Ingeniørhøjskolen Aarhus Universitet – Maskinteknik
Eksamenstermin: Vinter 2021/ 2022
Prøve i: M4STI1 Statistik for Ingeniører
Dato: 07.01.2022 ULH

OPGAVE 1

a. Udfyld nedenstående tabel med antal kuvetter, idet nødvendige mellemregninger anføres.

	Maskine A	Maskine B	I alt
T Transparent kuvette	560 - 8 = 552	768 - 552 = 216	**768**
M Mat kuvette	**8**	32 - 8 = 24	800 - 768 = 32
I alt	0.70 · 800 = 560	800 - 560 = 240	**800**

b. Beregning sandsynlighederne for:

Den fremstillede kuvette er transparent, $P(\mathrm{T})$: $P(\mathrm{T}) = \frac{768}{800} = \mathbf{0.96}$

Kuvetten er fremstillet på maskine A og er transparent, $P(\mathrm{A} \cap \mathrm{T})$: $P(\mathrm{A} \cap \mathrm{T}) = \frac{552}{800} = \mathbf{0.69}$

Kuvetten er transparent og er fremstillet på maskine B, $P(\mathrm{T} \cap \mathrm{B})$: $P(\mathrm{T} \cap \mathrm{B}) = \frac{216}{800} = \mathbf{0.27}$

Kuvetten er fremstillet på maskine A og er mat, $P(\mathrm{A} \cap \mathrm{M})$: $P(\mathrm{A} \cap \mathrm{M}) = \frac{8}{800} = \mathbf{0.01}$

Kuvetten er mat og er fremstillet på maskine B, $P(\mathrm{M} \cap \mathrm{B})$: $P(\mathrm{M} \cap \mathrm{B}) = \frac{24}{800} = \mathbf{0.03}$

c. Beregning sandsynlighederne for:

Kuvetten er transparent, når den er fremstillet på maskine B, $P(\mathrm{T}|\mathrm{B})$:

$$P(\mathrm{T}|\mathrm{B}) = \frac{216}{240} = \mathbf{0.90}$$

Kuvetten er mat, når den er fremstillet på maskine B, $P(\mathrm{M}|\mathrm{B})$:

$P(\mathrm{M}|\mathrm{B}) = \frac{24}{240} = \mathbf{0.10}$ **eller** $P(\mathrm{M}|\mathrm{B}) = 1 - P(\mathrm{T}|\mathrm{B}) = 1 - 0.90 = \mathbf{0.10}$

Kuvetten er fremstillet på maskine A, når den er mat, $P(\mathrm{A}|\mathrm{M})$:

$$P(\mathrm{A}|\mathrm{M}) = \frac{8}{32} = \mathbf{0.25}$$

Kuvetten er mat, når den er fremstillet på maskine A, $P(\mathrm{M}|\mathrm{A})$:

$P(\mathrm{M}|\mathrm{A}) = \frac{8}{560} = \mathbf{0.014}$

OPGAVE 2

a. Man kan anvende en Poisson-fordeling, fordi man tæller antal støbefejl pr. m^2 og kender det forventede antal støbefejl pr. m^2 (λ).

Sandsynlighedsfunktionen (tæthedsfunktion) for Poisson-fordelingen er givet ved:

$$P(Y=y)=\begin{cases}\frac{\lambda^y}{y!}\cdot e^{-\lambda} & \text{for } y=0,1,2,\dots \text{ og } \lambda>0\\ 0 & \text{ellers}\end{cases} \quad \text{dvs.} \quad P(Y=y)=\begin{cases}\frac{4^y}{y!}\cdot e^{-4} & \text{for } y=0,1,2,\dots\\ 0 & \text{ellers}\end{cases}$$

b. Bestemmelse af fordelingens:
Middelværdi: $\mu=\lambda=\mathbf{4}$
Varians: $\sigma^2=\lambda=\mathbf{4}$
Standardafvigelse: $\sigma=\sqrt{\lambda}=\sqrt{4}=\mathbf{2}$

c. Bestem sandsynligheden for 3 støbefejl.

$$P(Y=3)=\frac{4^3}{3!}\cdot e^{-4}=\frac{64}{6}\cdot e^{-4}=\mathbf{0.1954}$$

I MATLAB: poisspdf(3,4) = 0.1954

d. Bestem sandsynligheden for at antallet af støbefejl er mindre eller lig med 6.

