Northumberland Knowledge

Know Guide

How to Analyse Data

- November 2012 -

This page has been left blank

About this guide

The Know Guides are a suite of documents that provide useful information about using data and information supplied via the Northumberland Knowledge website.

This guide outlines how to analyse data effectively. To make effective use of the data available on the Northumberland Knowledge website you may need to analyse it in order to obtain the information you require. This guide outlines some basic methods you can use to calculate the most commonly used statistical measures. It also provides some guidance on the best ways to display data and how to select samples.

For definitions of terms used see the Glossary of Terms Know Guide.

Different types of data

When analysing data it is helpful to know what type of information you are dealing with as this will allow you to decide what you are able to do with it and how best to present it. Data can be **quantitative**, that is it can be measured or counted; or it can be **qualitative**, i.e. it describes a characteristic such as colour, make of car or records attitudes or opinions.

Quantitative data can be further described as **discrete,** that is data based on counts, e.g. number of people, or it can be **continuous** – measured on a scale, e.g. temperature, weight, height.

Qualitative data can be **nominal** – data divided into categories to be counted, e.g. colours, where these categories are not ordered or weighted, or **ordinal** - data categorised into a sequence or order, for example, if people were asked to rate their opinion of a service on a scale of 1 to 5 with 1 being poor and 5 being excellent.

How to calculate common statistical measures

How to work out the mean

The mean (or the most commonly used type of average) is one of the most frequently used calculations when analysing data. To work out the mean add all the numbers in a dataset together and then divide the total by how many numbers there are in the dataset. For example:

If the ages of five friends are 26, 29, 31, 34 and 35, their mean age is (26+29+31+34+35)/5 = 31.

There are advantages and disadvantages to using the mean. Although it is a commonly used and easily understood calculation it can be affected by extreme values in the dataset.

How to find the mode

The mode is the number that occurs most often in a set of data. So for example, in the set of numbers 1,2,3,5,5,5,8,9,10,11, the number 5 is the mode. It is possible to have more than one mode in a dataset if there are two or more numbers that occur most often. To find the mode a frequency distribution is usually created to tally the number of times each value occurs in a set of data.

The advantages of the mode are that it is easy to understand and is not affected by extreme values. However, not all datasets have modes.

How to find the median

The median is the middle number in a set of data when the data is put in order by size. To find the median arrange the numbers by size order and find the middle number (if there is an odd number of values) or if there is an even number of values, find the middle two, add them together and divide by two. So, using the numbers in the previous example the median would be:

$$\frac{5+5}{2} = 5$$

The median is easy to understand and is not overly influenced by extreme values in the dataset.

How to work out the range

The range is the difference between the lowest and highest value in a set of data. To find it subtract the lowest value in the data from the highest, for example in the set of numbers above, subtract 1 from 11 to find the range which is 10.

The range is easy to understand and to work out, however it is affected by extreme values and while it shows the spread of the data, it does not show the shape.

How to work out quartiles and the inter-quartile range

A quartile is one of four equal groups that a dataset can be divided into. The lower quartile is the position below which 25% of the data is located while 75% of the data is below the upper quartile. The difference between the upper and lower quartiles is called the inter-quartile range. The inter-quartile range contains the middle 50% of the data.

For the set of numbers above the upper and lower quartiles fall between values. To calculate where they fall:

Position of lower quartile = ¼*(n+1) n = the number of values
= ¼*(10+1) = 2.75

Position of upper quartile = ¾*(n+1)
= ¾ *(10+1) = 8.25

The exact values of the lower and upper quartiles are calculated by adding the value of the lower number to the specified fraction of the difference between the two observations. Therefore:

Value of lower quartile = 2 + 0.75*(3-2) = 2.75

Value of upper quartile = 9 + 0.25*(10-9) = 9.25

So the inter-quartile range would be 9.25-2.75 = 6.5

Unlike the range, the inter-quartile range is not affected by extreme values. However, it is more difficult to calculate and does not use all of the data.

