CHAPTER 5. HYPOTHESIS TESTING (TWO SAMPLES)

Hypothesis testing: Two samples

- Comparison of means (independent samples)
- Paired data test (dependent samples)
- Selecting the right statistical test

Statistisk Dataanalyse

Chapter 5: Assignments (Hypothesis testing: Two samples)

1. The average relative humidities (%) for two cities, are: 72.9 (city A) and 70.8 (city B), based on 25 measurements of relative humidity in each city. The standard deviations of these measurements are: 2.5 (city A) and 2.8 (city B). Based on the samples, can it be concluded with 95% confidence level that the relative humidity in the two cities is significantly different?

- **Vi starter allerførst med at nedskrive de værdier som vi kender:**
- X1 = 72,9 og X2 = 70,8
- N = 25
- S1 = 2,5 og S2 = 2,8
- **Genkend den rigtige test**
- Vi kan se, at fordi vi har at gøre med at måle den relative hedebølge mellem to byer som er (adskillige) fra hinanden, så har vi at gøre med two-sample independent t-test.
- **Vi udvælger den rigtige formel for sandheden af null-hypotesen i independent t-test.**

$$\frac{(\overline{x_1} - \overline{x_2}) - (\mu_1 - \mu_2)}{s_{\overline{x_1} - \overline{x_2}}} \approx t_{N1+N2-2}$$

- Værdierne indsættes i formlen på del.

$$\frac{(72{,}9 - 70{,}8) - (0 - 0)}{s_{\overline{x_1} - \overline{x_2}}}$$

- Vi mangler at indsætte værdier til standardafvigelsen i tælleren og derfor anvendes følgende formel:

$$s_{\overline{x_1} - \overline{x_2}} = \sqrt[2]{\frac{(N1-1)\cdot S1^2 + (N2-.1)\cdot S2^2}{N1+N2-2}} \cdot \sqrt[2]{\frac{1}{N1} + \frac{1}{N2}}$$

- Nu indsættes værdierne inde i selve formlen.

$$\sqrt[2]{\frac{(25-1)\cdot 2{,}5^2 + (25-.1)\cdot 2{,}8^2}{25+25-2}} \cdot \sqrt[2]{\frac{1}{25} + \frac{1}{25}} = \sqrt{2} \cdot 0{,}5363581 \approx 0{,}7585249$$

- Nu tager vi værdien fra Standardafvigelsen og derved indsætter det inde i selve formlen fra tidligere.

$$\frac{(72{,}9 - 70{,}8) - (0 - 0)}{0{,}758} \approx 2{,}770449$$

- **Nu skal vi finde vores frihedsgrad og derefter den kritiske værdi for konfidensintervallet.**
- **Vi beskæftiger os med 95%=0,05 og skal derfor anvende følgende for frihedsgraden ved højre side af den første formel.**

$$t_{N1+N2-2}$$
$$t_{25+25-2} = t_{48}$$

- **Gennem vores Student t-tabel, kan vi se at vi har fået følgende kritisk værdi ved two-tailed t-test.**

$$t_{48}\ og\ 95\% = 2{,}0106$$

- Vi kan se, at vi får følgende konfidensinterval: -2,01 < 2,77 < 2,01.

- Derfor kan vi afvise null-hypotesen og acceptere alternativ hypotesen, hvor der er signifikant forskel i hedebølge/varme ved de to byer.

2. The dean of a university claims that the average scores in Software Engineering education, of those students that were educated in public high schools is higher than the average scores of those students that were educated in private high schools. A sample of the 50 students from each group is randomly selected. The average scores of the students from public schools are 8.6 (out of 10) and with a standard deviation equal to 3.3. The average scores of the students from private schools are 7.9 (out of 10) and with a standard deviation equal to 3.3. Are we 90% confident that the statement of the dean is true?

- **Vi starter allerførst med at nedskrive de værdier som vi kender:**
- X1 = 8,6 og X2 = 7,9
- N = 10
- S1 = 3,3 og S2 = 3,3
- **Genkend den rigtige test**
- Vi kan se, at fordi vi sammenligner middelværdi og standardafvigelsen mellem to skoler, hvor den ene er kommune/stats drevet og hvor den anden er privat drevet, kan vi derfor sige at der er snak om en independent t-test.
- **Vi udvælger den rigtige formel for sandheden af null-hypotesen i independent t-test.**

$$\frac{(\overline{x_1} - \overline{x_2}) - (\mu_1 - \mu_2)}{s_{\overline{x_1} - \overline{x_2}}} \approx t_{N1+N2-2}$$

- Værdierne indsættes i formlen på del.

$$\frac{(8{,}6 - 7{,}9) - (0 - 0)}{s_{\overline{x_1} - \overline{x_2}}}$$

- Vi mangler at indsætte værdier til standardafvigelsen i tælleren og derfor anvendes følgende formel:

$$s_{\overline{x_1} - \overline{x_2}} = \sqrt[2]{\frac{(N1-1)\cdot S1^2 + (N2-.1)\cdot S2^2}{N1+N2-2}} \cdot \sqrt[2]{\frac{1}{N1} + \frac{1}{N2}}$$

