M4STI1 2021F Opgave 2 - Hypotesetest af komponenttykkelse fra to producenter

```
clc; clear; close all;

% Indlæsning af data:
D = xlsread('Data_M4STI1_2021F.xlsx', 'C:D');
```

a. Hypoteser

H0: mu_A - mu_B = delta0 = 0

Ha: mu_A - mu_B != 0

Her er mu_A og mu_B populationsmiddelværdi for tykkelsen af komponenten fra hhv. producent A og B. delta0 er den formodede forskel, som her er 0. '!=' betyder 'forskellig fra', så vi vælger en tosidet test, da vi skal undersøge, om der er forskel på tykkelsen fra de to producenter.

```
delta0 = 0;
```

b. Formel for teststørrelsen

Teststørrelsen for en hypotesetest af middelværdi for to uafhængige stikprøver er:

t0 = (y_A_streg - y_B_streg - delta0) / (s_p*sqrt(1/n_A + 1/n_B))

Teststørrelsen t0 er t-fordelt med n_A + n_B - 2 frihedsgrader

I formlen er y_A_streg og y_B_streg stikprøvegennemsnit for stikprøverne fra hhv. producent A og B.

delta0 er den formodede forskel på de to populationers middelværdi. Den er her 0.

s_p er puljet standardafvigelse, dvs. et estimat for de to populationers fælles standardafvigelse.

n_A og n_B er stikprøvestørrelserne for de to stikprøver.

c. Kritisk region

```
alfa = 0.05;                    % 5 % signifikansniveau
n_A = 12;                       % Stikprøvestørrelse fra producent A
n_B = 15;                       % Stikprøvestørrelse fra producent B
df = n_A + n_B - 2              % Frihedsgrader
```

```
df = 25
```

```
t_alfahalve = tinv(1-alfa/2, df)    % t_alfahalve = 2.0595
```

```
t_alfahalve = 2.0595
```

Vi forkaster nulhypotesen, hvis teststørrelsen t0 er større end **2.0595** eller mindre end **-2.0595**.

d. Beregning af teststørrelsens værdi

```
Producent = D(:,1);
Tykkelse  = D(:,2);
Tykkelse_A = Tykkelse(1:n_A)            % Stikprøven fra producent A
```

```
Tykkelse_A = 12×1
   22.3200
   21.7100
   21.5400
   22.0900
   22.0300
   22.0800
   21.8500
   21.6400
   22.3300
   21.7900
      ⋮
```

```
Tykkelse_B = Tykkelse(n_A+1:n_A+n_B)    % Stikprøven fra producent B
```

```
Tykkelse_B = 15×1
   22.2400
   22.4900
   22.5000
   22.4700
   22.3600
   22.1000
   22.4400
   22.6500
   22.6800
   22.5400
      ⋮
```

```
y_A_streg = mean(Tykkelse_A)            % Stikprøvegennemsnit fra producent A
```

```
y_A_streg = 21.9667
```

```
y_B_streg = mean(Tykkelse_B)            % Stikprøvegennemsnit fra producent B
```

```
y_B_streg = 22.4147
```

```
var_A = var(Tykkelse_A)                 % Stikprøvevarians for stikprøve A
```

```
var_A = 0.0722
```

```
var_B = var(Tykkelse_B)                 % Stikprøvevarians for stikprøve B
```

```
var_B = 0.0295
```

```
% Den puljede varians var_p er et vægtet gennemsnit af de to stikprøvers varianser:
var_p = ((n_A - 1)*var_A + (n_B - 1)*var_B) / (n_A + n_B - 2)
```

```
var_p = 0.0483
```

```
s_p = sqrt(var_p)
```

```
s_p = 0.2197
```

```
% Teststørrelsens værdi:
t0 = (y_A_streg - y_B_streg - delta0) / (s_p*sqrt(1/n_A + 1/n_B))
```

```
t0 = -5.2647
```

Teststørrelsens værdi er t0 = **-5.2647**

e. Konklusion

Teststørrelsens værdi på t0 = -5.2647 er mindre end den nedre, kritiske grænse på -t_alfahalve = -2.0595. Derfor forkaster vi nulhypotesen. Der er altså forskel på tykkelsen fra de to producenter.

```
pvalue = 2*tcdf(t0, df)
```

```
pvalue = 1.8815e-05
```

