Løsningsforslag M4STI1 2016F

Opgave 1

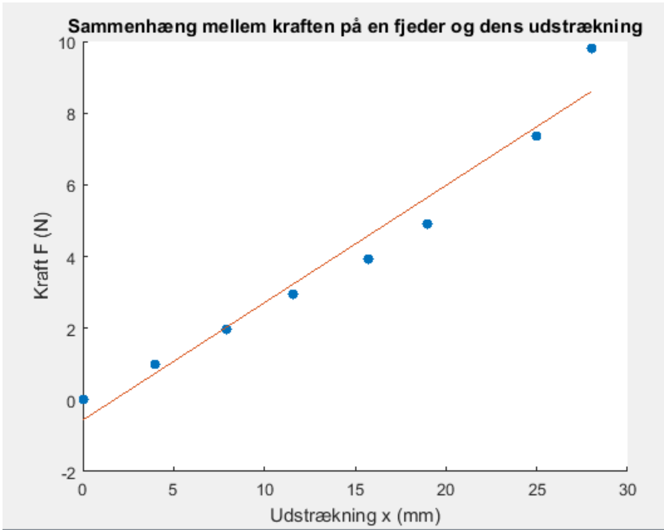
a. Jeg laver en lineær regression med x som regressor og F som respons. F er beregnet som loddernes masse, omregnet til kg, ganget med tyngdeaccelerationen. Funktionen mdl = fitlm(x,F) benyttes, og her er resultatet:

```
                   Estimate       SE        tStat       pValue  
                   ________    ________    _______    __________

    (Intercept)    -0.56444     0.44975     -1.255       0.25614
    x1              0.32732    0.026941      12.15    1.8901e-05


Number of observations: 8, Error degrees of freedom: 6
Root Mean Squared Error: 0.704
R-squared: 0.961,  Adjusted R-Squared 0.954
F-statistic vs. constant model: 148, p-value = 1.89e-05
```

Estimatet for fjederkonstanten k er proportionalitetsfaktoren mellem F og x, d.v.s.
k = 0.32732 N/mm.

b. P-værdien for bestemmelsen af fjederkonstanten er 1.8901e-05, så vi er næsten sikker på, at værdien ikke er nul. Til gengæld er p-værdien på 0.25614 for koefficienten b0 = -0.56444, så det er meget sandsynligt, at den i virkeligheden er 0. Det er jo også det, vi forventer af Hookes lov: b0 er 0 og b1 er en positiv konstant (fjederkonstanten). (Man kan faktisk godt tvinge MatLab til at holde b0=0, hvilket ville være mere korrekt, men det har vi ikke været inde på i kurset).
R-squared er 0.961 og Adjusted R-Squared er 0.954, så modellen forklarer næsten al variationen. Det er ikke så overraskende for en fysisk lov.

c. Figuren viser sammenhængen mellem målt udstrækning og den påvirkede kraft (blå punkter). Desuden vises regressionslinjen (rød linje). Trods de fine statistikker i regressionsanalysen, f.eks. høj R-i-anden, så lader linjen ikke til at fitte særligt godt til punkterne.



d. For at undersøge for 'unormale' punkter beregnes hat-diagonaler og studentiserede residualer (R-Student). De vises i følgende tabel:

x	F	Hat-diagonal	R-Student
0	0	0.4076	1.0498
4	0.981	0.2683	0.3624
7.9	1.962	0.1776	-0.0849
11.6	2.943	0.1327	-0.4095
15.7	3.924	0.1297	-0.9878
19	4.905	0.163	-1.2065
25	7.3575	0.3052	-0.4127
28	9.81	0.4158	5.1391

Løftestangspunkter (leverage) er unormale værdier x-retningen, og det måles med hat-diagonalen. Grænsen beregnes som lev_limit = 2*(c+1)/n, hvor c er antal regressorvariable og n er antal observationer. Her er c = 1 og n = 8, så limit = 0.5. Ingen punkter har en hat-diagonal over 0.5, så der er ingen løftestangspunkter.
Outliers er unormale værdier i y-retningen, og grænsen er 3 for den numeriske værdi af R-Student. Den sidste observation med F = 9.81 har |R-Student| = 5.1391 > 3, så det er en outlier. Da dette punkt ikke er et løftestangspunkt er der ingen influenspunkter.