- Nu indsættes værdierne inde i selve formlen.

$$\sqrt[2]{\frac{(10-1)\cdot 3{,}3^2 + (10-.1)\cdot 3{,}3^2}{10+10-2}} \cdot \sqrt[2]{\frac{1}{10} + \frac{1}{10}} = \frac{3{,}381494}{\sqrt{5}} \approx 1{,}51225$$

- Nu tager vi værdien fra Standardafvigelsen og derved indsætter det inde i selve formlen fra tidligere.

$$\frac{(72{,}9 - 70{,}8) - (0 - 0)}{1{,}51225} \approx 1{,}388659$$

- **Nu skal vi finde vores frihedsgrad og derefter den kritiske værdi for konfidensintervallet.**
- **Vi beskæftiger os med 90%=0,10 og skal derfor anvende følgende for frihedsgraden ved højre side af den første formel.**

$$t_{N1+N2-2}$$
$$t_{10+10-2} = t_{18}$$

- Gennem vores Student t-tabel, kan vi se at vi har fået følgende kritisk værdi ved two-tailed t-test.

$$t_{98}\ og\ 90\% = 1{,}734$$

- Vi kan se, at vi får følgende konfidensinterval: -1,73 < 1,38 < 1,73.
- Ved 90% konfidensniveau kan vi sige, at der er (ingen) bevis på at dekanens påstand sandt og derved kan det siges at der er ingen forskel i middelværdi på karakterindsats mellem den kommunedrevet skole og privatskolen.

3. We would like to know if the concentration of a compound in two brands of yogurt is different. We select 50 bottles of each type. The average concentration in one of the brands is 88.42 mg/L and in the other one is 80.61 mg/L. The standard deviations of the populations are 5.62 mg/L and 4.83 mg/L, respectively. Can we be 95% confident that there is a significant difference among the two brands? What about 99% confident?

- Vi starter allerførst med at nedskrive de værdier som vi kender:
- X1 = 88,42 og X2 = 80,61
- N = 50
- S1 = 5,62 og S2 = 4,83
- Genkend den rigtige test
- Fordi vi sammenligner middelværdier af to forskellige Yoghurt Brands med hinanden, kan vi derfor sige at den korrekte test til denne opgave er Independent t-test.
- Vi udvælger den rigtige formel for sandheden af null-hypotesen i independent t-test.

$$\frac{(\overline{x_1} - \overline{x_2}) - (\mu_1 - \mu_2)}{s_{\overline{x_1} - \overline{x_2}}} \approx t_{N1+N2-2}$$

- Værdierne indsættes i formlen på del.

$$\frac{(88{,}42 - 80{,}61) - (0 - 0)}{s_{\overline{x_1} - \overline{x_2}}}$$

- Vi mangler at indsætte værdier til standardafvigelsen i tælleren og derfor anvendes følgende formel:

$$s_{\overline{x_1} - \overline{x_2}} = \sqrt[2]{\frac{(N1-1)\cdot S1^2 + (N2-.1)\cdot S2^2}{N1+N2-2}} \cdot \sqrt[2]{\frac{1}{N1} + \frac{1}{N2}}$$

- Nu indsættes værdierne inde i selve formlen.

$$\sqrt[2]{\frac{(50-1)\cdot 5{,}62^2 + (50-.1)\cdot 4{,}83^2}{50+50-2}} \cdot \sqrt[2]{\frac{1}{50} + \frac{1}{50}} \approx 1{,}052063$$

- Nu tager vi værdien fra Standardafvigelsen og derved indsætter det inde i selve formlen fra tidligere.

$$\frac{(88{,}42 - 80{,}61) - (0 - 0)}{1{,}052063} \approx 7{,}42351$$

- Nu skal vi finde vores frihedsgrad og derefter den kritiske værdi for konfidensintervallet.
- Vi beskæftiger os med 95%=0,05 og skal derfor anvende følgende for frihedsgraden ved højre side af den første formel.

$$t_{N1+N2-2}$$
$$t_{50+50-2} = t_{98}$$

- Gennem vores Student t-tabel, kan vi se at vi har fået følgende kritisk værdi ved two-tailed t-test.

$$t_{98}\ og\ 95\% = 1{,}98$$

- Vi kan se, at vi får følgende konfidensinterval: -1,98 < 7,42 < 1,98.
- Fordi vores middelværdi ligger uden for konfidensintervallet, kan vi sige at der er en signifikant forskel mellem de to Brands af Yoghurt.

