LECTURE 4. HYPOTHESIS TESTING ONE SAMPLE

Hypothesis testing: One sample

- What is inferential statistics?
- Hypothesis testing for the mean
- Confidence Intervals for the mean
- Type error I and type error II

Chapter 4: Assignments
1. 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 Which one of these 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 bruger det vi har lært fra Præsentationen til at opskrive de værdier som vi skal finde.
- N = antal data / dataene længde = 7 (det ved vi godt.)
- S = Standardafvigelsen =?
- X = Middelværdi af Sample =?
- MU = Populations Middelværdi =?
- Vi starter med at finde Middelværdien og derefter Standardafvigelsen.

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

Subtraktion	7041,429-5460	7041,429-5900	7041,429-6090	7041,429-6310	7160-7041,429	8440-7041,429	9930-7041,429
Svar	1581,429	1141,429	951,429	731,429	118,571	1398,571	2888,571
Kvadrering	$\mathbf{1581,429^2 \approx 2500918}$	$\mathbf{1141,429^2 \approx 1302860}$	$\mathbf{951,429^2 \approx 905217,1}$	$\mathbf{731,429^2 \approx 534988,4}$	$\mathbf{118,571^2 \approx 14059,08}$	$\mathbf{1398,571^2 \approx 1956001}$	$\mathbf{2888,571^2 \approx 8343842}$

Summen	$\mathbf{2500918 + 1302860 + 905217,1 + 534988,4 + 14059,08 + 1956001 + 8343842 \approx 1,555789 \cdot 10^7}$
Standardafvigelse	$\sqrt[2]{\frac{\mathbf{1,555789 \cdot 10^7}}{\mathbf{7-1}}} \approx \mathbf{1610,274}$

- Nu skal vi finde den frihedsgraden og derved også den kritiske værdi.
- Vi kan se, at vores N-1 = 7-1 og at vi har en konfidensniveau på 99%=0,01
- Og derved kan vi se ude fra vores Student T-test tabel, at det er 3,7074.
- Nu skal vi anvende følgende formel til at finde frem til om populations middelværdien ligger mellem intervallerne.

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

- Vi kan i vores tilfælde se, at eftersom vi har udregnet vores værdier, kan vi bare indsætte dem inde i formlen.

$$7041,429 - 3,7074 \cdot \frac{1610,274}{\sqrt[2]{7}} = 7041,429 - \frac{5969,93}{\sqrt{7}} \approx 4785,008$$

$$7041,429 + 3,7074 \cdot \frac{1610,274}{\sqrt[2]{7}} = 7041,429 + \frac{5969,93}{\sqrt{7}} \approx 9297,85$$

- I vores tilfælde kan vi se at vores interval bliver 4785,008 < 7041,429 < 9297,85
- Derfor er vores svar A den rigtige 😊

2. A job advisor claims that the average salary for engineers is 24.000 euros/year. Ten engineers are randomly selected, and the mean of their salaries is 23450 euros/year and a standard deviation of 400 euros/year. Is there evidence (at 95% confidence level) to reject the statement of the job advisor?

- Vi kan løse denne opgave på to måder: Den simple og den lange.

SIMPEL-UDGAVE

- Kigger man på den simple udgave, kan man se at vi har 10 tilfældige udvalgte ingeniører som har en samlet løn på 23,450 euro pr. år. Denne samlet løn er svarende til den gennemsnitlige løn pr. år for alle ingeniører og derved kan vi fra et logisk synspunkt afvise job advisors statement og bekræfte alternativ-hypotesen.
- HUSK: Alternativ Hypotesen bekræftes fordi vi kan se at gennemsnitsløn for Populationen og Sample er anderledes.

LANGE-UDGAVE

- Vi starter med at nedskrive de værdier som vi kan se fra opgaven.
- N = tilfældige udvalgte ingeniører = 10
- S = 400 euro pr. år
- X = 23450
- MU = 24,000
- Nu anvendes følgende formel for at indplacere vores populations middelværdi i konfidensintervallet.

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

$$23450 - 2,2622 \cdot \frac{400}{\sqrt[2]{10}} = -\frac{904,88}{\sqrt{10}} + 23450 \approx 23163,85$$

$$23450 + 2,2622 \cdot \frac{400}{\sqrt[2]{10}} = \frac{904,88}{\sqrt{10}} + 23450 \approx 23736,15$$

- Derfor kan vi se at vores konfidensinterval bliver følgende: 23163,85 < 24000 < 23736,15
- Vi kan se, at 24000 ligger udenfor intervallet, derfor afviser vi null-hypotesen og bekræfter alternativ-hypotese.

