Ingeniørhøjskolen Aarhus Universitet – Maskinteknik
Eksamenstermin: Sommer 2022
Prøve i: M4STI1 & MH4STI Statistik for Ingeniører
Dato: 08.06.2022 ULH

OPGAVE 1

a. Udfyld antal i nedenstående tabel, idet nødvendige mellemregninger medtages.

	Maskine A	Maskine B	I alt
In Intakt metalskive	600 - 12 = <u>588</u>	**291**	900 - 21 = <u>879</u>
D Defekt metalskive	0.02 · 600 = <u>12</u>	300 - 291 = <u>9</u>	12 + 9 = <u>21</u>
I alt	**600**	900 - 600 = <u>300</u>	**900**

b. Beregning sandsynlighederne for:

Metalskiven er fremstillet på maskine A, P(A): $P(\mathrm{A}) = \frac{600}{900} = \mathbf{0.6667}$

Metalskiven er fremstillet på maskine B, P(B):
$P(\mathrm{B}) = \frac{300}{900} = \mathbf{0.3333}$ **eller** $P(\mathrm{B}) = 1 - P(\mathrm{A}) = 1 - 0.6667 = \mathbf{0.3333}$

Metalskiven er fremstillet på maskine A og er intakt, P(A∩In): $P(\mathrm{A} \cap \mathrm{In}) = \frac{588}{900} = \mathbf{0.6533}$

Metalskiven er fremstillet på maskine B og er intakt, P(B∩In): $P(\mathrm{B} \cap \mathrm{In}) = \frac{291}{900} = \mathbf{0.3233}$

c. Beregning sandsynlighederne for:

Metalskiven er intakt: $P(\mathrm{In}) = \frac{879}{900} = \mathbf{0.9767}$

Metalskiven er defekt:
$P(\mathrm{D}) = \frac{21}{900} = \mathbf{0.02333}$ **eller** $P(\mathrm{D}) = 1 - P(\mathrm{In}) = 1 - 0.9767 = \mathbf{0.02333}$

Metalskiven er intakt, når den er fremstillet på maskine B, $P(\mathrm{In}|\mathrm{B})$: $P(\mathrm{In}|\mathrm{B}) = \frac{291}{300} = \mathbf{0.9700}$

Metalskiven er defekt, når den er fremstillet på maskine B, $P(\mathrm{D}|\mathrm{B})$:
$P(\mathrm{D}|\mathrm{B}) = \frac{9}{300} = \mathbf{0.03000}$ **eller** $P(\mathrm{D}|\mathrm{B}) = 1 - P(\mathrm{In}|\mathrm{B}) = 1 - 0.9700 = \mathbf{0.03000}$

d. Er der uafhængighed mellem hændelsen: metalskiven er intakt, og hændelsen: metalskiven er fremstillet på maskine B. Svaret skal begrundes.

Hændelserne In og B er uafhængige hvis og kun hvis $P(\mathrm{In}|\mathrm{B}) = P(\mathrm{In})$
$P(\mathrm{In}|\mathrm{B}) = 0.9700 \neq P(\mathrm{In}) = 0.9767,$ så NEJ der er ikke uafhængighed mellem de to hændelser.

Eller
Hændelserne In og B er uafhængige hvis og kun hvis $P(\mathrm{B} \cap \mathrm{In}) = P(\mathrm{B}) \cdot \mathrm{P(In)}$
$P(\mathrm{B} \cap \mathrm{In}) = 0.3233 \neq P(\mathrm{B}) \cdot \mathrm{P(In)} = 0.3333 \cdot 0.9767 = 0.3256$, så NEJ der er ikke uafhængighed mellem de to hændelser.

OPGAVE 2

a. Man kan anvende en eksponential-fordeling, fordi den beskriver tiden mellem hver udskiftning af bor (borets levetid).
Den gennemsnitlig tid mellem hver udskiftning (før der skal ske en udskiftning) af bor er $\mu = 3$ timer.

