SCATTERPLOTS AND ASSOCIATIONS

It is common for two data sets to have an association where one set of variables may influence the other. For example if you were to measure a random group of people's height and weight you would probably find a positive association, i.e. taller people are more likely to be heavier.

Drawing a scatterplot provides a visual representation of any trend or underlying pattern in the data. A scatterplot is drawn by treating the bivariate data as a series of coordinate pairs and plotting the pairs on a suitable set of axes.

The data used often comes from measurements of real-life situations, so the graph is quite often limited to the positive part of the scale.

To draw a scatterplot:

Step 1 Draw a suitable set of axes by noting the range of each set of figures. If it is sensible, start with zero and have between 8 and 12 convenient, evenly spaced points on each scale.

Step 2 Treat each data pair as a coordinate point and place a mark on the grid. Do not join the points.

Example 3

Draw a scatterplot of the following bivariate data set and describe any trend you see.

Person	A	B	C	D	E	F	G	H
Height (cm)	140	160	180	90	100	50	60	120
Weight (kg)	60	75	95	40	50	20	35	65

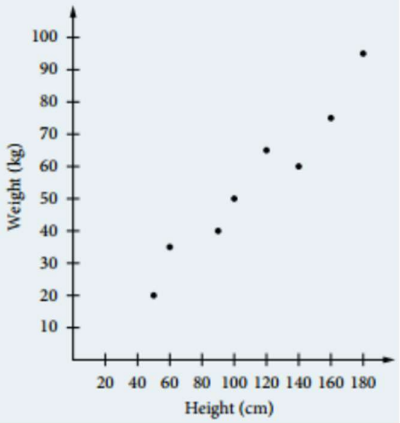
Solution

Draw a suitable set of axes by noting the range of each set of figures:

In this case Height is from 0 (cm) to 180 (cm) so start at zero and use 9 divisions of 20 (cm).

Weight varies from 0 (kg) to 95 (kg) so start at zero and use 10 divisions of 10 (kg) to cover the required range.

It appears that, in general, the taller the person is, the heavier they are.



Often one of the variables can explain the association and this is known as the explanatory variable, or independent variable. The other variable responds to a change in the independent variable and is known as the response variable, or dependent variable. In the case of height versus weight, a person's weight is more likely to be explained by their height than the other way around.

The horizontal axis is used for the independent variable and the dependent variable is on the vertical axis.

Examples:

- Temperature of the body responds to, or depends upon, the time spent in front of a heater.
- The number of ice-creams sold will respond to or depend on the temperature of the day.

Warning

Not all bivariate data sets have clear independent and dependent variables. They may both be dependent on a third variable. For example, a student's study scores in English and Music may appear related but both depend on other variables such as effort, intelligence, hours of practice and study.

Example 4

Assuming an association exists, identify the independent variable in each of the following pairs:

- age and wealth
- age and the number of offspring
- temperature and the number of people at the beach
- number of cigarettes smoked and chance of cancer
- the volume of petrol remaining in your tank and the distance you have driven.

Solution

Ask yourself which of the variables could cause or explain a change in the other:

Age and wealth: The independent variable is age.

Age and the number of offspring: The independent variable is age.

Temperature and the number of people at the beach: The independent variable is temperature.

Number of cigarettes smoked and chance of cancer: The independent variable is number of cigarettes smoked.

The volume of petrol remaining in your tank and the distance you have driven: The independent variable is distance you have driven.

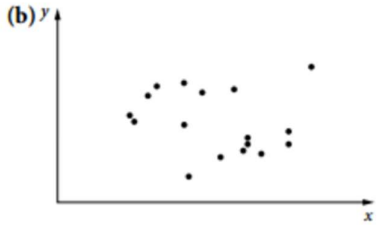
Linear trend

A single line that best represents the general pattern of the data is called a line of best fit.

If the general pattern of the data is roughly a straight line, a linear association exists.

Sometimes a pattern may exist but it is not a straight line. This is called a non-linear trend. Scatterplot **(a)** to the right is an example of a non-linear trend.

In other situations the points may be completely random and indicate no association at all. Scatterplot **(b)** to the right is an example of no association.