P-værdien på 0.0000188 viser også, at der er meget lille sandsynlighed for, at nulhypotesen er sand.

f. Antagelser

Vi har antaget tre ting:

1. De to stikprøver er uafhængige
2. De to stikprøver kommer fra en 'pæn' fordeling, så den centrale grænseværdisætning holder
3. De to stikprøver kommer fra populationer med samme varians

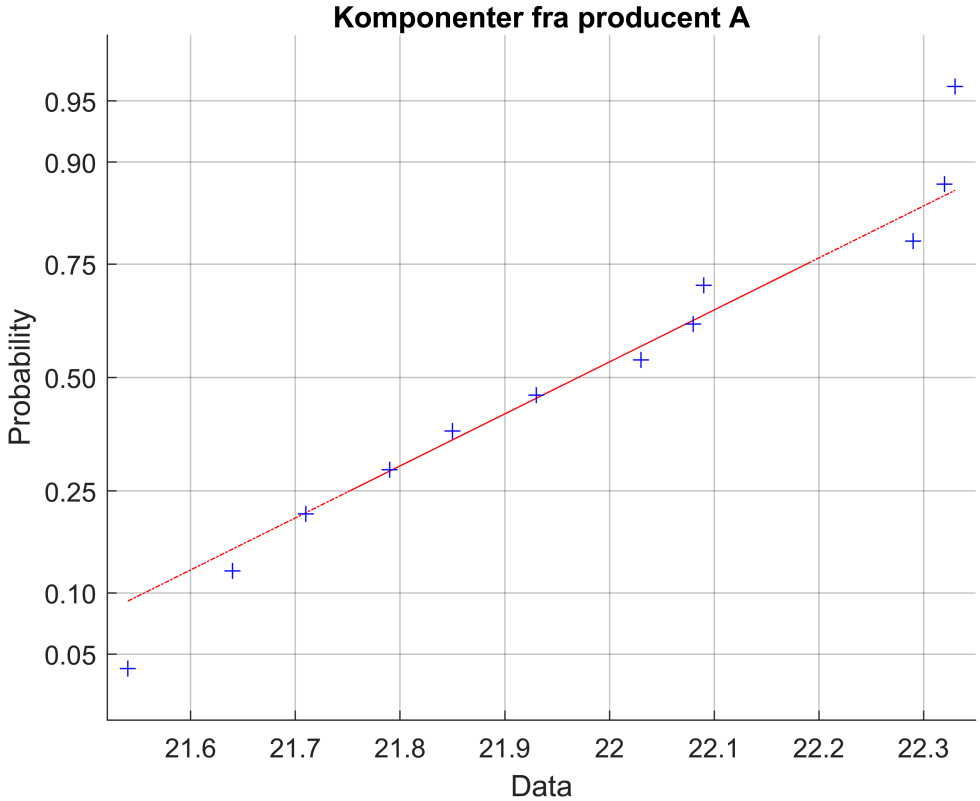
1. Uafhængighed

De målte komponenter kommer fra forskellige producenter, så vi må antage, at der ikke er sammenhæng mellem de enkelte observationer.

