Statistisk Dataanalyse

LESSON 8. STATICS IN R – PART 2

Class content:

1) Basic operations in R
2) Types of variables in R
3) Inferential statistics in R: Hypothesis testing + ANOVA

Lesson 8. Statistics in R - Part 2
1) We have measured the potato yield from 12 different farms. We know that the standard potato yield for the given variety is μ=20. x = [21.5, 24.5, 18.5, 17.2, 14.5, 23.2, 22.1, 20.5, 19.4, 18.1, 24.1, 18.5] Use R to reply the following questions: a. Does the population follow a normal distribution? b. What is the sample mean? c. What is the population mean μ (consider a 95% confidence level)? d. Is there evidence that the potato yield from these farms is significantly different than the standard yield?

A. Vi skal starte med at finde om der er en Normal Distribution, altså vi skal finde intervallerne med Standardafvigelse+Middelværdi, samt placering af Middelværdi i Midten

B. What is the sample Mean?

- Vi starter allerførst med at finde Middelværdien og derefter finder vi standardafvigelsen.

Summen af Potato Yield Sample	$\mathbf{21,5+24,5+18,5+17,2+14,5+23,1+22,21+20,5+19,4+18,1+24,1+18,5=242,11}$
Middelværdi af Potato Yield Sample	$\mathbf{\frac{242,11}{12}\approx 20,17583}$

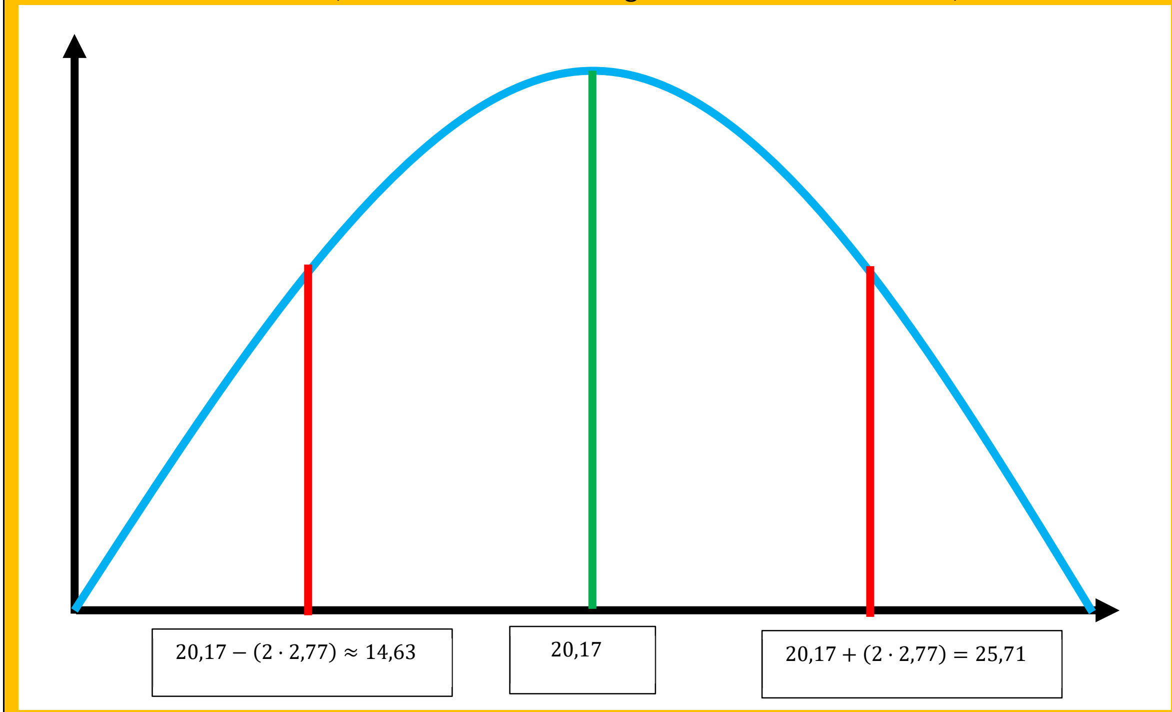
- Vi kan dermed se, at selve Middelværdi / Mean for sample er 20,17.
- Nu skal vi finde Standardafvigelsen.

