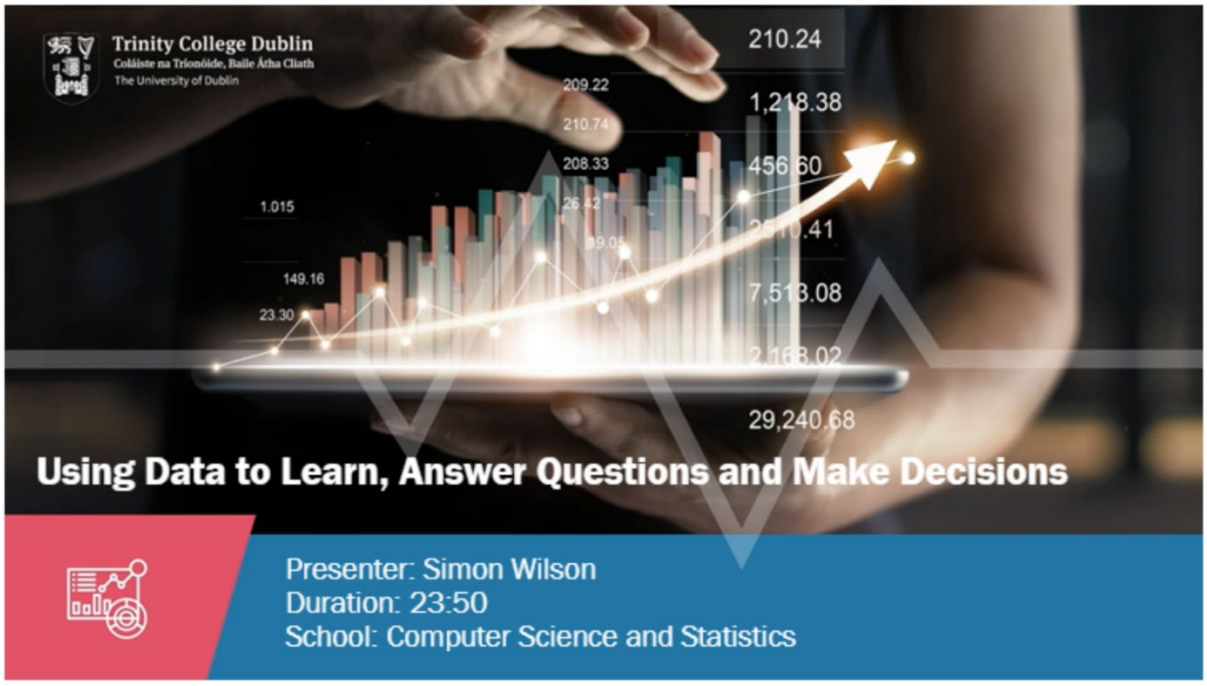
Using Data to Learn, Answer Questions and Make Decisions

Contents

Slide 1: Introduction



Welcome to this presentation on “Using data to learn, answer questions and make decisions”. My name is Simon Wilson and I will lead this presentation.

Before we start to look in detail at different statistical methods, in this presentation we take a step back and consider the place of statistical methods in the broader context of using data to answer questions and make decisions. We also look at some issues that permeate almost all statistical methods – such as causality versus correlation, and population versus sample. We conclude with a look at the fundamental reason why we have to use statistical methods at all, that is to say the existence of random variation.

Slide 2: Section 1: The Statistical Process

The world's oldest professional body for the statistics community, The Royal Statistical Society, states "Our vision is a world where data are at the heart of understanding and decision-making". This quote encapsulates the purpose of statistical methods. At their core is the idea to turn data, of whatever form, into knowledge. That knowledge can be in the form of providing evidence for a hypothesis or about the value of a quantity of interest. Built around this central notion are several other issues that lead in and out of that core idea, as we see next.

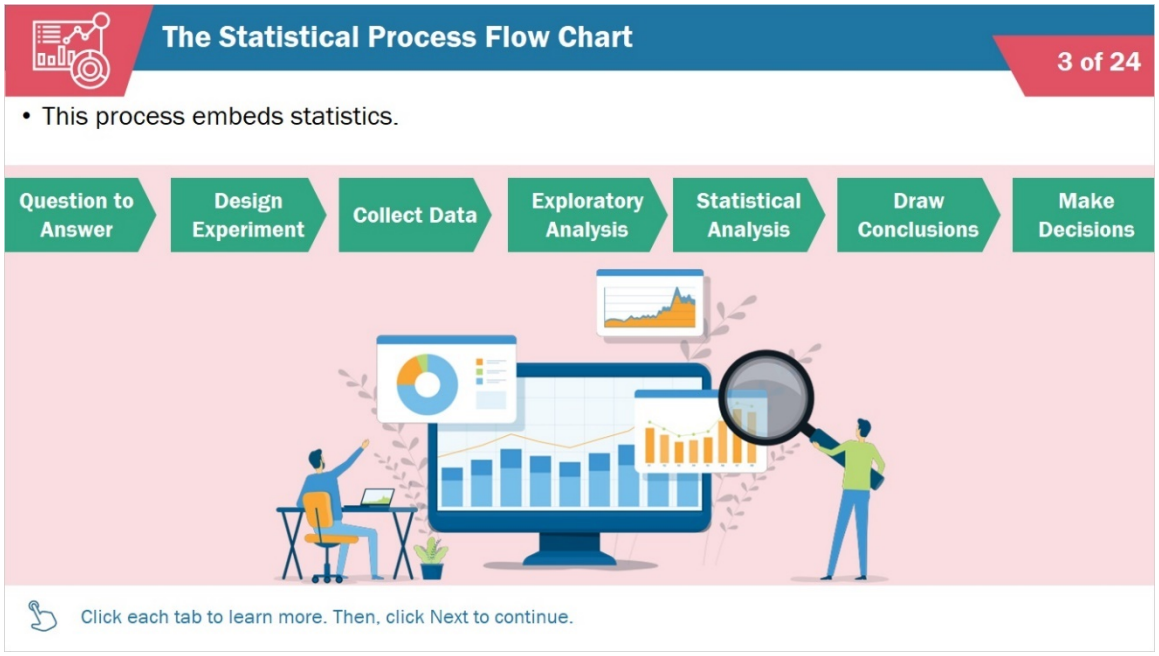
Reference(s):

1. Royal Statistical Society. (2021). RSS - What we do. https://www.rss.org.uk/about/what-we-do-(1)/

Image(s):

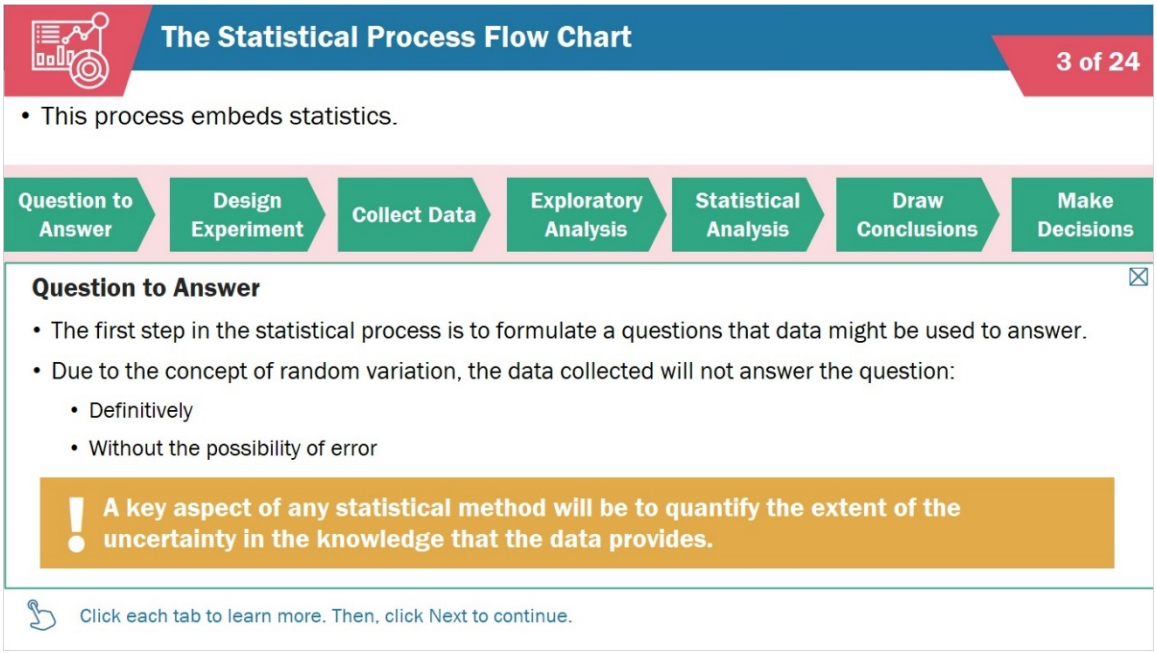
1. Royal Statistical Society. (2021). Royal Statistical Society Logo. [Illustration]. https://www.rss.org.uk/

Slide 3: Statistical Process Flow Chart



Here is a representation of the process that embeds statistics. Click the tabs to go through each stage of the process in turn. When you are ready, click Next to continue.

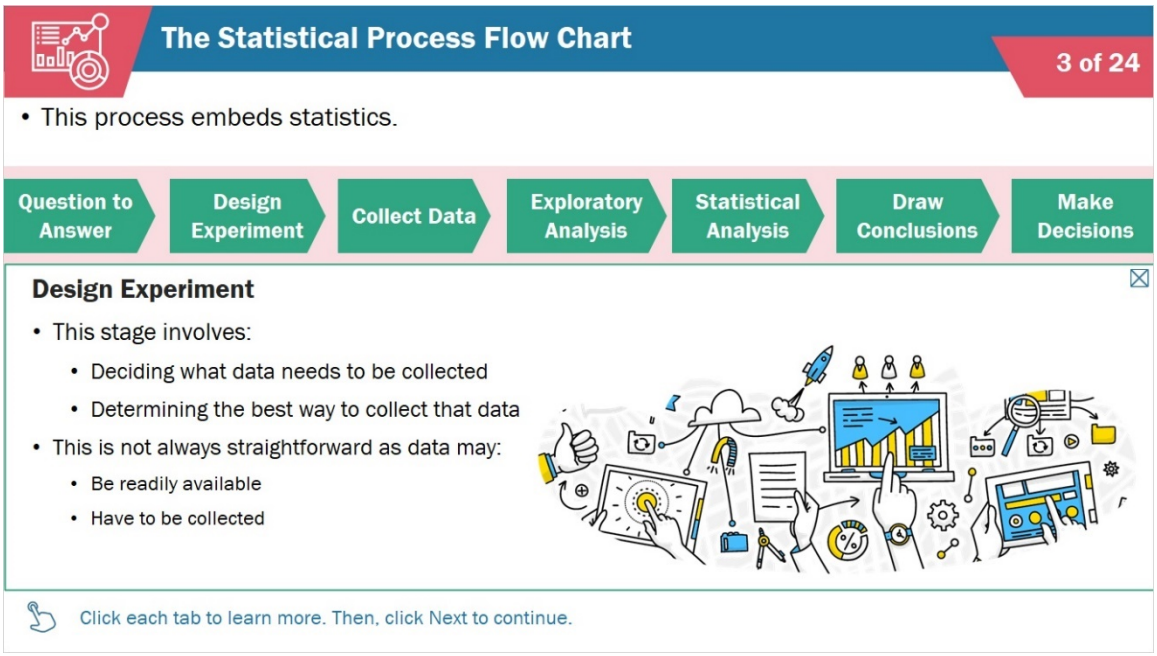
Tab 1: Question to Answer



The first step is to formulate a question that data might be used to answer, or at the least give some evidence towards answering. The concept of random variation, also known as chance variation, means roughly, that real data exhibit random and unpredictable variation. We will discuss this later, however this concept means that almost always, the data that we collect will not answer the question definitively and

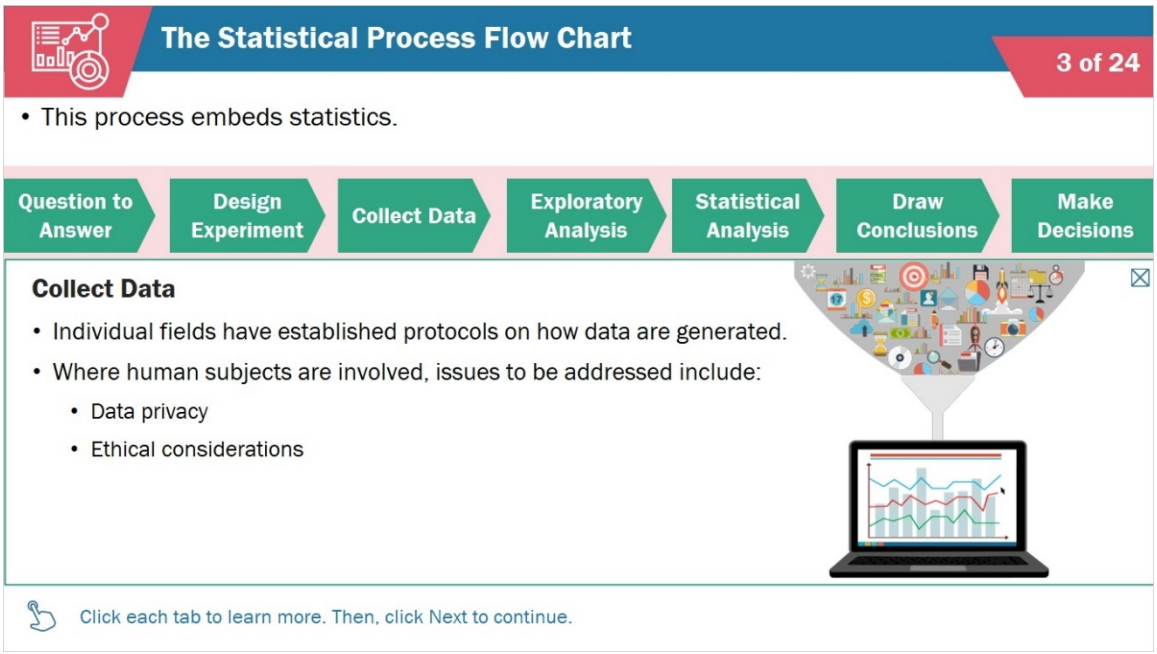
without the possibility of error. A key aspect of any statistical method will be to quantify the extent of the uncertainty in the knowledge that the data provides.

