# Eksamen - Sandsynlighed og Statistik 2021

Af Jesper Bertelsen

Indholdsfortegnelse

[Eksamen - Sandsynlighed og Statistik 2021 1](#_Toc155704142)

[Noter 1](#_Toc155704143)

[Binomialfordeling 1](#_Toc155704144)

[Normalfordeling 1](#_Toc155704145)

[T fordeling 2](#_Toc155704146)

[Opgave 1. Kvalitetskontrol af bremseskiver 3](#_Toc155704147)

[Opgave 2. Bremseskivens tykkelse 4](#_Toc155704148)

[Opgave 3. Brudstyrke af bremseskiver med forskellige legeringer 7](#_Toc155704149)

[Eksamen 2022 - Bølgeenergimaskiner 10](#_Toc155704150)

[Opgave 1 – Bølgers variabilitet 10](#_Toc155704151)

[Opgave 2 – Sammenhæng mellem bølgerne og bølgemaskinens effekt 12](#_Toc155704152)

[Opgave 3 – Sammenhæng mellem bølgerne og bølgemaskinens effekt 17](#_Toc155704153)

## Noter

### Binomialfordeling

Sandsynlighed:

Varians:

### Normalfordeling

Sandsynlighed:

### T fordeling

Konfidensinterval:

Prædiktionsinterval

## Opgave 1. Kvalitetskontrol af bremseskiver

En fabrik producerer bremseskiver til personbiler. Hver dag udtager man en stikprøve med 30 tilfældigt udvalgte bremseskiver blandt dagens produktion til grundig kvalitetskontrol. Historisk

set finder man 1.65 defekte bremseskiver i gennemsnit i den daglige stikprøve.

1. Hvor stor en procentdel af produktionen må formodes at være defekt?

Måske se på intervallet som man regner med at procenterne ligger indenfor.

Jeg har dog ikke blevet spurgt om at sige med en hvis procentsats, burde procenterne ligge indenfor.

Med en stikprøve af 30 dagligt, og med et gennemsnit på ”Succeser” ment på den måde, at det er de defekte som der bliver ledt efter, så det er en succes.

Der formodes at kunne antages, at stikprøverne er retvisende for hele produktionen.

Omkring 5,5% af produktionen vil da være defekt.

1. Hvad er sandsynligheden for, at ingen bremseskiver i dagens stikprøve er defekte?

Det er svært at sige. Jeg vil gå ud fra, at fordelingen er normal fordelt. En stikprøve på 30 virker også som en nogenlunde størrelse, og jo større de er, jo mere går de mod en normalfordeling.

Når det så er sagt, så kræver normalfordelingen en middelværdi og en spredning. Spredningen er ukendt så jeg kan kun løse det symbolsk.

Med middelværdien jeg har fået, vil ligningen da blive.

Det er den kumulerede normalfordeling fra -∞ -> 0. Den ville kunne simplificeres, men det betyder ikke noget for pointen. Jeg mangler stadigvæk konstanten sigma.

Hvad jeg ikke så var, at det er en binomialfordeling, for hver bremseskive kan beskrives binært, enten som en succes eller som en fiasko.

Binomialfordelingen følger formlen

Sandsynligheden for, at ingen bremseskiver vil være defekte i dagens stikprøve er derfor:

================

================

1. Hvad er sandsynligheden for, at der er 3 eller flere defekte i dagens stikprøve?

Sandsynligheden af det er den fulde sandsynlighed minus sandsynligheden for at få op til men ikke med 3.

=================

=================

1. Hvad er middelværdi, varians og standardafvigelse for antal defekte i de daglige stikprøver?

==================================

==================================

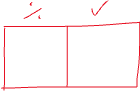
## Opgave 2. Bremseskivens tykkelse

Bremseskiveproducenten lover i specifikationerne, at bremseskivernes tykkelse er mindst 21.0 mm. For at sikre dette, tilstræber man at producere skiverne lidt tykkere, nemlig med en tykkelse på 21.8 mm. Den følgende tabel viser den målte tykkelse i en stikprøve på 28 bremseskiver:



Bremseskiveproducenten ønsker at undersøge, om den sande middelværdi for bremseskivernes tykkelse er 21.8 mm. Det vil de gøre ved at udføre en hypotesetest med et signifikansniveau på 5 %.

1. Opstil nulhypotese og alternativhypotese for testen.



Så der findes 3 scenarier.

