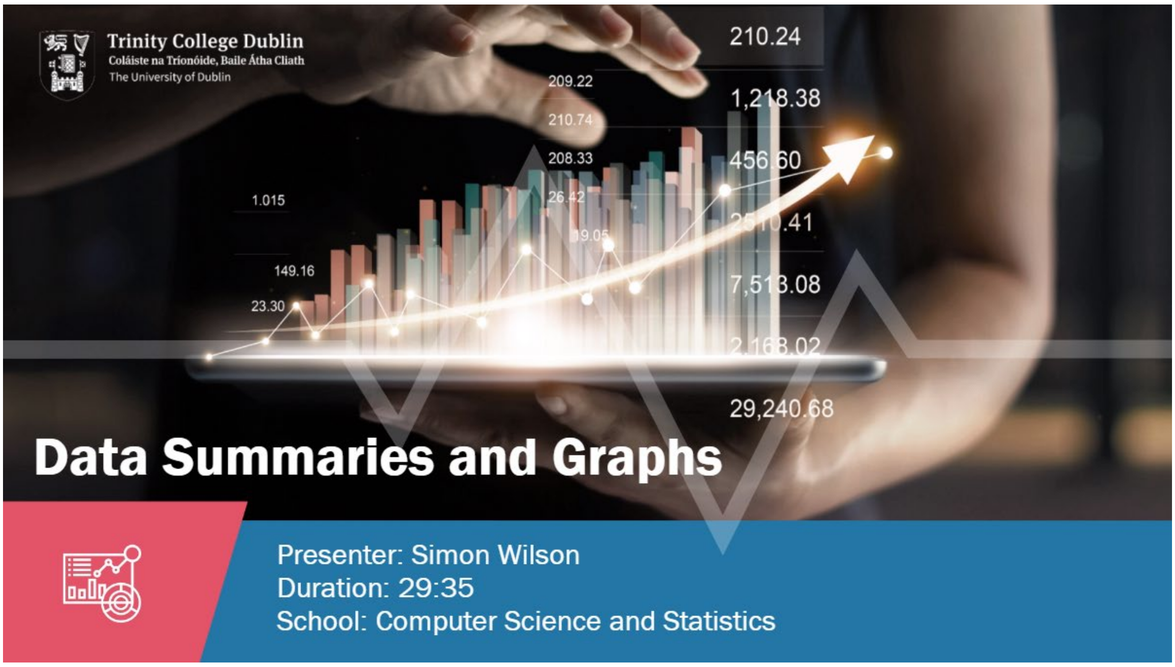
Data Summaries and Graphs

Slide 1: Introduction


210.24
209.22
1,218.38
210.74
208.33
456.60
1.015
26.42
2510.41
19.05
149.16
7,513.08
23.30
2,168.02
29,240.68

Trinity College Dublin
Coláiste na Tríonóide, Baile Átha Cliath
The University of Dublin