How to work out rates and percentages

Rates and percentages show a number as a proportion of another number. A percentage shows a number as a proportion of 100. It is calculated by dividing the numerator by the denominator and multiplying by 100. For example,

26/52 x 100 = 50%

Rates can be expressed as a proportion of 1,000, 10,000 etc and may be used, for example, to express birth rates, burglary rates.

To work out, for example, a burglary rate per 1,000 households, if there were 300 burglaries in an area of 15,000 households, the burglary rate would be:

300/15000 x 1000 = 20 burglaries per 1,000 households.

How to use confidence intervals

A confidence interval is used to represent uncertainty in data. It is a range of values in which it is likely that the true value of the characteristic being investigated (such as the mean) will occur. The most common interval used is the 95% confidence interval. This means that, across a whole dataset, the confidence interval is expected to contain the true values around 95 per cent of the time. For example, survey results indicate that 75 per cent ±5 per cent (at a 95 per cent confidence level) of library users are satisfied with their local library. Another way of putting this is that we are 95 per cent confident that between 70 and 80 per cent of library users are satisfied with their local library.

To work out a confidence interval it is necessary to know the sample mean $\bar{x}$, the size of the sample n and an estimate of the population standard deviation s.d. For a 95% confidence interval for the population mean the calculation would be:

$\bar{x} - \frac{1.96\ s.d.}{\sqrt{n}} \leq \mu \leq \bar{x} + \frac{1.96\ s.d.}{\sqrt{n}}$

$\frac{s.d.}{\sqrt{n}}$ = the standard error of the mean

To calculate a confidence interval of 90% replace 1.96 with 1.645 and for 99% 2.576.

How to work out standard deviation and variance

There are several measures of the dispersion of data. Variance and standard deviation are measures of the spread of the data around the mean. They indicate how the data is distributed, e.g. is it clustered around the mean or more widely dispersed.

The standard deviation is the square root of the variance. The variance is the average of the squared differences from the mean. The standard deviation and variance can be difficult to calculate manually.

To calculate the standard deviation, subtract the mean from each value in the dataset. This gives you a measure of the distance of each value from the mean. Square each of these differences and add all of the squares together. Divide the sum of the squares by the number of values in the data set. This is the variance. To find the standard deviation calculate the square root of this number.

The calculations are represented in the following ways:

(Population) Variance = $\sigma^2 = \frac{\sum(x - \mu)^2}{N}$ (sample) variance = $s^2 = \frac{\sum(x - \bar{x})^2}{n - 1}$

(Population) standard deviation = $\sigma = \sqrt{\sigma^2}$ (sample) standard deviation = $s = \sqrt{s^2}$

x = each value in the dataset
The mean is represented by the symbol μ for the mean of a population and $\bar{x}$ for a sample mean.
$\sum$ = sum
N = number of values
n = number of values in the sample

How to display data

Displaying data using a graph, chart or map can make the information easier to understand by summarising the important features and helping to identify trends and patterns. There are many different ways to display data and the method used should be selected according to the type of data to be displayed and the audience for whom the data is intended. Programs such as Excel allow charts to be easily produced from data.

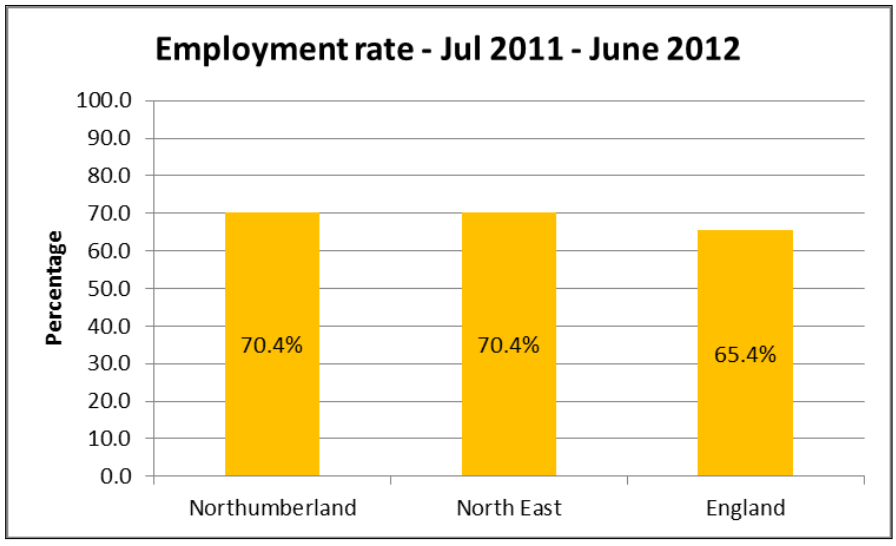
Charts and graphs should be clearly labelled and the data source should be acknowledged. The scale used should be appropriate and should not give a misleading impression of the data. Charts should be kept simple; don't use too many colours or special effects as this may confuse the message you are trying to convey.