1. Middelværdien er ikke 21,8mm & den opfylder hellere ikke en minimum tykkelse på 21. Det vil være falsk marketing at love minimum 21mm tykkelse.
2. Middelværdien er ikke 21,8mm men den opfylder at tykkelsen er minimum 21mm.
3. Middelværdien er 21,8mm som ønsket for kunne love minimum 21mm.
4. Opstil en formel for teststørrelsen og angiv hvilken fordeling, den følger. Beregn kritisk interval for testen.

Her er det ikke binært, succes eller fiasko. Det er hellere ikke noget pr. givent tidsenhed.

28 antal er en okay repræsentativ stikprøve som jeg vil mene vil kunne blive repræsenteret kontinuert. Derfor mener jeg, at fordelingen vi skal have fat i, er

*Normalfordelingen:*

Men standard afvigelsen & middelværdien er ukendt og kan kun beskrives ved estimatorer. Dermed må det være en

*T - fordeling:*

Som jeg skal bruge.

Til t fordelingen har jeg ikke brug for at estimerer middelværdien, frihedsgraden er dermed

df = n - 1

Med det her kan jeg så finde estimeringsfejlen:

Hvor anden del af produktet er standardafvigelsen og kan findes.

Figure : 95% Konfidensinterval

Et billede, der indeholder tekst, skærmbillede, software, Multimediesoftware

Automatisk genereret beskrivelse

Med 95% sikkerhed kan jeg derfor sige, at tykkelsen på bremseskiverne vil være mellem 21,28mm & 22,11mm.

1. Beregn teststørrelsens værdi og konkludér på testen.

Jeg afviser nulhypotesen, hvis



Så nulhypotesen afvises, middelværdien er ikke lige med

Nu tester jeg så alternativ hypotesen:



Så alternativhypotesen gælder. Testen viser, at tykkelsen fra produktionen er større end 21mm.

1. Oplys hvilke antagelser, der er gjort i hypotesetesten, og vurdér om antagelserne er rimelige på baggrund af data.
2. Beregn et 99 % konfidensinterval for den sande tykkelse af bremseskiverne.



Med 99% sikkerhed kan jeg sige, at tykkelsen på bremseskiverne er



1. Beregn et 95 % prædiktionsinterval for skivernes tykkelse.

Prædiktionsintervallet giver et estimat for hvilket interval observationerne ligger indenfor med 95% sikkerhed, den følger formlen:

Fra stikprøven ved jeg, at:

Så



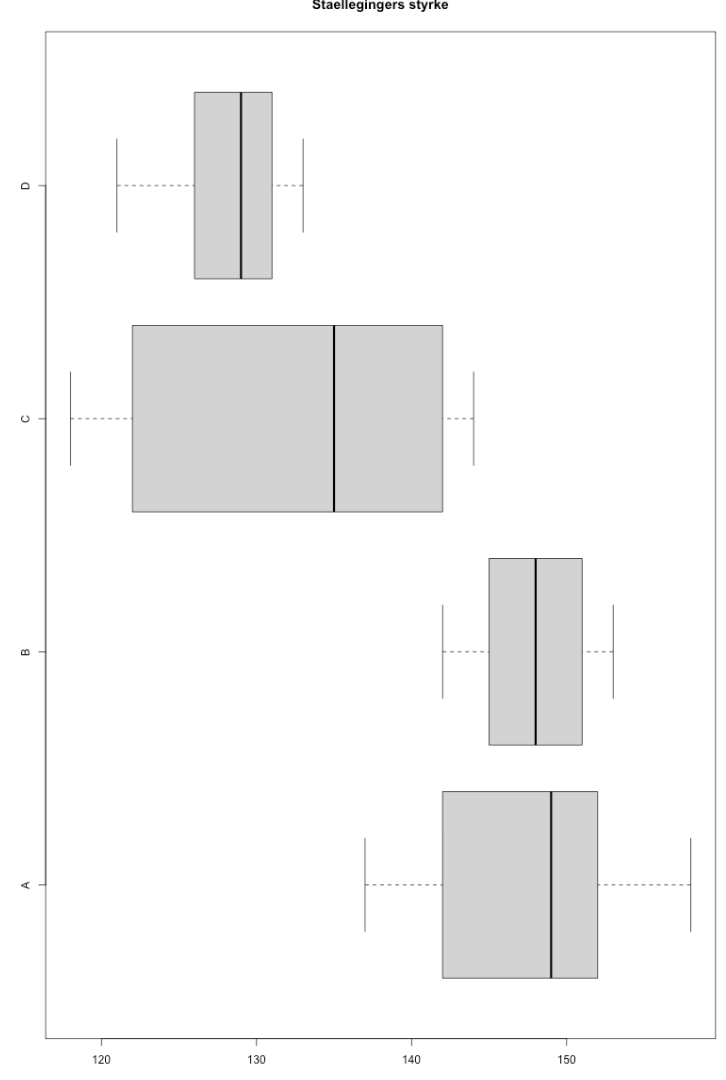
Med 95% sikkerhed, kan jeg da sige, at stikprøvens observationer ligger indenfor:

## Opgave 3. Brudstyrke af bremseskiver med forskellige legeringer

Bremseskivefabrikken vil gerne øge styrken af de producerede bremseskiver. Derfor eksperimenterer man med forskellige stållegeringer. I et forsøg har man målt brudstyrken af bremseskiver støbt med 4 forskellige legeringer, markeret som A, B, C og D. Resultatet vises i følgende tabel:

Et billede, der indeholder tekst, skærmbillede, Font/skrifttype, nummer/tal

Automatisk genereret beskrivelse

1. Lav et parallelt boksplot (kassediagram) for de fire legeringer. Hvad fortæller boksplottet?

Af boksplottet ses der to tydelige tabere. Type C & type D, som har de laveste brudstyrker.

Type A & B har de største brudstyrker. Type A har den observation med størst brudstyrke, men har også en stor spredning. Type B er på den anden side et mere sikkert valg.

Hvis de bedste typer blev valgt efter størst brudstyrke, men også blev lagt vægt på mindst mulig spredning, så vil min liste være:   
  
1. B

2. A

3. C

4. D

Selvom D har en mindre spredning, så er den bare noget svagere end C, derfor er den sidst.

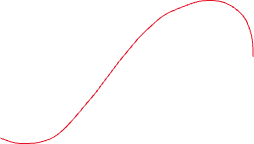
1. Undersøg med en variansanalyse, om der er statistisk belæg for at påstå, at der er forskel på de fire legeringer på 5 % signifikansniveau.

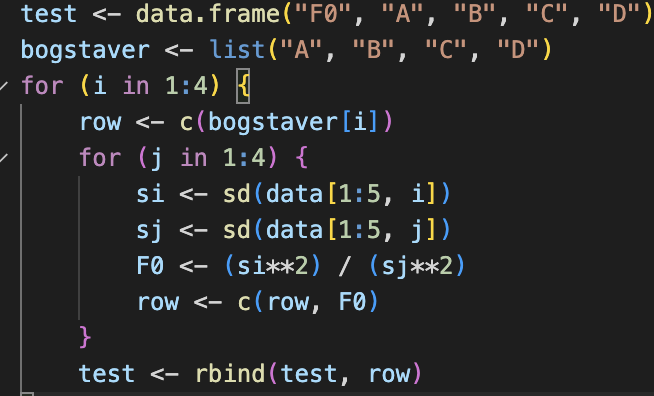
Så vores nulhypotese sætter vi til at være:

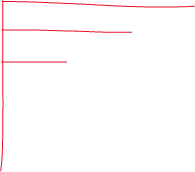
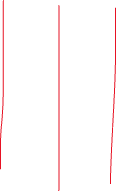
Og med det så forventer vi at teststørrelsen:

For alle kombinationer af .

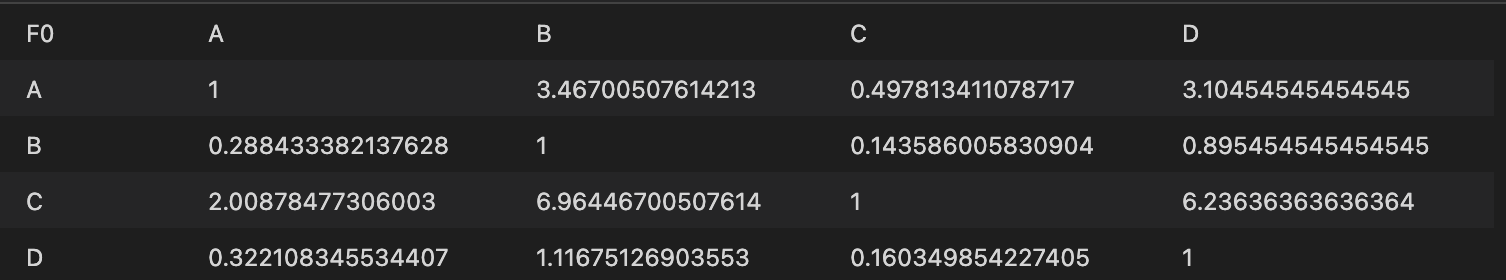
Til det har jeg lavet et script til at beregne alle teststørrelserne og sætte dem i en kasse:







Jeg itererer gennem kolonnerne i hver række, og laver så en tabel:





På diagonalen ses der etere, som er forventet efter som det er spredningen i , de er altså de samme.

Hvis jeg stikprøverne skulle være de samme, så ville jeg se, at nogle af dem ville være tætte på 1.

Det er tilfældet i et enkelt tilfælde.

Udelukkende ud fra variansanalysen, så ses B & D ud til at kunne være fra samme stikprøve. De ser også ens ud på boxplottet, spredningsmæssigt, men deres middelværdi er dog noget anderledes.

Frihedsgraderne for hvert stikprøve er

Hvis i r

Så må jeg skulle forkaste, at de ikke er ens.



Ud fra variansen kan jeg altså kun forkaste, at B & C ikke er stikprøver af samme legering.

1. Lav en parvis sammenligning af legeringerne med Tukey HSD metoden. Hvilke par af legeringer er forskellige på 5 % signifikansniveau?
2. Hvilke antagelser er der gjort for residualerne i analysen? Vurdér om antagelserne holder.
3. Hvilken legering ville du anbefale at bruge på baggrund af eksperimentet? Begrund dit svar.

# Eksamen 2022 - Bølgeenergimaskiner

Et billede, der indeholder sky, vand, udendørs, hav

Automatisk genereret beskrivelse

Et eksempel på en bølgeenergimaskine

I et kystnært land som Danmark er det naturligt, at der forskes i vedvarende energikilder som vind- og bølgekraft. Vi har mulighed for at høste masser af gratis energi fra vind og bølger. Danmark har været foregangsland indenfor begge energikilder, og vi er førende på verdensplan indenfor produktion af vindmøller. Imidlertid har det vist sig at være en gigantisk ingeniørmæssig udfordring at konstruere en bølgeenergimaskine (også kaldt en bølgemaskine), der kan udvinde energien fra bølgerne i det barske havmiljø.

## Et billede, der indeholder tekst, skærmbillede, Font/skrifttype, nummer/tal Automatisk genereret beskrivelseOpgave 1 – Bølgers variabilitet

Selv i stabilt vejr er der stor variation i enkeltbølgers højde. En gruppe ingeniørstuderende vil undersøge hvordan højden af enkeltbølger fordeler sig. De registrerer 329 bølgers højde på det samme sted i løbet af 10 minutter på en dag med stille vejr. Tabellen nedenfor viser antal bølger fordelt efter højde i intervaller på 20 cm. F.eks. kan man se, at de studerende talte 22 bølger i intervallet mellem 1.0 og 1.2 meters bølgehøjde.

*Omregn antal bølger i hvert interval til sandsynligheden for at få en bølge indenfor intervallet.*

Med r har jeg fået lavet en forbedret tabel som også fortæller sandsynligheden for hver kategori.

Den følger formlen:

Et billede, der indeholder tekst, skærmbillede

Automatisk genereret beskrivelseEt billede, der indeholder tekst, skærmbillede, Font/skrifttype

Automatisk genereret beskrivelse



Figure 1: Script til at lave matricen i r

Figure 2: Resultat af matricen

*Beregn den kumulerede sandsynlighedsfordeling for bølgehøjden. Hvad er sandsynligheden for, at den næste bølge er højere end 0.8 m?*

Til det udvider jeg min matrix endnu mere.

Jeg har lavet en funktion til at lave matrixen, så jeg bare skal indsætte listen med værdierne, antal rækker, antal søjler og dens søjlenavne.  
I *r* er der en funktion til at lave den kumulerede frekvens, så jeg ikke har brug for at iterere igennem værdierne, det gør funktionen for mig.

Kumulationen følger formlen:

Hvor er de kumulerede procenter til *i*, og *p* er procenterne til *i*.

