Løsningsforslag M4STI1 2016E

Opgave 1

Producenten af GPS-enheden har oplyst, at afstanden i meter fra den ønskede til den virkelige position kan beskrives med en eksponentialfordeling med $\lambda = 2.22$. Det svarer til, at GPS-enheden i gennemsnit rammer $\mu = {}^{1}/_{\lambda} = 0.45$ meter fra målet.

Beregn følgende under forudsætning af, at GPS-producentens oplysninger er korrekte:

a. *Sandsynligheden for at dronen lander indenfor en meter fra målet:*

   Vi skal bruge cdf (cumulated distribution function) for eksponentialfordelingen, d.v.s. expcdf:
   expcdf(1.0, 0.45) = <u>0.8916</u>

b. *Sandsynligheden for at dronen lander mellem 1 og 2 meter fra målet:*

   expcdf(2.0, 0.45) - expcdf(1.0, 0.45) = <u>0.0966</u>

c. *Sandsynligheden for at dronen lander mere end 2 meter fra målet:*
   Det er det samme som 1 minus sandsynligheden for at den lander under 2 meter fra målet:

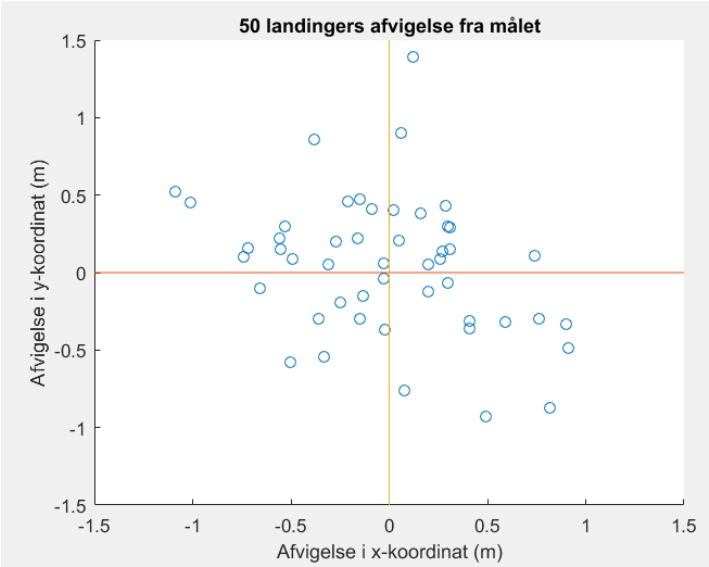
   1 - expcdf(2.0, 0.45) = <u>0.0117</u>

MatLab kode til opgave 1

```
%% E2016 Opg 1: Dronens landingspræcision
clear all; close all; clc;

mu = 0.45
lambda = 1/mu

%% a
p_under1m = expcdf(1,mu)

%% b
p_mellem1og2m = expcdf(2,mu) - expcdf(1,mu)

%% c
p_over2m = 1 - expcdf(2,mu)
```

Opgave 2

a. *Lav en figur, der plotter delta_y mod delta_x. Hvad kan figuren fortælle dig om data?*



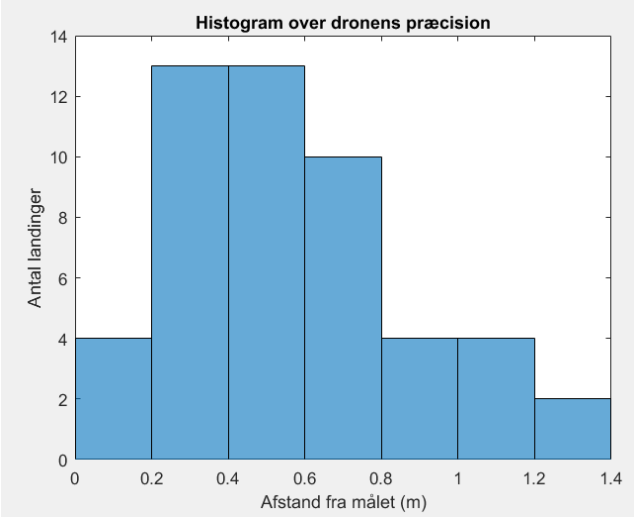
Punkterne lader til at være nogenlunde tilfældigt fordelt. Der er måske en tendens til at være mindre variation i øverste, højre og i nederste, venstre kvadrant.