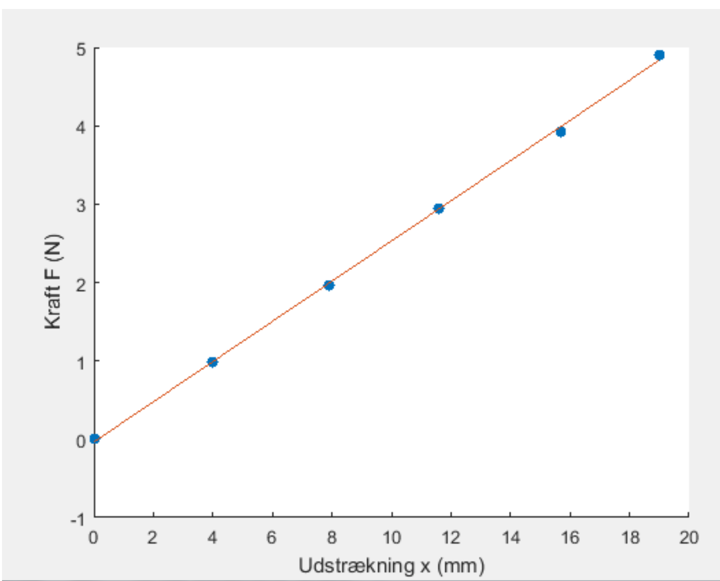
e. Analysen i delspørgsmål d. viste, at det sidste punkt er en outlier. Desuden viser figuren i delspørgsmål c. at det næstsidste punkt heller ikke lader til at passe til den lineære model.
Det er klart, at Hookes lov kun gælder indenfor et interval, der er begrænset af, hvor langt fjederen kan trækkes ud. Når fjederen nærmer sig sin maksimale længde vil en ekstra kraftpåvirkning ikke resultere i en yderligere forlængelse med proportionalitetsfaktoren k. Derfor er det i orden at udelukke punkter, der lader til at ligge udenfor intervallet, hvor Hookes lov gælder, til bestemmelsen af fjederkonstanten.

En ny lineær regression, hvor de sidste to punkter er udeladt giver følgende resultat:

```
                   Estimate        SE          tStat        pValue
                  _________    _________    _________    __________

    (Intercept)   -0.034997     0.038755     -0.90305       0.41756
    x1              0.25644    0.0033141       77.379    1.6718e-07


Number of observations: 6, Error degrees of freedom: 4
Root Mean Squared Error: 0.053
R-squared: 0.999,  Adjusted R-Squared 0.999
F-statistic vs. constant model: 5.99e+03, p-value = 1.67e-07
```

Nu bestemmes fjederkonstanten således til:
k = 0.25644 N/mm.

Det ses, at modellen nu forklarer næsten al variationen i data (R^2 og Adjusted er begge 0.999). Regressionslinjen fitter målepunkterne næsten perfekt:



