Statistisk Dataanalyse 2023

Instruktør: Vivek Misra

Exercise Class NR5

Solutions to the Tasks

Task 1 - Description

- The average relative humidities (%) for two cities, are: 72.9 (city A) and 70.8 (city B), based in 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?

Task 1 - Solution

- Solution: First we will collect the informations about values, in which we need:

Values
$x_1 = 72{,}9$
$x_2 = 70{,}8$
$N = 25$
$s_1 = 2{,}5$
$s_2 = 2{,}8$

Task 1 - Solution

- Solution: First we need to identify the right statistical test. We can see, that the we have two samples from two cities, and therefore we will chose two-sample independent t-test.
- We will use the formula below:

$$\frac{(\overline{x_1} - \overline{x_2}) - (\mu_1 - \mu_2)}{s_{\overline{x_1} - \overline{x_2}}}$$

- We can see, that the values are inserted inside the formula:

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

Task 1 - Solution

- Now we just need to calculate the lower part of the bracket $s_{\overline{x_1}-\overline{x_2}}$.

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

- Now we will insert the values inside the formula:

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

Task 1 - Solution

- After calculating the standard-deviation values, we will insert them in the lower part of the bracket.

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

- Now we need to calculate the degrees of freedom, and thereafter the critical value for the confidence interval. Because we are working with a confidence level of 90%=0,10, therefore we will intersect the confidence level with the result of the critical value.

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

Task 1 - Solution

- We can see, that we have got the critical value to be
$$t_{48}\ \&95\% = 2{,}0106$$
- From this we can see, that the interval is -2,01<2,77<2,01.
- The alternative hypothesis is acepted, which means that there is a significant difference in the humidity percentage of the two cities.

Task 2 - Description

- 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?

Task 2 - Solution

- Solution: First we will collect the informations about values, in which we need:

Values
$x_1 = 8{,}6$
$x_2 = 7{,}9$
$N = 50$
$s_1 = 3{,}3$
$s_2 = 3{,}3$

- We can see, that we are comparing a Government-Driven School with a Privat School. In the next page we will use the same formula again!

Task 2 - Solution

- Solution: First we need to identify the right statistical test. We can see, that we are comparing two different schools with eachother and therefore we are talking about a two-sample independent t-test.
- We will use the formula below:

$$\frac{(\overline{x_1} - \overline{x_2}) - (\mu_1 - \mu_2)}{s_{\overline{x_1} - \overline{x_2}}}$$

- We can see, that the values are inserted inside the formula:

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

Task 2 - Solution

- Now we just need to calculate the lower part of the bracket $s_{\overline{x_1}-\overline{x_2}}$.

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

- Now we will insert the values inside the formula:

$$\sqrt[2]{\frac{(50-1)\cdot 3{,}3^2 + (50-.1)\cdot 3{,}3^2}{50+50-2}} \cdot \sqrt[2]{\frac{1}{50}+\frac{1}{50}} = 0{,}66$$

Task 2 - Solution

- After calculating the standard-deviation values, we will insert them in the lower part of the bracket.

$$\frac{(72{,}9 - 70{,}8) - (0 - 0)}{0{,}66} \approx 1{,}06$$

- Now we need to calculate the degrees of freedom, and thereafter the critical value for the confidence interval. Because we are working with a confidence level of 90%=0,10, therefore we will intersect the confidence level with the result of the critical value.

$$t_{N1+N2-2}$$
$$t_{50+50-2} = t_{48}$$

Task 2 - Solution

- We can see, that we have got the critical value to be
$$t_{48}\ \&90\% = 1{,}734$$
- From this we can see, that the interval is -1,73<1,38<1,73.
- The null-hypothesis is accepted on the basis of 90% confidence level.
- The reason is clear, and that is because there is no evidence that the deans statement is true, because from the statistical analysis we can see that there is no difference between the means of the grades between the Government-School and the Privat-School.

Task 3 - Description

- 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?