b. *Beregn hvor langt hver af de 50 landinger har afveget fra målet og vis resultatet i et histogram. Hvordan stemmer histogrammet overens med GPS-leverandørens påstand om, at afvigelsen følger en eksponentialfordeling?*

Afstanden fra målet, r, beregnes som:

$$r = \sqrt{(\text{delta\_x})^2 + (\text{delta\_y})^2}$$

Nedenfor vises et histogram over de 50 landingers præcision. Histogrammet ligner ikke eksponentialfordelingen særligt godt, da vi ville forvente flest landinger i den laveste kategori (her 0 - 0.2m) og færre og færre landinger, jo længere væk fra målet, vi kommer.



De studerende har på baggrund af eksperimentet med de 50 landinger en mistanke om, at dronen ikke kan overholde den lovede præcision. Med andre ord har de mistanke om, at den i gennemsnit lander mere end 0.45 meter fra målet. De vil bruge de 50 landinger som en stikprøve til en hypotesetest af dronens sande præcision med et signifikansniveau på 5%.

c. *Opstil nulhypotese og alternativhypotese. Vil du vælge en-sidet eller to-sidet test? Begrund dit valg.*

H_0: $\mu = \mu_0 = 0.45$
H_a: $\mu > \mu_0$

Jeg vælger en en-sidet hypotesetest opad, da de studerende har mistanke om at præcisionen er *over* 0.45 m.

d. *Opstil en formel for teststatistikken. Angiv hvilken fordeling den følger.*

$$t_0 = \frac{\bar{r} - \mu_0}{s/\sqrt{n}}$$

Her er $\bar{r}$ den gennemsnitlige afstand fra målet i stikprøvens 50 landinger, s er stikprøvens standardafvigelse og n er stikprøvestørrelsen på 50. Teststatistikken t_0 er t-fordelt med n-1 frihedsgrader.

e. *Beregn den kritiske region for testen, beregn teststatistikkens værdi og konkludér på hypotesetesten.*

Da vi har en ensidet hypotesetest opad, forkaster vi nulhypotesen, hvis teststatistikken t_0 er større end den kritiske værdi, t_α. Her er α signifikansniveauet på 0.05.

$$t_\alpha = \text{tinv}(1 - \alpha, n - 1) = 1.6766$$

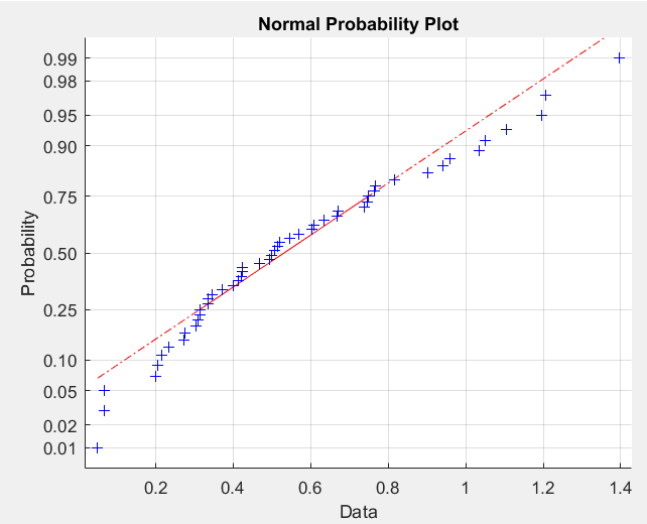
Vi beregner stikprøve-middelværdien til 0.56 (altså over 0.45) og stikprøve-standardafvigelsen til 0.32. Dermed bliver t_0 beregnet til 2.43.

Da teststatistikken er større end den kritiske grænse kan vi forkaste nulhypotesen på 5% signifikansniveau. Med andre ord er dronens præcision større end de påståede 0.45 meter.

