

# Laboratory 03

## Table of contents

<b>1</b>	<b>Reply on the feedback of Lab 2</b>	<b>1</b>
1.1	<i>Normally</i> on Line 4, Section 3.2 Question 2, Page 5 . . . . .	1
1.2	About the cats . . . . .	2
<b>2</b>	<b>Main scenario story (Context)</b>	<b>2</b>
<b>3</b>	<b>Main scenario quests (Objectives)</b>	<b>2</b>
<b>4</b>	<b>Solutions</b>	<b>3</b>
4.1	A peek on the dataset . . . . .	3
4.2	Q1: Distribution of the pain scores . . . . .	4
4.2.1	Q1a: Describe the distribution . . . . .	4
4.2.2	Q1b: One-sample $t$ test on this variable . . . . .	6
4.3	Q2: Mean of pain score tested . . . . .	7
4.4	Q3: 95% confidence interval for the mean pain score . . . . .	9
4.5	Q4: Summarizing the findings . . . . .	10
4.6	Q5: 95% CI for Switch players' mean pain score . . . . .	10

## 1 Reply on the feedback of Lab 2

Thanks for the feedback on my Lab 2 Assignment! Here's my response:

### 1.1 *Normally* on Line 4, Section 3.2 Question 2, Page 5

“normally” have higher median?

What I mean is that based on the sample and the box-plot, we can generally infer that “cat people tend to have a higher life satisfaction score” than fish owners.

Apologies for the earlier miswording and ambiguity in the interpretation.

## 1.2 About the cats

They usually come out in the early morning and late at night, as cats *normally* do. However, they are all around at various times. I even once witnessed the police and firefighters at our faculty, rescuing a young cat trapped on the roof!

## 2 Main scenario story (Context)



Figure 1: Warning - Health and Safety

There are reports of increases in injuries related to playing games consoles. These injuries were attributed mainly to muscle and tendon strains. A researcher collected data from 120 participants who played on a Nintendo Switch or watched others playing. The outcome was a pain score from 0 to 10, where 0 is no pain and 10 is severe pain. The data are in `switch.sav`.

## 3 Main scenario quests (Objectives)

1. Describe the distribution of the pain scores. Do you think the one-sample  $t$  test is suitable for this variable?
2. A pain score of 2 is considered as minor pain. Test whether the mean pain score is equal to 2 (two tailed, 5% level) and obtain the corresponding effect size.
3. Obtain a 95% confidence interval for the mean pain score.
4. Summary your findings from the previous questions in several sentences.

5. (Extra credit) Obtain a 95% confidence interval for the mean pain score of those who played on a Nintendo Switch. That is, exclude those who only watched others playing. (Hint: You learned how to exclude cases in Laboratory Assignment 1.)

## 4 Solutions

### 4.1 A peek on the dataset

As usual, I load modules that I may need in this laboratory assignment, then the dataset to my RAM and check attributes of the given dataset.

```
1 import pandas as pd
2
3 # Load the dataset
4 switch = pd.read_spss('./datasets/switch.sav')
```

```
1 # Descriptions
2 print(f'Shape: \n', switch.shape, '\n')
3 print(f'Columns: \n', switch.columns, '\n')
4 print(f'First 5 rows: \n', switch.head(5), '\n')
5 print(f'Describe the column `injury`: \n', switch.describe(), '\n')
```

Shape:  
(120, 5)

Columns:  
Index(['id', 'athlete', 'stretch', 'switch', 'injury'], dtype='object')

First 5 rows:

	id	athlete	stretch	switch	injury
0	yty	Athlete	Stretching	Playing switch	2.0
1	wel	Athlete	Stretching	Playing switch	2.0
2	qfs	Athlete	Stretching	Playing switch	1.0
3	oln	Athlete	Stretching	Playing switch	2.0
4	wxi	Athlete	Stretching	Playing switch	0.0

Describe the column `injury`:

	injury
count	120.000000
mean	2.891667
std	1.994934

min	0.000000
25%	2.000000
50%	2.000000
75%	4.000000
max	10.000000

## 4.2 Q1: Distribution of the pain scores

### 4.2.1 Q1a: Describe the distribution

#### Answer

To describe the distribution of the pain scores, I use histogram with a kernel density estimation curve as shown in Figure 2 as well as measurements (mean, mode and median) reflect central tendency (see Table 1).

Table 1: Mean, mode and median

Measurement	Value
Mode	2.00
Mean	2.89
Median	2.00

According to the graph:

1. Most of the observations are clustered around the lower pain scores (between 1 and 4), we can say that the distribution of pain scores is positively skewed rather than a perfect normal distribution.
2. There is a noticeable peak at a score of 2, which means the most frequent score is around 2.
3. A long tail extends to the higher scores, indicating the frequency of pain scores gradually decreases as the scores increase.