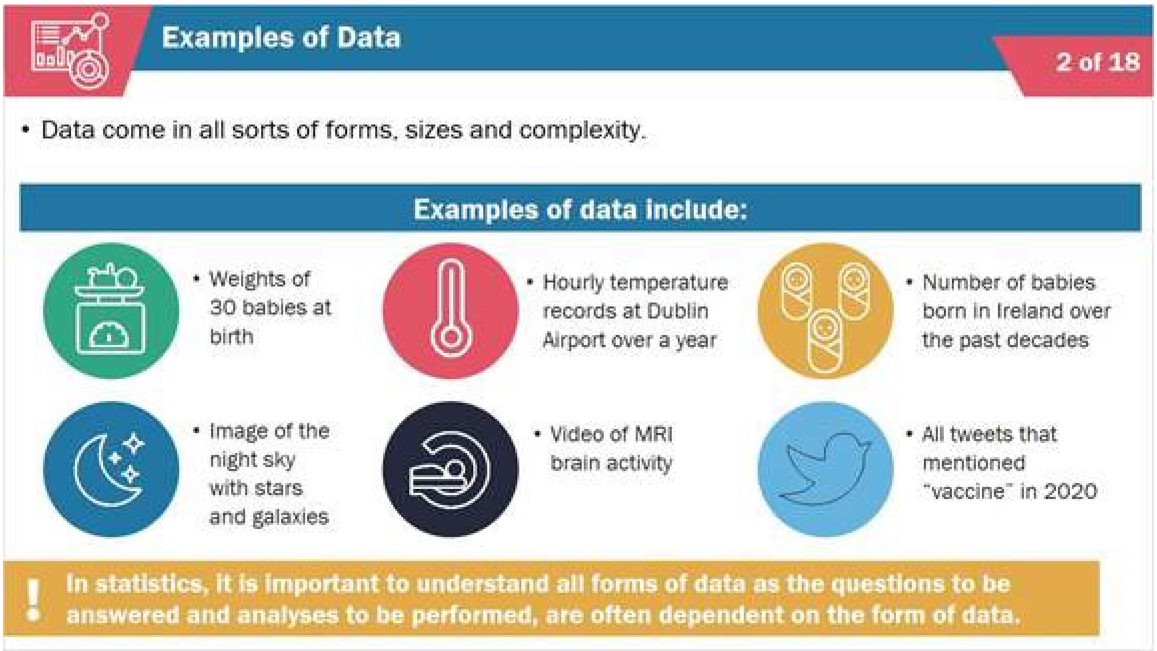
Data Summaries and Graphs

Presenter: Simon Wilson
Duration: 29:35
School: Computer Science and Statistics

Welcome to this presentation on “Data Summaries and Graphs” that will look at different data types and how to provide information from data in the form of summaries and graphs.

These are some of the first tasks of any statistical analysis of a new data set because they can give a lot of insight into properties of the data that might be important for what you’d like to learn from them. We’ll also try to discern some general principles around good data visualisation as well as the advantages and disadvantages of different data summaries and graphs.

Slide 2: Examples of Data



Examples of Data

2 of 18

- Data come in all sorts of forms, sizes and complexity.

Examples of data include:

- Weights of 30 babies at birth
- Hourly temperature records at Dublin Airport over a year
- Number of babies born in Ireland over the past decades
- Image of the night sky with stars and galaxies
- Video of MRI brain activity
- All tweets that mentioned “vaccine” in 2020

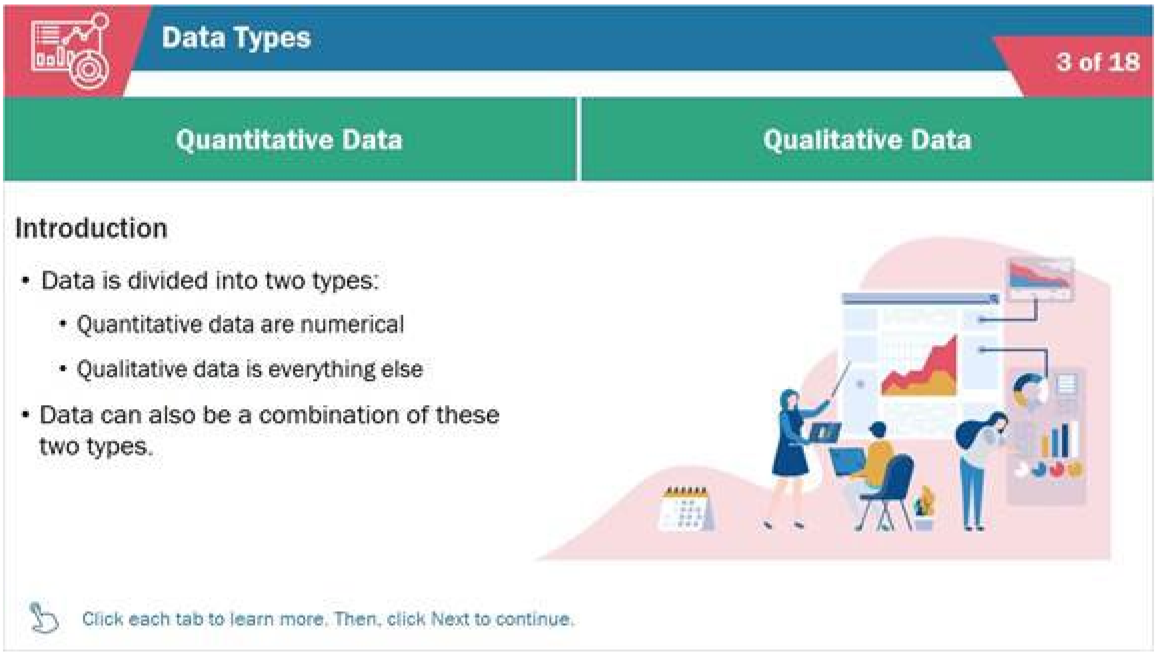
! In statistics, it is important to understand all forms of data as the questions to be answered and analyses to be performed, are often dependent on the form of data.

