Office for National Statistics

wjec cbac

The Office for National Statistics guide to social and economic research.

Advanced **Skills Challenge Certificate** (Welsh Baccalaureate)

Acknowledgements

This handbook has been designed by the Office for National Statistics (ONS) to help students develop the social and economic research skills required for the individual project requirement of the Advanced Skills Challenge Certificate. The ONS is the National Statistics Institute for the UK and our core business is the collection, analysis and dissemination of data that informs decision making across government as well as used widely by academia, charities and businesses.

This guide was the brainchild of Fiona Dawe, a Social Researcher and Head of the Economic Statistics External Engagement and Capability Team. She wanted to share the social and economic research knowledge and expertise that ONS has to help students and teachers alike feel more confident in their statistical and research skills. By working closely with Caroline Morgan and Emma Vincent at the WJEC we have produced a comprehensive resource, to guide you through conducting your own independent research project. Not only do we explain how to conduct and analyse your own research but we also explain the reasons behind the different methods and techniques to give you a greater understanding of how they can be best used in your research project.

All of the content in this guidance has been written by staff based at ONS Headquarters in Newport, South Wales. This work was undertaken as a voluntary project on top of their business as usual responsibilities. The contributors are all members of one of the civil service analytical professions:

Government Social Research; Government Economic Service; Government Statistical Service; and Government Operational Research. They work in different areas of ONS including external engagement; sector and financial accounts; well-being; public policy analysis; digital services and technology. We would like to say a huge thank you to Amy Brownbill, Philip Leake, Oliver Woodcock, and Rhian Jones for helping to produce and refine the content for this handbook.

Throughout the process of creating this guidance we have sought advice from local schools. We would like to thank Caerleon Comprehensive, St Teilos Church in Wales and Bassaleg schools for their expertise and recommendations.

Finally, we would like to extend a huge thank you Andrew Budd and Grace Ellins, of the ONS Design Team, who found time in their busy workload to turn our draft into this online document.

Charlotte Deeley (Editor)
Georgia Tasker-Davies (Project Manager)

Contents

Introduction

01

About This Handbook

How to Use the Handbook

The purpose of the individual project is to enable students like you, to develop a variety of skills through carrying out an independent research activity. The area you choose to research will be entirely up to you, but this is an excellent opportunity to undertake a piece of work with a focus on your future educational or career aspirations. As we mention later, if you are planning on applying to university, an independent project is a great addition to any personal statement.

In this book, we will guide you through the process of conducting your own research project. This will include advice on the best data collection methods for the type of research you want to carry out, the appropriate type of data analysis to carry out to answer your research question, how to interpret this data, and finally how to present your data to enable your reader to understand your findings.

Using examples of social and economic research carried out by the Office for National Statistics (ONS) we will explain how to conduct research that will allow you to produce quality statistics, and what you should look for when assessing the quality and reliability of data sets.

What is Social Research?

According to the Economic and Social Research Council, social research can be defined in a very broad sense as research into society, the way people behave, and what influence this has on the world around us.

By undertaking social research, we can learn about the world outside our own experience and start to understand the way in which our society works.

The results from well conducted social research can be used by governments to help create policies, local authorities to find out what is important to their constituents, and non-government organisations to help them target those most in need.

Social research covers a range of disciplines, encompassing both qualitative and quantitative methods. Social researchers will implement a range of techniques to collect, analyse, and interpret data. This handbook will gently guide you through how to conduct your own high-quality research. If some of the terms we have used so far aren't familiar to you, don't worry. We will guide you through all of this too!

Any terms you see in **bold** will be defined in our glossary at the back of this guide.

How Research Methods Can Help You in the Future

Social research methods give you the skills to plan, design, conduct, manage and report on a research project. This handbook will describe and explain the methods required to effectively conduct and present your own research.

Having the skills to understand social research will put you in a great position for a wide range of different career paths. By completing the individual project, you will have the opportunity to learn and apply research skills to a topic you are interested in. This will help prepare you for any research you may be required to do in your future studies, career, or even personal life.

If you are thinking of applying to university, an independent research project such as this will look great on your personal statement, and is a great topic for discussion at interviews.

The ONS conducts a lot of social research which is used by the UK government to inform a wide range of policies. A few examples of social research the government relies on to inform their decisions can be seen below;

Child health – Measures children's health and well-being including childhood, infant and perinatal mortality; Unexplained deaths in infancy; Childhood cancer survival; Children's well-being.

Drug use, alcohol and smoking – Measures smoking and drinking habits in Great Britain; Deaths related to drug poisoning and drug misuse; Deaths caused by diseases known to be related to alcohol consumption.

Education and childcare – Measures early years childcare; School and college education: Higher education and adult learning – including qualifications, personnel and safety and well-being.

Health and well-being – Analyses of social and economic data from government and other organisations to paint a picture of UK society and how it changes, including comparisons with other countries.

This list is by no means exhaustive (we could write an entire book on the social research the government utilizes), but you can see that social research spans across a lot of different subject areas. Even though most people don't realise, social research is impacting on all of our lives through the policies which guide our society.

About ONS

What is ONS?

ONS stands for the Office for National Statistics. We are the UK's largest producer of Official Statistics. Official statistics are statistics which are published by government agencies and public bodies, which can be used as a public good; In other words, they are freely available to all (including you!).

The ONS is a civil service department, that is independent from government. This means that it operates as an "arms-length body" from government, so is not headed by a Minister. Instead the head of ONS (the National Statistician) is answerable directly to parliament.

Responsible for collecting and publishing statistics on a huge variety of topics, reflecting many aspects of people's lives. This includes statistics relating to our economy, population and society at national, regional and local levels. ONS statistics can tell you an awful lot about life in the UK.

Almost every country has their own National Statistical Body, which follow strict guidelines at every stage of how statistics are produced. Because of this we can compare information between countries and know we are talking about the same thing. This is very important when drawing conclusions based upon data.

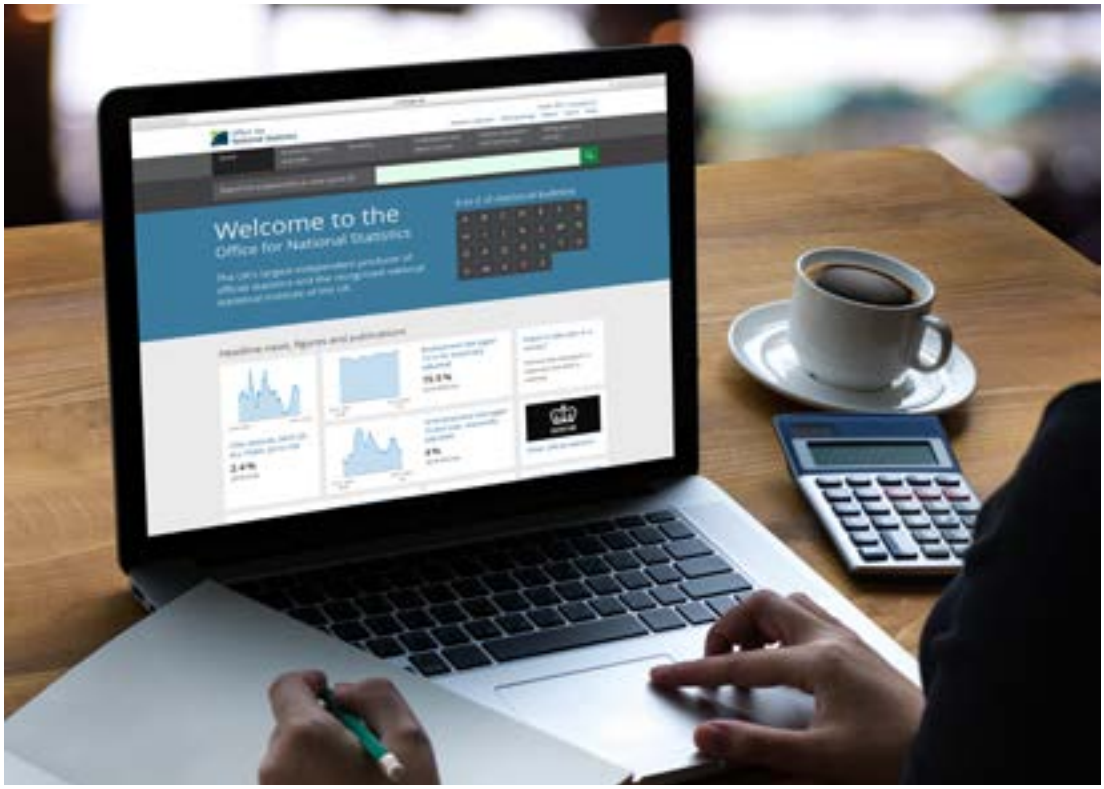
ONS statistics are regularly in the news. You have probably seen many news stories with ONS statistics as their source and not realised – keep an eye out and see how many you can spot this week.

How Can ONS Help You With Your Project?

ONS is a great starting point for any research project. Whether you want to investigate crime, culture, education, employment, health, social care, migration, well-being, or financial issues (and this is only a few examples!), ONS has data which can help direct you on your own independent research journey.

It's worth visiting to find out for yourself what we have to offer.

On our homepage, you can find headline news, figures, and publications which could inspire you. Alternatively, if you already have a research topic in mind, use the tabs at the top of the page, or our search function to find information relating to your area of interest.



Source: Office for National Statistics

It's wide variety of statistics make the ONS an excellent secondary data source. You can find out more about primary and secondary data in our dedicated sections on pages 22 and 24.

Here are just a few examples of questions you can answer with ONS statistics…

- What is the average life expectancy in my local area?
- What are the drinking habits of adults in Great Britain?
- How many people die from drug use/abuse in England and Wales?
- What influences people's vaping habits?
- How have the odds of surviving childhood cancer changed over the last 25 years?

We encourage you to have a look on the ONS website and see what interesting questions you can come up with!

If you require further support regarding social and economic research that is not covered in this handbook, you can contact welshbac.support@ons.gov.uk

Potential Careers at ONS

Why Join ONS?

ONS is a stimulating place to work. We are the UK's largest independent producer of official statistics and the recognised National Statistical Institute. Our statistics are in the media just about every day and used across government. Not only will you be making an impact on life in the UK, all our offices have a commitment to staff wellbeing and work life balance. As well as flexible working hours, ONS are soon to roll out regular wellbeing sessions across the organisation. Where this has been piloted, sessions have included having cuddles with guide dogs, yoga, meditation, a variety of exercise classes, and even mindful colouring!

ONS Offers Graduate Opportunities in Four Government Analytical Professions

- Government Economic Service
- Government Statistical Service
- Government Social Research
- Government Operational Research

Once you join one of the Government professions, there is the potential for your career to take you across government departments.

Economist – Government Economic Service (GES)

ONS is an exciting place for an economist to work, with the opportunity to apply economic theory to a range of practical issues. Our economists are involved in a range of work streams around the production, development and dissemination of UK economic statistics. You will be involved in stimulating and rewarding research and analysis that ensures our statistics are world class.

Statistician – Government Statistical Service (GSS)

From the economic future of the country to the provision of public services, statistics are vital to every imaginable area of our lives. Government play a pivotal role in government decision making. At ONS, you will join a team of data experts who collect, process and analyse the data underpinning our society and economy. You can expect to see your analysis hitting the headlines on a regular basis and being used by decision-makers in government and industry at the highest level.

Social Researcher – Government Social Research (GSR)

At ONS, you will work on a range of research, statistics and analysis projects primarily for ONS, but on occasions on behalf of other government departments. ONS uses a range of primary social research methods including large scale surveys, continuous and ad-hoc surveys and qualitative/quantitative methodologies to measure and understand society and the economy. We also undertake secondary data analysis and research of census, survey and administrative data.

Operational Researcher – Government Operational Research (GORS)

Government Operational Researchers support policy-making, strategy and operations. Operational Researchers look objectively at the complex problems departments face and apply a range of analytical and modelling techniques to not only help find better solutions to pressing issues but to also determine their potential impact before they're delivered. Better solutions, of course, mean better decisions and better policies.

Student Placement Schemes

Every year each Analytical profession offers the chance for undergraduate students to work at the heart of Government through a year-long paid placement. This is an exciting opportunity to gain professional and technical skills whilst working on some of the most interesting and pertinent problems currently facing the country. These placements start in summertime and are open to second year students looking to complete a placement before the start of their final year at university.

The professions also offer three-month summer placement opportunities. For more information on sandwich, summer and graduate opportunities in ONS and other government departments, access the links below:

- Economics
- Statistics
- Social Research
- Operational Research

In addition to the graduate roles described above, ONS have a variety of roles for people straight out of school too. Have a look at civil service jobs to see what is available in your area!

Conducting Your Own Research

A Brief Intro to Ethics

What are ethics? Ethics can be described as moral principles that govern a person's behaviour and actions. Ethics is concerned with what is good for both individuals and society as a whole. To be a good person we are expected to act morally, and as a researcher, we should also strive to conduct research morally.

Why does it matter? In the past, people have done horrible things in the name of research. Some of the most famous examples of this are the Nazi Human experiments. During World War 2, prisoners and the public were forcibly experimented on. They were not able to give **consent**. There was no concern for participants; many died or were horrifically injured in the process.

Establishing a code of ethics – After the war, the Nazi doctors who conducted the experiments stood trial to answer for their crimes in Nuremberg, Germany. The Nuremberg Code was established in 1947 following the trials.

This code is a set of research principles for ethical human research. It includes considerations such as **informed consent**, scientific validity (of researcher and the study), and avoidance of risks including pain and suffering.

Different research bodies will have different specific codes of conduct. However, according to the Economic and Social Research Council there are six key principles for ethical research:

- research should aim to maximise benefit for individuals and society and minimise risk and harm
- the rights and dignity of individuals and groups should be respected
- wherever possible, participation should be voluntary and appropriately informed
- research should be conducted with integrity and transparency
- lines of responsibility and accountability should be clearly defined

- independence of research should be maintained and where conflicts of interest cannot be avoided they should be made explicit

ONS adheres to the UK Statistics Authority Code of Practice. The Code provides producers of official statistics with the detailed practices they must commit to when producing and releasing official statistics.

The Code ensures that the statistics published by government serve the public. When producers of official statistics comply with the Code, it gives users of statistics and citizens confidence that published government statistics are of public value, are high quality and are produced by people and organisations that are worthy of trust. You can read more about the code here.

Wherever you see this logo, you can be sure that data is of good quality.

One of the main things that the code serves to do in terms of ethics is to reduce the burden on respondents. This means that if we have other means of collecting the data, we should make use of it. Otherwise we could be asking people to give us information they already provide elsewhere. A good example of this is ONS's recent shift to using administrative data – this means people spend less time responding to questionnaires, as we can get the information from other government departments who already collect this information.

Most organisations have their own codes governing research they conduct. For example, the NHS have the NHS Health Research Authority (HRA).

The HRA have several functions, including protecting the rights, safety, dignity and well-being of research participants; and facilitating and promoting ethical research that is of potential benefit to participants, science and society.

All ethics boards, no matter what the research area, strive to ensure that participants are protected and that research adds value to the area of investigation.

Reading Scientific Articles

The following guide is specific to journal articles; however, the principles can be applied to anything you read for research.

In order to come up with a good research project, you will have to do a lot of reading. One skill that will be of use throughout university or your future career, is how to read papers quickly and efficiently. Here are some hints and tips on how to do this.

Most journal articles will have what is called an **abstract** at the beginning. This is a quick summary of the information contained in the report. This is usually a brief version of the aims, what was done, what was found, and conclusions. By reading the abstract, you can quickly decide whether a paper is relevant to what you are trying to research, and whether or not you should read it.

(!) Do not just read the abstract if you are going to reference a paper.

There could be important information embedded in the body of the report which needs to be looked at in closer detail. The abstract is only usually around 250 words, this is not enough for all the important information to be in there.

Helpfully, reports are divided up into different sections (see 'How to Structure a Report' – page 14 for in depth look at the different sections). This means that if you know what you are looking for, you won't have to read a whole paper.

Think about what information you want, and which section of the report this would be in. You can then just read those sections!

For example: You will usually want to know what the **aim/ hypothesis** of a paper is – this tends to be found towards the end of the introduction.

If you are looking for what the researcher found, you could either look in the results section – this will likely be numerical/clinical findings, or in the discussion – the results will be put in context and discussed in the context of the paper.

The discussion will also be where you can find the **conclusions**, and what the researchers think their conclusions means in terms of the wider area of research.

If you want to know exactly how an author conducted their study, either to find a method to replicate, or just out of interest, this can be found in the methods section.

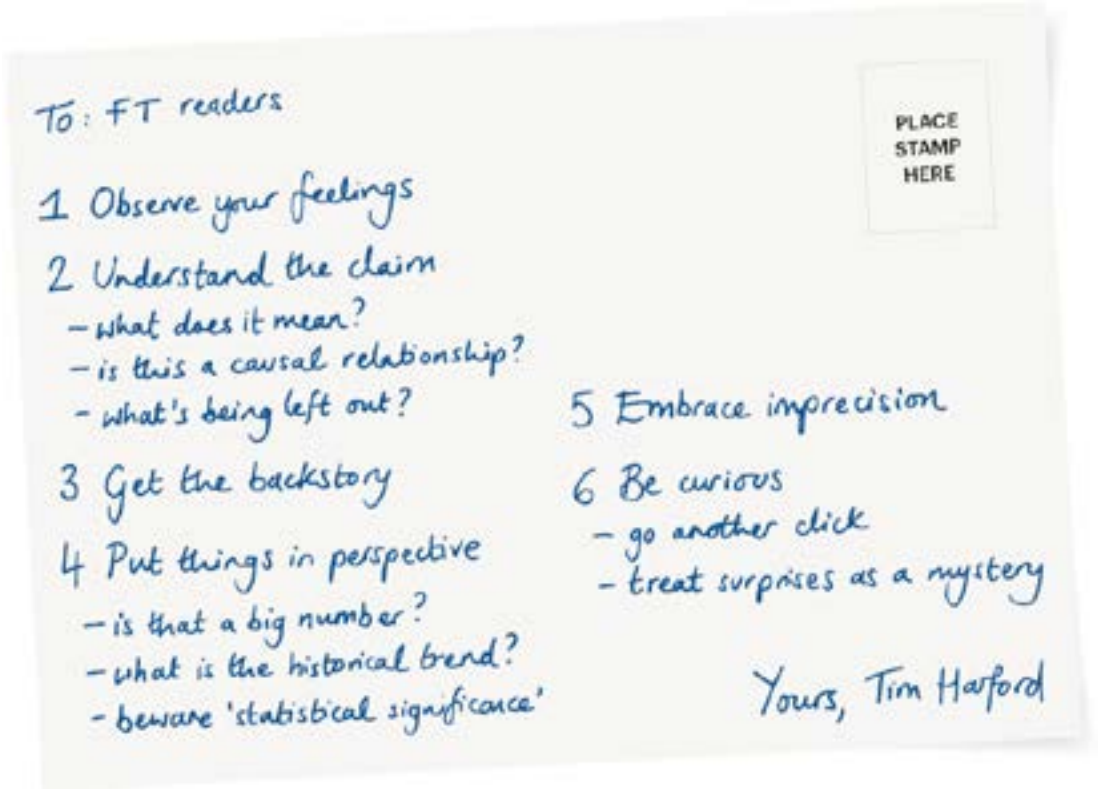
An important thing to consider when reading a paper is how does the data look? And importantly, do you really believe the findings?

