

Home  
Overview  
(/study/app)  
122-  
cid-  
754029/k

TOPIC 4  
PROBABILITY AND STATISTICS



(https://intercom.help/kognity)



SUBTOPIC 4.1  
COLLECTION OF DATA AND SAMPLING

4.1.0 The big picture

4.1.1 Data types and sources

Table of  
contents

4.1.2 Sampling

4.1.3 Checklist

Notebook

4.1.4 Investigation

Glossary

6d

Reading  
assistance

(x)  
Student  
view



Show all topics





Overview  
(/study/app  
122-  
cid-  
754029/k

4.1 Teacher view

## Collection of data and sampling

### Index

- The big picture
- Data types and sources
- Sampling
- Checklist
- Investigation

4. Probability and statistics / 4.1 Collection of data and sampling

# The big picture

Statistics is probably the most used and least understood branch of mathematics in our world today. Statistics is the study of what data says. Anything that you can measure or count is data, and statistics provides the tools you can use to generalise huge amounts of data and predict future outcomes with a high degree of accuracy. Statistics also provides a variety of tests and checks to help you make sure that your generalisations and predictions are valid. The beauty of statistics is that you can apply it to nearly any context, which is why statistics experts are highly sought after to exercise their skills in economics, politics, scientific research, sports and many other industries.

While statistics may well be the most practically important topic we study in this course for almost everybody, this does not detract from other areas of mathematics, in particular calculus, which is the bread and butter in different professions. And it says nothing about the beauty of the various branches of mathematics. However, the case for a good grounding in statistical concepts is solid. See, for example, this [TED talk by Arthur Benjamin](#)

([http://www.ted.com/talks/arthur\\_benjamin\\_s\\_formula\\_for\\_changing\\_math\\_education#t-157800](http://www.ted.com/talks/arthur_benjamin_s_formula_for_changing_math_education#t-157800)).

### Making connections

Think about how much you use data in your other Diploma courses. Among numerous other applications, you analyse:



Student  
view

- census information, demographics and economic data in the human sciences



Overview  
(/study/app  
122-  
cid-  
754029/k

- media ratings and sales data to determine what is popular in the arts
- experimental data and measurements in the natural sciences.

Visualisation is a big component in statistics, even more so in this era of what some have called 'big data'. If you want to see some pleasing examples of this, go through this playlist on [TED](#) ↗  
([http://www.ted.com/playlists/56/making\\_sense\\_of\\_too\\_much\\_data](http://www.ted.com/playlists/56/making_sense_of_too_much_data)).

## 🔑 Concept

In this subtopic you will use statistics to compile data into charts and graphs that make it possible to recognise **patterns** and draw **generalisations** about the data. How can you use statistics to show that your conclusions are **valid** ?

## ❖ Theory of Knowledge

The writer Mark Twain popularised the phrase ‘There are three kinds of lies: lies, damned lies, and statistics.’ Statistics are powerful in that they can turn quantitative numbers into qualitative understanding. However, as pointed out by Twain and in the video below, they can also mislead. It is important to be precise in the **scope** and **application** of statistics.

This leads to a knowledge question, ‘Given that statistics can provide multiple truths depending on numbers and analysis selected, is a singular truth possible or is all truth a matter of the knower’s **perspective** ?’



Student  
view

Home  
Overview  
(/study/app/  
122-  
cid-  
754029/)

## Ben Goldacre: Battling Bad Science



4. Probability and statistics / 4.1 Collection of data and sampling

# Data types and sources

Section

Student... (0/0)

Feedback



Print (/study/app/m/sid-122-cid-

Assign

754029/book/data-types-and-sources-id-26223/print/)

## Collecting and categorising data

Statistical analysis is the **analysis of data**. Part of statistics is the collection and organisation of data that describe certain characteristics of a group, called the **population**. We generally call this **descriptive statistics**. Often the population we want to study can be quite large, making it difficult or even impossible to collect data for the entire group. In these instances, we collect a **sample**. The sample should represent the population and its characteristics. Ideally, a sample is obtained by randomly selecting members of the population. You will learn more about sampling in section 4.1.2 (/study/app/m/sid-122-cid-754029/book/sampling-id-26224/).