#### Solution

```

1 injury = switch['injury']
2
3 # Calculate measurements of central tendency
4 injury_mean = injury.mean()
5 injury_mode = injury.mode()[0]
6 injury_median = injury.median()
7
8 # Tell the result
9 print(f'Central Tendency: \n')
```

```
10 print(f'Mean: ', injury_mean)
11 print(f'Mode: ', injury_mode)
12 print(f'Median: ', injury_median)
```

Central Tendency:

Mean: 2.8916666666666666

Mode: 2.0

Median: 2.0

```
1 import matplotlib.pyplot as plt
2 import seaborn as sns
3
4 # Plot the histogram
5 injury_hist = sns.histplot(switch, x='injury', stat='count', bins=10
    ↪ ,kde=True)
6 # Dashed line for Mean, Median and Mode
7 injury_hist.axvline(injury_mean, color='blue', linestyle='--', linewidth=1)
8 injury_hist.axvline(injury_median, color='red', linestyle='--',
    ↪ linewidth=1)
9 # Set title and labels
10 injury_hist.set_title('Distribution of the pain scores')
11 injury_hist.set_xlabel('Pain score (out of 10)')
12 # Show the plot
13 plt.show()
```

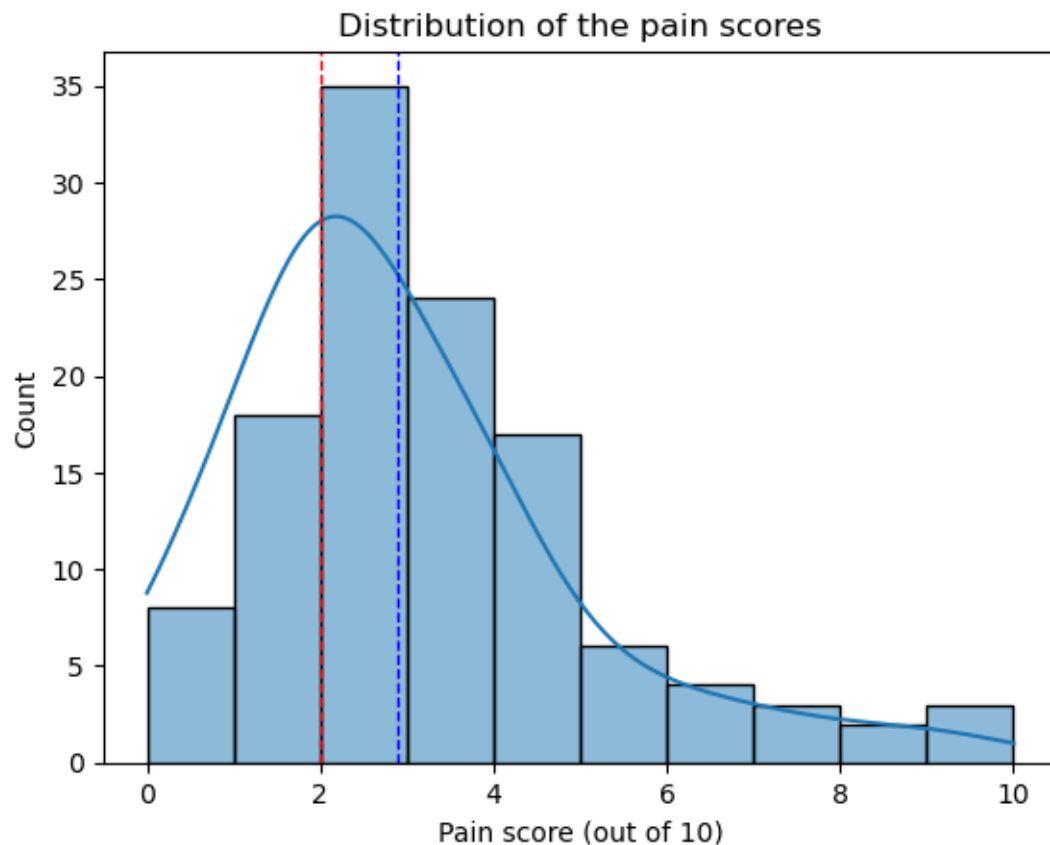


Figure 2: Distribution of the pain scores

#### 4.2.2 Q1b: One-sample $t$ test on this variable

##### Answer

Recall back the the slides in the lecture notes:

One-sample  $t$  test requires that:

- Sample mean describes central tendency.

