

# Laboratory 04

# Table of contents

1	Context	1
2	Objectives	1
3	Solutions	1
3.1	A quick dataset check-up	1
3.2	Q1: Data distribution	2
3.3	Q2: The (s)catterplot	3
3.4	Q3: The Pearson's $r$ of <code>dinner_time</code> and <code>meow</code>	4
3.5	Q4: Test $H_0$ at the 5% level	5
3.6	Q5: 95% CI of $\rho$	5
4	Final thoughts: Cats good, humans bad.	6

## 1 Context

( , - . \_ , ' ( | \ - / |  
` - . - ' \ ) - ` ( , o o )  
` - \ = " ! -

Open the data in the file `catterplot.sav`. These data measure two variables:

- `dinner_time`: the time since last feeding a cat.
- `meow`: how loud their purr is.

## 2 Objectives

1. Describe the distribution of `dinner_time`.
2. Draw a scatterplot of `meow` against `dinner_time` with the LOESS curve. Do you think the two variables are linearly related?
3. Compute the correlation coefficient of `dinner_time` and `meow`. Does the correlation coefficient suggest a strong linear relationship?
4. Test the null hypothesis that the two variables are not linearly related at the 5% level.
5. (Extra credit) Obtain a 95% confidence interval for the population correlation coefficient of the two variables.

### 3 Solutions

### 3.1 A quick dataset check-up

```
1 import pandas as pd
2 catterplot_df = pd.read_spss('./datasets/catterplot.sav')
3
4 print(f'Description on the dataset: \n', catterplot_df.describe())
```

Description on the dataset:

	dinner_time	meow
count	78.000000	78.000000
mean	9.865385	8.217949
std	6.068159	3.747371
min	1.000000	2.000000
25%	5.000000	5.000000
50%	9.000000	8.000000
75%	14.000000	12.000000
max	24.000000	15.000000

### 3.2 Q1: Data distribution

Answer

Table 1: Mean, mode and median

Measurement	Value
Mean	9.87
Median	9.00
Mode	5.00

Table 1 lists measurements on the central tendency of this data, the mean (9.87) is slightly higher than the median (9.00) but much higher than the mode (5.00). The distribution is right-skewed as shown by Figure 1, and a longer tail is found on the right side. There is a clear peak around 5, which indicates most cats in this sample were fed with in 5 hours.

**Solution**

To describe the distribution of the dataset, I will again use the mean, median, and mode to assess central tendency, along with a histogram to visualize the distribution.

```
1 def centrality(df):
2     mean = df.mean()
3     median = df.median()
4     mode = df.mode()[0]
5     return mean, median, mode
6
7 cats_mean, cats_median, cats_mode = centrality(catterplot_df['dinner_time'])
8 print(f'Central Tendency: \n'
9       f'Mean: {cats_mean}, \n'
10      f'Median: {cats_median} \n'
11      f'Mode: {cats_mode}')
```

Central Tendency:

Mean: 9.865384615384615,

Median: 9.0

Mode: 5.0

```
1 import matplotlib.pyplot as plt
2 import seaborn as sns
3
4 # Create a histogram to reveal the distribution.
5 dinner_time_hist = sns.histplot(catterplot_df, x='dinner_time', kde=True, bins=23)
6 # Set titles and labels for x- and y- axis
7 dinner_time_hist.set_title('Distribution of the time since last feeding a cat')
8 dinner_time_hist.set_xlabel('last_feed: Time elapsed since the last feed')
9 dinner_time_hist.set_ylabel('Frequency')
```

```

10 # Display the plot
11 plt.show()

