Løsningsforslag M4STI1 2018E

Opgave 1 – Hydrauliske pumper til eksoskelettet

a. *Beregn stikprøvens gennemsnit $\bar{y}$ og standardafvigelse s. Vurdér ud fra resultaterne, om leverandørens specifikationer virker troværdige*

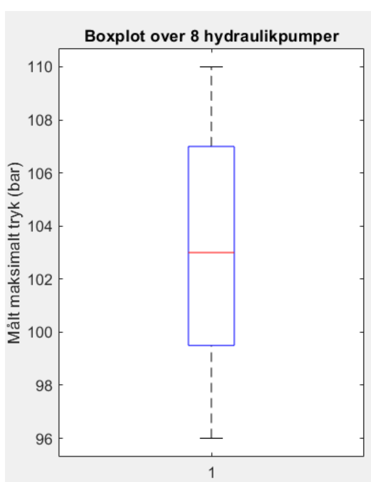
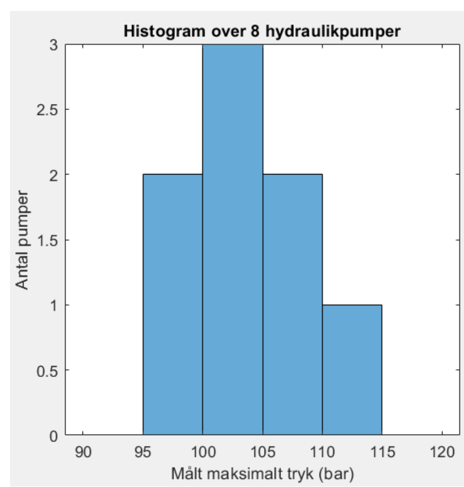
Stikprøvens gennemsnit og standardafvigelse beregnes med MatLab funktioner, hhv. mean() og std():

$$\bar{y} = \mathbf{103.1250}$$
$$s = \mathbf{4.9407}$$

Disse værdier ligger under begge oplyste værdier, hhv. 105 og 5.0, men dog ret tæt på. Derfor virker leverandørens specifikationer umiddelbart troværdige.

b. *Præsentér data fra stikprøven med et stem-and-leaf diagram, et boksplot og et histogram. Hvad fortæller diagrammerne om stikprøven?*

Her er stikprøven visualiseret som hhv. stem-and-leaf diagram, boksplot og histogram:

<table>
<tr>
<td>9 | 6 7
10 | 2 2 4 6 8
11 | 0
key: 36|5 = 365
stem unit: 10
leaf unit: 1</td>
<td>
</td>
<td>
</td>
</tr>
<tr>
<td>Stem-and-leaf diagram</td>
<td>Boksplot</td>
<td>Histogram</td>
</tr>
</table>

Diagrammerne viser alle, at stikprøven kommer fra en 'pæn' fordeling, der godt kunne ligne normalfordelingen. Den er nogenlunde symmetrisk med et enkelt toppunkt og uddøende haler. Stem-and-leaf diagrammet har kun tre stammer, med flest blade i den midterste stamme. Boksplottet har median 103, symmetrisk 'Interkvartil range' omkring medianen og symmetriske koste. Histogrammet kan måske virke usymmetrisk, men det skyldes en enkelt observation, der giver den øverste søjle. Når stikprøvestørrelsen er lille, som her (8), så kan tilfældigheder påvirke diagrammernes udseende.

c. *Er stikprøven plausibel?*

Det følger af den centrale grænseværdisætning (CGS), at $\bar{y}$ er normalfordelt med middelværdi μ og standardafvigelse $\sigma/\sqrt{n}$, hvis stikprøvestørrelsen n er tilstrækkelig stor.
Her er $\mu = 105$, $\sigma = 5.0$ og $n = 8$.

Når $\bar{y}$ ifølge CGS er normalfordelt følger det, at

$$z_0 = \frac{\bar{y}-\mu}{\sigma/\sqrt{n}} = \frac{103.125-105}{5.0/\sqrt{8}} = -1.0607$$

er *standard* normalfordelt. Det betyder, at vi under antagelse af CGS kan regne ud hvor sandsynlig stikprøven er vha. standard normalfordelingen.

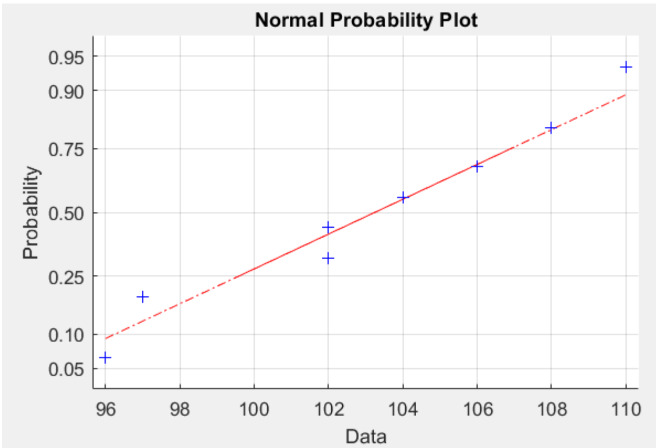
Sandsynligheden for at få en stikprøve med $z_0 = -1.0607$ eller mere ekstrem svarer til at få en z_0, der enten er mindre end -1.0607 eller er større end 1.0607 (dvs. længere væk fra middelværdien 0 end $z_0 = -1.0607$. Det kan formuleres enklere som sandsynligheden for at få $|z_0| > 1.0607$:

$$p = 2 \cdot (1 - \text{normcdf}(|z_0|)) = \mathbf{0.2888}$$

Med andre ord: Hvis leverandørens specifikationer af pumpernes middelværdi og standardafvigelse er korrekte, så vil vi forvente at få en stikprøve som den pågældende, eller mere ekstrem, i knap 29% af tilfældene. Det er ikke usandsynligt, så vi må tro på specifikationerne.

