Løsningsforslag M4STI1 2018F

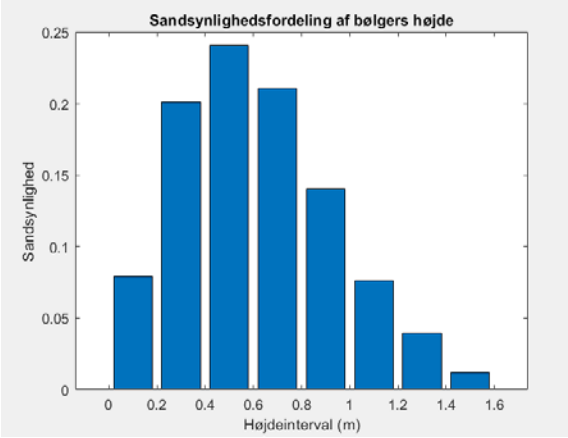
Opgave 1 – Bølgers variabilitet

a. *Omregn antal bølger i hvert interval til sandsynligheden for at få en bølge i intervallet. Lav et diagram, der viser sandsynlighedsfordelingen af bølgerne.*

Sandsynligheden for en bølge i et givet interval estimeres som antal observerede bølger i intervallet divideret med det totale antal. I tabellen nedenfor vises grænserne for hvert interval i de første to kolonner og i den tredje kolonne vises den beregnede sandsynlighed for en bølge i intervallet:

0	0.2000	0.0793
0.2000	0.4000	0.2012
0.4000	0.6000	0.2409
0.6000	0.8000	0.2104
0.8000	1.0000	0.1402
1.0000	1.2000	0.0762
1.2000	1.4000	0.0396
1.4000	1.6000	0.0122

Nedenfor vises sandsynlighedsfordelingen med et histogram:

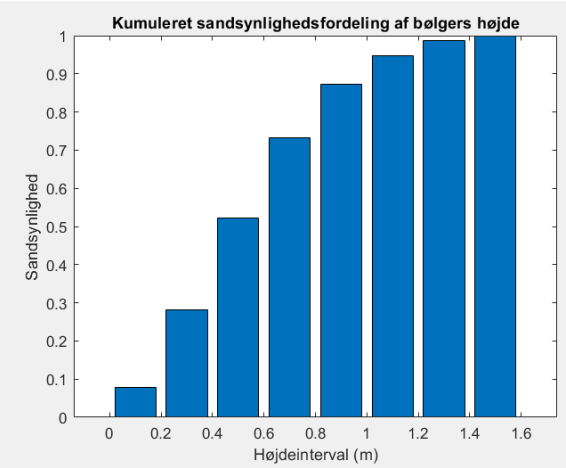


b. *Beregn den kumulerede sandsynlighedsfordeling for bølgehøjden.*

I tabellen nedenfor vises den kumulerede sandsynlighed i den tredje kolonne. Som før viser de første to kolonner intervalgrænserne:

0	0.2000	0.0793
0.2000	0.4000	0.2805
0.4000	0.6000	0.5213
0.6000	0.8000	0.7317
0.8000	1.0000	0.8720
1.0000	1.2000	0.9482
1.2000	1.4000	0.9878
1.4000	1.6000	1.0000

Den kumulerede sandsynlighedsfordeling kan vises i et diagram (selv om der ikke er blevet bedt om det):



c. *Hvad er sandsynligheden for at den næste bølge er højere end 1.0 m?*

Der er flere måder at beregne dette på.
Ud fra observationerne: Der blev optalt 25 + 13 + 4 = 42 bølger over 1 meter ud af de 328 bølger, så sandsynligheden er 42/328 = 0.128.
Ud fra sandsynlighedsfordelingen: Summen af sandsynligheder i intervallerne over 1.0 m er 0.0762 + 0.0396 + 0.0122 = 0.128.
Ud fra den kumulerede sandsynlighedsfordelingen: P(H > 1.0) = 1 - P(H <= 1.0) = 1 - 0.872 = **0.128**.

d. *Beregn den gennemsnitlige bølgehøjde (antag at alle bølgerne i et interval har samme højde, nemlig intervallets midterværdi).*

Den gennemsnitlige bølgehøjde beregnes ved at summere for hvert interval sandsynligheden for en bølge i intervallet gange bølgehøjden (intervallets midterværdi). D.v.s.:
H_middel = 0.0793*0.10 + 0.2012*0.30 + … + 0.0122*1.50 = **0.6159**.