Data come in all sorts of forms, sizes and complexity. Within your own experience, you can think about data that you come across, perhaps in your career, or have read about, or indeed are generating yourself as you go about your daily life. Some examples could be:

- Weights of 30 babies at birth
- Hourly temperature records at Dublin Airport over a year
- The number of babies born in Ireland in the last decades
- Or you could have more complicated data such as:
- The image of a part of the night sky containing stars and galaxies; or
- Video footage of brain activity from an MRI scanner; or how about
- All tweets that mention the word “vaccine” in 2020

In statistics, it is important to understand these different forms of data because the sorts of questions that we try to answer with them, and the sorts of analyses that we can perform, are often very dependent on the form of the data.

Slide 3: Data Types



Data Types

3 of 18

Introduction

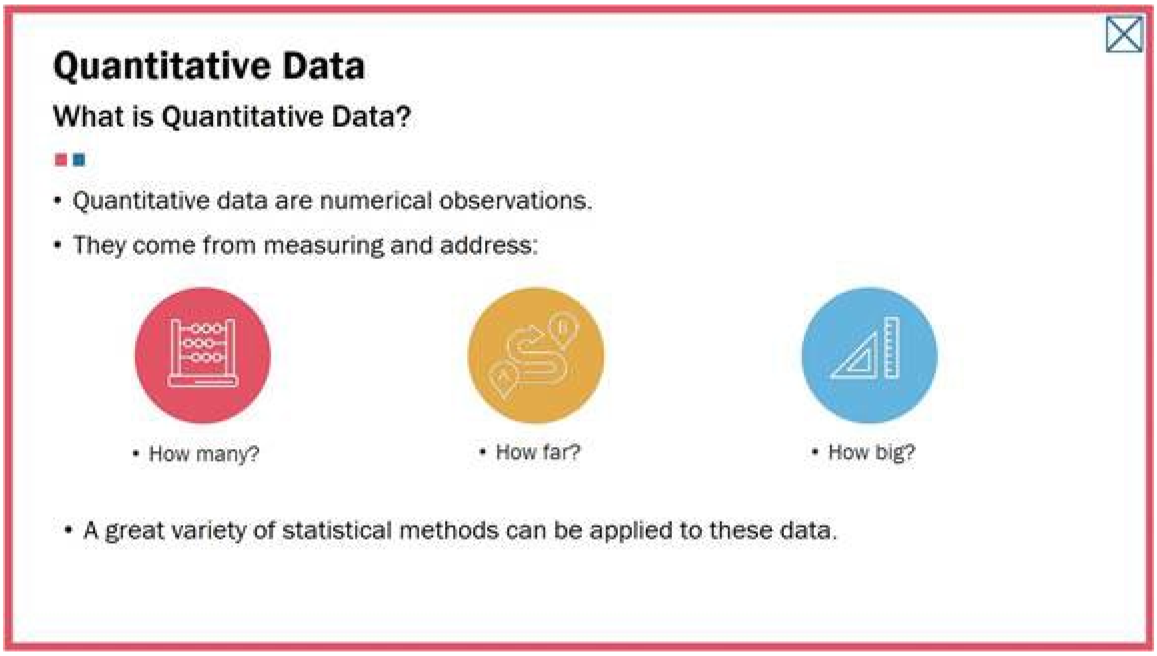
- Data is divided into two types:
  - Quantitative data are numerical
  - Qualitative data is everything else
- Data can also be a combination of these two types.

Click each tab to learn more. Then, click Next to continue.

In statistics, the big division in types of data is between what we call quantitative and qualitative. Broadly speaking, quantitative data are numerical, and qualitative data is everything else. We can have data that is a combination of these types as well.