d. *Hvad kan du konkludere om pumperne på baggrund af stikprøven, og under hvilke antagelser?*

Stikprøven underbygger leverandørens specifikationer. I delspørgsmål a. så vi, at stikprøvens gennemsnit og standardafvigelse er tæt på de oplyste værdier. I delspørgsmål c. så vi, at stikprøven er sandsynlig for en population med $\mu = 105$ bar og $\sigma = 5.0$ bar. Det konkluderede vi under antagelse af den centrale grænseværdisætning, som forudsætter, at $n = 8$ er tilstrækkelig stor. I delspørgsmål b. så vi, at stikprøven kommer fra en pæn fordeling: symmetrisk, med et enkelt toppunkt og ingen outliers. Derfor er $n = 8$ stor nok til at den centrale grænseværdisætning er opfyldt. Dette bekræftes yderligere af et normalfordelingsplot, der viser at observationerne ligger nogenlunde på en ret linje:



e. *Beregn et 99 % konfidensinterval for middelværdien*

Konfidensintervallet beregnes med formlen:

$$\bar{y} \pm t_{\alpha/2,n-1} \cdot s \cdot \sqrt{1/n}$$

hvor $\alpha = 0.01$.

99% konfidensinterval: **[97.0; 109.2]**

f. *Beregn et 99 % prædiktionsinterval*

Prædiktionsintervallet beregnes med formlen:

$$\bar{y} \pm t_{\alpha/2,n-1} \cdot s \cdot \sqrt{1 + 1/n}$$

99% prædiktionsinterval: **[84.8; 121.5]**

g. *Beregn stikprøvestørrelsen, hvis 99 % konfidensintervallet skal have en total bredde på højst 6.0 bar*

Bemærk: Denne opgave viste sig at være formuleret, så den var for svær at regne. Det blev der taget hensyn til i bedømmelsen.

Problemet er, at der står i opgaven, at man skal benytte stikprøven og se bort fra leverandørens specifikationer af middelværdi og standardafvigelse for pumperne.

Hvis der ikke havde stået det, kunne vi benytte populations-standardafvigelsen $\sigma = 5.0$ og formlen: $n \geq \left(\frac{z_{\alpha/2} \cdot \sigma}{B}\right)^2$

Her er B signifikansintervallets halve bredde, altså $B = 6.0/2 = 3.0$.

$z_{\alpha/2}$ beregnes som $z_{\alpha/2}$=$\mathrm{norminv}(1-\alpha/2) = 2.5758$. På den måde kunne stikprøvestørrelsen beregnes til $n \geq 18.4303$. Da n er et heltal skal vi runde op til $\boldsymbol{n = 19}$. Men som sagt, benytter vi her leverandørens specifikationer $\sigma = 5.0$, som vi ikke må ifølge opgaveformuleringen.

Hvis vi i stedet benytter stikprøvestandardafvigelsen $s = 4.9407$, skal vi bruge t-fordeling i stedet for standard normalfordeling: $t_{df,\alpha/2} \; = \; \mathrm{tinv}(1-\alpha/2, df)$.

Problemet er at værdien af $t_{\alpha/2}$ afhænger af stikprøvestørrelsen n, for antal frihedsgrader df er $n-1$. Derfor kan vi ikke udtrykke stikprøvestørrelsen som funktion af bredden (vi kan ikke isolere n i formlen).

I stedet kan man beregne signifikansintervallets halve bredde B for alle værdier af stikprøvestørrelse, n = 1, 2, …, indtil $B \leq 3.0$. B beregnes med formlen:

$$B = t_{df,\alpha/2} \cdot \frac{s}{\sqrt{n}}$$

Det viser sig, at for $n = 22$ bliver $B = 2.9692 < 3.0$.

Dermed er det korrekte svar $\boldsymbol{n = 22.}$

MatLab kode til opgave 1:

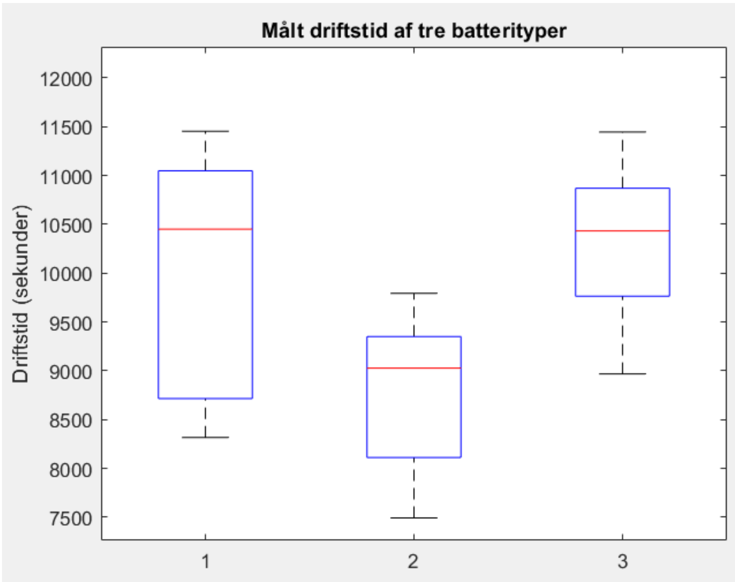
```
%% Eksamen M4STI1 2018E opgave 1: Hydrauliske pumper til eksoskelettet
clear; close all; clc; format compact;

%% Oplysninger fra opgaveformuleringen
mu = 105      % Oplyst populations-middelværdi
sigma = 5.0   % Oplyst populations-standardafvigelse

%% Indlæs og behandl data
Y = xlsread('Data_M4STI1_2018E.xlsx','A:A')
n = size(Y,1)

%% a: Stikprøve-middelværdi og -standardafvigelse
y_streg = mean(Y)   % y_streg = 103.1250
s = std(Y)          % s = 4.9407

% Stikprøve-middelværdien på y_streg = 103.1250 er en anelse under den
% oplyste middelværdi på 105 bar. Tilsvarende er
% stikprøve-standardafvigelsen s = 4.9407 en anelse under den oplyste værdi
% på 5.0 bar. Specifikationerne ser umiddelbart troværdige ud.


%% b: Stem-and-leaf diagram, boksplot og histogram
stemleafplot(Y)

figure(1)
boxplot(Y)
title('Boxplot over 8 hydraulikpumper');
ylabel('Målt maksimalt tryk (bar)');

figure(2)
histogram(Y,[90:5:120])
title('Histogram over 8 hydraulikpumper');
xlabel('Målt maksimalt tryk (bar)');
ylabel('Antal pumper');
% Diagrammerne viser alle, at stikprøven kommer fra en 'pæn' fordeling, der
% godt kunne ligne normalfordelingen. Den er nogenlunde symmetrisk med et
% enkelt toppunkt og uddøende haler. Boxplottet har median 103, symmetrisk
% 'Interkvartil range' omkring medianen og symmetriske koste.


%% c: Er stikprøven plausibel?
% Det følger af den centrale grænseværdisætning, at y_streg er
% normalfordelt med middelværdi mu og standardafvigelse sigma/sqrt(n),
% hvis n er tilstrækkelig stor. Derfor er z0 = (y_streg-mu)/(sigma/sqrt(n))
% standard normalfordelt:
z0 = (y_streg-mu)/(sigma/sqrt(n))           % z0 = -1.0607

% Sandsynligheden for at få en stikprøve med z0 = -1.0607 eller mere
% ekstrem svarer til at få |z0| > 1.0607 (d.v.s. enten mindre end -1.0607
% eller større end 1.0607):
P_stikproeve = 2*(1 - normcdf(abs(z0)))     % P_stikproeve = 0.2888

% Det er altså slet ikke usandsynligt, at få en stikprøve som vores, hvis
% leverandørens specifikationer er korrekte. Det kan forventes i knap 29%
% af stikprøver.


%% d: Konklusion og antagelser
% Stikprøven underbygger leverandørens specifikationer.
% I delspørgsmål a. så vi at stikprøvens gennemsnit og standardafvigelse
% er tæt på de oplyste værdier.
% I delspørgsmål c. så vi, at stikprøven er sandsynlig for en population
% med mu = 105 bar og sigma = 5.0 bar. Det konkluderede vi under antagelse
% af den centrale grænseværdisætning, som forudsætter, at n = 8 er
% tilstrækkelig stor.
```

```
% I delspørgsmål b. så vi, at stikprøven kommer fra en pæn fordeling:
% symmetrisk, med et enkelt toppunkt og ingen outliers. Derfor er n = 8
% stor nok til at den centrale grænseværdisætning er opfyldt. Dette bekræftes
% yderligere af et normalfordelingsplot, der viser at observationerne
% ligger nogenlunde på en ret linje:
normplot(Y)


%% e: 99% konfidensinterval for populationsmiddelværdien mu
% Når vi ikke kan bruge leverandørens oplyste standardafvigelse sigma, må
% vi beregne konfidensintervallet, hvor stikprøvestandardafvigelsen s
% indgår. Formlen er:
% y_streg +/- t_alfahalve*s*sqrt(1/n)

alfa = 0.01
t_alfahalve = tinv(1-alfa/2, n-1)       % t_alfahalve = 3.4995
ki_bredde = t_alfahalve*s*sqrt(1/n)     % ki_bredde = 6.1129
ki_lav = y_streg - ki_bredde            % ki_lav = 97.0121
ki_hoej = y_streg + ki_bredde           % ki_hoej = 109.2379
% 99% konfidensinterval: [97.0; 109.2]


%% f: 99% prædiktionsinterval:
pi_bredde = t_alfahalve*s*sqrt(1 + 1/n) % pi_bredde = 18.3388
pi_lav = y_streg - pi_bredde            % pi_lav = 84.7862
pi_hoej = y_streg + pi_bredde           % pi_hoej = 121.4638
% 99% prædiktionsinterval: [84.8; 121.5]


%% g: Stikprøvestørrelse, så konfidensintervallets bredde er højst 6.0
% N.B. Denne opgave viste sig at være formuleret, så den var for svær at
% regne. Det blev der taget hensyn til i bedømmelsen. Problemet er, at der
% står i opgaven, at man skal benytte stikprøven og se bort fra leverandørens
% specifikationer af middelværdi og standardafvigelse for pumperne.

% Hvis der ikke havde stået det, kunne vi benytte populations-
% standardafvigelsen sigma=5.0 og formlen: n_ny = (z_alfahalve*sigma/B)^2
% Så kunne opgaven løses sådan:
B = 6.0/2                           % Med B = 3.0 bliver den totale
                                    % intervalbredde 6.0
z_alfahalve = norminv(1-alfa/2)     % z_alfahalve = 2.5758
n_ny = (z_alfahalve*sigma/B)^2      % n_ny = 18.4303
% Vi runder op til nærmeste heltal, så intervalbredden er højst 6.0
n_ny_min = ceil(n_ny)               % n_ny_min = 19

% Men som sagt benyttes her leverandørens specifikationer sigma=5.0.
% Hvis vi benytter stikprøvestandardafvigelsen s=4.9407, skal vi bruge
% t-fordeling i stedet for standard normalfordeling:
% t_alfahalve = tinv(1-alfa/2, n-1).
% Problemet er at t_alfahalve afhænger af stikprøvestørrelsen n. Derfor kan
% vi ikke udtrykke stikprøvestørrelsen som funktion af bredden (vi kan
% ikke isolere n i formlen).

% Her er den korrekte (men for svære) løsning af opgaven:
% I stedet for at udtrykke stikprøvestørrelsen som funktion af bredden, så
% løber jeg igennem en løkke, hvor bredden beregnes for større og større
% stikprøvestørrelse, indtil bredden er kommet under de ønskede (B = 3.0):
n_ny = 0;                 % Jeg starter med mindst mulig stikprøvestørrelse
bredde = 99999;           % Bredden sættes til en tilpas stor startværdi
while bredde > B
    n_ny = n_ny + 1;
    t_alfahalve = tinv(1-alfa/2, n_ny);
    bredde = t_alfahalve*s*sqrt(1/n_ny);
    % Variablen 'bredde' angiver konfidensintervallets halve bredde når
    % stikprøvestørrelsen er n_ny. Nu skal vi teste om bredde er kommet
    % under B, så vi er færdige, eller om skal fortsætte i løkken og forøge
    % stikprøvestørrelsen.
end
```

```
bredde      % bredde = 2.9692
n_ny        % n_ny = 22
% Altså: når n_ny kommer op på 22, så kommer den halve bredde under 3.0, så
% den totale bredde af konfidensintervallet kommer under 6.0
```

