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.

I tilfældet, har vi skrevet kommandoen op på følgende måde

```
Havard <- c(5460,5900,6090,6310,7160,8440,9930)
shapiro.test(Havard)
res.ttest <- t.test(Havard)
res.ttest
```

Resultatet bliver følgende:

```
data:  Havard
t = 11.569, df = 6, p-value = 2.509e-05
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 5552.174 8530.683
sample estimates:
mean of x
 7041.429
```

I tilfældet, kan det ses at vi har en p-værdi som er under 0,05.
Det betyder at det er signifikant forskel!

Vores middelværdi ligger i konfidensintervallet og passer godt!

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 se, at selve job adviser siger at gennemsnitlige løn for alle ingeniører er 24000 Euro pr. år.
Men vores sample har allerede en gennemsnitligt løn som er 23450 Euro pr. år!
Derfor kan vi naturligvis afvise job advisors statement!

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 starter med at skrive følgende kommando op!

```
"Opgave 3"
Children <- c(7540,8421,8560,7412,8953,7859,6098)
shapiro.test(Children)
res.ttest <- t.test(Children,conf.level=0.99)
res.ttest
```

Oven på skal det understreges, at shapiro.test er normal distribution og res.ttest er one sample t-test. Udefra dette kan vi sige, at vi får følgende interval!

```
        One Sample t-test

data:  Children
t = 21.845, df = 6, p-value = 6.012e-07
alternative hypothesis: true mean is not equal to 0
99 percent confidence interval:
 6505.022 9164.407
sample estimates:
mean of x
 7834.714
```

Vores konfidensinterval bliver følgende: 6505,022 < 7834,714 < 9164,07

3. 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 starter med at finde N-værdien. Vores N-værdi er vores population størrelsen som er 50. Men vi skal også finde standardafvigelsen og i tilfældet, kan det ses at det er 0,8.

1. Brug formlen fra undervisningen.

- Vi burger formlen fra undervisningen, som er følgende:

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

$$\bar{x} = 5{,}6$$

$$t_{N-1} = t_{50-1} = t_{49}$$

2. Udregn midterste standardafvigelse divideret med mængde.

- Vi divider standardafvigelsen med mængden!

$$\frac{s}{\sqrt[2]{N}} = \frac{0{,}8}{\sqrt[2]{50}} = \frac{0{,}8}{\sqrt{7}} \approx 0{,}3023716$$

3. Find summen af middelværdien og divisionsbrøkerne sammen
- Hvis man tilføjer middelværdien og s/n med hinanden, så ville man få 5,9

$$5{,}6 + 0{,}3 \approx 5{,}9$$

4. Find subtraktionen af middelværdien og divisionsbrøkerne sammen
- I dette tilfælde, finder vi kun ud af hvilken del af ligningen har den samme metode som vi brugte tidligere.
- I tilfælde, finder vi den nedre del af intervallet gennem subtraktionen.

$$5{,}6 - 0{,}3 = 5{,}3$$

5. Conclude the Interval
6. Konkludere Intervallet
- I tilfældet, kan det ses at vi har fundet intervallet til at være: (5,3;5.9).
- Vi kan på grund af svaret sige, at på baggrund af sample med 50 elever, kan vi være 90% til 99% sikker på at middelværdien af alle cykler er mellem det givne interval over (5.3;5.9). Derfor er svarmulighed B, den rigtige svar.

4. 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 opskriver værdierne på følgende måde i R og derved løser det.

```
"Opgave 5"
Computer <- c(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
shapiro.test(Computer)
res.ttest <- t.test(Computer,conf.level=0.90)
res.ttest
```

- Vi kan se, at kommando kommer ud som svar her:

```
        One Sample t-test

data:  Computer
t = 4.2169, df = 29, p-value = 0.0002214
alternative hypothesis: true mean is not equal to 0
90 percent confidence interval:
  6.621856 15.559478
sample estimates:
mean of x
 11.09067
```