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

Vi har antaget den centrale grænseværdisætning for at kunne ræsonnere om teststatistikken t_0. Antagelsen er, at stikprøven kommer fra en 'pæn' fordeling. Vi kan teste, om antagelsen holder med et stem-and-leaf plot og med et normalfordelingsplot, som vist nedenfor:

```
  0 | 5 7 7
  1 |
  2 | 0 1 2 3 7 8 0
  3 | 1 1 1 4 4 4 7
  4 | 0 1 2 2 2 7 9
  5 | 0 1 1 2 5 7 0
  6 | 1 3 7 7
  7 | 4 5 5 6 7
  8 | 2
  9 | 0 4 6
 10 | 3 5
 11 | 1 0
 12 | 1
 13 | 0
key: 36|5 = 3.65
stem unit: 0.10
leaf unit: 0.01
```



Stem-and-leaf plottet viser en nogenlunde pæn fordeling med et toppunkt og uddøende haler. Normalfordelingsplottet viser en nogenlunde ret linje. Vigtigst er det dog, at stikprøvestørrelsen på n = 50 gør, at den centrale grænseværdisætning holder, også for mindre pæne fordelinger.

g. *Beregn et to-sidet 95% konfidensinterval for dronens præcision. Hvad fortæller konfidensintervallet?*

Med 95% sikkerhed er dronens sande præcision i konfidensintervallet [0.47; 0.65]. Bemærk, at 0.45 ligger udenfor konfidensintervallet.

MatLab kode til opgave 2

```
%% E2016 Opg 2: Hypotesetest om dronens landingspræcision
clear all; close all; clc;

% Indlæs data
M = xlsread('Data_M4STI1_2016E.xlsx', 'A:B')
delta_x = M(:,1)
delta_y = M(:,2)

%% a
figure(1)
hold on;
scatter(delta_x, delta_y)
axis([-1.5 1.5 -1.5 1.5])
plot([-1.5; 1.5], [0; 0])   % vandret referencelinje gennem (0,0)
plot([0; 0], [-1.5; 1.5])   % lodret referencelinje gennem (0,0)
title('50 landingers afvigelse fra målet');
xlabel('Afvigelse i x-koordinat (m)');
ylabel('Afvigelse i y-koordinat (m)');
hold off;
% Punkterne lader til at være nogenlunde tilfældigt fordelt.
% Der er måske en tendens til at være mindre variation i øverste, højre og
% i nederste, venstre kvadrant.


%% b
r = sqrt(delta_x.^2 + delta_y.^2)

figure(2)
histogram(r,7)
title('Histogram over dronens præcision');
xlabel('Afstand fra målet (m)');
ylabel('Antal landinger');
% Histogrammet ligner ikke eksponentialfordelingen særligt godt, da første
% søjle burde være højest.


%% Hypotesetest
mu0 = 0.45
alfa = 0.05
n = 50

%% c
% Skridt 1
% H0: mu = mu0
% Ha: mu > mu0
% Formodningen er, at dronen er mere upræcis end 0.45m, så vi vil ikke
% forkaste nulhypotesen, hvis stikprøven er mindre end 0.45. Derfor ensidet
% hypotesetest.

%% d
% Skridt 2
% t0 = (r_streg - mu0)/(s/sqrt(n))
% t0 er t-fordelt med n-1 frihedsgrader
% r_streg er den gennemsnitlige afstand fra målet i de 50 flyvninger.
% s er stikprøve-standardafvigelsen.
```

```
%% e
% Skridt 3
t_alfa = tinv(1-alfa,n-1)
% Kritisk region. Vi forkaster H0, hvis t0 > t_alfa
% t_alfa = 1.6766.

% Skridt 4
r_streg = mean(r)
s = std(r)
t0 = (r_streg - mu0)/(s/sqrt(n))

% Skridt 5
% Vi får en stikprøve-middelværdi på r_streg = 0.56, altså over den
% formodede populationsmiddelværdi på mu0 = 0.45. Hypotesetesten giver, at
% teststatistikken t0 = 2.4309 er større end den kritiske grænse på
% t_alfa = 1.6766. Derfor kan vi forkaste nulhypotesen på 5% signifikans-
% niveau.

%% f. Antagelser
% Vi har antaget den centrale grænseværdisætning for at kunne ræsonnere om
% teststatistikken t0. Antagelsen er, at stikprøven kommer fra en 'pæn'
fordeling.
% Vi kan teste om den holder med et stem-and-leaf plot
% og med et normalfordelingsplot:
stemleafplot(r,-2)
normplot(r)
% Stem-and-leaf plottet viser en nogenlunde pæn fordeling med et toppunkt
% og uddøende haler. Normalfordelingsplottet viser en nogenlunde ret linje.
% Vigtigst er det dog, at stikprøvestørrelsen på n = 50 gør, at den
% centrale grænseværdisætning holder, også for mindre pæne fordelinger.

%% g: 95 pct. konfidensinterval

t_alfahalve = tinv(1-alfa/2, n-1)
CI_bredde = t_alfahalve*s/sqrt(n)

CI_nedre = r_streg - CI_bredde
CI_oevre = r_streg + CI_bredde

% Vi får CI_nedre = 0.4689 og CI_oevre = 0.6494, så 95%
% konfidensintervallet er (0.4689; 0.6494).
% Det betyder, at vi er 95% sikre på, at dronens sande landings-præcision
% ligger imellem 0.47 og 0.65 meter. Vi kan bemærke, at den oplyste
% præcision på 0.45 meter ligger udenfor dette interval.
```