Opgave 2 - Batterier til eksoskelettet

a. *Lav og kommenter et parallelt boksplot, der viser driftstiden for hver batteritype*



Batteritype 1 og 3 har nogenlunde samme median, hvor type 2 ligger tydeligt lavere. Boksplottene for type 2 og 3 ser meget ensartede ud, bortset fra niveauet, hvor type 3 ligger højere end type 2. Type 1 ligger på niveau med type 3, men lader til at have lidt større variation. Der er dog ikke stor forskel på afstanden imellem kostenes ender på type 1 og 3. Da hvert boksplot er tegnet på baggrund af kun 5 observationer er det svært at konkludere, alene ud fra boksplottet, om der er reelle forskelle på de tre batterityper.

b. *Undersøg i en variansanalyse med 5 % signifikansniveau, om batterierne har samme driftstid*

Variansanalysen laves med MatLab funktionen anovan(). Den viser følgende ANOVA tabel:

Source	Sum Sq.	d.f.	Mean Sq.	F	Prob>F
X1	6.6572e+06	2	3.3286e+06	2.9	0.0938
Error	1.3765e+07	12	1.14708e+06		
Total	2.04222e+07	14			

Variansanalysen har F = 2.9 og en tilhørende p-værdi på 0.0938. Det vil sige, at vi kan ikke afvise på 5 % signifikansniveau, at alle 3 batterityper opfører sig ens (da 0.0938 > 0.05). Med andre ord: Bedømt ud fra variansanalysen kan alle tre batterityper have samme forventede driftstid.

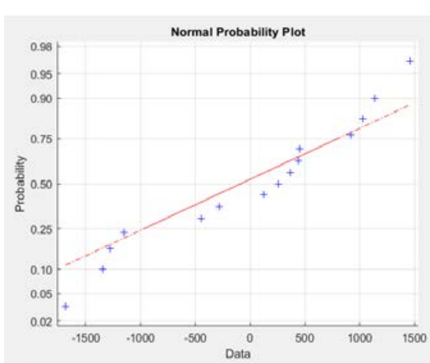
c. *Lav en parvis sammenligning af batterierne, stadig med 5 % signifikansniveau*

Variansanalysen viser, at type 1 ikke er signifikant forskellig fra hverken type 2 eller 3. Type 2 og type 3 er forskellige på 5% signifikansniveau (p-værdi 0.0417). Dette ses af både matricen c og grafikken, der begge er output fra MatLab funktionen multcompare():