```

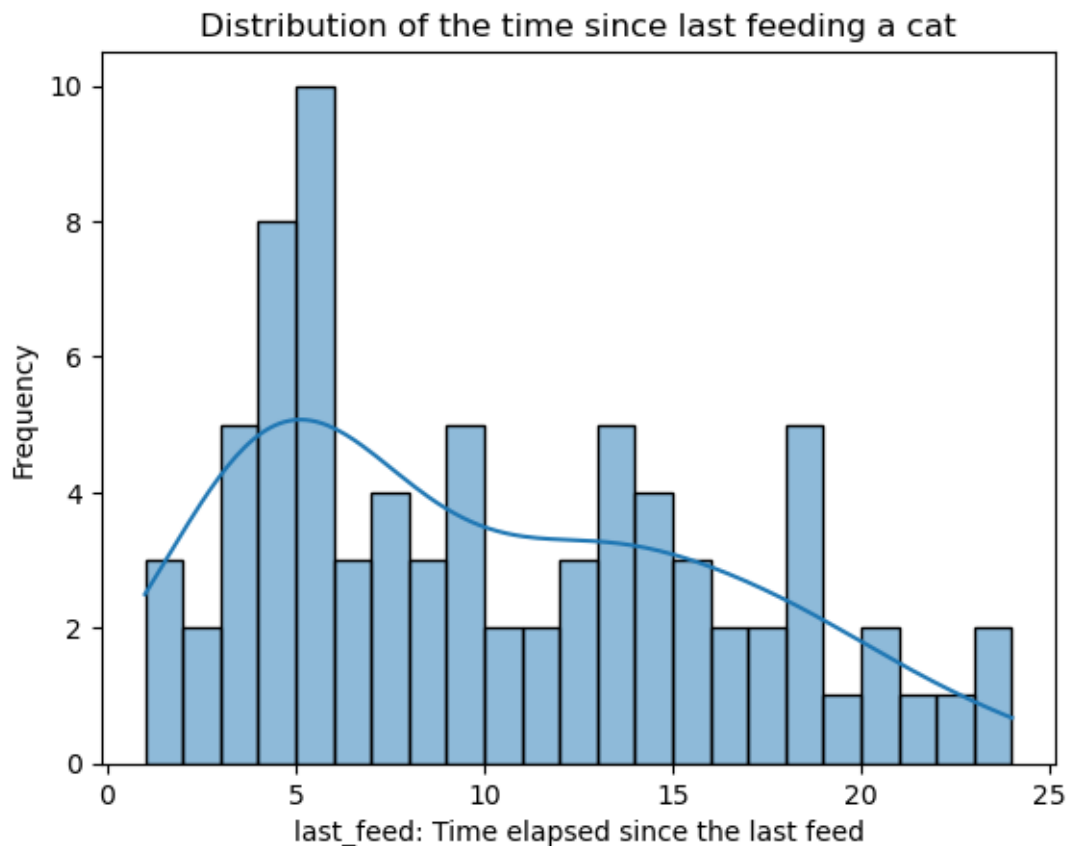


Figure 1: Distribution of Dinner Time

### 3.3 Q2: The (s)catterplot

#### Answer

The scatter plot in Figure 2 shows that the data, at least, forms a perfect shape of a cat! However, the scattered points and the gentle slope of the LOESS curve suggest that there is not a strong linear relationship between the time since the last feeding (`dinner_time`) and the loudness of purring (`meow`). The scatter plot also indicates a weak correlation between these two variables.

#### Solution

```

1 import matplotlib.pyplot as plt
2 import seaborn as sns
3
4 catterplot = sns.regplot(catterplot_df, x='dinner_time', y='meow', fit_reg=True,
5   ↪ lowess=True)
6 catterplot.set_title('The (S)catterplot')
7 catterplot.set_xlabel('last_feed: Time elapsed since the last feed')
8 catterplot.set_ylabel('meow: Loudness of Purring')
9 plt.show()

