UNISON

Open College

BLOCK HIR.103
STATISTICAL TECHNIQUES

UNISON Open College, 1 Mabledon Place, London WC1H 9AJ

Contents

A. Introduction

Statistics are concerned with summarising associations between variables (variables are simply things that vary, such as the type of houses a landlord has) and provides techniques for analysing data. A statistic is a statement about a set of data.
Examples of these are –

750,000 households live in fuel poverty in Scotland.

1.3 % of dwellings failed the fitness standard in Scotland in 1996.

In Scotland in 1996, 80% of households living in the private furnished rented sector had moved in the past 2 years. In contrast, only 17% of owners with a mortgage had moved.

1. Data Presentation and Descriptive Analysis

When a study is conducted, information or data is collected. Suppose a housing planner is interested in the age and type of dwellings and the tenure of the household in their local authority area. A small survey could be conducted by randomly selecting dwellings from the Council Tax register and a surveyor could assess the dwelling. The data could be set out as shown below.

Example 1

Dwelling	Date of construction	Dwelling type	Housing Costs (£)	Tenure	Number of rooms
1	Pre-1919	Detached house	23	Owner occupied	10
2	Post-1982	Detached house	57	Private rented	8
3	1965-1982	Tower block	32	Owner occupied	3
4	1965-1982	4-in-1 block	29	LA	4
5	1919-1944	Semi-detached house	33	LA	6
6	1965-1982	Tower block	27	LA	4
7	Post-1982	Tenement	34	HA	3
8	1945-1964	4-in-1 block	27	HA	4
9	Pre-1919	Tenement	35	Owner occupied	4
10	1919-1944	Terraced house	32	Owner occupied	6

Note: Housing costs are weekly mortgage payments for owners and weekly rents for HA and LA tenants.

The characteristics that have been collected are known as variables. How a variable is presented and summarised depends on the type of variable.

B. Data Characteristics

1. Types of Variable

There are two types of variable: **quantitative** or **qualitative** variables.

Quantitative variables are those where the values consist of numbers. Examples of these would be average numbers of rooms per property or average rents.

Qualitative variables use words to describe categories of measurement and have no numerical significance. They can be **binary**, which means that they have only two categories (such as a flat or house), **nominal**, which means that they can have several categories (such as tenure – owner-occupied, local authority rented, housing association rented or private rented), or they can be **ordinal**, which means that the variables are in some order, although it is not numerical (such as Council Tax Band – A, B, C, D, E, F, G, H).

2. Summarising Variables

In the example above, only the first 10 cases have been shown, and it is easy to read what the data is showing. In reality, most surveys will have more than 100 (usually much more) and so a means of summarising the data is needed. The best way to organise this task is to examine each variable in turn.

Qualitative variables are summarised by counting the frequency of cases in each category. This tells us how often something occurs or how 'frequent' it is. The proportion of cases in each category is also calculated. These proportions (or percentages) are known as **relative frequencies** and are calculated by dividing the number in the category by the total number and multiplying that answer by 100. The relative frequency tells us how often one variable appears in relation to another. These frequencies can be shown in a number of ways. The best method is to present them in a table.

Example 2

The relative frequency for Owner Occupation in this example = (37254/63035)*100=0.591*100=59.1%.

The total frequency should always be given. Percentages should rounded up and given to no more than one decimal place because they are easier to read if presented that way and do not affect the accuracy of your data. Occasionally percentage totals do not add to exactly 100 but 99.9% or 100.1% because of rounding. This is known as **rounding error** and is acceptable.

An important issue to note when collecting and presenting data is the issue of 'missing cases'. Usually when information is collected not all of the desired data can be found. For example, if someone is asked to participate in a survey, they answer some of the questions but not all of them. This situation is known as having missing data. A survey may have interviewed 100 people, and you may have all of the information about whether they were male or female. However, a few of the respondents may not have wanted to tell you their age. When you are showing results in a table, you should show the number of missing cases along with the number of valid responses. This will allow the reader to judge for themselves the validity of the questions with the missing cases. As a rule of thumb, if less than half the survey answered a particular question, then too much information is lost and the responses you have are not a good representation of the general picture.
Missing values, then, should be represented in a table if they exist, but should not be used in analysis of results.