The sample mean is slightly higher than the mode and median (see Table 1 and Figure 2, red dashed lines for the median and mode, blue for the mean) since the data is right-skewed. However, they are fairly close to each other, so the sample mean can still represent central tendency.

- Scores in the sample are randomly selected from the population

According to the description, the data was “collected from 120 participants who played on a Nintendo Switch or watched others playing.” For the sake of this assignment, I will assume that the participants were randomly selected from patients worldwide to fulfill the random sampling assumption.

- Either  $N$  is large or  $X$  follows a normal distribution

Given the right-skewed distribution as seen in Figure 2, the data may violate the assumption of normality required for the one-sample  $t$ -test. However, the *Central Limit Theorem* suggests that if the sample size is large (typically  $N > 30$ ), the sampling distribution of the sample mean tends to approach normality. Therefore, despite the skewed distribution, the sample size ( $N = 120$ ) makes the one-sample  $t$ -test acceptable in this case.

Additionally, question 2 specifically asks for a one-sample  $t$ -test without requiring further preprocessing of data (e.g., a log transformation), which further supports the applicability of the one-sample  $t$ -test to this data. If the data were unusable, there would be no reason to include the following questions.

In conclusion, a one-sample  $t$ -test is suitable for this dataset.

#### 4.3 Q2: Mean of pain score tested

##### Answer

1.  $p$  value

$$p \approx 3.11 \times 10^{-6}$$

At a 5% significance level ( $\alpha = 0.05$ ),  $p < 0.001$ , we reject the null hypothesis. The mean pain score is significantly different from 2 (a minor pain) at the 5% level.

2. The Cohen's  $d$  value

$$d \approx 0.45$$

The Cohen's  $d$  value indicates a medium effect. This suggests that the difference between the mean pain score ( $M = 2.89$ ) and the a minor pain ( $\mu_{hyp} = 2$ ) is meaningful in practical terms.

##### Solution

Given  $N = 120$ ,  $M \approx 2.89$ ,  $SD \approx 1.99$ ,  $\mu_{hyp} = 2$ , the standard error  $SE_M$  is:

$$SE_M = \frac{SD}{\sqrt{N}} \approx 0.18$$

```

1 from math import sqrt
2
3 # Standard Error Mean
4 # Note: I can use injury.sem() directly to get the result,
5 # but I shall calculate by my own for this assignment.
6
7 injury_sem = injury.std(ddof=1) / sqrt(120)
8 print(f'Standard Error Mean: ', injury_sem)

```

Standard Error Mean: 0.1821117309227567

With  $SE_M \approx 0.18$ , the  $t$  ratio is:

$$t = \frac{M - \mu_{hyp}}{SE_M} \approx 4.90$$

```

1 # t statistic
2 injury_t = (injury.mean() - 2) / injury.sem()
3 print(f't: ', injury_t)

```

t: 4.8962615540943375

Unfortunately, I can't calculate the  $p$ -value on hand, so in this part I'll call `scipy.stats.t` for help. the degree of freedom ( $df$ ) is:

$$df = N - 1 = 120 - 1 = 119$$

With  $t \approx 4.90$  and  $df = 119$ , then use survivor function to reach the  $p$ -value:

$$p \approx 3.11 \times 10^{-6}$$

```

1 import scipy.stats as stats
2
3 # 119 is the degree of freedom; Two-sided times two
4 injury_p = stats.t.sf(injury_t, 119) * 2
5
6 print(f'p: ', injury_p)

```

p: 3.1051091723547962e-06



The  $p$ -value is much smaller than 0.001 ( $p < 0.001$ ), the null hypothesis should be rejected.

I also did a sanity check with the ready-to-use function `scipy.stats.ttest_1samp`:

```
1 # A san-check on my calculation result:
2
3 injury_ttest_1samp = stats.ttest_1samp(injury, 2, alternative='two-sided')
4
5 print(f't: ', injury_ttest_1samp.statistic, '\n'
6       'df: ', injury_ttest_1samp.df, '\n'
7       'p-value: ', injury_ttest_1samp.pvalue)
```

```
t: 4.8962615540943375
df: 119
p-value: 3.1051091723547962e-06
```

The Cohen's  $d$  value is:

$$d = \frac{M - \mu_{hyp}}{SD} = \frac{t}{\sqrt{N}} \approx 0.45$$

```
1 # Effect Size d
2
3 injury_d = injury_t / sqrt(120)
4 print(f'Cohen\'s d:', injury_d)
```

```
Cohen's d: 0.4469654834371283
```

#### 4.4 Q3: 95% confidence interval for the mean pain score

##### Answer

Based on the sample of  $N = 120$  pain scores, with  $M \approx 2.89$  and  $SD \approx 1.99$ , the 95% CI for pain scores is  $[2.53, 3.25]$ .