Matlab kode for opgave 1:

```
%% Opgave 1: Hookes lov til bestemmelse af fjederkonstant
clc; clear all; close all;

%%Indlæsning af data
M = xlsread('Data_M4STI1_2016F.xlsx', 'A:B')
```

```matlab
% Omregning til F og x (Newton og mm)
m = M(:,1)/1000     % Loddernes masse målt i kg
F = m*9.81          % Loddernes tyngdekraft på fjederen (kg*m/s2 = N)
x = M(:,2)          % Fjederens forlængelse i mm

%% a: Lineær regression
mdl = fitlm(x,F)    % Lineær regression, hvor F udtrykkes som funktion af x
% Fjederkonstanten k aflæses som hældningskoefficienten, så k = b1 = 0.32732


%% b: Forklaring af modellens statistikker
% P-værdien for bestemmelsen af fjederkonstanten er 1.8901e-05, så vi er
% næsten sikker på, at værdien ikke er nul. Til gengæld er p-værdien på
% 0.25614 for koefficienten b0 = -0.56444, så det er meget sandsynligt, at
% den i virkeligheden er 0. Det er jo også det, vi forventer af Hookes lov:
% b0 er 0 og b1 er en positiv konstant (fjederkonstanten).
% (Man kan faktisk godt tvinge MatLab til at holde b0=0, hvilket ville være
% mere korrekt, men det har de studerende ikke lært)
% R-squared er 0.961 og Adjusted R-Squared er 0.954, så modellen forklarer
% næsten al variationen. Det er ikke så overraskende for en fysisk lov.



%% c: Figurer
y_hat = -0.56444 + 0.32732*x

figure(1)
hold on;
scatter(x,F, 'Filled')
plot(x, y_hat)
title('Sammenhæng mellem kraften på en fjeder og dens udstrækning');
xlabel('Udstrækning x (mm)');
ylabel('Kraft F (N)');
hold off;

% Alternativt kan en af MatLabs indbyggede funktioner bruges
figure(2)
plot(mdl)

%% d: Unormale punkter
lev = mdl.Diagnostics.Leverage;     % hat diagonal
rst = mdl.Residuals.Studentized;    % R-Student
resultat = [x, F, lev, rst]         % Jeg samler det hele i en tabel

c = 1           % Antal regressorvariable
n = size(x,1)   % Antal observationer

lev_limit = 2*(c+1)/n

% Grænsen for leverage er lev_limit = 0.5. Ingen punkter har lev > 0.5, så
% der er ingen løftestangspunkter (leverage)
% Grænsen for outlier er 3. Et enkelt punkt har numerisk værdi over 3,
% nemlig den sidste observation med F = 9.81. Det har |rst| = 5.1391 > 3,
```

```matlab
% så det er en outlier.
% Da dette punkt ikke er et løftestangspunkt er der ingen influenspunkter.


%% e: Bedre bestemmelse af fjederkonstanten
% Det er klart, at Hookes lov kun gælder indenfor et interval, der er
% begrænset af, hvor langt fjederen kan trækkes ud. Når fjederen nærmer sig
% sin maksimale længde vil en ekstra kraftpåvirkning ikke resultere i en
% yderligere forlængelse med proportionalitetsfaktoren k. Derfor er det i
% orden at udelukke punkter, der lader til at ligge udenfor intervallet,
% hvor Hookes lov gælder.

% Ud fra figuren i c) og analysen af outliers i d) tyder det på, at de
% sidste to målinger er udenfor intervallet, hvor Hookes lov gælder.
% Derfor vil jeg udelukke de sidste to punkter og bestemme fjederkonstanten
% ud fra de første 6.

x2 = x(1:6,:)
F2 = F(1:6,:)
mdl2 = fitlm(x2, F2)

y_hat2 = -0.034997 + 0.25644*x2

figure(3)
hold on;
scatter(x2, F2, 'Filled')
plot(x2, y_hat2)
xlabel('Udstrækning x (mm)');
ylabel('Kraft F (N)');
hold off;

% Alternativ
figure(4)
plot(mdl2)

% Nu bestemmes fjederkonstanten til k = 0.25644 N/mm.

% Ikke overraskende er modellen nu endnu bedre. P-værdien for bestemmelsen
% af fjederkonstanten er 1.6718e-07, så vi er næsten sikker på,
% at værdien ikke er nul. Til gengæld er p-værdien på
% 0.41756 for koefficienten b0 = -0.034997, så det er meget sandsynligt, at
% den i virkeligheden er 0, som forventet af Hookes lov.
% Både R-squared og Adjusted R-Squared er 0.999, så modellen forklarer
% næsten al variationen.
```

Opgave 2

a. Sandsynligheden for en specifik værdi med en kontinuert stokastisk variabel er altid nul:
P(Y = 4.0) = 0.

b. P(Y>4.1) = 1 - P(Y<4.1) = 1 - normcdf(4.1, 4.0, 0.1) = 1 - 0.8413 = 0.1587.

c. Når 99% af fjedrene skal ligge i det søgte interval, så vil 0.5% ligge over intervallet og 0.5% vil ligge under. Den nedre grænse for intervallet, v_nedre, kan findes som, hvor 0.5% af sandsynlighedsmassen ligger under:

    v_nedre = norminv(0.005, 4.0, 0.1) = 3.7424.

    Den øvre grænse for intervallet, v_oevre, kan tilsvarende findes som den værdi, hvor 99.5% ligger under (og derfor 0.5% ligger over):

    v_oevre = norminv(1-0.005, 4.0, 0.1) = 4.2576.

    Med andre ord ligger 99% af fjedrenes fjederkonstanter imellem 3.7424 og 4.2576.