Example 3

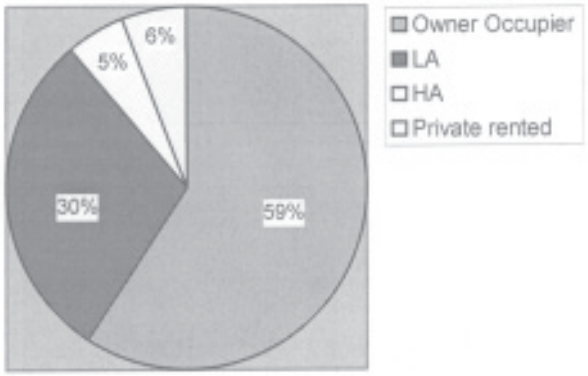
In a survey of tenants' satisfaction, 100 tenants participated. Age was asked of all tenants but six people declined to give their age. The age groups are shown in the table below:

Age group	Count	Percentage
<25	5	5
25-39	26	28
40-59	33	35
60+	30	32
Total	94	100
Missing	*6*	

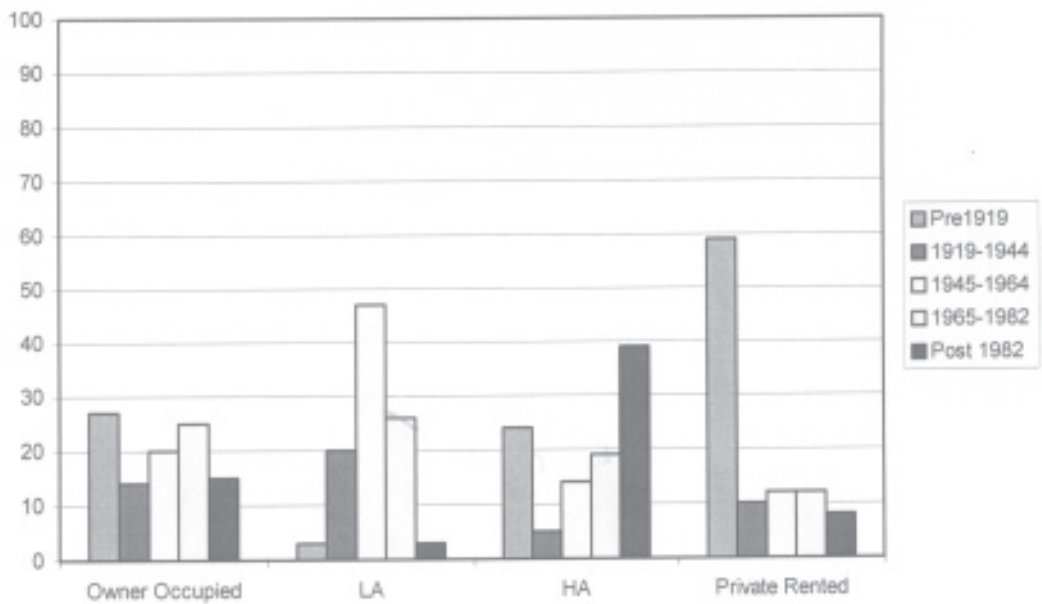
Note: The numbers of missing values are shown in the table but the total used for calculating the percentage is 94 and does not include the missing values.

As well as in tables, qualitative variables can also be presented as pie-charts or bar-charts. There are advantages to diagrams as they are easy to read and give an immediate picture of relative frequency. In the example below, we show a pie chart and a bar chart (or graph). In both cases, you can see that there are advantages in presenting your data in a pictorial form. Doing this gives you the advantage of making an immediate impact by illustrating the information in a way that brings out the important points or information. The pie chart is a circle which is divided into sectors to represent each item or variable. Each sector of the circle should have an area equal to the value of the variable. Pie charts are useful when, as in the example, there are a few items which make up proportions of a whole and where the proportions are more important than the numerical values. This would contrast with the use of tables, as in the above example, where numerical values are an important part of the information you need to show. Tables give actual numbers and are better for binary variables. The bar chart is among the most popular forms of pictorial representation. As you can see, our example shows tenure by age of dwelling. It is the height of each bar which gives the information and makes tenure comparison easy.