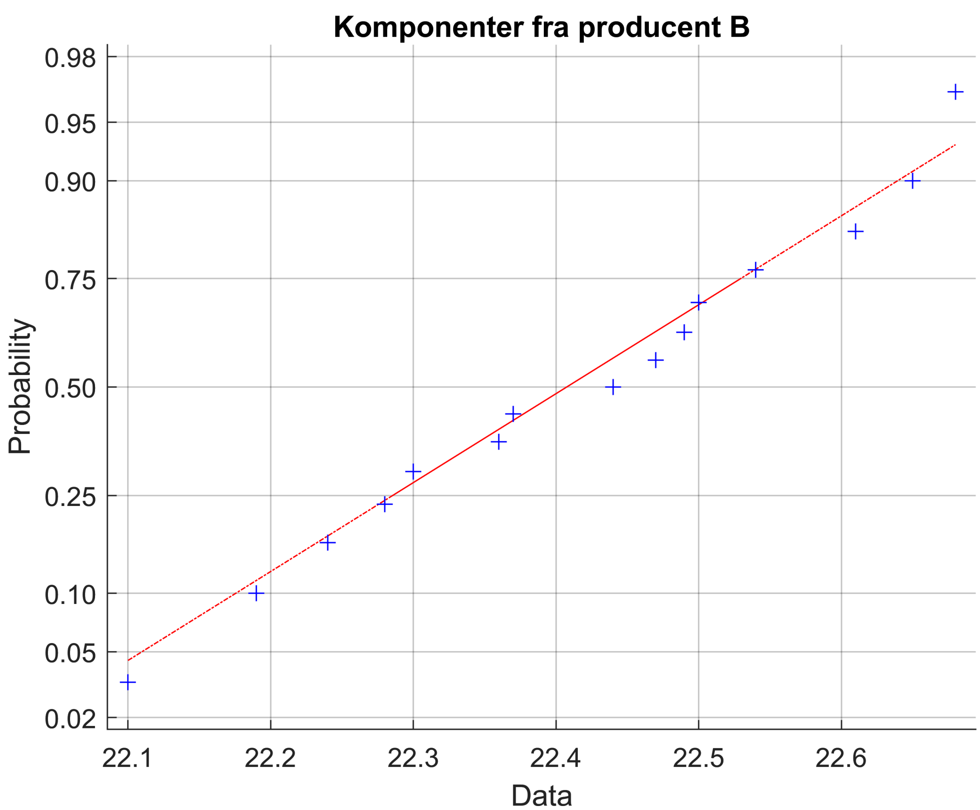
2. Den centrale grænseværdisætning

Vi kan teste dette med normalfordelingsplots

```
normplot(Tykkelse_A)
title('Komponenter fra producent A')
```



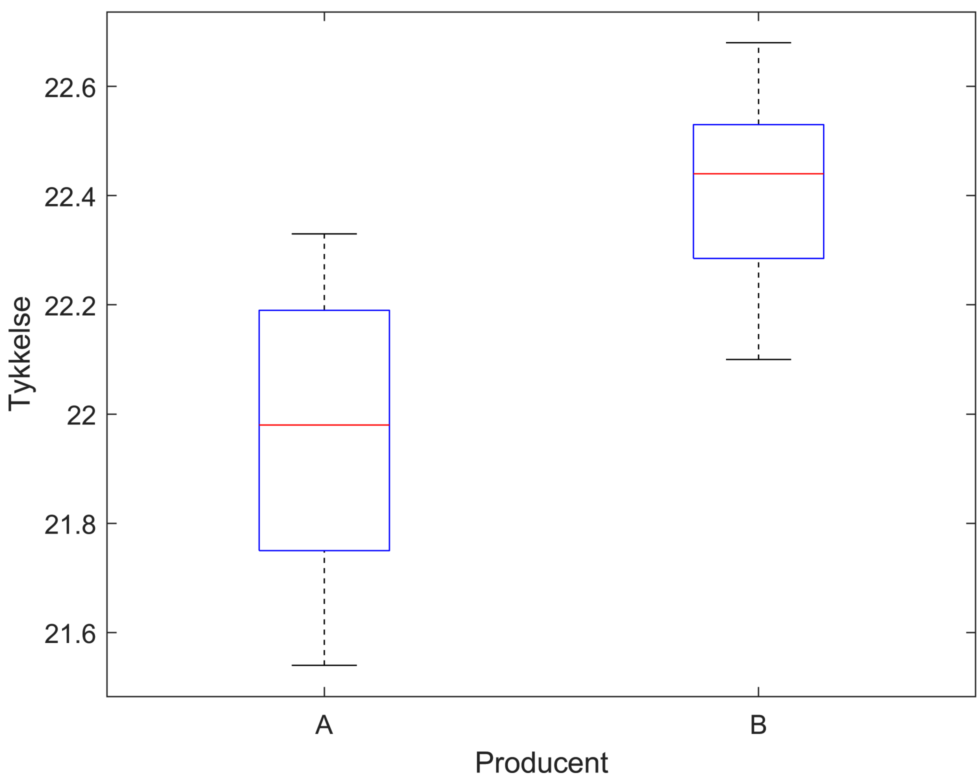

```
normplot(Tykkelse_B)
title('Komponenter fra producent B')
```