Opgave 3

Eksperiment med målt opladningsgrad (%) efter flyvning af en kilometer med forskellige belastninger.

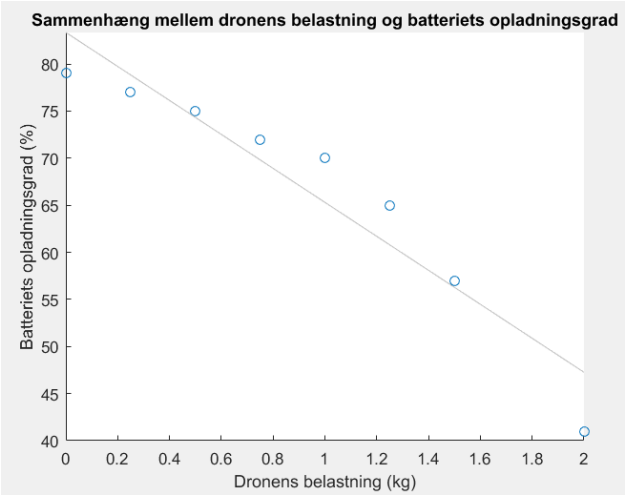
a. *Lav en lineær regression, der viser batteriets opladningsgrad som funktion af dronens belastning. Forklar ved hjælp af regressionsanalysens statistikker, om modellen er god.*

Den simple lineære regression laves med MatLab funktionen fitlm , som giver følgende resultat:

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

    (Intercept)     83.353     2.6316     31.674    6.5809e-08
    x1             -18.045     2.3914    -7.5457    0.00028102


