Correlation and Regression

QL 1.1

Learning Outcomes

By the end of this session, you should be able to...

1. Draw a scatter plot for a set of ordered pairs.

2. Compute the correlation coefficient.

3. Compute the equation of the regression line.

4. Apply the newly acquired knowledge to calculate correlation and regression in a dataset using numpy and python library.

Relationship among variables

- Effect of dust on respiratory health?

- Study hours ⟷ Student's score on an exam

- Is coffee related to heart damage?

- A businessperson wants to know whether the sale for a given month is related to the amount of advertising the firm does that month

Correlation

- **Correlation** is a method to determine whether a relationship between variables exists.

Scatterplot and Correlation

- A scatterplot is a graph of the ordered (x, y) of numbers consisting of the independent variable x and dependent variable y.
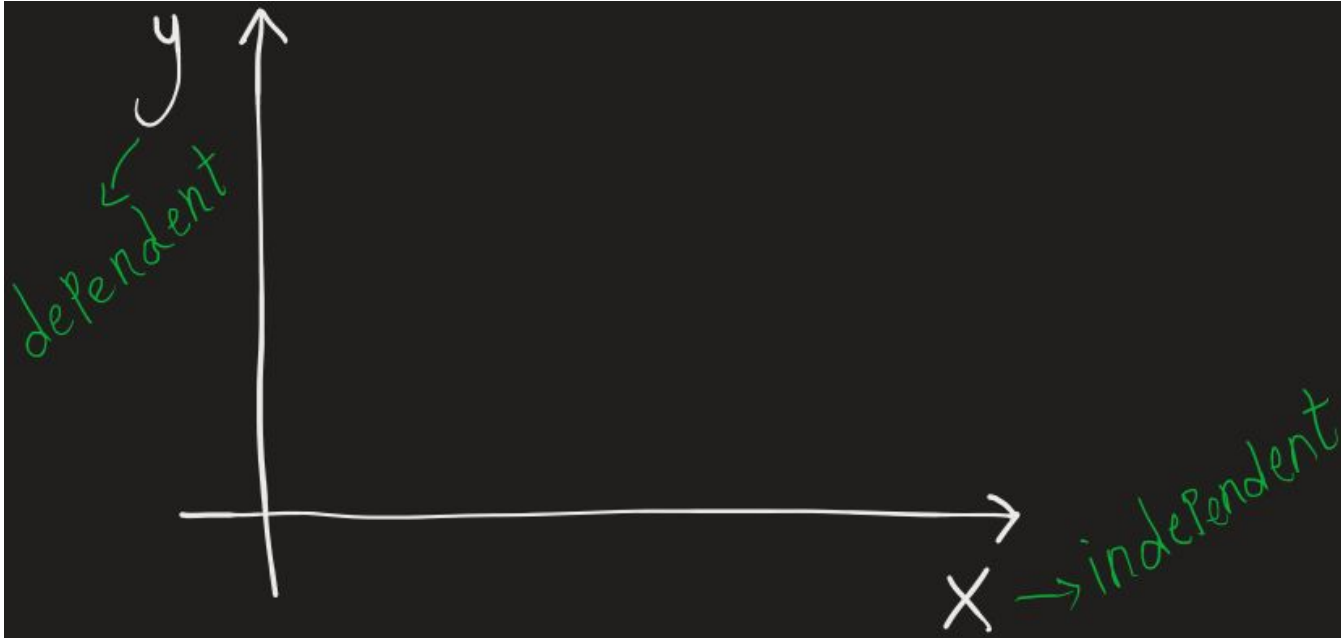
The following table shows the hours of study for each students and their corresponding grades. Plot the data using scatterplot.

Student	Hours of study x	Grade y (%)
A	6	82
B	2	63
C	1	57
D	5	88
E	2	68
F	3	75

Click here for scatterplot activity.

Absences and Final Grades

5 min

The correlation coefficient computed from the data measures the **strength** and **direction** of a **linear** relationship between two variables.