$21,5-20,17 \approx 1,33$	$24,5-20,17 \approx 4,33$	$20,17-18,5 \approx 1,67$	$20,17-17,2 \approx 2,97$	$20,17-14,5 \approx 5,67$	$23,1-20,17 \approx 2,93$
$22,21-20,17 \approx 2,04$	$20,5-20,17 \approx 0,33$	$20,17-19,4 \approx 0,77$	$20,17-18,1 \approx 2,07$	$24,1-20,17 \approx 3,93$	$20,17-18,5 \approx 1,67$

- Nu skal vi kvadrere subtraktionerne og derefter finde summen af dem. Til sidst divideres de med antallet minus 1, hvor vi så i kvotienten får udgivet vores Standardafvigelse.

$1,33^2 \approx 1,7689$	$4,33^2 = 18,7489$	$1,67^2 = 2,7889$	$2,97^2 \approx 8,8209$	$5,67^2 = 32,1489$	$2,93^2 \approx 8,5849$
$2,04^2 = 4,1616$	$0,33^2 \approx 0,1089$	$0,77^2 = 0,5929$	$2,07^2 \approx 4,2849$	$3,93^2 = 15,4449$	$1,67^2 = 2,7889$
Summen af Kvadraterne	$1,7689+18,7489+2,7889+8,8209+32,1489+8,5849+4,1616+0,1089+0,5929+4,2849+15,4449+2,7889 \approx 100,2435$				
Standardafvigelse	$\sqrt[2]{\frac{100,2435}{12+1}} \approx 2,776876$				

- Vi kan dermed se, at vores Standardafvigelse for Potato Yield er 2,77.



- Vi kan dermed udefra vores ”håndregning” se, at ved 95 konfidensniveau er der snak om en Normal Distribution, efter den udregnede middelværdi passer ind i de udregnede intervaller:
- Interval: -14,63 < 20,17 < 25,71

C. Nu skal vi anvendes vores Viden fra One Sample T-test, hvor vi skal sammenligne vores Sample Middelværdi med Populations Hypotese Middelværdi

- Vi kan i vores tilfælde se, at vi skal anvende følgende formel som er nedenfor.

$$\boldsymbol{\bar{x} - t_{N-1} \cdot \frac{s}{\sqrt[2]{N}} < \mu < \bar{x} + t_{N-1} \cdot \frac{s}{\sqrt[2]{N}}}$$

- Vi skal lige starte med at finde selve den kritiske værdi gennem frihedsgraden.
- I vores tilfælde, kan det ses at vi har følgende frihedsgradsværdi: 12-1=11
- Så har vi Konfidensniveau som er: 95%=0,05.
- Udefra dette, anvendes Student T-test tabellen og følgende kritiske værdi fås: 2,17.

$$\mathbf{20,17 - 2,17 \cdot \frac{2,77}{\sqrt[2]{12}} = 20,17 - \frac{3,00545}{\sqrt{3}} \approx 18,4348}$$

$$\mathbf{20,17 + 2,17 \cdot \frac{2,77}{\sqrt[2]{12}} = \frac{3,00545}{\sqrt{3}} + 20,17 \approx 21,9052}$$

- Nu får vi dannet følgende Interval for Populationen af One Sample.

$$\mathbf{18,43 < 20,17 < 21,90}$$

- Fordi vores Populationsmiddelværdi passer inde i Intervallerne for One-Saple, kan vi derfor bekræfte null-hypotesen som siger at der er ingen signifikant forskel mellem Potato Yield og Standard Yield 😊

2) A researcher wants to see if a vitamin included in the diet changes the cholesterol. Six subjects were pretested at Week 0, and then they took the vitamin during 6-weeks. After the 6-weeks period, their cholesterol level was measured again. Using R, can we conclude (with 95% confidence level) that the cholesterol level has been changed? Assume the variable is approximately normally distributed.

Subject	1	2	3	4	5	6
Week 0	215	239	208	190	172	244
Week 6	184	160	201	188	169	219

3) Repeat the previous exercise using now a 90% confidence level.