Number of observations: 8, Error degrees of freedom: 6
Root Mean Squared Error: 4.22
R-squared: 0.905,  Adjusted R-Squared 0.889
F-statistic vs. constant model: 56.9, p-value = 0.000281
```

Med R-squared = 0.905 og Adjusted R-Squared = 0.889 forklarer modellen en meget stor del af variationen i data. Begge koefficienter er signifikant forskellige fra 0, da de har p-værdier tæt på 0. Der er således negativ korrelation i data (p-værdi =0.00028102). At dømme ud fra regressionsanalysen alene, er det en god model.

b. *Lav en figur, der illustrerer data og regressionsmodellen. Kommentér figuren.*



Figuren viser, at selv om regressionsanalysens statistikker var gode, så er det ikke en lineær sammenhæng imellem belastning og opladningsgrad. For små belastninger op til ca. et kg ser sammenhængen lineær ud, men så begynder belastningen at være hårdt ved batteriet.

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

Tabellen nedenfor viser resultatet af undersøgelsen af 'unormale' datapunkter:

Belastning (kg)	Opladningsgrad (%)	Hat-diagonal	R-student
0.000	79	0.3885	-1.4282
0.250	77	0.2632	-0.4743
0.500	75	0.1779	0.1600
0.750	72	0.1328	0.5197
1.000	70	0.1278	1.2426
1.250	65	0.1629	1.1086
1.500	57	0.2381	0.1775
2.000	41	0.5088	-3.8379

Vi kan se i tabellen resultat, at kun det sidste punkt med belastningen 2 kg har studentiseret residual med numerisk værdi over grænseværdien 3, nemlig -3.8379. Dette punkt er derfor en outlier.

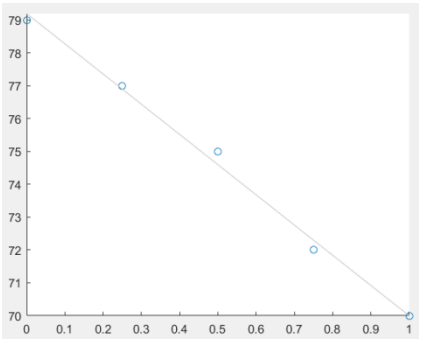
Grænseværdien for løftestangspunkt beregnes til $\frac{2(k+1)}{n} = \frac{2(1+1)}{8} = 0.5$. Det er også kun det sidste punkt med belastning på 2 kg, der har hat-diagonal over 0.5. Dette punkt har en hat-diagonal på 0.5088, så det er et løftestangspunkt. Da det samme punkt både er outlier og løftestangspunkt, så er det et indflydelsespunkt. De andre punkter er normale.

d. *En typisk pizza, som dronen skal fragte, er en deep pan pizza på 630 gram. Giv dit bedste bud på opladningsgraden, når dronen har fragtet sådan en pizza 1 kilometer, og begrund dit bud.*

Man kunne forsøge sig med at transformere data for at finde en lineær sammenhæng for hele datasættet, men da kurven er fint lineær indtil en belastning på 1 kg, og 0.63kg befinder sig på den lineære del, så laver jeg blot en lineær regression på de første fem datapunkter.

	Estimate	SE	tStat	pValue
(Intercept)	79.2	0.24495	323.33	6.5239e-08
x1	-9.2	0.4	-23	0.00018003

Number of observations: 5, Error degrees of freedom: 3
Root Mean Squared Error: 0.316
R-squared: 0.994, Adjusted R-Squared 0.992
F-statistic vs. constant model: 529, p-value = 0.00018



Den forventede opladningsgrad efter transport af pizzaen kan beregnes fra regressionslinjen:
opladn_630 = 79.2 - 9.2*0.63 = <u>73.4</u>

MatLab kode til opgave 3

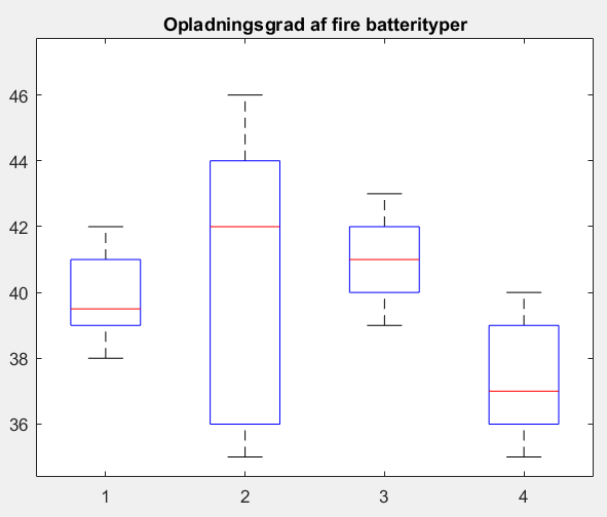
```
%% E2016 Opg 3: Batteriets opladningsgrad
clear all; close all; clc;

M = xlsread('Data_M4STI1_2016E.xlsx', 'D:E')

vaegt = M(:,1)
opladn = M(:,2)

%% a. Lineær regression
mdl = fitlm(vaegt, opladn)
% Med R^2 = 0.905 og R^2_adjusted = 0.889 forklarer modellen en meget
% stor del af variationen i data. Begge koefficienter er signifikant
% forskellige fra 0, så der er negativ korrelation i data (p-værdi =
% 0.00028102. At dømme ud fra regressionsanalysen alene, er det en god
% model.

%% b: Figur
figure(1)
scatter(vaegt,opladn)
lsline
title('Sammenhæng mellem dronens belastning og batteriets opladningsgrad');
xlabel('Dronens belastning (kg)');
ylabel('Batteriets opladningsgrad (%)');

% Figuren viser, at selv om regressionsanalysens statistikker var gode, så
% er det ikke en lineær sammenhæng imellem belastning og opladningsgrad.
% For små belastninger op til ca. et kg ser sammenhængen lineær ud, men så
% begynder belastningen at være hårdt ved batteriet.