Tab 2: Design the Experiment



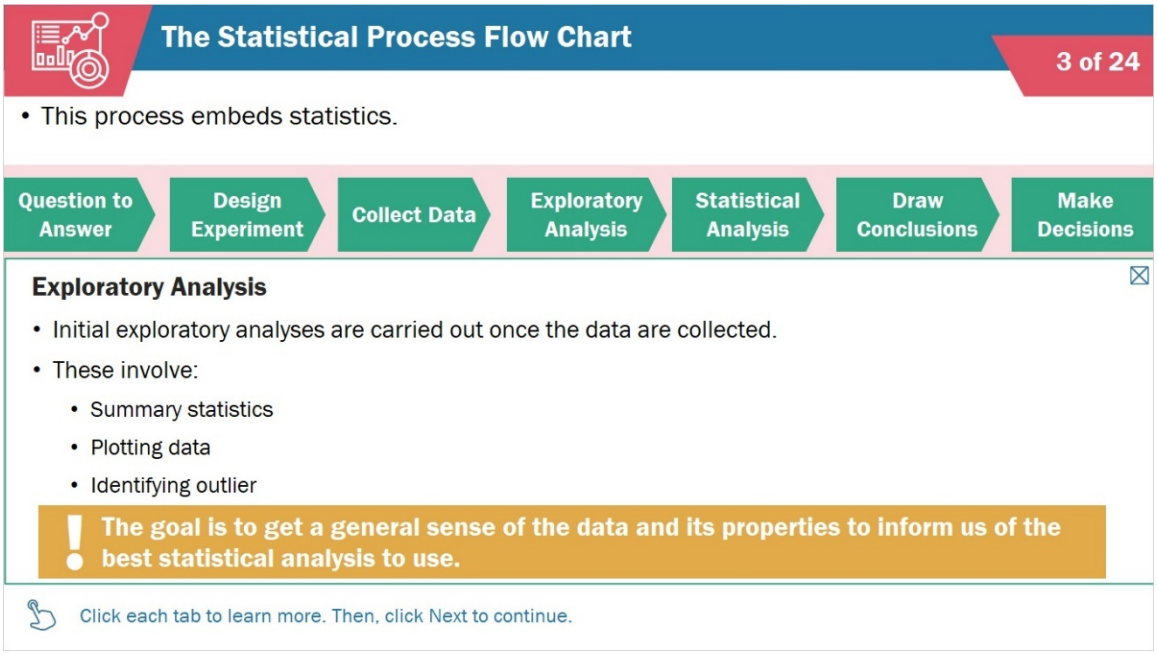
Step two is to design the experiment. This is often an overlooked part of the process. However, in essence, what it involves is deciding what data needs to be collected and then determining the best way to collect that data to answer your question. In some cases, this might be straightforward but often it is not. The data may be readily available or, it will have to be collected. Later in this presentation, we will look at the use of data that has been collected in a rigorous way versus data that has not, for example data that has already been collected for some other purpose.

Tab 3: Collect Data



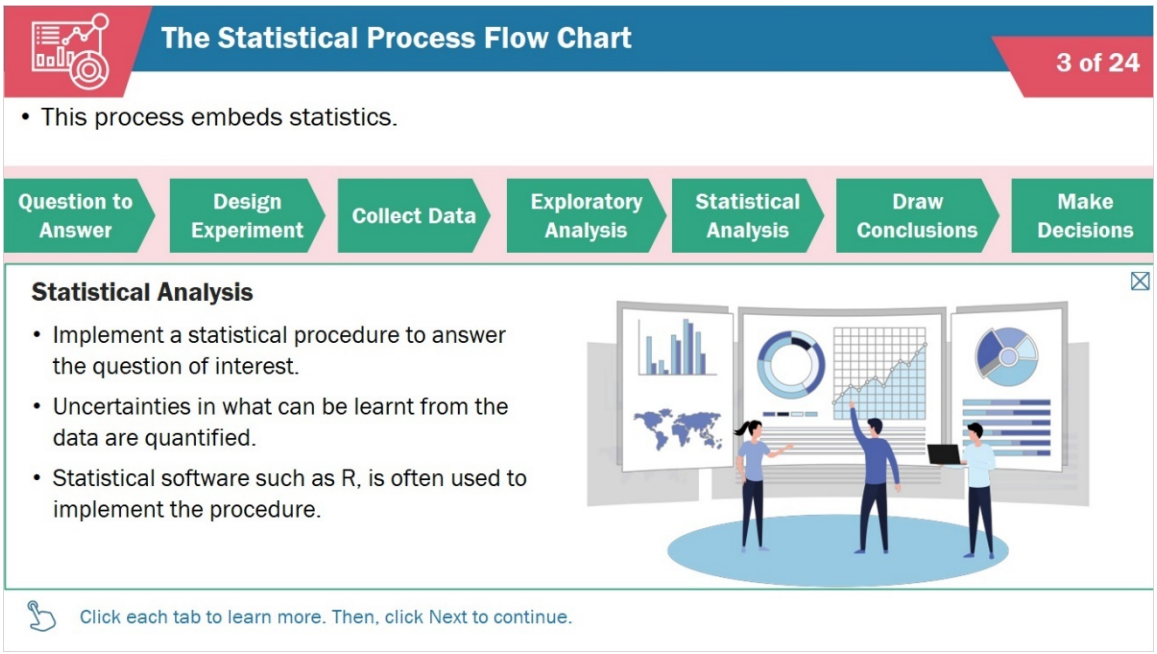
The third step is to collect data. In many fields, there are well-established protocols on how data are generated. This is of course particularly the case where human subjects are involved, with issues of data privacy and other ethical considerations that must be addressed.

Tab 4: Exploratory Analysis



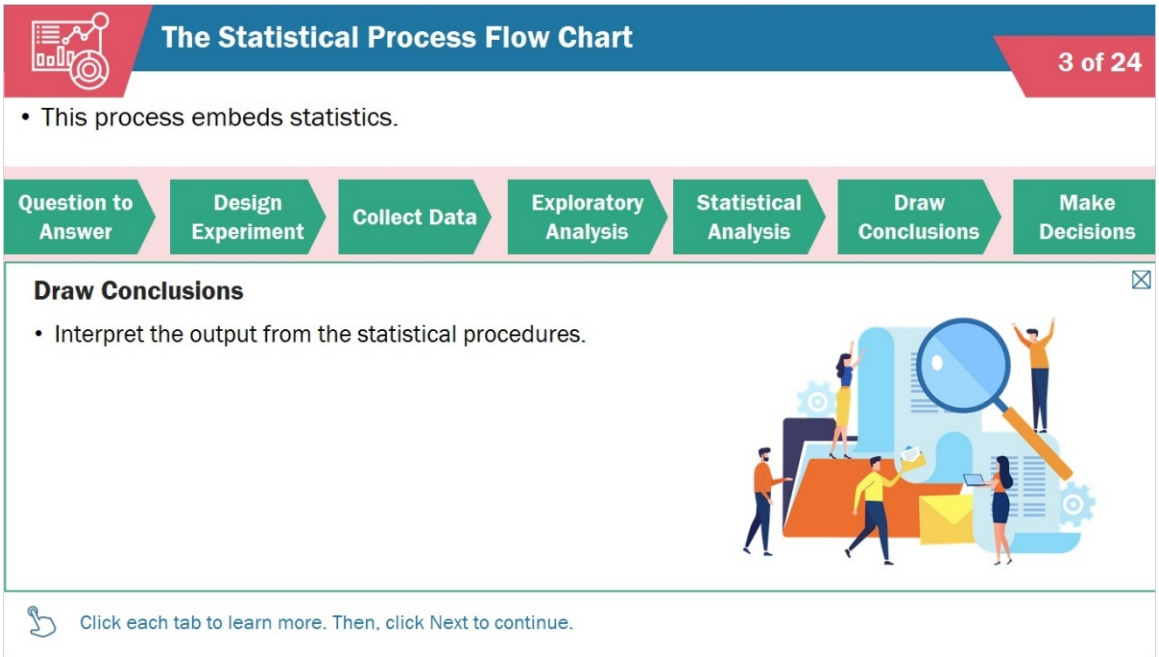
Step four is exploratory analysis, the point that one might typically think of as 'statistics' begins. Once the data are collected, initial exploratory analyses are done. These usually involve the calculation of summary statistics, graphing the data and identifying outlier observations. We want to make sure that the outlier observations are not the result of some error in the data collection process. The goal of exploratory analysis is to get a general sense of the data and its properties. This is very important for the next stage as it informs us about the best statistical analysis to use.

Tab 5: Statistical Analysis



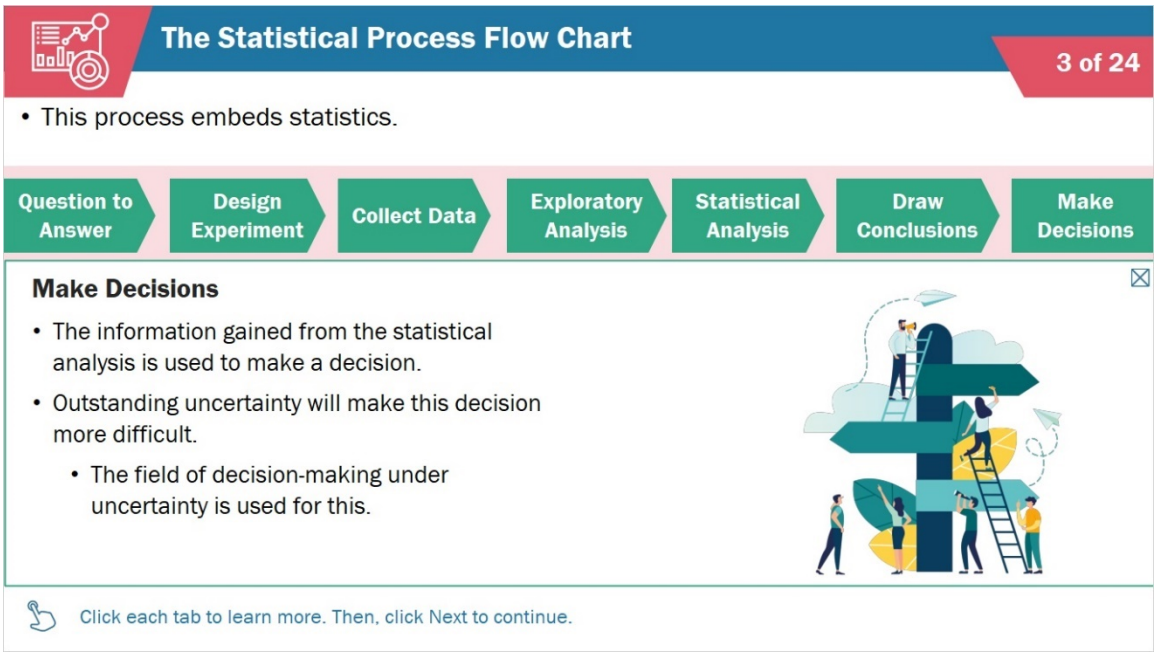
Step five is the statistical analysis and it is here that a statistical procedure is implemented to answer the question of interest and, as we've mentioned, uncertainties in what we can learn from the data are quantified. Since this element is so important, we will further break it down shortly. Statistical software, such as the statistical software called R that is used on this course, is often used to implement the procedure.

Tab 6: Draw Conclusions



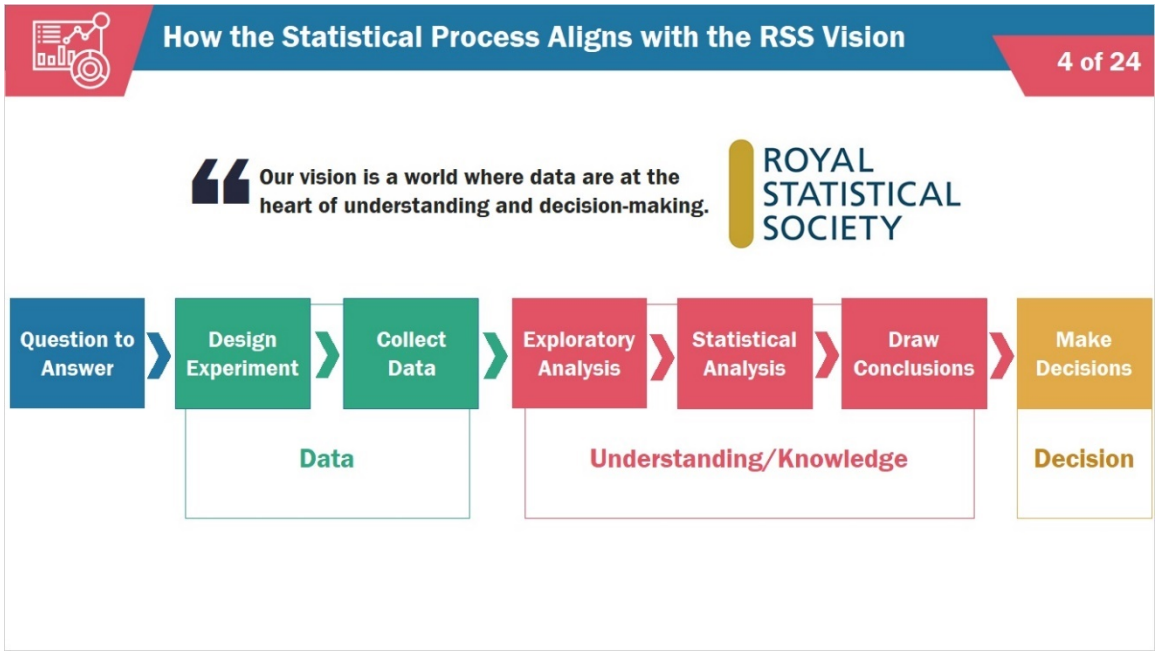
Step six is to draw conclusions. Statistical procedures produce an output that has to be interpreted.