Task 3 - Solution

- Solution: First we will collect the informations about values, in which we need:

Values
$x_1 = 88{,}42$
$x_2 = 80{,}61$
$N = 50$
$s_1 = 5{,}62$
$s_2 = 4{,}83$

- In the next page, we need to use the right statistical test.

Task 3 - Solution

- Solution: First we need to identify the right statistical test. We can see, that we are comparing two different Yoghurt Brands with eachother. Therefore we will conduct a two-sample independent t-test.
- We will use the formula below:

$$\frac{(\overline{x_1} - \overline{x_2}) - (\mu_1 - \mu_2)}{s_{\overline{x_1} - \overline{x_2}}}$$

- We can see, that the values are inserted inside the formula:

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

Task 3 - Solution

- Now we just need to calculate the lower part of the bracket $s_{\overline{x_1}-\overline{x_2}}$.

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

- Now we will insert the values inside the formula:

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

Task 3 - Solution

- After calculating the standard-deviation values, we will insert them in the lower part of the bracket.

$$\frac{(72{,}9 - 70{,}8) - (0 - 0)}{1{,}052063} \approx 7{,}42351$$

Now we need to calculate the degrees of freedom, and thereafter the critical value for the confidence interval. Because we are working with a confidence level of 95%=0,05, therefore we will intersect the confidence level with the result of the critical value.

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

Task 3 - Solution

- We can see, that we have got the critical value to be
$$t_{98}\ \&95\% = 1{,}98$$
- From this we can see, that the interval is -1,98<7,42<1,98.
- The null-hypothesis is rejected and the alternative hypothesis is valid.
- The reason is clear, and that is because the mean is placed outside the confidence level, which essentially means that there is a significant difference between the two Brands of Yoghurt.

Task 4 - Description

- 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

Task 4 - Solution

- Solution: First we will collect the informations about values, in which we need:

Values
$x_1 = ?$
$x_2 = ?$
$N = 20$
$s_1 = ?$
$s_2 = ?$

- In the next page, we need to use the right statistical test.

Task 4 - Solution

- Solution: Because we are comparing an athletes jump before a training program and after training, we can then use a two-sample dependent t-test due to change over time!
- First we will calculate the mean before the program and after.

Sum of Before Program	$3,7+3,3+3,2+4,0+4,2+4,2+3,7+5,0+4,5+4,0+3,0+2,7+3,2+3,2+4,7+4,0+4,2+4,2+3,8 \approx 72,8$
Mean of Before Training Program	$X2 = \frac{72,8}{19} \approx 3,831579$
Sum of After	$4,0+3,7+3,2+3,7+4,7+4,3+4,7+4,0+5,0+4,8+4,2+3,3+2,8+3,0+3,0+4,7+4,3+4,5+4,5+4,0 \approx 80,4$
Mean of After Training Program	$X1 = \frac{80,4}{19} \approx 4,231579$

Task 4 - Solution

- **To the right we have shown that the variance is equal to 1,31.**
- Now we will process further, and use the right formula for two sample dependent t-test.

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

Now we need to calculate the standard deviation according to the formula shown below:

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

The valuers are inserted inside.

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

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	

Task 4 - Solution

- Now, after calculating the standard deviation. We will implement the two-sample dependent t-test formula as we showed earlier.

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

- Now the values will be insert inside:

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

- In the next page, the degrees of freedom and t-test value will be found.

Task 4 - Solution

- Now we need to calculate the degrees of freedom, and thereafter the critical value for the confidence interval. Because we are working with a confidence level of 95%=0,05, therefore we will intersect the confidence level with the result of the critical value.

$$t_{N-1}$$
$$t_{20-1} = t_{19}$$

Task 4 - Solution

- We can see, that we have got the critical value to be
$$t_{19}\ \&\ 95\% = 2{,}1009$$
- From this we can see, that the interval is -2,10<7,10<2,10.
- The null-hypothesis is rejected and the alternative hypothesis is valid.
- The reason is clear, and that is because there is a significant difference between the means of the jump before the training program and after the training program
- And therefore we can state, that the training program has an effect, since we have calculated a significant difference between the means!

Tak for idag!

Instruktør: Vivek Misra