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

$$P(Y = y) = \begin{cases} \lambda \cdot e^{-\lambda y} & \text{for } y > 0, \ldots \text{og } \lambda > 0 \\ 0 & \text{ellers} \end{cases}$$

b. Bestemmelse af det gennemsnitlige antal udskiftning af bor pr. time:
$\lambda = \mathbf{\frac{1}{3}} = 0.3333$

c. Bestemmelse af fordelingens:
Middelværdi: $\mu = \frac{1}{\lambda} = \frac{1}{\frac{1}{3}} = \mathbf{3}$
Varians: $\sigma^2 = \frac{1}{\lambda^2} = \frac{1}{\left(\frac{1}{3}\right)^2} = \mathbf{9}$
Standardafvigelse: $\sigma = \frac{1}{\lambda} = \mathbf{3}$

d. Bestemmelse af sandsynligheden for at boret skal udskiftes, når det har været i brug i 2 timer:
$P(Y \leq 2) = \mathbf{0.4866}$

I MATLAB: expcdf(2,3) = 0.4866

e. Bestemmelse af sandsynligheden for at der går mere end 6 timer før boret skal udskiftes:
$P(Y \geq 6) = 1 - P(Y \leq 6) = 1 - 0.8647 = \mathbf{0.1353}$

I MATLAB: $1 - \text{expcdf}(6{,}3) = 1 - 0.8647 = 0.1353$

f. Bestemmelse sandsynligheden for at der går mellem 2 og 4 timer før boret skal udskiftes:

$P(2 \leq Y \leq 4) = P(Y \leq 4) - P(2 \leq Y) = \text{expcdf}(4{,}3) - \text{expcdf}(2{,}3)$ = 0.7364 - 0.4866 = **0.2498**

OPGAVE 3

En virksomhed fremstiller powerbanks til opladning af blandt andet mobiltelefoner. Varedeklarationen for virksomhedens powerbanks angiver, at en powebank har en kapacitet på 30Ah (ampere timer). Virksomheden ønsker, at undersøge om middelkapaciteten af de fremstillede powerbanks afviger fra de ønskede 30 Ah.

Der udtages en tilfældig stikprøve på 20 powerbanks. De målte kapaciteter i Ah er indført i nedenstående tabel:

30.0	30.4	30.3	29.6	29.6	29.2	30.2	28.6	29.7	30.5
29.4	29.9	29.8	30.4	30.0	29.5	28.9	31.1	30.2	29.5

Opgaven drejer sig om at undersøge ved hjælp af en hypotesetest, om middelværdien af kapaciteten for de fremstillede powerbanks afviger fra de ønskede 30 Ah, når der vælges et signifikansniveau på 5%.

a. Nulhypotese og alternativ hypotese for hypotesetesten.

$H_0: \mu = 30$
$H_a: \mu \neq 30$

Der vælges en to-sidet test, da kapaciteten for de fremstillede powerbanks ikke må afvige fra de ønskede 30 Ah.

b. Formel for teststørrelsen (teststatistikken), og angiv hvilken fordeling den følger.

Da populationsvariansen ikke er kendt, så bestemmes standardafvigelsen, *s*, ud fra stikprøven.

Teststørrelsen, $t = \frac{\bar{y}-\mu}{s/\sqrt{n}}$ som er *t*-fordelt med *n*-1frihedsgrader

c. Bestemmelse af det kritiske område for testen.

Signifikansniveau er $\alpha = 0.05$

Der testes på $n = 20$ powerbanks og antal frihedsgrader er: $df = n - 1 = 20 - 1 = 19$

Testen er to-sidet, så H_0 afvises, hvis $|t| > t_{n-1,\frac{\alpha}{2}}$

Værdien af den kritiske værdi $t_{n-1,\frac{\alpha}{2}} = t_{19,0.025}$ findes vha. MATLAB:

$t_{19,0.025} = tinv(1\text{-}\frac{\alpha}{2}\text{,n-1}) = tinv(1 - 0.025{,}20 - 1) = tinv(0.975{,}19) = 2.0930$

Dvs. det kritiske område: H_0 afvises, hvis $\boldsymbol{|t| > t_{19,0.025} = 2.0930}$

OPGAVE 3 fortsat

d. Beregning teststørrelsens (teststatistikkens) værdi, idet nødvendige mellemregninger medtages.