3. 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 skal udregne konfidensintervallet for middelværdien af børn født i en by.
- Vi starter allerførst med at opskrive den værdi som vi kender fra opgaven.
- N = antal data / dataenes længde = 7 (det ved vi godt.)
- S = Standardafvigelsen =?
- X = Middelværdi af Sample =?
- MU = Populations Middelværdi =?
- Vi starter med at finde Middelværdien og derefter Standardafvigelsen.

Sum	$7540 + 8421 + 8560 + 7412 + 8953 + 7859 + 6098 = 54843$
Middelværdi	$\frac{54843}{7} \approx 7834,714$

Subtraktion	7834,714-7540	8421-7834,714	8560-7834,714	7834,714-7412	8953-7834,714	7859-7834,714	7834,714-6098
Svar	294,714	586,29	725,29	422,714	1118,286	24,286	1736,714
Kvadrering	$29.714^2 \approx 882,9218$	$586,29^2 \approx 343736$	$725,29^2 \approx 526045,6$	$422,714^2 \approx 178687,1$	$1118,286^2 \approx 1250564$	$24,286^2 \approx 589,8098$	$1736,714^2 \approx 3016176$

Summen	$882,9218 + 343736 + 526045,6 + 178687,1 + 1250564 + 589,8098 + 3016176 \approx 5316681$
Standardafvigelse	$\sqrt[2]{\frac{5316681}{7-1}} = \frac{\sqrt{1772227}}{\sqrt{2}} \approx 941,336$

- Nu skal vi finde den frihedsgraden og derved også den kritiske værdi.
- Vi kan se, at vores N-1 = 7-1 og at vi har en konfidensniveau på 99%=0,01
- Og derved kan vi se ude fra vores Student T-test tabel, at det er 3,7074.
- Nu skal vi anvende følgende formel til at finde frem til om populations middelværdien ligger mellem intervallerne.

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

$$7834,714 - 3,7074 \cdot \frac{941,336}{\sqrt[2]{7}} = 7834,714 - \frac{3489,909}{\sqrt{7}} \approx 6515,652$$

$$7834,714 + 3,7074 \cdot \frac{941,336}{\sqrt[2]{7}} = 7834,714 + \frac{3489,909}{\sqrt{7}} \approx 9153,776$$

- I vores tilfælde kan vi se at vores interval bliver 6515,652 < 7834,714 < 9153,776

4. A survey of 50 students found that the mean age of their bicycles was 5.6 years. Assuming the standard deviation of the population is 0.8 years, which of these statements is correct?

A. Based on the sample of 50 students, we can be 90% confident that the mean age of all bicycles is between 5.297 and 5.903.

B. Based on the sample of 50 students, we can be 99% confident that the mean age of all bicycles is between 5.297 and 5.903.

C. A and B are correct.

D. None of them is correct.

- Vi skal starte med at finde de nedskrevne værdier som er i teksten.
- N = antal elever = 50
- S = antal år = 0,8
- X = Middelværdi af cyklernes alder = 5,6
- MU = er ikke beskrevet i teksten = 0.
- Fordi vi ikke har fået opgivet en datasæt, kan vi derfor gå direkte til sagen med frihedsgraden og der finder den kritiske værdi.
- Vi kan se, at vores N-1 = 50-1 og at vi har en konfidensniveau på 90%=0,10.
- Student t-test viser at two-tailed med 90% er: +-1,6766
- Student t-test viser at two-tailed med 99% er: +-2,68
- Nu skal vi anvende følgende formel til at finde frem til om samples middelværdien ligger mellem intervallerne.
- Vi kan se, at vores Udregning med 90% bliver følgende:

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

$$5,6 - 1,6766 \cdot \frac{0,8}{\sqrt[2]{50}} = 5,6 - \frac{0,268256}{\sqrt{2}} \approx 5,410314$$

$$5,6 + 1,6766 \cdot \frac{0,8}{\sqrt[2]{50}} = 5,6 + \frac{0,268256}{\sqrt{2}} \approx 5,789686$$