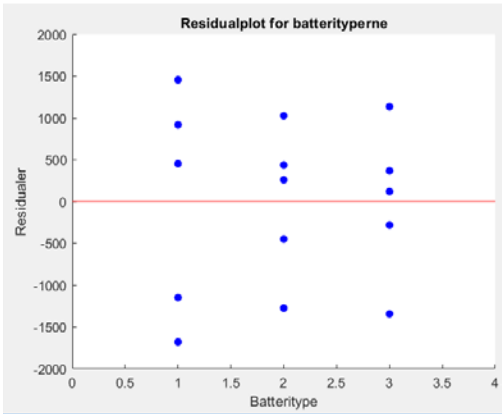
```
c =
        1         2    -246.07     1229.8     2705.7    0.094494
        1         3    -1789.9       -314     1161.9     0.65126
        2         3    -3019.7    -1543.8    -67.932    0.041746
```

d. *Hvilke antagelser er der gjort for residualerne i variansanalysen? Undersøg om antagelserne er overholdt*

Det er en antagelse i variansanalysen, at residualerne er normalfordelte med middelværdi 0 og med samme varians for alle batterityper. Vi kan teste, om residualerne er normalfordelte med et normalfordelingsplot:



Residualerne ligger nogenlunde på en ret linje, så den antagelse ser ud til at holde. Antagelsen om varianshomogenitet kan vi teste visuelt med et plot af residualerne:



Residualerne for type 2 og 3 er meget ensartede, men de er lidt større for type 1. Vi kan undersøge med en Bartlett's test, om der er signifikant forskel:

Group	Count	Mean	Std Dev
1	5	9997.2	1351.55
2	5	8767.4	885.59
3	5	10311.2	911.2
Pooled	15	9691.9	1071.02

Bartlett's statistic	0.85733
Degrees of freedom	2
p-value	0.65138

Bartlett's testen viser, at vi ikke kan forkaste nulhypotesen om at variansen er ens for de tre typer. P-værdien på 0.65138 fortæller, at vi kan forvente residualer som vores, eller mere ekstreme, i 65% af tilfældene, hvis de tre batterityper har samme varians. Derfor konkluderer vi, at de har samme varians, så antagelsen om varianshomogenitet holder.

e. *Hvilken type batteri vil du vælge som det bedste? Argumentér for dit valg*

For at vælge den bedste batteritype har jeg beregnet minimum, maksimum, middelværdi og median for hver type. Valget står mellem type 1 og 3, der begge har høje driftstider, sammenlignet med type 2. Type 1 har både det højeste maksimum (11453 s mod 11446 s for type 3) og den højeste median (10451 s mod 11446 s). Men type 1 har også det laveste bundniveau sammenlignet med type 3 (8317 s mod 8968 s). Forskellene i maksimum og median er dog meget små. Type 3 har den højeste middelværdi (10311 mod 9997 for type 1). Jeg ville vælge type 3, fordi det lader til at give mere ensartet driftstid end type 1. Bartlett's testen viste godt nok at denne forskel i varians ikke var statistisk signifikant, men i mangel af bedre kriterier foretrækker jeg batteritype 3.