$$P(Y\le 6)=P(Y=0)+P(Y=1)+P(Y=2)+P(Y=3)+P(Y=4)+P(Y=5)+P(Y=6)=$$

0.0183 + 0.0733 + 0.1465 + 0.1954 + 0.1954 + 0.1563 + 0.1042 = **0.8893**

I MATLAB: poisscdf(6,4) = 0.8893

eller

```
x = 0:1:6
```

```
x = 1×6
     0     1     2     3     4     5     6
```

```
poisspdf(x, lambda)
```

```
ans = 1×6
    0.0183    0.0733    0.1465    0.1954    0.1954    0.1563    0.1042
```

e. Bestem sandsynligheden for at antallet af støbefejl er større eller lig med 7.

$$P(Y\ge 7)=1-P(Y\le 6)=1-0.8893=\mathbf{0.1107}$$

I MATLAB: $1-\text{poisscdf}(6,4)$ = 0.1107

f. Bestem sandsynligheden for at antallet af støbefejl ligger mellem 2 og 5.

$$P(2<Y<5)=P(Y=3)+P(Y=4)=0.1954+0.1954=\mathbf{0.3907}$$

$P(2<Y<5)=P(Y\le 4)-P(Y\le 2)=\text{poisscdf}(4,4)-\text{poisscdf}(2,4)$ = 0.6288 - 0.2381 = **0.3907**

OPGAVE 3

a. Beregning af sandsynlighederne:

Observationer:	Afdeling A	Afdeling B	I alt
Menu 1	33	60	93
Menu 2	9	10	19
I alt	42	70	112

Menu 1 vælges: $P(M1) = \frac{93}{112} = \mathbf{0.8304}$

Menu 2 vælges: $P(M2) = \frac{19}{112} = \mathbf{0.1696}$

eller $P(M2) = 1 - P(M1) = 1 - 0.8304 = \mathbf{0.1696}$

En ansat er fra afdeling A: $P(A) = \frac{42}{112} = \mathbf{0.3750}$

En ansat er fra afdeling B: $P(B) = \frac{70}{112} = \mathbf{0.6250}$

eller $P(B) = 1 - P(A) = 1 - 0.3750 = \mathbf{0.6250}$

b. Beregning af det forventede antal observationer, under antagelse af uafhængighed mellem menuvalg og afdeling:
Antal observationer i alt: $n = 112$

$E_{M1,A} = n \cdot P(M1) \cdot P(A) = 112 \cdot 0.8304 \cdot 0.3750 = 34.88 \approx \mathbf{35}$
$E_{M1,B} = n \cdot P(M1) \cdot P(B) = 112 \cdot 0.8304 \cdot 0.6250 = 58.13 \approx \mathbf{58}$
$E_{M2,A} = n \cdot P(M2) \cdot P(A) = 112 \cdot 0.1696 \cdot 0.3750 = 7.123 \approx \mathbf{7}$
$E_{M1,B} = n \cdot P(M2) \cdot P(B) = 112 \cdot 0.1696 \cdot 0.6250 = 11.87 \approx \mathbf{12}$

Forventede observationer:	Afdeling A	Afdeling B	I alt
Menu 1	35	58	93
Menu 2	7	12	19
I alt	42	70	112

c. Nulhypotese og alternativ hypotese for hypotesetesten.

H_0: valg af menu er uafhængig af afdeling
H_1: valg af menu er ikke uafhængig af afdeling

d. Formel for teststørrelsen (teststatistikken), og angivelse af hvilken fordeling den følger.

Teststørrelsen er beregnes ved: $\chi_0^2 = \sum_{i=1}^{r} \sum_{j}^{c} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$

og er χ^2-fordelt med (r-1)(c-1) frihedsgrader.

O_{ij}: angiver observation i tabellens i'te række og j'te kolonne.

E_{ij}: angiver forventet observation i tabellens i'te række og j'te kolonne.

OPGAVE 3 fortsat

r er antal rækker, og *c* er antal kolonner.

Da vi har en 2x2 kontingenstabel, så er antallet af frihedsgrader: $df = (2-1)(2-1) = \mathbf{1}$

e. Bestemmelse den kritiske værdi og angivelse af det kritiske område for testen, når der vælges et signifikansniveau på 5%.