If you don't think the quality of a report is very good – maybe the author misses out important information, or makes conclusions they can't back up with the data they produced – it might be a good idea to leave it and try and find a better-quality paper to reference.

As always, practice makes perfect. The more you read journal articles, the easier you will find it to find the information you are looking for, and judge whether a paper is high quality.

Top Tips from an Economist

Tim Harford (an economist) gave some top tips to readers which are easy to remember when you are looking at data and research.



How to Structure a Report

Being asked to write an independent project can be daunting if you have never done a piece of work like this before. When you have a big word count to cover, you can easily ramble and get away from the point you are trying to make. By sticking to a defined structure, you can clearly and precisely present your work to get the highest marks. Below we will outline a structure for how most scientific papers and journal articles are written. Becoming familiar with this structure now could help you in any future higher education or work.

Abstract – Very often when writing a scientific piece of work, you will be required to write an abstract at the beginning. This is a quick summary of the information contained in the report. This usually contains a brief version of the aims, what was done, what was found, and conclusions. By referring to an abstract, a reader should be able to understand your research in very general terms. Bear in mind, that an abstract is only ever around 250 words! This can be tricky to get the hang of, so it may be worth having someone who isn't familiar with what you are doing read it through to see if they can tell what your research project is investigating.

TOP TIP

Write this section last!

You will then have all the information required to write an informative abstract.

Introduction – This one is pretty much common sense! You are introducing your reader to what your report is all about, and letting them know why they should continue reading. You don't have to write this section first – it can be easier to write your introduction having already written up your later sections. For example, the method can change in the process of a study. It is also helpful to have the results written up too, as the actual results you get can be very different from what you initially expect. By writing your introduction last, you can shape your writing so that it flows well with the rest of the report.

An introduction usually starts quite generally, with broad context for your research, getting more specific to your individual project as you go through (a bit like a funnel). Ideally in your introduction you want to demonstrate an understanding of the theoretical background and context for your project. This way you can justify the research you are undertaking, and explain how it fits in with other work available on the subject. You need to clearly set out your aims and objectives. From this you then lead into your particular question, what you will do, and what you expect to find.

Rational for Research Methods – This is where you describe exactly what you did, the materials you used, participants and anything else you can think of that would help someone replicate your study exactly. This can be further broken down into subsections to keep you on track. This could be participants, materials, and procedure. This can be tricky when you first start, as it is difficult to distinguish what is relevant.

Let's say you want your participants to complete a questionnaire for your project. What is important to include in your methods section for someone to be able to replicate your method exactly?

Results – This section is simply to give the reader a description of your results. You do not need to explain them at this point. You can use figures, tables, anything that is appropriate to best demonstrate what your results are showing. You can draw the reader's attention to particular patterns or trends your results follow, or data that you think is particularly important. For more detail please see the data presentation chapter.

Discussion – A discussion has a couple of aims. These are to explain the results of your study, and look at what these results mean in the context of the research area you have been investigating. Here you will need to interpret and explain your results, look at whether the question you raised in your introduction has been answered, show how the results you've produced relate to existing literature, look at the significance of your results, and discuss areas for future research that you think could be useful considering what you have found.

Referencing/Appendices – Throughout your report you will need to provide references when you have included an idea or theory that is not your own. In your report, you will need to put the name of the author from which you have borrowed the information, and the date of the publication. This can either be in the text, in brackets, or using footnotes. There are many referencing styles. You will need to check with your teacher which one they want you to follow, and use the recommended guidelines for this system. At ONS they mostly use the footnote system.

At the end of the report you will need a reference list. This is a list of all the references you have in the main body of your report, organised in alphabetical order. The format for this will also follow specific rules laid out by the referencing system you are using.

An appendix contains any information which is not 100% necessary to explain your findings, but could be a useful tool for your reader. As mentioned before, if you have created a novel questionnaire, this could be included here, so that the reader can see exactly what was being asked. Information to be included in an appendix could also include figures, tables, charts, graphs of results, transcripts of interviews, pictures, maps, drawings, or anything else that would be required for replication of your study that is not otherwise available. Appendices are lettered, and are referenced in the text as such. For example – for the full questionnaire, please see Appendix A.

Guide to Writing a Report

Number one tip for writing a good report is before you start, do your homework!

What research has gone before? What do you believe on the topic? What do you want to say? How will you do this?

When you are writing a report for the first time (or the first couple of times) it is a good idea to keep in mind the following points. The purpose of a report is to describe your study, tell the reader what you did, what you found, and how this relates to existing theories and literature.

To do this you will need to:

- explain what your question was, and why it's important

- describe exactly what you did, and the sample you used (the more detail the better, as this will make it easier for your study to be replicated)
- What did you find? Appropriate summary statistics reported
- What might this mean? Use the context of previous literature and the contribution your findings make to this
- limitations, there will always be some. These could be a different sample or sampling method you would have used if you could, a different experimental methodology or improvements you could have made to your methodology in hindsight
- future directions. Based on your findings, what would you do next/recommend someone else could do to follow on from your project?

Your hypotheses should be expected and previously justified. In other words, don't just make up your hypotheses out of thin air, they should be based on existing theories and literature.

TOP TIPS

- Be simple, be concise and work out your angle/story.
- Only write what you've got – don't over interpret your findings.
- Think about your reader (especially their boredom levels).
- Back up statements with references (or note as opinion) – but don't just cite everything you read. Each point needs to be considered and add to your overall report.
- Writing style – always past tense with a passive voice (focus on what was done rather than who did it).

Qualitative vs Quantitative Research

Simply put, the difference between qualitative and quantitative research is that quantitative research will result in numerical data, while qualitative research will give you non-numerical data.

Quantitative Data

Quantitative data tends to stem from closed questions, meaning when you ask a question you expect a specific piece of information back between a certain range.

EXAMPLE

Quantitative data on earnings

ONS have an annual survey of hours and earnings. This is an annual survey which "provides information about the levels, distribution and make-up of earnings and hours paid for employees by sex and full-time and part-time working."

Source: Annual Survey of Hours and Earnings, Low Pay and Annual Survey of Hours and Earnings Pension Results, Office for National Statistics

This is quantitative data as the survey is asking the recipient for specific pieces of 'numerical' information back, such as the number of hours worked, and earnings.

The results of this survey can be found on the Nomis website (a service provided by the ONS for free access on up to date labour statistics).

Quantitative data can be **aggregated** to high levels, meaning that data from a large amount of people can be summarised with a few statistics.

Qualitative Data

Qualitative data is data which can't be measured numerically/ quantified. For example, the colour of your hair can be recorded but not measured.

Qualitative data usually comes from more open questions like "what are your opinions on…?"

You can collect qualitative data to complement your quantitative data. It may be that you want to delve down and get more of an understanding behind the quantitative data. You can conduct qualitative research to learn the reasoning behind people's answers.

There is not Always a 'Best' Data Type

Both quantitative and qualitative data can be used in conjunction with each other; it is good practice to find some reliable secondary quantitative data (more details on page 22) due to you probably not having the resources to conduct statistically sound quantitative research yourself. For this you need to have collected data from a very large group of people. There is lots of data available through ONS website. Then you can conduct your own qualitative research, aiming to discover the underlying reasoning behind the quantitative data.

LEARNING OBJECTIVES

By the end of this chapter students should feel able to:

- ✓ Describe what social research is.
- ✓ Explain what ONS stands for and what they do.
- ✓ Understand how the principles of ethics apply to their research project.
- ✓ Describe and explain the difference between quantitative and qualitative data.

Data Collection

02

There are two main methods of collecting data: primary and secondary research.

Explained simply, primary research involves collecting your own data, while secondary research involves gathering data from existing sources. We'll explore both in this chapter.

Primary Research

Primary research, sometimes called 'field research', involves gathering and using new data that has not been collected before. For example, surveys using questionnaires or interviews with groups of people in a focus group.

Sources of Primary Data

The source of your primary data is the **population sample** from which you collect the data. The first step in the process is determining your target population and the type of sample you wish to use. The type of sample you use will depend on the population you are targeting. Go to the section on sampling to find out more.

Collection of Primary Data

There are many methods of collecting primary data (observed or collected directly from first-hand experience) with a few examples explained below:

Questionnaires

Questionnaires are a popular means of collecting data, but are difficult to design and often require many rewrites before an acceptable questionnaire is produced.

The ONS use questionnaires for a number of different surveys including the family expenditure survey, labour force survey and the census.

Examples of census questionnaires can be found on this website. By using existing high quality questionnaires, like those produced by ONS, you can be sure that the questions you are asked have been validated. There is also the added advantage of having existing results you can compare your answers to. If you are collecting your own data it still counts as primary research, even if the questionnaire is not your own!

TOP TIP

Don't just use yes/no questions…

They can't really tell you that much, at advanced level you should be looking to explore questions in more depth.

Interviews

Interviewing is a technique that is primarily used to gain an understanding of the underlying reasons and motivations for people's attitudes, preferences or behaviour. Interviews can be undertaken on a personal one-to-one basis or in a group.

Which is Best?

Table 1: Advantages and Disadvantages of Interviews and Questionnaires

	Advantages	**Disadvantages**
Questionnaires	Can be used as a method or as a basis for interviewing or a telephone survey Can be posted, emailed or sent via text Can cover many people Wide geographic coverage Inexpensive	Questions should be relatively simple Often have a low response rate No control over who completes it Problems with incomplete questionnaires Time delay while waiting for responses to be returned
Interviews	More likely to receive accurate information Good response rate Possible to follow-up on questions	Need to set up interviews which could be time consuming. Take longer to conduct Geographic limitations Can be costly

Secondary Research

Secondary research involves gathering existing data that has already been produced. For example, researching the internet, newspapers, journal articles, and company reports.

TOP TIP

Many people actually start by conducting secondary research, and then use this to inform their primary research. The Office of National Statistics (ONS) website is a good place to start.

Sources of Secondary Data

There are a variety of secondary data sources which can be found online. However, you must be careful where you obtain your data from. Most government departments will give you reliable data that you can use in your analysis.

The ONS as a Data Source

The ONS produces a range of data on a variety of topics. We've included a couple of examples below, but please have a look on the website for data on your subject of choice.

The Labour Force Survey (LFS)

The LFS covers all aspects of people's work, including the education and training needed to equip them for work, the jobs themselves, job-search for those out of work and income from work and benefits. LFS quarterly datasets are provided to government departments, approved researchers and the public.

The sample is made up of approximately 40,000 responding UK households and 100,000 individuals per quarter. The LFS is intended to be representative of the entire population of the UK.

Data from the LFS is available freely on the ONS website.

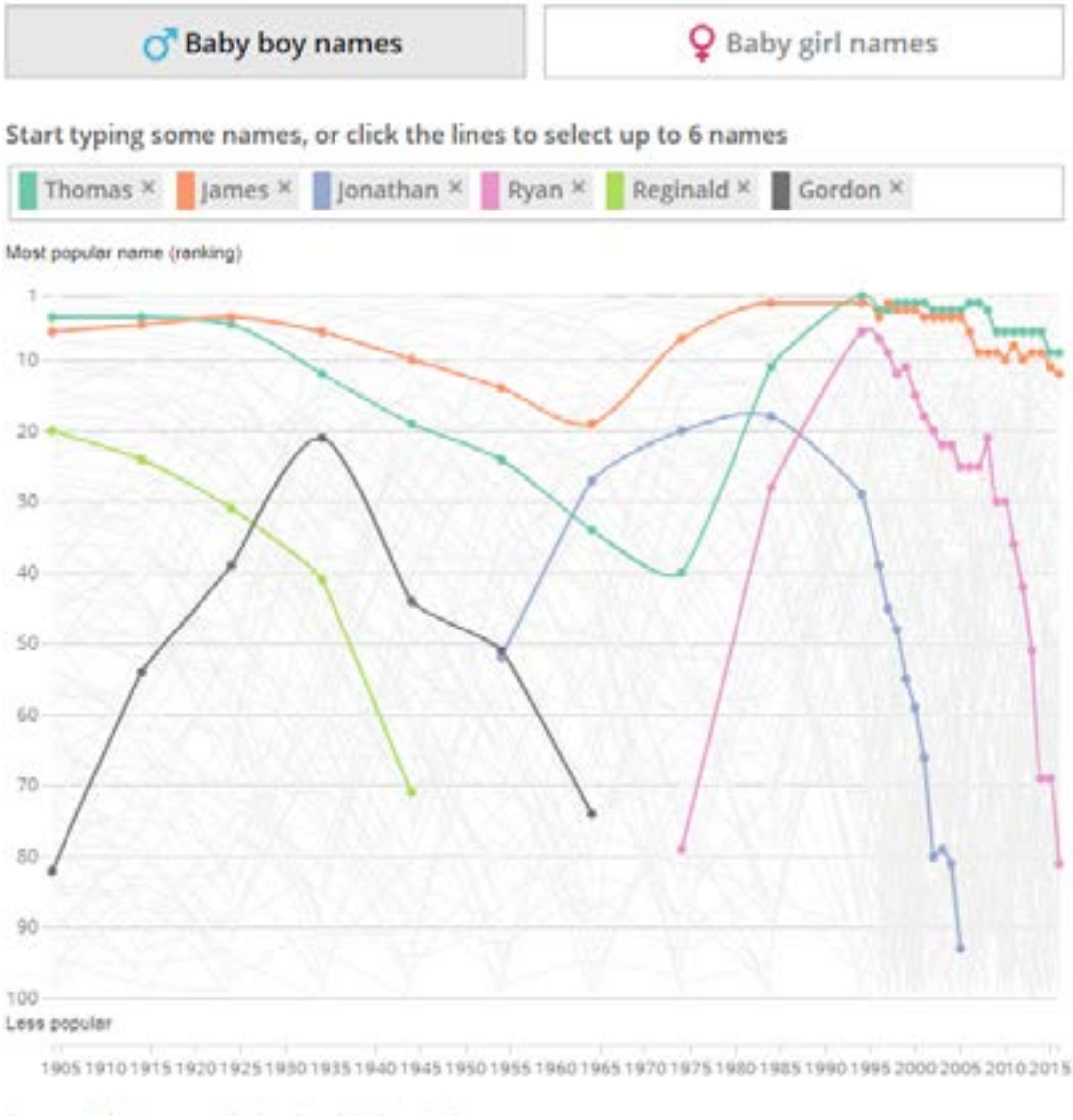
Examples of how the ONS uses LFS data are;

- estimating employment
- estimating unemployment
- estimating economic inactivity

Baby Names

One of the most popular datasets produced by the ONS is baby names for both girls and boys.

Figure 1: Trends in Male Baby Names 1904-2016



Source: [Baby Names – England and Wales](), **Office for National Statistics**

Baby name statistics come from final annual births registration data and represent all live births occurring in England and Wales in a specific calendar year (including a very small number of late registrations).

Statistics for baby names for both boys and girls are available free on the ONS website.

Finding Data Sources Through Journal Papers and News Articles

As well as looking on government department websites, you can also find data sources through research papers or news articles.

For example, say you were interested in finding out how much people earn depending on what university they graduated from, or what subject they studied. You could firstly see what research has been done on this already by looking on the internet. You find that the BBC has written an article on this subject.

The degrees that make you rich... and the ones that don't

By Dr Jack Britton
Institute for Fiscal Studies

17 November 2017 | 334 Share

GETTY IMAGES

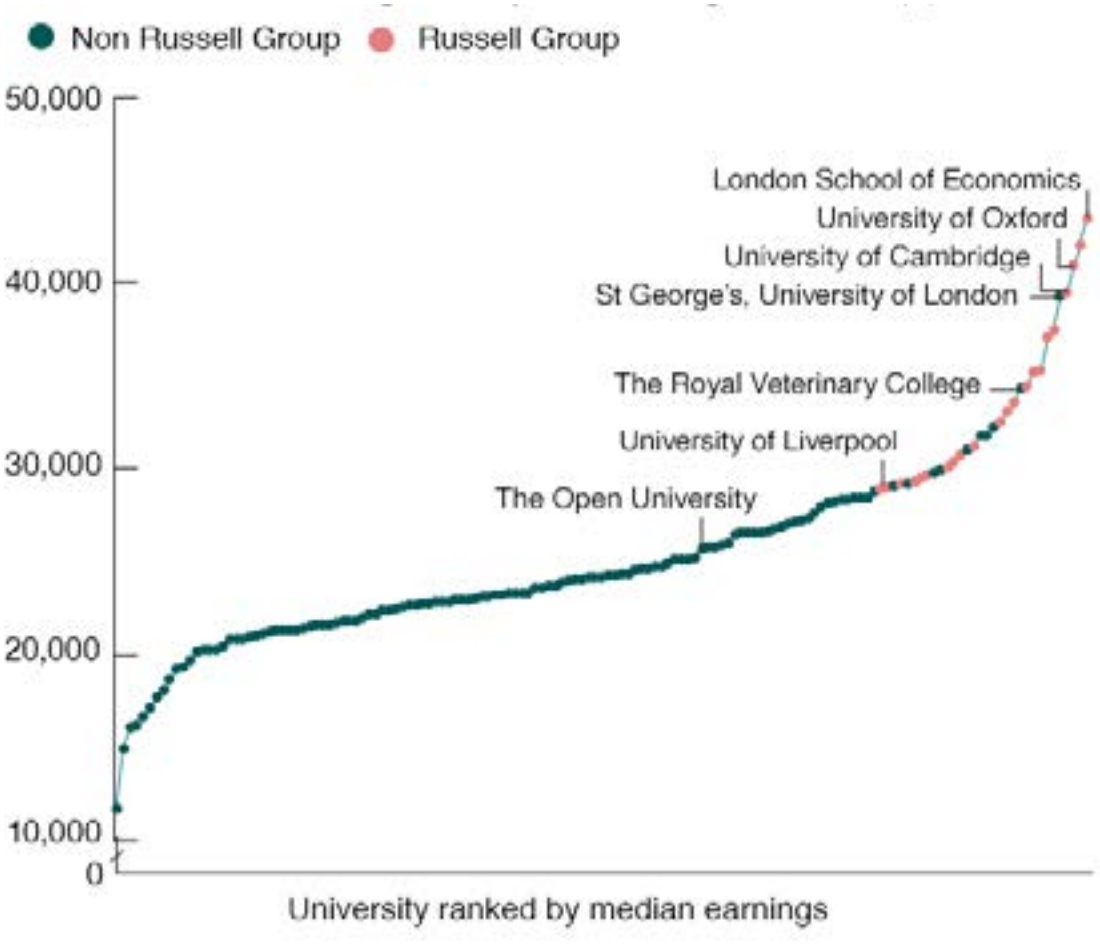
Hundreds of thousands of young people are in the process of applying to university, in time for a 2018 start. Their choices can make a huge difference to future earnings.

