x}}$. At $\hat{\beta}$ er negativ betyr i praksis at estimert forventet endring for hele året avtar med økende verdi på x (betyr også at det er negativ korrelasjon mellom x og y).
Et spredningsplott av dataene med den estimerte regresjonslinja lagt inn er vist under.



b) Fra datautskriften ser vi at $r^2 = \underline{\underline{0.17}}$, dvs bare en veldig liten andel av variasjonen i dataene er forklart ved regresjonslinja.
Det er sammenheng mellom x og Y dersom $\beta \neq 0$. Dvs vi må teste:

$$H_0: \ \beta = 0 \qquad \text{mot} \qquad H_1: \ \beta \neq 0$$

Siden det i oppgaven ikke er oppgitt noe nivå for testen bruker vi det mest vanlige nivået, $\alpha = 0.05$. En p-verdi for testen er gitt i datautskriften, og p-verdien=0.17 er større enn nivået $\alpha = 0.05$ som betyr at $\underline{\underline{\text{vi forkaster ikke } H_0}}$. Dvs dataene gir ikke grunnlag for å konkludere at det er sammenheng mellom endringen av indeksen de fem første dagene og endringen hele året. Vi kan ikke bruke endringen av indeksen de første fem dagene til å predikere endringen for hele året!
(Alternativt til å bruke p-verdi kan vi utføre testen ved å sjekke om testobservatoren $T = \frac{\hat{\beta}}{SE(\hat{\beta})}$ for stigningstallet β faller utenfor forkastningsgrensene. I datautskriften er denne regnet ut til å bli -1.48. Forkastningsgrense for en tosidig test med $\alpha = 0.05$ og $n - 2 = 11$ er $t_{\alpha/2,n-2} = t_{0.025,11} = 2.201$ og $-t_{0.025,11} = -2.201$. Dvs vi er ikke utenfor forkastningsgrensene og forkaster ikke H_0.)

Oppgave 45

a)

$$H_0: \ \mu_X = \mu_Y \qquad \text{mot} \qquad H_1: \ \mu_X \neq \mu_Y$$

Siden vi har en uparet sammenligning av to utvalg med ukjent varians baserer vi testen på (pensum/formelsamling)

$$T = \frac{\bar{X} - \bar{Y}}{S_p\sqrt{\frac{1}{n_X} + \frac{1}{n_Y}}} \sim t(n_X + n_Y - 2)$$

Med nivå 5%, dvs $\alpha = 0.05$, tosidig test og $n_x = 12$ og $n_Y = 9$ forkaster vi H_0 dersom $T \leq -t_{0.025,19} = -2.093$ eller dersom $T \geq t_{0.025,19} = 2.093$.
For å regne ut testobservatoren må vi først regne ut s_p:

$$\begin{aligned} s_p^2 &= \frac{(n_x - 1)s_x^2 + (n_y - 1)s_y^2}{n_x + n_y - 2} = \frac{(n_x - 1)\frac{1}{n_x - 1}\sum_{i=1}^{n_x}(x_i - \bar{x})^2 + (n_y - 1)\frac{1}{n_y - 1}\sum_{i=1}^{n_y}(y_i - \bar{y})^2}{n_x + n_y - 2} \\ &= \frac{\sum_{i=1}^{n_x}(x_i - \bar{x})^2 + \sum_{i=1}^{n_y}(y_i - \bar{y})^2}{n_x + n_y - 2} = \frac{7206.25 + 2912}{12 + 9 - 2} = 532.5395 \\ s_p &= \sqrt{532.5395} = 23.08 \end{aligned}$$

Observert verdi på testobservatoren blir da: $t = \frac{5391/12 - 4284/9}{23.08\sqrt{\frac{1}{12} + \frac{1}{9}}} = -2.63$
Siden $-2.63 < -2.093$ blir konklusjonen at vi $\underline{\underline{\text{forkaster } H_0}}$. Dataene gir grunnlag for å konkludere at det er forskjell i forventet startlønn mellom menn og kvinner i denne bransjen (og siden $\bar{y} = 4284/9 = 476 > \bar{x} = 5391/12 = 449$ tyder disse tallene på av kvinner generelt har høyest lønnsnivå).

Oppgave 46

a) La $\mu_D = E(D)$. Siden $D = X - Y$ (tid før kurs minus tid etter kurs) gjør kurset arbeiderne mer effektive dersom $\mathrm{E}(D) > 0$. Dvs vi skal teste:

$$H_0: \ \mu_D \leq 0 \qquad \text{mot} \qquad H_1: \ \mu_D > 0$$

Siden vi har direkte målinger av differanser i tidsbruk person for person kan vi basere oss på disse differansemålingene $D_1, \ldots, D_n$ og bruke en vanlig ett-utvalgs t-test. kan Dersom H_0 er korrekt er