Signifikansniveau er $\alpha = 0{,}05$

Da vi har en 2x2 kontingenstabel, så er antallet af frihedsgrader: $df = (2-1)(2-1) = \mathbf{1}$

Værdien af den kritiske værdi $\chi^2_{\alpha,df} = \chi^2_{0.05,1}$ findes vha. MATLAB:

$\chi^2_{0.05,1} = chi2inv(1-\alpha, df) = chi2inv(1-0.05,1) = chi2inv(0.95,1) = \mathbf{3.8415}$

Dvs. det kritiske område: H_0 afvises, hvis $\boldsymbol{\chi^2_0 > \chi^2_{0.05,1} = 3.8415}$

f. Beregning af teststørrelsens (teststatistikkens) værdi. Mellemregninger skal fremgå.

$$\chi^2_0 = \frac{(33-34.88)^2}{34.88} + \frac{(60-58.13)^2}{58.13} + \frac{(9-7.123)^2}{7.123} + \frac{(10-11.87)^2}{11.87} = 0.9507 \approx \mathbf{0.95}$$

g. Konklusion på hypotesetesten.

Teststørrelsen (teststatistikken) $\chi^2_0 = 0.9507$ er mindre end den kritiske værdi $\chi^2_{0.05,1} = 3.8415$

Dvs. H_0-hypotesen kan ikke forkastes. Medarbejderens valg af frokostmenu er uafhængig af den afdeling, hvor medarbejderen er ansat.

Dette spørges der ikke om:

- *p*-værdien kan findes vha. MATLAB:

  $p - værdi = 1 - chi2cdf(0.9507,1) = 0.3295$

- Kontrol med funktionen chi2cont:

```
[h, pval, chi2_0] = chi2cont(O, alpha)
```

```
h = 0
pval = 0.3295
chi2_0 = 0.9508
```

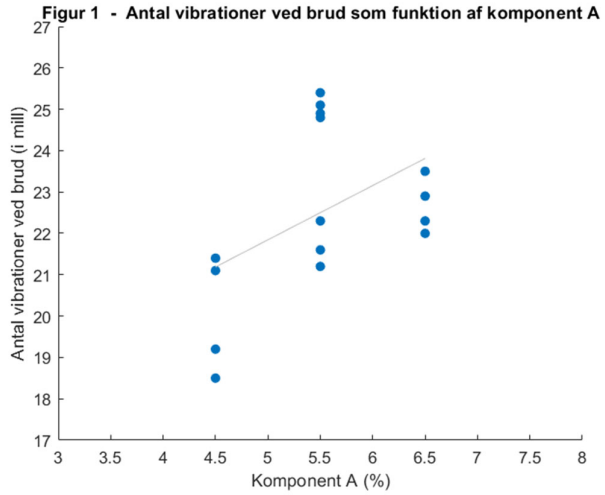
h angiver om nulhypotesen kan forkastes (0: nej, 1: ja)
pval er *p*-værdien for chi-i-anden testen
chi2_0 er teststatistikken

OPGAVE 4

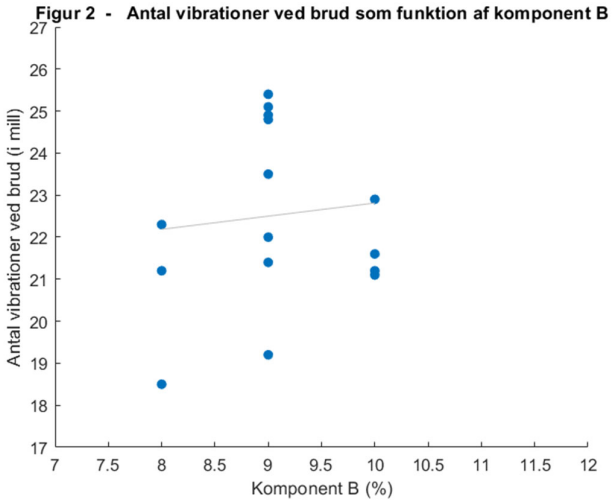
Et metal tilsættes to komponenter, komponent A og komponent B, for at fremstille en metallegering med forbedret resistens mod træthedsbrud. Mængden af de tilsatte komponenter angives i masseprocent (%).
Der udføres vibrationstest i et standardudstyr, og antallet af vibrationer ved brud måles i enheden million (M) vibrationer.
Der udføres en forsøgsrække med 17 målinger. De målte data er indført i tabel, hvor komponent A betegnes A, komponent B betegnes B, og antal af vibrationer ved brud betegnes V.

a. I MATLAB laves der laves et scatterplot af data:
med komponent A som regressorvariabel og antal vibrationer, V, som responsvariabel, figur 1
med komponent B som regressorvariabel og antal vibrationer, V, som responsvariabel, figur 2

Scatterplot:



Scatterplot viser en del variation af data, men generelt vokser antal vibrationer, V, når mængden af komponent A vokser. Hermed er sammenhængen mellem komponent A og antal vibrationer, V, positivt korreleret.