In practice, it is not easy to ensure that your sample is representative of the population, and several methodologies have been developed to create representative samples. Here are two cases of rather bad sampling:



Student  
view



- Conclusions drawn about the height of students of a particular school by measuring the heights of the basketball team are likely to be false.
- Predictions about a pending political vote based on making random calls between 9 a.m. and 11 a.m. and asking people who they will vote for are also likely to be false, as not everybody can answer calls between these times.

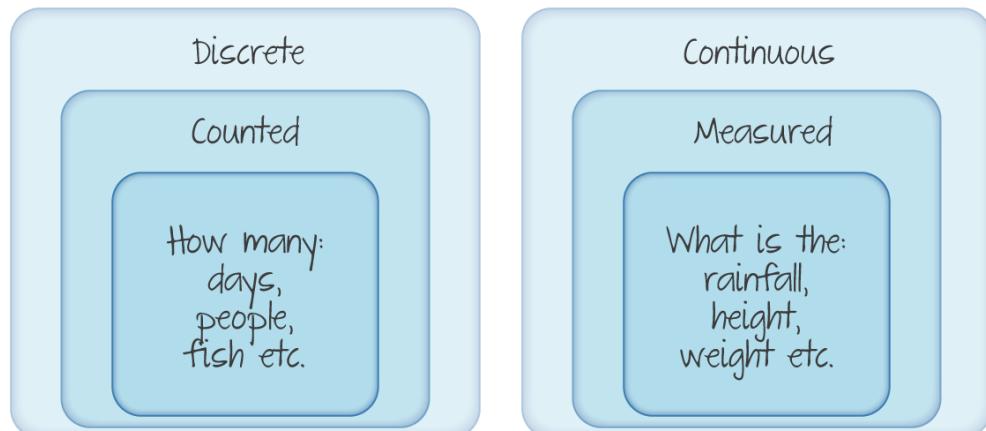
When the sample data is well chosen and described, an analysis will allow us to draw conclusions about the population based on this sample. We call this making **statistical inferences**.

### ❗ Exam tip

For examination purposes in Papers 1 and 2 , the given data will be treated as the full population and not as a selected sample of the population.

Not all data is numerical. Categorical data is based on qualitative observations, such as eye colour. Numerical data can be divided into two groups: discrete and continuous. The rule of thumb is:

1. Discrete data is data that can be *counted* ; it gives the number of times something occurs or the number of items that exist. For example, you could ask 'How many times was the bus late this month?'
2. Continuous data is data that is *measured* . However, the values of the actual data cannot be determined exactly, and the data may be limited to a range. Rather than asking 'Is the donut I bought weigh 155g?', you could ask 'Is the donut I bought weigh between 153g and 157g?'





Overview  
(/study/app  
122-  
cid-  
754029/k

## Discrete and continuous data.

More information

The image is a diagram comparing discrete and continuous data using two nested rectangles. On the left side, it illustrates 'Discrete' data labeled with 'Counted' and provides examples like "How many: days, people, fish etc.". On the right side, it illustrates 'Continuous' data labeled with 'Measured' and includes examples like "What is the rainfall, height, weight etc." The diagram visually differentiates between the two data types with text explanations inside the shapes.

[Generated by AI]

### ✓ Important

The distinction in statistical questions between discrete and continuous data is important: discrete data may take on exact values, while continuous data lies in a range. We will return to this when studying discrete and continuous random variables and their probability distributions.

As you collect data there are two questions that you can ask yourself:

1. Are you getting data that is accurate?
2. Is there anything unusual about the data you have found?

## Reliability

The first question is primarily concerned with the **reliability** of your data. Reliable data is free of errors and bias. While no research produces data that is perfect, reliable data comes from individuals who do their best to perform their research carefully, responsibly and without attempting to manipulate the results to support a particular conclusion. Data can be unreliable when individuals are careless when collecting or recording information.