- Vi kan se, at fordi vi skal sammenligne medikationen fra Uge 0 til Uge 6, kan vi derfor sige at det er Two-Sample Dependent T-Test.
- Vi starter med at danne følgende Tabel for at kunne finde de nødvendige værdier til formlen nedenfor.

x_1	x_2	$x_1 - x_2$	$(x_1 - x_2)^2$
215	184	$215 - 184 = 31$	$31^2 = 961$
239	160	$239 - 160 = 79$	$79^2 = 6241$
208	201	$208 - 201 = 7$	$7^2 = 49$
190	188	$190 - 188 = 2$	$2^2 = 4$
172	169	$172 - 169 = 3$	$3^2 = 9$
244	219	$244 - 219 = 25$	$25^2 = 625$
		$31 + 79 + 7 + 2 + 3 + 25 = 147$	$961 + 6241 + 49 + 4 + 9 + 625 = 7889$
$\frac{147}{6} = \frac{49}{2} = 24,5$			

- Nu anvendes følgende formel for Two-Sample Dependent T-test.

$$\frac{(\overline{x_1} - \overline{x_2}) - (\overline{\mu_1} - \overline{\mu_2})}{\frac{s_{\overline{x_1} - \overline{x_2}}}{\sqrt[2]{N}}}$$

- Vi skal starte med at finde Standardafvigelsen gennem følgende Formel.

$$s_{\overline{x_1}-\overline{x_2}} = \sqrt[2]{\frac{\sum(\overline{x_1}-\overline{x_2})^2 - \frac{(\sum(\overline{x_1}-\overline{x_2}))^2}{N}}{N-1}}$$

$$\sqrt[2]{\frac{7889 - \frac{147^2}{6}}{6-1}} = \frac{7\cdot\sqrt{35}}{\sqrt{2}} \approx 29{,}2831$$

- Nu indsætter vi alle de udregnede værdier inde i formlen for Dependent T-Test.

$$\frac{24{,}5-0}{\frac{29{,}28}{\sqrt[2]{6}}} = \sqrt{6}\cdot 0{,}8367486 \approx 2{,}049607$$

- Nu skal vi finde selve den kritiske værdi og frihedsgraden.
- Vi kan se, at frihedsgraden er: N-1=6-1=5 og vi har konfidensniveauet = 95%=0,05.
- Vi kan se ude fra Student T-test at vores kritiske værdi bliver således: -2,57
- Vores Konfidensintervaller bliver således: -2,57 < 2,04 < 2,57.
- Fordi vores Middelværdi passer inde i Konfidensintervallet, kan vi derfor sige acceptere vores null-hypotese som betyder at der er ingen forskel i kolesterol-niveau før undersøgelsen og efter. Derfor er kolostrol-niveauet ikke ændret.

- Men hvis man kigger overimod den anden opgave, kan det ses at vi har fået oplyst et andet konfidensniveau og dette kan vi bruge til at udregne en nyt kritisk værdi.

- Kritisk værdi = (N-1 ; 90%=0,10) og dette svarer til (6-1 ; 0,10) som er = 2,015 i two-tailed.

- Derfor kan vi danne følgende Konfidensinterval: -2,015 < 2,04 < 2,015

- Vi kan derfor sige, at fordi vores Middelværdi ligger uden for Konfidensintervallerne, kan vi derfor med 90% sige at der er sket en ændring på Kolostrol-Niveauet i løbet af de 6 Uger eftersom Null-Hypotesen er blevet afvist og Alternativ Hypotesen med forskelligheden er blevet accepteret 😊

4) We would like to know if the concentration of a compound in two brands of yogurt is different. We select 20 bottles of Brand A and 20 bottles of Brand B. The results are shown in the excel file "yogurt.xslx".
   a) What is the appropriate test to use to respond our research question?
   b) What is the main assumption to be tested before performing the test? After importing the data to R, test the assumption using the appropriate method.
   c) Can you conclude whether the compound's concentration in the two brands of yogurts is significantly different?

i. Vi kan se, at fordi vi sammenligner to forskellige Yoghurt-Brands med hinanden er der snak om Two-Sample Independent T-test.

ii. Inden man importerer data ned, er det nødvendigt at kigge efter hvad Konfidensniveauet er og om begge Brands er normal fordelt. Vi kan se, at fordi vi ikke har fået oplyst nogen som helst Konfidensniveauer, kan vi sige at der er snak om 9%=0,05.

iii. Her anvender vi Two-Sample Independent T-test formlen nedenfor:

$$\frac{\overline{\boldsymbol{x}} - \boldsymbol{\mu}}{\frac{\boldsymbol{s}}{\sqrt[2]{\boldsymbol{N}}}}$$

- I tilfældet, kan det ses at vi har følgende formel for vores Pooled Standardafvigelse.

$$s_{\overline{x_1}-\overline{x_2}} = \sqrt[2]{\frac{(20-1)\cdot 11{,}73 + (20-1)\cdot 8{,}45}{20+20-2}} \cdot \sqrt[2]{\frac{1}{20}+\frac{1}{20}} = \frac{3{,}176476}{\sqrt{10}} \approx 1{,}00449$$

- Nu indsættes værdierne inde i Two-Sample Independent T-test formlen.