Scatterplot viser en del variation af data. Der er en sammenhæng mellem komponent B og antal vibrationer, V, men den er ikke lineær.
Antallet af vibrationer er højt ved 9 % af komponent B men lavere ved 8 % og 10 % af komponent B.
En polynomial sammenhæng kan muligvis være et bedre bud.

OPGAVE 4 fortsat

b. Multipel lineær regressionsanalyse, der beskriver antallet af vibrationer, V, som funktion af tilsat mængde af komponent A og af komponent B,
og opskrivning regressionsligningen: $V = b_0 + b_1 A + b_2 B$

Der udføres en multipel lineær regression i MATLAB vha: mdl = fitlm([A ,B], V)

Og der fås følgende output:

```
mdl =
Linear regression model:
    y ~ 1 + x1 + x2

Estimated Coefficients:
                   Estimate      SE       tStat      pValue
                   ________    _______    ______    ________
    (Intercept)     12.469       7.35     1.6964     0.11192
    x1              1.3125     0.69538     1.8875    0.080004
    x2              0.3125     0.69538     0.4494     0.66002

Number of observations: 17, Error degrees of freedom: 14
Root Mean Squared Error: 1.97
R-squared: 0.212,  Adjusted R-Squared: 0.0993
F-statistic vs. constant model: 1.88, p-value = 0.189
```

Dermed fås regressionsligningen:

$$\boldsymbol{V = 12.469 + 1.3125 \cdot A + 0.3125 \cdot B}$$

c. Vurdering om modellen beskriver observationerne godt ud fra regressionsanalysens statistikker (f.eks. R^2, F og p-værdi). Der anvendes et signifikansniveau, $\alpha = 0.05$ ved vurderingen.

$R^2 = 0.212$, dvs kun 21.2% af variationen forklares af modellen.

F = 1.88 har en p-værdi på $0.189 > 0.05$. Med en nulhypotese, H_0, der siger, at data er ukorrelerede, en alternativ hypotese, H_a, der siger, at data er korrelerede, og med et signifikansniveau, α, på 5%, så kan H_0 ikke forkastes. Dvs. data er ukorrelerede, begge koefficienter b_1 og b_2 kan være 0 samtidigt.

Koefficienterne $b_1 = 1.3125$ og $b_2 = 0.3125$ har en p-værdi på henholdsvis 0.080 og 0.660, som begge er større end 0.05. Med en nulhypotese, H_0, der siger, at koefficienten er 0, en alternativ hypotese, H_a, der siger, at koefficienten er forskellig fra 0, og med et signifikansniveau, α, på 5%, så kan H_0 ikke forkastes. b_1 og b_2 er ikke signifikant forskellige fra 0. (Skæring med y-aksen b_0=12.469 har en p-værdi på 0.112 > 0.05 er derfor heller ikke signifikant forskellig fra 0).
Alt i alt beskriver modellen data dårligt.

OPGAVE 4 fortsat

d. Udvidelse af modellen ved inddragelse af kvadratled og interaktionsled:

$$V = b_0 + b_1 A + b_2 B + b_{11} A^2 + b_{22} B^2 + b_{12} AB$$

hvor b_0, b_1, b_2, b_{11}, b_{22} og b_{12} er konstanter.

Der udføres en multipel lineær regression i MATLAB, hvor der anvendes Wilkinson notation:

```
mdl1 = fitlm([A, B], V, 'y ~ x1 + x2 + x1^2 + x2^2 +  x1:x2 ')
```

Og der fås følgende output:

```
mdl1 =
Linear regression model:
    y ~ 1 + x1*x2 + x1^2 + x2^2

Estimated Coefficients:
                   Estimate      SE        tStat      pValue
                   ________    _______    _______    _________

    (Intercept)    -233.08      59.593    -3.9112    0.0024292
    x1              28.232      8.6834     3.2513    0.0077177
    x2              38.849      11.394     3.4096    0.0058295
    x1:x2             -0.5     0.62004    -0.8064       0.4371
    x1^2           -2.0382      0.6035    -3.3772    0.0061728
    x2^2           -1.9882      0.6035    -3.2944     0.007149

Number of observations: 17, Error degrees of freedom: 11
Root Mean Squared Error: 1.24
R-squared: 0.754,  Adjusted R-Squared: 0.642
F-statistic vs. constant model: 6.74, p-value = 0.00416
```

Kun en koefficient er ikke signifikant forskellig fra 0, da den har en *p*-værdi større end 0.05. Det er $b_{12} = 0.5$. b_{12} har en *p*-værdi = 0.437 > 0.05. Jeg vælger derfor at fjerne interaktionsleddet med koefficienten b_{12} fra modellen.