Et billede, der indeholder tekst, skærmbillede, Font/skrifttype

Automatisk genereret beskrivelse

Figure 3: Ny opsætning af matrice



Et billede, der indeholder tekst, skærmbillede

Automatisk genereret beskrivelse

Figure 4: Resultat af den nye matrice



Jeg går ud fra, at det er opgaven spørger om, at sandsynligheden er skarpt større end 0.8, da vil sansynligheden være:

====================================

====================================

*Beregn den gennemsnitlige bølgehøjde (antag at alle bølgerne i et interval har samme højde, nemlig intervallets midterværdi).*

Så her skal vi vægte observationerne, med deres højder.

Beregningen følger formlen:

I *r* ser den sådan her ud:

hoejdeGennemsnit <- sum((hoejderMidte \* observeret)) / n

Hvor jeg tager brug af sprogets indbyggede sum funktion.

==================

==================

Slut på den her opgave, videre til den næste.

## Opgave 2 – Sammenhæng mellem bølgerne og bølgemaskinens effekt

Et billede, der indeholder tekst, skærmbillede, nummer/tal, Font/skrifttype

Automatisk genereret beskrivelseDen effekt, som en bølgemaskine kan producere, afhænger af:

1. Bølgernes højde og
2. Afstanden imellem bølgetoppene, d.v.s. bølgelængden.

Som vi så på i opgave 1, er der stor variation på bølgernes form i havet, selv under roligt og stabilt vejr. Derfor bliver bølgemaskiner typisk testet i mindre skalamodeller i et bassin, hvor man kan skabe ensartede bølgetog og måle den producerede effekt.

Tabellen til højre viser data for den målte effekt (𝐸𝐸), som en bestemt bølgemaskine producerer, når den rammes af et bølgetog med en fastlagt bølgelængde (𝐿𝐿) og bølgehøjde (𝐻𝐻).

*Lav et plot for hver af de uafhængige variable (hhv. bølgelængde og bølgehøjde), der viser om der er korrelation mellem variablen og den producerede effekt. Diskutér kort dine plots.*

Så her ønskes der et plot af effekten til længder og højder (L, E) & (H, E), og så en korrelation mellem variablen og den producerede effekt. Kan observationen beskrives som en ligning?

Her er tanken at lave noget lineært regression på den, så jeg kan finde en sammenhængen på de uafhængige variabler overfor effekten.

Til lineær regression bruger man mindste kvadraters metode.   
På formlen

Vil vi prøve at beskrive observationen på en lineær måde.   
Konstanterne kan beskrives som:

Hvor variablerne med baren over, er observationernes middelværdi.

*R* har en bekvem funktion, som finder alt hvad jeg har brug for, for mig:

fortæller vigtige ting om den lineære regression af y beskrevet ud fra x.

Lineære regression for :

Et billede, der indeholder skærmbillede, tekst, linje/række, Kurve

Automatisk genereret beskrivelseEt billede, der indeholder tekst, skærmbillede, Font/skrifttype, software

Automatisk genereret beskrivelseMed *R*’s funktion abline, kan man så indsætte den regression man får fra lm, så den selv indsætter koefficienterne som på formlen ovenfor, og laver den til en linje.   
  
Jeg laver en funktion, så det er nemmere at lave to plots:

Figure 6: Funktion som skal plotte (x, y) & en korrelation mellem dem

Figure 5: (l, e) & dens korrelation

Et billede, der indeholder skærmbillede, linje/række, diagram, Kurve

Automatisk genereret beskrivelseSom der ses, så er der ikke en tydelig lineær korrelation mellem bølgelængder og effekten.   
Men effekten vil vi måske kunne beskrive ud fra bølgehøjden, for trods nogle små afstikkere, så ser korrelationen nogenlunde tilfredsstillende ud, ikke den bedste, men okay.

Figure 7: (h, e) & dens korrelation

*Lav en multipel lineær regressionsanalyse, der beskriver produceret effekt som funktion af bølgelængde og bølgehøjde. Skriv regressionsligningen op.*

Til denne lineære analyse prøves effekten at beskrives som følge af højden i en bølge og længden af en bølge.

Et billede, der indeholder tekst, skærmbillede, Font/skrifttype

Automatisk genereret beskrivelseMen effekten kan også være påvirket af en interaktion mellem højden og længden. Hvis højden er påvirket af længden eller omvendt.