Dette gøres vha MATLAB:

Først beregnes:

Stikprøvens middelværdi: $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i = \frac{1}{20}\sum_{i=1}^{20} y_i = 29.84$

Stikprøvens varians: $s^2 = \frac{n\sum_{i=1}^{n} y_i^2 - (\sum_{i=1}^{n} y_i)^2}{n(n-1)} = \frac{20\cdot 17815 - 596.8^2}{20\cdot(20-1)} = 0.3457$

Stikprøvens standardafvigelse: $s = \sqrt{0.3457} = 0.5879$

I MATLAB:
$\bar{y} = mean(y) = 29.84$

$s^2 = var\ (y) = 0.3457$

$s = std(y) = 0.5879$

Teststørrelsen (teststatistikken): $t = \frac{\bar{y}-\mu}{s/\sqrt{n}} = \frac{29.84-30}{0.5879/\sqrt{20}} = -1.2170 \approx -1.22$

e. Konklusion på hypotesetesten.

Teststørrelsen (teststatistikken) $t = -1.22$, dvs $|t| = 1.22 < t_{19,0.025} = 2.0930$ og dermed ligger *t* ikke i det kritiske område, men i H_0-accept-området

Dermed kan $\boldsymbol{H_0}$**-hypotesen ikke forkastes** på baggrund af stikprøven. **Kapaciteten for de fremstillede powerbanks afviger ikke fra de ønskede 30 Ah.**

Dette spørges der ikke om:

p-værdien kan findes vha. MATLAB:

$p - værdi = 2\cdot\left(1 - tcdf(abs(t), n-1)\right) = 2\cdot\left(1 - tcdf(1.2170{,}19)\right) = 0.2385$

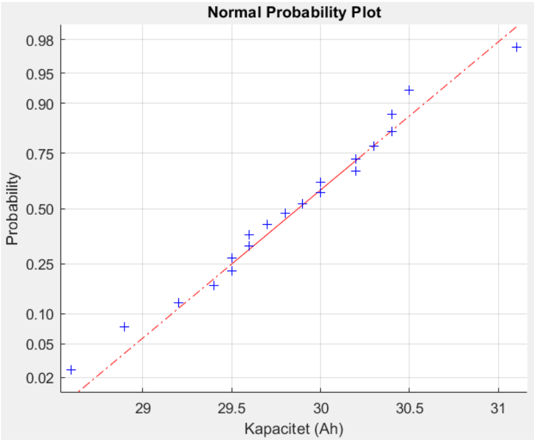
f. Hvilke antagelser er der foretaget for at udføre hypotesetesten?
Er disse antagelser rimelige på baggrund af data?
Vurderingen af data skal foregå på baggrund af normalfordelingsplot og f.eks. stem-and-leaf-plot.

Det er antaget, at den centrale grænseværdisætning gælder. Det betyder at teststørrelsen (teststatistikken) følger en *t*-fordeling, hvis *n* er tilstrækkelig stor.
Hvor stor *n* skal være afhænger af, hvor pæn den fordeling stikprøven kommer fra er.

Der udføres følgene plot i MATLAB:

OPGAVE 3f fortsat

Normalfordelingsplot: normplot(y)



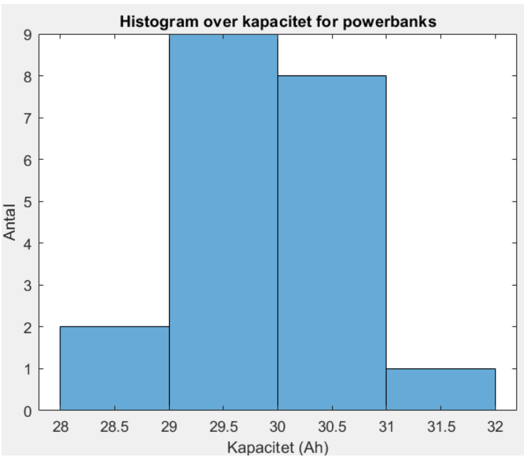
Normalfordelingsplottet viser en pæn lineær sammenhæng. Dette støtter antagelsen om en pæn fordeling, dvs. at stikprøvens fordeling ligner en normalfordeling.

Stem-and-leaf plot: stemleafplot(y,-1)

```
  28 | 6 9
  29 | 2 4 5 5 6 6 7 8 9
  30 | 0 0 2 2 3 4 4 5
  31 | 1
key: 36|5 = 36.5
stem unit: 1.0
leaf unit: 0.1
```

Stem-and-leaf plottet viser, at data kommer fra en pæn fordeling med en enkelt top, nogenlunde symmetrisk og med hurtigt uddøende haler.