Der udføres en ny multipel lineær regression i MATLAB:

```
mdl2 = fitlm([A, B], V, 'y ~ x1 + x2 + x1^2 + x2^2')
```

Og der fås følgende output:

OPGAVE 4d fortsat

```
mdl2 =
Linear regression model:
    y ~ 1 + x1 + x2 + x1^2 + x2^2

Estimated Coefficients:
                   Estimate      SE        tStat      pValue
                   ________    _______    _______    _________

    (Intercept)    -208.33      50.332    -4.1392    0.0013729
    x1              23.732      6.5553     3.6203    0.0035126
    x2              36.099      10.712     3.3699    0.0055711
    x1^2           -2.0382     0.59464    -3.4275    0.0050083
    x2^2           -1.9882     0.59464    -3.3435    0.0058506

Number of observations: 17, Error degrees of freedom: 12
Root Mean Squared Error: 1.22
R-squared: 0.739,  Adjusted R-Squared: 0.652
F-statistic vs. constant model: 8.51, p-value = 0.00171
```

Nu er alle koefficienter for hældningskoefficienterne signifikant forskellige fra 0. De har alle en *p*-værdi som er mindre end 0.05. (Det ses også at skæring med *y*-aksen er forskellig fra 0).

$F = 8.51$ har en *p*-værdi på $0.0017 < 0.05$. Dvs. data er korrelerede.

$R^2 = 0.739$ og er kun reduceret lidt fra den fulde udvidelse af modellen, hvor $R^2 = 0.754$.

Modellen forklarer altså 74% af variationen. Scatterplots fra opgave 4a figur 1 og figur 2 viser også en del variation i data.

R^2_{adj} som justerer for antal parametre i modellen og dermed mindsker risiko for overfitning er:
$R^2_{adj} = 0.652$ og er øget lidt fra den fulde udvidelse af modellen, hvor $R^2_{adj} = 0.642$.

Jeg vælger denne model.

e. Opskrivning af ligningen for den foretrukne model: $V = b_0 + b_1 A + b_2 B + b_{11} A^2 + b_{22} B^2$ er:

$$\boldsymbol{V = -208.33 + 23.732 \cdot A + 36.099 \cdot B - 2.0382 \cdot A^2 - 1.9882 \cdot B^2}$$

f. Undersøgelse for "unormale" datapunkter, dvs. løftestangspunkter, outliers og indflydelsespunkter. Svaret begrundes.

For at undersøge for "unormale" punkter beregnes hat-diagonaler og studentiserede residualer, (*rst*), i MATLAB:

OPGAVE 4f fortsat

```
lev = mdl2.Diagnostics.Leverage;
rst = mdl2.Residuals.Studentized;
```

og resultaterne samles i en tabel.

Tabel fra MATLAB:

resultat = 17×6 table

	Nr	Komponent A	Komponent B	Antal_vibrationer	Hat diagonal Lev	Studentiseret residual rst
1	1	4.5000	10	21.1000	0.4342	2.1495
2	2	4.5000	8	18.5000	0.4342	-0.2543
3	3	4.5000	9	19.2000	0.3092	-2.0415
4	4	4.5000	9	21.4000	0.3092	0.3373
5	5	5.5000	8	22.3000	0.3092	0.1945
6	6	5.5000	9	25.1000	0.1579	0.6124
7	7	5.5000	9	24.8000	0.1579	0.3479
8	8	5.5000	9	24.9000	0.1579	0.4351
9	9	5.5000	9	25.1000	0.1579	0.6124
10	10	5.5000	8	21.2000	0.3092	-0.8715
11	11	5.5000	10	21.2000	0.3092	-1.5878
12	12	5.5000	10	21.6000	0.3092	-1.1128
13	13	5.5000	9	25.4000	0.1579	0.8887
14	14	6.5000	9	22.0000	0.3092	-1.7876
15	15	6.5000	10	22.9000	0.4342	0.9852
16	16	6.5000	8	22.3000	0.4342	1.0149
17	17	6.5000	9	23.5000	0.3092	-0.1596