$$T = \frac{\bar{D} - 0}{S_D/\sqrt{n}} \sim t(n-1)$$

Med signifikansnivå 5%, dvs $\alpha = 0.05$, forkaster vi H_0 dersom $T \geq t_{0.05,9} = 1.833$.
Observert: $t = \frac{37/10}{\sqrt{268.1/9}/\sqrt{10}} = 2.14$. Siden $2.14 > 1.833$ blir konklusjonen at vi forkaster H_0.
Dataene gir grunnlag for å konkludere at forventet differanse i tidsbruk er positiv, dvs forventet tidsbruk etter kurset er lavere enn før kurset.
Testen bygger på antagelse om at differansemålingene $D_1, \ldots, D_n$ er uavhengige og normalfordelte.
Den typen forsøk som er gjort her med målinger av tidsbruk før og etter på de samme personene kalles en paret sammenligning.
En alternativ måte å undersøke effekten av kurset på kunne være å la en gruppe personer gjennomgå kurset og så la denne gruppen personer og en gruppe personer som ikke har gjennomgått kurset utføre samme type arbeidsoppgave og måle tidsbruken. Fra dette kunne man teste om det er forskjell mellom hvor lang tid kursgruppen og gruppen som ikke har hatt kurs bruker på oppgaven. Dette kalles en uparet sammenligning.
Den viktigste fordelen med paret sammenligning er at man får fjernet effekten av den tilfeldige variasjonen i tidsbruk mellom personer. Dermed kan man klare seg med å teste kurset på færre personer for å se eventuell effekt. En annen fordel med paret sammenligning er at problemet reduserer seg til et vanlig ett-utvalgsproblem (siden man kan basere seg på differansemålingene, D-ene). En mulig ulempe med den parete sammenligningen i denne situasjonen kan være dersom arbeidsoppgaven er en type oppgave arbeiderne sjelden utfører og de dermed har en læringseffekt fra første gang de utfører oppgaven som medfører at det vil tendere til å gjøre oppgaven raskere den andre gangen selv uten å gjennomgå et kurst.

Oppgave 47

a) Første skritt i testen er å regne ut forventet antall i hver celle i tabellen under antagelsen om lik fordeling i gruppene. Vi trenger da også rad- og kolonnesummene. Disse og de forventede verdiene er gitt i tabellen under (de forventede verdiene står i parantesene i hver celle). For eksempel finner vi forventet verdi for kombinasjonen "kvinner" og "sosialistisk" som: $510 \cdot (355/1000) = 29.3$ og for kombinasjonen "menn" og "stemte ikke" som: $490 \cdot (204/1000) = 100.0$. Testobservatoren blir da:

	sosialistisk	borgerlig	stemte ikke	Totalt
kvinner	203 (181.1)	211 (224.9)	96 (104.0)	510
menn	152 (174.0)	230 (216.1)	108 (100.0)	490
Totalt	355	441	204	1000

$$\begin{aligned}
\mathcal{Q} &= \sum_{\text{alle celler}} \frac{(\text{observert-forventet})^2}{\text{forventet}} \\
&= \frac{(203-181.1)^2}{181.1} + \frac{(211-224.9)^2}{224.9} + \frac{(96-104.0)^2}{104.0} \\
&\quad + \frac{(152-174.0)^2}{174.0} + \frac{(230-216.1)^2}{216.1} + \frac{(108-100.0)^2}{100.0} \\
&= 8.44
\end{aligned}$$

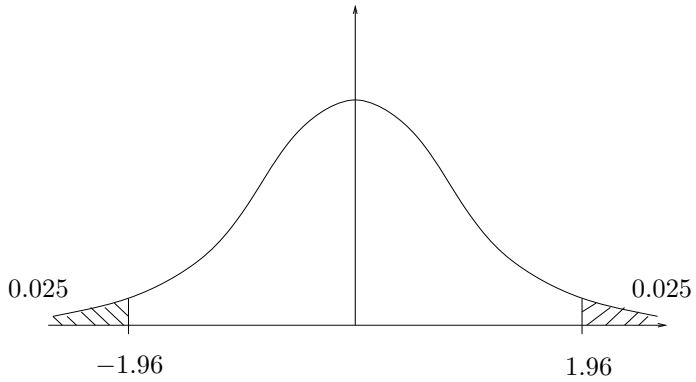
Denne verdien skal vi sammenligne med 5% kvantilen i kjikvadratfordelingen med parameter (frihetsgrader) $(r-1)\cdot(k-1) = (2-1)\cdot(3-1) = 2$. Fra tabellen på side 539 i boka finner vi at denne kvantilen har verdi 5.99. Siden $\mathcal{Q} = 8.44 > 5.99$ blir konklusjonen av vi forkaster nullhypotesene om lik stemmefordeling for menn og kvinner. Dvs, vi kan konkludere at det var en generell forskjell mellom menn og kvinner i hvordan de stemte.
Forutsetningen for testen er at forventet antall observasjoner i hver celle er minst 5. Dette ser vi er oppfylt her.