Tab 7: Make Decisions



The final step is to make decisions. Often, answering the question from data is not the final step. We will want to use this knowledge to make a decision. Key to this is that the outstanding uncertainty will make this decision more difficult. The field of decision making under uncertainty is used for this.

Slide 4: How the Statistical Process Aligns with the RSS Vision

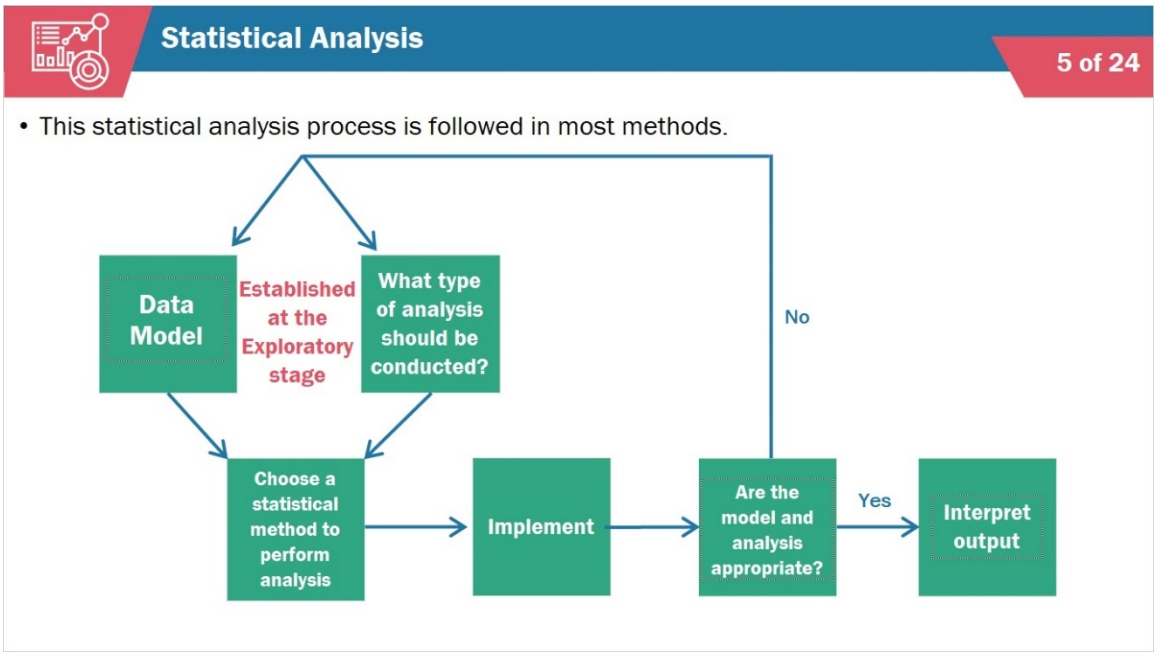


Before we move on, let's see how this process aligns with the Royal Statistical Society Statement. The elements of our process are divided into the 3 parts that the statement alludes to: data, understanding or knowledge, and decision.

Reference(s):
1. Royal Statistical Society. (2021). *RSS - What we do.* https://www.rss.org.uk/about/what-we-do-(1)/

Image(s):
1. Royal Statistical Society. (2021). *Royal Statistical Society Logo.* [Illustration]. https://www.rss.org.uk/

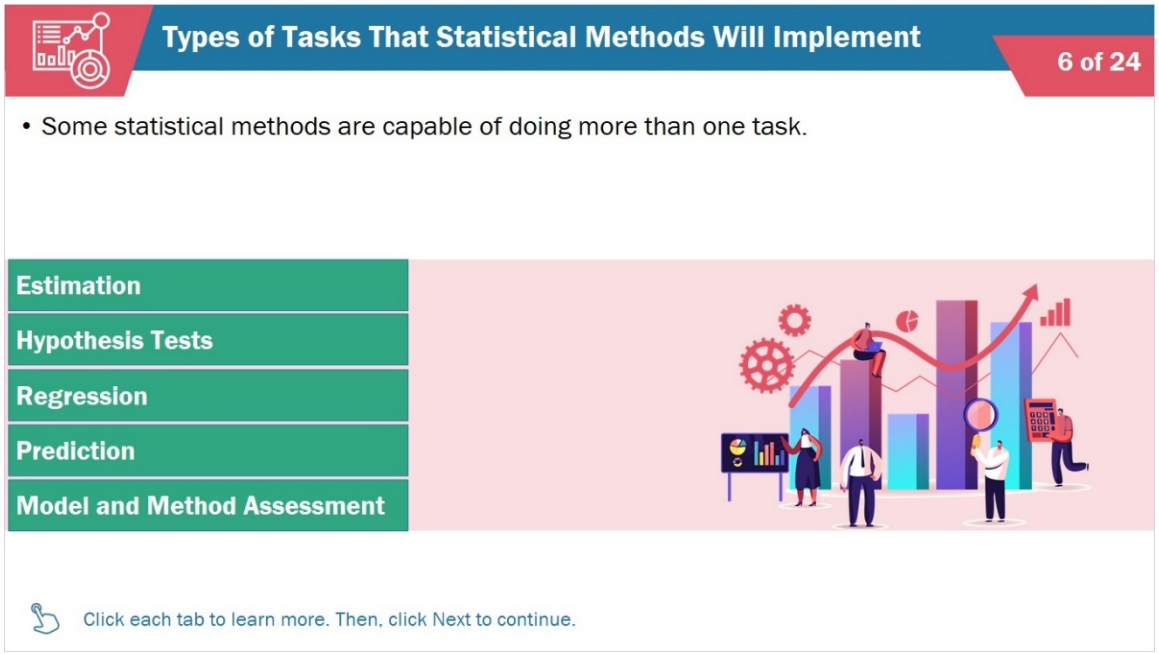
Slide 5: Statistical Analysis



We are now going to spend some time focussing on the statistical analysis stage of the process.

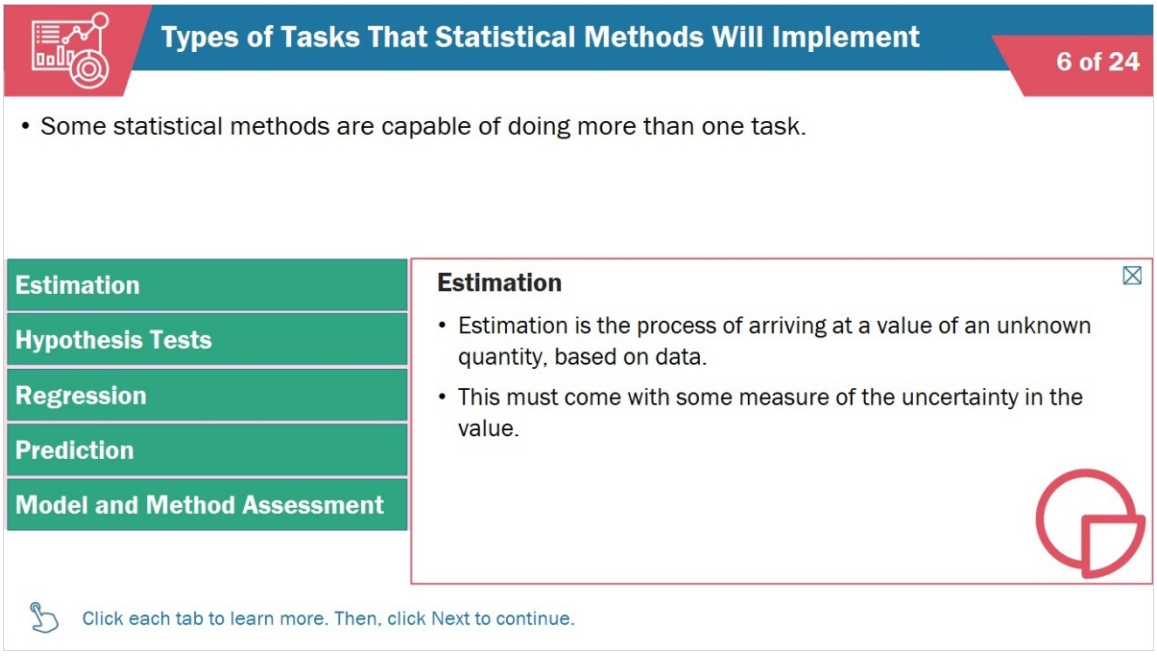
Within a statistical analysis, there is a process that one can see being followed in most of the methods that we look at. The exploratory analysis, the stage prior to this, typically informs us about a data model – that is descriptions of how data are distributed and what type of analysis we decide to conduct. This allows us to choose the appropriate statistical method to perform, which we then implement. The output of the statistical method, or some associated analysis, can help us in deciding post hoc whether this has been an appropriate analysis and, if not, how we might improve on it. The stage ends with an input into the next stage where conclusions are drawn, having been able to interpret the output of the analysis.

Slide 6: Types of Tasks That Statistical Methods Will Implement



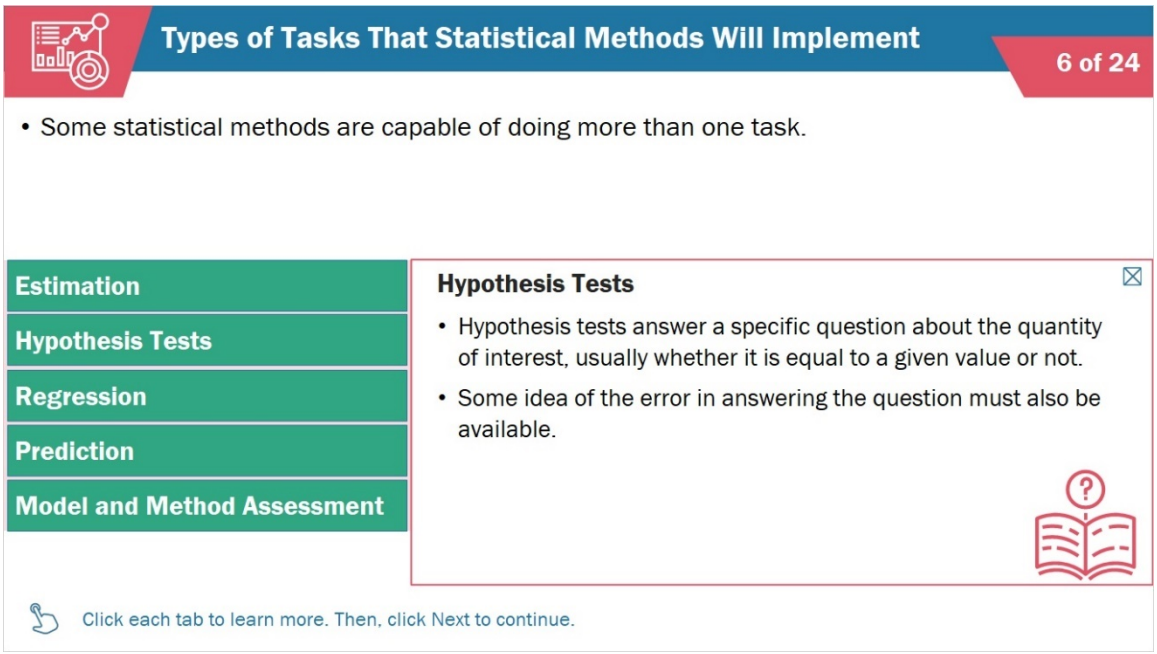
Now we look at the different types of tasks that statistical methods will typically implement. One method may be capable of doing more than one of these tasks. Click the tabs to learn about each task. When you are ready, click Next to continue.