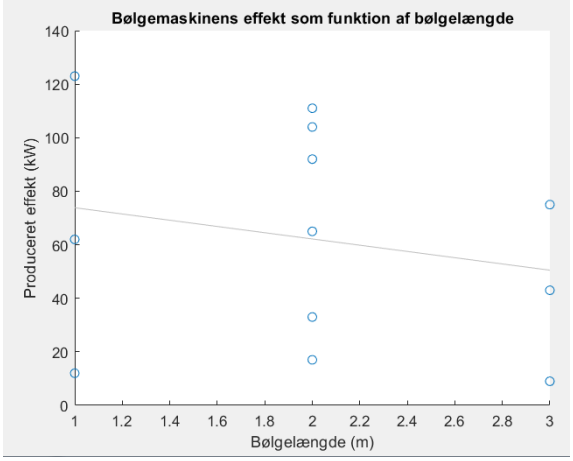
MatLab kode til opgave 1

```
%% M4STI1 2018F Opgave 1: Bølgers variabilitet
clc; clear; close all; format compact;

%% Indlæs og behandl data
D = xlsread('Data_M4STI1_2018F.xlsx','A:C')

H_fra = D(:,1)      % Startværdi for interval af bølgehøjde
H_til = D(:,2)      % Slutværdi for interval af bølgehøjde
```

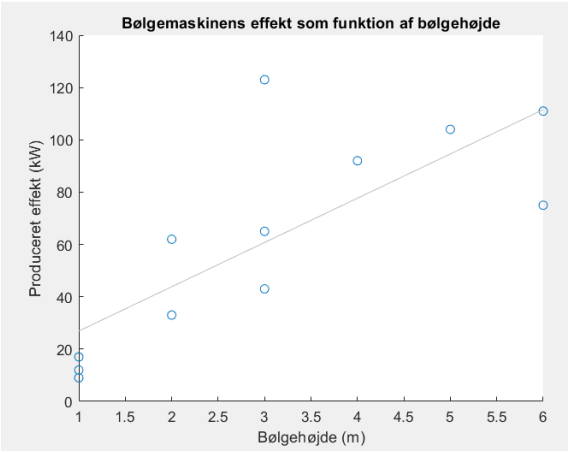
```
O = D(:,3)          % Observeret antal bølger i intervallet
n = sum(O)          % Antal bølger ialt
k = size(O,1)       % Antal intervaller

%% a: Sandsynlighedsfordeling for bølgehøjder
P_hoejde = O/n              % Sandsynlighedsfordelingen af bølgehøjder
test = sum(P_hoejde)        % Summen skal være 1

res = [H_fra, H_til, P_hoejde]

H_midt = (H_fra + H_til)/2  % Midterværdier for intervallerne

figure(1)
bar(H_midt, P_hoejde)
title('Sandsynlighedsfordeling af bølgers højde')
xlabel('Højdeinterval (m)')
ylabel('Sandsynlighed')


%% b: Kumuleret sandsynlighedsfordeling
P_kumul = zeros(k,1)                    % Initialisering med nuller
for i=1:k
    P_kumul(i) = sum(P_hoejde(1:i));    % Det i-te element i P_kumul er sum
end                                     % af de hidtidige sandsynligheder
P_kumul

res = [H_fra, H_til, P_kumul]

% Der bedes ikke om en grafisk præsentation, men her er den:
figure(2)
bar(H_midt, P_kumul)
title('Kumuleret sandsynlighedsfordeling af bølgers højde')
xlabel('Højdeinterval (m)')
ylabel('Sandsynlighed')


%% c: Sandsynligheden for at en bølge er over 1.0 m
% Dette findes som summen af sandsynligheder fra det sjette interval
% (1.0 - 1.2) til det sidste interval:
P_o1 = sum(P_hoejde(6:8,1))     % P_o1 = 0.1280
P_o1_alt = 1 - P_kumul(5,1)     % Alternativ beregning med den kumulerede
                                % sandsynlighedsfordeling


%% d: Bølgernes gennemsnitshøjde
% Jeg har allerede beregnet intervallernes midterværdi:
% H_midt = (H_fra + H_til)/2
H_middel = (O'*H_midt)/n
% Bølgernes gennemsnitshøjde beregnes til H_middel = 0.6159 m = 0.62 m
```

Opgave 2 – Sammenhæng mellem bølgerne og bølgemaskinens effekt

a. *Lav et plot for hver af de uafhængige variable (hhv. bølgelængde og bølgehøjde), der viser om der er korrelation mellem variablen og den producerede effekt. Diskutér dine plots kort.*



Figuren viser, at der er faldende effekt med stigende bølgelængde, altså en negativ korrelation. Det er forventeligt, for når bølgelængden er høj, kommer der færre bølger per tidsenhed, som bølgemaskinen kan udnyttes energi af. Plottet viser stor variation i data som følge af, at bølgehøjden tilsyneladende også har stor indvirkning på den producerede effekt.