Oppgave 48

La X= humusinnhold.
a)

$$\begin{aligned}
P(\text{stanse}) = P(X > 13) &= 1 - P(X \le 13) = 1 - P(\frac{X-10}{\sqrt{2}} \le \frac{13-10}{\sqrt{2}}) \\
&= 1 - P(Z \le 2.12) = 1 - 0.983 = \underline{\underline{0.017}} \\
P(8 \le X \le 12) &= P(X \le 12) - P(X < 8) = P(Z \le \frac{12-10}{\sqrt{2}}) - P(Z < \frac{8-10}{\sqrt{2}}) \\
&= P(Z \le 1.41) - P(Z < -1.41) = 0.9207 - 0.0793 = \underline{\underline{0.8414}}
\end{aligned}$$

b) Vi skal her finne et 95% spredningsintervall:



$$\begin{aligned} P(10-a \le X \le 10+a) &= 0.95 \\ P(\frac{-a}{\sqrt{2}} \le \frac{X-10}{\sqrt{2}} \le \frac{a}{\sqrt{2}}) &= 0.95 \\ P(\frac{-a}{\sqrt{2}} \le Z \le \frac{a}{\sqrt{2}}) &= 0.95 \end{aligned}$$

Siden 2.5% kvantilen i standard normalfordeling er $z_{0.025} = 1.96$ får vi ønsket sannsynlighet ved å sette $\frac{a}{\sqrt{2}} = 1.96$ som gir $a = 1.96 \cdot \sqrt{2} = \underline{\underline{2.77}}$. Dvs det er 95% sannsynlighet for at humusinnholdet ligger i intervallet [10-2.77, 10+2.77]=$\underline{\underline{[7.23, 12.77]}}$. (Kontroller gjerne svaret!)

c)
- Flere enkeltforsøk som resulterer i "suksess" eller ikke "suksess" - flere dager hvor produksjonen må stanses eller ikke.
- Sannsynligheten for "suksess" er den samme i alle enkeltforsøk - samme sannsynlighet $p = 0.017$ hver dag for at produksjonen må stanses.
- Uavhengige enkeltforsøk - uavhengig fra dag til dag om produksjonen må stanses.
- Et bestemt antall forsøk - et bestemt antall, $n = 20$, dager.

Dvs, betingelsene for binomisk fordeling er oppfylte, og vi har dermed at Y="antall dager produksjonen må stanses i løpet av 20 arbeidsdager" vil være $\underline{\underline{\text{binomisk fordelt med parametre } n = 20 \text{ og } p = 0.017, Y \sim \text{Bin}(20, 0.017).}}$

d)
$$\begin{aligned} P(Y \ge 1) &= 1 - P(Y = 0) = 1 - \binom{20}{0} 0.017^0 (1-0.017)^{20} = 1 - 0.710 = \underline{\underline{0.29}} \\ \text{E}(Y) &= np = 20 \cdot 0.0.017 = \underline{\underline{0.34}} \end{aligned}$$

e)
$$\begin{aligned} \hat{\mu} &= \underline{\underline{\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i}} \\ \text{E}(\hat{\mu}) &= \text{E}(\frac{1}{n}\sum_{i=1}^{n} X_i) = \frac{1}{n}\text{E}(\sum_{i=1}^{n} X_i) = \frac{1}{n}\text{E}(X_1 + X_2 + \ldots + X_n) \\ &= \frac{1}{n}(\text{E}(X_1) + \text{E}(X_2) + \ldots + \text{E}(X_n)) = \frac{1}{n}(\mu + \mu + \ldots + \mu) = \frac{1}{n}(n\mu) = \underline{\underline{\mu}} \end{aligned}$$

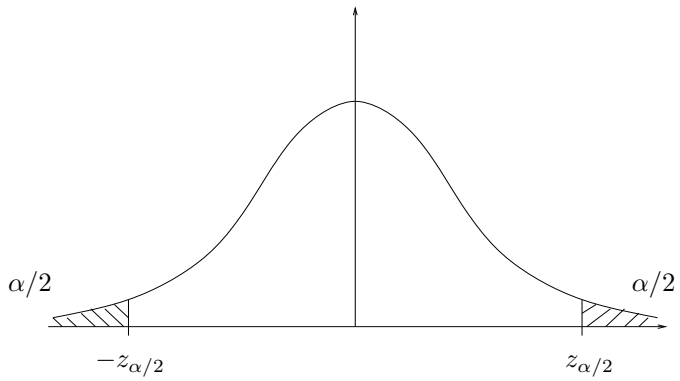
dvs estimatoren $\hat{\mu}$ er forventningsrett siden $\text{E}(\hat{\mu}) = \mu$.

$$\begin{aligned}
\mathrm{Var}(\hat{\mu}) &= \mathrm{Var}(\frac{1}{n}\sum_{i=1}^{n} X_i) = \frac{1}{n^2}\mathrm{Var}(\sum_{i=1}^{n} X_i) = \frac{1}{n^2}\mathrm{Var}(X_1 + X_2 + \ldots + X_n) \\
&= \frac{1}{n^2}(\mathrm{Var}(X_1) + \mathrm{Var}(X_2) + \ldots + \mathrm{Var}(X_n)) = \frac{1}{n}(\sigma^2 + \sigma^2 + \ldots + \sigma^2) \\
&= \frac{1}{n^2}(n\sigma^2) = \frac{\sigma^2}{n} = \frac{2}{10} = \underline{\underline{0.2}} \\
\hat{\mu} &= \bar{x} = \frac{1}{10}(7.3 + 7.3 + 7.6 + 8.3 + 11.1 + 10.1 + 6.8 + 7.5 + 8.1 + 6.3 + 9.6) = \underline{\underline{8.27}}
\end{aligned}$$

f) Situasjonen her er normalfordeling med ukjent μ og kjent σ.