Tab 1: Estimation



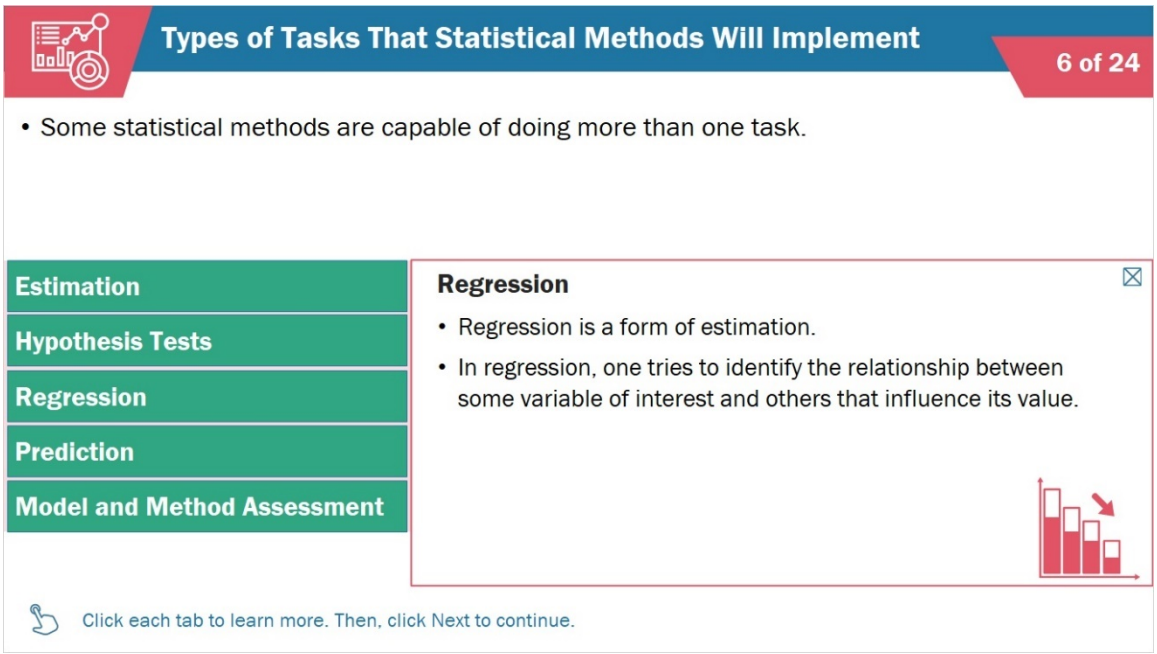
Estimation is the process of arriving at a value of an unknown quantity that we are interested in, based on data. As a simple example, we could use data on the heights of a group of female adults to produce an estimate of the mean height of the entire population. Crucially for a statistical method, this must come with some measure of the uncertainty in the value; in other words, some characterisation of the possible error or difference between this estimate and the actual value.

Tab 2: Hypothesis Tests



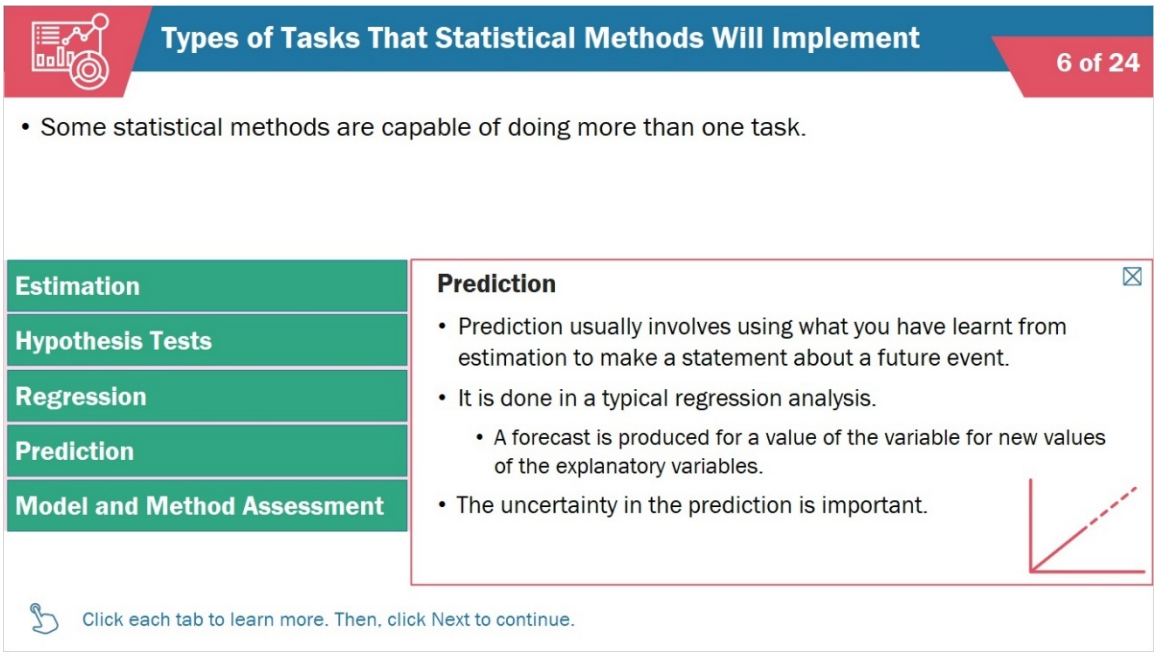
Hypothesis tests answer a specific question about the quantity of interest, usually whether it is equal to some given value or not. As with estimation, the characterising property of a statistical approach is that some idea of the error in answering the question must also be available.

Tab 3: Regression



Regression is, strictly speaking, a form of estimation but is such an important task in statistics that we list it separately. In regression, one tries to identify the relationship between some variable of interest and others that influence its value.

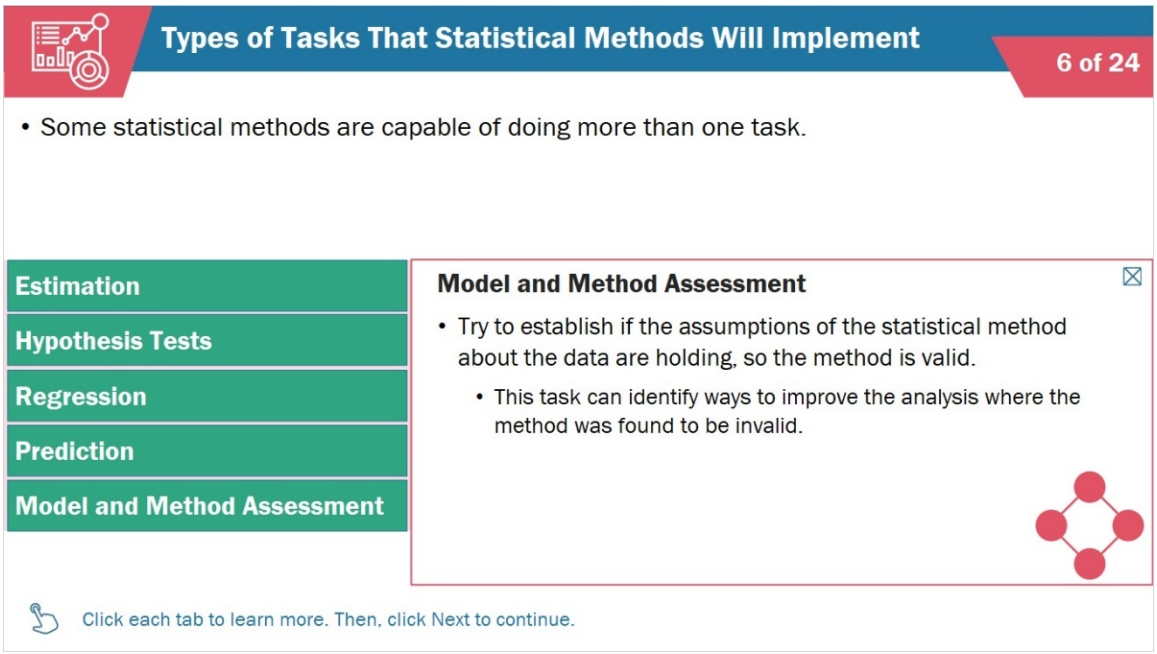
Tab 4: Prediction



Prediction is a task that usually involves using what you have learnt from estimation to make a statement about a future event. It is also a task done in a typical regression analysis, where one produces a forecast for a value of the variable for new values of the explanatory variables that may influence it.

The uncertainty in the prediction is also important.

Tab 5: Model and Method of Assessment



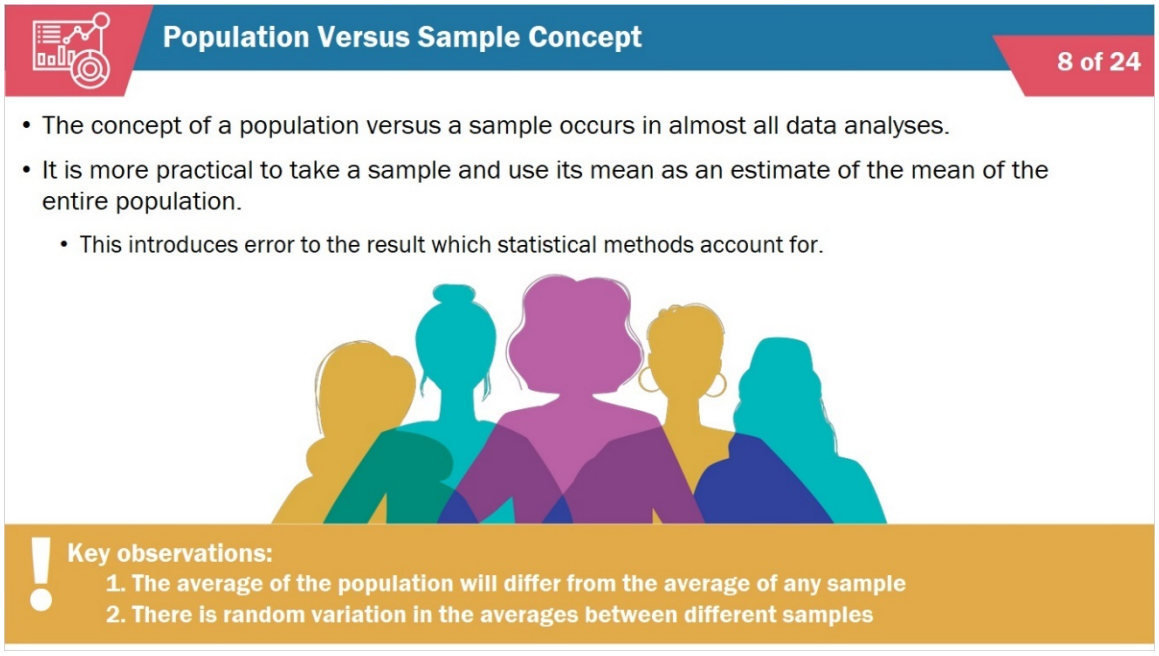
Finally, model and method assessment. In this task, one tries to establish if the assumptions of the statistical method about the data are holding, or at least close enough to holding so the method is valid. Ideally, in this task, one can also identify

ways to improve the analysis in the case where the conclusion is that the method was not valid.

Slide 7: Section 2: Other Important Concepts in Statistical Analyses

The rest of this presentation looks at a set of different concepts that we can encounter when implementing any of the statistical methods in the course.

Slide 8: Population Versus Sample Concept

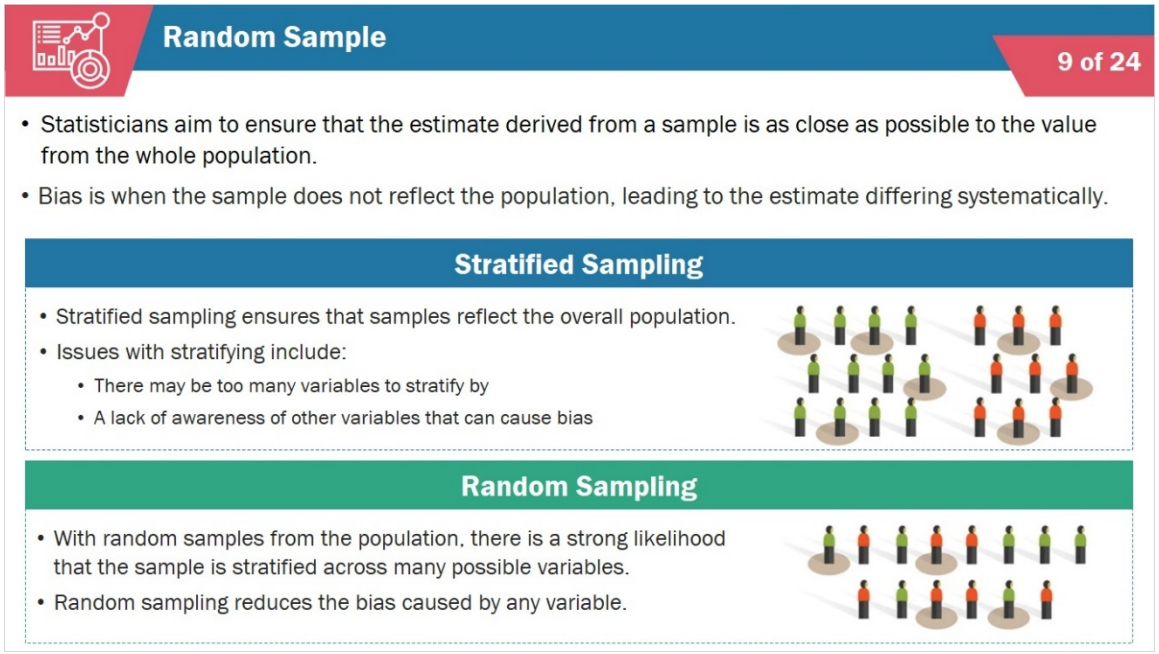


The concept of a population versus a sample occurs in almost all data analyses and is best illustrated with an example. Here we use the example of the population of all adult women in Ireland and trying to answer the question 'what is the mean height of adult women in Ireland?'. There is an obvious, but impractical, way to do this, which is to measure the heights of every adult woman in the country and take the average of those measurements. What we have to do in practice is sample a small fraction of all women, a sample of them that is to say. Then we can use the mean of the sample as an estimate of the mean of the entire population.

Key is the following observation. Suppose we took a sample of 100 women and got their average height. Now repeat the experiment, taking a sample of another 100 women. Would the average height of the second sample be the same? Almost certainly not! In fact, we can state two things: firstly, the average of the population will almost certainly differ from the average of *any* sample and secondly, there is random variation in the averages between different samples.

Almost never in statistics do we have data on an entire population that would allow us to answer a question exactly, either because it is not practical (as here) or just impossible to do even in principle. We are relying on a sample that we hope is representative of that population and using properties of it to provide an estimate. The trade-off with practicality is the introduction of error into what we are learning. Statistical methods account for this uncertainty in some way.