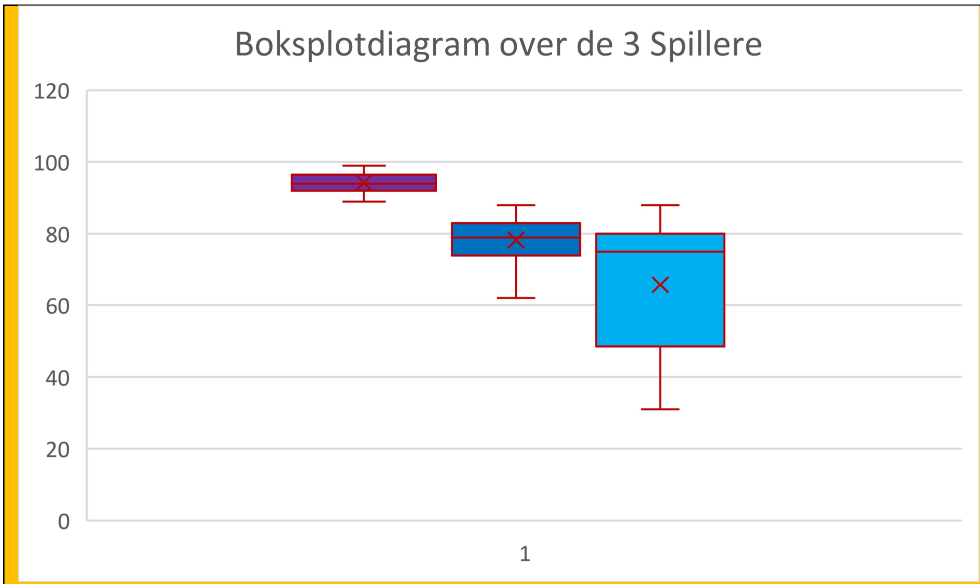
$$\frac{\mathbf{(69,95-52,58)}}{\frac{\mathbf{1}}{\sqrt[2]{\mathbf{20}}}} = \sqrt{\mathbf{5}} \cdot \mathbf{34,74} \approx \mathbf{77,681}$$

- Da vi har fundet vores værdier, kan vi derfor finde vores frihedsgrader og Kritiske værdi.
- Frihedsgrad: N+N-2 = 20+20-2=38
- Kritiske værdi ved Two-Tailed: (38 ; 95%=0,05) = 2,02
- Vi kan udefra det danne følgende Interval: -2,02 < 77,68 < 2,02

- Fordi vi vores Middelværdi ikke passer ind til vores kritiske Konfidensintervaller, kan vi derfor sige med 95%, at de to Yoghurt Brands er signifikant forskellige gennem bekræftelse af alternativ hypotesen.

5) We want to compare the scores obtained by three professional basketball players. The data with all the scores obtained by the players in the pre-season games is available in the csv file "basketball.csv".
   a. Construct a boxplot for each of the players, to better visualize the data.
   b. When using the appropriate statistical test, is there a significant difference among the scores obtained by each of the players?
   c. In case there is a difference, which player/players obtained a higher or lower score than the other/others?

- A. For at kunne løse denne opgave på den "håndfulde-måde", kan vi gå inde under Indsæt-Menuen og derved finde Boksplot ikonet under Histogram. Derefter kan Boksplotgrafen dannes som følgende:



- Vi kan se, at den Blå-Boksplot har den øverste Størsteværdi sammenlignet med de andre Boksplotgrafer. Men fordi denne Boksplotgraf tilhører Michael, kan vi derfor sige at Michael er ham som er den bedste spiller blandt de 3 nævnte.

- B. Til sådan type opgaver, plejer vi oftest at anvende One-Way ANOVA især med LSD-Test hvor man finder ud af hvordan specifikt værdierne adskiller sig i forhold til Intervaller.

- Vi kan se, at vi har startet med at finde Middelværdierne for de 3 Spillere.

Summen af Michael	$93+94+92+97+90+97+99+92+94+96+89+92+92+99+94+96+96+91+93+94+99=1979$
Middelværdi	$\frac{1979}{21} \approx 94,2381$

Summen af Damon	$81+75+72+79+83+88+83+79+71+62+88+74+79+74+76+73+79=1316$
Middelværdi	$\frac{1316}{21}=\frac{188}{3} \approx 62,66667$

Summen af Allen	$83+84+81+78+73+77+40+80+35+88+78+35+80+55+67+80+58+75+42+61+31=1381$
Middelværdi	$\frac{1381}{21} \approx 65,7619$

- Nu skal vi til at udregne selve den totale middelværdi, som er vores "sum of mean" og det betyder at vi lægger alle middelværdierne sammen til en.