Pie chart showing the % dwellings in each tenure.



Bar graph of tenure by age of dwelling.



The bar chart above shows the percentage of age of dwelling by each tenure. It gives an immediate impression that most private renters live in pre-1919 dwellings and very few local authority tenants live in pre-1919 dwellings or post-1982 dwellings. This bar chart gives an example of comparing two qualitative variables – a nominal one and an ordinal one. Four separate bar charts for each tenure could have been produced.

Quantitative variables can be presented in a graph or by summarising the actual numbers. Graphical representations are useful as they give a more immediate picture of how the variable looks for the whole data set. If the variable has whole number categories (for example, the number of bedrooms) then a bar chart or table is a good way of presenting the data. If the variable is continuous (for example, house price data), a histogram should be drawn. To do this, split the variable into groups or intervals and then count the number of cases in each group.

Example 4

Housing costs

Housing Costs (£)	Frequency	Relative frequency
0-14.99	107	5.1
15-29.99	713	33.8
30-44.99	703	33.3
45-59.99	297	14.1
60-74.99	126	6.0
75-89.99	64	3.0
90-104.99	55	2.6
105-119.99	33	1.6
120-134.99	6	0.3
135-150	8	0.4
Total	2,112	100

3. Rules for Histograms

1. Each observation should only be able to be assigned to one interval. End points of the intervals should be specified to make sure this happens. If above the intervals had been 40-50,000 and 50-60,000, which interval would £50,000 go in?
2. Choose the number of intervals in a sensible way. Usually 6-10 intervals are sufficient, although more can be used with bigger sample sizes. Too few intervals may mask peaks in the distribution, too many may result in a strange shape.
3. Use relative frequencies (percentages) instead of actual frequencies so that histograms of different sample sizes can be compared.
4. The scale on the relative frequency axis must start at zero (0).

C. Measures of Central Tendency and Dispersion

1. Central Tendency

When we have collected a set of data, we often try to find a single figure which best represents the results. The most reasonable figure to use here is a central or middle mark. In statistical terms, we are trying to find a measure of **central tendency**.

There are three types of averages we can use – the **mean**, **median** and **mode**.

The mean is the most often used measure of central tendency. It is an arithmetical average. It is calculated by adding up all the values and dividing by the number of cases. From Example 4, the mean value of the housing costs for the whole survey is £32.62 per week. The mean is substantially affected by extreme values and so is therefore best used when the distribution is approximately symmetrical. In this example, a couple of households pay more than £150 per week for their housing costs, where most other values are less than £50. This would affect our mean figure and cause us to ask whether it was an accurate measure to use.

The median is the middle value in our data; half of the cases are below the median and half of the cases are above the median. It is found by putting all the values in increasing order, and if the sample size is odd, then the median is the middle value; if the sample size is even, then the median is halfway between the two middle observations. Median housing costs is £30 per week. The median is not affected by extreme values and is more representative of the centre than the mean for distributions that are skewed (see graphs below).

The third type of average is the mode. The mode is the most common value – in our example it would be the housing cost that occurs most often. However, the mode is not often used as a measure of central tendency, for a number of reasons. First, what do we do if there were two housing costs with the same high frequency? Which one would we choose? Second, there will be occasions when the mode clearly does not represent a central mark. We may have the highest single number of households in the lowest housing cost category, but this may be well short of an 'average' housing cost and thus inappropriate for us to use.

2. Dispersion

As well as the average value, the spread of the data around this value is important to us in having a full knowledge of what our data is telling us.

There are several ways for us to measure the spread or the dispersion of our data.