Der er en mere tydelig positiv korrelation mellem bølgehøjde og effekt. Det er ikke overraskende, at bølgemaskinen kan producere højere effekt, desto højere bølgerne er. Der lader til at være en stærkere virkning af bølgehøjde end af bølgelængde, og der er mindre variation i data for dette plot.

b. *Lav en multipel lineær regressionsanalyse, der beskriver produceret effekt som funktion af bølgelængde og bølgehøjde. Skriv regressionsligningen op.*

Resultatet af regressionsanalysen vises her:

```
Estimated Coefficients:
                   Estimate       SE        tStat       pValue
                   ________    _______    ________    __________

    (Intercept)      51.19      18.113      2.8261      0.019848
    Laengde         -24.737     8.3794     -2.9522      0.016164
    Hoejde           19.606     3.3781      5.8038    0.00025826

Number of observations: 12, Error degrees of freedom: 9
Root Mean Squared Error: 19.8
R-squared: 0.799,  Adjusted R-Squared 0.754
F-statistic vs. constant model: 17.9, p-value = 0.000732
```

Dermed er regressionsligningen:

Effekt = 51.19 - 24.737*Laengde + 19.606*Hoejde

c. *Forklar v.h.a. regressionsanalysens statistikker og dine plots fra delopgave a), om modellen beskriver observationerne godt.*

Det er en ganske god model. Alle koefficienterne er signifikant forskellige fra 0 på 95 % signifikansniveau, da p-værdierne er tæt på 0 (under 0.05). I ANOVA testen er F=17.9, som giver en p-værdi på 0.000732. Det er altså meget usandsynligt, at hverken bølgelængde eller højde har en effekt på den producerede effekt. Desuden er R-squared 0.799 og Adjusted R-squared er 0.754, så modellen forklarer en pæn del af variationen. Men der er plads til forbedringer.

d. *Undersøg om der er 'unormale' datapunkter, d.v.s. outliers, løftestangs- eller indflydelses-punkter.*

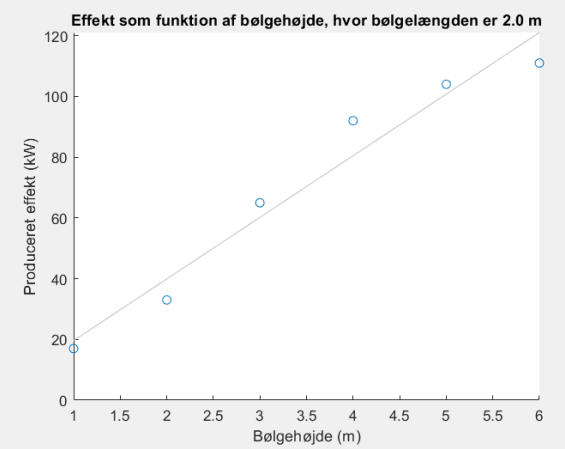
Resultatet af undersøgelsen vises i følgende tabel, hvor kolonnerne er hhv. observationsnummer, bølgelængde, bølgehøjde, effekt, leverage og studentiseret residual.
Studentiseret residual får vi fra MatLab som mdl.Residuals.Studentized, og den numeriske værdi skal være under 3.
Leverage er hat-diagonalen, som vi får fra MatLab som mdl.Diagnostics.Leverage. Her skal værdien være under lev_limit = 2*(k+1)/n = 0.5 (k = 2 er antal regressorer, og n = 12 er antal observationer).

1.0000	1.0000	1.0000	12.0000	0.3086	-2.7008
2.0000	1.0000	2.0000	62.0000	0.2551	-0.2030
3.0000	1.0000	3.0000	123.0000	0.2599	3.1069
4.0000	2.0000	1.0000	17.0000	0.2101	-0.2326
5.0000	2.0000	2.0000	33.0000	0.1176	-0.4066
6.0000	2.0000	3.0000	65.0000	0.0835	0.2232
7.0000	2.0000	4.0000	92.0000	0.1079	0.6128
8.0000	2.0000	5.0000	104.0000	0.1906	0.2263
9.0000	2.0000	6.0000	111.0000	0.3317	-0.4945
10.0000	3.0000	1.0000	9.0000	0.4708	0.8499
11.0000	3.0000	3.0000	43.0000	0.2664	0.4052
12.0000	3.0000	6.0000	75.0000	0.3978	-1.3323

Der er ingen løftestangspunkter, da alle værdier for leverage er under lev_limit på 0.5. Dog er punkt nr. 10 med Laengde = 3.0 og Hoejde = 1.0 tæt på, da leverage = 0.4708. Der er en enkelt outlier, for punkt

nr. 3 med Laengde = 1.0 og Hoejde = 3.0 har |rst| = 3.1069, som er over grænsen på 3. Der er ingen indflydelsespunkter, for det ville kræve, at samme punkt er både løftestangspunkt og outlier.