Slide 9: Random Sample



That brings us to the idea of a random sample and why, when designing an experiment to collect data, so much emphasis is placed on the importance of trying to get one.

We would like, as much as we can, to ensure that the estimate that we derive from a sample is as close as possible to the value from the whole population. One issue that is very important in this regard is bias, that is, the sample does not reflect the population in some way and that leads to the estimate differing systematically. If we go back to our

example of height from the previous slide, if average height changes with age, then we should try to ensure that the profile of ages in our sample agrees with the profile in the population. If our sample only contained, for example, women under 40 years of age then the estimate could be expected to be higher or lower than the population average.

There is a way to ensure this, by stratified sampling. Simply, we make sure that we sample women by age in proportion to what is in the population e.g. if 10% of adult women are aged 18-25 in the population then we make sure 10% of our sample are also women in this age group, and so on.

The problem now is the number of variables that we might want to stratify by. In our example, one might imagine others such as ethnicity. Very quickly, one finds that we want to stratify into more groups than the size of the sample that one is planning to take, so stratification cannot even be done. Another problem is that there may be other variables that can cause a bias that do not even occur to us.

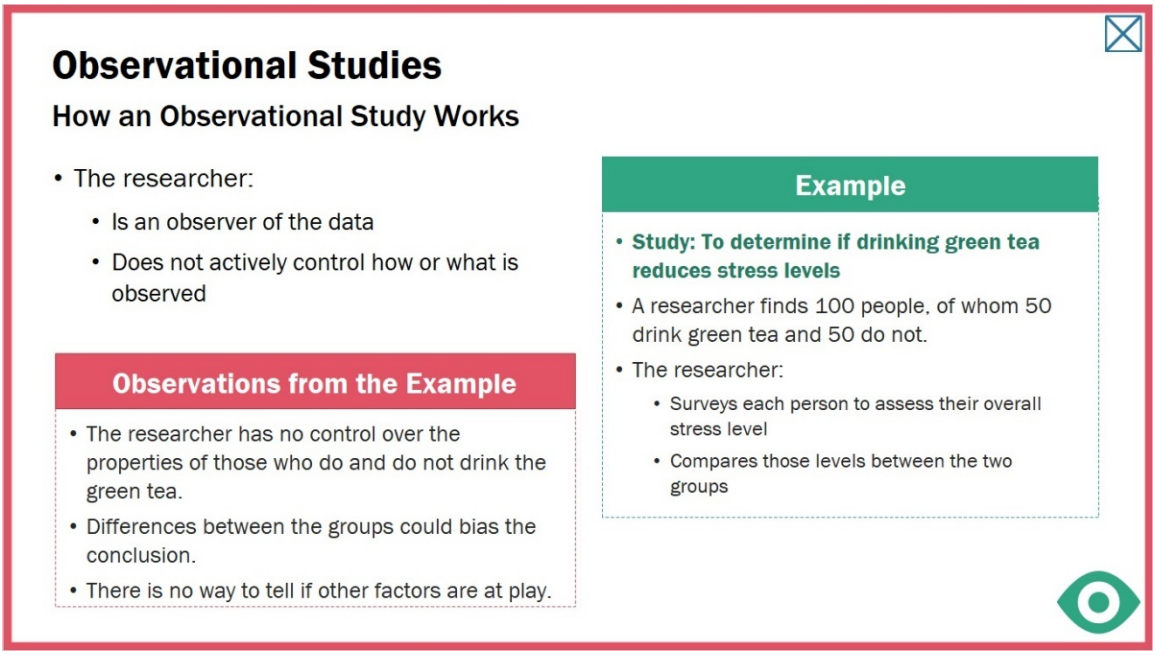
In a random sample, the idea is that, by randomly picking a sample from the population, one will more or less have stratified across any possible variable to match the population. It's not quite as good as stratified sampling, because the randomness of the sampling means that there is variation from the proportion that you want ideally. Nevertheless, the enormous advantage of random sampling is its ability to reduce the bias caused by any variable, not just the ones that you managed to think about.

Slide 10: Observational Studies Versus Designed Experiments

The issue of observational studies versus designed experiments has become increasingly important as very large data sets in all sorts of fields become available, and the means to analyse them have also become easier to access. Often, there is a rush to analyse these data and draw conclusions without thought to how the data were collected and hence, if the statistical method used is appropriate.

Click the tabs to learn more about the difference between these two types of studies. When you are ready click Next to continue.

Tab 1: **Observational Studies**

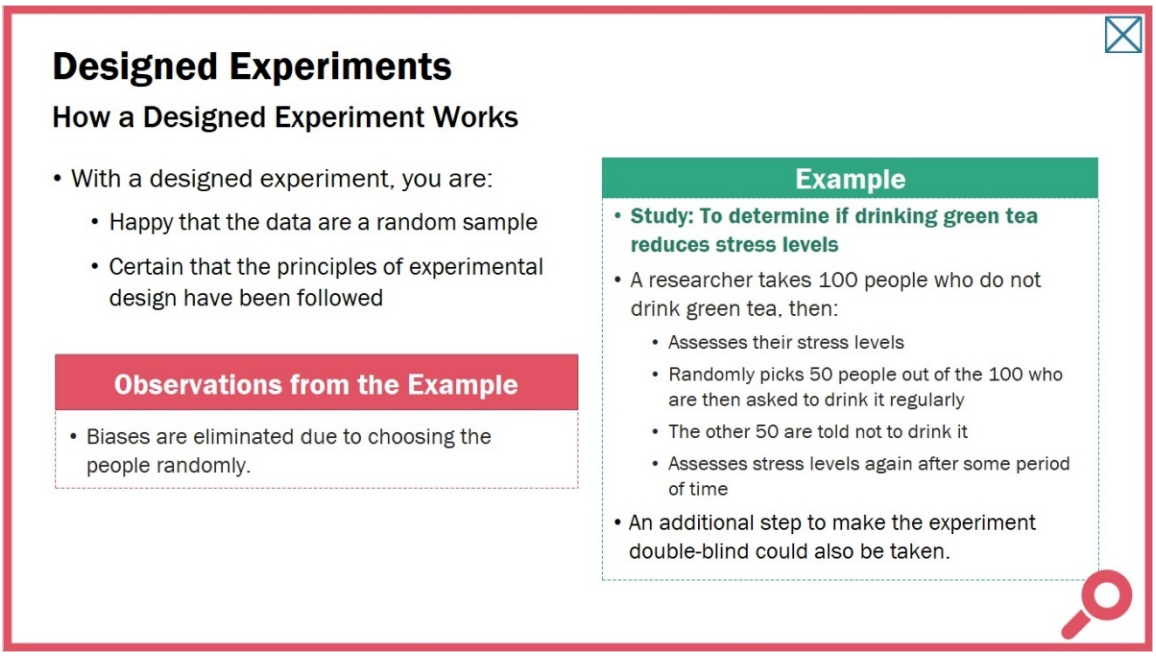


In so-called observational studies the researcher, or person doing the data analysis, is just an observer of the data and does not actively control how or what is observed.

Let's look at an example of a study where we want to determine if drinking green tea reduces stress levels.

The researcher finds 100 people, 50 of whom drink green tea and 50 who don't. The researcher surveys each person to assess their overall stress level and compares those levels between the two groups. It seems a reasonable way to answer the question about whether green tea reduces stress, but the researcher had no control over the properties of the people who do and do not drink green tea. Differences between the two groups (for example, the people drinking tea could lead a less stressful lifestyle more generally) could severely bias the conclusions, and there is no way to tell if any difference is due to the green tea or other factors.

Tab 2: Designed Experiments



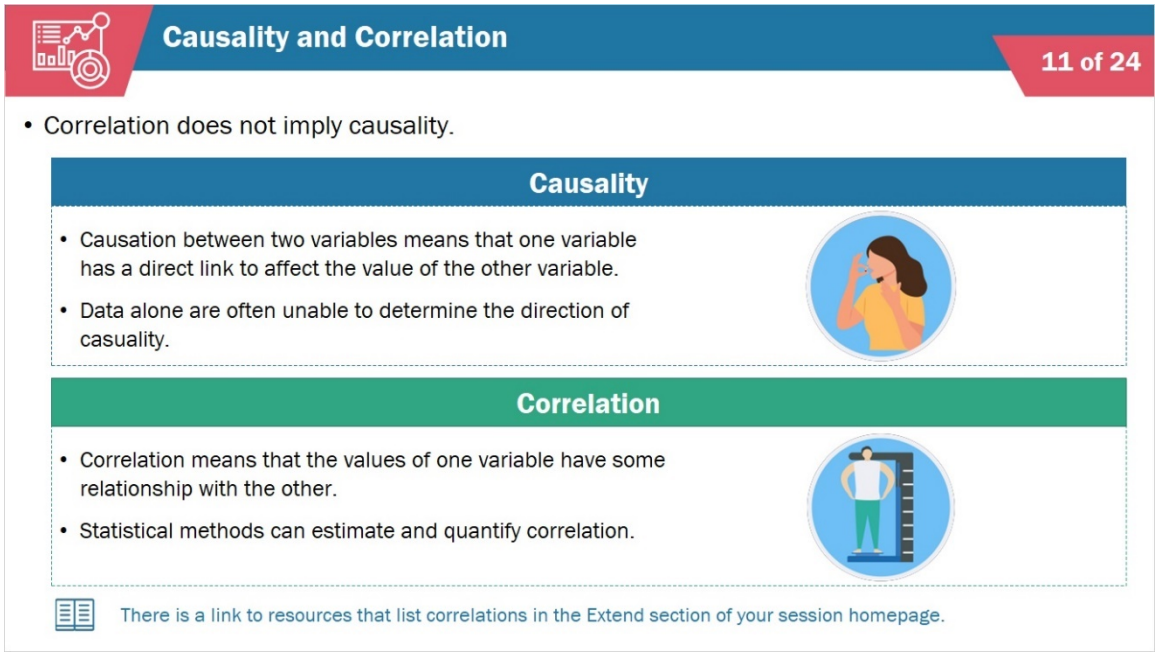
In a well-designed experiment, we are happy that the data are a random sample or have followed the principles of experimental design.

Let's look at how our study can be carried out in this way. A researcher takes 100 people who don't drink green tea and assesses their stress levels. 50 out of the 100 are randomly chosen and asked to drink the green tea regularly. The other 50 are told not to drink it. After a period of time, the stress levels of each group are assessed again. Here we have a more properly designed experiment.

Note that, because we randomly pick people to start drinking tea, we are comparing tea and non-tea drinking on the same individuals and so we are eliminating all of the biases that arose from using different people in the observational study. Note that an extra step would be a 'double blind' experiment, where none of the individuals in the study know whether they are drinking green tea or not (for example, everyone is given a similar looking and tasting drink each day, but only 50 of them are actually drinking green tea).

The point here is that many data sets come from observational studies, and not necessarily well-designed ones, and so one has to be very aware of potential biases that can influence an analysis.

Slide 11: Causality and Correlation

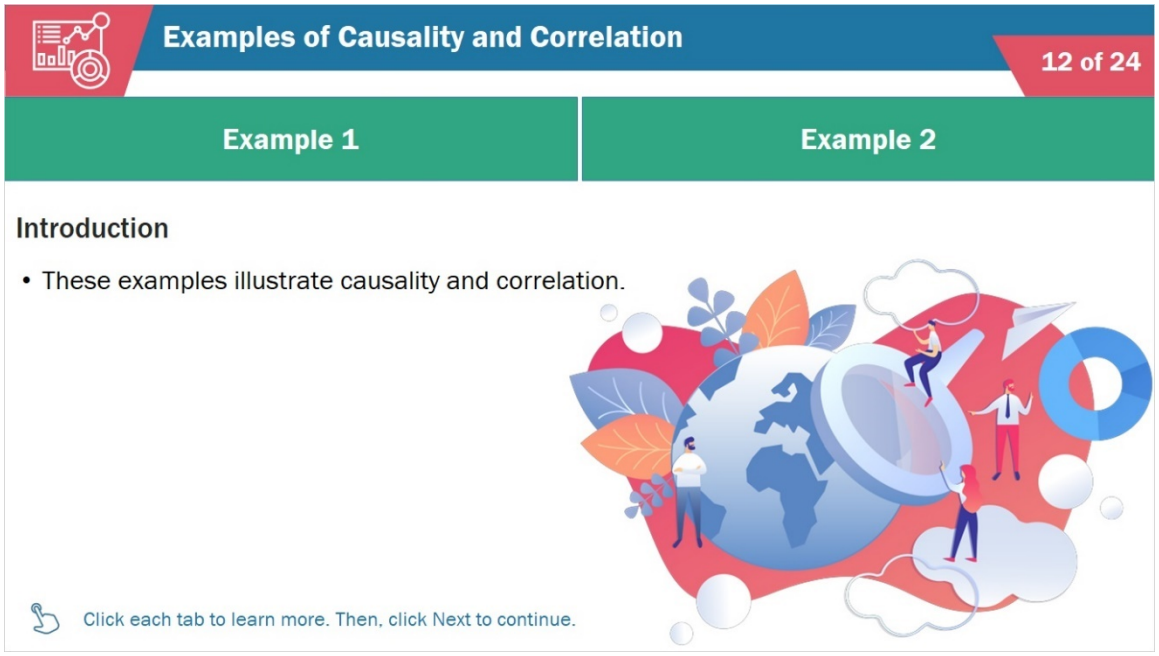


That causality is not the same as correlation is now a well-worn phrase. To be clear, causation between two variables means that one of the variables has a direct link to affect the value of the other. Taking an aspirin to stop a headache or switching on central heating to warm up a house, are examples. Correlation on the other hand is a weaker concept; it just means that the values of one variable have some relationship with the other. An example would be an individual's height and weight.