Types of charts

Bar charts/histograms

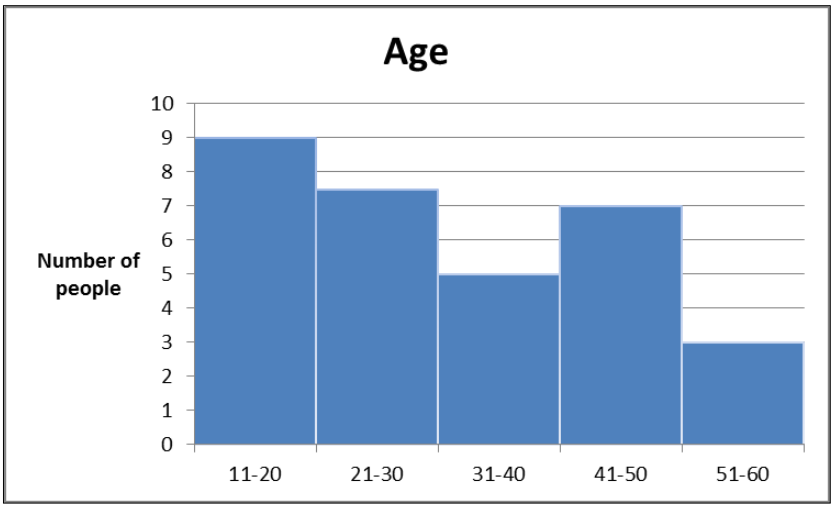
A bar chart can be used to display and shows relationships between data in terms of size. Bar charts can be useful to show discrete or qualitative data which is displayed using bars that do not touch, while a histogram shows continuous data and the bars do touch. Histograms are often used to illustrate the distribution of data.

Example – bar chart



Source: Annual Population Survey via NOMIS

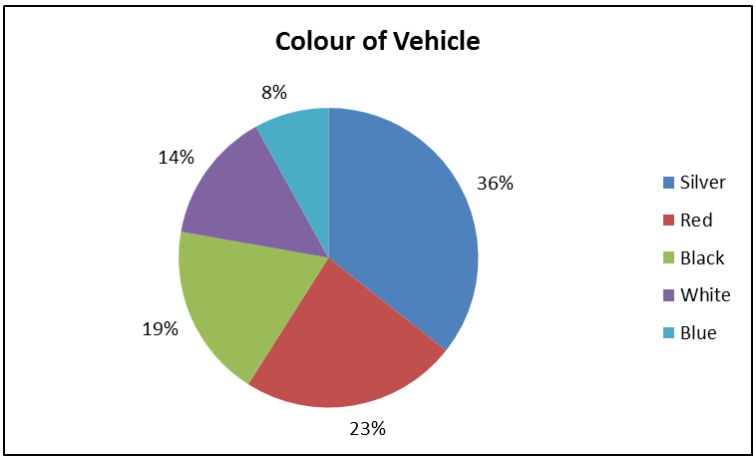
Example - histogram



Pie charts

Pie charts are useful for representing data where it is important to show individual values as a proportion of the whole population.