Student  
view

Bias is another type of unreliability. Consider an online advertisement that says, ‘What do you think of Politician X’s plan to combat terrorism? Click here to complete a survey.’ In this example, bias can occur accidentally because people are more likely to participate if they have a very strong opinion about that politician or about the issue. This means that a large portion of a population probably goes unstudied because they simply do not care enough to take the survey. Bias could also be more blatant if this survey were only advertised on a website run by supporters of Politician X or if the question were worded like this: ‘What do you think of trusted Politician X’s clever plan to combat terrorism?’

### Be aware

When collecting data from other sources, especially online sources, you must consider whether the individuals providing the data might be trying to promote a particular agenda. If they are, that doesn’t automatically make it unreliable data, but it should make you be more careful when using it to draw conclusions.

## Outliers

The second question is concerned with the concept of outliers . Outliers are data values that are very different to the rest of the data. They can occur for a number of reasons. An outlier might simply be a naturally occurring extraordinary value, such as a student who did not study scoring 20% on an exam. On the other hand, it may be a value that is the result of abnormal circumstances, such as a runner who takes over a minute to finish the 100 m dash because they twist an ankle and limp to the finish line. You will learn how to identify outliers and show them on a box-and-whisker plot in subtopic 4.2 (/study/app/m/sid-122-cid-754029/book/the-big-picture-id-26227/).

### Be aware

Some outlying data items may be an error in the sample. The context of the data is important in deciding how to interpret it.

Reliable data can be very challenging to find, especially when it might make one individual or group look bad compared to another. For example, the leadership of a particular country might be tempted to report skewed data about poverty levels, unemployment and voter participation. This is why there are independent organisations that collect objective data about a variety of factors worldwide.

One such organisation is the World Health Organization (<https://www.who.int/gho/en/>), which provides data concerning numerous health-related issues. Other organisations that are not independent have also produced resources that have been generally recognised as reliable, like the United States CIA publication The World Factbook (<https://www.cia.gov/the-world-factbook/>), which contains updated information and data for 267 world entities.

## 3 section questions ▾

4. Probability and statistics / 4.1 Collection of data and sampling

# Sampling

Section

Student... (0/0)

 Feedback

 Print (/study/app/m/sid-122-cid-

754029/book/sampling-id-26224/print/)

Assign

## Sampling data from a population

### When collecting all the data is not feasible

Researchers often explore problems related to a population that is extremely large. This makes it very difficult, or even impossible, to collect data about every member. Thankfully, statistics allows us to draw very accurate conclusions about a population simply by collecting data about a subset of the group, called a sample. The practice of selecting this subset is called **sampling**.

One example of the difference between a sample and the population is seen when governments hold an election. Numerous researchers ask people who they will vote for as the election day approaches, but they clearly cannot ask every voter. They use a

Home  
Overview  
(/study/app  
122-  
cid-  
754029/k

sample. On election day, the government collects the official votes from the entire voting public (or at least those who choose to vote). They are collecting data from the entire population.

Another example of sampling could be seen in a study of the impact of tainted water on the people of a particular region of the world. It would be impossible to find and speak with every person who lives in the region, let alone get their consent to run medical tests on them to collect data for the entire population you want to study. However, you could travel throughout the region to find a sample of people who vary in age, size and sex in order to test the effects on a representative cross-section of the population.

## How large a sample do you need?

There are some rules for how large a sample should be for specific statistical tests, but it is hard to generalise. As long as the sample represents the relevant features of the full population, it is a good sample. The larger the sample, the better it represents the full population. There is no set rule regarding how big your sample should be. However, you will notice that on reading a published statistical study, it will probably state how much data was collected and the method of collection. It seems a little unfair, but it is unlikely that you will be criticised for having too much data . However, you may be criticised for not having enough data.

What can make this challenging is that a data set of 50 entries or more will be sufficient for one type of statistical study, while another type might require a data set of at least 100 data entries to yield an informative test and another might require a data set of 200. As you learn about each type of study, pay attention to acceptable sample sizes and how the sample size can affect the reliability of results.