Click the tabs to learn more about quantitative and qualitative types of data. When you are ready, click Next to continue.

Tab 1: Quantitative Data



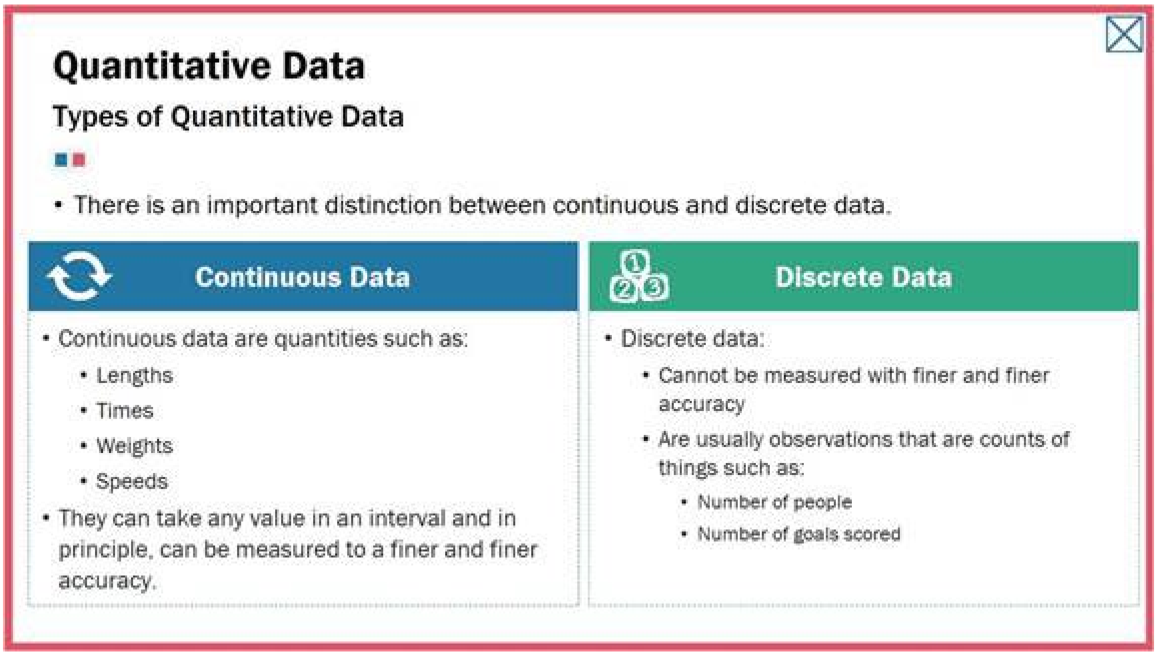
Quantitative Data

What is Quantitative Data?

- Quantitative data are numerical observations.
- They come from measuring and address:
  - How many?
  - How far?
  - How big?
- A great variety of statistical methods can be applied to these data.

Quantitative data are numerical observations and usually come from measuring something and so are answering a question like ‘how many?’, ‘how far?’, ‘how big?’, and so on. These sorts of data have the greatest variety of statistical methods that can be applied to them.

Tab 1.1: Types of Quantitative Data



Quantitative Data

Types of Quantitative Data

- There is an important distinction between continuous and discrete data.