r: **Sample** correlation coefficient

ρ: **Population** correlation coefficient

-1 < Correlation Coefficient < +1

Strong negative linear relationship

No linear relationship

Strong positive linear relationship

−1   0   +1
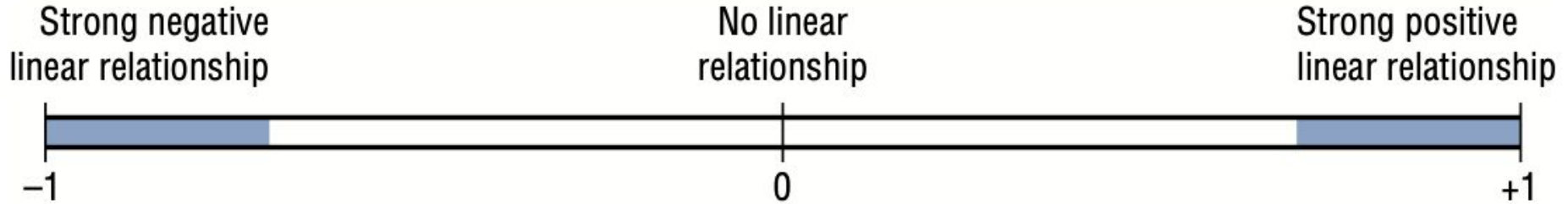
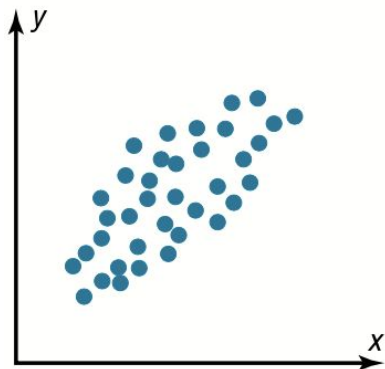**Range of Values for the Correlation Coefficient**

Do not memorize this formula!

Formula for the Correlation Coefficient r

$$r = \frac{n(\Sigma xy) - (\Sigma x)(\Sigma y)}{\sqrt{[n(\Sigma x^2) - (\Sigma x)^2][n(\Sigma y^2) - (\Sigma y)^2]}}$$
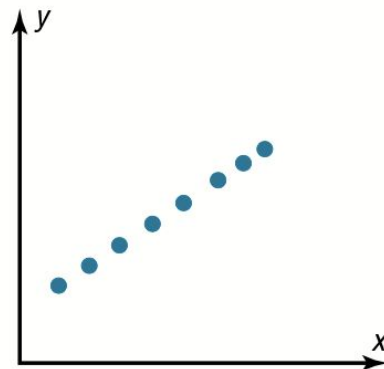
where n is the number of data pairs.
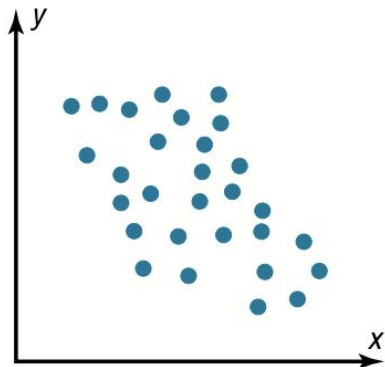
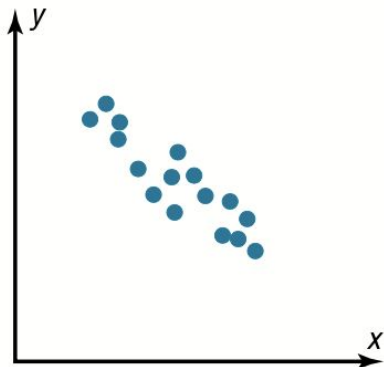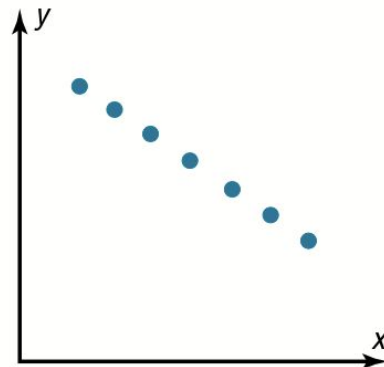# Correlation Coefficient and Scatterplot