MatLab kode for opgave 2:

```
%% Opgave 2: Sandsynligheder om fjedre
clc; clear all; close all;

%% Baggrundsoplysninger
mu_0 = 4.0       % Producentens oplyste middelværdi
sigma_0 = 0.1    % Producentens oplyste spredning

%% a
% Sandsynligheden for en specifik værdi er per definition 0,
% da normalfordelingen er en kontinuert fordeling
% P(k = 4.0) = 0

%% b
% Sandsynligheden for at en fjeder har fjederkonstant k > 4.1 er 1 minus
% sandsynligheden for at den har k < 4.1
P_under4komma1 = normcdf(4.1, mu_0, sigma_0)
P_over4komma1 = 1 - P_under4komma1

% P(k > 4.1) = 1 - P(k < 4.1) = 0.1587

%% c
% Når 99% af fjedrene skal ligge i det søgte interval, så vil 0.5% ligge
% over intervallet og 0.5% vil ligge under.

alpha = 0.01
alphahalf = alpha/2
```

```
% Intervallets nedre værdi:
v_low = norminv(alphahalf, mu_0, sigma_0)

% Intervallets øvre værdi:
v_high = norminv(1-alphahalf, mu_0, sigma_0)

% Det ønskede interval er (3.7424, 4.2576)
```

Opgave 3

a. Formlen for beregning af stikprøvestørrelsen er:

$$n = (z\_alphahalf*sigma\_0/B)^2$$

hvor sigma_0 er den påståede spredning på 0.1 N/mm, B er intervalbredden på 0.05 N/mm og z_alphahalf er beregnet som

z_alphahalf = -norminv(alpha/2)

med alpha = 0.05, da et 95% konfidensinterval svarer til at signifikansniveauet alpha er 5% = 0.05. Vi får z_alphahalf = 1.96 og dermed n = 15.3658. Da stikprøvestørrelsen skal være et heltal, runder vi for en sikkerheds skyld op til n = 16.
Ved at vælge en stikprøvestørrelse på 16 er vi mindst 95% sikre på, at den sande middelværdi ligger i intervallet y_streg +/- 0.05, hvor y_streg er stikprøvemiddelværdien.

b. Hypotesetesten går ud på at undersøge, om variansen i virkeligheden er større end det, som fabrikken har oplyst. Fabrikken har oplyst at spredningen er 0.1 N/mm, så variansen er var_0 = 0.01 $(N/mm)^2$.

H0: Variansen er lig med 0.01 $(N/mm)^2$

Ha: Variansen er større end 0.01 $(N/mm)^2$

Det er altså en ensidet test.

c. Teststatistikken for test af varians i en enkelt stikprøve er

chi2_0 = (n - 1)*s2/var_0,

hvor er antal observationer i stikprøven, var_0 er den påståede varians og s2 er stikprøvevariansen. Teststatistikken følger en chi-i-anden fordeling med n-1 frihedsgrader.

d. Da vi har valgt en ensidig hypotesetest opadtil forkaster vi nulhypotesen, hvis teststatistikken er større end den kritiske værdi, chi2_kritisk, som har den egenskab, at arealet under kurven for chi2 fordelingen over den kritiske værdi er netop signifikansniveauet alpha=0.05. Dermed er den kritiske værdi

chi2_kritisk = chi2inv(1 - alpha, n - 1) = 30.1435.

Vi forkaster nulhypotesen, hvis teststatistikken chi2_0 er over chi2_kritisk.

Stikprøvevariansen beregnes til 0.0287, så teststatistikken er

chi2_0 = (n-1)*s2/var_0 = 54.5000.

Teststatistikken chi2_0 er 54.5000 og chi2_kritisk er 30.1435. Da teststatistikken er over den kritiske værdi kan vi forkaste nulhypotesen. Variansen er altså i virkeligheden over producentens angivne værdi. Med andre ord er fjedrene ikke så ensartede, som producenten har lovet.

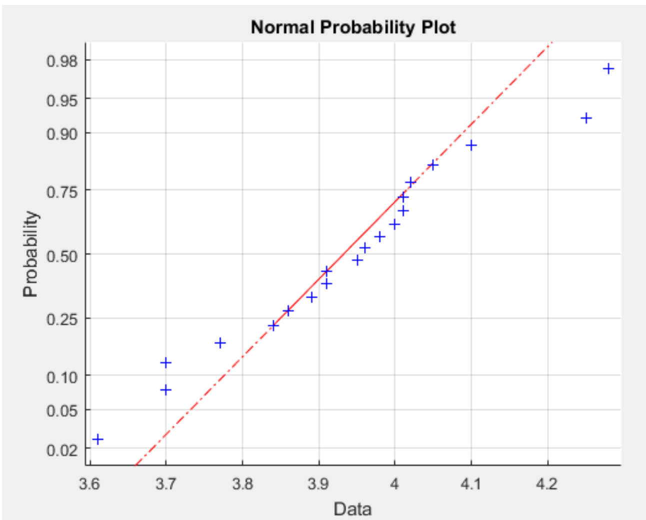
e. 95% konfidensintervallet for den sande varians er intervallet mellem værdierne CI_low og CI_high, som bestemmes her:

CI_low = (n - 1)*s2/chi2inv(1-alpha/2, n-1) = 0.0166

CI_high = (n - 1)*s2/chi2inv(alpha/2, n-1) = 0.0612

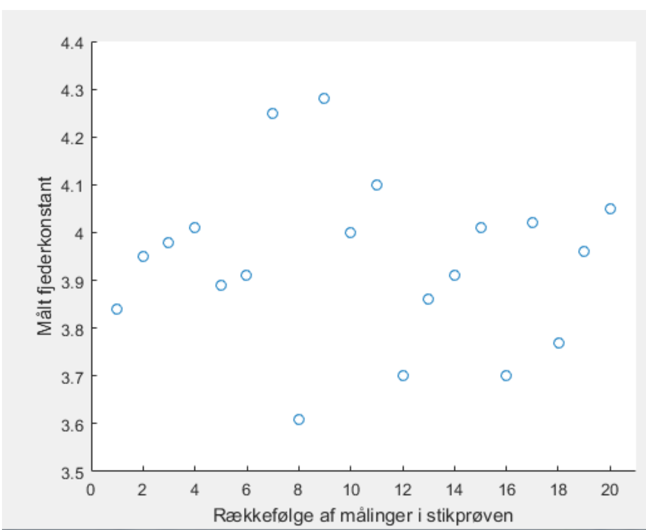
95% konfidensintervallet er (0.0166, 0.0612). Med andre ord er vi 95% sikre på, at variansen ligger mellem 0.0166 og 0.0612. Bemærk, at den lovede varians på 0.01 ligger udenfor dette interval.

f. Hypotesetesten for varians antager, at stikprøven er tilfældig og kommer fra en normalfordeling. Producenten har oplyst, at fjedrene er normalfordelt, men vi kan teste det med et normalfordelingsplot:



Resultatet af normplot er, at stikprøvens observationer ikke ligger særligt pænt på en linje. Derfor kan vi være bekymrede for, om hypotesetestens konklusioner holder.

Vi kan teste om stikprøven er tilfældig ved at plotte målingerne efter rækkefølge:



Der er ikke nogen tydelig sammenhæng mellem målingerne, så vi har ingen grund til at tro, at stikprøven ikke er tilfældig.

g. Hypotesetest af om middelværdien er lavere end de lovede 4.0 N/mm.

   H0: Middelværdien er 4.0 N/mm

   Ha: Middelværdien er under 4.0 N/mm

h. Vi har fået oplyst populations-spredningen sigma_0 = 0.1 N/mm, men den foregående hypotesetest viser, at denne værdi ikke er korrekt. Derfor er det bedst at gå ud fra, at vi ikke kender spredningen, så den må estimeres med stikprøvens spredning. Derfor er teststatistikken:

   t0 = (y_streg - mu_0)/(s/sqrt(n)),

   hvor y_streg og s er hhv. middelværdi og spredning for stikprøven, mu_0 er den påståede middelværdi på 4.0 N/mm, og n er stikprøvestørrelsen.

   Teststatistikken t0 er t-fordelt med n-1 frihedsgrader.

i. Da vi har en ensidet hypotesetest nedadtil skal vi finde den kritiske værdi, t_kritisk, så vi forkaster nulhypotesen, hvis teststatistikken bliver mindre end den kritiske værdi.

   t_kritisk = tinv(alpha, n-1)

   Vi skal forkaste H0 hvis t0 < t_kritisk.

