Master of Data Science Online Programme
Course: Basic Statistics
SGA #3: How to test a drug

Student: Andrei Batyrov (Fall 2022)
arbatyrov@edu.hse.ru

Higher School of Economics (HSE)
National Research University
Faculty of Computer Science

June 6, 2023

Contents

Listings

Problem 1

You perform a clinical trial to study new drug. You have 20 volunteers with some disease. You randomly split all the volunteers into two groups (10 volunteers in each): the treatment group and the control group. Volunteers in the treatment group receive the new drug, volunteers in the control group receive placebo (pills that looks like a drug but do not have active substance). You conclude that new drug is effective if people who take the drug will recover faster (on average) than people in the control group. If your drug is effective, you will invest in its production, otherwise you will look for another drug. Assume that you obtained the following data (disease duration in days).

Control group, days	Treatment group, days
6	7
7	6
7	6
5	5
7	5
8	6
8	7
7	5
7	5
7	8

Describe this problem in terms of statistical hypothesis testing framework.

1. How would you model your data in terms of random variables?
2. State the null hypothesis and the alternative.
3. Will your alternative be one-sided or two-sided? Why?
4. What kind of statistical test will you use? Why this test? Use this test (apply Python if necessary and provide your code), analyse the results and provide a conclusion in mathematical and real-life terms.
5. Would you invest into production of this drug?

Solution

Let X_C be the random variable "Disease duration in days in the control group", that is $x_C = (6, 7, 7, 5, 7, 8, 8, 7, 7, 7) \sim X_C$ with mean value $\bar{x}_C$. Let X_T be the random variable "Disease duration in days in the treatment group", that is $x_T = (7, 6, 6, 5, 5, 6, 7, 5, 5, 8) \sim X_T$ with mean value $\bar{x}_T$. Since both random variables describe some parameter of randomly chosen independent participants, due to Central Limit Theorem, it is reasonable to believe that they will have normal distribution. However, since the size of both samples is rather small – just 10 people both – it safer to consider a distribution with tails heavier than those of normal distribution, such as Student's t-distribution [3]. So, we will assume that $X_C, X_T \sim t$-distribution.

We are interested in evaluating the effect of the new drug, which is reduced disease duration, therefore our null hypothesis would be "No effect or negative effect", that is the disease duration has not changed or even increased, and the alternative hypotheses would be "Positive effect", that is the disease duration has decreased:

$\mathcal{H}_0 : \bar{x}_T \geq \bar{x}_C$, disease duration has not changed or increased,

$\mathcal{H}_1 : \bar{x}_T < \bar{x}_C$, disease duration has decreased,

We will also fix the significance level – probability of Type I error of reporting the desired effect that is not present in fact – $\alpha = 0.05$ [6]. First let's find the mean values for both groups to see if there is any difference between them:

$$\begin{aligned} \bar{x}_C &= \frac{6+7+7+5+7+8+8+7+7+7}{10} = 69/10 = 6.9, \\ \bar{x}_T &= \frac{7+6+6+5+5+6+7+5+5+8}{10} = 60/10 = 6.0. \end{aligned} \tag{1}$$

We can see that the disease duration has changed by $6.0 - 6.9 = -0.9$ days as compared to the control group, that is in treatment group the disease duration has decreased by 0.9 days. To evaluate if this result is significant, we need to find the probability of observing such or even smaller (larger absolute value) difference, provided that $\mathcal{H}_0$ holds, i.e. p-value [2]. Assuming that the volunteers for both control and treatment groups were chosen randomly, we will consider these random variables as independent. In this case we will need to use two-sample one-tailed t-test [4]. We cannot use paired t-test in our case: observations are not paired, since they come from different independent groups – the same volunteer cannot be in the control and treatment group at the same time. We use one-tailed t-test here because we are interested in catching the positive effect to determine if we should invest in the production, so zero or negative effect fall into the null hypothesis. Now, let's consider the t-score statistics of the t-distribution that describes the difference of the means of the samples x_T and x_C [5]:

$$t(x_T, x_C) = \frac{\bar{x}_T - \bar{x}_C}{\sqrt{\frac{Var_+(x_T)}{n} + \frac{Var_+(x_C)}{m}}}, \tag{2}$$

where Var_+ is the unbiased variance of a sample, n is the treatment sample size, and m is the control sample size. First, let's find the variances:

$$\begin{aligned} Var_+(x_T) &= \frac{(7-6.0)^2 + (6-6.0)^2 + (6-6.0)^2 + (5-6.0)^2 + (5-6.0)^2 + \\ &+ (6-6.0)^2 + (7-6.0)^2 + (5-6.0)^2 + (5-6.0)^2 + (8-6.0)^2}{10-1} = \frac{10}{9} = 1.(1), \\ Var_+(x_C) &= \frac{(6-6.9)^2 + (7-6.9)^2 + (7-6.9)^2 + (5-6.9)^2 + (7-6.9)^2 + \\ &+ (8-6.9)^2 + (8-6.9)^2 + (7-6.9)^2 + (7-6.9)^2 + (7-6.9)^2}{10-1} = \frac{23}{30} = 0.7(6), \end{aligned} \tag{3}$$