- Med 90% Konfidensinterval har vi indsat samples middelværdi og fået: 5,41 < 5,6 < 5,76.
- Vi kan dermed bekræfte null-hypotesen
- Vi kan se, at vores Udregning med 99% bliver følgende:

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

$$5,6 - 2,68 \cdot \frac{0,8}{\sqrt[2]{50}} = 5,6 - \frac{0,4288}{\sqrt{2}} \approx 5,296793$$

$$5,6 + 2,68 \cdot \frac{0,8}{\sqrt[2]{50}} = 5,6 + \frac{0,4288}{\sqrt{2}} \approx 5,903207$$

- Med 99% Konfidensinterval har vi indsat samples middelværdi og fået: 5,29 < 5,6 < 5,90.
- Vi kan dermed bekræfte null-hypotesen
- Vi kan derfor sige, at vores svarmulighed med C er korrekt, da A og B er rigtige.

5. The following data represent a sample of the assets (in millions of euros) of 30 motherboards computer hardware manufacturers in Europe. Find the 90% confidence interval of the mean.

12.23	16.56	4.39
2.89	1.24	2.17
13.19	9.16	1.42
73.25	1.91	14.64
11.59	6.69	1.06
8.74	3.17	18.13
7.92	4.78	16.85
40.22	2.42	21.58
5.01	1.47	12.24
2.27	12.77	2.76

- Vi skal starte med at finde de nedskrevne værdier som er i teksten.
- N = antal elever = 30
- S = ikke gengivet på forhånd = ?
- X = Ikke gengivet på forhånd = ?
- MU = Ikke gengivet på forhånd = ?
- Vi kan starte med, at udregne Middelværdien og derefter Standardafvigelsen.

Summen	$12.23 + 2,89 + 13,19 + 73,25 + 11,59 + 8,74 + 7,92 + 40,22 + 5,01 + 2,27 + 16,56 + 1,24 + 9,16 + 1,91 + 6,69 + 3,17 + 4,78 + 2,42 + 1,47 + 12,77 + 4,39 + 2,17 + 1,42 + 14,64 + 1,06 + 18,13 + 16,85 + 21,58 + 12,24 + 2,76 \approx 332,72$
Middelværdien	$\frac{332,72}{30} \approx 11,09067$

Subtraktion	1,14	5,47	6,7	8,2	9,85	8,92	2,1	1,93	9,6
Subtraktion	62,16	9,18	3,55	0,5	4,4	10,03	2,35	7,92	7,0
Subtraktion	3,17	6,31	5,76	29,13	8,67	10,49	6,08	9,62	1,1
Subtraktion	8,82	1,68	8,33						
Kvadrering	1,2996	29,9209	44,89	67,24	97,0225	79,5664	4,41	3,7249	93,508
Kvadrering	3863,866	84,2724	12,6025	0,25	19,36	100,6009	5,5225	62,7264	49,561
Kvadrering	10,0489	39,8161	33,1776	848,5569	75,1689	110,0401	36,9664	92,5444	1,322
Kvadrering	77,7924	2,8224	69,3889						
Sum	6017,991								

Standardafvigelse	$\sqrt[2]{\frac{6017,991}{\sqrt[2]{30}}} = 33,147$

- Nu skal vi finde selve frihedsgraden og derefter den kritiske værdi.
- Vi kan se, at vi beskæftiger os med konfidensniveau på 90%=0,10.
- Derfor ender vi med at skrive såden: N-1=30-1 og 90%=0,10.
- Udefra vores two-tailed Student T-test kan vi se at vi får kritisk værdi til: +-1,6991
- Nu skal vi bare indsætte de kendte værdier inde i formlen.

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

$$11{,}09067 - 1{,}6991 \cdot \frac{33{,}147}{\sqrt[2]{30}} = 11{,}09067 - \frac{56{,}32007}{\sqrt{30}} \approx 0{,}8080795$$

$$11{,}09067 + 1{,}6991 \cdot \frac{33{,}147}{\sqrt[2]{30}} = \frac{56{,}32007}{\sqrt{30}} + 11{,}09067 \approx 21{,}37326$$

- Nu kan vi se at fordi vi fik vores sample middelværdi til at være 11,09, kan vi derfor plotte det inde i følgende udregnede konfidensinterval: 0,80 < 11,09 < 21,37.
- Derfor kan vi bekræfte vores null-hypotese 😊