(a) r = 0.50
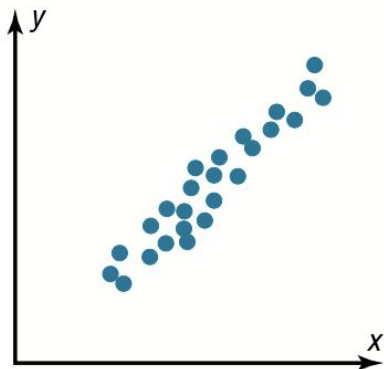
(b) r = 0.90

(c) r = 1.00

(d) r = −0.50

(e) r = −0.90

(f) r = −1.00

Break 10 mins
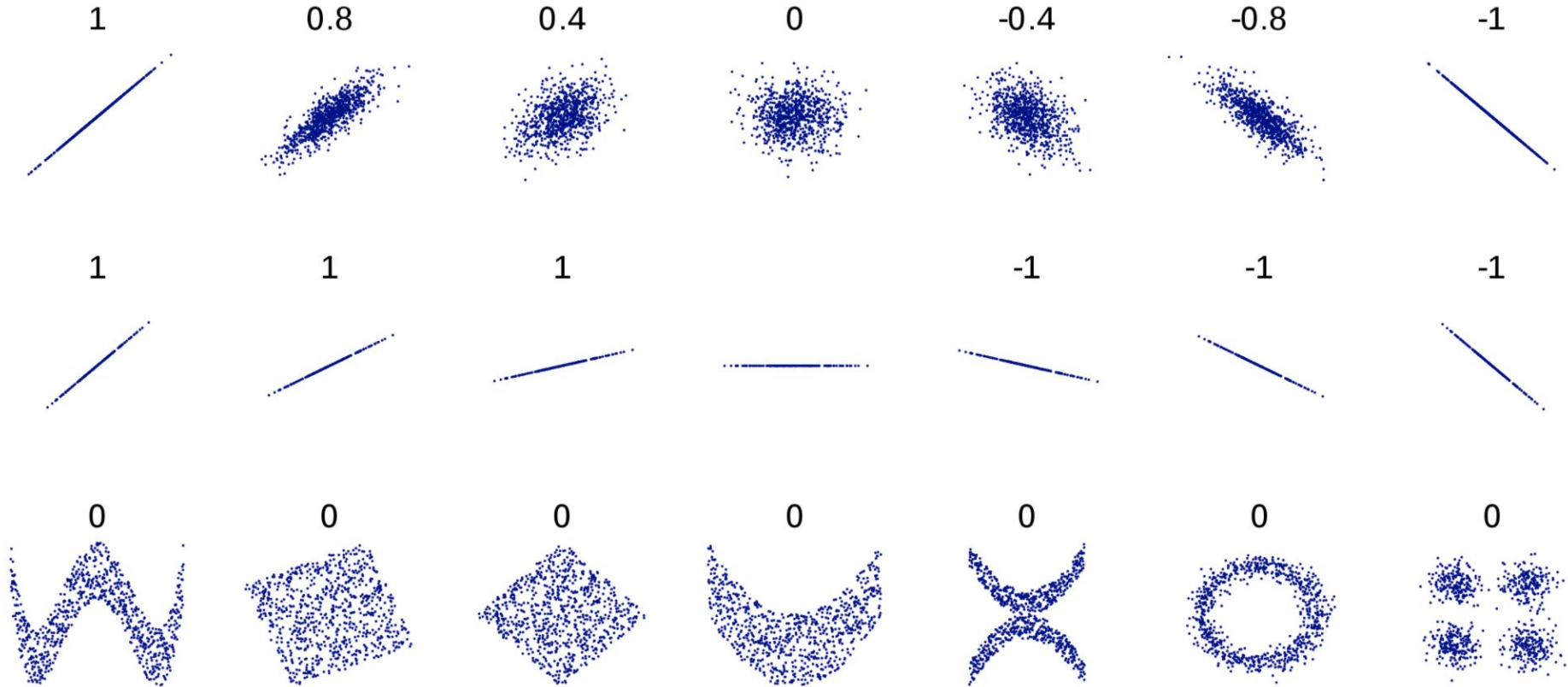
Correlation Coefficient and Scatterplot

The following table shows the hours of study for each students and their corresponding grades. Compute the correlation coefficient.

Student	Hours of study x	Grade y (%)
A	6	82
B	2	63
C	1	57
D	5	88
E	2	68
F	3	75

Absences and Final Grades

5 mins

Click here for the activity.

(go to the activity 2)

Correlation and Causation

5 mins

Assume there is a high correlation between two variables. Then what we can say about the two variables?

Discuss it in groups.

Correlation & Causation

1. There is a **direct** cause-and-effect relationship between the two variables. That is, x causes y.

2. There is a **reverse** cause-and-effect relationship between the variables. That is, y causes x.

3. The relationship between the variables may be caused by a third variable (lurking variable)

4. There may be a complexity of interrelationships among many variables.

5. The relationship may be coincidental.

Regression

You are studying the relationship between two variables:

1. You collect data

2. Construct a scatterplot

3. Compute the correlation coefficient

4. Test the significance of the relationship (* p-value, out of ql1.1 scope)

5. If the value of corr. Coef. is significant, then the next step is to determine

   the equation of the **regression line.**

Regression Line

The purpose of the regression line is to enable the researcher

to see the trend and make predictions. (inferential statistics)

Note: Determining the regression line when r is not significant

and then making prediction using the regression line are

meaningless.