Now, the t-score is

$$t(x_T, x_C) = \frac{6.0 - 6.9}{\sqrt{\frac{10/9}{10} + \frac{23/30}{10}}} = -\frac{9}{10\sqrt{\frac{169}{900}}} = -\frac{27}{13} \approx -2.077. \tag{4}$$

We are interested in the left tail of the distribution ($\bar{x}_T < \bar{x}_C \Rightarrow \bar{x}_T - \bar{x}_C < 0$). The p-value at $t(x_T, x_C)$ for the left tail is $P(T \leq t(x_T, x_C))$. Recalling that as per [1] $P(T \leq t(x_T, x_C))$ is the value of cumulative distribution function at $t(x_T, x_C)$,

$$p\text{-value}(t(x_T, x_C)) = P(T \leq t(x_T, x_C) | \mathcal{H}_0) = CDF(t(x_T, x_C)), \tag{5}$$

where CDF is the cumulative distribution function of the t-distribution. CDF for t-distribution cannot be expressed in elementary functions and is usually computed with numerical methods. Let's use the Python's library Scipy to find the value of $CDF(t(x_T, x_C))$ for t-distribution with degrees of freedom [3] for two samples $\nu = n + m - 2 = 10 + 10 - 2 = 18$:

$$p\text{-value}(t(x_T, x_C) = -2.077) = CDF(-2.077) \approx 0.026 = 2.6 \times 10^{-2}. \tag{6}$$

Now, comparing the obtained result (6) with the significance level $\alpha = 5 \times 10^{-2}$, it is easily seen, that p-value $\approx 2.6 \times 10^{-2} < \alpha = 5 \times 10^{-2}$, which means that we are well within the critical region of our t-distribution and we control the probability of Type I error. In this case, we are prone to reject $\mathcal{H}_0$ in favour of $\mathcal{H}_1$. The difference in the disease duration in the control and treatment groups is statistically significant. We may claim that people who received the new drug in the treatment group recovered faster by 0.9 days on average than those who did not receive the new drug in the control group, and this faster recovery was not a fluke.

Let's also confirm the calculations by running a two-sample one-tailed t-test using the Python's library Scipy – see Listing 1. The result of the script is shown in Listing 2. The `equal_var` parameter of the function `test_ind` is determined by equality or non-equality of the samples' variances.

Further Considerations

We have just shown that the 0.9-day faster recovery is a statistically significant result by passing the two-sample one-sided t-test for $\alpha = 5 \times 10^{-2}$. However, we should not be overoptimistic. It is obvious, that in our scenario we should carefully fix and and control the probability of Type I error – we should minimise the probability of reporting a positive effect from the new drug, while such

effect is absent in fact ($\mathcal{H}_0$ is true). We can notice that by fixing $\alpha = 10^{-2}$, we will fail our t-test having p-value of 2.6×10^{-2}, which may mean that 0.9-day reduction is not a sufficient difference for decision-making.

Another issue is that intuitively it seems that having 10 participants in each group is not enough to make a decision about investing or not into production of a drug. We need more data. At the same time, more data will have more diversity – the sample variance may increase. And we may well fail the t-test in this case. Indeed, it is easy to check that by introducing 2 more participants in the treatment group with values 2 and 10 days, i.e. new $x_T = (7, 6, 6, 5, 5, 6, 7, 5, 5, 8, \mathbf{2}, \mathbf{10})$, the variance will increase from 1.(1) to 3.(81), while the mean stays the same 6.0. However, now the difference (-0.9) days is no longer statistically significant. See Listing 3 for the result of the Python script with updated x_T. It may seem that having 2 and 10 days is unusual. However, it is quite possible: a person might have a genetically stronger immune system and thus recovers much faster than an average person or vice versa. Also, it might be an error or typo in medical records.

In conclusion, though as per the data provided in the problem we are prone to invest into production of the new drug, in real life before making an investment decision, we need to collect more data and probably set more strict requirements for the significance level. Also, though not considered in this problem, but in real life we may want to set the level of and control the probability of Type II error [6] – obtaining a false negative result – when we do *not* report the positive (in our case) effect of the new drug, but this effect actually exists. This means that we will lose an investment opportunity – we are too pessimistic and cautious. So, in any case, we need to control both Type I and Type II errors in real life scenarios. This is about health safety after all!

Answer

1. Since both random variables describe some parameter of randomly chosen independent participants, due to Central Limit Theorem, it is reasonable to believe that they will have normal distribution. However, since the size of both samples is rather small – just 10 people both – it safer to consider a distribution with tails heavier than those of normal distribution, such as Student's t-distribution. So, we will assume that $X_C, X_T \sim t$-distribution.

2. The null hypothesis is "No effect or negative effect", that is the disease duration has not changed or even increased. The alternative hypotheses is "Positive effect", that is the disease duration has decreased.

3. The alternative hypothesis is one-sided because we are interested in catching the positive effect to determine if we should invest in the production, so zero or negative effect fall into the null hypothesis.