e. *Lader der til at være en lineær sammenhæng mellem bølgehøjde og effekt, når bølgelængden er fast* $L = 2.0\ m$*?*



Bedømt ud fra plottet ovenfor lader sammenhængen mellem bølgehøjde og effekt nærmest til at være S-formet. Det tyder på, at der skal en vis bølgehøjde til for at bølgemaskinen kan udnytte energien, og der lader til at være et niveau, hvor bølgemaskinen ikke kan hente mere energi ud, selv om bølgerne bliver større. En simpel lineær regression viser dog, at en lineær model er en god beskrivelse af data:

```
Estimated Coefficients:
                   Estimate      SE        tStat       pValue
    (Intercept)    -0.66667    8.3514    -0.079827     0.94021
    x1               20.286    2.1444       9.4597    0.00069657

Number of observations: 6, Error degrees of freedom: 4
Root Mean Squared Error: 8.97
R-squared: 0.957,  Adjusted R-Squared 0.947
F-statistic vs. constant model: 89.5, p-value = 0.000697
```

Estimatet for hældingskoefficienten b1=20.286 er signifikant (pValue = 0.00069657). R-Squared og Adjusted R-Squared er over 0.9 (hhv. 0.957 og 0.947), så modellen forklarer en stor andel af variationen. Ikke desto mindre kunne en S-formet model vise sig at være bedre.

f. *Lav en transformation af bølgehøjden med funktion:* $H_l = 1/(1 + \exp(k - H))$
hvor H *er bølgehøjden og* k *er en konstant, der er karakteristisk for bølgemaskinen. For den pågældende bølgemaskine har man estimeret, at* $k = 2.5\ m$*. Lav en simpel lineær regression med* H_l *som regressorvariabel og* E *som responsvariabel.*

```
Estimated Coefficients:
                   Estimate      SE       tStat       pValue
                   ________    ______    _______    __________

    (Intercept)    -9.1154     3.1056    -2.9351      0.042597
    x1              122.39     4.3717     27.996     9.6844e-06

Number of observations: 6, Error degrees of freedom: 4
Root Mean Squared Error: 3.09
R-squared: 0.995,  Adjusted R-Squared 0.994
F-statistic vs. constant model: 784, p-value = 9.68e-06
```

Regressionsanalysen viser, at den lineære model mellem H_l og E er virkelig god.
Hældningskoefficienten er signifikant og næsten al variation forklares af modellen (R-squared = 0.995).

g. *Brug den lineære regression fra delopgave f) til at skrive et udtryk for E som funktion af H. Beregn den forventede effekt med en bølgehøjde på 4.7 m?*

Først skrives ligningen for E som funktion af H_l med koefficienterne fra delopgave f):

Effekt = -9.1154 + 122.39*Hoejde_l

Så indsættes udtrykket for H_l (Hoejde_l):

Effekt = -9.1154 + 122.39*1/(1 + exp(k - Hoejde))

Nu kan effekten, der svarer til en bølgehøjde på 4.7 m beregnes ved at indsætte Hoejde = 4.7 i ligningen:

Effekt_0 = -9.1154 + 122.39*1/(1+exp(k - 4.7)) = 101.0661

Den forventede effekt af bølger på 4.7 m er 101.0661 kW = **101 kW.**

MatLab kode til opgave 2

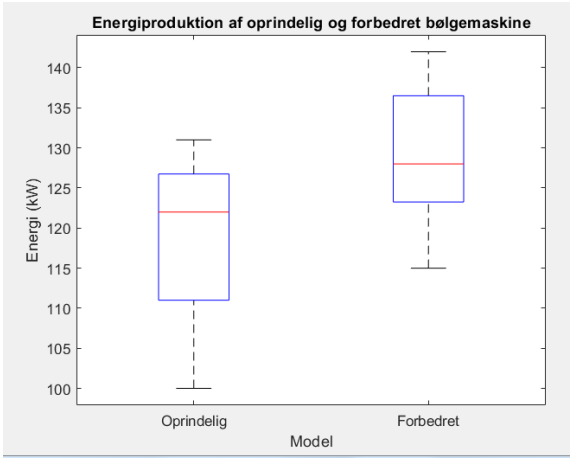
```
%% M4STI1 2018F Opgave 2: Bølgemaskinens effekt afhængig af bølgerne
clc; clear; close all; format compact;

%% Indlæs og behandl data
D = xlsread('Data_M4STI1_2018F.xlsx','E:G')

Laengde = D(:,1)    % Bølgelængde
Hoejde = D(:,2)     % Bølgehøjde
Effekt = D(:,3)     % Produceret effekt
n = size(D,1)       % Antal observationer i datasættet


%% a: Plots for de uafhængige variable
figure(1)
scatter(Laengde,Effekt)
lsline                      % Regressionslinje
title('Bølgemaskinens effekt som funktion af bølgelængde')
xlabel('Bølgelængde (m)')
ylabel('Produceret effekt (kW)')
% Figur 1 viser, at med stigende bølgelængde er der faldende effekt, altså
% en negativ korrelation. Det er forventeligt, for når bølgelængden er høj,
% kommer der færre bølger per tidsenhed, der kan udnyttes energi af.
% Der er stor variation i data som følge af, at bølgehøjden tilsyneladende
% også har stor indvirkning på den producerede effekt.
```

```
figure(2)
scatter(Hoejde, Effekt)
lsline                      % Regressionslinje
title('Bølgemaskinens effekt som funktion af bølgehøjde')
xlabel('Bølgehøjde (m)')
ylabel('Produceret effekt (kW)')
% Der er en positiv korrelation, så bølgemaskinen kan producere højere
% effekt, desto højere bølgerne er. Der lader til at være en stærkere
% virkning af bølgehøjde end af længde, og der er mindre variation i data
% for dette plot.


%% b: Multipel lineær regression
mdl = fitlm([Laengde Hoejde], Effekt, ...
    'ResponseVar', 'Effekt', ...
    'PredictorVars', {'Laengde', 'Hoejde'})

% Regressionsligning:
%   Effekt = 51.19 - 24.737*Laengde + 19.606*Hoejde


%% c: Fortolkning af regressionsanalysen
% Det er en ganske god model. Alle koefficienterne er signifikant
% forskellige fra 0 på 95 % signifikansniveau, da p-værdierne er tæt på 0
% (under 0.05). I ANOVA testen er F=17.9, som giver en p-værdi på 0.000732.
% Det er altså meget usandsynlig at hverken bølgelængde eller højde har en
% effekt på den producerede effekt. Desuden er R-squared 0.799
% og Adjusted R-squared er 0.754, så modellen forklarer en pæn del af
% variationen. Men der er plads til forbedringer.


%% d: Unormale punkter
lev = mdl.Diagnostics.Leverage;     % hat diagonal
rst = mdl.Residuals.Studentized;    % R-Student
nr = (1:n)';
resultat = [nr, Laengde, Hoejde, Effekt, lev, rst]            % Jeg samler det hele til en
resultattabel

k = 2;                  % Der er to regressorer
lev_limit = 2*(k+1)/n   % lev_limit = 0.5
% Der er ingen løftestangspunkter, da alle værdier for lev er under
% lev_limit på 0.5. Dog er punkt nr. 10 med Laengde = 3.0 og Hoejde = 1.0
% tæt på, da lev = 0.4708.
% Der er en enkelt outlier, for punkt nr. 3 med Laengde = 1.0 og
% Hoejde = 3.0 har |rst| = 3.1069, som er over grænsen på 3.
% Der er ingen indflydelsespunkter, for det kræver at samme punkt er både
% løftestangspunkt og outlier.


%% e: Er sammenhængen lineær for fast bølgelængde L = 2.0?
% Datasættet reduceres til de 6 observationer med Laengde = 2.0:
Hoejde = Hoejde(4:9,:)
Effekt = Effekt(4:9,:)

figure(3)
scatter(Hoejde, Effekt)
lsline
title('Effekt som funktion af bølgehøjde, hvor bølgelængden er 2.0 m')
xlabel('Bølgehøjde (m)')
ylabel('Produceret effekt (kW)')
% Plottet viser en sammenhæng, der måske bedre kan beskrives som S-formet
% end som lineær. Der lader til at skulle en vis bølgehøjde til for at
% danne energi, og når bølgelængden bliver tilstrækkelig stor klinger den
% producerede effekt af.

% Jeg tester med en simpel lineær regression:
mdl2 = fitlm(Hoejde, Effekt)
```

```
% Det er en god model, hvor hældingskoefficienten b1=20.286 er signifikant
% (pValue = 0.00069657). R-Squared og Adjusted R-Squared er over 0.9 (hhv.
% 0.957 og 0.947), så modellen forklarer en stor andel af variationen.
% Men på plottet ser punkterne ud til at følge et S-formet forløb.


%% f: Transformation af bølgehøjden
k = 2.5
Hoejde_l = 1./(1 + exp(k - Hoejde))
mdl3 = fitlm(Hoejde_l, Effekt)
plot(mdl3)


%% g: Effekt som funktion af bølgehøjde
% Regressionsligning:
% Effekt = -9.1154 + 122.39*Hoejde_l
% Effekt = -9.1154 + 122.39*1/(1 + exp(k - Hoejde))
Hoejde_0 = 4.7
Effekt_0 = -9.1154 + 122.39*1/(1+exp(k - Hoejde_0))
% Den forventede effekt af bølger på 4.7 m er 101.0661 kW = 101 kW
```

