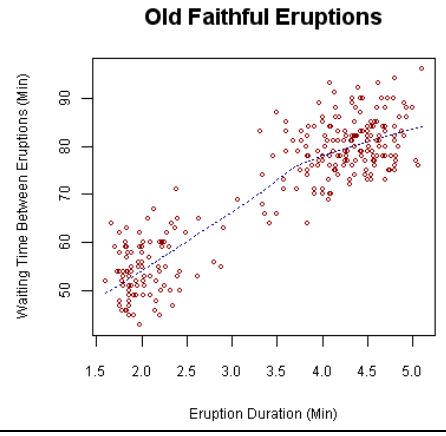
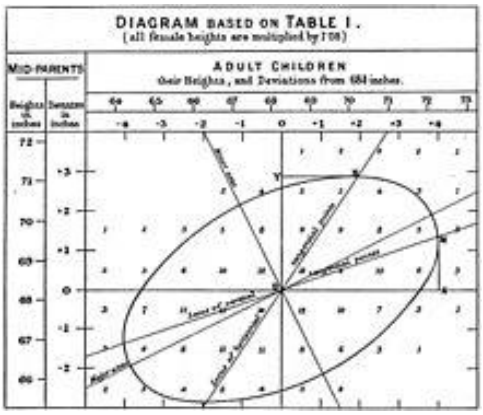
Two-Dimensional Data Analysis

I. Key mathematical terms

Terms	Symbol	Chinese translation
Scatter plot/diagram		
Independent variable		
Dependent variable		
Correlation coefficient		
Least square method		
Regression line		

II. Scatter plot

Scatter plot is a graphical representation of data points. It can show the relationship between two variables. In the diagram, each point represents values for both variables. Here are some examples that occur in our daily lives.

1. Waiting time between eruptions and the duration of the eruption (The Old Faithful Geyser in Yellowstone National Park)	2. Heights of adults and children by Sir Francis Galton.
	

https://zh.wikipedia.org/wiki/%E6%95%A3%E5%B8%83%E5%9C%96#/media/File:Oldfaithful3.png

https://www.researchgate.net/figure/Galtons-smoothed-correlation-diagram-for-the-data-on-heights-of-parents-and-children_fig15_226400313

How to draw the scatter plot?

If you gathered some data and want to use a scatter plot to understand the relationship between the data. You can follow the steps below to progressively create the scatter plot.

Step 1: Collect data

Collect pairs of data where a relationship is suspected.

Step 2: Choose the units for the axes

The units for the axes will depend on the data you are investigating, which could be scores, lengths, weights, etc.

Step 3: Plot Points

Plot points on the selected axes according to the units. For each pair of data, place a dot or a symbol on the corresponding coordinate.

Step 4: Observe trends

Observe the pattern of points to see if a relationship is obvious.

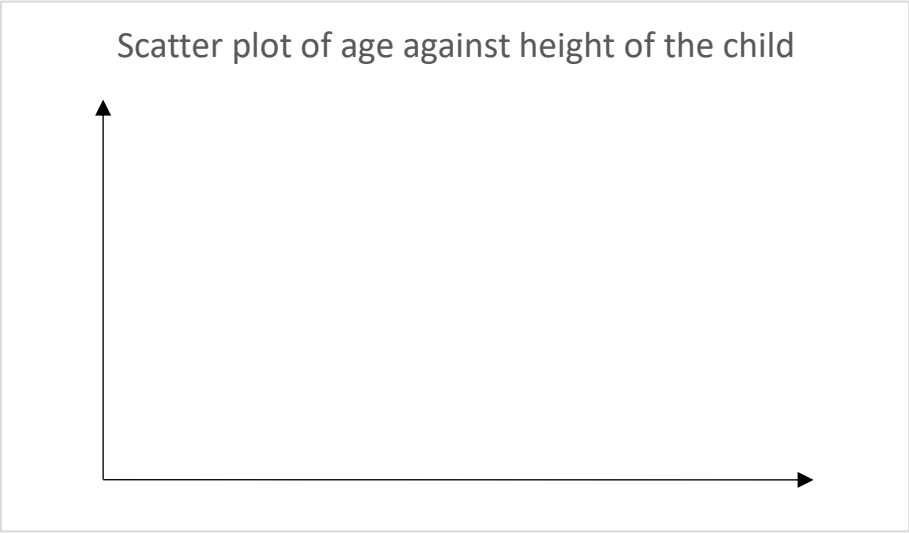
<key>

To determine the relationship of the data, we'll introduce correlation coefficient and regression line later.

Example 1

The following data gives the age of the child in years and height. Please draw the scatter plot of age of the child against height of the child (in centimeters).