Histogram: histogram(y[28:1:32])



Histogrammet viser, at data kommer fra en pæn fordeling med en enkelt top, nogenlunde symmetrisk og med hurtigt uddøende haler.

Normalfordelingsplot, stem-and-leaf plot og/ eller histogram viser alle, at stikprøven kommer fra en pæn fordeling, der godt kunne ligne en normalfordeling. **Altså holder antagelsen.**

OPGAVE 4

På en bilfabrik ønsker man en høj-kvalitetslakering af bilerne i sportsvogns-produktionen.
Der udvælges tre forskellige typer forbehandlinger af metallet: forbehandling 1, 2 og 3. Efter lakering foretages en slidtest.
De tre forskellige forbehandlinger giver følgende resultater, hvor høje værdier afspejler en god slidresistens, som dermed sikrer den ønskede høje kvalitet af lakeringen.

Forbehandling	Slidresistens				
1	93.1	94.6	92.9	93.8	95.5
2	91.6	93.1	94.2	92.4	93.2
3	97.8	96.5	96.8	97.2	97.1

a. Der udføres boxplot i MATLAB:

Forbehandling 1 og 2 har nogenlunde samme median, mens medianen for forbehandling 3 ligger tydeligt højere.

Interquartile range (IQR) er nogenlunde det samme for forbehandling 1 og 2. For forbehandling 3 er IQR betydelig mindre.

Boksene for forbehandling 1 og 2 ligger nogenlunde på samme niveau, dog ligger boksen for forbehandling 2 lidt lavere. Boksen for forbehandling 3 ligger tydeligt højere end de to andre. Nedre kvartil for forbehandling 3 ligger endvidere over øvre koste for forbehandling 1 og 2.

Variationen for forbehandling 1 og 2 er nogenlunde den samme, mens for forbehandling 3 er variationen noget mindre.
Da hvert boxplot er tegnet ud fra kun af 5 observationer er det vanskeligt at konkludere, om der er en reel forskel på de tre forbehandlinger.

b. Denne delopgave drejer sig om at undersøge ved hjælp af en variansanalyse, om der er en forskellig effekt af forbehandlingerne på slidresistensen, dvs. om der er forskel på middelværdierne af slidresistensen for de tre forbehandlinger. Der vælges et signifikansniveau på 5%.

1. Nulhypotese og alternativ hypotese for variansanalysen.
Lad $\tau_i = \mu - \mu_i$ være effekt af den i'te forbehandling.
$H_0: \tau_1 = \tau_2 = \tau_3 = 0$
$H_a: \tau_i \neq 0$ for mindst én forbehandling $i = 1, 2, 3$

2. Nulhypotese og alternativ hypotese formuleres med ord.
H_0: Der er ingen effekt af forbehandlingen, dvs. ingen afvigelse fra den overordnede middelværdi. (Forskellene i boxplot i opgave 4a er tilfældige).
H_a: Mindst én forbehandling har effekt, dvs. giver afvigelse fra den overordnede middelværdi.

OPGAVE 4b fortsat

3. Der udføres variansanalysen anovan i MATLAB.

Kommando: `[p,table,stats] = anovan(slidresistens, forbehandling)`

Analysis of Variance

Source	Sum Sq.	d.f.	Mean Sq.	F	Prob>F
X1	47.0813	2	23.5407	30.13	2.09789e-05
Error	9.376	12	0.7813		
Total	56.4573	14			

Variansanalysen har en F-værdi på 30.13 og en tilhørende p-værdi på $2.10 \cdot 10^{-5}$.

4. Konklusion på variansanalysen.

Da p-værdien på $2.10 \cdot 10^{-5}$ er mindre end signifikansniveauet på 0.05, så forkastes H_0, og H_a accepteres. Dvs. at mindst en af forbehandlingerne har en effekt på slidresistensen.

5. Hvilke antagelser skal være opfyldt for residualerne ved variansanalysen?

- Residualerne skal være normalfordelte
- Residualerne bør være tilfældige og ikke afhænge af f.eks. forbehandling eller observationsrækkefølge
- Residualerne bør være ensartede