$$\hat{\mu} = \overline{X}$$

$$Z = \frac{\overline{X} - \mathrm{E}(\overline{X})}{\sqrt{\mathrm{Var}(\overline{X})}} = \frac{\overline{X} - \mu}{\sqrt{\sigma^2/n}} = \frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \sim \mathrm{N}(0,1)$$



$$\begin{aligned}
P(-z_{\alpha/2} \leq \frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \leq z_{\alpha/2}) &= 1 - \alpha \\
P(-z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \leq \overline{X} - \mu \leq z_{\alpha/2}\frac{\sigma}{\sqrt{n}}) &= 1 - \alpha \\
P(\overline{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \leq \mu \leq \overline{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}) &= 1 - \alpha
\end{aligned}$$

Dvs et $(1-\alpha)100\%$ konfidensintervall for μ er gitt ved:

$$\underline{\underline{[\overline{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}}, \overline{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}]}}$$

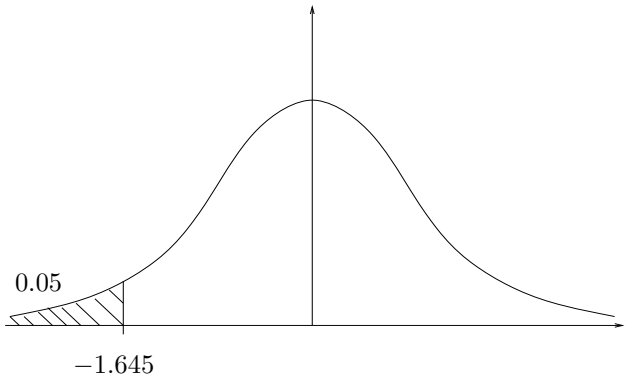
Med $\alpha = 0.05$ blir $z_{\alpha/2} = z_{0.025} = 1.96$, og med $\bar{x} = 8.27$, $n = 10$ og $\sigma = \sqrt{2}$ blir 95% konfidensintervallet:

$$[8.27 - 1.96\frac{\sqrt{2}}{\sqrt{10}}, 8.27 + 1.96\frac{\sqrt{2}}{\sqrt{10}}] = \underline{\underline{[7.39\ , 9.15]}}$$

g) $H_0: \ \mu \geq 10$ mot $H_1: \ \mu < 10$

Situasjonen her er normalfordeling med ukjent μ og kjent σ. Dersom H_0 er korrekt er da

$$Z = \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}} = \frac{\bar{X} - 10}{\sigma/\sqrt{n}} \sim N(0,1)$$



Med signifikansnivå 5%, dvs $\alpha = 0.05$, forkaster vi H_0 dersom $Z \leq -z_{0.05} = -1.645$.
Observert: $z = \frac{8.27-10}{\sqrt{2}/\sqrt{10}} = -3.87$
Siden $-3.87 < -1.645$ blir konklusjonen at vi $\underline{\text{forkaster } H_0}$. Dataene gir grunnlag for å konkludere at forventet humusinnhold er redusert, dvs renseanlegget har effekt.
(Dvs gjennomsnittet på 8.27 er så mye lavere enn 10 at vi her kan konkludere at den reelle forventningsverdien μ (gjennomsnittsverdien av veldig mange humusmålinger) må være lavere enn 10.)

p-verdien for denne ensidige testen blir p-verdi$=P(Z \leq z_{obs}) = P(Z \leq -3.87) = \underline{0}$.

h) For beregning av styrke så bruker vi formelen fra forelesningsnotatene om hypotesetesting (eller regel 6.18/6.15 i boka). Merk at vi har har en test av typen hvor $H_1: \ \mu < \mu_0$, og vi skal da bruke $\gamma(\mu) = P(Z \leq -z_\alpha + \frac{\mu_0 - \mu}{\sigma/\sqrt{n}})$. Med $\sigma = \sqrt{2}$, $n = 10$, $z_\alpha = z_{0.05} = 1.645$ og $\mu_0 = 10$ blir beregningen:

$$\gamma(9) = P(Z \leq -1.645 + \frac{10-9}{\sqrt{2}/\sqrt{10}}) = P(Z \leq 0.59) = \underline{0.72}$$

Dette betyr i praksis at dersom $\mu = 9$ er det 72% sannsynlighet for at testen i punkt g) vil gi forkastning (når $n = 10$).
En styrke på 90% betyr at $1 - \beta = 0.90$ eller $\beta = 0.10$ (der β =P(type II feil)). Da er $z_\beta = z_{0.10} = 1.282$. Formelen på formelarket/forelesningsnotatene gir oss da at nødvendig utvalgsstørrelser blir

$$n = \frac{(z_\beta + z_\alpha)^2\sigma^2}{(\mu_0 - \mu)^2} = \frac{(z_{0.1} + z_{0.05})^2\sigma^2}{(10-9)^2} = \frac{(1.282 + 1.645)^2 2}{(10-9)^2} = 17.1$$

Dvs de må gjøres minst $\underline{18}$ målinger for å få styrke på minst 0.90 (som er det samme som en sannsynlighet for type II feil på maks 0.10).