Summen af Spillernes Middelværdi	$94,23+62,66+65,76 \approx 222,65$
Totale Middelværdi af Spillerne	$\frac{222,65}{3} \approx 74,21667$

- Nu skal vi udregne summen af kvadraterne for alle Spillere.

- Dette betyder, at vi tager den enkelte dataværdi og trækker det fra den totale middelværdi af kunstgøderne. Vi viser den enkelte eksempel herhenne.

Sum of Square Total	$(93-74{,}21)^2+(94-74{,}21)^2+(92-74{,}21)^2+(97-74{,}21)^2+(90-74{,}21)^2+(97-74{,}21)^2+(99-74{,}21)^2+(92-74{,}21)^2+(94-74{,}21)^2+(96-74{,}21)^2+(89-74{,}21)^2+(92-74{,}21)^2+(92-74{,}21)^2+(99-74{,}21)^2+(94-74{,}21)^2+(96-74{,}21)^2+(96-74{,}21)^2+(91-74{,}21)^2+(93-74{,}21)^2+(94-74{,}21)^2+(99-74{,}21)^2+(81-74{,}21)^2+(75-74{,}21)^2+(74{,}21-72)^2+(79-74{,}21)^2+(83-74{,}21)^2+(88-74{,}21)^2+(83-74{,}21)^2+(79-74{,}21)^2+(74{,}21-71)^2+(74{,}21-62)^2+(88-74{,}21)^2+(74{,}21-74)^2+(79-74{,}21)^2+(74{,}21-74)^2+(76-74{,}21)^2+(74{,}21-73)^2+(79-74{,}21)^2+(83-74{,}21)^2+(84-74{,}21)^2+(81-74{,}21)^2+(78-74{,}21)^2+(74{,}21-73)^2+(77-74{,}21)^2+(74{,}21-40)^2+(80-74{,}21)^2+(74{,}21-35)^2+(88-74{,}21)^2+(78-74{,}21)^2+(74{,}21-35)^2+(80-74{,}21)^2+(74{,}21-55)^2+(74{,}21-67)^2+(80-74{,}21)^2+(74{,}21-58)^2+(75-74{,}21)^2+(74{,}21-42)^2+(74{,}21-61)^2+(74{,}21-31)^2\approx 17994{,}4$
Sum of Square Between for Michael	$12\cdot(94{,}23-74{,}21)^2\approx 4809{,}605$
Sum of Square Between for Damon	$12\cdot(62{,}6-74{,}21)^2\approx 1617{,}505$
Sum of Square Between for Allen	$12\cdot(65{,}7-74{,}21)^2\approx 0{,}0845$
Sum of Total Square Between	$4809{,}605+1617{,}505+0{,}0845\approx 6427{,}194$

- Nu skal vi udregne Sum Square Residual og dette er muligt ved at trække Summen af Square-Total fra Summen af Sum of Total Square Between

SS-Total - SS-Between	$\mathbf{17994,4-6427,194\approx 11567,21}$

- Nu skal vi udregne de adskillige ting som er vist i tabellen på Præsentationen.

X	Sum of Squares	d.f.	Mean	F-Ratio
Factor	SS-Between = 6427,194	Antal gruppe - 1 => 3-1 = 2	Dividere Total Square Between med 2 $\frac{6427{,}194}{2}=3213{,}597$	Dividere Mean Faktor med Mean Residual $\frac{3214{,}597}{350{,}5215}\approx 9{,}170898$
Residual	$\mathbf{17994,4-6427,194\approx 11567,21}$	$35-2=33$	Dividere SS-Between Residual med df-Residual	

			$\frac{11567,21}{33} \approx 350,5215$	
Total	SS-Total = 17994,4	Antal dataværdier => 36-1 = 35		

- Nu, hvor frihedsgraden er blevet fundet, skal vi derfor finde den kritiske værdi som er markeret med blå.

k-1, N-k => (3-1, 36-1)	(3,26)

- Vi kan se, at selve F-Ratio er større end den Kritiske Værdi og derfor kan vi sige at det er følgende:
- F-Ratio > F-Tabel
- Fordi F-Ratio er Større end F-Tabel, kan vi afvise Null-Hypotesen, hvilket betyder at der er en forskel mellem middelværdier af de 3 Spillere i Basketball.