Age of the child	3	4	5	6	7	8	9	10
Height of the child	70	83	95	110	115	122	130	138

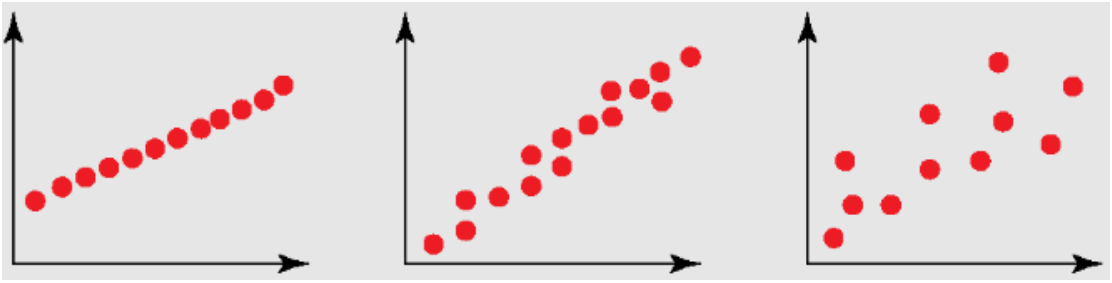


How would you describe the relationship between these data points?

In the previous question, how did you describe your scatter plot? In mathematics, we use positive correlation (strong/weak), negative correlation (strong/weak), or zero correlation to describe the relationship between data. Let's take a look at all the situations together.

Positive, negative and zero correlation

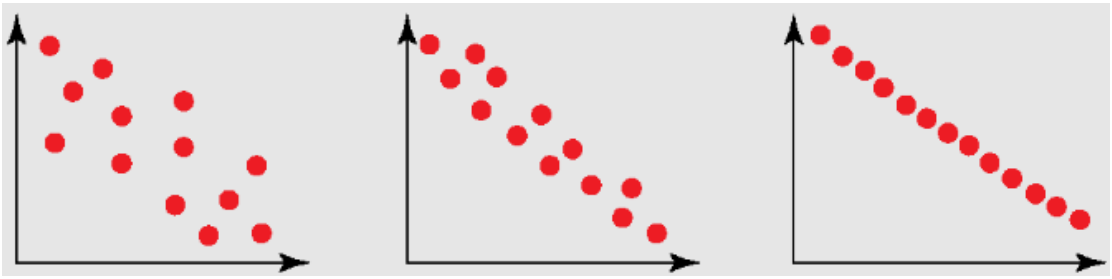
When the points in a scatter plot are distributed in a trend that gradually moves from the **bottom left to the top right**, we call this type of relationship positive correlation.



strong ⟶ weak

Positive correlation

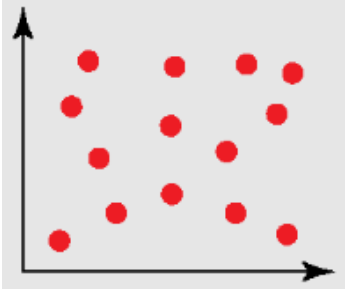
When the points in a scatter plot are distributed in a trend that gradually moves from the **top left to the bottom right**, we call this type of relationship negative correlation.



weak ⟵ strong

Negative correlation

The closer the points are to the trend line, the stronger we consider the correlation of the data to be; conversely, the farther the points are from the trend line (or we cannot find the trend line), the weaker the correlation is. For example:

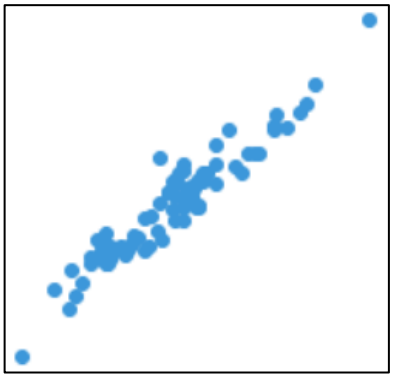


Zero correlation (unrelated)

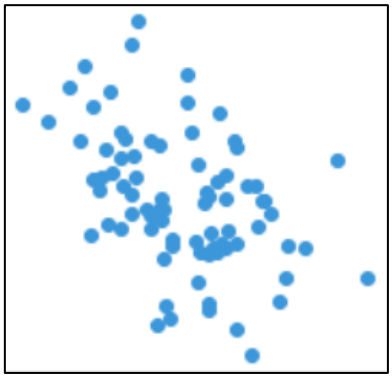
Example 2

Please try to classify the following scatter plots as strong/weak positive correlation, strong/weak negative correlation, or zero correlation.