Løftestangspunkter (leverage) er ”unormale” værdier i x-retningen, og det måles med hat-diagonalen. Grænsen beregnes vha. formlen:

$\text{lev}_{\text{limit}} = \frac{2(c+1)}{n}$, hvor c er antal regressorvariable, og n er antal observationer.
Der er 4 regressorvariable i den valgte model: A, B, A^2 og B^2.
Dvs.

$$\text{lev}_{\text{limit}} = \frac{2(4+1)}{17} = 0.5882$$

I tabellen ses ingen punkter med hat diagonal > 0.5882. **Dvs. der er ingen løftestangspunkter.**

Outliers er ”unormale” værdier i y-retningen, og det måles på den numeriske værdi af det studentiserede residual, *rst*. Grænsen er: $|rst| > 3$.

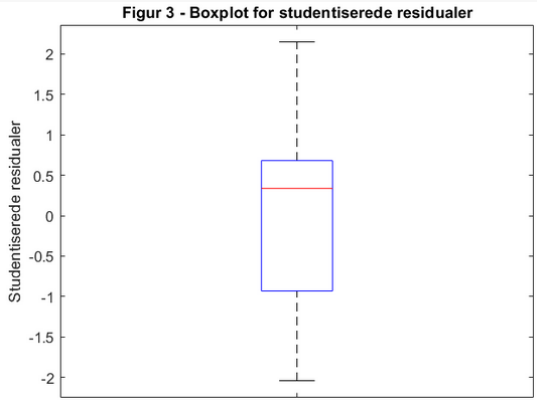
I tabellen ses ingen punkter med $|rst| > 3$. **Dvs. der er ingen outliers.**

Et punkt skal være både et løftestangspunkt og outlier for at være et indflydelsespunkt. Der er derfor **ingen indflydelsespunkter**.

OPGAVE 4 fortsat

g. Der laves et box-plot, et histogram og et normalfordelingsplot for de studentiserede residualer, og plottene kommenteres.

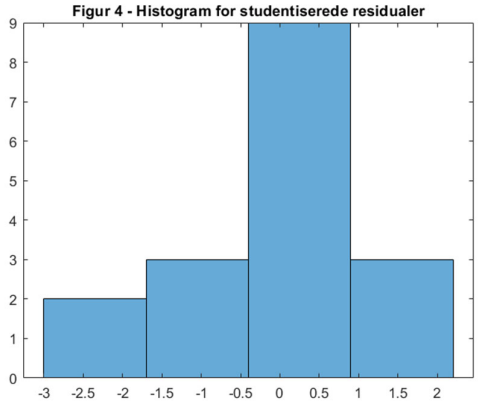
Boxplot `boxplot(rst)`



Boxplottet har en median på 0.3373.
Det interkvartile range ligger ikke symmetrisk omkring medianen, den nedre del er væsentlig bredere end den øvre del. Dvs. fordelingen af data er venstre skæv. Kostene er nogenlunde lige lange. Der er ingen outliers.
Dvs. data, de studentiserede residualer, kommer fra en nogenlunde pæn fordeling.

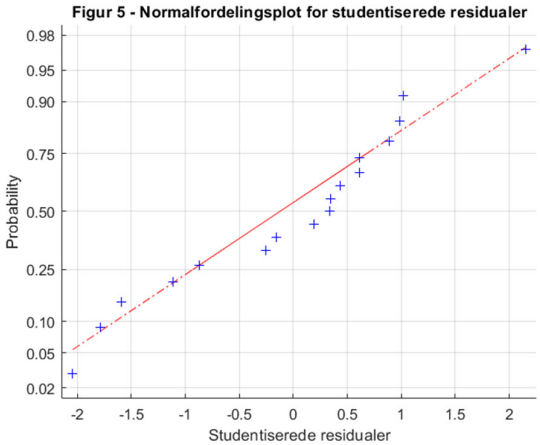
I MATLAB: Median = median(rst) = 0.3373

Histogram: `histogram(rst,4)`



Histogrammet viser, at formen er venstre skæv, der er en top, og der er hurtigt uddøende haler. Dvs. data, de studentiserede residualer, kommer fra en nogenlunde pæn fordeling.

Normalfordelingsplot `normplot(rst)`



Normalfordelingsplottet viser en tendens til S-formet kurve. Dvs. de studentiserede residualer ikke følger en normalfordeling.

De tre plot viser, at de studentiserede residualer kommer fra en nogenlunde pæn fordeling. Den er venstre skæv, og den følger ikke en normalfordeling.