Opgave 3 – Forbedring af en bølgemaskine

a. *Lav og kommenter et parallelt boksplot, der viser energiproduktionen for de to skalamodeller.*



Boksplottet viser at energiproduktionen for den forbedrede model generelt ligger højere. Både median, det interkvartile range (blå kasse) og koste ligger højere for den forbedrede model. De to bokse ser ensartede ud, så variationen lader til at være af samme størrelsesorden. Der ser ikke ud til at være outliers.

b. *Man ønsker at slå fast med et signifikansniveau på 5 %, om energiproduktionen er højere med den modificerede model end med den oprindelige. Opstil nulhypotese og alternativhypotese for denne hypotesetest.*

Vi ønsker at afgøre med hypotesetesten, om populationsmiddelværdien for den forbedrede model er signifikant højere end for den oprindelige model. Hvis vi kalder populationsmiddelværdierne for hhv. mu_o og mu_f, så er nul- og alternativhypotese:

H0: mu_o - mu_f = 0

Ha: mu_o - mu_f < 0

c. *Opstil formlen for teststatistikken og beregn dens værdi. Angiv hvilken fordeling den følger.*

Vi laver en t-test for to uafhængige stikprøver med ukendt populationsvarians. Teststørrelsen t0 har følgende formel:

t0 = (y_o_streg - y_f_streg)/(s_pooled*sqrt(1/n_o + 1/n_f))

hvor y_o_streg og n_o er stikprøvemiddelværdi og stikprøvestørrelse for den oprindelige model, y_f_streg og n_f er stikprøvemiddelværdi og stikprøvestørrelse for den forbedrede model, og hvor s_pooled er den puljede standardafvigelse. Til at beregne s_pooled skal vi bruge stikprøve-standardafvigelserne s_o og s_f.

n_o = 11
n_f = 13
y_o_streg = 119.3636
y_f_streg = 129.4615
s_o = 9.7598
s_f = 8.5500
s_pooled = sqrt(((n_o - 1)*s_o^2 + (n_f - 1)*s_f^2)/(n_o + n_f - 2)) = 9.1198
t0 = (y_o_streg - y_f_streg)/(s_pooled*sqrt(1/n_o + 1/n_f)) = **-2.7028**

Teststørrelsen er t-fordelt med df = n_o + n_f - 2 = 22 frihedsgrader.

d. *Beregn den kritiske region for testen og konkludér på hypotesetesten.*

Sådan som jeg har formuleret hypoteserne har jeg en ensidig test nedadtil, hvor vi forkaster nulhypotesen, hvis t0 < t_alfa, hvor t_alfa er den kritiske værdi, beregnet som
t_alfa = tinv(alfa,df)
Da alfa = 0.05 og df = 22 fås t_alfa = **-1.7171**.

Vi har altså t0 = -2.7028 og t_alfa = -1.7171. Da teststørrelsen t0 er mindre end den kritiske værdi t_alfa, forkaster vi nulhypotesen. Det er altså lykkedes de studerende at få den nye model til at producere mere energi end den oprindelige.

e. *Beregn et 95 % konfidensinterval for forskellen på middelværdi af modellernes energiproduktion.*

Konfidensintervallet bestemmes som forskellen af stikprøvenes middelværdier plus/minus B, hvor B beregnes som følger:
B = tinv(1-alfa/2,df)*s_pooled*sqrt(1/n_o + 1/n_f) = 7.7483
Dermed kan konfidensintervallets grænser beregnes:
KI_lav = (y_o_streg - y_f_streg) - B
KI_hoej = (y_o_streg - y_f_streg) + B

Forskellen i middel energiproduktion er med 95% sikkerhed indenfor konfidensintervallet [-17.8462; -2.3496]. Det, at 0 ikke tilhører konfidensintervallet, understreger resultatet af hypotesetesten: Forskellen i modellernes middel energiproduktion er mindre end 0, så den forbedrede model er bedre.

