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
Julius Durmann, Markus Ewert, Yutong Chao, Dr. Mete Ahunbay

Exercise T1.1 *Gas consumption*

According to the information supplied by the manufacturer of a certain type of car, its gas consumption in city traffic is approximately normally distributed with expected value $\mu = 9.5\,\ell/100km$. The standard deviation $\sigma = 2.5\,\ell/100km$ is commonly known (to the general public and the manufacturer). In order to review the manufacturers prediction, a consumer organization has performed a test on 25 cars which yielded the following result:

Average gas consumption: $\bar{x} = 10.5\,\ell/100km.$

Check the manufacturers statement with a suitable test for significance levels $\alpha = 0.05$ and $\alpha = 0.01$.

Solution

1) i) One sample, ii) σ_X known

2) The null hypothesis is $H_0 : \mu_x = \mu_0 = 9.5$ and states that information supplied by the manufacturer is correct, whereas the alternative hypothesis $H_1 : \mu_x \neq \mu_0 = 9.5$ states that the information supplied by the manufacturer is **not** correct.

3) Gauss test:

$$z_0 = \frac{\bar{x} - \mu_0}{\sigma_0}\sqrt{n} = \frac{10.5 - 9.5}{2.5}\sqrt{25} = 2$$

4) a) $\alpha = 0.05$
   b) $\alpha = 0.01$

5) Since it is two-sided test we use $\alpha/2$ to find the critical value:
   a) $1 - \frac{\alpha}{2} = 1 - 0.025 = 0.975, \quad z^c = z_{0.975} \approx 1.96$
   b) $1 - \frac{\alpha}{2} = 1 - 0.005 = 0.995, \quad z^c = z_{0.995} \approx 2.58$

6) Test decision:
   a) $z_0 = 2 > z^c = 1.96$. Thus, H_0 is rejected.
   b) $z_0 = 2 < z^c = 2.58$. Thus, H_0 is **not** rejected.

Alternative solution, using the p-value criterion instead of the test statistics criterion:

5) Calculating p value corresponding to the test statistic:

$$\frac{p}{2} = 1 - \phi(z_0) \approx 1 - 0.97725 = 0.02275 \approx 0.023$$

(Note: since it is a two sided test, what we get from the test statistic is $p/2$.)

6) We compare $p/2$ with $\alpha/2$:

   a) $p/2 \approx 0.023 < 0.025 = \alpha/2 \;\Rightarrow\; p < \alpha$. Thus, H_0 is rejected.

   b) $p/2 \approx 0.023 > 0.005 = \alpha/2 \;\Rightarrow\; p > \alpha$. Thus, H_0 is not rejected.

Exercise T1.2 *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) $|t_0| = 3.3734 > 2.262 = t^c_{0.975;9}$

$\Longrightarrow$ 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 T1.3 *Masks during Covid19*

In the context of the COVID-19 pandemic, 8 children and 10 adults were asked how many hours per day they wear a mask. The following table shows their answers. The hypothesis is "On average, adults wear their mask longer per day than children". 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
Adult	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	N	N	N	N	N	N	N	N

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) $H_1 : \mu_a > \mu_c$ on average, adults wear their mask longer vs. $H_0 : \mu_a \leq \mu_c$ on average, adults wear their mask shorter or equally long:

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

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

$$t_0 = \frac{\bar{x}_a - \bar{x}_c - \mu_0}{s_{\bar{a}-\bar{c}}} \text{ with } s_{\bar{x}-\bar{w}} = \sqrt{\frac{s_a^2}{n_a} + \frac{s_c^2}{n_c}},$$

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

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

Further,

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

and

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

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

$$df = \frac{\left(s_{\bar{a}-\bar{c}}^2\right)^2}{\frac{s_a^4}{n_a^2(n_a-1)} + \frac{s_c^4}{n_c^2(n_c-1)}}.$$

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, adults wear their mask longer per day than children.

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

```
adult = [4, 2, 3, 5, 7, 2, 7, 3, 5, 2]
child = [2, 1, 5, 3, 1, 3, 2, 3]

result = stats.ttest_ind(adult, child, alternative="greater", equal_var=False)

result

TtestResult(statistic=1.9494276330540574, pvalue=0.03470640093813483, df=15.637129
```