4. Assuming that the volunteers for both control and treatment groups were chosen randomly, we will consider these random variables as independent. In this case we will need to use two-sample one-tailed t-test. We cannot use paired t-test in our case: observations are not paired, since they come from different independent groups – the same volunteer cannot be in the control and treatment group at the same time. We use one-tailed t-test here because we have one-sided alternative hypothesis.

5. We have passed the t-test with p-value $= 2.6 \times 10^{-2} < \alpha = 5 \times 10^{-2}$ and rejected the null hypothesis in favour of the alternative hypothesis – the disease duration has decreased by 0.9 days in patients taking the new drug, and this reduction is statistically significant. Based on the given data, we may want to invest into production of this drug. However, in real life scenario, we need to gather more data and decrease the significance level to reduce the health safety and investment risks.

```python
from scipy.stats import t, ttest_ind
import numpy as np
'''
1) Set up statistical hypothesis testing framework
H_0: x_treatment_bar >= x_control_bar, no or negative effect
H_1: x_treatment_bar < x_control_bar, positive effect
'''
H_1 = 'less'
x_control = np.array([6, 7, 7, 5, 7, 8, 8, 7, 7, 7])
x_treatment = np.array([7, 6, 6, 5, 5, 6, 7, 5, 5, 8])
#x_treatment = np.array([7, 6, 6, 5, 5, 6, 7, 5, 5, 8, 2, 10])
sample_size = len(x_control)
alpha = 0.05 # Type I error probability
print('Input data:')
print(f'x_control = {list(x_control)}')
print(f'x_treatment = {list(x_treatment)}')
print(f'x_control_bar = {x_control.mean():.2f}')
print(f'x_treatment_bar = {x_treatment.mean():.2f}')
print(f'x_treatment_bar - x_control_bar = {(x_treatment.mean() - x_control.mean()):.2f}')
print(f'x_control_var = {x_control.var(ddof=1):.3f}')
print(f'x_treatment_var = {x_treatment.var(ddof=1):.3f}')
print()
'''
2) Two-sample one-tailed t-test for treatment and control groups (independent)
'''
equal_var = x_control.var(ddof=1) == x_treatment.var(ddof=1)
if equal_var:
    print('Run two-sample one-tailed t-test for equal variances:')
else:
    print('Run two-sample one-tailed t-test for non-equal variances:')
t_stat, p_val = ttest_ind(x_treatment, x_control, equal_var=equal_var, alternative=H_1)
t_crit = t.ppf(alpha, df=2 * sample_size - 2)
print(f't-score = {t_stat:.3f}')
print(f't-crit = {t_crit:.3f}')
print(f'p-value = {p_val:.3f}')
print(f'alpha = {alpha}')
print('Reject H_0') if p_val < alpha else print('Do not reject H_0')
```

Listing 1: t-test Python script

```
Input data:
x_control = [6, 7, 7, 5, 7, 8, 8, 7, 7, 7]
x_treatment = [7, 6, 6, 5, 5, 6, 7, 5, 5, 8]
x_control_bar = 6.90
x_treatment_bar = 6.00
x_treatment_bar - x_control_bar = -0.90
x_control_var = 0.767
x_treatment_var = 1.111

Run two-sample one-tailed t-test for non-equal variances:
t-score = -2.077
t-crit = -1.734
p-value = 0.026
alpha = 0.05
Reject H_0
```

Listing 2: Result of running t-test Python script

```
Input data:
x_control = [6, 7, 7, 5, 7, 8, 8, 7, 7, 7]
x_treatment = [7, 6, 6, 5, 5, 6, 7, 5, 5, 8, 2, 10]
x_control_bar = 6.90
x_treatment_bar = 6.00
x_treatment_bar - x_control_bar = -0.90
x_control_var = 0.767
x_treatment_var = 3.818

Run two-sample one-tailed t-test for non-equal variances:
t-score = -1.432
t-crit = -1.734
p-value = 0.086
alpha = 0.05
Do not reject H_0
```

Listing 3: Result of running t-test with updated x_T

References

[1] Ilya Schurov. *Cumulative distribution function (CDF)*. Faculty of Computer Science, Higher School of Economics. URL: https://smartedu.hse.ru/mod/page/0/756817.

[2] Ilya Schurov. *Introducing p-value*. Faculty of Computer Science, Higher School of Economics. URL: https://smartedu.hse.ru/mod/page/0/798133.

[3] Ilya Schurov. *T-distribution*. Faculty of Computer Science, Higher School of Economics. URL: https://smartedu.hse.ru/mod/page/0/798145.

[4] Ilya Schurov. *Two-sample t-test, part 1*. Faculty of Computer Science, Higher School of Economics. URL: https://smartedu.hse.ru/mod/page/0/798161.

[5] Ilya Schurov. *Two-sample t-test, part 2*. Faculty of Computer Science, Higher School of Economics. URL: https://smartedu.hse.ru/mod/page/0/798162.

[6] Ilya Schurov. *Type I and type II errors*. Faculty of Computer Science, Higher School of Economics. URL: https://smartedu.hse.ru/mod/page/0/798130.