- C. Derfor skal vi nu anvende LSD-Testen for at kunne finde ud af "specifikt", hvilken Spiller i Basketball adskiller sig fra den ene og om der er andre som ligner det samme.

- Det skal understreges, at for at kunne finde intervallerne i LSD, anvendes følgende formel:

$$\overline{x_1} + \frac{\sqrt[2]{2}}{2} \cdot t_{dfresiduals} \cdot \sqrt[2]{\frac{MS_{residuals}}{n_k}}$$

F-Ratio > F-Tabel			LSD-Intervaller
Kunstgødningerne	Middelværdi	Nedre Grænse	Øvre Grænse
Michael	94,23	$94,23 - \frac{\sqrt[2]{2}}{2} \cdot 3,26 \cdot \sqrt[2]{\frac{350,52}{12}} = 94,23 - \frac{17,61909}{\sqrt{2}} \approx 81,77142$	$94,23 + \frac{\sqrt[2]{2}}{2} \cdot 3,26 \cdot \sqrt[2]{\frac{350,52}{12}} = 94,23 + \frac{17,61909}{\sqrt{2}} \approx 106,6886$
Damon	62,6	$62,6 - \frac{\sqrt[2]{2}}{2} \cdot 3,26 \cdot \sqrt[2]{\frac{350,52}{12}} = 62,6 - \frac{17,61909}{\sqrt{2}} \approx 50,14142$	$62,6 + \frac{\sqrt[2]{2}}{2} \cdot 3,26 \cdot \sqrt[2]{\frac{350,52}{12}} = 62,6 + \frac{17,61909}{\sqrt{2}} \approx 75,05858$

Allen	65,7	$65{,}7-\frac{\sqrt[2]{2}}{2}\cdot 3{,}26 \cdot \sqrt[2]{\frac{350{,}52}{12}} = 65{,}7-\frac{17{,}61909}{\sqrt{2}} \approx 53{,}24142$	$65{,}7+\frac{\sqrt[2]{2}}{2}\cdot 3{,}26 \cdot \sqrt[2]{\frac{350{,}52}{12}} = 65{,}7+\frac{17{,}61909}{\sqrt{2}} \approx 78{,}15858$
(95%)			

Vi kan her til sidst konkludere, at der er med 95% signifikant forskel mellem de 3 Basketball Spillere. Dette kunne vi ses ved at F-Ratio var større end F-Tabel og dette resulterede at vi under LSD-testen kunne se at Michaels score var anderledes sammenlignet med Damon og Allen, da de havde næsten den samme score i LSD-Intervallet. Alt i alt, kan vi sige at Michael er den bedste spiller, da han har scoret flest i Basketball 😊

6) According to the Harvard Business Review (in the article: "How to Spend Way Less Time on Email Every Day"), the average professional checks his/her emails 15 times per day. The data represent a sample of the number of times/years, that 7 employees in a company check their emails:

5460 5900 6090 6310 7160 8440 9930

Use R to find out: which one of the following statements is correct?

A. We can be 99% confident that the mean number of times that the employees of this company check their email each year is between 4785 and 9298.

B. We can be 99% confident that the mean number of times that the employees of this company check their email is not significantly different from that of the "average professional".

C. None of the previous responses is correct.

D. A and B are correct.

- Vi skal starte med at finde Middelværdien for den Email-datasæt som vi har fået udgivet.

Summen af Email	$5460+5900+6090+6310+7160+8440+9930=49290$
Middelværdi af Email	$\frac{49290}{7}\approx 7041{,}429$

- Vi kan se, at vi har i Svar A har fået angivet Konfidensintervaller: 4785 < mu < 9298, hvilket betyder at hvis den udregnede Middelværdi indsættes er: 4785 < 7041,429 < 9298
- Derfor kan vi sige, at Svar A er rigtigt!

- Hvorimod i Svar B, kan det ses at vi har fået angivet en Konfidensniveau på 99%, hvilket betyder at der er 1% risiko for at vi begår en type 1 error, dette vil sige at vi afviser en sandt null-hypotese.

- Derfor kan vi sige, at Svar B er rigtigt, da vi kan med 99% garanti sige at der er ingen forskel i den gennemsnitlige tjekning af email hos medarbejderne, sammenlignet med den gennemsnitlige professionelle.