##### Solution

Given  $c = 1.96$  for a 95% confidence interval and  $SE_M \approx 0.18$  as calculated in the last section, a 95% confidence interval of the mean pain score is:

$$[M - c \times SE_M, M + c \times SE_M] \approx [2.53, 3.25]$$

```

1 # CI for two-tailed t-statistics
2 def confidence_interval(alpha, mean, sem, df):
3     c = stats.t.interval(1 - alpha, df)[1]
4     ci_upper = mean + (c * sem)
5     ci_lower = mean - (c * sem)
6     print(f'CI (Lower): ', ci_lower)
7     print(f'CI (Upper): ', ci_upper)
8     return str(f'[{ci_lower}, {ci_upper}]')
9
10 injury_ci = confidence_interval(0.05, injury_mean, injury_sem, 119)
11 print(injury_ci)

```

```

CI (Lower):  2.53106725077079
CI (Upper):  3.252266082562543
[2.53106725077079, 3.252266082562543]

```

```

1 # And scipy.stats.t does the same thing.
2 stats.t.interval(confidence=0.95,
3                 df=119,
4                 loc=injury_mean,
5                 scale=injury_sem)

```

```

(2.53106725077079, 3.252266082562543)

```

## 4.5 Q4: Summarizing the findings

### Answer

A one-sample  $t$ -test is conducted to reveal whether mean pain score for a sample of  $N = 120$  patients differed from the minor pain with a score of 2. For this example,  $M = 2.89$ ,  $SD = 1.99$  and  $SE_M = 0.18$ . The 95% CI for  $M$  was  $[2.53, 3.25]$ . The result was  $t(119) = 4.90$ ,  $p < 0.001$ , two tailed. The effect size is  $d = 0.45$  by Cohen's standards, which represents a medium effect. The difference between the sample mean ( $M = 2.89$ ) and the score of minor pain (2) is statistically significant using  $\alpha = 0.05$ , two tailed.

## 4.6 Q5: 95% CI for Switch players' mean pain score

### Answer

The 95% confidence interval for the mean pain score of those who played on a Nintendo Switch is  $[3.14, 4.33]$ .

## Solution

1. Check the structure of column switch then apply the filtering:

```
1 # Check how many people get injured while playing
2 print(f'Variables in the column switch: \n',
   ↪ switch['switch'].value_counts())
3 # Filter out all switch players
4 injury_ns = switch[switch['switch'] == 'Playing switch']['injury']
5 # And a sanity check
6 print(f'Filtered data: \n', injury_ns.describe())
```

Variables in the column switch:

```
switch
Playing switch      60
Watching switch     60
Name: count, dtype: int64
```

Filtered data:

```
count      60.000000
mean        3.733333
std         2.313312
min         0.000000
25%         2.000000
50%         3.500000
75%         5.000000
max         10.000000
Name: injury, dtype: float64
```

2. Calculating the CI:

Given  $N_{player} = 60$ , then the  $df_{player} = N_{player} - 1 = 59$ ,

Based on the data we also have  $M_{player} \approx 3.73$  and  $SE_{M_{player}} \approx 0.30$

The 95% confidence interval for the mean pain score of those who played on a Nintendo Switch is:

$$[M - c \times SE_M, M + c \times SE_M] \approx [3.14, 4.33]$$

```
1 injury_ns_mean = injury_ns.mean()
2 injury_ns_sem = injury_ns.sem()
3 injury_ns_dregf = len(injury_ns) - 1
4
5 print(f'Sample size: {len(injury_ns)}, \n'
6       f'Degree of Freedom: {injury_ns_dregf}, \n'
7       f'Mean: {injury_ns_mean}, \n')
```

```

8         f'Standard Error: {injury_ns_sem}')
```

```

9
```

```

10 injury_ns_ci = confidence_interval(0.05, injury_ns_mean, injury_ns_sem,
    ↪ injury_ns_dregf)
```

```

11 print(f'\nThe 95% CI for Switch players: \n', injury_ns_ci)
```

Sample size: 60,  
 Degree of Freedom: 59,  
 Mean: 3.7333333333333334,  
 Standard Error: 0.29864729557842784  
 CI (Lower): 3.1357414752023387  
 CI (Upper): 4.3309251914643285

The 95% CI for Switch players:  
 [3.1357414752023387, 4.3309251914643285]