From looking at the article we can see that the data behind it came from the 'Institute for Fiscal Studies'.

We can now go onto the Institute for Fiscal Studies website to find the data behind this article, which may be helpful for our own research.

Figure 2: Graduate Earnings by University

Median annual earnings five years after graduation (£)



Source: Institute for Fiscal Studies

As well as news articles, existing data sources can be found by reading research journals.

Another example, looking at this journal article 'Impacts of Family Background on Educational Attainments' written by John Ermisch and Marco Francesconi, you can see the data they used by just reading the abstract:

> "The analysis uses new data matching parents and their young adult children to study the impact of family background on young people's educational attainments. The data is derived from the first seven years (1991–97) of the British Household Panel Study. Parents' educational attainments are found to be very strongly associated with their children's educational attainments, and for an important part of the population these associations can be given a causal interpretation. In addition, young adults who experience single parenthood as children and those who come from families in the bottom income quartile have significantly lower educational attainments."

Should You Use Primary or Secondary Research?

It depends on what you're trying to achieve. In a perfect world, you'd use both! Have a look at the pros and cons of each in table 2.

Table 2: Advantages and Disadvantages of Primary and Secondary Research

	Positives	**Negatives**
Primary research	You can tailor your data collection to your research question	Can be time consuming Might not have the resources to collect a large sample of data
Secondary research	Easily available Saves costs to the researcher Will provide essential background and help to clarify or refine research problem	Because this data has already been collected it may not be the most relevant data for the research question Potentially unreliable (depending on the source)

Sampling

We use sampling when we want to find something out about a population, and collecting data for the whole population isn't feasible. Instead we select a sub-group, called a "sample", to represent the population.

If you're conducting your own primary research for your project, or drawing on existing sources through secondary research, you will need to know about sampling.

What is a Sample?

A sample is a segment of the research population which can be any size smaller than the total population. Data collected from a sample of the research population is then used to represent the whole population.

What is the Research Population?

The "population" is all of the subjects that we are interested in finding out about. The population will vary depending on your research needs:

- if the research is investigating an issue that is relevant to everyone in the UK – for example, "what is the average hours a week spent watching TV?", the research population would be everyone who lives in the UK (and has a TV)
- if your research question was more specific, for example: "what is the average height of women in the UK?", the research population would be all women in the UK

Why Would You Use a Sample?

Often it is not possible to collect information from everyone in the research population because:

- the population is very large
- it would cost a lot of money to collect information from everyone
- it would take too much time to collect information from everyone
- managing the data collected from everyone would be very difficult – it would take a lot of time to enter, clean and analyse the data

EXAMPLE

What is the average height of women in the UK?

It would take a lot of time and cost a lot of money to observe the heights of all of the women in the UK (the research population). Instead we can take a sample of the research population and then use this to get an accurate estimation of average height.

What Does ONS Do?

ONS very rarely collects information from everyone in the research population because it is too expensive to do this. Nearly all ONS surveys are carried out using a sample that has been selected to be representative of the research population. Information about the sample and how we selected the people or business can be found in the methodology information that we provide alongside our statistical releases on the ONS website.

The exception is the ten-yearly **census**, which is a count of the entire population.

The census is an attempt to count all people and households in the UK. It is conducted every 10 years, and it is the only survey which provides a detailed picture of the entire population. It covers everyone at the same time, and asks the same core questions. This means that you can use this data to compare different parts of the country or different groups of people. For a research project, using data from the census is a great source.

Most ONS survey samples are random probability samples – this means that they are designed to enable us to know how accurate our statistics are and to understand the limitations and bias in the sample design.

If you want the results from your survey to be statistically robust, you'll need to use a form of random probability sampling. This means you will be able to make precise statements about your data in relation to the total research population.

Simple Random Sampling

A random sample gives the researcher an unbiased representation of the population. A simple random sample means every member of the population has an equal chance of being selected for the sample, which is important for being able to come to conclusions about the total population.

A classic example of a simple random sample is drawing a name from a hat – everyone has the same chance of their name being drawn from the hat. Another way you could perform a simple random sample is by assigning a number to every member of the population and then getting a computer to randomly generate as many numbers as you want for your sample, from the population range. Then the corresponding population members will form your sample.

There is a risk that this sample won't capture the true views of the population. To get accurate estimations of the population, we need the sample to be representative of the total population.

EXAMPLE

A class of 14 students, half male, half female are asked what their favourite sport is.

You pick a random sample of 6 students to represent the class, by pulling their names out of a hat. Your sample is made up of 5 male and 1 female.

This is a random sample – but it is not representative of the gender split in the class – does this matter?

For this research question, it is likely that this sample (being predominantly male) will give a biased result. So a simple random sample may not be appropriate for this research question.

But if the research question had been "what time did you wake up this morning?", it is less likely to have any associated gender bias and therefore it would not matter that the sample was not representative of the gender split in the class.

When the sample is not representative of the research population, you have what is known as **sampling error**. You can eliminate sampling error by increasing the size of the sample and making it more representative.

Thinking about the above example, if you wanted to know the average height of women in the UK, it would take a lot of time and cost a lot of money to observe the heights of the whole population; instead we can take a sample of the population and use this to get a pretty accurate estimation (as long as the sample is representative, meaning the sample accurately reflects the population) of the population's average height.

Stratified Random Sampling

Stratified sampling involves splitting the population up in to sub-groups and using a proportionate representation of each subgroup in the sample. For example, if you had a population with 20% of people under 30, 50% between the age of 31 and 64 and 30% 65+, you would use the same proportions of ages in your sample. This helps make your sample representative of the total population.

(!) Every member of the population has to belong to a subgroup, and they cannot belong to more than one.

To find the sample size of each subgroup you take the size of the total sample and divide it by the size of the population, you then multiply that answer by the size of the corresponding population sub-group.

EXAMPLE

You want to interview a sample of the residents that live on your street, finding out their views on the local councils proposal to ban on-street parking. To get a representative insight into the populations view, you decide to use a stratified sample of the population split up into house types. You know that 55% of residencies are flats, 30% terrace houses, 10% semi-detached and 5% detached. There are 580 residencies on your street. You only have the resources to interview 60 residencies. The table below shows the calculations for working out the number of each house type to be included in the sample.

Table 3: How to Calculate Stratified Random Sample to be Representative of House Type

House type	Population	Calculation	Number in sample
Flat	0.55*580=319	(60/580)*319	33
Terrace	0.3*580=174	(60/580)*174	18
Semi-detached	0.1*580=58	(60/580)*58	6
Detached	0.05*580=29	(60/580)*29	3

Stratified sampling can be advantageous as it reduces **selection bias**. Selection bias causes one group to be picked more often for the sample, for example choosing mostly one race for the sample. However, you need to be able to identify every member of the population for you to be able to find the required sample size for each sub-group of the population. A stratified sample may prove more time consuming and costly to conduct compared to a systematic or simple random sample, however you are likely to achieve a more representative sample.

What Does ONS Do?

ONS use stratified sampling for their monthly business survey, information about sampling techniques can be found in the relevant Quality and Methodology Information (QMI) publication. In this case, stratified sampling is used for smaller businesses and then for larger businesses, who have a certain number of employees, the whole population is used. The monthly business survey is used to collect information on monthly turnover of UK businesses within the production and service sector. As can be seen below, the sample only accounts for just over 2% of the population, meaning it is important to gain a sample that represents the population as well as it can.

Survey Information

Selection criteria: businesses from various industrial sectors and regions in the UK

- Population: approx. 1,450,000
- Sample: approx. 32,000
- Frequency: monthly
- Dispatch date: 23rd
- Return-by date: 7th

EXAMPLE

We have a high school in which we want to find out how much exercise the students get outside of school. Instead of sampling the whole population, we decide to interview a maximum of 75 students. We have decided to use a stratified sample to ensure that our sample is representative of the whole school.

As you can see in the table below, there are a total of 2603 students in the school, split up into year groups and sex.

Table 4: Total Population

Year	Male	Female	Sum
7	299	290	589
8	252	256	508
9	204	206	410
10	288	305	593
11	241	262	503
Sum	1284	1319	2603

To appropriately reflect students across all year groups and sexes in our sample, we need to work out each subgroup as a proportion of the whole school. You do this by dividing the number in each subgroup by the total number in the whole population. You then multiply this by your chosen sample size (in this example 75) to get the number of participants you will need to choose from that subgroup.

Now that we have our sample proportions worked out, we can randomly select the correct number of members from each subgroup of the population

Once we have all the data from the sample, we can perform all required analysis

Sample size: 75

Table 5: Sample Population

Year	Male	Female	Sum
7	9	8	17
8	7	7	14
9	6	6	12
10	8	9	17
11	7	8	15
Sum	37	38	75

Systematic Random Sampling

A systematic sample is a method in which a segment of the population is taken using a random starting point and a fixed interval. The fixed interval is calculated by dividing the size of the population by the size of the sample that you want. So, if the population size was 50,000 and you wanted a sample of 1,000, the fixed interval would be 50 (50,000/1,000). Your sample would include observations of every 50th member of the population. The starting point is randomly selected so that each member of the population has an equal chance of being selected for the sample.

EXAMPLE

We want to find the average bmi of students in a college. There are 500 students in the college, meaning the population size is 500. However, there is only time to measure the bmi of 20 people. Therefore, we will have to use a sample of the population to estimate the average bmi of students in this college.

In this example, we are going to use a systematic sample of 20 as a representation of the total population. A random number between 1 and 25 was generated, with 3 as the outcome. The 3rd student in the population is therefore going to be our first sample member.

Then to find the rest of the members of our sample, the fixed interval 25, found by dividing the size of the population (500) by the sample size (20), needs to be added to find the second student that will make up our sample.

We then repeatedly jump 25 members of the population until the sample size is reached.

Next, we find the bmi of the 20 students selected for the sample, as seen below, and from this we can find the average bmi for the sample and use this as our estimated population bmi.

Table 6: BMI for Systematic Random Sample

Sample member	Corresponding population member	BMI
1	3	21.54
2	28	22.22
3	53	19.09
4	78	20.45
5	103	19.20
6	128	24.17
7	153	21.66
8	178	20.70
9	203	20.33
10	228	24.01
11	252	24.05
12	278	22.47
13	303	23.51
14	328	22.56
15	353	21.78
16	378	23.51
17	403	19.95
18	428	24.03
19	453	23.04
20	478	20.05

The average bmi of the sample is: **21.91** (this is found by summing the bmi's for every member of the population, then dividing by 20 – the number of people in the sample).

In this case, the average bmi of the total population was 21.69.

As you can see from the above example, the sample mean was a pretty accurate estimate of the population mean. A systematic sample is an improvement over a simple random sample as it spreads the sample over the whole population. However, this only holds if the population is random and there is no hidden pattern in the population. If this is the case, there is a high chance of sampling error.

For example, if the population was set out so that every member alternated between being male and female, and the fixed interval was an even number, the whole sample would consist of only males or only females depending on whether the random starting point landed on a male or female.

Systematic samples are relatively easy and low-cost compared to other sampling methods, also they can be used for large samples, so long as a list can be created of the whole population.

What Does ONS Do?

One example of where ONS use systematic sampling is in its 'Wealth and Assets Survey', this survey measures the well-being of households and individuals in terms of their assets, savings and debt and planning for retirement. Systematic sampling is used in conjunction with stratified sampling to form the sample for this survey. Systematic sampling is used as the second part of the sampling process, where 26 addresses are picked for each primary sampling unit (PSU). To find each PSU in the first place, stratified samples were taken from the small users' postcode address file. For more information on the Wealth and Assets Survey, see link.

Sample Size

You can have a sample size of anything up to the population size. You should aim to make your sample as large as is feasible to do so. Larger samples will give a better representation of the population and will give you more accurate results/interpretation.

There are four things you need to know to determine your sample size:

- the size of the population
- the confidence interval (the range that the value of a parameter falls into, given a specific probability – more info on page 56)
- the confidence level (the probability that the value of the parameter falls within the range of the confidence interval)
- how much variance you would expect the sample to produce

Go to the data analysis chapter for more information on confidence intervals and the variance.

Online sample size calculators are available for example Survey Monkey. Why not try it? Enter the population size, confidence interval and confidence level.

Studies With Small Sample Sizes

In statistics, when it comes to sample size, bigger is usually better.

So why not sample the whole population? Dealing with large samples can become an issue practically; as more participants means more time and money is required to conduct the study and analyse the results. This is why using existing data sets can be really useful. Places like ONS do the work to produce large data sets from huge samples, that no one person would be able to do – like the census.

Due to time constraints, an issue you may face in your own research is relying on a small sample size. As long as you are aware of the limitations of this, and can express what this means in relation to the conclusions you are drawing from your data, this is not a problem!

What are the issues of having a small sample? Small samples can be problematic, as the smaller the sample, the less information is contained within it. The chances of finding a significant difference in data is small if the sample is small. So, what does it mean if our findings don't produce a **significant** difference? Basically, we can't disprove the null hypothesis, which means we can't say anything definitively about the results.

Statistical Significance

We conduct significance testing to know if we should reject the **null hypothesis**. If we do not have enough information to be able to say for sure that the thing we are altering is what is causing the difference, then we have to say it is not significant.

What does "not significant" mean? It means that we have failed to demonstrate that there is evidence against the null hypothesis. For example, if we are doing a test to see if a new medicine works better than a placebo, the null hypothesis would be that there will be no difference in outcomes for patients treated with the new drug compared to the placebo.

This is often misinterpreted as meaning that there is no difference between the two treatments. In reality, all it means is we do not have enough evidence to prove that there is a difference if it is present.

If there really is no difference between two populations, the probability of a significant difference being found between two samples is 0.05, whatever the sample size.

Accuracy and Reliability of Sources

Accurate and **reliable** data is very important for analysis. Government departments (especially in the UK) offer accurate and reliable statistics and a great place to start is by looking on the Office for National Statistics website. We know that statistics are accurate and reliable if they have been accredited national statistics by the Office for Statistics Regulation and use the following quality mark;

As well as the ONS, you can find statistics across all departments by looking on the GOV.UK website.

There are a few private sector companies that offer free data but the level of detail won't be as easily accessible as you will often have to pay for the data. An example of data produced by the private sector is the purchasing managers index (PMI). This list below gives some sources of reliable data;

- Office for National Statistics
- GOV.UK
- European Central Bank
- Institute of Fiscal Studies website
- The Bank of England
- The Organisation for Economic Co-operation and Development (OECD)
- Welsh Government
- Stats Wales

Unreliable Data

When using secondary data always be aware that there may be some bias involved. The ONS uses statistical methodology to reduce bias so that users know that data produced by ONS is reliable and accurate.

However, some official statistics in different countries may not be as heavily regulated and governments may have an agenda to produce different statistics. It is always worth looking for the methodology of how a statistic is produced to check it's accuracy and reliability. For example all ONS publications come with Quality and Methodology Information (QMI) report. An example of this for our workforce jobs publication can be found here.

LEARNING OBJECTIVES

By the end of this chapter students should feel able to:

- ✓ Understand the difference between Primary and Secondary research.
- ✓ Decide which type of data is most suitable for their research question.
- ✓ Understand the difference between different sampling techniques and when it is appropriate to use them.
- ✓ Be able to evaluate the accuracy and reliability of different types of source material.

Data Analysis

03

Correlation

What is Correlation?

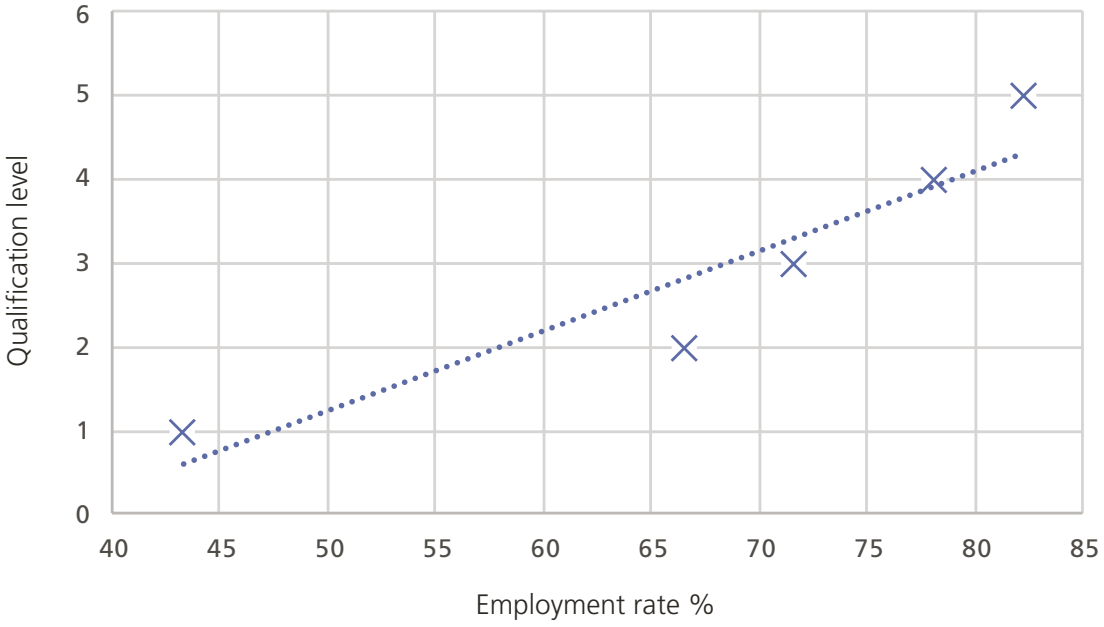
Correlation allows us to measure the relationship between two variables (i.e. workers education and wages).

Correlation is expressed as a number; correlation always lies between -1 and 1. It is important that you know how to interpret this number, so the properties of correlation are explained below.

Positive Correlation

Positive numbers indicate a **positive correlation** between two variables, in the example below qualification level and employment rate are the two variables. From the chart below as the employment rate increases, the qualification level gets higher. Therefore, as one increases the other increases. The line has a positive gradient therefore we say that there is a positive correlation.