i) Vi estimerer σ^2 med utvalgsvariansen (som enklest regnes ut med å legge tallene inn på kalkulatoren): $s^2 = \frac{1}{n-1}\sum_{i=1}^{10}(x_i - \bar{x})^2 = \frac{1}{9}\sum_{i=1}^{10}(x_i - 8.27)^2 = \underline{2.35}$ (Dersom kalkulatoren din bare regner ut s må du opphøyde s-verdien i andre for å finne s^2.)
Vi er nå i situasjonen med normalfordeling med ukjent μ og ukjent σ og $(1-\alpha)100\%$ konfidensintervallet for μ er da gitt ved

$$[\bar{X} - t_{\alpha/2,n-1}\frac{S}{\sqrt{n}}, \quad \bar{X} + t_{\alpha/2,n-1}\frac{S}{\sqrt{n}}].$$

Med $\alpha = 0.05$ blir $t_{\alpha/2,n-1} = t_{0.025,9} = -2.262$, og med $\bar{x} = 8.27$, $n = 10$ og $s = \sqrt{2.35} = 1.53$ blir 95% konfidensintervallet:

$$[8.27 - 2.262\frac{1.53}{\sqrt{10}}, 8.27 + 2.262\frac{1.53}{\sqrt{10}}] = \underline{\underline{[7.18\ , 9.36]}}$$

j) Siden målingene er normalfordelte kan vi finne et konfidensintervall for σ ved å ta utgangspunkt i resultatet at $(n-1)\frac{S^2}{\sigma^2} \sim \chi(n-1)$ og regne videre som forklart i forelesningsnotatene. Vi får da at et $(1-\alpha)100\%$ konfidensintervall for σ er gitt ved

$$[\sqrt{\frac{(n-1)S^2}{\chi_{\alpha/2,n-1}}}, \sqrt{\frac{(n-1)S^2}{\chi_{1-\alpha/2,n-1}}}]$$

Med $\alpha = 0.05$ og $n = 10$ finner vi i tabell E.6 i boka at $\chi_{\alpha/2,n-1} = \chi_{0.025,9} = 19.02$ og $\chi_{1-\alpha/2,n-1} = \chi_{0.975,9} = 2.70$. Med $s^2 = 2.35$ får vi da

$$[\sqrt{\frac{(10-1)\cdot 2.35}{19.02}}, \sqrt{\frac{(10-1)\cdot 2.35}{2.70}}] = \underline{\underline{[1.05,\ 2.80]}}$$

k)

$$H_0: \ \mu \geq 10 \qquad \text{mot} \qquad H_1: \ \mu < 10$$

Når vi har normalfordeling med ukjent σ tar vi utgangspunkt i

$$T = \frac{\bar{X} - \mu_0}{S/\sqrt{n}} = \frac{\bar{X} - 10}{S/\sqrt{n}} \sim t(n-1)$$

Med signifikansnivå 5%, dvs $\alpha = 0.05$, forkaster vi H_0 dersom $T \leq -t_{0.05,9} = -1.833$. Observert: $t = \frac{8.27-10}{\sqrt{2.35}/\sqrt{10}} = -3.57$
Siden $-3.57 < -1.833$ blir konklusjonen fremdeles at vi $\underline{\underline{\text{forkaster } H_0}}$. Dataene gir grunnlag for å konkludere at forventet humusinnhold er redusert, dvs renseanlegget her effekt.

Oppgave 49

a) Vi har her en uparet sammenligning av to utvalg. Et $(1-\alpha)100\%$ konfidensintervall for $\mu_X - \mu_Y$ er da gitt ved

$$\left[\bar{X} - \bar{Y} - t_{\alpha/2,n_X+n_Y-2} S_p \sqrt{\frac{1}{n_X} + \frac{1}{n_Y}},\ \bar{X} - \bar{Y} + t_{\alpha/2,n_X+n_Y-2} S_p \sqrt{\frac{1}{n_X} + \frac{1}{n_Y}}\right]$$

Merk at det fra oppgitt informasjon er lett å regne ut s_p:

$$\begin{aligned} s_p^2 &= \frac{(n_x-1)s_x^2 + (n_y-1)s_y^2}{n_x+n_y-2} = \frac{(n_x-1)\frac{1}{n_x-1}\sum_{i=1}^{n_x}(x_i-\bar{x})^2 + (n_y-1)\frac{1}{n_y-1}\sum_{i=1}^{n_y}(y_i-\bar{y})^2}{n_x+n_y-2} \\ &= \frac{\sum_{i=1}^{n_x}(x_i-\bar{x})^2 + \sum_{i=1}^{n_y}(y_i-\bar{y})^2}{n_x+n_y-2} = \frac{30.920+28.401}{10+10-2} = 3.2956 \\ s_p &= \sqrt{3.2956} = 1.815 \end{aligned}$$

Videre har vi $t_{\alpha/2,n_X+n_Y-2} = t_{0.025,18} = 2.101$, $\bar{x} = 288.0/10 = 28.80$, $\bar{y} = 260.7/10 = 26.07$, og vi får da følgende 95% konfidensintervall for $\mu_x - \mu_y$:

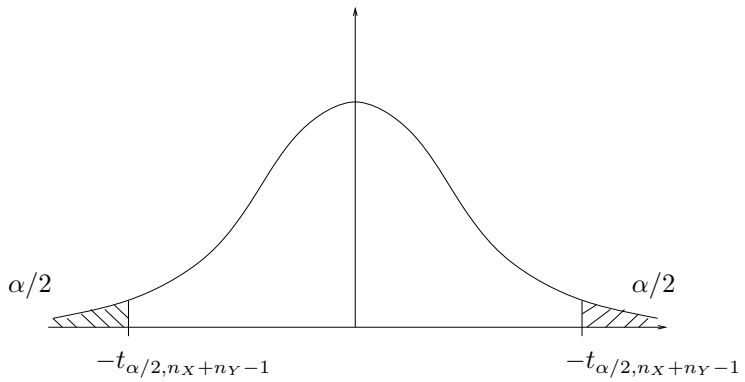
$$\left[28.80 - 26.07 - 2.101 \cdot 1.815\sqrt{\frac{1}{10}+\frac{1}{10}}, 28.80 - 26.07 + 2.101 \cdot 1.815\sqrt{\frac{1}{10}+\frac{1}{10}}\right] = \underline{\underline{[1.0, 4.4]}}$$

b)

$$H_0: \ \mu_X = \mu_Y \qquad \text{mot} \qquad H_1: \ \mu_X \neq \mu_Y$$

Siden vi har en uparet sammenligning av to utvalg med ukjent varians baserer vi testen på

$$T = \frac{\bar{X} - \bar{Y}}{S_p\sqrt{\frac{1}{n_X}+\frac{1}{n_Y}}} \sim t(n_X + n_Y - 2)$$



Med nivå 5%, dvs $\alpha = 0.05$, tosidig test og $n_x = n_Y = 10$ forkaster vi H_0 dersom $T \leq -t_{0.025,18} = -2.101$ eller dersom $T \geq t_{0.025,18} = 2.101$.
Observert: $t = \frac{28.8-26.07}{1.815\sqrt{\frac{1}{10}+\frac{1}{10}}} = 3.36$
Siden $3.36 > 2.101$ blir konklusjonen at vi $\underline{\underline{\text{forkaster } H_0}}$. Dataene gir grunnlag for å konkludere at det er forskjell i forventet bremselengde mellom dekkene.
Siden vi allerede har laget et 95% konfidensintervall kunne vi ha lest resultatet av testen rett ut av konfidensintervallet. Siden $\mu_X - \mu_Y = 0$ ikke er inneholdt i konfidensintervallet vil vi på 5% nivå forkaste den tosidig hypotesetest av nullhypotesen $\mu_X - \mu_Y = 0$ mot alternativet $\mu_X - \mu_Y \neq 0$.

Oppgave 50

a) Siden $\mu_D = \mu_X - \mu_Y$ har markedsføringstiltaket positiv effekt dersom $\mu_D > 0$ (da er $\mu_x > \mu_Y$, dvs forventet overskudd er større med tiltaket enn uten). Dvs vi skal teste:

$$H_0: \ \mu_D \leq 0 \qquad \text{mot} \qquad H_1: \ \mu_D > 0$$

Siden vi her har en parvis sammenligning basere vi oss på differansemålingene $D_1, \ldots, D_n$ og bruke en vanlig ett-utvalgs t-test. kan Dersom H_0 er korrekt er

$$T = \frac{\bar{D} - 0}{S_D/\sqrt{n}} \sim t(n-1)$$

Med signifikansnivå 5%, dvs $\alpha = 0.05$, forkaster vi H_0 dersom $T \geq t_{0.05,7} = 1.895$. Fra oppgitt informasjon får vi at $\bar{d} = 20/8 = 2.5$ og $S_D^2 = \frac{1}{n-1}\sum_{i=1}^n (D - \bar{D})^2 = 246/7 = 35.14$ slik at fra observerte data får vi: $t = \frac{2.5}{\sqrt{35.14/\sqrt{8}}} = 1.19$. Siden $1.19 < 1.895$ blir konklusjonen at vi forkaster ikke H_0. Dataene gir ikke grunnlag for å konkludere at tiltaket har positiv effekt.
Testen bygger på antagelsene at differansemålingene $D_1, \ldots, D_n$ er uavhengige og normalfordelte.

b) Merk at siden vi her gjør parvise sammenligninger så baserer vi oss på differansene og bruker vanlige ett-utvalgsmetoder på differansemålingene. Med antagelsen om uavhengighet, normalfordeling og at $\sigma_D^2 = 35.14$ kan vi bruke formelen for utvalgsstørrelse for test om μ fra formelarket/forelesningsnotatene. En styrke på 90% betyr at $1 - \beta = 0.90$ eller $\beta = 0.10$ (der β =P(type II feil)). Da er $z_\beta = z_{0.10} = 1.282$. Nødvendig utvalgsstørrelser blir da

$$n = \frac{(z_\beta + z_\alpha)^2\sigma^2}{(\mu_0 - \mu)^2} = \frac{(z_{0.1} + z_{0.05})^2\sigma^2}{(\mu_0 - \mu)^2} = \frac{(1.282 + 1.645)^2 \cdot 35.14}{(0-5)^2} = 12.04$$

Dvs de må test ut på minst 13 butikker for å få styrke på minst 0.90. Siden de har gjort tilnærmingen å anta at variansen er kjent og lik utvalgsvariansen fra punkt a), mens variansen egentlig er ukjent, bør de teste ut på en del flere butikker.