We can measure the range, which is the difference between the lowest and highest value (for example, rents can range from £10 to £187.50, giving a range of £177.50 per week). The range is not usually recommended though as it depends on only two values which may both be extreme values. And it does not, of course, tell us anything about the general spread of rents in our area.

We could use an interquartile range, which is not affected by extreme values. We saw earlier that the median cuts the data into halves. The quartiles simply cut the data into quarters. The interquartile range is the difference between the lower and upper quartile. The reason that the interquartile range is used is that, unlike the range, it is not going to be affected by one particularly high or low extreme value, and it thus may represent the spread of the distribution more appropriately.

However, using quartiles does not use all the information available from the data and a number of measures of spread have been developed which attempt to use all data.

Most of these use the mean as the 'central' position and compare the rest of the data with the mean to see how far each level of rent varies or deviates from it.

The next two measures of spread involve averaging the distance from the mean of all the values. Each value of the variable will differ from the mean by a certain amount, and if the differences are large (i.e. the histogram is wide), then the spread is also large. To calculate the variance:

calculate the difference between each value and the mean and square it;

sum all of these squared values; and then

divide by (n-1), where n is the number in the sample.

The variance is the average of the squared differences and is not in the same units as the data. This makes the interpretation of the size of the variance difficult. A measure which is in the same units as the data is the standard deviation. This is the square root

of the variance. The standard deviation is affected by extreme value and data skew, but in general 95% of the values will lie within two standard deviations of the mean.

D. Sampling Theory

1. Comparing Two Variables

One of the main aims of carrying out a survey is to establish association between variables. A common type of survey in housing is a tenant's satisfaction survey. The aim of this survey is to find out how satisfied tenants are with their landlord (either LA or HA) and any reasons that some tenants are more satisfied than others. The main variable of interest here is satisfaction with landlord. This main variable is known as the **dependent variable** or **response variable**. Other variables that are collected are known as **independent variables** or **explanatory variables**. Examples of these are area in which tenant lives, size of dwelling, type of dwelling, age of tenant, gender of tenant, amount of rent tenant pays. The independent variables are used to explain the variation in the dependent variable.

For example, in a survey of local authority tenants in two different areas, the tenants in area A expressed more satisfaction in their landlord than those in area B. The explanatory variables collected in the survey will help to answer why that might be the case. It could be due to the fact that area A contains new low-rise houses with gardens whereas area B contains pre-1919 tenements and tower blocks. Or tenants in area B have to deal with many different housing officers and feel that their complaints or requests for maintenance have not been dealt with appropriately.

There are different ways to explore the relationships between variables, depending on the type of variable.

1.1 Binary vs binary and nominal vs nominal

How would you compare satisfaction and gender?

The answer is to construct a table: one variable forms the rows, the other the columns.

	Satisfied with landlord		
Gender	Yes	No	Total
Male	120 (54%)	100 46%)	220 (55%)
Female	120 (67%)	60 (33%)	180 (45%)
Total	240 (60%)	160 (40%)	400