Figure 3: Percentage Employment Rate Against Qualification Level

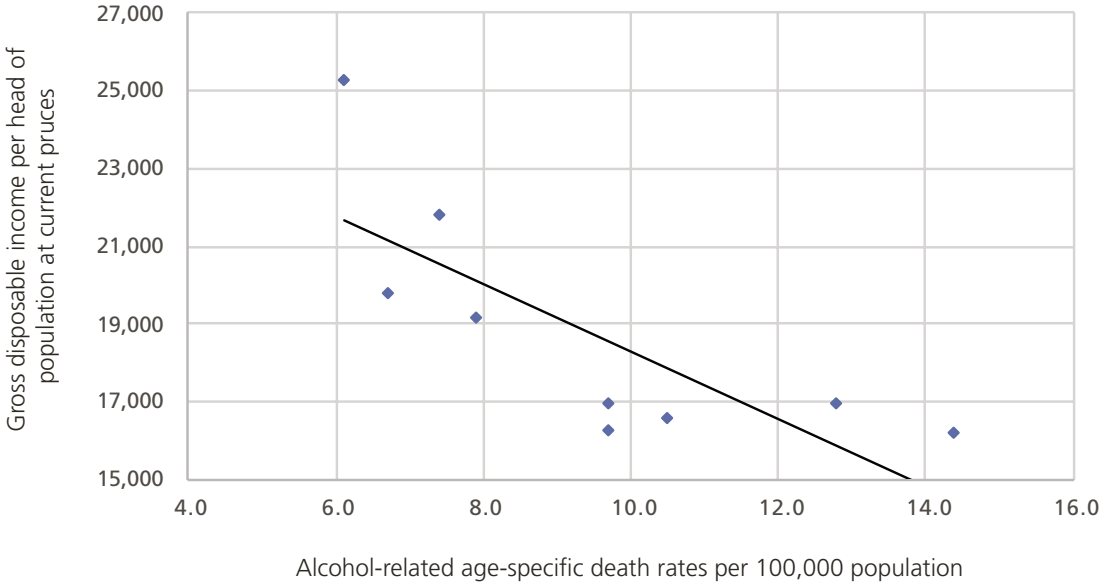


Source: Labour Force Survey, **Office for National Statistics**

Negative Correlation

Negative values indicate a **negative correlation**. From the chart below as the number of alcohol-related deaths increase in a region the lower the disposable income per head in that region. In other words, as one increases the other decreases. The line has a negative gradient, and we say there is negative correlation.

Figure 4: The relationship between regional gross disposable income per head of population at current prices and regional alcohol-related age-specific death rates per 100,000 population

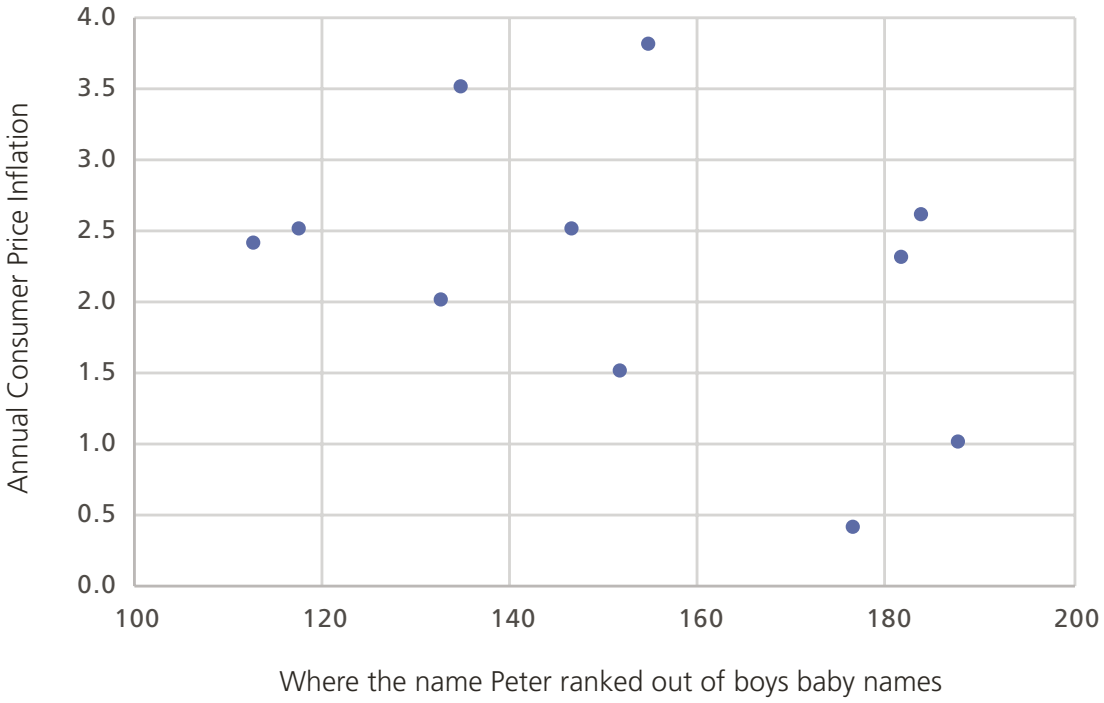


Sources: Health and social care, Gross disposable household income, **Office for National Statistics**

No Correlation

When correlation = 0 it indicates that the two variables are uncorrelated (no relationship). For example, from the chart below there is no correlation between the number of boy's named Peter and the consumer price inflation rate.

Figure 5: The Relationship Between the Popularity of the Name Peter and Consumer Price Inflation



Sources: Consumer price inflation times series, Baby names in England and Wales, **Office for National Statistics**

Properties of Correlation

- Larger positive values indicate stronger positive correlation and larger negative values indicate a stronger negative correlation.
- When correlation = 1 this indicates perfect positive correlation.
- When correlation = -1 this indicates perfect negative correlation.
- The correlation between Y and X (i.e. ice cream and temperature) is the same as the correlation between X and Y .
- The correlation between any variable and itself (for example, the correlation between Y and Y) is 1.

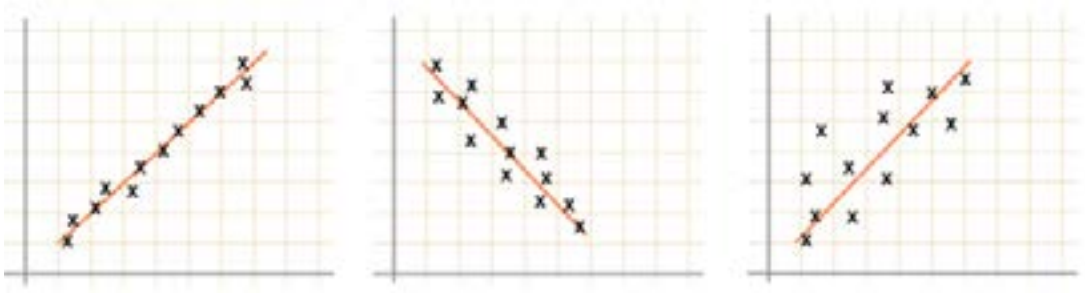
Lines of Best Fit

The 'line of best fit' is a line that goes through most of all the scatter points on a graph. The closer the points are to the line of best fit the stronger the correlation.

The line of best fit is drawn so the points are evenly distributed on either side of the line.

The first chart shows the points are closer to the line of best fit. This means it has a stronger correlation than the other two graphs.

Figure 6: Scatter diagrams illustrating how to draw a line of best fit



Source: BBC GCSE Bitesize

EXAMPLE

Say you're investigating the relationship between workers education and their likelihood of being employed. You can measure correlation between education and employment using Excel or any spreadsheet/statistical package.

Here is data from the ONS that shows the percentage of those who are employed for each level of qualification status.

Table 7: Exploration of Employment Rate and Educational Level

Employment Rate (percent)	Education Level	Education Qualification Label
82	5	Graduates
78	4	A Level or equivalent
72	3	A* to C grade GCSE or equivalent
67	3	Other qualifications
43	1	No qualifications

Source: Office for National Statistics

To work out the correlation in excel you use the command 'correl' and then select each column separately.

Figure 7: Excel Example of CORREL Function

Qualification	Education qualification level	%
Graduates	5	82
A level or equivalent	4	78
A* to C grade GCSE or equivalent	3	72
Other Qualifications	2	67
No Qualifications	1	43
	=CORREL(B8:B12,C8:C12)	

Source: Office for National Statistics

The correlation from this example is 0.93 which indicates a strong positive relationship between employment and the level of education. In other words, the higher the level of education the higher the likelihood of being employed.

So, in our education and employment example we found a positive relationship but what does it mean if our correlation number is negative? Let's suppose we are interested in investigating the relationship between an increase in vehicle tax and congestion on the roads. We find that the correlation (or relationship) between vehicle tax and congestion is negative. This means that the higher the vehicle tax the lower the amount of congestion on the roads. Also, it suggests that the lower the vehicle tax the greater the congestion on the roads will be.

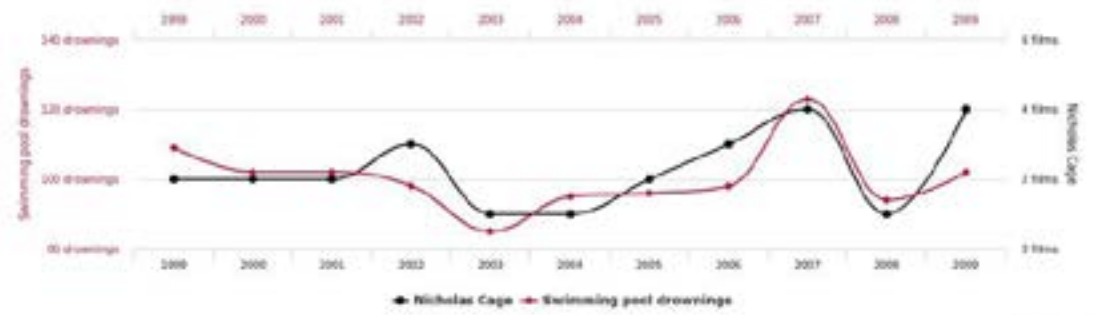
Correlation Doesn't Mean Causation!

In our education/employment example, we discovered that education and employment are correlated positively, indicating a positive relationship between the two. This finding of a positive correlation between education and employment could be interpreted as meaning that more education does directly influence the chances of getting employed. However, correlation doesn't always mean causality.

Looking at our education/employment example, one reason for the positive correlation could be that greater education means workers are more specialized and able gain employment easier. In this case, education would be an indirect cause, and specialized skills, a direct cause of employment.

The key message to take away is that correlations can be very suggestive but doesn't always establish **causality**. Take a look at the example below, there is high correlation between the number of people who drowned by falling into a pool and films Nicolas Cage appeared in. But watching Nicholas Cage films doesn't cause people to drown in a pool. So, although these two things are correlated, one doesn't cause the other. For other examples of correlation without causation see this website.

Figure 8: Number of People Who Drowned by Falling into a Pool Correlates With Films Nicolas Cage Appeared in



Sources: Tylervigen

Correlation Summary

Correlation is a common way of measuring the relationship between two variables. It is a number that can be calculated using Excel or any spreadsheet software package.

Correlation can be interpreted by using scatter diagrams. The sign of the correlation relates to the slope of a best fitting line and the magnitude of the correlation relates to how scattered the data points are around the best fitting line.

Correlations does not necessarily imply causality between two variables.

Standard Deviation

What is Standard Deviation?

Standard deviation is an important measure of spread or dispersion.

When looking at data measures such as the mean, median or mode they can hide the **variability** of the data so the standard deviation can work out the spread of the dataset.

For example, the two lists of numbers below are significantly different in nature but have the same; mean 17.1 (the average which is the total sum of the numbers divided by how many numbers there are), median 18 (the midpoint) and range 32 (which is the highest number minus the lowest number in the sample i.e. 33 - 1).

1 6 12 18 22 28 33

1 15 15 18 19 19 33

The standard deviation of the first list is significantly larger than the standard deviation of the second list. In other words, there is more spread around the mean in the first list of numbers.

Working Out the Standard Deviation

An easy way of working out the standard deviation is by using the function "STDEV" in excel.

This example below measures the standard deviation of real gross household disposable income per head.

Figure 9: Excel Example of STDEV Function on Household Income

	Real Gross Household Disposable Income per head £ Deflated, Market Prices Seasonally Adjusted	Real Gross Adjusted Household Disposable Income per head CRYB £ Deflated, Market Prices Seasonally Adjusted	Real Net Household Adjusted Disposable Income per head CRSF £ Deflated, Market Prices Seasonally Adjusted
2015 Q1	4768	5900	5018
2015 Q2	4857	5990	5094
2015 Q3	4972	6115	5207
2015 Q4	4902	6022	5110
2016 Q1	4814	5946	5041
2016 Q2	4844	5984	5071
2016 Q3	4879	6014	5095
2016 Q4	4774	5904	4981
2017 Q1	4721	5836	4921
2017 Q2	4793	5919	5002
2017 Q3	4802	5912	4984
2017 Q4	4776	5882	4956
=STDEV(A18:B57			

STDEV(number1, [number2],...)

Source: Economic well-being, Office for National Statistics

From the calculation the standard deviation is £120.40 which suggests that the dataset is spread.

The lower the standard deviation the more clustered the data is around the average (mean). The higher the standard deviation the greater the spread of the data and less clustered it is around the mean.

If we assume that the dataset is normally distributed (See page 45) then we can say that most people of the sample population (68%) have a gross household income per head within £120.40 of the average.

For a step by step guide to work out your standard deviation by using calculations please see Appendix A at the back of this Book.

EXAMPLE

How standard deviation can be beneficial...

Suppose you are looking for a new job and you see that people who work there earn on average £51,800. However, when taking the standard derivation of people's earnings at this company you find that the standard derivation is £31,925. This shows that the spread of the data is large, meaning you are unlikely to earn the average salary of £51,800 if you took the job. If we assume the data is normally distributed as explained below we can say that most people who work in this company (68%) have a salary within £31,925 of the average salary.

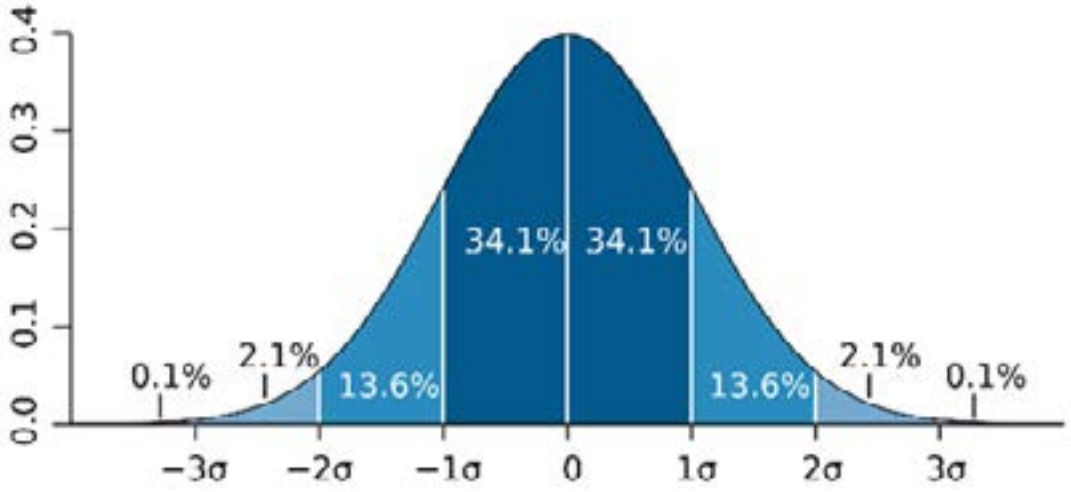
Distribution of Data

Normal Distributions

A **normal distribution**, sometimes called the bell curve, is a distribution that occurs naturally in different situations. Take for example people's height, with most people being of average height and with only a few being either very tall or very small.

A perfect normal distribution will look like the diagram below, with 50% of the sample falling below the average (in the middle) and 50% above the average.

Figure 10: Example of Normal Distribution Bell Curve



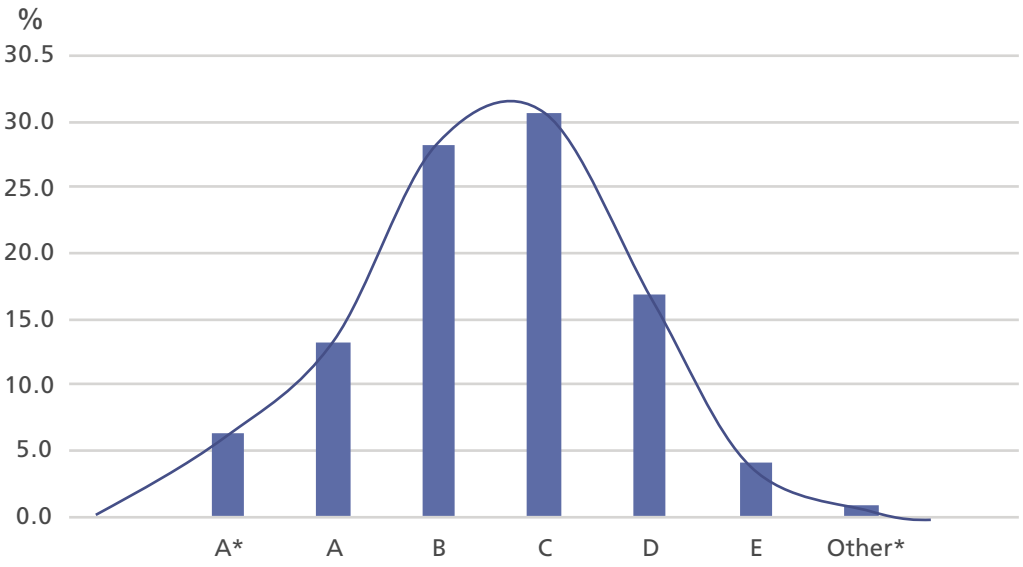
For normally distributed data we find generally;

- 68% of values are 1 standard deviation away from the mean
- 95% of values are 2 standard deviation away from the mean
- 99.7% of values are 3 standard deviation away from the mean

So, going back to our salary example we found that one standard deviation was £31,925 so we can say that 68% of the sample would fall within £31,925 of the average salary.

The normal distribution below shows the scores of people who did GCSE English in 2017. We can see from the data that there is a resemblance to the bell curve with the highest percentage of students obtaining the average grade C (31%) and fewer students obtaining the outliers at grade A* and Other (6.3% and 0.3% respectively).