### ! Exam tip

If you are collecting data for your IA (internal assessment), it is important to explain whether your data represents a sample or the population. If it is a sample, you should explain how you chose it. This is a great opportunity to demonstrate personal engagement and incorporate critical reflection.

X  
Student  
view



## Sampling methods

Overview

(/study/app  
122-  
cid-  
754029/k)

Researchers can choose from different methods to select a sample. The choice is important, as sampling can be the determining factor of whether conclusions are valid or invalid and whether research is biased or objective. It is important that your sample represents the population, or you may not be able to draw valid conclusions about the population. For example, if you want to determine what students in your school think about a certain issue, it would not be appropriate to construct a sample entirely from your own year group. That would be an example of sample bias because it would over-represent the opinions of your year group and under-represent the opinions of other year groups in the population.

One characteristic that every good sample has is that it is **random**. This means that each potential data point has the same probability of being chosen. If I wish to measure the heights of all the year 12 students, then each student in year 12 should have an equal probability of being chosen for the survey. I won't, for example, deliberately choose the really tall person and the really small person – they would be chosen only if they are picked by a random process.

There are several ways to select your sample. Some are shown in the table below.

Types of sampling.

Type	Description	Examples
Simple	Achieving randomness by a simple, completely random process.	You choose items out of a hat or you use a random number generator to pick items from a list.
Convenience	Choosing a sample based on how easy it is to find the data.	You ask the first twenty people who walk past you to answer your survey.
Systematic	If data is listed, selecting a random starting point and then choosing the rest of the sample at a consistent interval in the list.	You roll a die and get a 6, so you start with the 6th item and then choose every 10th item in the list after that.

Student  
view

Type	Description	Examples
Quota	Choosing a sample that is only comprised of members of the population that fit certain characteristics.	You want a sample of 50, but they must all be ladies between the ages of 16 and 25. The selection is not random, a convenience method can be used.
Stratified	Choosing a random sample in a way that the proportion of certain characteristics matches the proportion of those characteristics in the population.	The population is 45% male and 55% female, so you make a random sample of 100 people that has 45 men and 55 women in it.

## Activity

Brainstorm with your classmates the ways in which each type of sampling could introduce bias or avoid it. In which kinds of research is it particularly important to avoid bias?

## Theory of Knowledge

Consider how sampling connects with the ethics. How can a researcher construct a sample ethically or unethically? What implications does sampling have on the validity of research?

## 1 section question

4. Probability and statistics / 4.1 Collection of data and sampling

# Checklist

## Section

Student... (0/0)

 Feedback

 Print (/study/app/m/sid-122-cid-

754029/book/checklist-id-26225/print/)

 Assign



Student  
view



Overview

(/study/app

122-

cid-

754029/k



## What you should know

By the end of this subtopic you should be able to:

- identify whether data is discrete or continuous
- recognise what factors can make a data source reliable or biased
- identify possible outliers in a data set
- list different types of random sampling
- understand how to create a sample that is representative of the population.

4. Probability and statistics / 4.1 Collection of data and sampling

# Investigation

---

**Section**

Student... (0/0)

Feedback



Print (/study/app/m/sid-122-cid-

754029/book/investigation-id-26226/print/)

**Assign**

Even when sampling is done correctly, how you interpret and apply the data is another area in which ethical lines are often crossed. In relation to natural science, neuroscientist Dr Molly Crockett discusses how the media like to make definitive claims about various products and lifestyle habits that are supported with ambiguous research.

Molly Crockett: Beware neuro-bunk

Student  
view

Overview  
(/study/app  
122-  
cid-  
754029/k

After watching the video, explore a variety of resources to see if you can find other medical claims, not related to neuroscience, that have been proven false. See if you can find an example in which the data was valid but those making the claim were misusing it. Did the data they had point to a different conclusion? Was the data simply inconclusive?

---

### Rate subtopic 4.1 Collection of data and sampling

Help us improve the content and user experience.



Student  
view