```

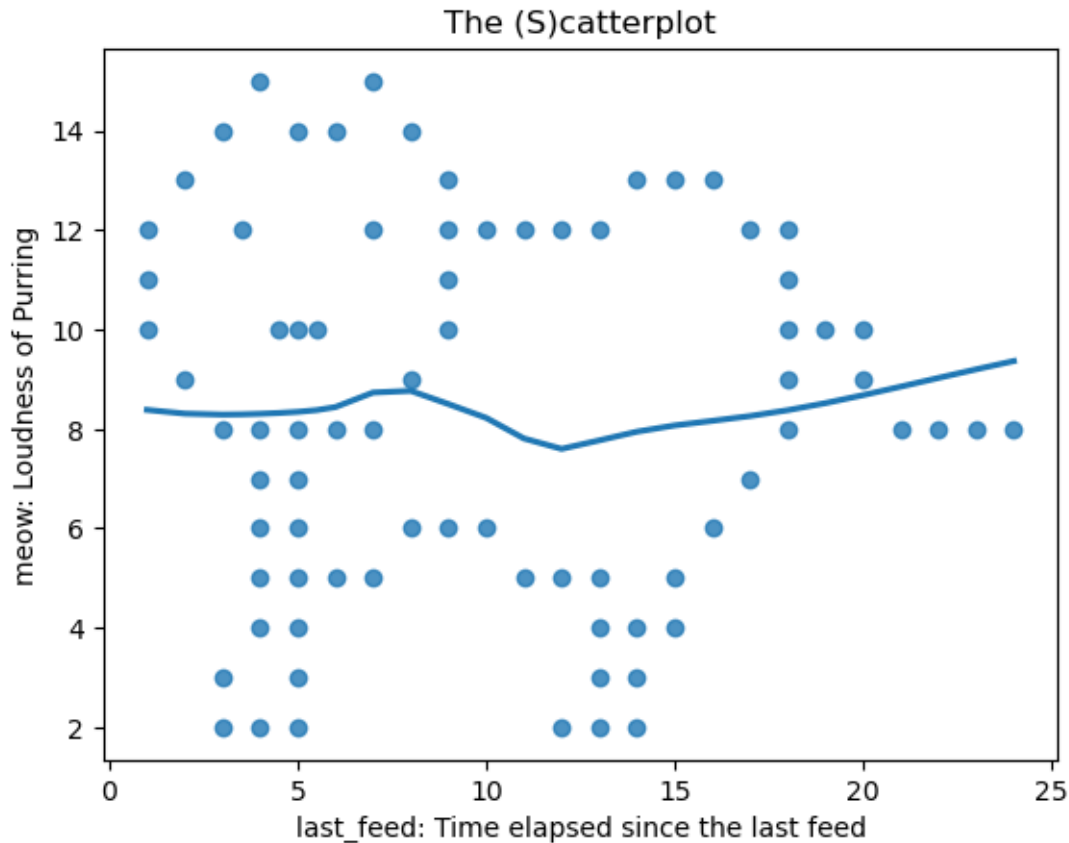


Figure 2: The (s)catterplot on meow against dinnertime.

### 3.4 Q3: The Pearson's $r$ of `dinner_time` and `meow`

**Answer**

$$r \approx -6.12 \times 10^{-3}$$

The pearson's correlation coefficient  $r \approx -6.12 \times 10^{-3}$  is very close to zero, suggesting there may be no linear relationship between `dinner_time` and `meow` in this dataset. However, it's important to note that the data lacks further details. Specifically, we don't know whether it was collected from a single cat fed multiple times (78 times in this case) or from 78 different cats. Although the correlation is statistically weak, we cannot definitively conclude that there is no relationship between the given variables.

In conclusion, the pearson's  $r$  suggests there is no strong linear relationship between the two variables.

**Solution**

To calculate the pearson's  $r$ , I use the formula<sup>1</sup>:

$$r = \frac{\sum Z_X Z_Y}{N - 1}$$

```

1 import scipy.stats as stats
2
3 pearson_r_cats, p_cats = stats.pearsonr(catterplot_df.meow, catterplot_df.dinner_time)
4
5 print(f'Correlation Coefficient\n'
6       f'Pearson-r: {pearson_r_cats}\n'
7       f'p-value: {p_cats}')
```

<sup>1</sup>I just put the formula here for ... fun? I didn't actually calculated on hand this time. :P

Correlation Coefficient  
Pearson-r: -0.006117549224989473  
p-value: 0.9576068830096687

### 3.5 Q4: Test $H_0$ at the 5% level

#### Answer

$p \approx 0.9576$

The  $p$ -value is much greater than the significance level  $\alpha = 0.05$ , therefore, we fail to reject the null hypothesis ( $H_0$ ) that `dinner_time` and `meow` are not linearly related. This means that there is insufficient evidence to suggest a linear relationship between `dinner_time` and `meow` at the 5% significance level, the sample does not indicate a significant linear correlation between the two variables.

#### Solution

The null hypothesis ( $H_0$ ) and the alternative hypothesis ( $H_1$ ) are formulated as follows:

- $H_0 : \rho = 0$ : no linear relationship.
- $H_1 : \rho \neq 0$ : linear relationship exists.