In the Extend element of this session, there are links to resources that list surprising correlations that would certainly appear to have no causal relationship.

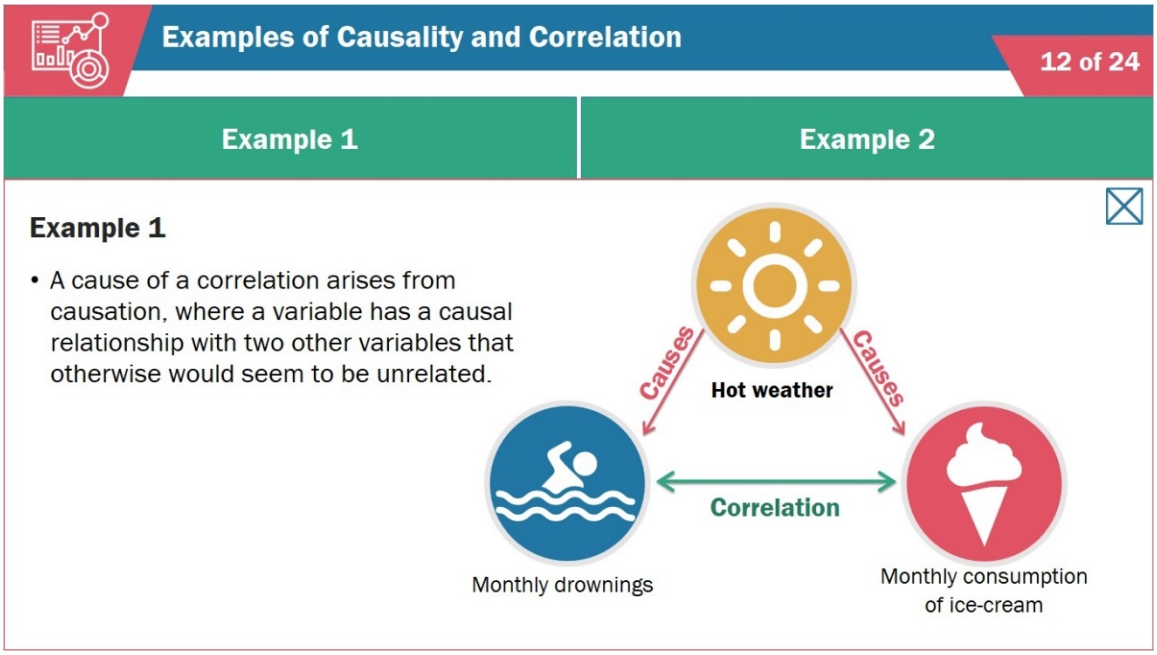
Statistical methods, particularly in regression, can estimate and quantify correlation but the issue of causality is far more challenging. In particular, the data alone are often unable to determine the direction of the causality. In this course, we'll focus on correlation and just be aware that that is not the same as identifying a causal relationship.

Slide 12: Examples of Causality and Correlation



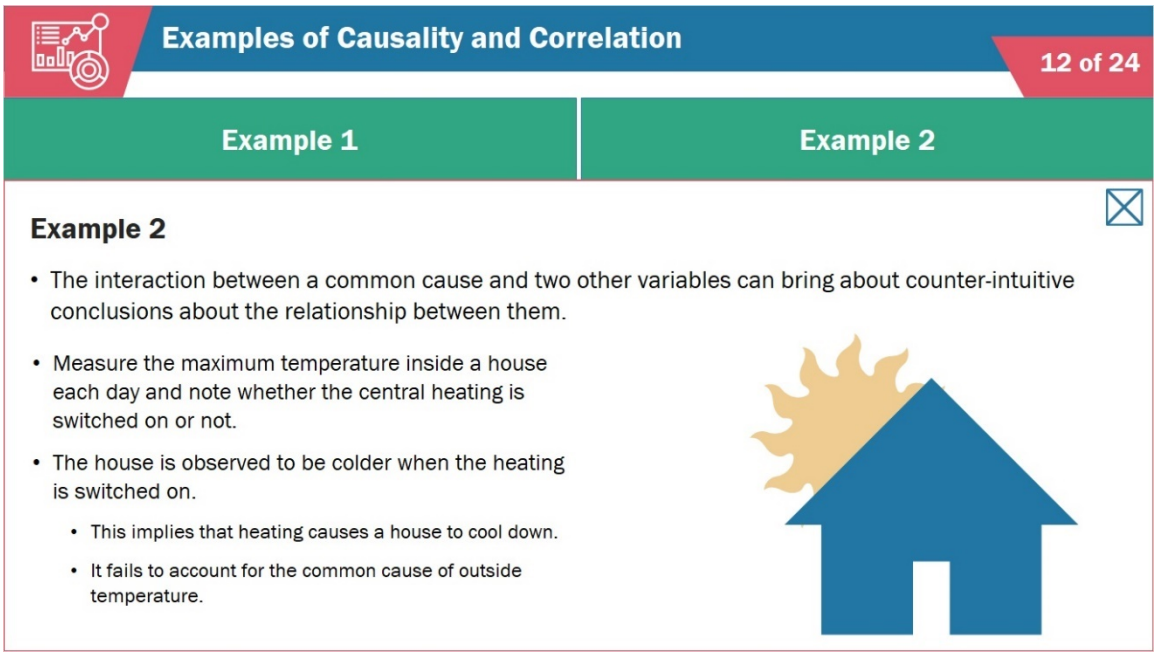
Let's take a look at two examples that illustrate causality and correlation. Click the tabs to learn more. When you are ready, click Next to continue.

Tab 1: Example 1



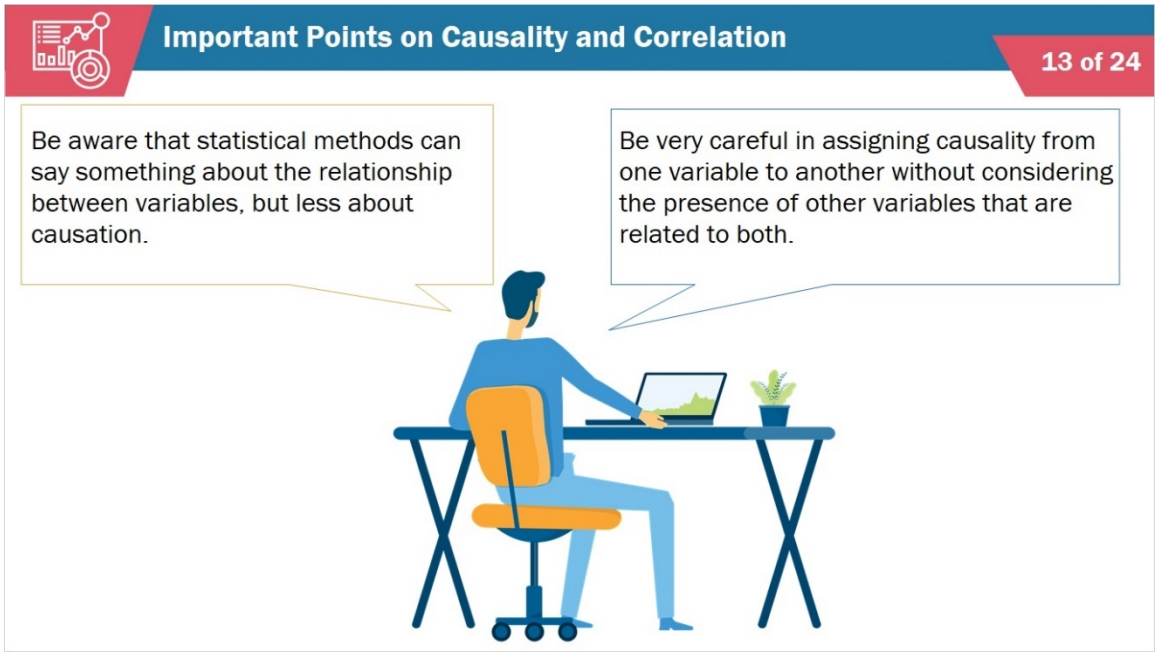
Often, one cause of a correlation arises from causation, where a variable has a causal relationship with two other variables that otherwise would seem to be unrelated. The example of ice cream consumption and drownings is a classic one. Both have a cause in hotter weather which causes more people both to swim, and hence drown, and eat ice cream at the same time. As a result, there is correlation between the two.

Tab 2: Example 2



The example on this slide illustrates that the interaction between a common cause and two other variables can bring about counter-intuitive conclusions about the relationship between them. The maximum temperature reached inside a house is measured each day, as well as whether the central heating is switched on that day or not. It is observed that the house is colder when the heating is switched on. Hence, central heating causes a house to cool down. In this case, the correlation between central heating and the temperature of a house is mistakenly identified as the reverse of what we know to be the causal relationship, because it fails to account for the common cause of outside temperature which is causing both the temperature of the house to change, as well as the decision to turn on the central heating. Given the outside temperature, we would see that the heating does indeed warm up the house.

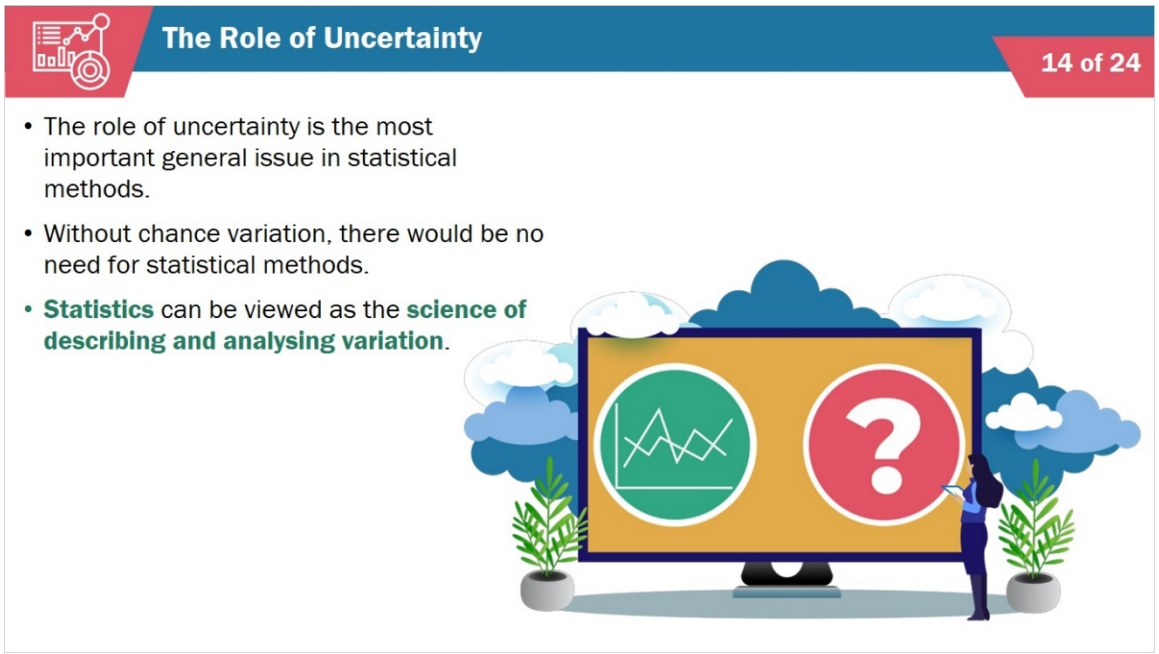
Slide 13: Important Points on Causality and Correlation



So, what do you need to take away from this issue? There are two important points:

- Be aware that statistical methods in general can say something about the relationship between variables but less about causation
- One must be very careful in assigning causality from one variable to another without considering the presence of other variables that are related to both

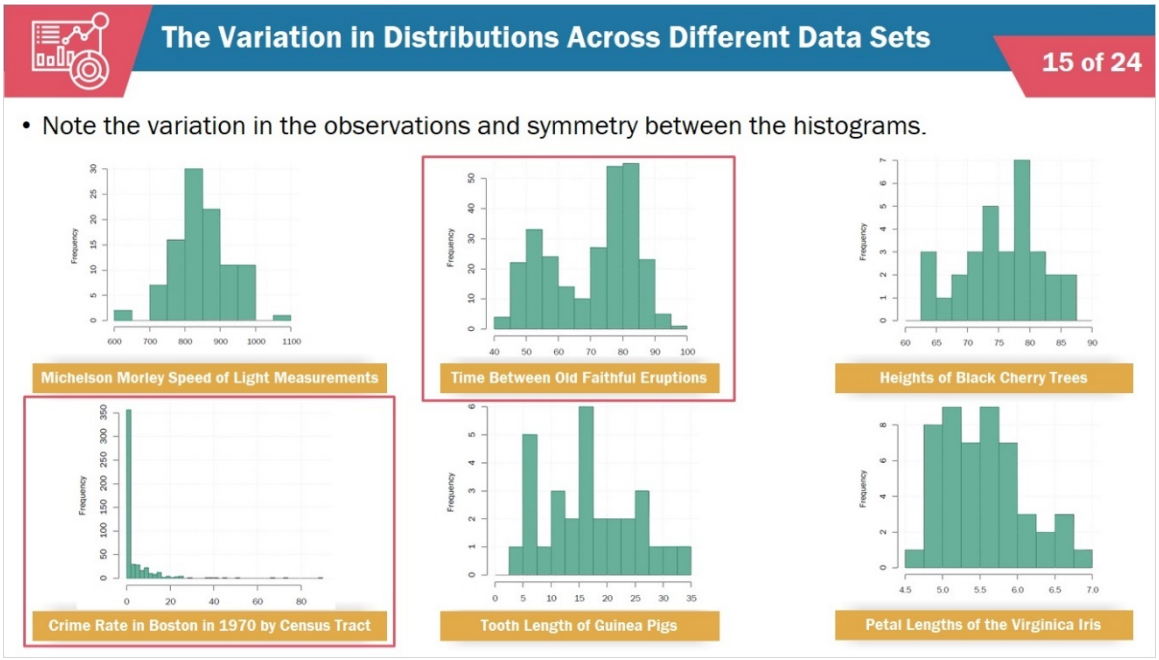
Slide 14: The Role of Uncertainty



We have left the most important general issue in statistical methods to last. It is fair to say that without chance variation – the fact that observations of a quantity will very likely

differ in unpredictable ways – there would be no need for statistical methods at all! Indeed, statistics can be viewed as the science of describing and analysing variation.