%% c. 'Unormale' datapunkter
lev = mdl.Diagnostics.Leverage;     % hat diagonal
rst = mdl.Residuals.Studentized;    % R-Student
resultat = [M, lev, rst]            % Jeg samler det hele til en resultattabel
som table 6.42

k = 1 % Antal regressorvariable
n = size(M,1) % Antal observationer
lev_limit = 2*(k+1)/n

% Vi kan se i tabellen resultat, at kun det sidste punkt med belastningen 2 kg
% har studentiseret residual |rst|>3, nemlig 3.8379. Dette punkt er derfor
% en outlier.
% Det er også kun det sidste punkt med belastning på 2 kg, der har leverage
% over grænsen på lev_limit = 0.5. Dette punkt har lev = 0.5088.
% Da punktet både er outlier og løftestangspunkt, så er det et
% indflydelsespunkt.

%% d. Pizza på 630 gram

% Man kunne forsøge sig med at transformere data for at finde en lineær
% sammenhæng for hele datasættet, men da kurven er fint lineær indtil en
% belastning på 1 kg, og 0.63kg befinder sig på den lineære del, så laver
% jeg blot en lineær regression på de første fem datapunkter.

vaegt_red = vaegt(1:5,:)
opladn_red = opladn(1:5,:)
fitlm(vaegt_red, opladn_red)
```

```
figure(5)
scatter(vaegt_red, opladn_red)
lsline

% Resultatet ser pænt ud og regressionslinjen følger data.
% Modellen har en meget høj R^2 på 0.994 og R^2_adjusted på 0.992.
% Begge koefficienter er signifikant forskellige fra 0.
% Vi kan dermed beregne opladningsgraden efter en km flyvning med pizzaen
% på 0.63 kg:
opladn_630 = 79.2 - 9.2*0.63

% Resultatet er 73.4 pct.
% Havde vi brugt den oprindelige model, ville vi få 72.0 pct.:
opladn_630_oprindelig = 83.353 - 18.045*0.63
```

Opgave 4

Test af kvaliteten af fire typer batterier til dronen, hver med 6 gentagelser.

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



Batteritype 1, 3 og 4 har nogenlunde ensartet form på boksplottet (ensartet varians), mens batteritype 2 har meget større variation. Type 2 har den højeste median, efterfulgt af henholdsvis 3, 1 og 4.

b. *Undersøg i en variansanalyse med signifikansniveau 5%, om der er forskel på batterierne.*

Variansanalysen kan laves med MatLab funktionerne anova1 og anovan. Her er brugt anovan, så jeg også får residualer ud. Analysen har F = 2.4 og en tilhørende p-værdi på 0.098, som vist nedenfor. Det vil sige, at vi ikke kan afvise på 5% signifikansniveau, at alle 4 batterityper opfører sig ens (da 0.098 > 0.05). De observerede forskelle, der f.eks. ses i boksplottet er måske tilfældige.

Analysis of Variance

Source	Sum Sq.	d.f.	Mean Sq.	F	Prob>F
X1	51.5	3	17.1667	2.4	0.098
Error	143	20	7.15		
Total	194.5	23			

c. *Lav en parvis sammenligning af batterierne.*

MatLab kommandoen multcompare giver følgende resultat:
Type 1 er ikke forskellig fra nogen andre typer på 5% signifikansniveau
Type 2 og type 4 er forskellige

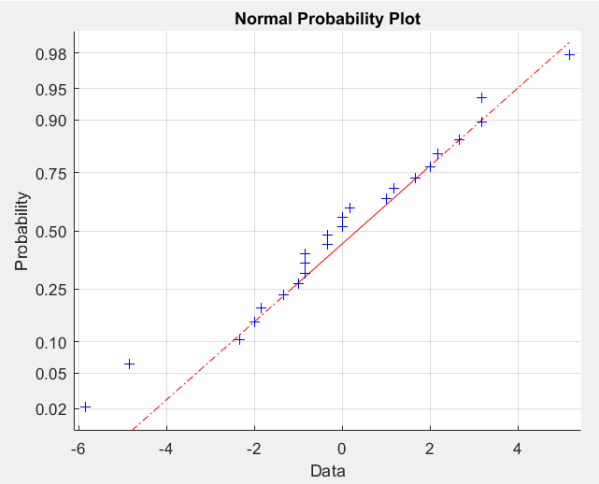
Type 3 og type 4 er forskellige
Type 4 er forskellig fra type 2 og 3.

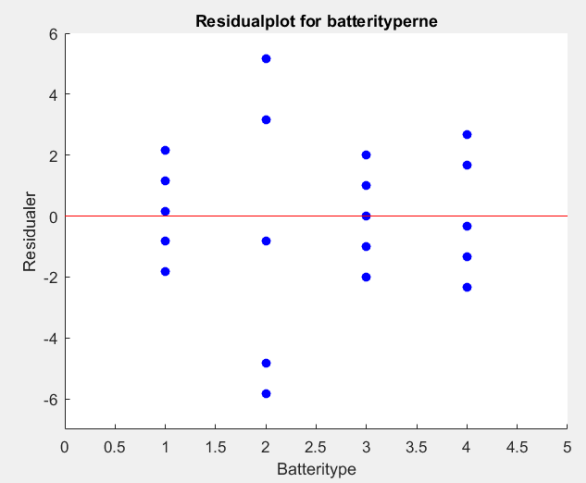
d. *Hvilket batteri vil du vælge som det bedste? Argumentér for dit valg.*

Valget står mellem type 2 og type 3, selv om de to ikke er signifikant forskellige. Type 2 har den højeste median, mens type 3 har det højeste gruppegennemsnit på 41.0, mod type 2's 40.8. Det er ikke meget forskel, så når alt kommer til alt vælger jeg type 3, fordi denne batteritype er mere ensartet end type 2, som man også kan se i det parallelle boksplot.

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

Det er en antagelse, 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 MatLab funktionen anovan.





I figuren til venstre ses normalfordelingsplottet. Det ses, at punkterne ligger nogenlunde på en ret linje. I figuren til højre ses residualplottet, hvor det er tydeligt, at residualerne er ensartede for batteritype 1, 3 og 4, men de er markant større for batteritype 2. Som vi så i boksplottet er der større variation i de målte opladningsgrader for type 2. Derfor er der også større variation i residualerne. Desværre må vi sige, at selv om antagelsen om normalfordelte residualer ser ud til at holde, så holder antagelsen om ensartet varians ikke. Derfor kan vi ikke stole på resultatet af vores variansanalyse.

MatLab kode til opgave 4