4. We want to know whether or not a certain training program is able to increase the maximum long jump of athletes. We recruit a simple random sample of 20 long jump athletes and measure each of their maximum long jump. Then, we have each athlete use the training program for one month and then measure their maximum long jump again at the end of the month. These are the results (below). Does the training program have any effect on the maximum long jump? (use level of significance = 0.05)

Athlete	Maximum long jump before training program	Maximum long jump after training program
1	3.7	4.0
2	3.3	3.7
3	3.2	3.2
4	4.0	3.7
5	4.2	4.7
6	4.2	4.3
7	4.7	4.7
8	3.7	4.0
9	5.0	5.0
10	4.5	4.8
11	4.0	4.2
12	3.0	3.3
13	2.7	2.8
14	3.2	3.0
15	3.2	3.0
16	4.7	4.7
17	4.0	4.3
18	4.2	4.5
19	4.2	4.5
20	3.8	4.0

Statistisk Dataanalyse

- Vi starter allerførst med at nedskrive de værdier som vi kender:
- X1 = ? og X2 = ?
- N = længden af datasættet af før og efter (begge kolonner) = 19
- S = ? (Bliver udregnet i Excel-tabel, inspireret fra Præsentationen)
- Genkend den rigtige test
- Fordi vi sammenligner athleternes spring før og efter stævningen, skal vi derfor anvende dependent t-test fordi vi sammenligner eller følger ting over tid.
- Fordi vi ikke har fået direkte at vide hvad middelværdien og standardafvigelsen er, derfor skal vi selv regne det ud.

Summen af Before	$3,7 + 3,3 + 3,2 + 4,0 + 4,2 + 4,2 + 3,7 + 5,0 + 4,5 + 4,0 + 3,0 + 2,7 + 3,2 + 3,2 + 4,7 + 4,0 + 4,2 + 4,2 + 3,8 \approx 72,8$
Middelværdi af Before	$X2 = \frac{72,8}{19} \approx 3,831579$
Summen af After	$4,0 + 3,7 + 3,2 + 3,7 + 4,7 + 4,3 + 4,7 + 4,0 + 5,0 + 4,8 + 4,2 + 3,3 + 2,8 + 3,0 + 3,0 + 4,7 + 4,3 + 4,5 + 4,5 + 4,0 \approx 80,4$
Middelværdi af After	$X1 = \frac{80,4}{19} \approx 4,231579$

- Nu udregner vi de relevante værdier til standardafvigelsen som findes i nævneren.

X1	X2	X1-X2	(X1-X2)^2
3,7	4	0,3	0,09
3,3	3,7	0,4	0,16
3,2	3,2	0	0
4	3,7	0,3	0,09
4,2	4,7	0,5	0,25
4,2	4,3	0,1	0,01
4,7	4,7	0	0
3,7	4	0,3	0,09
5	5	0	0
4,5	4,8	0,3	0,09
4	4,2	0,2	0,04
3	3,3	0,3	0,09
2,7	2,8	0,1	0,01
3,2	3	0,2	0,04
3,2	3	0,2	0,04
4,7	4,7	0	0
4	4,3	0,3	0,09
4,2	4,5	0,3	0,09
4,2	4,5	0,3	0,09
3,8	4	0,2	0,04
		4,3	**1,31**
		0,226315789	

- Vi udvælger den rigtige formel for sandheden af null-hypotesen i dependent t-test.
- I tilfældet kan det ses at vi anvender følgende formel.

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

- For standardafvigelsen kan det ses at der skal anvendes et andet formel i dependent test, sammenlignet med independent test.

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

- Vi indsætter værdierne inde selve formlen og får følgende resultat:

$$\sqrt[2]{\frac{1{,}31 - \frac{4{,}3}{19}}{19 - 1}} \approx 0{,}2453664$$

- Nu tages Standardafvigelsesværdien og indsættes i hovedformlen.

$$\frac{(4{,}23 - 3{,}83) - (0)}{\frac{0{,}2453664}{\sqrt[2]{19}}} = \sqrt{19} \cdot 1{,}630215 \approx 7{,}105943$$

- Nu skal vi finde vores frihedsgrad og derefter den kritiske værdi for konfidensintervallet.
- Vi beskæftiger os med 95%=0,05 og skal derfor anvende følgende for frihedsgraden ved højre side af den første formel.

$$t_{N-1}$$
$$t_{19-1} = t_{18}$$

- Gennem vores Student t-tabel, kan vi se at vi har fået følgende kritisk værdi ved two-tailed t-test.

$$t_{18}\ og\ 95\% = 2{,}1009$$

- Vi kan se, at vi får følgende konfidensinterval: -2,10 < 7,10 < 2,10.
- Vi kan derfor sige, at der er en forskel fordi vores værdi ligger uden for konfidensintervallet.
- Derfor kan vi sige, at der er et bevis på at træningsprogrammet har en effekt, siden der er forskel i før og efter hop af atleterne før og efter deltagelse i træningsprogrammet. 😊

Statistisk Dataanalyse