Continuous Data

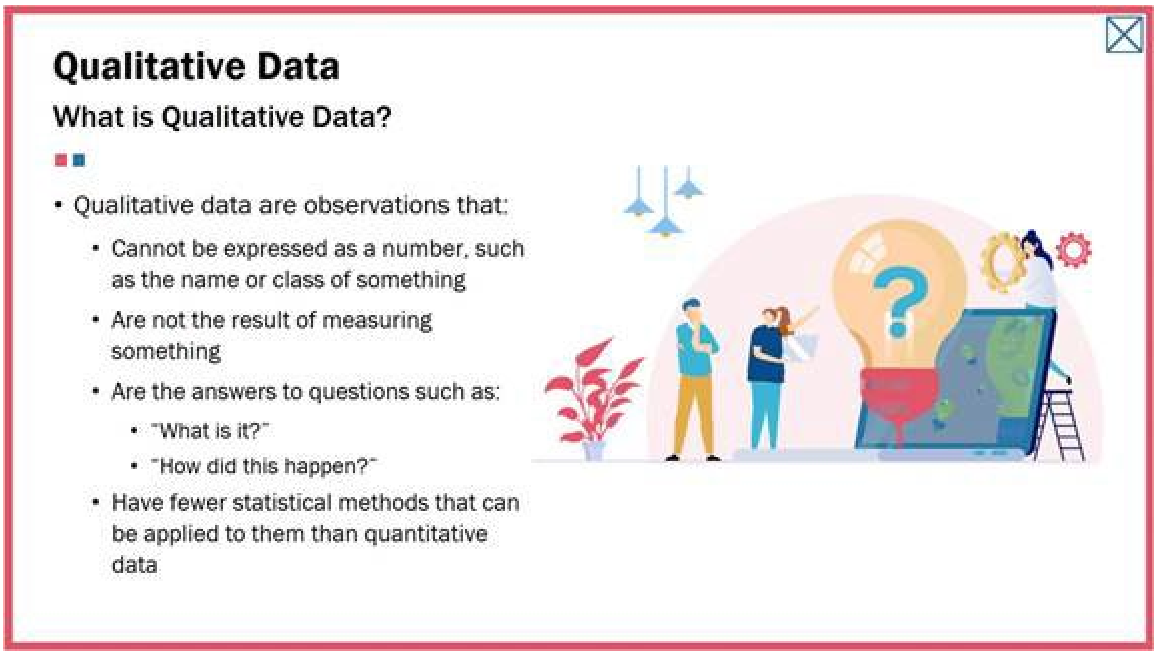
- Continuous data are quantities such as:
  - Lengths
  - Times
  - Weights
  - Speeds
- They can take any value in an interval and in principle, can be measured to a finer and finer accuracy.

Discrete Data

- Discrete data:
  - Cannot be measured with finer and finer accuracy
  - Are usually observations that are counts of things such as:
    - Number of people
    - Number of goals scored

Within quantitative data, an important distinction is between continuous and discrete data. Continuous data are quantities like lengths, times, weights, speeds and so on. What makes data on these quantities continuous is that they can take any value in an interval and, at least in principle, they can be measured to a finer and finer accuracy. In contrast, discrete data cannot be measured with finer and finer accuracy. They are usually observations that are counts of things such as a number of people, number of goals scored in a game etc.