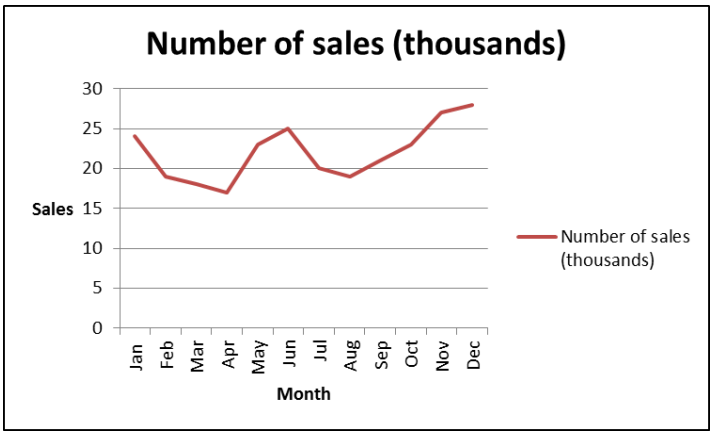
Example



Line graphs/run charts

These types of charts are useful for showing data representing change over time, for example, unemployment rates, as they can show trends and cycles in the data. They can also be used to compare data from different subsets of data over time, for example comparing Northumberland data with the North East or national figures.

Example



Tables

Where there is a lot of detailed data to convey a table may be the most appropriate way of displaying the information. As with charts, tables should be clearly labelled and sources acknowledged.

Example

Population Summary

Region	All People	Males		Females	
		Count	%	Count	%
Northumberland	316,028	154,124	48.8	161,904	51.2
North East	2,596,886	1,269,703	48.9	1,327,183	51.1
England	53,012,456	26,069,148	49.2	26,943,308	50.8

Source: ONS 2011 Census Population Estimates

How to select a sample

A sample is a subset of the overall population which is to be investigated. It may not be possible to collect information from every member of the population, therefore a sample is used to make estimates. However, to do this a sample must be representative of the population. To objectively choose a sample of a population it is useful to have a sampling strategy.

The most commonly used strategies are **simple random sampling** and **stratified sampling**. Both are types of probability sampling. A simple random sample is the most straightforward method of sampling. Every member of the population has the same chance of being included in the sample.

Simple random sampling

To carry out simple random sampling each member of the sampling frame (a list of all items from which the sample can be drawn) are assigned a random number, then ordered using the number. The required number for the sample is then selected. This type of sample is often the most suitable as it is possible to estimate how close the sample average is to the population average.

Each member of the population has the same chance of being selected and therefore the sample should be representative of the target population. Although this is a simple method of sampling it can result in an unbalanced, unrepresentative sample and may be expensive if the population is spread over a large geographic area and visits to sample members are planned.

Stratified sampling

A stratified sample can be used to improve the precision of a random sample. The population is split into separate layers or strata. These could be for example, age groups or gender. The strata must be different from each other and not overlap. A random sample is taken from each layer. A stratified sample ensures that data is collected from all key sub groups in the population.

Stratified samples can be proportionate or disproportionate. A proportionate sample is the simplest as the same sampling fraction is used to select from each stratum so that the sample from each is proportionate to the size of the stratum population. When selecting a disproportionate sample, a different proportion may be selected from some strata than from others. A disproportionate sample may be used when there is more variability in one stratum than another.

Sample size

The larger the sample size the more reliable the results will be, although the fraction of the population is not important. However, the sample should be large enough so that the sample sizes for the smallest sub groups can be said to accurately describe them. When deciding on a sample size you will need to consider how confident you need to be in the results. Most survey samples will be

calculated on a 95 per cent confidence level but some health surveys may use a 99 per cent confidence level and 90 per cent is usually the lowest confidence level that would be recommended.

It is not only the sample size that will affect the accuracy of results. A high response rate, using a high quality sampling frame and ensuring suitable survey methods are used will also increase accuracy.

Northumberland Knowledge

The Know Guide was produced by the Policy and Research Team, Northumberland County Council
knowledge@northumberland.gov.uk

November 2012