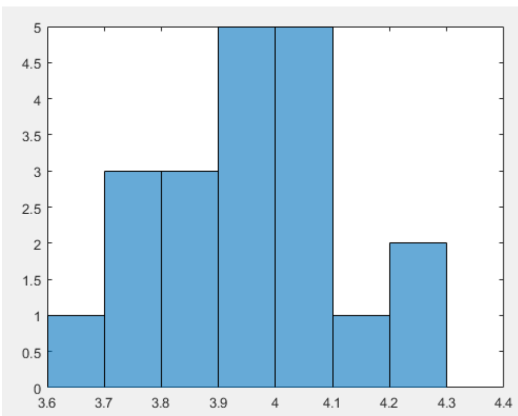
   Teststatistikken t_0 beregnes til -1.5843, mens den kritiske værdi t_kritisk er -1.7291. Vi kan altså ikke forkaste H0 på 5% signifikansniveau. Selv om stikprøvens middelværdi er 3.94, og altså under den lovede værdi på 4.0, så giver stikprøven ikke bevis nok til at kunne sige, at den sande middelværdi er under 4.0.

j. I hypotesetesten for middelværdi har vi blot antaget den centrale grænseværdisætning. Det er en mildere antagelse end for hypotesetest af varians, hvor vi antog at stikprøven kom fra en normalfordeling. Her er det tilstrækkeligt, at stikprøven kommer fra en 'pæn' fordeling, dvs. en nogenlunde symmetrisk fordeling med et enkelttoppunkt og uddøende haler. Da stikprøvestørrelsen på 20 er forholdsvis høj holder antagelsen, selv hvis fordelingen ikke er særlig pæn. Vi kan teste antagelsen om en pæn fordeling med stem-and-leaf plot og med et histogram.

   Stem-and-leaf plot:

```
  36 | 1
  37 | 0 0 7
  38 | 4 6 9
  39 | 1 1 5 6 8
  40 | 0 1 1 2 5 0
  41 |
  42 | 5 8
key: 36|5 = 3.65
stem unit: 0.10
leaf unit: 0.01
```

Histogram:



Begge diagrammer viser, ligesom normplottet i delspørgsmål f, at fordelingen er pæn, så antagelsen holder.