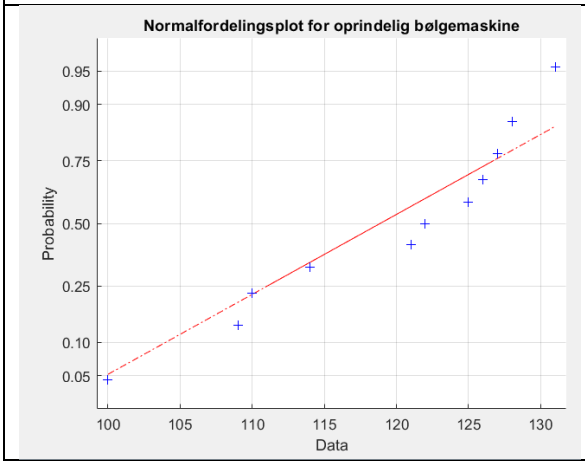
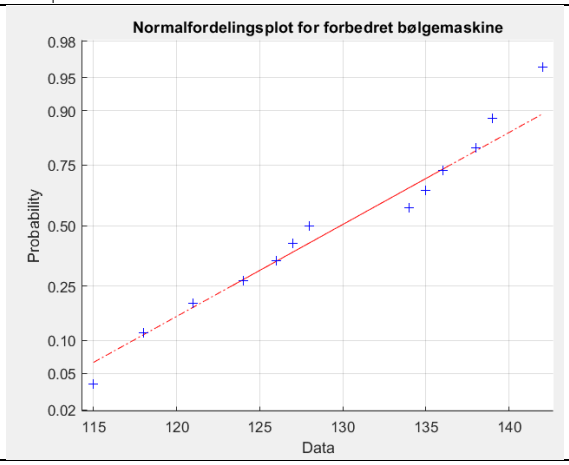
f. *Diskutter hvordan boksplot, hypotesetest og konfidensinterval stemmer overens.*

Boksplot, hypotesetest og konfidensinterval giver det samme billede: Som nævnt under delopgave a) viser boksplottet, at der generelt er højere energiproduktion for den forbedrede bølgemaskine-model. Medianen ligger højere. Hypotesetesten i delopgave b) til d) viser, at der er signifikant forskel på

middelværdierne, så den forbedrede model faktisk er en forbedring. 95% konfidenstintervallet for forskellen af middelværdier i delopgave e) indeholder ikke 0, så det understøtter, at middelværdien er størst for den forbedrede model.

g. *Oplys hvilke antagelser, der er gjort i hypotesetesten, og om antagelserne er rimelige på baggrund af data.*

Vi har antaget den centrale grænseværdisætning for hver stikprøve. Desuden har vi antaget, at de to stikprøver har samme varians. Vi kan teste den centrale grænseværdisætning med stem-and-leaf plots og med normalfordelingsplots:

Oprindelig model	**Forbedret model**
10 | 0 9 11 | 0 4 12 | 1 2 5 6 7 8 13 | 1 key: 36|5 = 365 stem unit: 10 leaf unit: 1	11 | 5 8 12 | 1 4 6 7 8 13 | 4 5 6 8 9 14 | 2 key: 36|5 = 365 stem unit: 10 leaf unit: 1
	

Begge stem-and-leaf plots viser 'pæne' fordelinger med et enkelt toppunkt, nogenlunde symmetrisk fordeling og hurtigt uddøende haler. Normalfordelingsplots viser også, at fordelingerne for stikprøverne ligner normalfordelingen. Dog er data for den oprindelige bølgemaskine ikke helt så overbevisende som for den forbedrede. De forholdsvis pæne fordelinger kombineret med rimeligt store stikprøvestørrelser gør, at vi kan være trygge ved, at antagelsen om den centrale grænseværdisætning holder.

Antagelsen om at de to stikprøver har samme varians underbygges af det parallelle boksplot fra delopgave a), der viser to ensartede bokse. Vi har desuden beregnet stikprøve-standardafvigelserne til værdier i samme størrelsesorden, nemlig hhv. 9.76 og 8.55.