MatLab kode for opgave 2:

```
%% M4STI1 2018E opgave 2: Variansanalyse af forskellige batteriers driftstid
clear; close all; clc; format compact; format shortG;

%% Indlæs og behandl data
M = xlsread('Data_M4STI1_2018E.xlsx','C:D')

batteritype = M(:,1)
driftstid  = M(:,2)

%% a: Parallelt boxplot for hver batteritype
figure(1)
boxplot(driftstid, batteritype)
title('Målt driftstid af tre batterityper');
ylabel('Driftstid (sekunder)');

% Batteritype 1 og 3 har nogenlunde samme median, hvor type 2 ligger
% tydeligt lavere.
% Boxplottene for type 2 og 3 ser meget ensartede ud, bortset fra niveauet,
% hvor type 3 ligger højere end type 2. Type 1 ligger på niveau med type 3,
% men lader til at have lidt større variation. Der er dog ikke stor forskel
% på afstanden imellem kostenes ender på type 1 og 3.
% Da hvert boxplot er tegnet på baggrund af kun 5 observationer er det
% svært at konkludere, alene ud fra boxplottet, om der er reelle
% forskelle på de tre batterityper.


%% b: Variansanalyse (ANOVA)
[p,table,stats] = anovan(driftstid, batteritype)
% Variansanalysen har F = 2.9 og en tilhørende p-værdi på 0.0938. Det vil
% sige, at vi kan ikke afvise på 5 % signifikansniveau, at alle 3
% batterityper opfører sig ens (da 0.0938 > 0.05). Med andre ord:
% Bedømt ud fra variansanalysen kan alle tre batterityper have samme
% forventede driftstid.


%% c: Parvis sammenligning af batterityperne med LSD metoden
% Signifikansniveau alfa:
alfa = 0.05
[c,m] = multcompare(stats,  'Alpha',alfa,  'CType','lsd')

% Type 1 er ikke forskellig fra hverken type 2 eller 3
% Type 2 og type 3 er forskellige på 5% signifikansniveau (p-værdi 0.0417)


%% d: Holder antagelserne for residualerne?
% Det er en antagelse i variansanalysen, at residualerne er normalfordelte
% med middelværdi 0 og med samme varians for alle batterityper.
% Vi kan teste, om residualerne er normalfordelte med et
% normalfordelingsplot. Residualerne hentes i stats objektet, der er output
% fra anovan:
resid = stats.resid
```

```
figure(2)
normplot(resid)

% Residualerne ligger nogenlunde på en ret linje, så den antagelse ser ud
% til at holde. Antagelsen om varianshomogenitet kan vi teste visuelt med
% et plot af residualerne:

figure(3)
hold on;
scatter(batteritype, resid, 'b', 'filled')
plot([0; 4], [0; 0], 'r')   % vandret referencelinje gennem (0,0)
title('Residualplot for batteritypeerne');
xlabel('Batteritype');
ylabel('Residualer');
axis([0 4 -2000 2000])
hold off;

% Residualerne for type 2 og 3 er meget ensartede, men de er lidt større
% for type 1. Vi kan undersøge med en Bartlett's test, om der er
% signifikant forskel:

vartestn(driftstid, batteritype)

% Bartlett's testen viser, at vi ikke kan forkaste nulhypotesen om at
% variansen er ens for de tre typer. P-værdien på 0.65138 fortæller, at vi
% kan forvente residualer som vores, eller mere ekstreme i 65% af
% tilfældene, hvis de tre batterityper har samme varians. Derfor
% konkluderer vi, at de har samme varians, så antagelsen om
% varianshomogenitet holder.


%% e: Min foretrukne type batteri
% For at beregne minimum, maksimum, median og middelværdi for hver
% batteritype opdeler jeg data i array'et A med en kolonne for hver type:
A = [driftstid(1:5), driftstid(6:10), driftstid(11:15)]
maksimum = max(A,[],1)
medianer = median(A,1)
minimum = min(A,[],1)
middelv = mean(A,1)

% Valget står mellem type 1 og 3, der begge har høje driftstider,
% sammenlignet med type 2.
% Type 1 har både det højeste maksimum (11453 s mod 11446 s for type 3) og
% den højeste median (10451 s mod 11446 s). Men type 1 har også det laveste
% bundniveau sammenlignet med type 3 (8317 s mod 8968 s).
% Forskellene i maksimum og median er dog meget små.
% Type 3 har den højeste middelværdi (10311 mod 9997 for type 1).
% Jeg ville vælge type 3, fordi det lader til at give mere ensartet
% driftstid end type 1. Bartlett's testen viste godt nok at denne forskel
% i varians ikke var statistisk signifikant, men i mangel af bedre kriterier
% foretrækker jeg batteritype 3.
```

Opgave 3 - Kvinders og mænds foretrukne design

a. *Beregn sandsynlighederne $P(A)$, $P(A^c)$, $P(B)$, $P(B^c)$ på baggrund af data fra brugerundersøgelsen*

$P(A)$ er sandsynligheden for, at en tilfældig testperson foretrækker designet SL1. I alt 17 (6 kvinder og 11 mænd) ud af de 37 testpersoner foretrækker SL1. De udgør $17/37 = 45.95$ %.
De resterende $37 - 17 = 20$ personer foretrækker SL2, så $P(A^c) = 20/37 = 54.05$ %.
Der er 19 kvinder blandt de 37 testpersoner, så $P(B) = 19/37 = 51.35$ %. De resterende er mænd, så $P(B^c) = 1 - 0.5135 = 48.65$ %.
Opsummering:

$P(A) = \mathbf{0.4595}$
$P(A^c) = \mathbf{0.5405}$
$P(B) = \mathbf{0.5135}$
$P(B^c) = \mathbf{0.4865}$

b. *Beregn sandsynlighederne $P(A \cap B)$ og $P(A \mid B)$*

$P(A \cap B)$ er sandsynligheden for fælleshændelsen mellem A og B. Dvs. sandsynligheden for at en tilfældig testperson er både kvinde og foretrækker SL1. Der er 6 ud af de 37 testpersoner, der opfylder dette. Derfor er sandsynligheden $6/37 = 16.2$ %:

$P(A \cap B) = 6/37 = \mathbf{0.1622}$

$P(A \mid B)$ er sandsynligheden for A givet B. D.v.s. sandsynligheden for at en af de 19 kvindelige testpersoner foretrækker SL1. Der er 6 ud af de 19 kvindelige testpersoner, der foretrækker SL1, så sandsynligheden er $6/19 = 31.6$ %:

$P(A \mid B) = 6/19 = \mathbf{0.3158}$

c. *Lader A og B til at være uafhængige hændelser?*

Definitionen på, at A og B er uafhængige er, at $P(A) = P(A \mid B)$. Da $P(A) = 0.46$ og $P(A|B) = 0.32$ kunne det godt tyde på, at de to hændelser ikke er uafhængige - værdierne er ikke tæt på hinanden. Altså: 46 % af testpersonerne foretrækker SL1, men blandt kvinderne er det kun 32 %. Der lader til at være kønsforskelle. Tilsvarende kan man beregne, at $P(B) = 0.51$, men $P(B \mid A) = 0.35$.

d. *Beskriv Venn diagrammets hvide område med ord og dets grå område med symboler*

Det hvide område i Venn diagrammet er indenfor 'boblen' A, så det angiver testpersoner, der foretrækker SL1. Samtidig er det hvide område udenfor 'boblen' B, så testpersonerne i det hvide område er ikke kvinder. Med andre ord: Det hvide område angiver de mænd, der foretrækker SL1. Dem er der 11 af.