De to normalfordelingsplots ligger nogenlunde pænt på en linje, så antagelsen holder

3. Samme populationsvarians

```
boxplot(Tykkelse, Producent, 'Labels',{'A','B'})
xlabel('Producent')
ylabel('Tykkelse')
```



Det kan godt se ud til, at der er større spredning på data fra producent A end fra producent B. Både kassens højde (IQR) og afstanden mellem kostenes ender er større for boksplot A end for B.

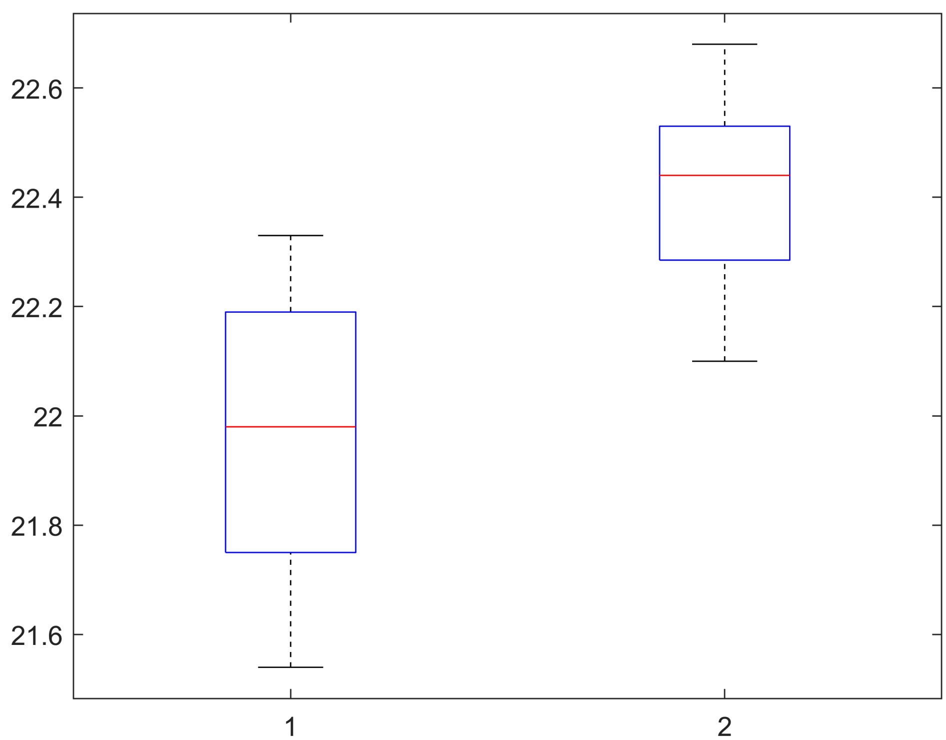
Vi kan undersøge med en Bartlett's test, om der er signifikant forskel på de to populationers varianser, således at forskellene vi ser i boksplottet bare skyldes tilfældigheder.

```
vartestn(Tykkelse, Producent)
```

Group Summary Table

Group	Count	Mean	Std Dev
1	12	21.9667	0.26871
2	15	22.4147	0.17167
Pooled	27	22.2156	0.21971

Bartlett's statistic	2.38318
Degrees of freedom	1
p-value	0.12265




```
ans = 0.1226
```

Testen viser heldigvis, at det ikke kan afvises, at populationerne har samme varians. Bartlett's testen har nulhypotesen at de to populationers varianser er ens. Teststørrelsens værdi er 2.38 og den tilhørende p-værdi er 0.123. Hvis de to populationers varianser er ens, vil vi altså se forskelle som i boksplottet (eller endnu større) i over 12 % af tilfældene. Det er ikke usandsynligt.

Vi kan altså konkludere at antagelserne holder, så der er forskel på komponenttykkelserne fra de to producenter.