Figure 11: GCSE English Grades



Source: Department of Education

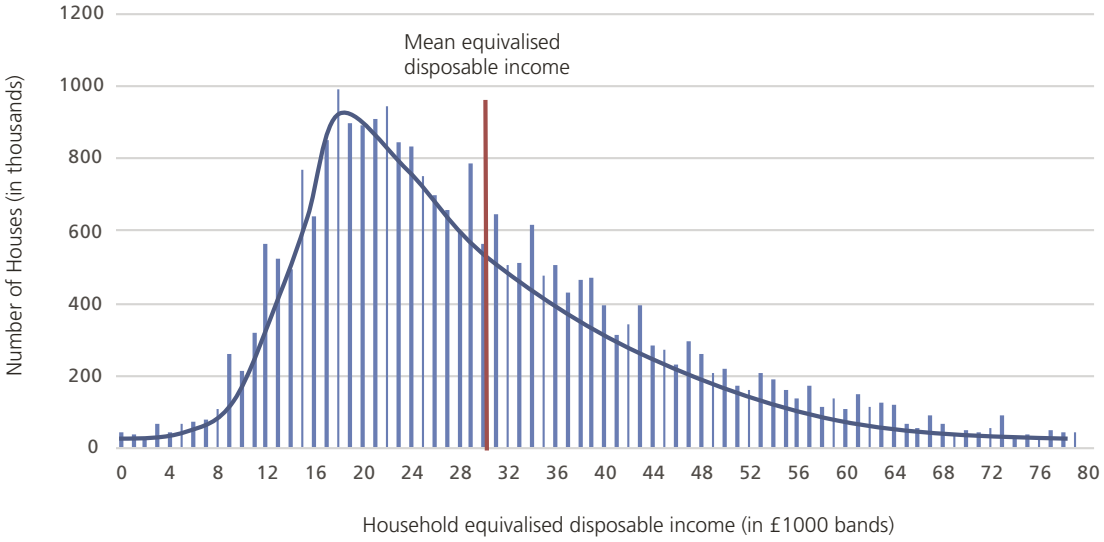
***Includes ungraded, no award (absent/declined) and pending.**

Skewed Data

When the data isn't normally distributed the data may be **skewed** either negatively or positively. This means that there is distortion in the data and it is not normally distribution.

Skewness characterises the degree of asymmetry of a distribution around its mean. The diagram below shows a positive skew as the long tail is on the positive side of the peak and the mean is on the right of the peak value.

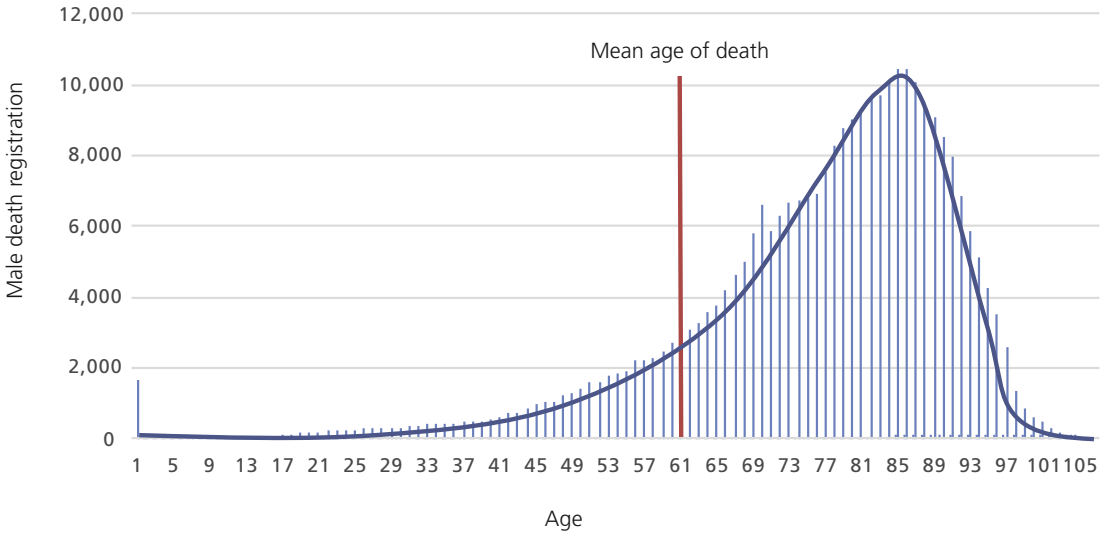
Figure 12: Distribution of UK Household Disposable Income, Financial Year Ending 2017



Source: [Household disposable income and inequality in the UK: financial year ending 2017](), **Office for National Statistics**

Negative skewness indicates a distribution with an asymmetric tail extending toward more negative values. The diagram below shows a negative skew with the long tail at the negative side of the peak and the mean on the left-hand side of the peak.

Figure 13: Male Death Registration by Single Year of Age, UK, 2016



Source: [Deaths by single year of age tables, UK](), **Office for National Statistics**

Problems With Skewed Data

When data is normally distributed the average lies directly in the middle of the data and there is an equal 50/50 split of data above and below the average. In other words, the average is the same as the median (the midpoint of the data set). Therefore, when we do analysis we are able to say that the majority of the sample population (68.2%) are within one standard deviation of the average (mean).

However, when the dataset is skewed you have to be careful with the analysis. For example, figure X shows that the mean equalised disposable income for households in the UK was £30,000 a year. However, this average isn't caused by the majority and is skewed by a few households having a large annual income. This is the sort of thing you have to bear in mind when making any conclusions based upon a data set that is not normally distributed.

Confidence Intervals

What is a Confidence Interval?

Statistics are often meant to reflect a large population, but statisticians or analysts rarely have access to data for every single member of that population. To estimate statistics for a large population, statisticians will use a sample (which is designed to reflect the makeup of the whole population). To check that their sample represents a true reflection of the true population they can use confidence intervals. A confidence interval is a range of values, worked out from sample statistics, that is likely to contain the value of an unknown population value.

For example, suppose that you randomly sample individuals and ask them how many hours a week they do exercise. The confidence interval calculated is 2 – 6 hours. The confidence interval indicates that you can be 95% confident that the mean of the entire population falls within this range and the average person exercises for at least 2 hours and at the most 6 hours a week.

There are a range of confidence intervals that analysts use, although the most frequently used are the 99%, 95% and 90% confidence intervals.

What Does ONS Do?

ONS calculate their confidence intervals to help the users see how accurate the data they are using in their own analysis.

For example, the ONS produces confidence intervals for 'Travel trends' (annual estimates of travel and tourism visits to the UK and associated earnings and expenditure between the UK and the rest of the world) at the 95% confidence interval. The data table below shows the estimate for each of the categories and their 95% confidence interval.

Let's take the 'Number of visits to the UK' which is estimated to be 37,609,000 people in 2016. The 95% confidence interval says that we can be 95% confident that 37,609,000 lies within 2.2% +/- % of the true population estimate.

Figure 14: International Passenger Survey Confidence Intervals for 2017 Estimates

	Estimate	Relative 95% confidence interval (+/- % of the estimate)
Overseas visitors to the UK		
Number of visits ('000s)	39,214	2.3%
Total earnings (£million)	24,507	3.3%
Number of visitor nights ('000s)	284,781	3.0%
UK residents going abroad		
Number of visits ('000s)	72,772	1.8%
Total expenditure (£million)	44,840	2.5%
Number of visitor nights ('000s)	743,469	2.5%

Source: Travel Trends: Confidence intervals, Office for National Statistics

Another example is working out the 95% confidence interval for internet access for households and individuals. From the table below each of the survey estimates fall within the lower and upper limit of the 95% confidence interval. So for example, by using the table below we can state with 95% confidence that 40.9 million people used the internet on a daily basis during 2017.

Figure 15: household internet access with 95% confidence intervals

	Lower limit		Survey estimate		Upper limit	
	Millions	%	Millions	%	Millions	%
2015	38.3	76	39.3	78	40.3	80
2016	41.0	81	41.8	82	42.5	84
2017	40.0	78	40.9	80	41.7	82
2018	43.2	84	44.1	86	44.9	87

Source: Internet access – households and individuals: 95% confidence intervals, **Office for National Statistics**

Properties of Confidence Intervals

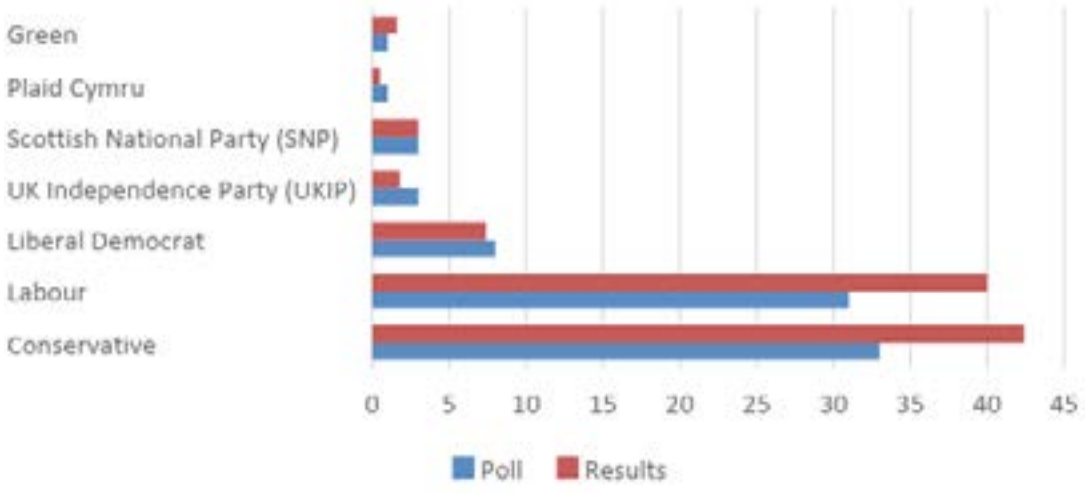
The larger the sample size the closer the representative sample is to the true population. Therefore, the confidence interval (the lower and upper bounds) will be narrower and you can be more certain that your sample estimate represents the true population sample.

How do we Use Confidence Intervals and How Can They be Made More Accurate?

Analysts/Statisticians use confidence intervals to work out if their data is representative of the population. Therefore, the larger the sample size the more likely it is to be representative of the population and the smaller the confidence interval will be.

For example, in the run-up to the UK's general election last year YouGov surveyed 2130 adults in the UK asking them who they would vote for. The results from the poll suggested that Conservatives were the most popular with 33% voting Conservatives, closely followed by voters for Labour who had 31% of the sample, then the Liberal Democrats with 8% of the sample. The actual results on the 8th of June were similar with most people voting Conservative and Labour. However, the percentage of people who voted both Conservative and Labour were larger and different at 42.4% and 40.0% respectively. This suggests for the polls to be more accurate than their sample size needs to be larger to be more representative of the population. A larger sample would also have a narrower confidence interval.

Figure 16: The % of People That Voted for the Different Parties in the 2017 UK General Election



Source: BBC and YouGov

Standardisation of Data

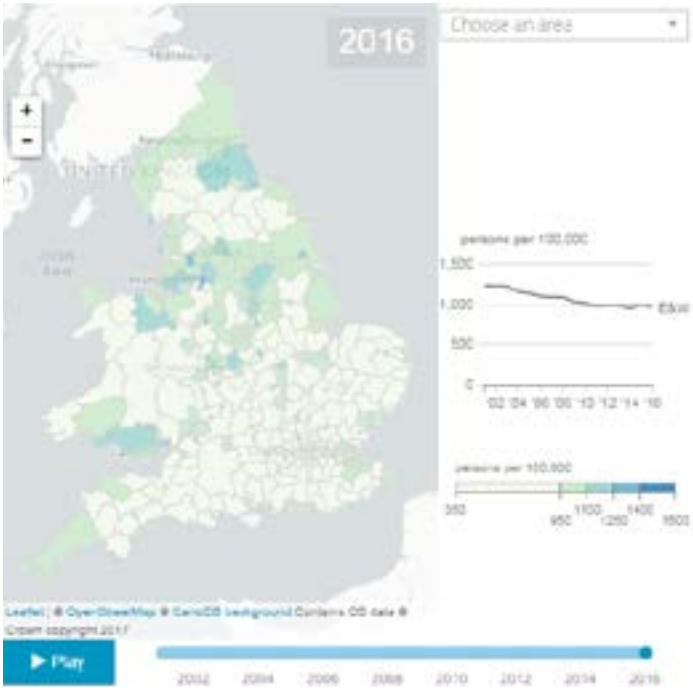
Age Standardisation

Age standardisation is a technique used to allow populations to be compared when the age profiles of the populations are different. For example, if we compared two countries health costs we may find that country A has greater health costs then country B. However, country A has a higher percentage of elderly people. Using the assumption that more elderly people access healthcare than younger people then looking at the comparisons of health costs between country A and country B wouldn't be fair. By standardising the age of the population, we can eliminate the effect that the older generation has on the results to make better and accurate comparisons between the two countries.

EXAMPLE

The ONS use age standardisation when comparing the mortality rate between local authority districts in the UK.

Figure 17: Mortality Rates Across the UK



Source: Office for National Statistics

To Calculate Age Standardisation

To adjust for age, a standard population must be selected. Standard populations have been developed for specific countries and regions. ONS use the European Standard Population (ESP) when comparing population statistics across Europe.

This is the equation the ONS uses to calculate the age-standardised rate. The population is split up into age categories for each sex, so there are two groups for each age category and all these groups are used in the formula below.

Age-standardised rate = (Pkmk)Pk

Where;

=”the sum of”

Pk = Standard population in sex/age group k

mk = Observed mortality rate (deaths per 100,000 persons) in sex/age group

k = age/sex group 0, 1-4, 5-9, …, 80-84, 85 years and over

Worked Example Using ONS Data

Using the data shown below we calculate the age standardisation.

Firstly, we work out the observed mortality rate, see below;

Figure 18: Excel Example of Calculating Age Standardisation

Age group	ESP	Deaths	Population	Age-specific	Age-standardised
<1	1,600	1,856	358,077	=C7/D7*100000	829,318.8
1-4	6,400	245	1,365,532	17.9	114,827.0
5-9	7,000	147	1,547,075	9.5	66,512.6
10-14	7,000	195	1,636,034	11.9	83,433.5
15-19	7,000	778	1,804,424	43.1	301,813.8
20-24	7,000	1,121	1,924,978	58.2	407,641.0
25-29	7,000	1,371	1,878,849	73.0	510,791.4
30-34	7,000	1,603	1,724,669	92.9	650,617.6
35-39	7,000	2,586	1,908,443	135.5	948,521.9
40-44	7,000	3,887	2,073,527	187.5	1,312,208.6
45-49	7,000	5,116	1,941,822	263.5	1,844,247.3
50-54	7,000	6,859	1,693,684	405.0	2,834,826.3
55-59	6,000	10,406	1,560,711	666.7	4,000,484.4
60-64	5,000	16,201	1,617,467	1,001.6	5,008,139.3
65-69	4,000	19,746	1,209,237	1,632.9	6,531,722.1
70-74	3,000	27,002	1,018,781	2,650.4	7,951,267.2
75-79	2,000	35,310	785,993	4,492.4	8,984,812.8
80-84	1,000	41,758	531,633	7,854.7	7,854,666.7
85+	1,000	61,875	399,177	15,500.6	15,500,642.6

Once we have the mortality rate (under age-specific column in above table) we can multiply this by the standard population. The ONS uses the European standard population (ESP) shown below.

Figure 19: European Standard Population Distributions

Age	Population	Age	Population
<1	1,600	45-49	7,000
01-04	6,400	50-54	7,000
05-09	7,000	55-59	6,000
10-14	7,000	60-64	5,000
15-19	7,000	65-69	4,000
20-24	7,000	70-74	3,000
25-29	7,000	75-79	2,000
30-34	7,000	80-84	1,000
35-39	7,000	85+	1,000
40-44	7,000		
		Total	100,000

Below shows ESP multiplied by the mortality rate to give you the age-standardised in the final column.

Figure 20: Excel Example of how to Calculate Age-Standardised Mortality Rates Using ESP

Age group	ESP	Deaths	Population	Age-specific	Age-standardised
<1	1,600	1,856	358,077	518.3	=E7*B7
1-4	6,400	245	1,365,532	17.9	114,827.0
5-9	7,000	147	1,547,075	9.5	66,512.6
10-14	7,000	195	1,636,034	11.9	83,433.5
15-19	7,000	778	1,804,424	43.1	301,813.8
20-24	7,000	1,121	1,924,978	58.2	407,641.0
25-29	7,000	1,371	1,878,849	73.0	510,791.4
30-34	7,000	1,603	1,724,669	92.9	650,617.6
35-39	7,000	2,586	1,908,443	135.5	948,521.9
40-44	7,000	3,887	2,073,527	187.5	1,312,208.6
45-49	7,000	5,116	1,941,822	263.5	1,844,247.3
50-54	7,000	6,859	1,693,684	405.0	2,834,826.3
55-59	6,000	10,406	1,560,711	666.7	4,000,484.4
60-64	5,000	16,201	1,617,467	1,001.6	5,008,139.3
65-69	4,000	19,746	1,209,237	1,632.9	6,531,722.1
70-74	3,000	27,002	1,018,781	2,650.4	7,951,267.2
75-79	2,000	35,310	785,993	4,492.4	8,984,812.8
80-84	1,000	41,758	531,633	7,854.7	7,854,666.7
85+	1,000	61,875	399,177	15,500.6	15,500,642.6
Total	100,000	238,062	26,980,113	35,615.8	65,736,495.0

Income Equivalisation

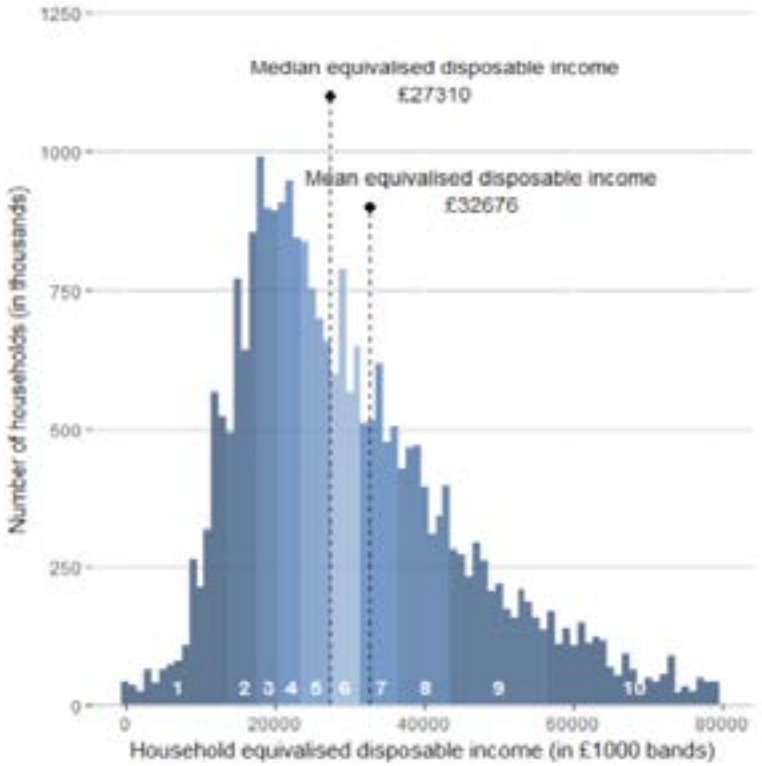
Income equivalisation or equivalised income refers to household income that has been recalculated to consider differences in household size and composition. For example, households with a lot of people are likely to need a higher income to achieve the same standard of living as households with less people.

Also, different ages of people living in a household will affect how much income people need as living costs for adults are normally higher than for children. After equivalisation is applied, households with the same equivalised income can be said to have a comparable standard of living.

EXAMPLE

The ONS use income equivalisation when looking at the distribution of UK household disposable income.