```
%% E2016 Opg 4: Variansanalyse af forskellige batteriers opladningsgrad
clear all; close all; clc;

M = xlsread('Data_M4STI1_2016E.xlsx', 'G:H')

batteri = M(:,1)
opladn  = M(:,2)

%% a
figure(1)
boxplot(opladn, batteri)
title('Opladningsgrad af fire batterityper');

% Batteritype 1, 3 og 4 har nogenlunde ensartet form på boksplottet, mens
% batteritype 2 har meget større variation. Type 2 har den højeste median,
% efterfulgt af henholdsvis 3, 1 og 4.


%% b
[p,table,stats] = anovan(opladn, batteri)
% Variansanalysen har F = 2.4 og en tilhørende p-værdi på 0.098. Det vil
% sige, at vi kan ikke afvise på 5% signifikansniveau, at alle 4
% batterityper opfører sig ens (da 0.098 > 0.05).


%% c
[c,m] = multcompare(stats,  'Alpha',0.05,  'CType','lsd')

% Type 1 er ikke forskellig fra nogen andre
% Type 2 og type 4 er forskellige
% Type 3 og type 4 er forskellige
% Type 4 er forskellig fra type 2 og 3.


%% d
% Valget står mellem type 2 og type 3, selv om de to ikke er signifikant
% forskellige. Type 2 har den højeste median, mens type 3 har det højeste
% gruppegennemsnit på 41.0, mod type 2's 40.8. Det er ikke meget forskel,
% så når alt kommer til alt vælger jeg type 3, fordi det er mere ensartet,
% som man også kan se i det parallelle boksplot.


%% e
% Det er en antagelse, 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)

figure(3)
hold on;
scatter(batteri, resid, 'b', 'filled')
```

```
plot([0; 5], [0; 0], 'r')   % vandret referencelinje gennem (0,0)
title('Residualplot for batterityperne');
xlabel('Batteritype');
ylabel('Residualer');
axis([0 5 -7 6])
hold off;
```