Tab 2: Qualitative Data



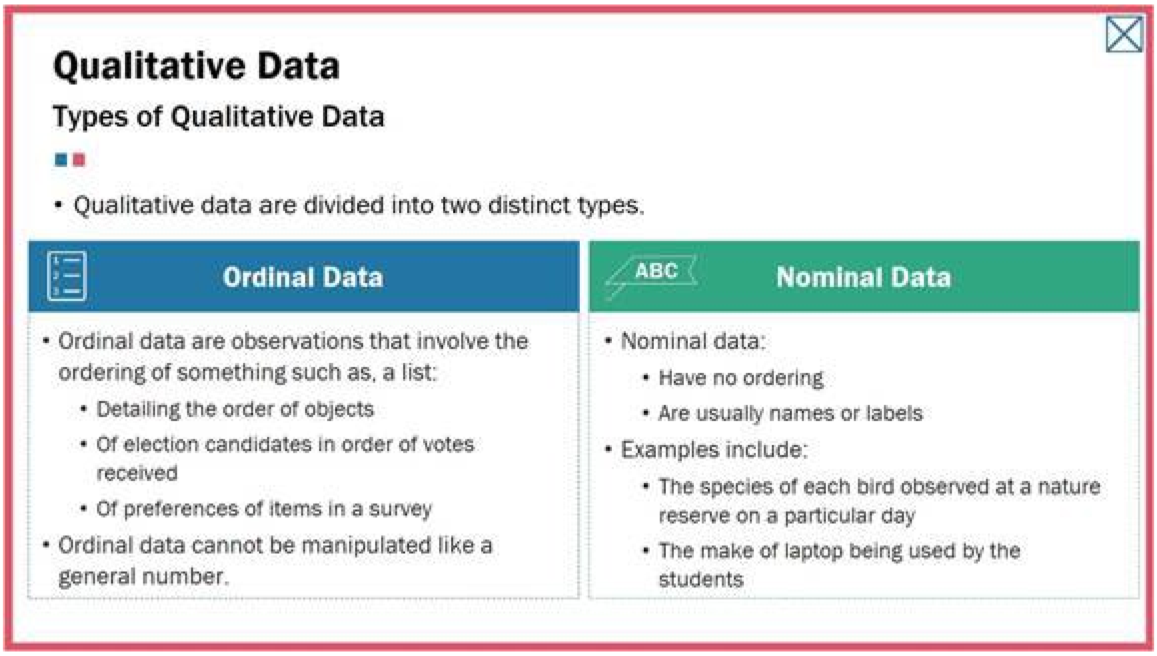
Qualitative Data

What is Qualitative Data?

- Qualitative data are observations that:
  - Cannot be expressed as a number, such as the name or class of something
  - Are not the result of measuring something
  - Are the answers to questions such as:
    - "What is it?"
    - "How did this happen?"
  - Have fewer statistical methods that can be applied to them than quantitative data

Qualitative data are observations that cannot be expressed as a number, such as the name or class of something. Such observations usually are not the result of measuring something and are the answers to questions such as 'what is it?' or 'how did this happen?'. These sorts of data have fewer statistical methods that can be applied to them, but these days that still means that there are impressive statistical tools that can be used to analyse, learn and make decisions based on such data.

Tab 2.1: Types of Qualitative Data



Qualitative Data

Types of Qualitative Data

- Qualitative data are divided into two distinct types.

Ordinal Data

- Ordinal data are observations that involve the ordering of something such as, a list:
  - Detailing the order of objects
  - Of election candidates in order of votes received
  - Of preferences of items in a survey
- Ordinal data cannot be manipulated like a general number.

Nominal Data

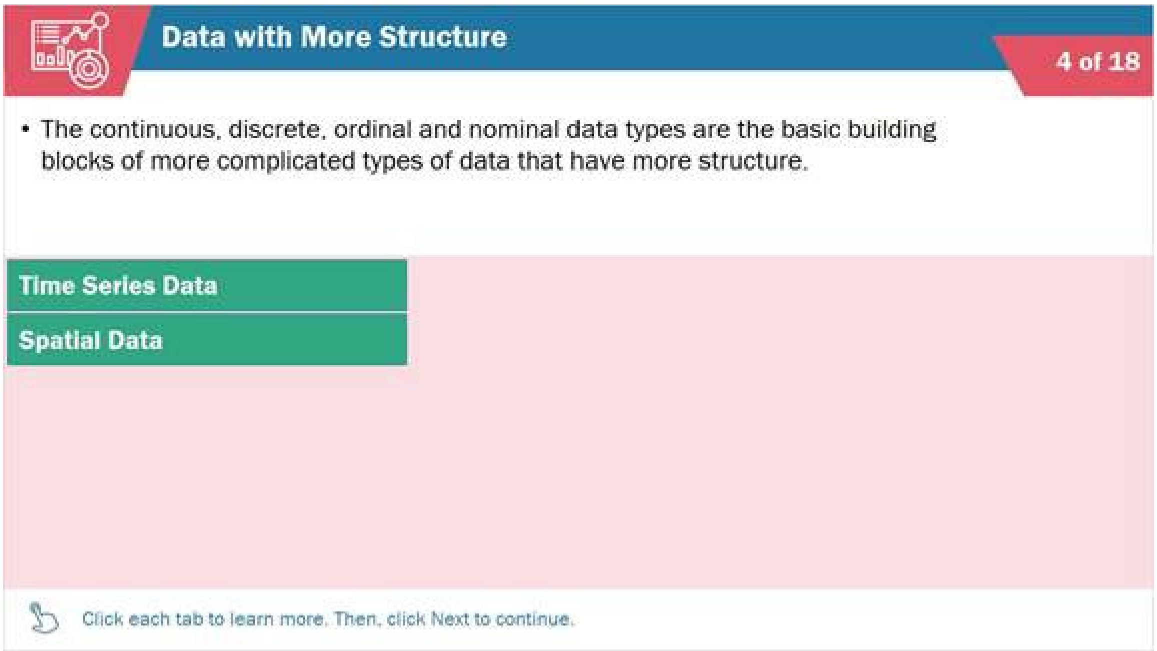
- Nominal data:
  - Have no ordering
  - Are usually names or labels
- Examples include:
  - The species of each bird observed at a nature reserve on a particular day
  - The make of laptop being used by the students

Qualitative data are also divided into two distinct types: Ordinal data and nominal data.