Figure 21: Distribution of UK Household Disposable Income



Source: [Household disposable income and inequality in the UK: financial year ending 2017](), **Office for National Statistics**

How to Calculate Equivalised Income

There are various scales available, although the OECD (Organisation for Economic Co-operation and Development) modified equivalence scale is used widely across Europe, which allows comparisons to be made between European countries. It adjusts household income to reflect the different resource needs of single adults, any additional adults in the household and children in various age groups.

To calculate the equivalised income using the modified OECD equivalence scale, each member in the household is first given an equivalence value. The modified OECD equivalence values are shown in the table X. Single adult households are taken as the reference group and are given a value of 1.

For larger households, each additional adult is given a small value of 0.5 to reflect the economies of scale achieved when people live together. Economies of scale arise when households share resources such as water and electricity, which reduces the living costs per person. Children under the age of 14 are given a value of 0.3 to take account of their lower living costs, children aged 14 and over are given a value of 0.5 because their living costs are assumed to be the same as an adult.

Table 8: Income Equivalence Values for Different Age Groups

Type of Household Member	Equivalence value
First adult	1.0
Additional adult	0.5
Child aged: 14 and over	0.5
Child aged: 0-13	0.3

The equivalence values for each household member are summed to give a total equivalence number for the household. For example, the total equivalence value of a married couple with three children, two of which are over 14 and one who is 11 would be:

Equivalence value = 1 + 0.5 + 0.3 + 0.5 + 0.5 = 2.8

The total equivalence value of 2.8 shows that the household needs nearly three times the income of a single adult household to achieve a comparable standard of living.

Once the equivalence value is calculated you than divide the total income for the household by the equivalence value. For example, if the household we explained above had an annual income of £40,000 their equivalised income would be £14285.71.

£40,000/2.8 = £14285.71

For a single adult household with an income of £40,000, the equivalised income remains at £40,000 because of the equivalence value for one adult is 1. This shows that a single adult household will have a higher standard of living than a larger household with the same level of income.

LEARNING OBJECTIVES

By the end of this chapter students should feel able to:

- ✓ Understand and calculate simple correlations.
- ✓ Understand and calculate standard deviation.
- ✓ Understand and recognise different distributions and confidence intervals of data.
- ✓ Be able to understand which form of analysis is best to answer their research question.

Data Presentation

04

Why Present Data?

While many people struggle to interpret numbers, visual aids like charts are a much more universal language.

Without them, it's easy to miss important details about data.

Look at these four datasets. They're actual and predicted test scores taken from four fictional classes.

Table 9: Actual and Predicted Results for Four Fictional Classes

Class A		Class B		Class C		Class D	
Actual Result	Predicted Result	Actual Result	Predicted Result	Actual Result	Predicted Result	Actual Result	Predicted Result
10	8.04	10	9.14	10	7.46	8	6.58
8	6.95	8	8.14	8	6.77	8	5.76
13	7.58	13	8.74	13	12.74	8	7.71
9	8.81	9	8.77	9	7.11	8	8.84
11	8.33	11	9.26	11	7.81	8	8.47
14	9.96	14	8.1	14	8.84	8	7.04
6	7.24	6	6.13	6	6.08	8	5.25
4	4.26	4	3.1	4	5.39	19	12.5
12	10.84	12	9.13	12	8.15	8	5.56
7	4.82	7	7.26	7	6.42	8	7.91
5	5.68	5	4.74	5	5.73	8	6.89

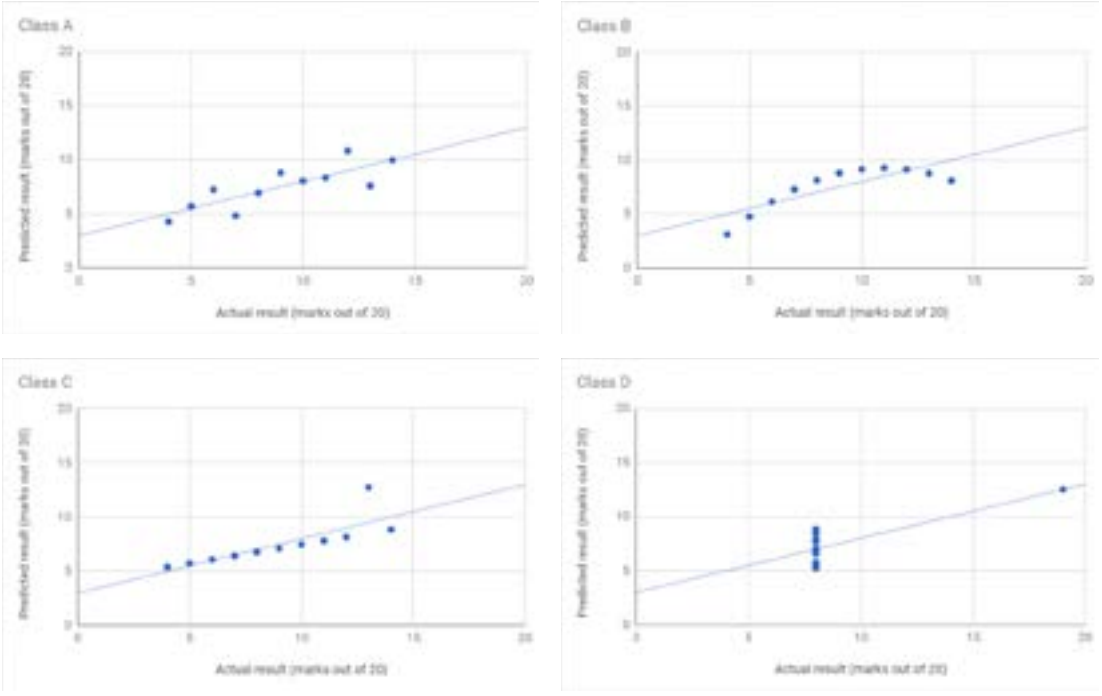
The mean (the average) actual and predicted results are the same for all four classes, and the variance (how far a set of numbers are spread out) of both results are also identical.

Table 10: Actual and Predicted Mean and Variance Values

Mean actual result	9
Variance of actual result	11
Mean predicted result	7.5
Variance of predicted result	4.1

Therefore, without visualising the results for each class, you'd be forgiven for thinking they're all the same!

Figure 22: Test Results From Four Fictional Classes Showing the Range of Distributions Despite Data Having the Same Mean and Variance.



Presenting (or visualising) data helps to do the following:

- give a fast overview or summary of a dataset
- communicate memorable or important stories in a dataset(s)
- reveal insight (that would otherwise be hidden)
- show errors or anomalies (data points which are very different to the rest of the sample) in a dataset

Which Form of Presentation is Appropriate?

Text, tables, graphs and maps represent a toolkit for statistical communication. Being able to spot when and where to use each is vital for communicating your conclusions effectively.

The most important question to ask yourself is what are you trying to show? What message do you want the reader to take away?

Sometimes, your conclusion might not require a chart or table, you can use text. In other cases, you'll need a series of charts.

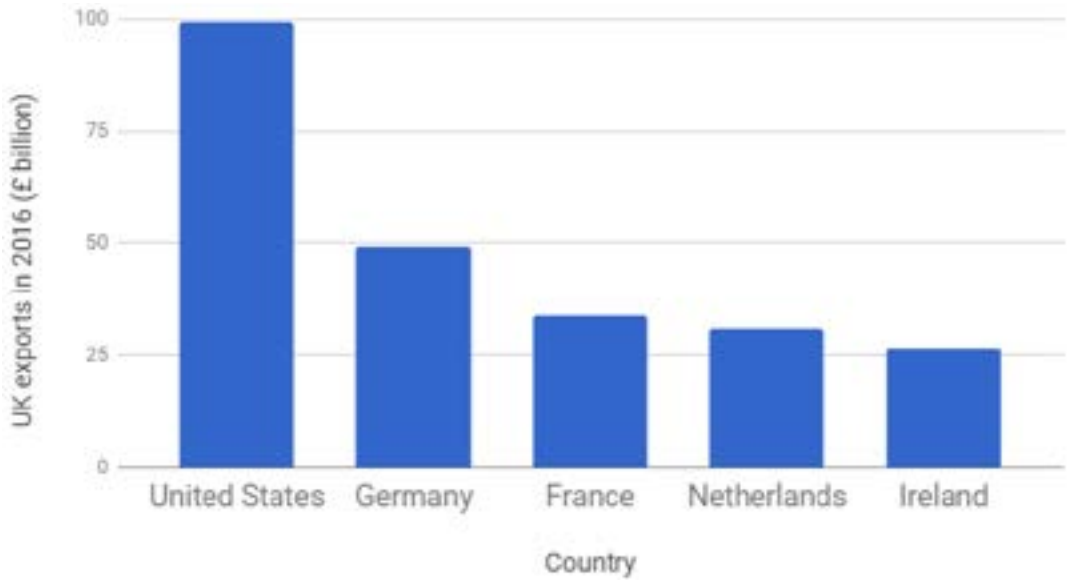
Before you decide which approach to take, you'll need to first establish and then prioritise the statistical relationships you're trying to show.

Let's use the example of UK trade with the rest of the world.

Say you want to show UK exports to its biggest partners – you want to compare the size or **magnitude** of exports.

As a general rule, bars are the best way to show magnitude (they could be horizontal or vertical).

Figure 23: Top 5 UK Export Partners Shown as a Bar Chart

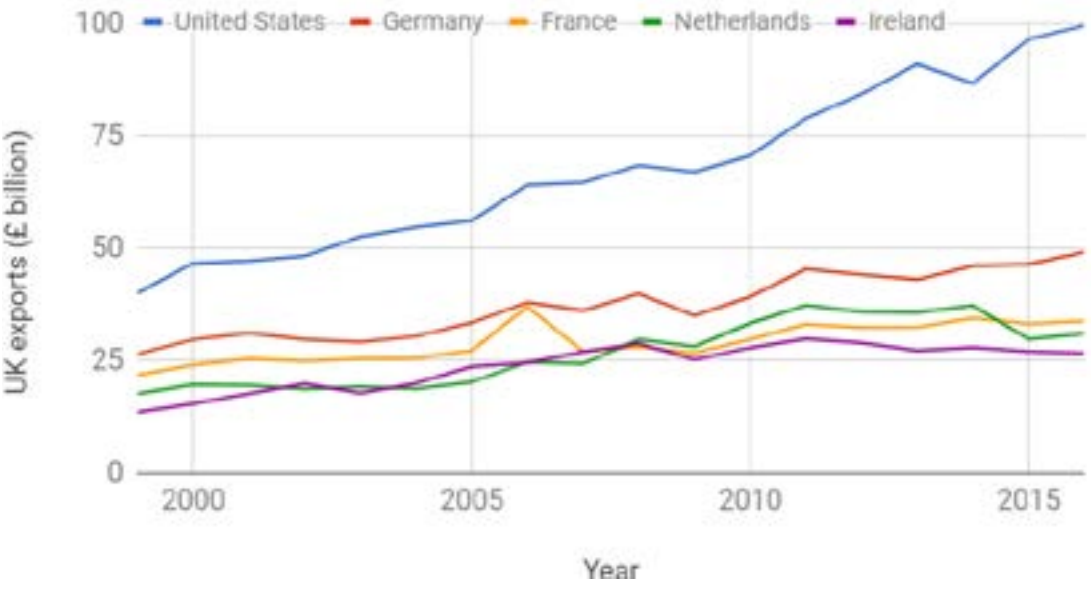


Source: Office for National Statistics

Now, we've shown the UK's top five export partners, but how have our exports to these countries **changed over time**?

A line chart is a good way of showing change over time.

Figure 24: Trend in UK Exports to top Five UK Export Partners from 1999-2016



Source: Office for National Statistics

Staying with our trade example, let's say we want to compare our exports to the EU with our exports to the rest of the world.

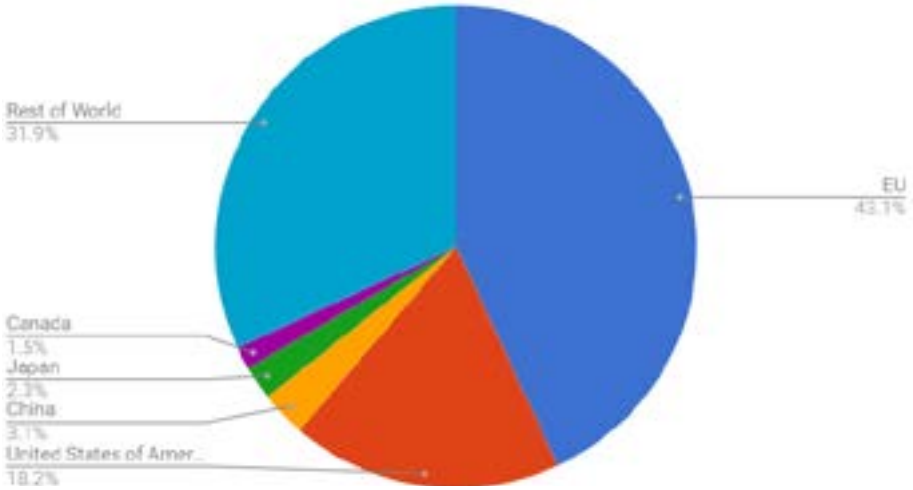
This time, however, we want to look at what **proportion** of our exports go to the EU and the rest of the world. Your first instinct is probably to use a pie chart.

Figure 25: Distribution Of UK Exports Between EU and Rest of the World

Source: Office for National Statistics

A pie chart works well in the case on the previous page because there are only two categories and the difference is clear to see. But what if we add countries to this breakdown?

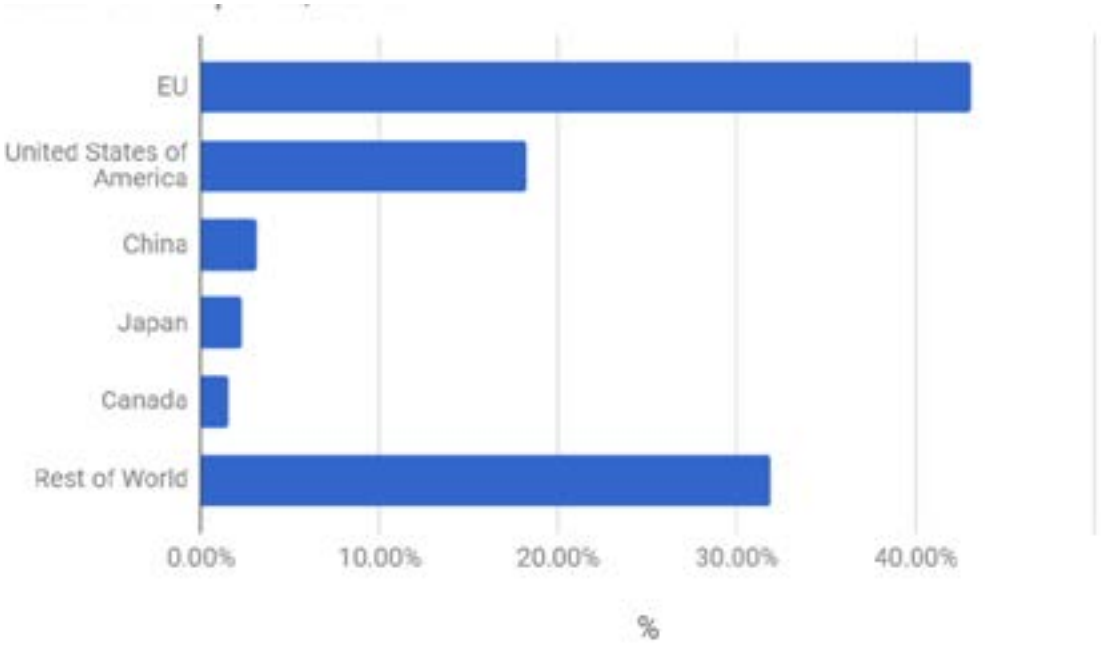
Figure 26: Pie Chart Distribution of UK Exports Between EU, Rest of the World, USA, China, Japan and Canada



Source: Office for National Statistics

In this case a pie chart does not clearly show the differences between the countries, instead a bar chart shows the differences more clearly.

Figure 27: Bar Chart Distribution of UK Exports Between EU, Rest of the World, USA, China, Japan and Canada



Source: Office for National Statistics

As a general rule, use a pie chart when there are large differences between values, and there are fewer than six categories.

Use a bar chart when there are more than five categories and you need to show small differences between values.

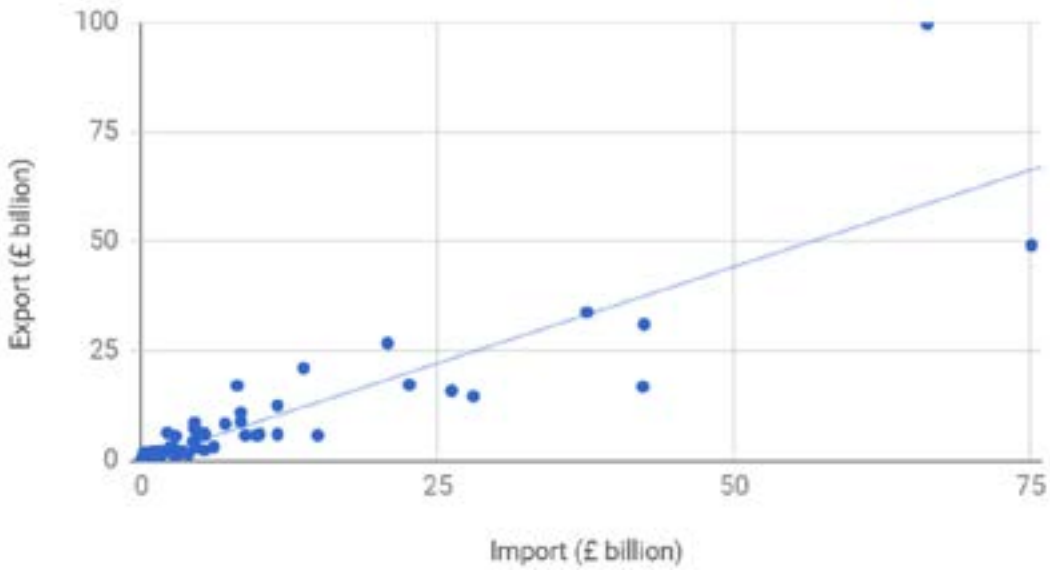
In your conclusion, you may want to show the relationship between things, to see whether they're **correlated**.

If things are positively correlated, they move in the same direction as each other – when one goes up, so does the other. If they're negatively correlated, they move in the opposite direction – when one goes up, the other goes down.

In our trade example, we've been focussing on exports. What if we add imports? Does the UK tend to import from and export to the same countries, or is there no relationship between the two?

We can use a scatter plot to show whether two things are correlated. In this chart, each dot represents a country, and we've plotted them according to UK exports and imports.

Figure 28: Correlation Between UK Imports and Exports



Source: Office for National Statistics

From this we can see that the UK imports and exports a lot from certain countries, and little from others. This suggests a positive correlation between imports and exports.