(1)

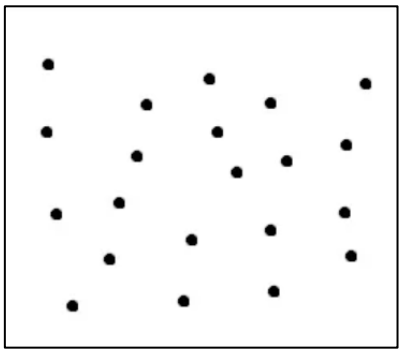


(2)



(3)

(4)



Now, you may know how to analyze data correlation using scatter plots. However, scatter plots are not precise enough. We hope to use quantitative indicators to describe the correlation of data, which can show positive and negative differences and can also have specific range limitations. Therefore, calculating the correlation coefficient is essential.

III. Correlation coefficient

The degree of association can be measured by a correlation coefficient (Pearson correlation coefficient), denoted by r. The correlation coefficient is measured on a scale that varies from 1 to -1.(specific range limitations) When all the points on the scatter plot align along a straight line that isn't vertical or horizontal, we refer to this case as the perfect correlation. When one variable increases as the other increases, the correlation is positive; when one decreases as the other increases, the correlation is negative. The complete absence of correlation (vertical straight line, horizontal straight line, symmetric graph, etc.) is represented by 0.

To find the correlation coefficient, you need the following formulas:

Correlation coefficient

Let $(x_1, y_1),\ (x_2, y_2),\ldots,\ (x_n, y_n)$ be n sets of two-dimensional data.

$$\mu_x = \frac{1}{n}(x_1 + x_2 + \ldots + x_n),\ \mu_y = \frac{1}{n}(y_1 + y_2 + \ldots + y_n)$$

Then the correlation coefficient between variables x and y is: $r_{xy} = \frac{S_{xy}}{\sqrt{S_{xx}}\sqrt{S_{yy}}}$.

$$S_{xx} = (x_1 - \mu_x)^2 + (x_2 - \mu_x)^2 + \ldots + (x_n - \mu_x)^2,$$
$$S_{yy} = (y_1 - \mu_y)^2 + (y_2 - \mu_y)^2 + \ldots + (y_n - \mu_y)^2,$$
$$S_{xy} = (x_1 - \mu_x)(y_1 - \mu_y) + (x_2 - \mu_x)(y_2 - \mu_y) + \ldots + (x_n - \mu_x)(y_n - \mu_y)$$

How to find the correlation coefficient?

The formula of the correlation coefficient is very complicated, to find the correlation coefficient efficiently, you can use the following steps:

Step 1: Find the sample means

Find the sample means of x and y. Calculate (μ_x, μ_y).

Step 2: Distance of each data point from its mean

Find the distance of each data point from its mean. Calculate $x_i - \mu_x,\ y_i - \mu_y$.

Step 3: Complete the top and the bottom of the coefficient equation

Find the following values:

$$S_{xy} = (x_1 - \mu_x)(y_1 - \mu_y) + (x_2 - \mu_x)(y_2 - \mu_y) + \ldots + (x_n - \mu_x)(y_n - \mu_y)$$
$$S_{xx} = (x_1 - \mu_x)^2 + (x_2 - \mu_x)^2 + \ldots + (x_n - \mu_x)^2$$
$$S_{yy} = (y_1 - \mu_y)^2 + (y_2 - \mu_y)^2 + \ldots + (y_n - \mu_y)^2$$

Step 4: Put the numbers into the formula

Put the numbers into the formula and finish the calculation. $r = \frac{S_{xy}}{\sqrt{S_{xx}}\sqrt{S_{yy}}}$

<key>

Please remember the steps above, which will be needed for future questions.

Example 3

The following table shows the price of bubble tea (x) and the sales volume (y). Find the correlation coefficient of x and y.

Price of bubble tea (x)	28	29	30	31	32
Sales volume (y)	11	12	10	8	9

<sol>

Step 1:

Step 2:

Step 3:

Step 4:

Properties of correlation coefficient

Once we have the formula for the correlation coefficient, we can observe the properties through the formula.

Property 1: Independence of units

The correlation coefficient is unitless, unaffected by changes in the unit of data.

Property 2: Positive/Negative

When $r > 0$, gives that data are positively correlated. When $r < 0$, gives that data are negatively correlated

Property 3: Bounded range

The correlation coefficient has bounded range $|r| \leq 1$.

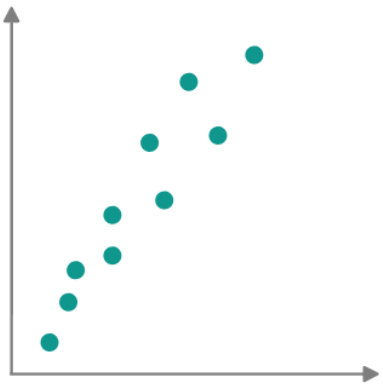
Property 4: Magnitude of strength

When $|r|$ is closer to 1, it indicates a stronger linear correlation. When it is closer to 0, it suggests a weaker linear correlation.