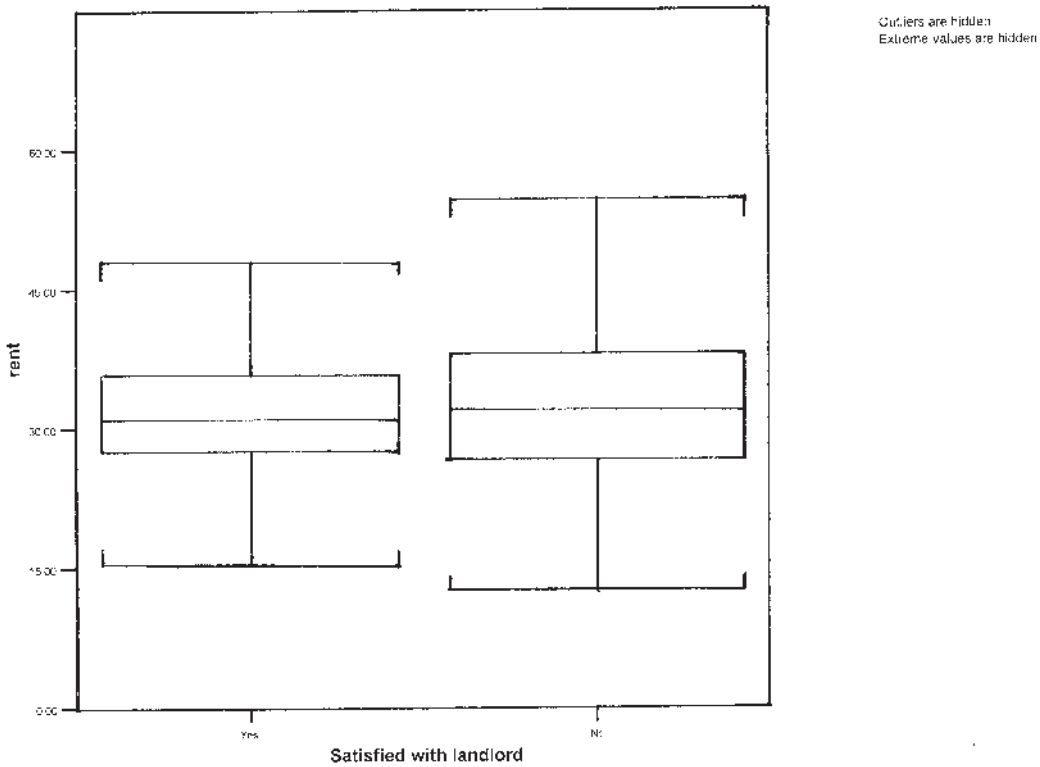
Totals and percentages should always be given.

This information can also be displayed as a bar chart – two bar charts side by side are drawn. As with all graphs an immediate comparison is available.

1.2 Continuous and binary/nominal

How would you compare rent levels with satisfaction?

One way would be to calculate the mean rent for the tenants who are satisfied and the mean rent for the tenants who are not satisfied. Histograms can be drawn – one for each group. Remember to use relative frequencies and same scale and number of intervals to allow comparison. Box plots can also be drawn showing interquartile range, mean and median for all levels of a binary or nominal variable. This allows an immediate appreciation of differences between groups.



1.3 Continuous and Continuous

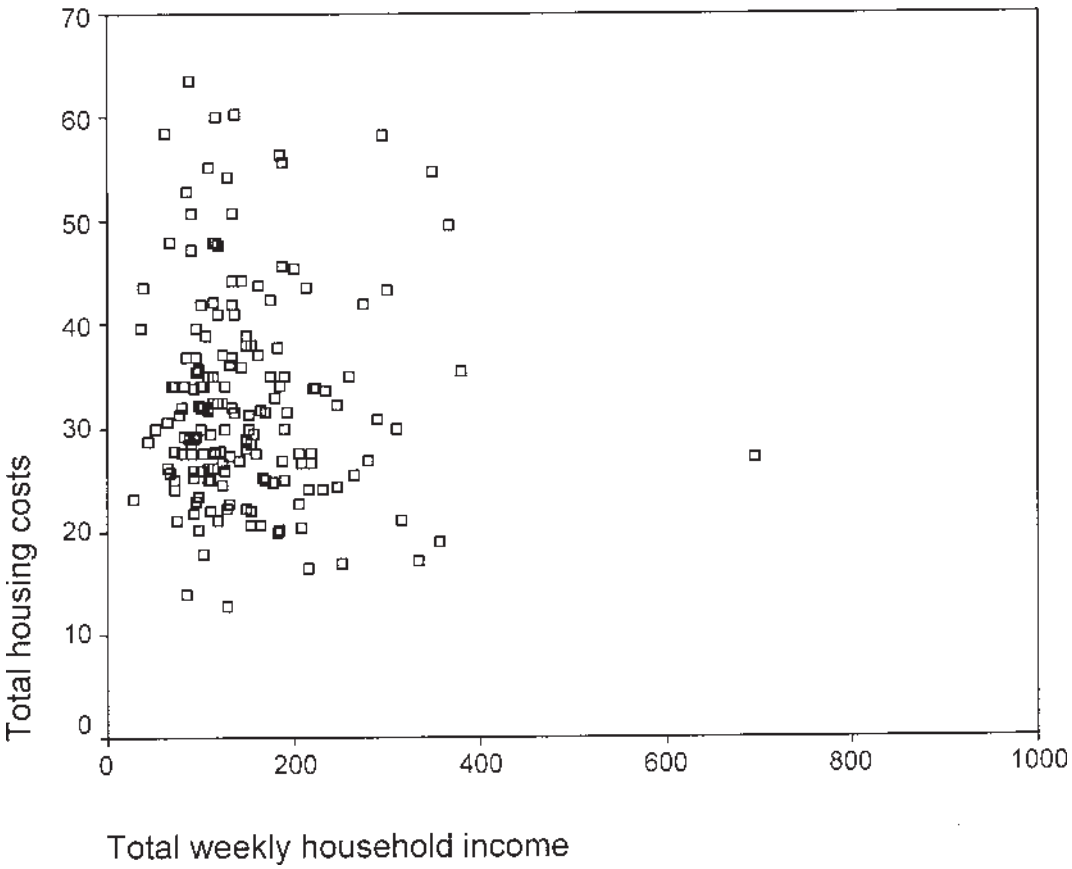
How would you compare rent with income?