Slide 15: The Variation in Distributions Across Different Data Sets

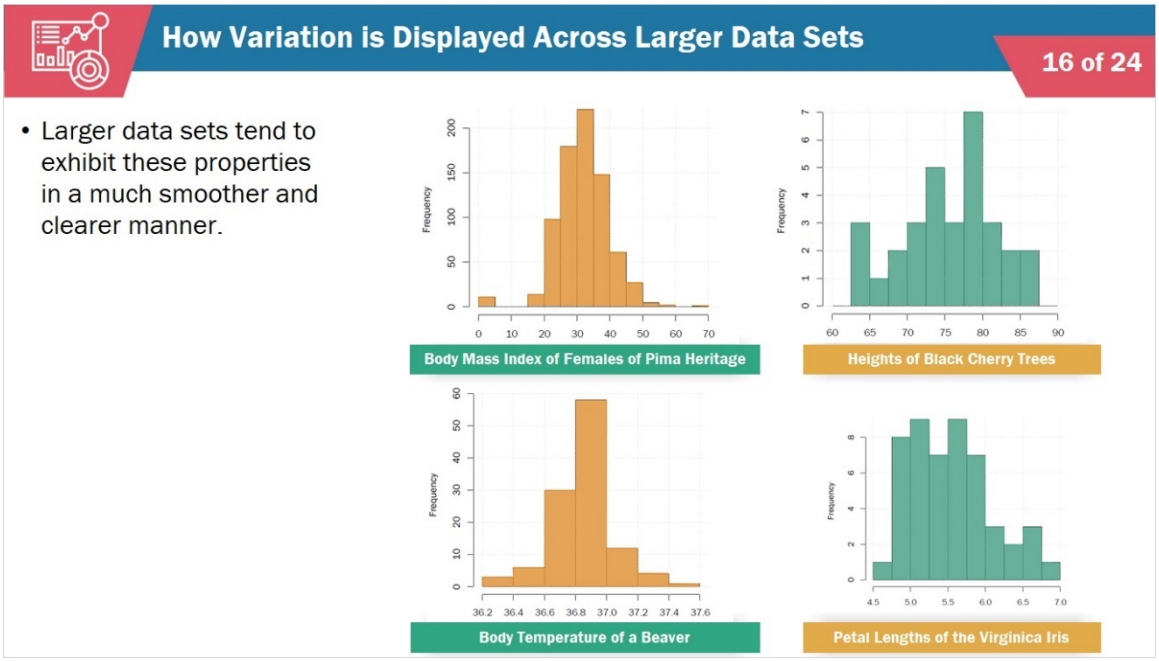


Here we see histograms of six different data sets from quite widely different situations. The first thing to note is that there is variation; not all of the observations have the same value!

There are some common characteristics of the distribution of observations among some of these sets. Several of them seem to have a central peak where the data are concentrated, with the frequency of observations decreasing symmetrically on either side of this peak. But such a characteristic is not universal; the Old Faithful data appears to have two peaks, and the Boston crime data does not have a symmetric shape at all.

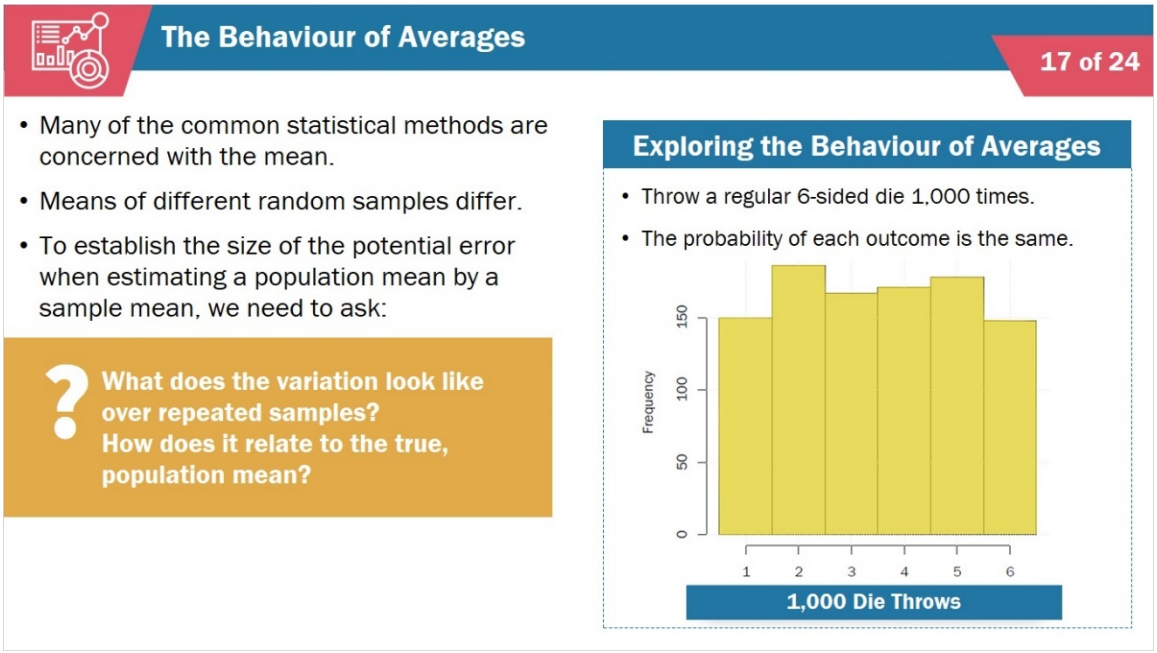
The six data sets are rather small in size and hence tend to be quite noisy making these characteristics not immediately obvious sometimes.

Slide 16: How Variation is Displayed Across Larger Data Sets



Larger data sets will tend to exhibit these properties in a much smoother and clearer manner, as is seen here with two more examples.

Slide 17: The Behaviour of Averages

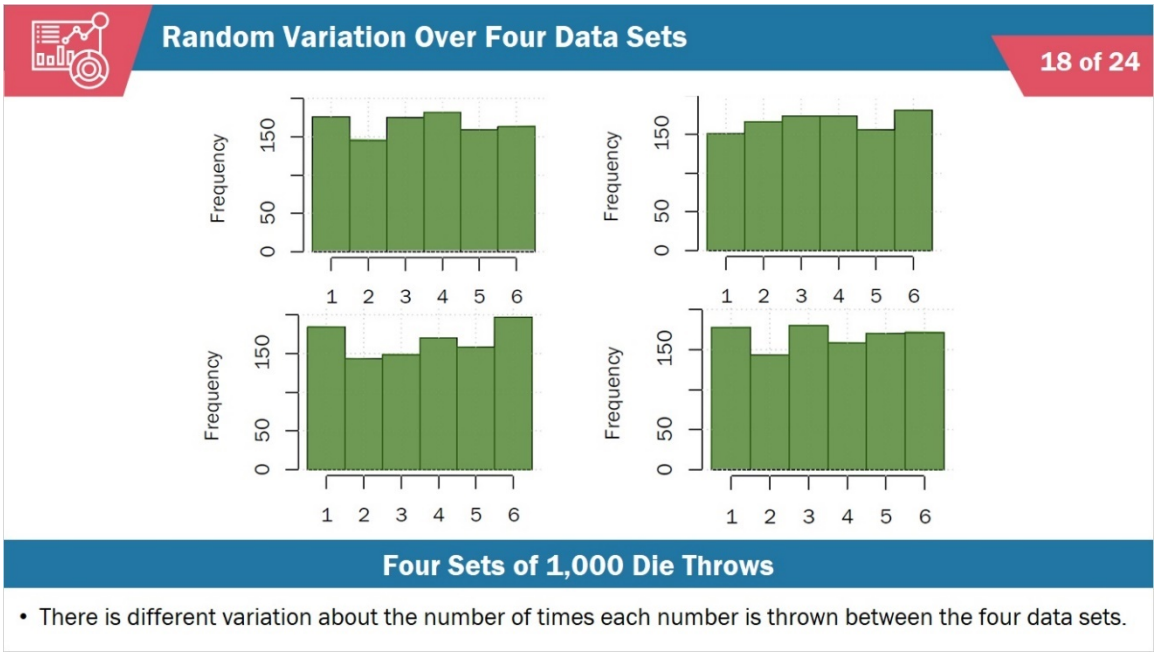


Since many of the statistical methods that we look at in this module concern the mean in some way, we now focus on properties of the random variation in a sample mean. If we are interested in estimating the mean of some property of a population (like height of adult women), a sensible way to do it is to take a (hopefully random) sample from the population and use the average of the sample as an estimate of the population mean.

As we've already noted, if we were to take another random sample then its mean would almost certainly be different from that of the first sample. What does this variation look like over repeated samples? How does it relate to the 'true' or population mean? These are important questions to ask because they tell us something about the size of the potential error when estimating a population mean by a sample mean.

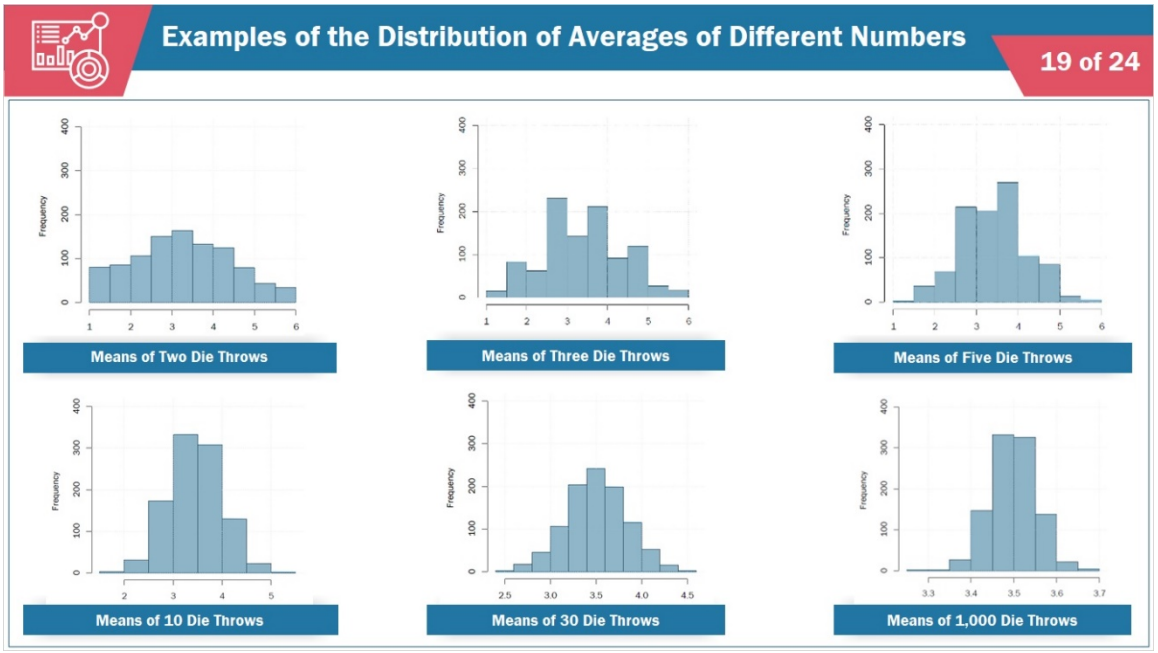
We explore this through throwing a regular 6-sided die. To start with, here is a histogram of the result of throwing the die 1,000 times. The probability of each outcome, one to six, is the same at one sixth, and so over the 1,000 throws we see that each number appears in roughly the same number of throws, although because each outcome is random, we do not get exactly the same number of throw for each outcome.

Slide 18: Random Variation Over Four Data Sets



As we've said, there is random variation in how the 1,000 throws turn out. Here are four more sets of 1,000 throws and of course, we see that broadly each number is occurring in the same amount but there is different variation about that general property between the four sets.