Given  $r \approx -0.0061$ ,  $p \approx 0.9576$  and  $\alpha = 0.05$ :

```
1 def test_hypothesis(p_value, alpha=0.05):
2     print(f'Given the p-value={p_value} and alpha={alpha}:')
3     if p_value < alpha:
4         print("Reject the null hypothesis: \nThere is evidence of a linear
5             ↪ relationship.")
6     else:
7         print("Fail to reject the null hypothesis: \nNo significant linear
8             ↪ relationship.")
9
10 test_hypothesis(p_cats, alpha=0.05)
```

Given the p-value=0.9576068830096687 and alpha=0.05:  
Fail to reject the null hypothesis:  
No significant linear relationship.

### 3.6 Q5: 95% CI of $\rho$

#### Answer

The 95% confidence interval for the population correlation coefficient  $\rho$  is approximately:

$$[-0.2283, 0.2167]$$

#### Solution

Given  $N = 78$ ,  $r = -6.12 \times 10^{-3}$ , the 95% confidence interval for the population correlation coefficient  $\rho$  is then calculated based on the Fisher  $z$ -transformation, steps of calculation is listed below and I wrapped the calculation procedure in a single function:

$z$ -transformation

$$z = \frac{1}{2} \ln\left(\frac{1+r}{1-r}\right) = \text{artanh}(r)$$

Standard error of  $z$

$$SE_z = \frac{1}{\sqrt{N-3}}$$

CI in the  $z$ -scale

$$z_{lower} = z - c \times SE_z$$

$$z_{upper} = z + c \times SE_z$$

Inverse Fisher transformation

$$r = \frac{\exp(2z) - 1}{\exp(2z) + 1} = \tanh(z)$$

Finally we get

$$r_{lower} = \tanh(z_{lower})$$

$$r_{upper} = \tanh(z_{upper})$$

```

1  from math import sqrt, tanh, atanh
2
3  n_cats = len(catterplot_df)
4
5  def ci_rho(n, pearson_r, alpha=0.05):
6      # z-transformation
7      z = atanh(pearson_r)
8      # SEM of z
9      se_z = 1 / sqrt(n - 3)
10     # CI in the z score in the 95% level.
11     c = stats.norm.ppf(1 - alpha / 2)
12     z_lower = z - c * se_z
13     z_upper = z + c * se_z
14     # Inverse Fisher transformation
15     r_lower = tanh(z_lower)
16     r_upper = tanh(z_upper)
17     return r_lower, r_upper
18
19 r_lower_cats, r_upper_cats = ci_rho(n_cats, pearson_r_cats)
20
21 print(f'0.95 Confidence Intervals of rho: \n'
22       f'Lower Limit: {r_lower_cats} \n'
23       f'Upper Limit: {r_upper_cats}')

```

```

0.95 Confidence Intervals of rho:
Lower Limit: -0.22833745967041358
Upper Limit: 0.21670822097748485

```

## 4 Final thoughts: Cats good, humans bad.

In Chinese social media, there's a meme called “Cats good, humans bad” (貓好人壞), which humorously suggests that cats are always right and any misbehavior is blamed on humans. My partner and I keep an American Shorthair (see Figure 3) in our rented place, and we call him *the Prince*, not because of his noble lineage, but due to his annoying behaviors that sometimes drive us crazy. For instance, when he desires food, he jumps on our bed and meows loudly, trying to wake us up in the early morning!

I think if I plot the `loudness` of our Prince's meowing, the peak would definitely appear BEFORE mealtime, as he meows to summon his servant.

And he eats better than we do! According to my partner, she steams duck breasts for our cat at low temperature to keep the meat juicy, a process that takes hours to prepare. Meanwhile, his two poor



Figure 3: The Prince

owners often resort to instant food from the supermarket or just grab a meal from the nearby Saizeriya. (Yes, there's a Saizeriya close to our block!)

Yet, we always remind ourselves that we are not serving the Prince to the best of our abilities. The kitty is always innocent!

貓好人壞.

Alright, it's good to be a cat like the Prince, who keeps meowing at the door, knowing that at the start and the end of the day, a two-legged monster will appear to feed him, regardless of how stressful and dangerous the world outside may be.

We all love him, without any doubt!