But remember that correlation doesn't equal causation.

Charting two things together just because they're correlated isn't always appropriate. The reader may assume that one causes the other, unless you tell them otherwise.

Look at the chart below – it shows a strong positive correlation between the divorce rate in the US state of Maine and consumption of margarine. But these things are totally unrelated, one certainly doesn't cause the other. For more information on correlation see the data analysis chapter.

Figure 29: Spurious Correlation

Visit tylervigen for more of these fun examples!

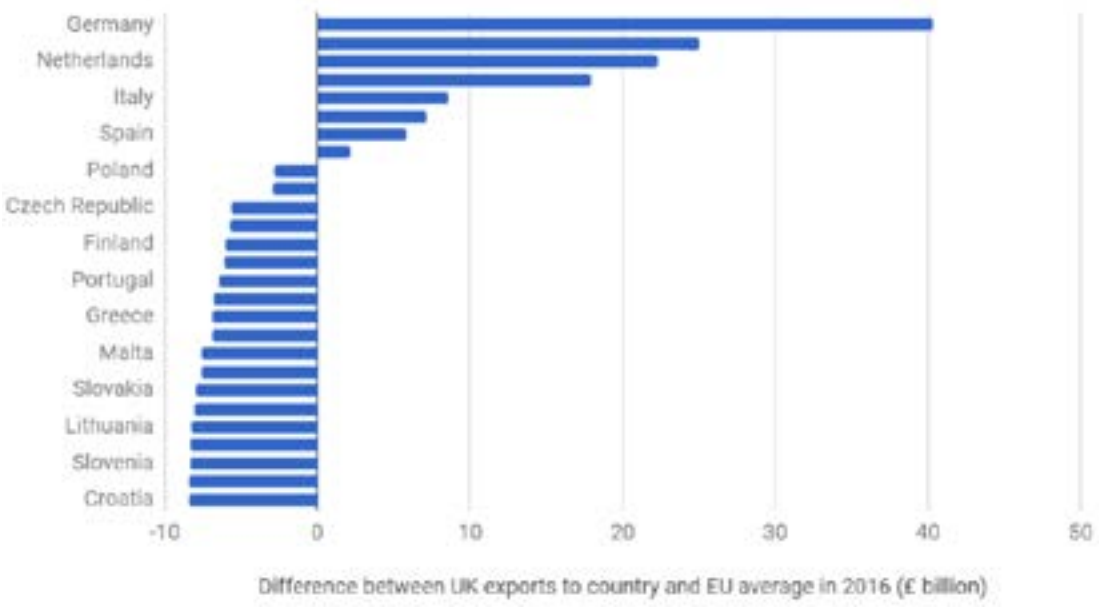
You can compare your variables by showing **deviation (spread)**.

Deviation is just the difference between a value and a fixed point, like zero or an average.

Using trade figures, you could compare UK exports to EU countries with the EU average.

Use bars to show deviation, and order the categories so that the chart is easier to read.

Figure 30: Difference Between UK Exports to Country and EU Average in 2016



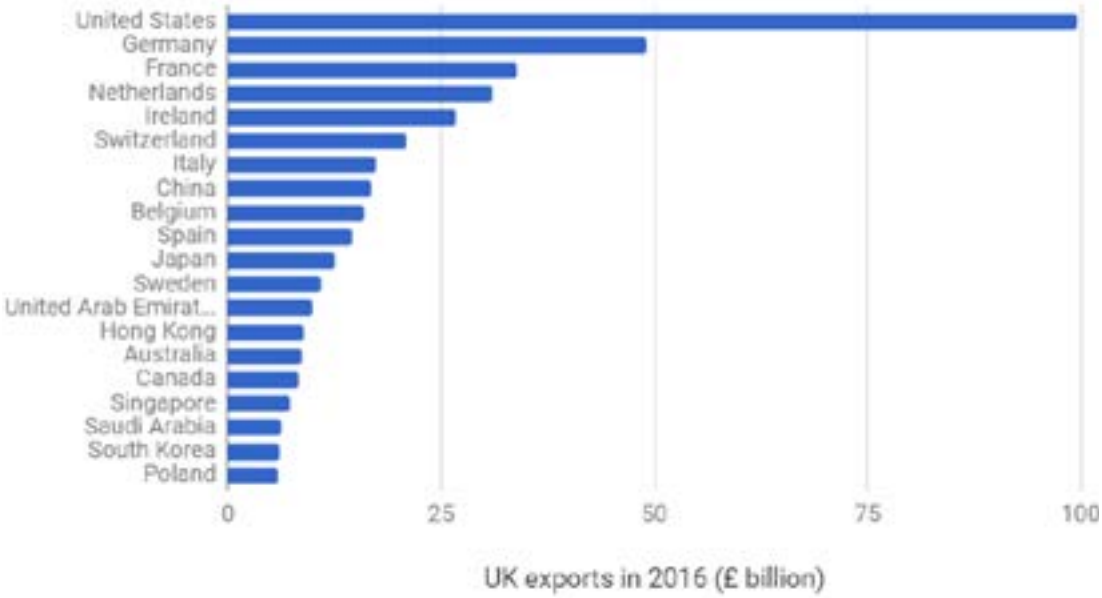
Source: Office for National Statistics

Similarly, **ranking** is an effective way of comparing different values.

Ranking is an item's position in an ordered list. If your class was ranked by Welsh Bacc score and you achieved the best score in the class, you would hold the top ranking. Sometimes an item's ranking may be more important than its value.

In our trade example, we could rank our top 20 export partners around the world. From this we can quickly see that we export most to the USA.

Figure 31: UK Exports to our Top 20 Trade Partners in £billion

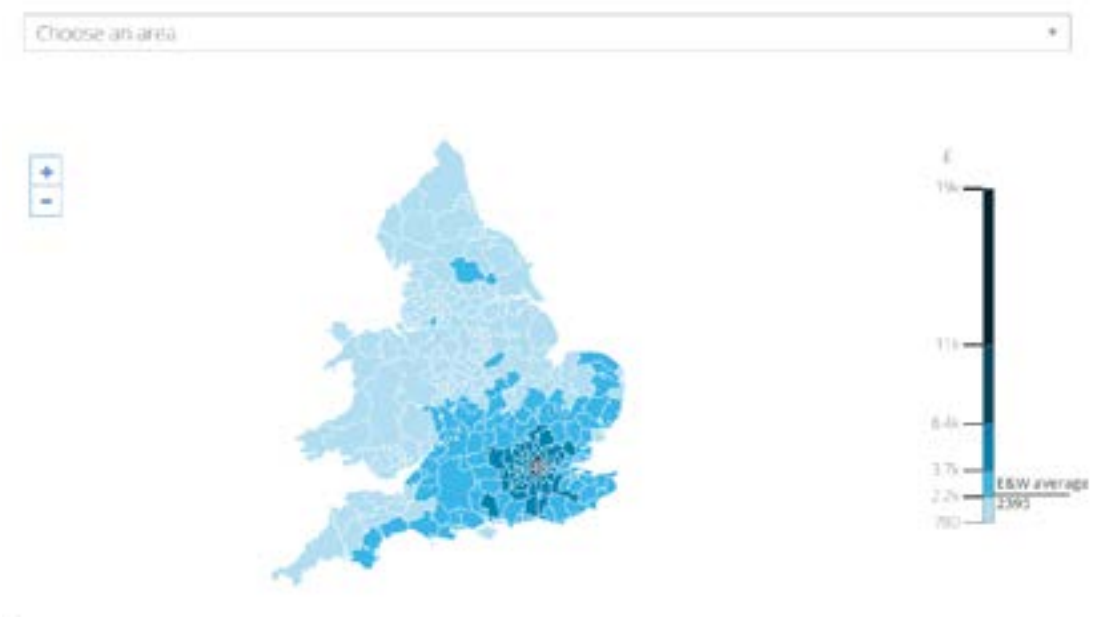


Source: Office for National Statistics

If you've got **spatial/geographical** data, you can use a map to show it.

Have a look at this map of house price per square metre in England and Wales.

Figure 32: Property Prices per Square Meter in England and Wales.



Source: House price per square meter and price per room, England and Wales, Office for National Statistics

A map is a great way of showing trends – you can clearly see from this that house prices rise in and around London.

ⓘ Beware of using maps for the sake of it though

Sometimes another type of chart will do a better job of showing the data (even if it looks less fancy!).

Here's an example showing change in public sector pay across different regions. Which do you prefer – the map or the bar chart?

Figure 33: Public Sector Employment Change Between 2008 and 2017

Source: Public Sector Employment, Office for National Statistics

What does ONS do?

As a regular publisher of articles, analysis and reports, ONS presents data in charts and tables all the time. ONS has published guidelines on chart type, titles and text and the use of tables.

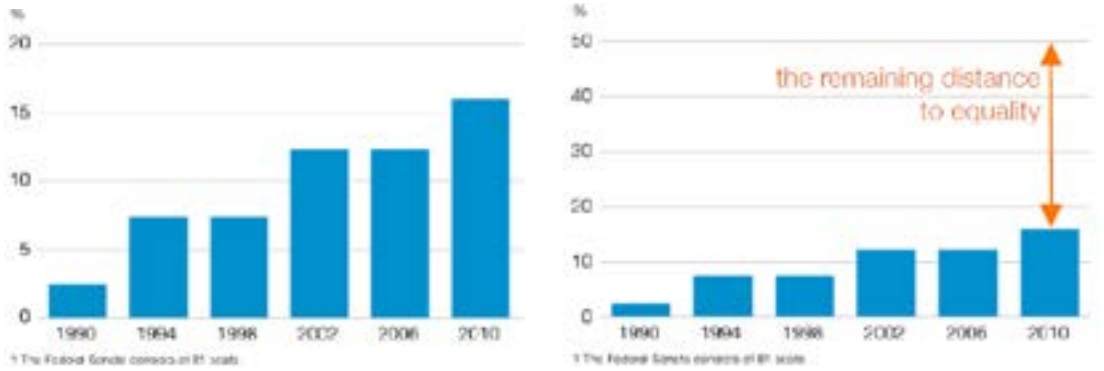
Using Annotation and Colour

Don't be afraid to use annotation and colour on your charts to help get your conclusion across.

There can be situations where a chart doesn't tell the full story on its own.

For example, the bar chart on the left shows the proportion of women in Brazil's senate. This is a perfectly good chart, but we can improve it using annotation.

Figure 34: Bar Chart Showing Female Representation in the Federal State

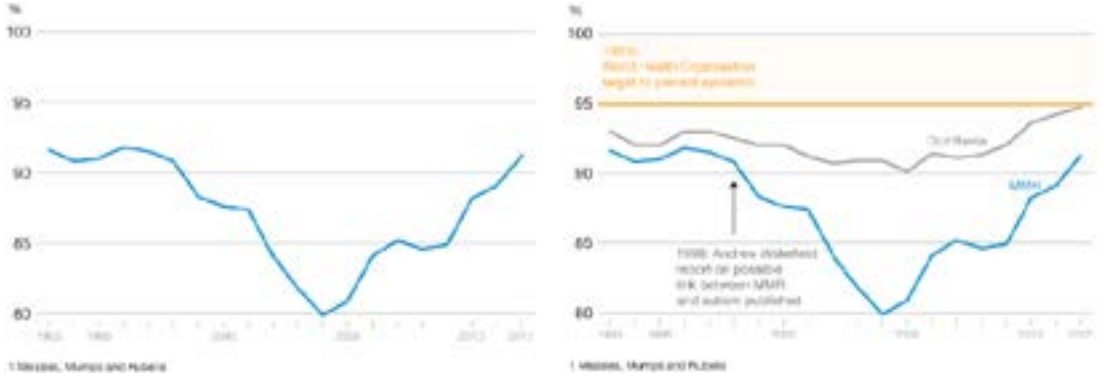


Source: Inter-Parliamentary Union

The underlying message is how far Brazil still has to go to reach equality, which is now clearly shown in the chart on the right.

Here's another example – the chart on the left shows the vaccination rate for the MMR jab. Again, there's nothing wrong with using a line chart to show change over time. But this doesn't say very much on its own.

Figure 36: Comparison Between Single Time Trend and with Relevant Comparison and WHO Target Rate

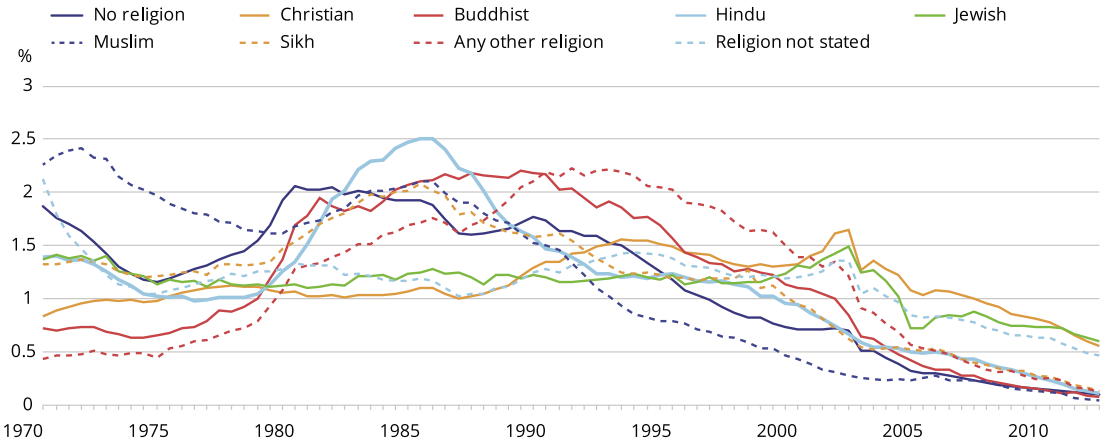


Source: NHSIC

On the right, we've compared MMR jabs with Diptheria, and provided extra context with the World Health Organisation target.

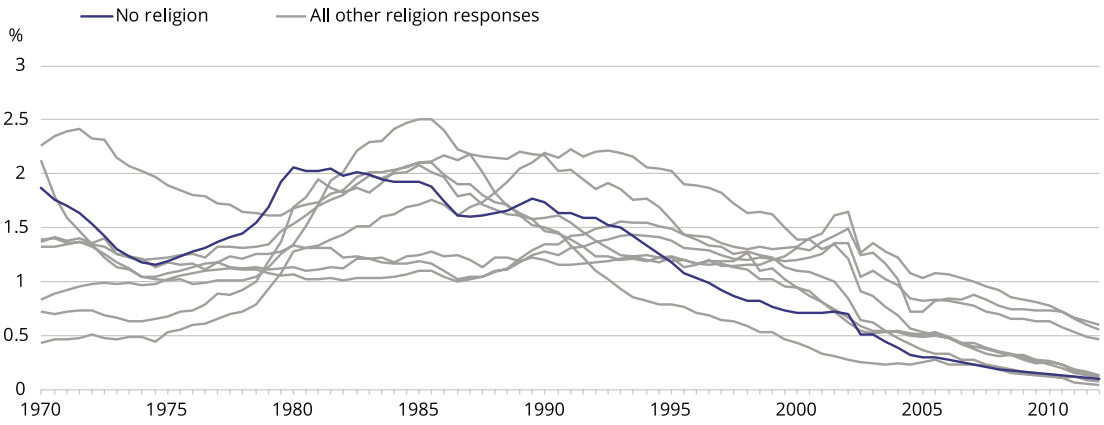
There are also situations where colour can be used to highlight a variable of interest. Take a look at the chart below.

Figure 37: Line Graph Showing Change in Religion in the UK Over Time.



It's very difficult to read because there's too much going on. But, you can use colour to highlight a particular trend.

Figure 38: Line Graph Showing Change in Religion in the UK Over Time, with Key Trend Highlighted



Using colour effectively doesn't mean using lots of bright colours. It's about making something readable and as easy to understand as possible.

What Does ONS Do?

ONS has published guidance on how to best use colour. Within these guidelines, there is also advice and links to further information about how to make your charts accessible. For example, this will ensure that someone who is colour-blind is able to read your chart as well as anyone else. It also takes into account consideration such as having to print charts in black and white, colour printing can be expensive, but you still need to be able to distinguish between colours!

Pitfalls to Avoid

Don't forget to label your graphs!

This includes giving them informative titles, and clearly labelling the axis.

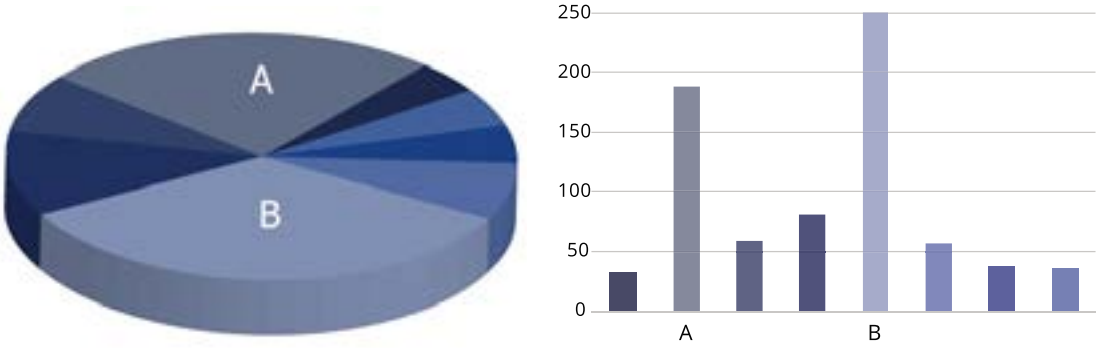
The most important rule when presenting data is don't misrepresent what your data is actually showing!

It's easy to get carried away when producing charts. You've got the freedom to do what you want with your chart – you can add colour, designs and even make it 3D.

(!) Try not to use 3D when creating charts. The false perspective will distort the data.

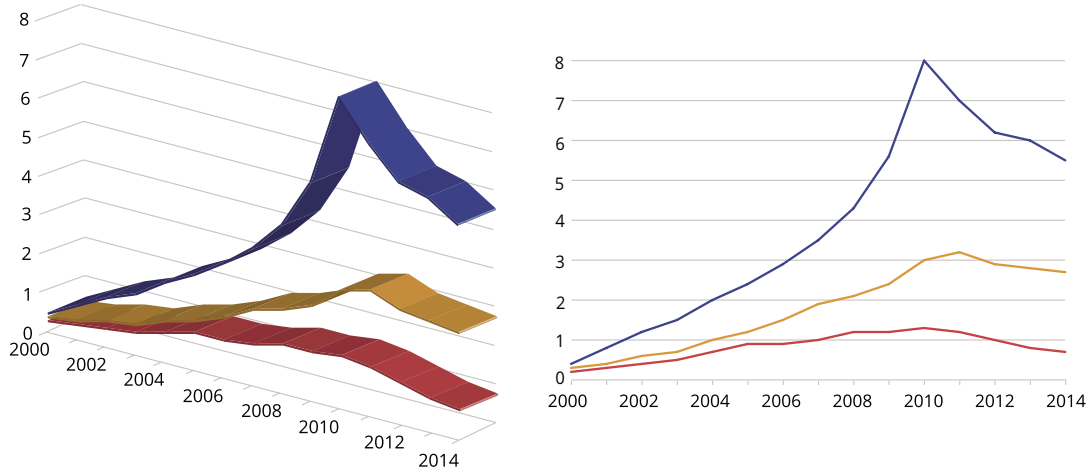
For example, categories A and B seem equal when plotted in 3D. However, category B is noticeably larger, as shown when plotted in 2D.

