Chair of Decision Sciences and Systems
TUM School of Computation, Information and Technology
Technical University of Munich

Business Analytics & Machine Learning
Tutorial sheet 1: Statistics – Solution

Prof. Dr. Martin Bichler, Prof. Dr. Jalal Etesami
Julius Durmann, Markus Ewert, Johannes Knörr, Yutong Chao

Exercise T1.1 *Effect of tax on consumption*

The following table contains data of 10 individuals' consumption levels before and after a tax increase, measured by an index value. High index values correspond to high consumption levels. The rows represent individuals' identifiers i, their index values prior to the tax increase a_i, and after the tax increase b_i.

i	1	2	3	4	5	6	7	8	9	10
a_i	27	31	23	35	26	27	26	18	22	21
b_i	40	36	43	34	25	41	32	29	21	36
$d_i = a_i - b_i$	-13	-5	-20	1	1	-14	-6	-11	1	-15

a) Perform a hypothesis test in order to find out whether there is a significant $(\alpha = 0.05)$ difference between consumption levels prior to the tax increase and consumption levels after the tax increase. Assume, that the difference is normally distributed.

b) Verify your result by applying `stats.ttest_rel()` in Python using the *SciPy* package.

Solution

a) 1) Two dependent samples

2) $\mu_D = \mu_{\text{before}} - \mu_{\text{after}}$

$$H_0 : \mu_{\text{before}} = \mu_{\text{after}} \iff H_0 : \mu_D = \mu_0 = 0$$
$$H_1 : \mu_{\text{before}} \neq \mu_{\text{after}} \iff H_1 : \mu_D \neq \mu_0 = 0$$

3) Paired t-test:

$$t_0 = \frac{\bar{d} - \mu_0}{s_d}\sqrt{n}, \qquad \bar{d} = \frac{1}{n}\sum_i d_i, \qquad s_d = \sqrt{\frac{\sum_i (d_i - \bar{d})^2}{n-1}}$$

Thus, the average difference is $\bar{d} = -8.1$. The standard deviation of differences is $s_d \approx 7.5931$. The test statistic now is $t_0 \approx \frac{-8.1-0}{7.5931}\sqrt{10} = -3.3734$.

4) $\alpha = 0.05$

5) $t^c_{1-\frac{\alpha}{2};n-1} = t^c_{0.975;9} = 2.262$ (see t-table)

6) $|t_0| = 3.3734 > 2.262 = t^c_{0.975;9}$

$\implies H_0$ is rejected, meaning the given data suggests that a tax increase does indeed have an effect on consumption.

b) Corresponding Python Code:

```
a = [27 ,31 ,23 ,35 ,26 ,27 ,26 ,18 ,22 ,21]
b = [40 ,36 ,43 ,34 ,25 ,41 ,32 ,29 ,21 ,36]

t_statistic, p_value = stats.ttest_rel(a, b)

print("t-statistic = ", t_statistic)
print("p-value = ", p_value)

d = [a[i] - b[i] for i in range(len(a))]

t_statistic, p_value = stats.ttest_1samp(d, 0)

print("t-statistic = ", t_statistic)
print("p-value = ", p_value)
```

$\Longrightarrow H_0$ is rejected.

Exercise T1.2 *Masks during Covid19*

In the context of the COVID-19 pandemic, 8 men and 10 women were asked how many hours per day they wear a mask. The following table shows their answers. The hypothesis is "On average, women wear their mask longer per day than men". It can be assumed, that the average time people wear their mask is normally distributed.

Individual no. (i)	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
Hours per day	4	2	3	5	7	2	7	3	5	2	2	1	5	3	1	3	2	3
Gender	f	f	f	f	f	f	f	f	f	f	m	m	m	m	m	m	m	m

a) Test the hypothesis "by hand" with a significance level of $\alpha = 0.05$ and 16 degrees of freedom.

b) Search for the corresponding functions in Python and use them to verify your result.

Solution

a) 1) i) Two samples, ii) independent

2) $H_1 : \mu_f > \mu_m$ on average, women wear their mask longer vs. $H_0 : \mu_f \leq \mu_m$ on average, women wear their mask shorter or equally long:

$$H_1 : \mu_D = \mu_f - \mu_m > \mu_0 = 0 \text{ and } H_0 : \mu_D = \mu_f - \mu_m \leq \mu_0 = 0$$

3) Here, we apply the Welch test. We have

$$t_0 = \frac{\bar{x}_f - \bar{x}_m - \mu_0}{s_{\bar{f}-\bar{m}}} \text{ with } s_{\bar{x}-\bar{w}} = \frac{s_f^2}{n_f} + \frac{s_m^2}{n_m},$$

where the formula is taken from the test manual's third step.

$$\bar{x}_f = \frac{4+2+3+5+7+2+7+3+5+2}{10} = 4,$$
$$\bar{x}_m = \frac{2+1+5+3+1+3+2+3}{8} = 2.5$$