- Fordi vores Svar A og B er begge rigtige, kan vi også derfor sige at vores Svar D er rigtigt!

7) The number of children born in 7 towns in a region is:

7540 8421 8560 7412 8953 7859 6098

Find the 99% confidence interval for the mean number of children born annually per town.

- Vi kan se, at Opgaven går ude på at anvende One-Sample T-Test Hypothesis Testing.
- Vi starter allerførst med at finde Middelværdien for selve de 7 byer.

Summen af 7 Byer	$\mathbf{7540+8421+8560+7412+8953+7859+6098 = 54843}$
Middelværdien af 7 Byer	$\mathbf{\frac{54843}{7} \approx 7834,714}$

- Nu skal vi finde Standardafvigelsen.

$\mathbf{7834,714 - 7540 \approx 294,714}$	$\mathbf{8421 - 7834,71}$ $\mathbf{\approx 586,286}$	$\mathbf{8560 - 7834,71}$ $\mathbf{\approx 725,286}$	$\mathbf{7834,714 - 7412 \approx 422,714}$	$\mathbf{8953 - 7834,71}$ $\mathbf{= 1118,28}$	$\mathbf{7859 - 7834,71}$ $\mathbf{\approx 24,286}$	$\mathbf{7834,714 - 6098 = 1736,71}$
Kvadrering						
$\mathbf{294,714^2 \approx 86856,3}$	$\mathbf{586,286^2 \approx 343731,}$	$\mathbf{725,286^2 \approx 526039,}$	$\mathbf{422,714^2 \approx 178687,}$	$\mathbf{1118,286^2 \approx 1250564}$	$\mathbf{24,286^2 \approx 589,809}$	$\mathbf{1736,714^2 \approx 3016176}$
Summen af Kvadraterne						
$\mathbf{86856,34+343731,3+526039,8+178687,1+1250564+589,8098+3016176 \approx 5402644}$						

- For, at kunne finde den endelige Standardafvigelse skal vi derfor dividere med (N+1).

$$\mathbf{\sqrt[2]{\frac{5402644}{7+1}} = \frac{\sqrt{1350661}}{\sqrt{2}} \approx 821,7849}$$

- Nu mangler vi bare at finde den kritiske værdi gennem frihedsgraden og konfidensniveauet.

$$\boldsymbol{Kritisk\ værdi = df(N-1)\ og\ K\%}$$
$$\boldsymbol{Kritisk\ værdi = df(7-1)\ og\ 99\% = 0,01}$$
$$\boldsymbol{Kritisk\ Værdi = 2,44}$$

- Nu skal vi sætte alle værdierne fra vores Udregning af Standardafvigelse og Middelværdi inde i Formel for Population Mean i One-Sample

$$\boldsymbol{\bar{x} - t_{N-1} \cdot \frac{s}{\sqrt[2]{N}} < \mu < \bar{x} + t_{N-1} \cdot \frac{s}{\sqrt[2]{N}}}$$
$$\mathbf{7834,714 - 2,44 \cdot \frac{821,7849}{\sqrt[2]{7}} = 7834,714 - \frac{2005,155}{\sqrt{7}} \approx 7076,837}$$
$$\mathbf{7834,714 + 2,44 \cdot \frac{821,7849}{\sqrt[2]{7}} = 7834,714 + \frac{2005,155}{\sqrt{7}} \approx 8592,591}$$

- Eftersom vi fået udregnet Intervallerne, fås Konfidensintervallerne for Populations Middelværdien til at se følgende ud: $\mathbf{7076,837 < 7834,714 < 8592,591}$
- Fordi vores Populations Middelværdi passer ind i Konfidensinterval, kan vi med 99% sikkerhed sige, at null-hypotesen er afkræftet 😊

8) We want to evaluate three different methods to lower the blood pressure of individuals that have been diagnosed with high blood pressure. Eighteen subjects are randomly assigned to three groups (6 per group): the first group takes medication, the second group exercises, and the third one follows a specific diet. After four weeks, the reduction in each person's blood pressure is recorded. Is there a significant difference among the reduction obtained from each of the three methods? If yes, which method was more effective?

Medication	Exercise	Diet
12	14	6
8	9	10
11	2	5
17	5	9
16	7	8
15	4	6

- Vi starter allerførst med at udregne Middelværdi for hver kolonne af Forsøg på Reducering af Blodtryk.