Figure 39: Illustration of how 3D Pie Charts Can Distort Values



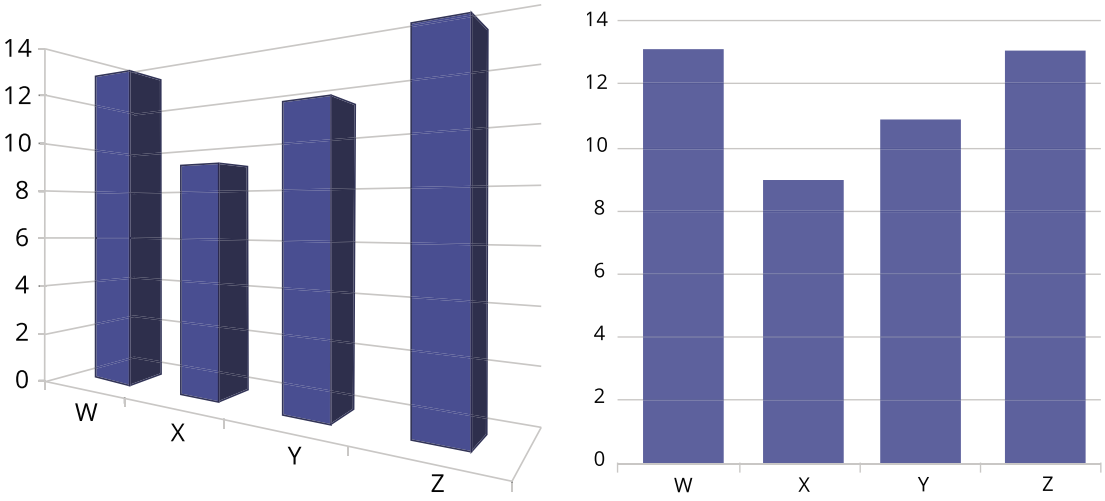
This doesn't just apply to pie charts. When plotted in 3D, the highest value appears to be around 7.9 in 2012. The same data in 2D clearly shows the highest value is 8 in 2010.

Figure 40: Illustration of How 3D Line Graphs can Distort Values



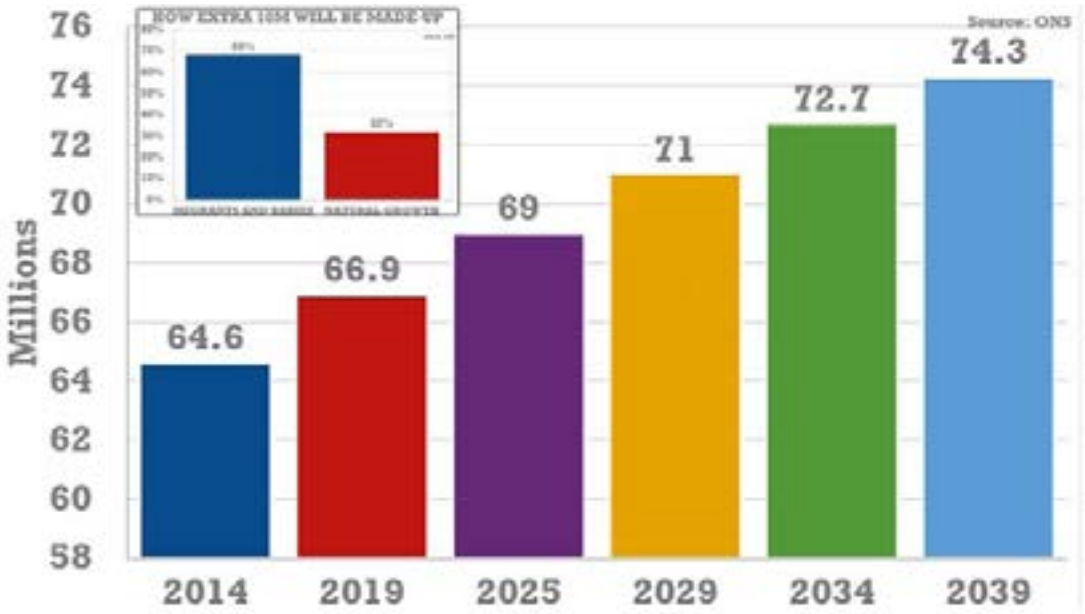
In 3D, bar W looks smaller than bar Z. However they're both equal.

Figure 41: Illustration of how 3D Line Graphs can Distort Values



Can You Spot any Issues with the Next Example?

Figure 42: problematic bar chart showing estimated population increase in Britain up to 2039



The shortened y-axis exaggerates differences between values – the bar representing 74.3 million is more than twice the size of the one showing 64.6 million. Where possible, avoid starting your y-axis above zero.

The bars are all coloured differently, for no reason. Don't use lots of bright colours for the sake of it; colour should be used sparingly to draw a reader's attention to a particular trend.

What Does ONS Do?

ONS has published advice on chart design, to ensure your reader understands the main messages in your chart.

LEARNING OBJECTIVES

By the end of this chapter students should feel able to:

- ✓ Understand which form of presentation is most appropriate for telling the story of their data.
- ✓ Recognise and understand how to interpret different forms of data presentation.
- ✓ Understand how data presentation can be adapted to different audiences.
- ✓ Recognise potentially misleading data visualisations.

Glossary

05

Abstract
A summary of the contents of a book, article, or speech

Accurate
Precise and representative of a true value

Aggregated
Data combined from several measurements

Aim
In intention to achieve a particular outcome

Appendices
Supplementary material at the end of a book, in more detail than is required in the body of text.

Causality
The relationship between cause and effect

Census
An official count or survey of an entire population

Conclusions
A judgement or decision reached by reasoning

Consent
Permission for something to happen or agreement to do something

Informed consent
Permission granted in full knowledge of the possible consequences, typically that which is given by a patient to a doctor for treatment with knowledge of the possible risks and benefits.

Correlated/Correlation
A mutual relationship or connection between two or more things.

Negative correlation
A relationship between two variables in which one variable increases as the other decreases, and vice versa

Positive correlation
Relationship between two variables in which both variables move together. i.e. One variable decreases as the other variable decreases, or one variable increases while the other increases.

Deviation
The amount by which a single measurement differs from a fixed value such as the mean

Discussion
A detailed breakdown of a topic in writing.

Gross Domestic Product (GDP)
Is a monetary measure of the market value of all the final goods and services produced in a period of time (as of May 2018 ONS moved from releasing this quarterly, to monthly!)

Hypothesis
A proposed explanation made on the basis of limited evidence as a starting point for further investigation.

Introduction
An opening to a piece of work which typically identifies the topic, arouses interest, and prepares the audience for the development of the thesis

Magnitude
A description of the size or scale of something

Normal distribution
An arrangement of a data set in which most values cluster in the middle of the range and the rest taper off symmetrically toward either extreme

Null hypothesis
A hypothesis that says there is no statistical significance between the two variables in the hypothesis. It is the hypothesis that the researcher is trying to disprove.

Population sample
Subset of subjects that is representative of the entire population

Proportion
The number or amount of a group or part of something when compared to the whole:

Qualitative
Qualitative data aims to give an understanding of underlying reasons, opinions, and motivations. It provides insights into the problem or helps to develop ideas or hypotheses for potential quantitative research. It can be used to uncover trends in thought and opinions, and dive deeper into the problem.

Quantitative
Numerical data or data that can be transformed into usable statistics. Quantitative data can be used to quantify attitudes, opinions, behaviours (to name a few) and generalize results from a larger sample. Quantitative Research uses measurable data to formulate facts and uncover patterns in research.

Quarter
Is one of the four three-month periods that make up a financial year. In the UK financial quarters do not line up with the calendar year, the financial year runs from 1 April to 31 March.

- 1st quarter: 1 April–30 June
- 2nd quarter: 1 July–30 September
- 3rd quarter: 1 October–31 December
- 4th quarter: 1 January–31 March

Ranking
Placing things in order based on a position in a hierarchy or scale

Rational for research methods
Provides the methods and procedures used in a research study or experiment. This allows other researchers to reproduce your experiment if they want and to assess alternative methods that might produce differing results.

Referencing
When referring to another's work you have to provide the original source of the information and give credit to the author.

Reliable
Overall consistency of a measure. A measure is said to have a high reliability if it produces similar results under consistent conditions

Results
Information obtained by an experiment or other scientific method.

Sampling error
Error in a statistical analysis arising from the unrepresentativeness of the sample taken

Selection bias
Selection bias is the bias introduced by the selection of individuals, groups or data for analysis in such a way that proper randomization is not achieved. This will mean that the sample is not representative of the intended population.

Significant
Likely that a relationship between two or more variables is cause by something other than chance.

Skewed
Data can be "skewed", meaning it tends to have a long tail on one side or the other. Data with no skew would be considered to be normally distributed.

Spatial data
Data or information that identifies the geographic location of features and boundaries. Spatial data is usually stored as coordinates and topology, and is data that can be mapped

Validate
To prove that something is correct

Variability/Variance
Gives a general idea of the spread of your data. Variability gives a way to describe how much data sets vary, and allows comparison to other data sets.

Appendices

06

Working out the Standard Deviation

The standard deviation is the square root (the number divided by itself) of the variance.

The variance is defined as:

The average of the squared differences from the mean.

To calculate the variance, follow these steps:

- work out the mean (the simple average of the numbers)
- then for each number: subtract the mean and square the result (the squared difference)
- then work out the mean of those squared differences

Example of Working out Standard Deviation

You have measured the heights of a group of people (in cms) and made sure that the sample is representative of the population (see section X for important information on sampling):

> The heights are: 132cm, 142cm ,144cm, 152cm, 153cm, 156cm, 160cm, 164cm, 168cm and 172cm.

Find out the mean, the variance, and the standard deviation.

Your first step is to find out the mean:

> Answer: Mean = 132 + 142 + 144 +152 + 153 + 156 + 160 + 164 + 168 + 17210 = 8045 = 154.3
>
> So, the mean = 154.3

To calculate the Variance, take each difference from the mean, square it, and then divide by the sample size – 1 (N-1):

$$= 497.3 + 151.3 + 106.1 + 5.3 + 1.7 + 2.9 + 32.5 + 94.1 + 187.7 + 313.39$$
$$= 154.7$$

So, the variance is 154.7

And the standard deviation is just the square root of Variance, so:

$$\sigma = 154.7$$
$$= 12.4 \text{ (to the nearest cm)}$$

We can now show which heights are within one standard deviation (12.4 cm) of the mean.

In this example, we have used a sample of the population therefore we have used N-1 when calculating the variance. If the whole population is used than we use only the sample size (N). The N-1 is the most commonly used because often statisticians/analysts have a sample that is representative of the population.

You can also use excel to work out the standard deviation by using the function "STDEV".

For the mathematical notation for standard deviation please see the annex.

Standard Deviation (Mathematical Notation)

The standard deviation of an entire population is known as σ (sigma) and is calculated using:

$$\sigma = (x - \mu)\sum N$$

Where x represents each value in the population, μ is the mean value of the population, ∑ is the summation (or total), and N is the number of values in the population.

The standard deviation of a sample is known as a S and is calculated using:

$$s = (x - x)\sum N - 1$$

Where x represents each value in the population, x is the mean value of the sample, ∑ is the summation (or total) and n -1 is the number of the values in the sample minus 1.

Steps to Calculating a Confidence Interval

The basic breakdown of how to calculate a confidence interval for a population mean is as follows:

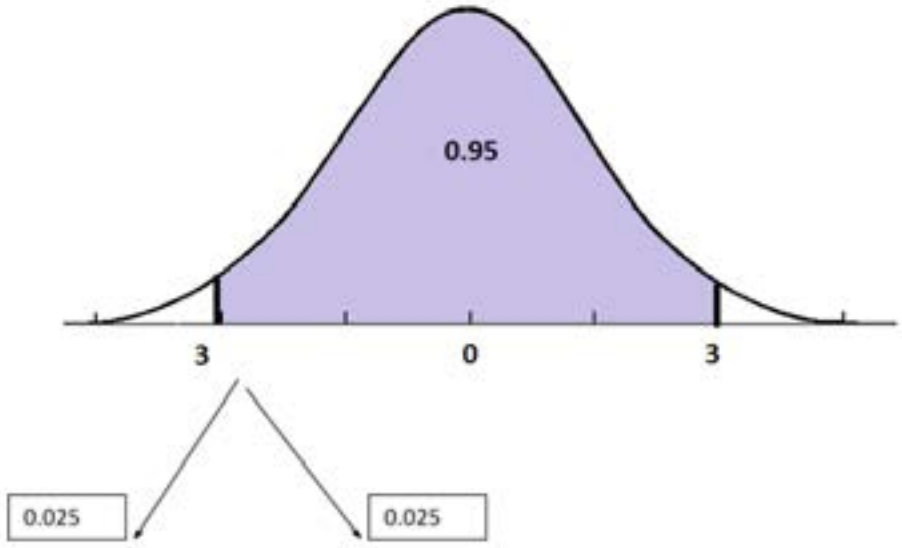
- identify the sample mean (divide the total of the sample by the number of data points (observations)), x
- identify whether the standard deviation for the whole population is known, σ, or unknown, s (see section on standard deviation for how to calculate)
- if standard deviation for the whole population is known then the z value is used as the critical value. A snapshot of a z table is shown below,

Table A: Standard Normal Probabilities (z table)

z	0	0.01	0.02	0.03	0.04	0.05	0.06
0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026
0.3	0.6179	0.6217	0.6255	0.6293	0.6333	0.6368	0.6406
0.4	0.6554	0.6561	0.6628	0.6664	0.6700	0.6736	0.6772
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7745
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315
1	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962
1.3	0.9032	0.9039	0.9066	0.9082	0.9099	0.9115	0.9131
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750
2	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846
2.2	0.9861	0.9864	0.9868	0.9871	0.9875	0.9878	0.9881

Let's suppose we want to use a z value for a 95% confidence level. By assuming a normal distribution (see figure X), we can work out the tails of the normal distribution by taking 0.95 away from 1 which equals 0.05.

Figure X: A normal Distribution With 95% of the Population Falling Within Three Standard Deviations From the Mean.



We then divide 0.05 by two to get the tails for both sides which equals 0.025. The area to the left of the required z value (shown by the arrow below) is 0.95 + 0.025 = 0.975.

So, we use the 0.975 figure to work out the z value from the table. We can see from Table A on page 96, that the z value is 1.9 + 0.06 so the z-value that we would use for a 95% confidence interval would be 1.96.

Once you have the z value you can then calculate the confidence interval.

Going back to our height example in the standard deviation section instead of sampling only 10 students we sample a whole class of 30 students. Say we want to work out the 95% confidence interval. We have worked out the standard deviation for the whole class which is 13.4 cm so we use the z values. For a 95% confidence interval, we have worked out previously that the z value is 1.96.

Firstly, we take the mean of the sample;

489930 = 163.3cm

Secondary we work out the standard deviation divided by the square root of the sample;

$$13.430$$
$$= 13.45.5$$
$$= 2.5$$

Next, we multiply 2.5 by the z value;

$$1.96 * 2.5 = 4.8 \text{ cm}$$

Finally, to get both sides of the confidence interval we subtract and add 4.8 cm from the mean;

$$163.3 - 4.8 = 158.5 \text{ cm}$$
$$163.3 + 4.8 = 168.1 \text{ cm}$$

So, we can say from this example we are 95% confident that the mean of the class falls within this range and the average person is between 158.5 cm and 168.1 cm tall.

If the population standard deviation is unknown and only a sample standard deviation is known or the sample size is below 30 then Student's t-distribution is used as the critical value (see table below).

This value is dependent on the confidence level (C) for the test and degrees of freedom (found in the first column of the table) is found by subtracting one from the number of observations, n – 1. The critical value is found from the t-distribution table.

For example, say we had a confidence level of 95% (this is the same as 0.95) and a sample size of 13, assuming a normal distribution we can work out the tails of the distribution by subtracting 0.95 from 1 which equals 0.05 (see figure X). We then divide 0.05 by two to

get each tail which equals 0.025. So, the degrees of freedom would be the sample size – 1 which equals 12 and we select 0.025 from the top row. The t-critical value that we would use in this example is highlighted by the green box and equals 2.17881.

df/p	0.4	0.25	0.1	0.05	0.025	0.01	0.005
1	0.32492	1	3.077684	6.313752	12.7062	31.82052	53.65674
2	0.288875	0.816497	1.885618	2.919886	4.30265	6.96456	9.92484
3	0.276671	0.764892	1.637744	2.353363	3.18245	4.5407	5.84091
4	0.270722	0.740697	1.533206	2.131847	2.77645	3.74695	4.60409
5	0.267181	0.726887	1.475884	2.015048	2.57058	3.36493	4.03214
6	0.264835	0.717558	1.439756	1.94318	2.44691	3.14267	3.70743
7	0.263187	0.711142	1.416924	1.414924	1.894579	2.36462	3.49948
8	0.261821	0.706387	1.396815	1.396815	1.859548	2.306	3.35539
9	0.260955	0.702722	1.383029	1.383029	1.833113	2.26216	3.25984
10	0.260185	0.699812	1.372184	1.372184	1.812461	2.22814	3.16924
11	0.259556	0.697445	1.36343	1.36343	1.795885	2.2089	3.10581
12	0.259033	0.695483	1.356217	1.356217	1.782288	2.17881	3.05454
13	0.258591	0.693829	1.350171	1.350171	1.770833	2.16037	3.01228
14	0.258213	0.692417	1.34503	1.34503	1.76131	2.14479	2.97684

Once you have the t-critical value you can work out the confidence interval by using the same method as explained before but using the t-critical value instead of the z-value.

Once the confidence interval is calculated we can see if the true mean of the population lies in the interval. If this is the case, we can state "There is a _% (depending on level of confidence) probability that the true mean of the population lies in the _% confidence interval".

Researchers usually present the 95% confidence intervals, although other intervals are possible (for example, 99% or 90% confidence intervals are sometimes presented).

Confidence Intervals (Mathematical Notation)

For a known standard deviation;

$$x - z\,\sigma n\ ,\ x + z\,\sigma n$$

For an unknown population standard deviation or sample size less than 30;

$$x - t\,sn\ ,\ x + t\,sn$$

Where x is the mean value of the sample, z is the z value, σ is the standard deviation of the whole population, t is the t-critical value, s is the sample standard deviation and n is the sample size (number of observations).