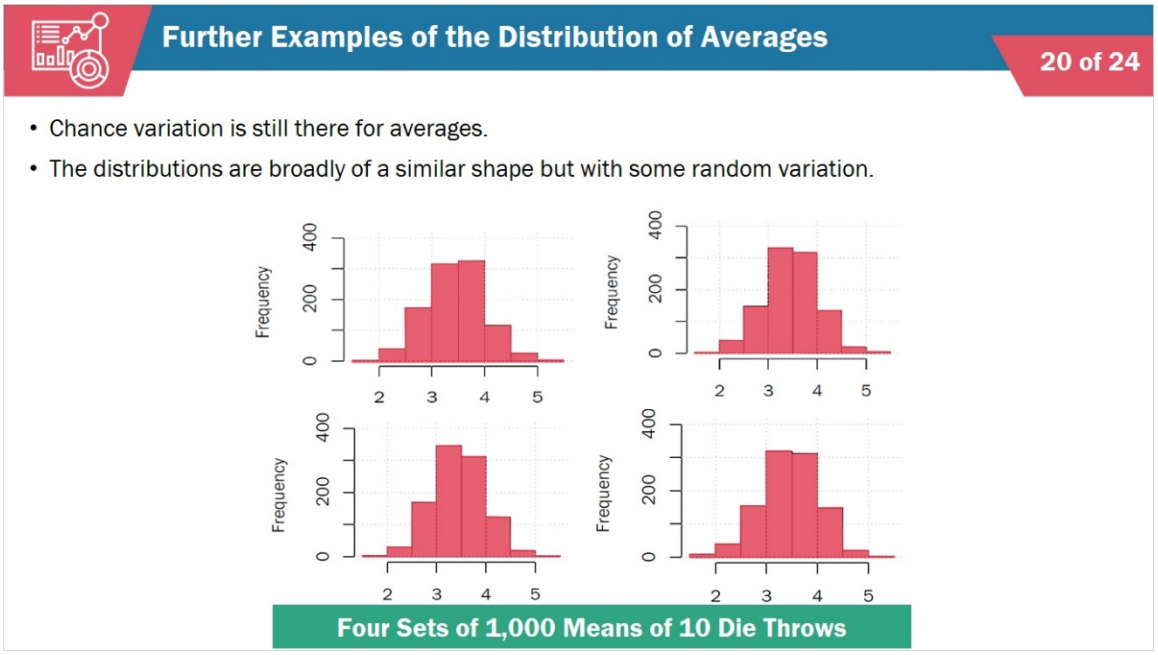
Slide 19: Examples of the Distribution of Averages of Different Numbers



What does the distribution of the average of two die throws look like? Or of three, 10, 1,000 throws? Let's think about the average of two throws; so throw a die twice, say getting a five and a one, and then take the average, which is three. Now repeat that a total of 1,000 times to get 1,000 means of two throws. Then repeat that idea but calculating the average of three, five, 10, 30 and then 1,000 throws. Let's draw histograms of the 1,000 averages in each case and see what we get.

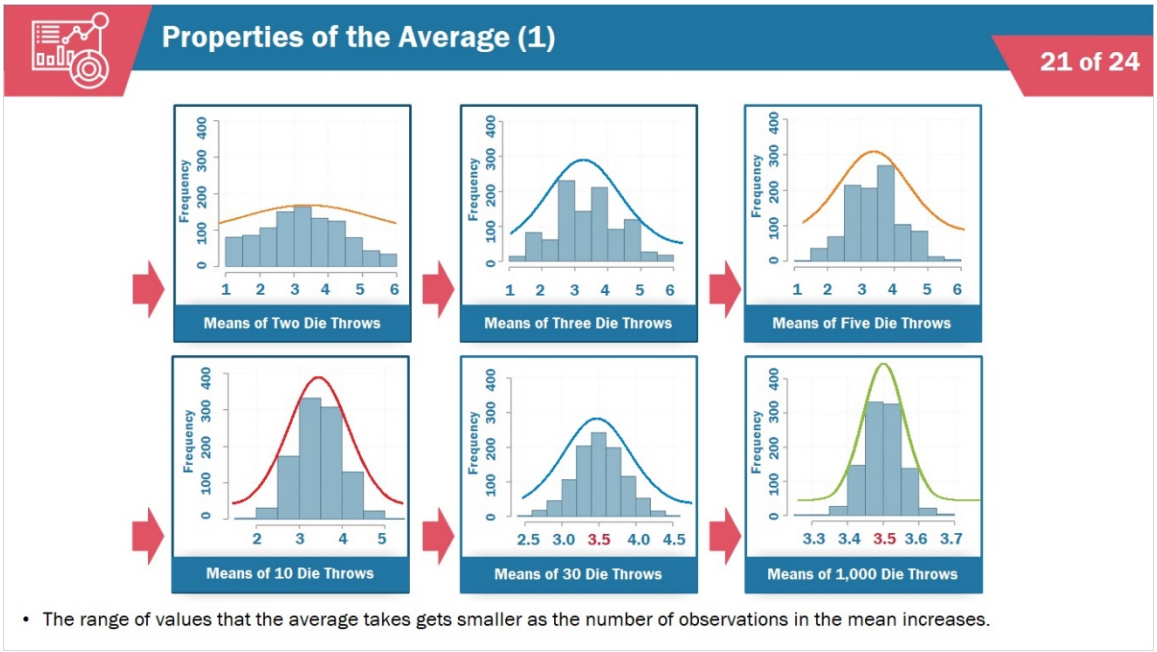
What can we say about the distributions of the means?

Slide 20: Further Examples of the Distribution of Averages



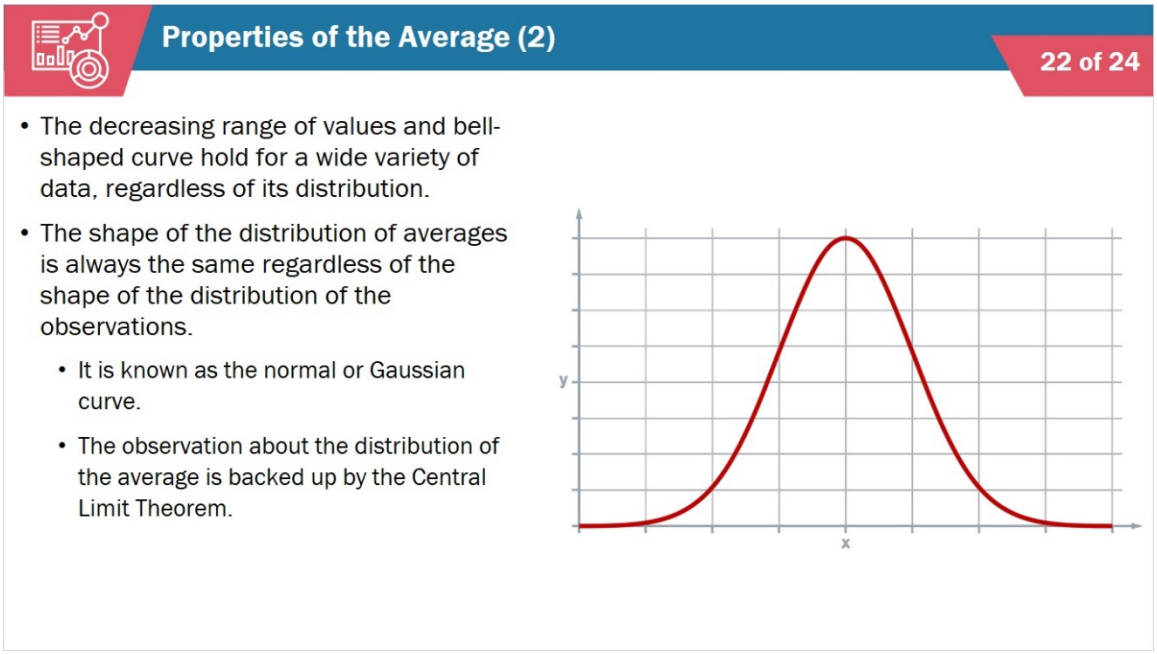
Chance variation is still there for averages. Here are four other sets of 1000 means (in this case of the averages of 10 throws). We see the distributions are broadly of similar shape but again, with some random variation.

Slide 21: Properties of the Average (1)



First note the range of values on the horizontal axis of each histogram. The range of values that the average takes gets smaller and smaller as the number of observations in that average increases, ultimately centering around 3.5. Second, the shape of the distribution looks more and more like a symmetric, bell-shaped curve. This is certainly very pronounced in the averages of 10 or more throws.

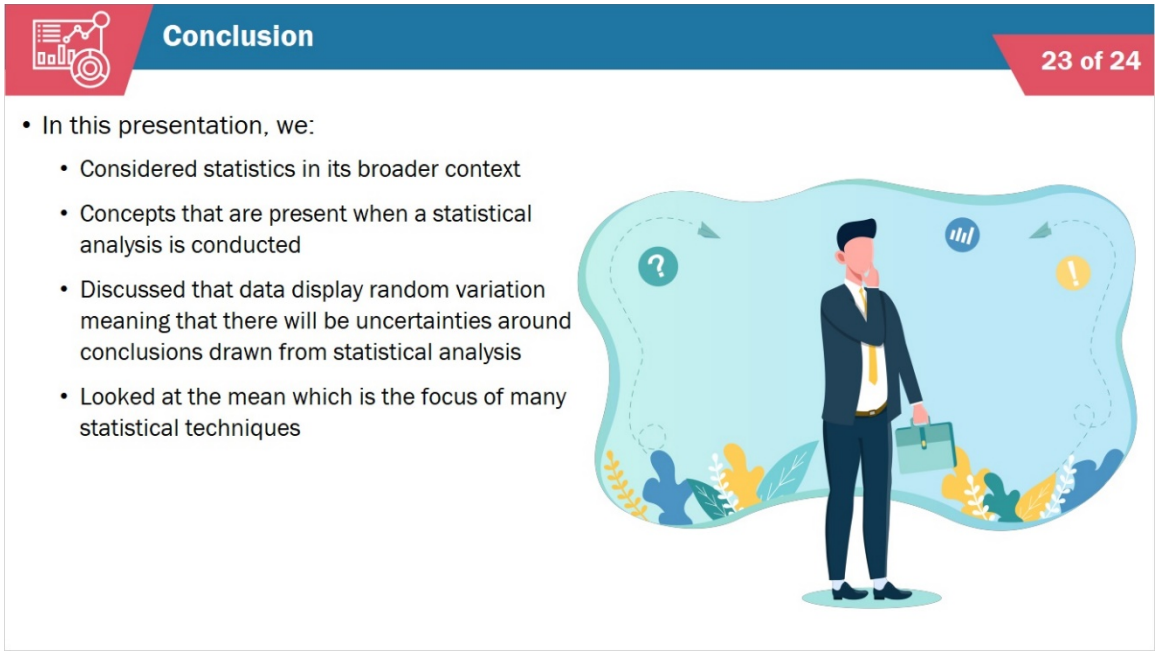
Slide 22: Properties of the Average (2)



Both of these properties hold for a wide variety of data, regardless of its distribution. For example, the distribution of die throws is a flat one across the six possible outcomes, but we would see very similar behaviour if we repeated these experiments for data coming from very different looking distributions. The more remarkable property is the shape of the distribution of averages – a distinctive 'bell-shaped' symmetric curve – is always the same regardless of the shape of the distribution of the observations.

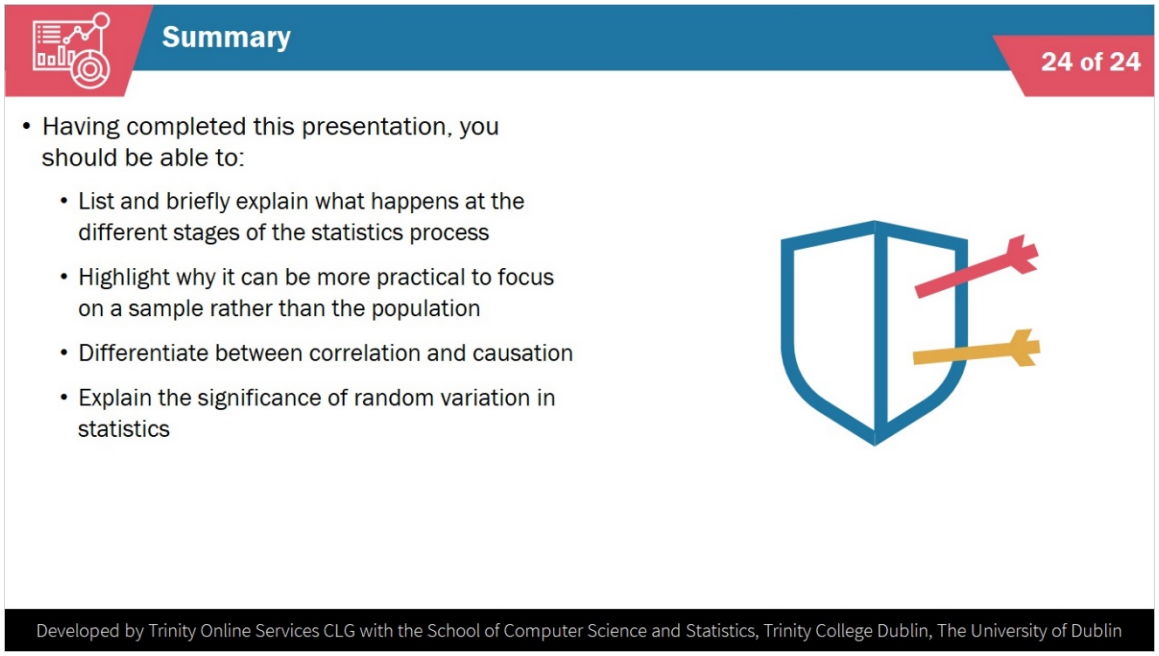
This curve is known as the normal or Gaussian curve and we discuss it in much more detail in Session 3 of this module as it plays a central role in many of the statistical techniques that we look at. It should also be noted that there is some solid mathematical theory to back up our observation about the distribution of the average (known as the Central Limit Theorem).

Slide 23: Conclusion



As was said at the start, the purpose of this presentation is to take a step back and consider statistics in its broader context, and some of the important concepts that are present almost always when we conduct a statistical analysis. This ended with the most important, and defining, concept of all, that data display random variation, and that as a consequence there will be uncertainties around any conclusion that we draw from a statistical analysis. Finally, we looked at the case of the average or mean, because that is the focus of many of the statistical techniques that we will look at in the rest of this module.

Slide 24: Summary



Having completed this presentation, you should be able to:

- List and briefly explain what happens at the different stages of the statistics process
- Highlight why it can be more practical to focus on a sample rather than the population
- Differentiate between correlation and causation
- Explain the significance of random variation in statistics