MatLab kode til opgave 3

```
%% M4STI1 2018F Opgave 3: Forbedring af en bølgemaskine
clc; clear; close all; format compact;

%% Indlæs og behandl data
D = xlsread('Data_M4STI1_2018F.xlsx','I:J')

% Første kolonne af D indeholder information om, om målingen er fra den
% oprindelige model (1) eller den forbedrede model (2)
% Anden kolonne indeholder målinger af energiproduktionen

n_o = 11  % Antal målinger på oprindelig model
n_f = 13  % Antal målinger på forbedret model

G = D(:,1)  % Gruppe (oprindelig eller forbedret model)
E = D(:,2)  % Energiproduktion

E_o = E(1:n_o)          % Energiproduktion målt på oprindelig model (o)
E_f = E(n_o+1:n_o+n_f)  % Energiproduktion målt på forbedret model (f)



%% a: Boksplot
figure(1)
boxplot(E, G, 'labels', {'Oprindelig', 'Forbedret'})
title('Energiproduktion af oprindelig og forbedret bølgemaskine');
xlabel('Model');
ylabel('Energi (kW)');

% Boksplottet viser at energiproduktionen for den forbedrede model generelt
% ligger højere. Både median, interkvartil range og koste ligger højere for
% den forbedrede model. De to bokse ser ensartede ud, så variationen lader
% til at være af samme størrelsesorden. Der ser ikke ud til at være outliers.


%% b: Hypoteser
alfa = 0.05   % Signifikansniveau
delta = 0     % Vi undersøger om der er forskel på stikprøvernes populations-
              % middelværdier, altså om forskellen er på delta = 0

% Hypoteser.
% H0: mu_o - mu_f = delta
% Ha: mu_o - mu_f < delta (N.B. Ensidig test, da vi formoder, at modellen er blevet forbedret)


%% c: Teststatistik
% Vi laver en t-test for to uafhængige stikprøver med ukendt populationsvarians
% t0 = (y_o_streg - y_f_streg - delta)/(s_pooled*sqrt(1/n_o + 1/n_f))

y_o_streg = mean(E_o)   % Stikprøvemiddelværdi for oprindelig model
y_f_streg = mean(E_f)   % Stikprøvemiddelværdi for forbedret model
s_o = std(E_o)          % Stikprøvestandardafvigelse for oprindelig model
s_f = std(E_f)          % Stikprøvestandardafvigelse for forbedret model

df = n_o + n_f - 2     % Frihedsgrader, df = 22
% s_pooled er den puljede stikprøvespredning, hvor de to stikprøvers
% spredning er vægtet sammen efter antal observationer:
s_pooled = sqrt(((n_o - 1)*s_o^2 + (n_f - 1)*s_f^2)/df)   % s_pooled = 9.1198
t0 = (y_o_streg - y_f_streg - delta)/(s_pooled*sqrt(1/n_o + 1/n_f))

% Teststatistikken er t-fordelt med df = n_o + n_f - 2 = 22 frihedsgrader
% Værdien er beregnet til t0 = -2.7028


%% d: Kritisk region og konklusion
% Skridt 3. Kritisk region
```

```
% Vi har en ensidig test nedadtil, hvor vi afviser nulhypotesen, hvis
% t0 < t_alfa, hvor
% t_alfa = tinv(alfa,df)

t_alfa = tinv(alfa,df)

% t0 = -2.7028 og t_alfa = -1.7171. Da teststatistikken t0 er mindre end den
% kritiske værdi t_alfa afviser vi nulhypotesen.
% Det er altså lykkedes de studerende at få den nye model til at producere
% mere energi end den oprindelige.

pvalue = tcdf(t0, df)
% Det bekræftes også af p-value = 0.0065, som er mindre end signifikans-
% niveauet alfa = 0.05

[h,p,ci,stats] = ttest2(E_o, E_f, 'Alpha',alfa, 'Tail','left')
% Vi kan få bekræftet resultaterne med MatLab funktionen ttest2


%% e: 95% konfidensinterval
KI_span = tinv(1-alfa/2,df)*s_pooled*sqrt(1/n_o + 1/n_f)
KI_lav  = y_o_streg - y_f_streg - KI_span
KI_hoej = y_o_streg - y_f_streg + KI_span
% Forskellen i middel energiproduktion er med 95% sikkerhed i
% konfidensintervallet [-17.8462; -2.3496]


%% f: Diskussion af boksplot, hypotesetest og konfidensinterval
% Boksplot, hypotesetest og konfidensinterval giver det samme billede:
% Som nævnt under a) viser boksplottet, at der generelt er højere
% energiproduktion for den forbedrede bølgemaskine-model. Medianen ligger
% højere.
% Hypotesetesten viser, at der er signifikant forskel på middelværdierne,
% så den forbedrede model faktisk er en forbedring.
% 95% konfidenstintervallet for forskellen af middelværdier indeholder
% ikke 0, så det understøtter, at middelværdien er størst for den
% forbedrede model.


%% g: Antagelser
% Vi har antaget den centrale grænseværdisætning for hver stikprøve.
% Desuden har vi antaget at de to stikprøver har samme varians.
% Vi kan teste den centrale grænseværdisætning med stem-and-leaf plots og
% med normalfordelingsplots:
stemleafplot(E_o)
stemleafplot(E_f)

figure(2)
normplot(E_o)
title('Normalfordelingsplot for oprindelig bølgemaskine')

figure(3)
normplot(E_f)
title('Normalfordelingsplot for forbedret bølgemaskine')

% Begge stem-and-leaf plots viser 'pæne' fordelinger med et enkelt
% toppunkt, nogenlunde symmetrisk fordeling og hurtigt uddøende haler.
% Normalfordelingsplots viser også, at fordelingerne for stikprøverne
% ligner normalfordelingen. Dog er data for den oprindelige bølgemaskine
% ikke helt så overbevisende som for den forbedrede.
% De pæne fordelinger kombineret med rimeligt store stikprøvestørrelser
% gør, at vi kan være trygge ved, at antagelsen om den centrale
% grænseværdisætning holder.

% Antagelsen om at de to stikprøver har samme varians underbygges af det
% parallelle boksplot, der viser to ensartede bokse. Vi har desuden
% beregnet stikprøve-standardafvigelserne til værdier i samme
% størrelsesorden, nemlig hhv. 9.76 og 8.55
```