Summen á Medikament	$\mathbf{12+8+11+17+16+15=79}$
Middelværdi	$\mathbf{\frac{79}{6}\approx 13,16667}$

Summen á Træning	$\mathbf{5+9+2+0+1+3=20}$
Middelværdi	$\mathbf{\frac{20}{6}=\frac{10}{3}\approx 3,333333}$

Summen á Kost	$\mathbf{6+10+5+9+8+0=38}$
Middelværdi	$\mathbf{\frac{38}{6}=\frac{19}{3}\approx 6,333333}$

- Nu skal vi til at udregne selve den totale middelværdi, som er vores "sum of mean" og det betyder at vi lægger alle middelværdierne sammen til en.

Summen af Forsøgernes Middelværdi	$\mathbf{13,16+3,33+6,33=22,82}$
Totale Middelværdi af Forsøgerne	$\mathbf{\frac{22,82}{3}\approx 7,606667}$

- Nu skal vi udregne summen af kvadraterne for alle Forsøg.
- Dette betyder, at vi tager den enkelte dataværdi og trækker det fra den totale middelværdi af kunstgøderne. Vi viser den enkelte eksempel herhenne.

Statistisk Dataanalyse

Sum of Square Total	$(12-7{,}61)^2+(8-7{,}61)^2+(11-7{,}61)^2+(17-7{,}61)^2+(16-7{,}61)^2+(15-7{,}61)^2+(5-7{,}61)^2+(9-7{,}61)^2+(2-7{,}61)^2+(0-7{,}61)^2+(1-7{,}61)^2+(3-7{,}61)^2+(6-7{,}61)^2+(10-7{,}61)^2+(5-7{,}61)^2+(9-7{,}61)^2+(8-7{,}61)^2+(6-7{,}61)^2 = 426{,}9578$
Sum of Square Between	$6\cdot(13{,}16-7{,}61)^2 = 184{,}815$
Sum of Square Between	$6\cdot(3{,}33-7{,}61)^2 \approx 109{,}9104$
Sum of Square Between	$6\cdot(7{,}33-7{,}61)^2 \approx 0{,}4704$
Sum of Total Square Between	$184{,}815+109{,}9104+0{,}4704 \approx 295{,}1958$

- Nu skal vi udregne Sum Square Residual og dette er muligt ved at trække Summen af Square-Total fra Summen af Sum of Total Square Between

SS-Total - SS-Between	$\mathbf{426,9578-295,1958=131,762}$

- Nu skal vi udregne de adskillige ting som er vist i tabellen på Præsentationen.

X	Sum of Squares	d.f.	Mean	F-Ratio
Factor	SS-Between = 295,1958	Antal gruppe - 1 => 3-1 = 2	Dividere Total Square Between med 2 $\frac{295{,}1958}{2} = 147{,}5979$	Dividere Mean Faktor med Mean Residual $\frac{147{,}5979}{8{,}235} \approx 17{,}92324$
Residual	$\mathbf{426,9578-295,1958=131,762}$	$18-2=16$	Dividere SS-Between Residual med df-Residual $\frac{131{,}76}{16} = 8{,}235$	
Total	SS-Total = 426,9589	Antal dataværdier => 18-1 = 17		

- Nu, hvor frihedsgraden er blevet fundet, skal vi derfor finde den kritiske værdi som er markeret med blå.

k-1, N-k	$(3{,}68)$

- Vi kan se, at selve F-Ratio er større end den Kritiske Værdi og derfor kan vi sige at det er følgende:
- F-Ratio > F-Tabel
- Fordi F-Ratio er Større end F-Tabel, kan vi afvise Null-Hypotesen, hvilket betyder at der er en forskel mellem middelværdier af de 3 Kunstgødninger.
- Derfor skal vi nu anvende LSD-Testen for at kunne finde ud af "specifikt", hvilken Forsøg adskiller sig fra den ene og om der er andre som ligner det samme.

F-Ratio > F-Tabel			LSD-Intervaller
Forsøg	Middelværdi	Nedre Grænse	Øvre Grænse

Medikament	13,16	11,00	15,31
Træning	3,33	1,17	5,48
Kost	6,33	4,17	8,48
(95%,12)			

Vi kan her til sidst konkludere, at der er en signifikant forskel mellem de 3 forsøger til at reducere blodtrykket. I tilfældet, kan det ses at Medicin/Medikamentet er det mest effektive, da det har det højeste middelværdi sammenlignet med de andre metoder. Vi kan se, at der er ingen effektivitet i brug af metoder som Træning og Kost, da de resulterer den samme effektivitet. Derfor kan det siges, at Medikament er noget anderledes.