Hændelserne beskrevet i mulighed **2** ($(A \cap B) \cup A^c$) og **3** ($B \cup A^c$) svarer begge til det grå område i Venn diagrammet. Det er dem, der enten er kvinder eller foretrækker SL2.

e. *Beregn hvor mange af de 37 testpersoner, man vil forvente for hver kombination af prototype-design og køn, hvis man antager, at de to faktorer (prototypedesign og køn) er uafhængige*

Generelt gælder formlen $P(A \cap B) = P(A\,|B) \cdot P(B)$ for to hændelser A og B, men hvis hændelserne er uafhængige, så gælder $P(A \cap B) = P(A) \cdot P(B)$. Under forudsætning af at A og B er uafhængige kan vi således beregne sandsynligheden for, at en tilfældig testperson både er kvinde og foretrækker SL1 til:

$$P(A \cap B) = P(A) \cdot P(B) = 0.4595 \cdot 0.5135 = 0.2360$$

Altså: Når 46.0 % af testpersonerne foretrækker SL1, og 51.4 % af testpersonerne er kvinder, så vil vi forvente at 23.6 % af testpersonerne er kvinder, der foretrækker SL1, hvis de to ting er uafhængige. Når der er 37 testpersoner, vil vi forvente $37 \cdot 0.2360 = 8.7297$ testpersoner, der er kvinder og foretrækker SL1 (stadig under forudsætning af at A og B er uafhængige hændelser).

Tilsvarende kan vi beregne det forventede antal testpersoner, der opfylder at de er kvinder og foretrækker SL2, mænd der fortrækker SL1 og mænd der foretrækker SL2:

```
E =
      8.7297        8.2703
       10.27        9.7297
```

f. *Opstil nulhypotese og alternativhypotese for hypotesetesten*

Vi skal lave en hypotesetest af typen kontingenstabel for at afgøre, om A og B er uafhængige hændelser. Hypoteserne opstilles:

H_0: Præference for prototypedesign og køn er uafhængige

H_a: Præference for prototypedesign og køn er *ikke* uafhængige

g. *Opstil en formel for teststørrelsen og angiv, hvilken fordeling den følger*

Teststørrelsen for denne type Chi-i-anden test er en sum af bidrag fra hver celle i matricen O. Hvert bidrag sammenligner det observerede antal (en celle i matricen O) med det forventede antal (den tilsvarende celle i matricen E), hvis nulhypotesen er sand. Teststørrelsen ${\chi^2}_0$ beregnes således:

$${\chi^2}_0 = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{\left(O_{ij} - E_{ij}\right)^2}{E_{ij}}$$

${\chi^2}_0$ er chi-i-anden fordelt med $(r-1)(c-1)$ frihedsgrader. Her er $r = c = 2$, da begge faktorer har to udfald (Designpræference: SL1 og SL2; køn: kvinde og mand). Med andre ord har vi en 2x2 kontingenstabel. Derfor er der 1 frihedsgrad.

h. *Beregn den kritiske region for testen, beregn teststørrelsens værdi og konkludér på hypotesetesten*

Den kritiske region beregnes med den inverse funktion til den kumulerede fordelingsfunktion for chi-i-anden fordelingen, dvs. MatLab funktionen chi2inv():

$$\chi^2_{\text{kritisk}} = \text{chi2inv}(1-\alpha, \text{df}) = \text{chi2inv}(0.95, 1) = \mathbf{3.8415}$$

Vi forkaster altså nulhypotesen, hvis teststørrelsen er over 3.8415.

Teststørrelsens værdi:

$${\chi^2}_0 = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{\left(O_{ij} - E_{ij}\right)^2}{E_{ij}} = \mathbf{3.2459}$$

Konklusion:
Teststørrelsen ${\chi^2}_0 = 3.2459$ er ikke større end den kritiske værdi, $\chi^2_{\text{kritisk}} = 3.8415$. Derfor kan vi ikke forkaste nulhypotesen på 5 % signifikansniveau. P-værdien er 0.07, så de forskelle i præferencer, vi har observeret mellem kønnene, er ikke tilstrækkeligt store til, at vi kan udelukke, at de er tilfældige. Vi må gå ud fra, at designpræference og køn er uafhængige.