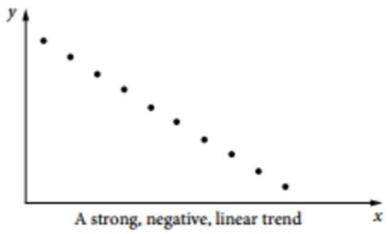
The pattern may show either a positive, increasing trend (going up as you move to the right) or a negative, decreasing trend (going down as you move to the right).

When analysing the scatterplot, look for the following characteristics:

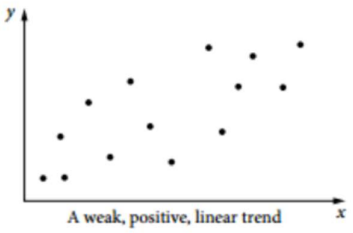
- Is there is an observable pattern?
- Does the pattern show a linear or non-linear association?
- Is the slope positive or negative?
- Identify any outliers, that is, single points that seem to be well outside the general pattern of the rest of the data. If the outliers are excluded from the data, the pattern will be easier to see.
- How well does the pattern fit represent the data? (Is it strong, moderate or weak?) A good fit represents a strong trend and a poor fit represents a weak trend.

SCATTERPLOTS AND ASSOCIATIONS

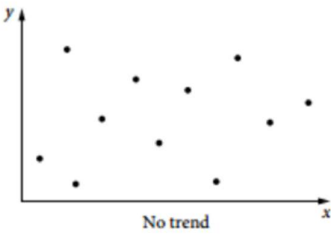
Examples are as follows:



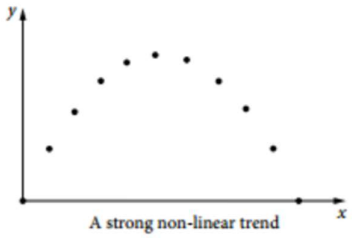
A strong, negative, linear trend



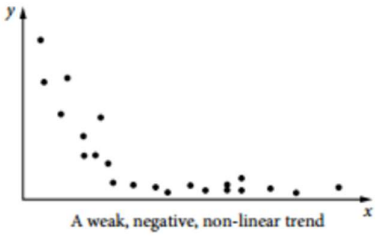
A weak, positive, linear trend



No trend



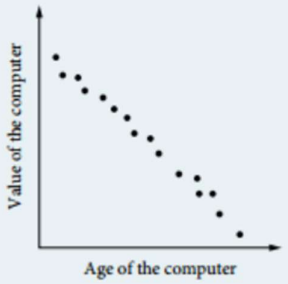
A strong non-linear trend



A weak, negative, non-linear trend

Example 5

Describe the associations between the variables represented in the following scatterplot.



Solution

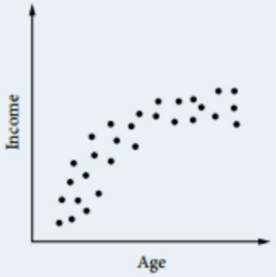
Is there a pattern? Check for linearity. Is the slope positive or negative and how well does the pattern represent the data?

There is a pattern, with a straight line and the slope is negative. The scatterplot shows a strong negative linear relationship between the age of the computer and the value of the computer.

Therefore, there is a strong, negative, linear association.

Example 6

Describe the associations, if any, in the following scatterplot.



Solution

Is there a pattern? Check for linearity. Is the slope positive or negative and how well does the pattern represent the data?

There is a pattern, which is non-linear and the slope is positive. The scatter plot shows a moderate positive linear relationship between age and income.

Therefore, there is a moderate, positive, non-linear association.

Example 7

Describe the associations, if any, in the following scatterplot.

Solution

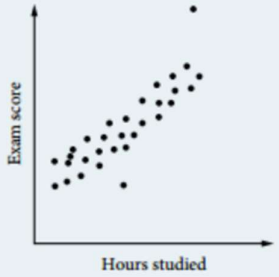
Is there a pattern? Check for linearity. Is the slope positive or negative and how well does the pattern represent the data?

There is no pattern. The scatter plot shows no relationship between Medicare number and income.

Therefore, there is no association between the variables.

Example 8

Describe the associations, if any, in the following scatterplot.



Solution

Is there a pattern? Check for linearity. Is the slope positive or negative and how well does the pattern represent the data?

(Note the presence of possible outliers in this graph.)

There is a pattern, which is linear and the slope is positive. The scatterplot shows a strong positive linear relationship between hours studied and exam score.

Therefore, there is strong, positive, linear association.