MatLab kode for opgave 3:

```
%% Opgave 3: Hypotesetests af fjedre
clc; clear all; close all;

%% Baggrundsoplysninger
mu_0 = 4.0       % Producentens oplyste middelværdi
sigma_0 = 0.1    % Producentens oplyste spredning


%% a: Stikprøvestørrelse
% Formlen for beregning af stikprøvestørrelsen er (4.2) s. 179 i V&K:
% n = (z_alphahalf*sigma_0/B)^2

alpha = 0.05     % Signifikansniveau (5%)
B = 0.05         % Ønsket bredde af konfidensinterval

z_alphahalf = -norminv(alpha/2)
```

```
n = (z_alphahalf*sigma_0/B)^2

% Resultatet er n = 15.3658, så vi runder op til n = 16.
% Ved at vælge en stikprøvestørrelse på 16 er vi mindst 95% sikre på,
% at den sande middelværdi mu ligger i intervallet y_streg +/- 0.05,
% hvor y_streg er stikprøvemiddelværdien.


%% Indlæsning af stikprøvedata
y = xlsread('Data_M4STI1_2016F.xlsx', 'D:D')
n = size(y,1)

%% b: Hypoteser for hypotesetest af fjedrenes varians
% Hypotesetesten foretages for varians. Vi har fået oplyst producentens
% værdi for spredning, så vi beregner variansen som kvadratet af
% spredningen:

var_0 = (sigma_0)^2  % var_0 er oplyst varians. Vi beregner den som
                     % sigma i anden, dvs. var_0 = (0.1 N/mm)^2

% Maskinfabrikkens mistanke er, at variansen i virkeligheden er større
% end oplyst af producenten, så det er en ensidig hypotesetest
% H0: var = var_0
% Ha: var > var_0


%% c: Teststatistikken
% Teststatistikken for test af varians i en enkelt stikprøve er
% chi2_0 = (n-1)*s2/var_0, hvor s2 er stikprøvevariansen
% Teststatistikken følger en chi-i-anden fordeling med n-1 frihedsgrader


%% d: Kritisk region, teststatistikken og konklusion
% Da vi har valgt en ensidig hypotesetest opadtil forkaster vi nulhypotesen,
% hvis teststatistikken er større end den kritiske værdi, chi2_kritisk, som
% har den egenskab, at arealet under kurven for chi2 fordelingen over den
% kritiske værdi er netop alpha. Dermed er den kritiske værdi
chi2_kritisk = chi2inv(1-alpha, n-1)

% Stikprøvevarians s2:
s2 = var(y)
% Teststatistik chi2_0:
chi2_0 = (n-1)*s2/var_0

% Konklusion:
% Teststatistikken chi2_0 er 54.5000 og chi2_kritisk er 30.1435.
% Da teststatistikken er over den kritiske værdi kan vi forkaste nulhypotesen.
% Variansen er altså i virkeligheden over producentens angivne værdi. Med
% andre ord er fjedrene ikke så ensartede, som producenten har lovet.

%% e: 95% konfidensinterval
% Nedre værdi:
```

```
CI_low = (n - 1)*s2/chi2inv(1-alpha/2, n-1)
% Øvre værdi:
CI_high = (n - 1)*s2/chi2inv(alpha/2, n-1)

% Konfidensinterval: (0.0166, 0.0612)
% Bemærk, at den oplyste varians på 0.01 ikke ligger i intervallet.

%% f: Antagelser
% Hypotesetesten for varians antager, at stikprøven er tilfældig og
% kommer fra en normalfordeling.
% Producenten har oplyst, at fjedrene er normalfordelt, men vi kan teste
% det med et normalfordelingsplot

figure(1)
normplot(y)

% Resultatet af normplot er, at stikprøvens observationer ikke ligger
% særligt pænt på en linje. Derfor kan vi være bekymrede for, om
% hypotesetestens konklusioner holder.
% Vi kan teste om stikprøven er tilfældig ved at plotte målingerne efter
% rækkefølge
x = [1:20];

figure(2)
scatter(x,y)
xlabel('Rækkefølge af målinger i stikprøven');
ylabel('Målt fjederkonstant');
axis([0 21 3.5 4.4]);
% Der er ikke nogen tydelig sammenhæng mellem målingerne, så vi har ingen
% grund til at tro, at stikprøven ikke er tilfældig.


%% g: Hypotesetest for fjedrenes middelværdi
% Maskinfabrikken har mistanke om, at middelværdien er lavere end 4.0,
% derfor vælger vi en ensidet test:
% H0: mu = mu_0
% Ha: mu < mu_0


%% h: Teststatistikken
% Vi har fået oplyst populations-spredningen sigma = 0.1 N/mm, men den
% foregående hypotesetest viser, at denne værdi ikke er korrekt. Derfor er
% det bedst atgå ud fra, at vi ikke kender spredningen, så den må estimeres
% med stikprøvens spredning. Derfor er teststatistikken:
% t0 = (y_streg - mu_0)/(s/sqrt(n)), hvor y_streg og s er hhv. middelværdi
% og spredning for stikprøven, og n er stikprøvestørrelsen.
% t0 er t-fordelt med n-1 frihedsgrader


%% i: Kritisk værdi, teststatistikken og konklusion

t_kritisk = tinv(alpha, n-1)
```

```
% Vi skal forkaste H0 hvis t0 < t_kritisk

y_streg = mean(y)
s = std(y)
t_0 = (y_streg - mu_0)/(s/sqrt(n))

% Konklusion:
% Teststatistikken t_0 er -1.5843, mens den kritiske værdi t_kritisk er -1.7291.
% Vi kan altså ikke forkaste H0 på 5% signifikansniveau. Selv om
% stikprøvens middelværdi er 3.94, og altså under den lovede værdi på 4.0,
% så giver stikprøven ikke bevis nok til at kunne sige, at den sande
% middelværdi er under 4.0.

%% j: Antagelser
% I hypotesetesten for middelværdi har vi blot antaget den centrale
% grænseværdisætning. Det er en mildere antagelse end for hypotesetest af
% varians, hvor vi antog at stikprøven kom fra en normalfordeling.
% Her er det tilstrækkeligt, at stikprøven kommer fra
% en 'pæn' fordeling, dvs. en nogenlunde symmetrisk fordeling med et enkelt
% toppunkt og uddøende haler. Da stikprøvestørrelsen på 20 er forholdsvis
% høj holder antagelsen, selv hvis fordelingen ikke er særlig pæn. Vi kan
% teste antagelsen om en pæn fordeling med stem-and-leaf plot og med et
% histogram.

stemleafplot(y,-2)

% Vi kan også teste om fordelingen er 'pæn' med et histogram:
histogram(y,7)

% Begge diagrammer viser, ligesom normplottet i f), at fordelingen er pæn.
```