MatLab kode for opgave 3:

```
%% Eksamen M4STI1 2018E opgave 3: Kønsspecifikke præferencer af designs
clear; close all; clc; format compact;

%% Data fra brugerundersøgelsen

O = xlsread('Data_M4STI1_2018E.xlsx','G:H')

% Hændelse A: Testpersonen foretrækker model SL1
% Hændelse B: Testpersonen er en kvinde

%% a. Sandsynligheder
% P(A) er sandsynligheden for at en tilfældig testperson foretrækker SL1.
% I alt 17 (6 kvinder og 11 mænd) ud af de 37 testpersoner foretrækker SL1.
% De udgør 17/37 = 45.95 %.
% I det følgende beregner jeg samlet antal af hvert køn og præference:
soejlesum = sum(O,1) % Sum for hver søjle (antal kvinder og mænd)
raekkesum = sum(O,2) % Sum for hver række (antal der foretrækker SL1 og SL2)
total = sum(sum(O))  % Total antal testpersoner

P_A = raekkesum(1)/total    % P_A  = 0.4595
P_Ac = raekkesum(2)/total   % P_Ac = 0.5405
P_B = soejlesum(1)/total    % P_B  = 0.5135
P_Bc = soejlesum(2)/total   % P_Bc = 0.4865


%% b. Sandsynligheder
% P(A n B) er sandsynligheden for fælleshændelsen mellem A og B. Dvs.
% sandsynligheden for at en tilfældig testperson er både kvinde og
% foretrækker SL1. Der er 6 ud af de 37 testpersoner, der opfylder dette.
% Derfor er sandsynligheden 6/37 = 16.2 %:
P_A_faelles_B = O(1,1)/total        % P_A_faelles_B = 0.1622

% P(A | B) er sandsynligheden for A givet B. D.v.s. sandsynligheden for at
% en af de 19 kvindelige testpersoner foretrækker SL1. Der er 6 ud af de 19
% kvindelige testpersoner, der foretrækker SL1, så sandsynligheden er
% 6/19 =  31.6 %:
P_A_givet_B = O(1,1)/soejlesum(1)   % P_A_givet_B = 0.3158


%% c.
% Definitionen på at A og B er uafhængige er, at P(A) = P(A|B). Da
% P(A) = 0.46 og P(A|B) = 0.32 kunne det godt tyde på, at de to hændelser
% ikke er uafhængige. Altså: 46 % af testdeltagerne foretrækker SL1, men
% blandt kvinderne er det kun 32 %. Der lader til at være kønsforskelle.
% Tilsvarende kan man beregne, at P(B) = 0.51, men P(B|A) = 0.35.


%% d.
% Det hvide område i Venn diagrammet angiver de mænd, der foretrækker SL1.
% Dem er der 11 af.
%
% Hændelserne beskrevet i mulighed 2 og 3 svarer begge til det grå område i
% Venn diagrammet. Det er dem, der enten er kvinder eller foretrækker SL2.


%% e. Beregn tabel over forventet antal, hvis der er uafhængighed, E
soejlefrekvens = soejlesum/total    % Frekvens for hver søjle
raekkefrekvens = raekkesum/total    % Frekvens for hver række
```

```
r = size(O,1)                    % Antal rækker
c = size(O,2)                    % Antal søjler
E = zeros(r,c);                  % Matricen E kommer til at indeholde
                                 % forventet antal. I første omgang
                                 % indeholder den 0'er i alle celler


% For hver celle i matricen E beregnes det forventede antal som det totale
% antal gange sandsynligheden for at være i den i-te række og j-te søjle.
% Denne sandsynlighed kan beregnes som sandsynligheden for at være i den
% i-te række gange sandsynligheden for at være i den j-te søjle, da de to
% ting antages at være uafhængige.
for i=1:r
    for j=1:c
        E(i,j) = total*raekkefrekvens(i)*soejlefrekvens(j);
    end
end
E


%% f. Opstil nul- og alternativ-hypoteser
% H0: Præference for prototypedesign og køn er uafhængige
% Ha: De er ikke uafhængige


%% g. Formel for teststørrelsen og dens fordeling
% chi2_0 = sum_i(sum_j( (O_ij - E_ij)^2/E_ij ))
% hvor O_ij er det observerede antal testpersoner, der foretrækker design i
% (SL1, SL2) og som har køn j (kvinde, mand), og hvor E_ij er det tilsvarende
% forventede antal under forudsætning af uafhængighed.
% Teststørrelsen chi2_0 er chi-i-anden fordelt med (r-1)*(c-1)
% frihedsgrader.


%% h. Beregn kritisk værdi, teststatistikken og konklusion på testen
% For at beregne kritisk værdi skal jeg bruge antal frihedsgrader
% Beregning af frihedsgrader df
df = (r - 1)*(c - 1)    % df = 1
% signifikansniveau 5%
alfa = 0.05;
chi2_kritisk = chi2inv(1-alfa,df)   % chi2_kritisk = 3.8415
% Vi forkaster H0, hvis chi2_0 > chi2_kritisk

% Beregning af teststatistikken chi2_0.
% Vi har de observerede antal O_ij i O og de forventede antal E_ij i E
chi2_0 = 0;
for i=1:r
    for j=1:c
        chi2_0 = chi2_0 + (O(i,j) - E(i,j))^2 / E(i,j);
    end
end
chi2_0  % chi2_0 = 3.2459

p_value = 1 - chi2cdf(chi2_0, df)   % p_value = 0.071601

% Konklusion:
% Teststørrelsen chi2_0 = 3.2459 er ikke større end den kritiske værdi,
% chi2_kritisk = 3.8415, så vi kan ikke forkaste nulhypotesen på
% 5 % signifikansniveau. P-værdien er 0.07, så de forskelle i præferencer,
% vi har observeret mellem kønnene er ikke tilstrækkeligt store til, at vi
```

```
% kan udelukke, at de er tilfældige.


%% Alternativ løsning af hypotesetest (til kontrol af løsning)
% Funktionen chi2cont er ikke i MatLab men kan downloades fra BB
[h, pval, chi2_0] = chi2cont(O, alfa)
% h angiver om nulhypotesen kan forkastes (0: nej, 1: ja)
% pval er p-værdien for chi-i-anden testen
% chi2_0 er teststørrelsen
```