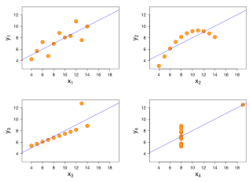
MDM4U: Mathematics of Data Management

How Are Data Related?

Scatter Plots and Linear Correlation

J. Garvin



Correlation

Sometimes we are interested in how changes in one variable relates to those of another.

Correlation is a numerical value that represents the degree to which two variables are related.

Correlation can also be viewed as a measure of how much one variable *depends* on another.

In this unit, we will talk about different types of correlations.

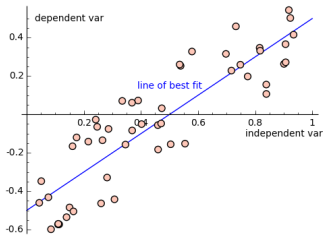
Scatter Plots

A *scatter plot* is a tool that can be used to visualize the relationship between two variables.

On the horizontal axis is the *independent variable*, and on the vertical axis is the *dependent variable*.

A *line of best fit* can be used to estimate a relationship between the variables.

Scatter Plots



Classification of Linear Correlations

If changes in the independent variable are proportional to those in the dependent variable, then the two variables demonstrate a *linear correlation*.

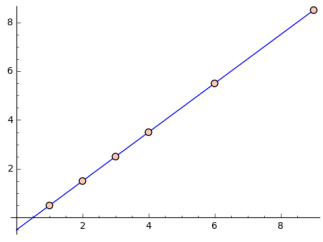
If the dependent variable increases as the independent variable increases, then the linear correlation is *positive*.

If the dependent variable decreases as the independent variable increases, then the linear correlation is *negative*.

In the previous example, the scatter plot shows a positive linear correlation.

Classification of Linear Correlations

If all data lie on the *line of best fit*, then the linear correlation is *perfect*.



In practice, it is rare to obtain a perfect linear correlation.

Classification of Linear Correlations

If most data are close to the line of best fit, we classify the linear correlation as *strong*.

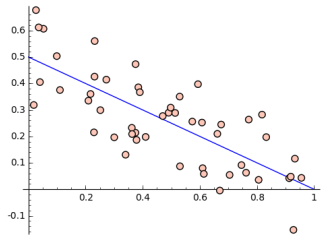
If most data are widely dispersed, the linear correlation is *weak*.

Linear correlations that are neither that close, nor that dispersed, are *moderate*.

Classification of Linear Correlations

Example

Classify the correlation in the following scatter plot.

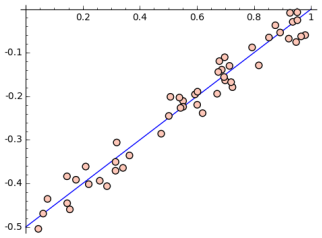


The scatter plot shows a moderate negative linear correlation.

Classification of Linear Correlations

Example

Classify the correlation in the following scatter plot.



The scatter plot shows a strong positive linear correlation.

Correlation Coefficient

The strength of a linear correlation can also be assigned a numerical value, known as the *correlation coefficient*.

Pearson's Coefficient of Correlation

The correlation coefficient of a linear relationship is given by $r = \dfrac{n\sum xy - (\sum x)(\sum y)}{\sqrt{[n\sum x^2 - (\sum x)^2]\,[n\sum y^2 - (\sum y)^2]}}$, where each of the n data has a coordinate (x, y).

Despite its appearance, this is relatively straightforward to calculate using a table.

We need the sums of each of the following: the x values, the y values, the products of each x-y pair, the x^2 values, and the y^2 values.

Correlation Coefficient

Example

Given the following data relating an object's length (cm) to its mass (kg), calculate the correlation coefficient and classify the correlation. Illustrate the data using a scatter plot.

Length	15	22	19	18	15	45	27	18	18	51
Mass	2.8	3.5	3.4	2.8	3.0	6.1	4.2	3.2	2.9	7.0

Correlation Coefficient

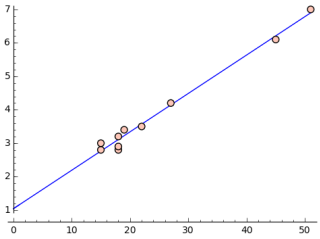
x	y	xy	x^2	y^2
15	2.8	42.0	225	7.84
22	3.5	77.0	484	12.25
19	3.4	64.6	361	11.56
18	2.8	50.4	324	7.84
15	3.0	45.0	225	9.00
45	6.1	274.5	2025	37.21
27	4.2	113.4	729	17.64
18	3.2	57.6	324	10.24
18	2.9	52.5	324	8.41
51	7.0	357.0	2601	49.00
248	38.9	1133.7	7622	170.99

Correlation Coefficient

$$\begin{aligned} r &= \frac{n\sum xy - (\sum x)(\sum y)}{\sqrt{[n\sum x^2 - (\sum x)^2]\,[n\sum y^2 - (\sum y)^2]}} \\ &= \frac{10 \cdot 1133.7 - 248 \cdot 38.9}{\sqrt{(10 \cdot 7622 - 248^2)(10 \cdot 170.99 - 38.9^2)}} \\ &\approx 0.9932 \end{aligned}$$

Thus, the data demonstrate a strong, positive linear correlation.

Correlation Coefficient



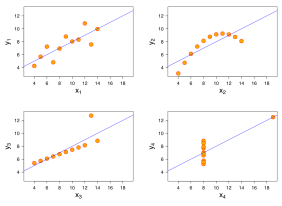
The correlation coefficient, while useful, is just a number and may or may not be relevant to the data at hand.

It is important to graph the data before any conclusions can be made about the significance of the correlation coefficient.

Correlation Coefficient

Example

Which graph has the greatest correlation coefficient?



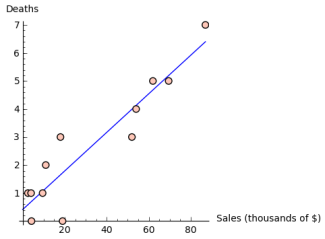
All have a correlation coefficient of approximately 0.816, even though all of the relationships are not linearly related.

Correlation and Causation

While the previous example illustrated a fairly common *cause-and-effect* relationship – that an increase in length results in an increase in mass – a high correlation coefficient does not necessarily imply causation.

Consider the following scatter plot comparing monthly ice cream sales and deaths due to drowning.

Correlation and Causation



Correlation and Causation

The data are relatively close to the line of best fit, and the correlation coefficient is 0.859, suggesting a strong, positive correlation between ice cream sales and deaths due to drowning.

It does not make sense, however, to conclude that an increase in ice cream sales causes an increase in the number of deaths.

Instead, the strong correlation is probably due to another factor – in this case, the time of year.

We will discuss external sources of bias in more detail later.

Questions?