c) Det vi i praksis skal gjøre her er å regne ut nødvendig utvalgsstørrelse for testen

$$H_0: \ p \leq 0.50 \qquad \text{mot} \qquad H_1: \ p > 0.5$$

for å oppnå en styrke på 90% (dvs $1 - \beta = 0.90$ som gir $\beta = 0.1$) for alternativet $p = 0.70$ når testen utføres med nivå $\alpha = 0.05$. Fra formelen i forelesingsnotatene om hypotesetesting/på formelarket får vi da at

$$\begin{aligned} n &= \frac{(z_\beta\sqrt{p(1-p)} + z_\alpha\sqrt{p_0(1-p_0)})^2}{(p_0 - p)^2} = \frac{(z_{0.1}\sqrt{0.7(1-0.7)} + z_{0.05}\sqrt{0.5(1-0.5)})^2}{(0.5-0.7)^2} \\ &= \frac{(1.282\sqrt{0.7(1-0.7)} + 1.645\sqrt{0.5(1-0.5)})^2}{(0.5-0.7)^2} = 49.7 \end{aligned}$$

Dvs de må teste ut på minst 50 butikker.

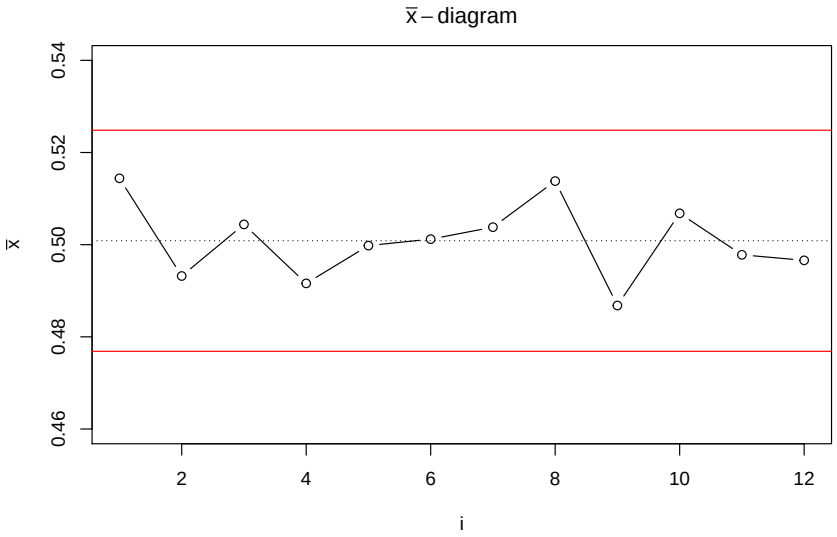
Den første fremgangsmåten bruker mer informasjon da størrelsene på overskuddene tas med i beregningene. Med den siste metoden brukes mye mindre informasjon da man bare ser på om overskuddet øker eller ikke - uten å ta hensyn til hvor mye det øker/minker. Siden den første metoden bruker mer informasjon klarer man seg med mange færre målinger. (Hvor mange målinger man trenger avhenger også av hvor store effekter man ønsker å kunne oppdage, jo mindre effekter jo flere målinger, men i scenarioene satt opp over ser vi i alle fall at man klarer seg med mange færre registreringer med første fremgangsmåte.) En fordel med den siste fremgangsmåten er at man ikke trenger å anta at differansene er normalfordelte. Videre er dataregistreringen litt enklere med den siste metoden (man trenger ikke vite størrelsen på differansene, bare om de er positive eller negative), men man må på den annen side teste ut tiltaket på mange flere butikker så det totale arbeidet blir likevel antagelig større. Dvs, med mindre differanse avviker klart fra normalfordeling blir totalvurderingen at den første fremgangsmåten er klart best.

Oppgave 51

a) Vi begynner med $\bar{x}$-diagrammet.

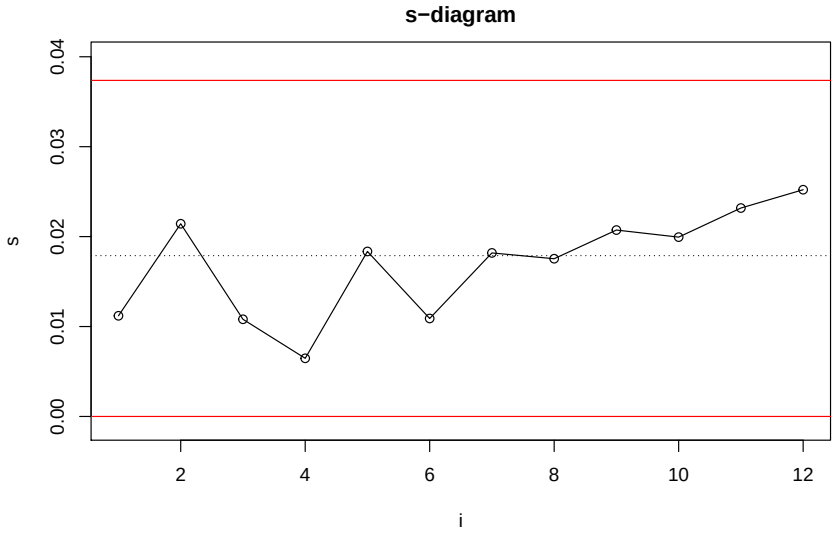
Senterlinje: $\bar{\bar{x}} = \frac{1}{12}\sum_{i=1}^{12}\bar{x}_i = \frac{1}{12}(0.514 + \cdots + 0.497) = 6.010/12 = 0.501$.
$\bar{s}^2 = \frac{1}{12}\sum_{i=1}^{12} s_i^2 = \frac{1}{12}(0.00012 + \cdots + 0.00064)) = 0.00386/12 = 0.000322$ og dermed blir $\bar{s} = \sqrt{0.000322} = 0.018$.
Øvre kontrollinje: $\bar{\bar{x}} + 3\frac{\bar{s}}{\sqrt{n}} = 0.501 + 3\frac{0.018}{\sqrt{5}} = 0.525$
Nedre kontrollinje: $\bar{\bar{x}} - 3\frac{\bar{s}}{\sqrt{n}} = 0.501 - 3\frac{0.018}{\sqrt{5}} = 0.477$