Example 4

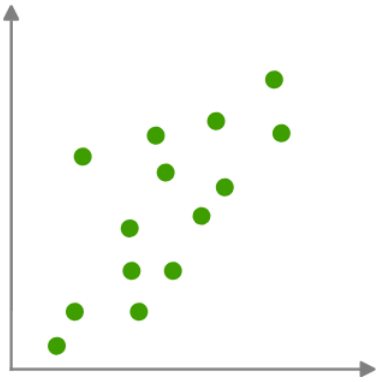
The scatter plots below correspond to three sets of data. Explain and compare the magnitudes of the correlation coefficients for these data sets.

(1)



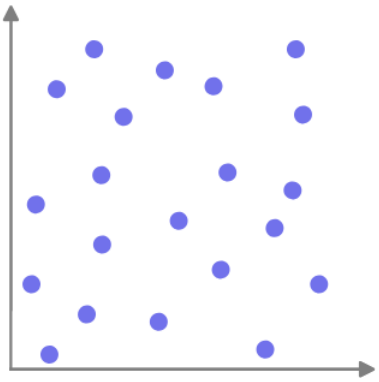
The points in the scatter plot roughly form a trend line with a positive slope.

(2)



The points in the scatter plot also form a trend line with a positive slope, but compared to the previous data points, they are more scattered.

(3)



The points in the scatter plot do not form a trend line, indicating almost no correlation.

Hence, we have the magnitudes of the correlation coefficients: $r_1 > r_2 > r_3$.

When solving certain problems, you do not necessarily have to calculate the correlation coefficient. Instead, you can estimate the correlation coefficient range by observing the graph trend or the distribution of data points. You can save time this way.

When you discover a linear correlation in the data, the next step is to find the most suitable line to describe the relationship between two variables precisely. Through this line, we can not only understand the relationship between the two variables, but also predict the corresponding y values by using x values. However, how do we find this line? We utilize the "Least square method" to determine this line.

IV. The best-fitting line

The best-fitting line refers to the straight line that best represents the relationship between two variables in a dataset. To assess whether this line is the best-fitting line we are looking for, let's first introduce some common notation:

* y_i denotes the observed response for experimental unit i (通常為資料點的 y 值)
* x_i denotes the predictor value for experimental unit i (通常為資料點的 x 值)
* $\hat{y}_i$ is the predicted response(fitted value) for experimental unit i, and $\hat{y}_i = ax_i + b$.

In general, when we use $\hat{y}_i = ax_i + b$ to predict the actual response y_i, we make a prediction error (residual error) of size: $e_i = y_i - \hat{y}_i$.

A line that fits the data best will be one for which the n prediction errors are as small as possible. To achieve this goal is to invoke the "least squares criterion," which minimizes the sum of the squared prediction errors.

* The equation of the best-fitting line is: $\hat{y}_i = ax_i + b$
* The residual error is: $e_i = y_i - \hat{y}_i$
* The squared residual error: $Q = (y_1 - \hat{y}_1)^2 + (y_2 - \hat{y}_2)^2 + \ldots + (y_n - \hat{y}_n)^2$

We minimize the equation for the sum of the squared residual error and get the least squares estimates for a and b.

*a: $a = \frac{S_{xy}}{S_{xx}} = r \times \frac{\sigma_y}{\sigma_x}$, b: $b = \mu_y - a\mu_x$

(r is the correlation coefficient, μ_x, μ_y are the mean and S_{xy}, S_{xx} are what we mentioned on page 5.)

Example 5

Use the least squared method to find the best-fitting line for the data points:

x	-3	-2	-1	1	2	3
y	-5	-3	0	3	4	1

Example 6

Suppose we have a group of data with the following conditions:

$$n=8,\ \mu_x=65,\ \mu_y=70,\ \sigma_x=10,\ \sigma_y=5,\ r_{xy}=0.8$$

(1) Find the equation of the line of best fit for the data.

(2) If $x=60$, predict the value of y.

Example 7

A company wants to launch a new product and is conducting surveys on different unit prices X and market demand quantities Y before it releases. The survey results are as follows:

X	6	7	8	9	10
Y	9	10	8	6	7

(1) Find the correlation coefficient of X and Y.

(2) Find the best-fitting line for the data points(For Y in terms of X).

(3) If the pricing at launch is set at 7.5, using the best-fitting line, estimate approximately how much market demand there will be.

<資料來源>

1. Scatter plot

https://d138zd1ktt9iqe.cloudfront.net/media/seo_landing_files/diksha-q-how-to-calculate-correlation-coefficient-01-1609233340.png

https://byjus.com/maths/scatter-plot/

https://www.atlassian.com/data/charts/what-is-a-scatter-plot

2. Correlation coefficient

https://www.mathsisfun.com/data/correlation.html

3. 南一書局數學（二）

製作者：國立臺灣師範大學附屬高級中學 蕭煜修