Figure 8: Lineær model til e = l + h

Først testes den ikke interaktive model:

Så testes den interaktive:

Et billede, der indeholder tekst, skærmbillede, Font/skrifttype

Automatisk genereret beskrivelse

Figure 9: Lineær model til e = l + h + lh

Der ses at forklaringsgraden: Multiple R-Squared er større i den interaktive model. Det må altså være en bedre model, og jeg kan derfor ikke forkaste, at der ikke er interaktion mellem bølgehøjde og bølgelængde.

Det er dermed den model jeg bruger.

…

*Forklar vha. regressionsanalysens statistikker og dine plots fra delopgave a., om modellen beskriver observationerne godt.*

*I resten af opgaven skal du kun bruge de 6 observationer, hvor bølgelængden er 2.0 m*

*Et billede, der indeholder tekst, skærmbillede

Automatisk genereret beskrivelseLader der til at være en lineær sammenhæng mellem bølgehøjde og effekt, når bølgelængden er fast 𝐿𝐿 = 2.0 m?*



For bølgelængden på 2m må jeg tage et udpluk af listerne fra indeks [3:9[



Et billede, der indeholder tekst, Font/skrifttype, skærmbillede, nummer/tal

Automatisk genereret beskrivelse

I dette interval lader der til, at der kan skrives en lineær sammenhængen mellem bølgehøjden og effekten.

*Lav en transformation af bølgehøjden med funktion:*

hvor 𝐻 er bølgehøjden og 𝑘 er en konstant, der er karakteristisk for bølgemaskinen. For den pågældende bølgemaskine har man estimeret, at 𝑘 = 2.5 m.  
Lav en simpel lineær regression med 𝐻𝑙 som regressorvariabel og 𝐸 som responsvariabel.

Her har jeg en variabel som jeg skal estimere. Hvis jeg kiggede ind i estimering af variabler, så kunne jeg godt finde ud af det.

*Brug den lineære regression fra delopgave e. til at skrive et udtryk for 𝐸𝐸 som funktion af 𝐻𝐻. Beregn den forventede effekt med en bølgehøjde på 4.6 m.*

## Opgave 3 – Sammenhæng mellem bølgerne og bølgemaskinens effekt

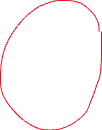
Et billede, der indeholder tekst, skærmbillede, Font/skrifttype, nummer/tal

Automatisk genereret beskrivelse

1. Lav og kommenter et parallelt boksplot, der viser energiproduktionen for de to skalamodeller.

Et billede, der indeholder Rektangel, skitse, diagram, kvadratisk

Automatisk genereret beskrivelse



De originale maskiner: 50% af maskinerne producerer , 25% af dem producerer & de sidste 25% producerer .

De modificerede har 25% af maskinerne til at producere , 50% af maskinerne til at producere & de sidste 25% producerer .

Tallene er estimater taget fra grafen.

Som konklusion på boxplottene, så kan det konkluderes at 75% af de modificerede bølgemaskiner producerer mere energi en halvdelen af de originale. Og en generel tendens er, at de modificerede er bedre.

1. Man ønsker at slå fast med et signifikansniveau på 5 %, om energiproduktionen er højere med den modificerede model end med den oprindelige. Opstil nulhypotese og alternativhypotese for denne hypotesetest.

Der ønskes at kunne siges med 95% sandsynlighed, at energiproduktionen er højere med de modificerede modeller end de oprindelige.

, som vi ønsker at forkaste.

, som vi ønsker skal være rigtigt for 95% af alle observationer.

1. Opstil formlen for teststørrelsen og beregn dens værdi. Angiv hvilken fordeling den følger.

Her skal formlen ses som, at teststørrelsen er af god størrelse for at være gældende, ellers så havde man skulle have brugt   
Teststørrelser af 12 er i den lave ende, men lad mig sige, at det duer.

Hvis så må jeg skulle forkaste nul hypotesen:

==============================================================================  
Så jeg må forkaste nulhypotesen, teststørrelserne er ikke lige så store.   
Da z0 er størst, så må jeg også kunne konkludere, at de modificerede maskiner er bedre end de oprindelige. ==============================================================================

1. Beregn den kritiske region for testen og konkluder på hypotesetesten.
2. Oplys hvilke antagelser, der er gjort i hypotesetesten, og om antagelserne er rimelige på baggrund af data.
3. Beregn et 95 % konfidensinterval for forskellen på materialernes middelværdi. Diskutér hvordan boksplot, hypotesetest og konfidensinterval stemmer overens.