Plott av $\bar{x}$-diagrammet er gitt under:



Siden punktene i $\bar{x}$-diagrammet holder seg godt innenfor grensene og vi ikke ikke ser noe bestemt mønster kan vi konkludere med at prosessen er stabil i forhold til forventningsverdien.

For å lage s-diagram har vi allerede senterlinja, $\bar{s} = 0.018$. Videre er (der vi finner B_L og B_U i tabell i forelesingsnotatene):
Øvre kontrollinje: $\bar{s}B_U = 0.018 \cdot 2.09 = 0.038$.
Nedre kontrollinje: $\bar{s}B_L = 0.018 \cdot 0 = 0$.
Plott av s-diagrammet:



Siden punktene i s-diagrammet holder seg godt innenfor grensene og vi ikke ikke ser noe bestemt mønster kan vi konkludere med at prosessen har stabil varians.

b) Toleransegrensene som stilles til prosessen er altså $T_L = 0.48$ og $T_U = 0.52$. Siden $\bar{\bar{x}} = 0.501 \approx (T_L + T_U)/2 = 0.50$ (vi kan si at prosessen er velsentrert) kan vi bruke den enkleste versjonen av kapabilitetsindeksen:

$$KI = \frac{T_U - T_L}{6\hat{\sigma}} = \frac{0.52 - 0.48}{6 \cdot 0.018} = 0.37$$

Siden $KI < 1.33$ kan vi konkludere med at prosessen ikke holder god nok kvalitet - variasjonen i volum er altfor stor i forhold til kravet som stilles! (Dersom du ser på enkeltmålingene i tabellen ser du at det er en god del målinger som faller utenfor intervallet $[0.48,\ 0.52]$. Merk også at det er feil i fasiten bak i boka på denne oppgaven.)

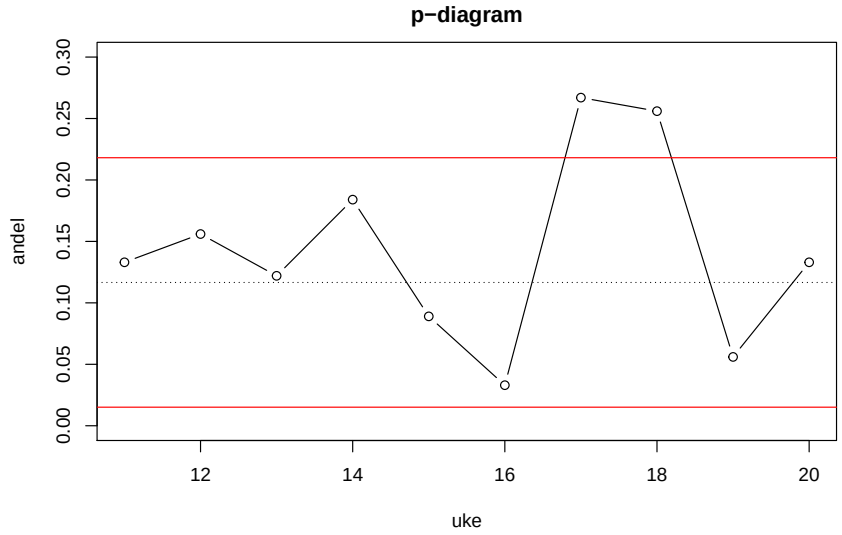
Oppgave 52

a) Senterlinje: $\bar{p} = (0.133 + 0.067 + \cdots + 0.144)/10 = \underline{\underline{0.117}}$.

Øvre kontrollinje: $\bar{p} + 3\sqrt{\bar{p}(1-\bar{p})/n} = 0.117 + 3\sqrt{0.117(1-0.117)/90} = \underline{\underline{0.219}}$.

Nedre kontrollinje: $\bar{p} - 3\sqrt{\bar{p}(1-\bar{p})/n} = 0.117 - 3\sqrt{0.117(1-0.117)/90} = \underline{\underline{0.015}}$.

Plott av p-diagrammet er gitt under:



Vi ser at i uke 17 og 18 er prosessen utenfor kontrollgrensene - i disse ukene har det skjedd noe som har medført en unormalt høy andel feilaktig ekspederte ordre.