A scatter diagram can be drawn to give a visual representation of the relationship between the two variables.

The dependent variable – rent - on the y (vertical) axis; and

the independent variable – income - on the x (horizontal) axis.

This allows you to see what happens to rent as income increases.



E. Regression and Correlation

Regression is a statistical tool that allows the exploration and quantification of the relationship between two continuous variables. Regression determines the mean value for the dependent variable for any given value of the independent variable. The starting point of regression analysis is to draw a scatter diagram. This allows you to assess visually the best relationship between the two variables. To use regression, the relationship should be a straight line; if this is not the case, regression cannot be used. The value of regression is that it allows us to predict how a change in one variable will affect another. If we do not find a relationship between two variables, then they are uncorrelated and a change in one of them cannot be used to predict a change in another. Correlation means that there is a relationship between two variables.

A straight line has a mathematical equation (regression equation) to describe it: y=a+bx, where y is the y-axis (dependent) and x is the x-axis (independent).

Example 5

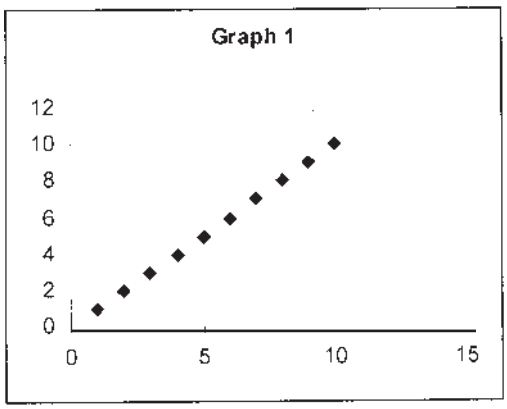
Housing costs = a +b * (income)

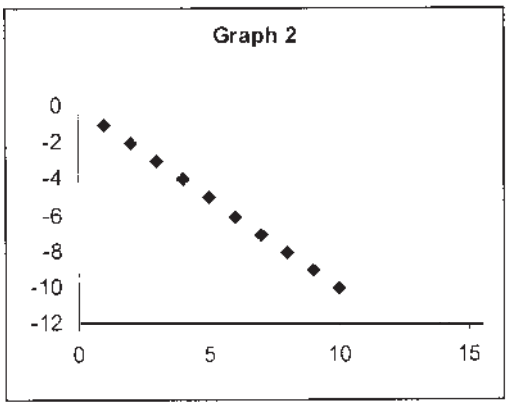
a is the value of the dependent variable when the independent variable is zero (the **intercept**), and b is the amount by which the mean dependent variable increases for every unit increase in the independent variable (the **gradient**).