Further,

$$\begin{aligned} s_f^2 &= \frac{(4-4)^2+(2-4)^2+(3-4)^2+(5-4)^2+(7-4)^2+(2-4)^2+(7-4)^2+(3-4)^2+(5-4)^2+(2-4)^2}{10-1} \\ &= \frac{0^2+(-2)^2+(-1)^2+1^2+3^2+(-2)^2+3^2+(-1)^2+1^2+(-2)^2}{9} \\ &= 3.778, \\ s_m^2 &= \frac{(-0.5)2+(-1.5)^2+2.5^2+0.5^2+(-1.5)^2+0.5^2+(-0.5)^2+0.5^2}{7} \\ &= 1.714 \end{aligned}$$

and

$$s_{\bar{f}-\bar{m}}^2 = \frac{3.778}{10} + \frac{1.714}{8} = 0.592 \quad \Rightarrow \quad s_{\bar{f}-\bar{m}} = 0.769$$

leading to

$$t_0 = \frac{1.5}{0.769} = 1.949$$

4) $\alpha = 0.05$

5) The degrees of freedom are $df = 16$. This is taken from the exercise, but can also be calculated via

$$df = \frac{\left(s_{\bar{f}-\bar{m}}^2\right)^2}{\frac{s_f^4}{n_f^2(n_f-1)} + \frac{s_m^4}{n_m^2(n_m-1)}}.$$

This results in $t_{0.95;16}^c = 1.746$ which is taken from the t-table.

6) $t_0 = 1.949 > 1.746 = t^c$

7) H_0 can be rejected. Regarding a significance level of $\alpha = 0.05$ it can be concluded that on average, women wear their mask longer per day than men.

b) The same result can be achieved by using Python as follows:

```
female = [4, 2, 3, 5, 7, 2, 7, 3, 5, 2]
male = [2, 1, 5, 3, 1, 3, 2, 3]

result = stats.ttest_ind(female, male, alternative="greater")

result

TtestResult(statistic=1.8650096164806276, pvalue=0.0403139296502071, df=16.0)
```

Exercise T1.3 *Research Methods*

You are a researcher investigating whether a new study technique improves the average test scores of students. You have collected data on the test scores of 15 students who used the new technique (group NT) and 15 students who did not (group OT). The following table contains the test scores, where i is the index of a student in a specific group. We can assume that both samples are normally distributed.

a) State the null hypothesis (H_0) and the alternative hypothesis (H_1) for this scenario.

b) Explain whether this is a one-sided or two-sided test and justify your choice.

i	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
NT_i	85	89	92	88	91	90	87	93	86	91	84	88	89	90	92
OT_i	79	81	75	82	77	80	78	84	76	80	78	83	82	79	85

c) Conduct the t-test in Python using the SciPy library to compare the means of the two groups using a significance level of $\alpha = 0.05$. You can leverage the provided notebook.

d) Given the test result, would you reject the null hypothesis H_0? Explain the result in the context of the research question.

e) Determine and interpret the corresponding 95% confidence interval in Python.

Solution

a) H_0: The new study technique does not improve the average test scores of students

$$H_0 : \mu_{NT} \leq \mu_{OT} \iff H_0 : \mu_{NT} - \mu_{OT} \leq 0$$

H_1: The new study technique improves the average test scores of students

$$H_0 : \mu_{NT} > \mu_{OT} \iff H_0 : \mu_{NT} - \mu_{OT} > 0$$

This is a *one-sided test* because we are only interested in whether the new technique has a positive effect on test scores.

b) Corresponding Python Code:

```
import numpy as np
from scipy import stats

# Data collection
scores_nt = np.array([85, 89, 92, 88, 91, 90, 87, 93, 86, 91, 84,
                      88, 89, 90, 92])
scores_ot = np.array([79, 81, 75, 82, 77, 80, 78, 84, 76, 80, 78,
                      83, 82, 79, 85])

# Conduct one-sided t-test
t_statistic, p_value = stats.ttest_ind(scores_nt, scores_ot,
                                       alternative='greater')

print(f'Test statistic: {t_statistic}, P-Value: {p_value}')
```

c) Based on the results of the one-sided t-test, the t-statistic is $t = 8.8798$, and the p-value is $p = 6.1914 \times 10^{-10}$.

We reject the null hypothesis because the p-value is less than the significance level ($\alpha = 0.05$). This indicates that the new study technique significantly improves the average test scores of students.

d) Determining the confidence interval:

```
n_nt = len(scores_nt)
n_ot = len(scores_ot)
mean_nt = np.mean(scores_nt)
mean_ot = np.mean(scores_ot)
std_dev_nt = np.std(scores_nt, ddof=1)  # We determine the sample standard
                                        # deviation, so ddof=1
std_dev_ot = np.std(scores_ot, ddof=1)  # We sum over N-ddof samples to
                                        # determine the standard deviation

# Pooled standard deviation
pooled_std = np.sqrt(((n_nt - 1) * std_dev_nt**2 + (n_ot - 1) * \
             std_dev_ot**2) / (n_nt + n_ot - 2))

# Calculate standard error of the difference in means (SE)
SE = pooled_std * np.sqrt(1/n_nt + 1/n_ot)

# Calculate t-critical value for 95% confidence level and 14 degrees of freedom
t_critical = stats.t.ppf(0.975, df=14)

# Calculate margin of error (MOE)
MOE = t_critical * SE

# Calculate the confidence interval
CI_lower = (mean_nt - mean_ot) - MOE
CI_upper = (mean_nt - mean_ot) + MOE

CI_lower, CI_upper
```

The 95% confidence interval for the difference in means is approximately (6.8767, 11.2566). This means we are 95% confident that the true difference in means falls within this interval and is greater than zero. Therefore, we can conclude that the new study technique significantly improves the average test scores of students.