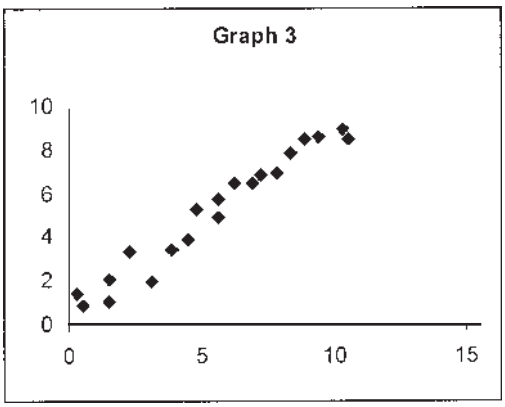
Regression analysis uses a technique called **least squares** to select a line which best fits the data. The best fitting line would be a line which passes closest to all the points on the scatter diagram. The easiest way to measure how close a line passes to a point is to measure the vertical distance between the line and the point. The least squares method finds a line which minimises the sum of the squared distances between the points and the line. There could be infinitely many lines which could be a best fit and so a computer is generally used to determine the line.

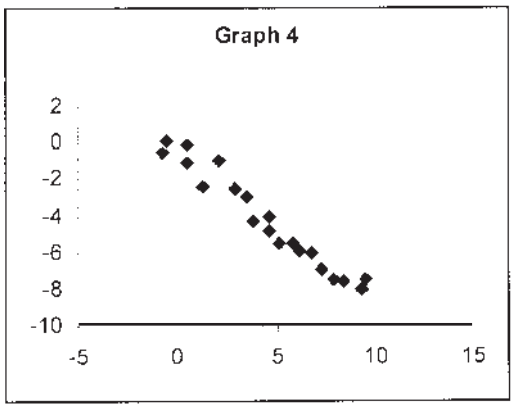
If the strength of the relationship between two continuous variables is required, then a **correlation coefficient** is calculated. The correlation coefficient determines how close to a straight line the relationship between the two variables is. A correlation coefficient can take values between –1 and +1. If the

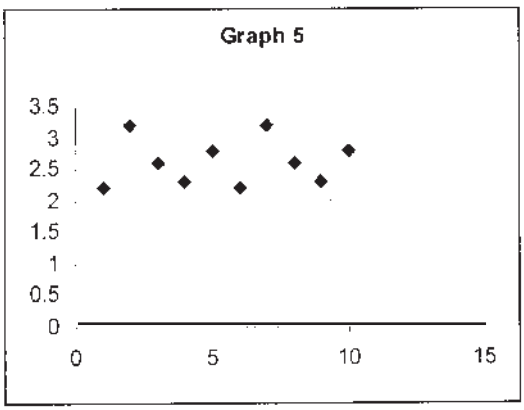
points lie exactly on a straight line, then the correlation will be 1. The graphs below show different types of relationships and the corresponding correlations.











In graph 1, we see a perfect relationship – as one variable increases, the other decreases. In the second, we also see a perfect relationship: as one variable increases, the other also increases. Graphs 3 and 4 show positive and negative correlation but not an exact relationship. This is because most of the points do not lie

on the straight line but some of the variation in one variable is explained by variation in the other. Graph 5 shows some relationship between the variables but is not a straight line and there is no correlation.

Even if we find a correlation between variables, this means only that they are associated with one another but not that one variable **causes** the other. Association is **not** causation. It is tempting to say when a relationship is found that one variable causes the other. Even if two variables are associated with each other, it is not necessarily the case that one variable is the cause and the other is the effect. If an association is found between the main variable of interest (called the **dependent variable**) and some individual explanatory variables, then the relationship between the explanatory variables should be examined. Any differences between the explanatory variables may help to explain the variation in the dependent variable.

For example, if tenants in area A are more satisfied than in area B, and women are more satisfied than men, we need to check whether there are more women in area A. Is it area or gender that is the main explanation for satisfaction?