Ordinal data are observations that involve the ordering of something. It could be a list with the order of objects, such as a list of election candidates in the order of the number of votes that they received. Or it could be a list of things with a preference to them, such as a customer feedback survey for a hotel where things like cleanliness, breakfast and so on are measured on a five-point scale “very bad”, “bad”, ”neutral”, “good” and “very good”. Ordinal data may have numbers associated with them like the scale I’ve just mentioned, which could equally have been given as one, two, three, four or five, with five being very good, but we cannot manipulate them like a number by say multiplying or dividing them as it makes no sense.

Nominal data have no such ordering in mind. They are usually just names or labels, for example the species of each bird observed at a nature reserve on a particular day, or the make of laptop used by the students doing this session.

Slide 4: Data with More Structure



Data with More Structure

4 of 18

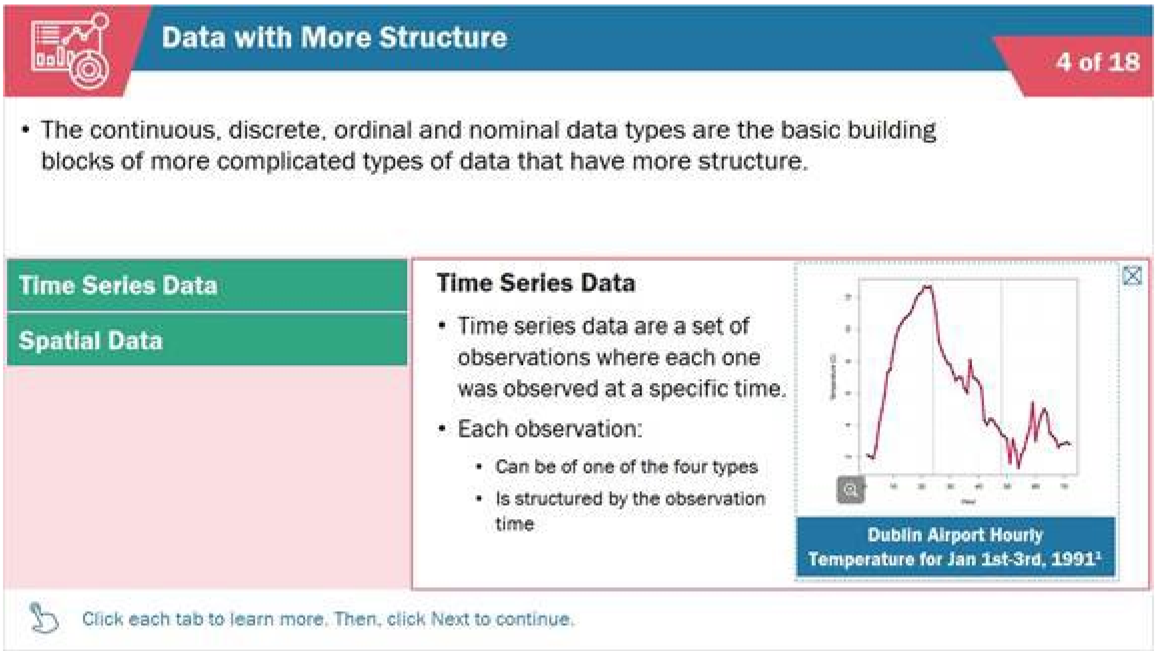
- The continuous, discrete, ordinal and nominal data types are the basic building blocks of more complicated types of data that have more structure.

Click each tab to learn more. Then, click Next to continue.

The four data types that we’ve just discussed – continuous, discrete, ordinal, nominal – can be thought of as the basic building blocks of more complicated types of data that have more structure such as time series and spatial data.

Click the tabs to find out about these data with more structure. When you are ready click Next to continue

Tab 1: Time Series Data



Data with More Structure

4 of 18

- The continuous, discrete, ordinal and nominal data types are the basic building blocks of more complicated types of data that have more structure.

Time Series Data

Spatial Data

Time Series Data

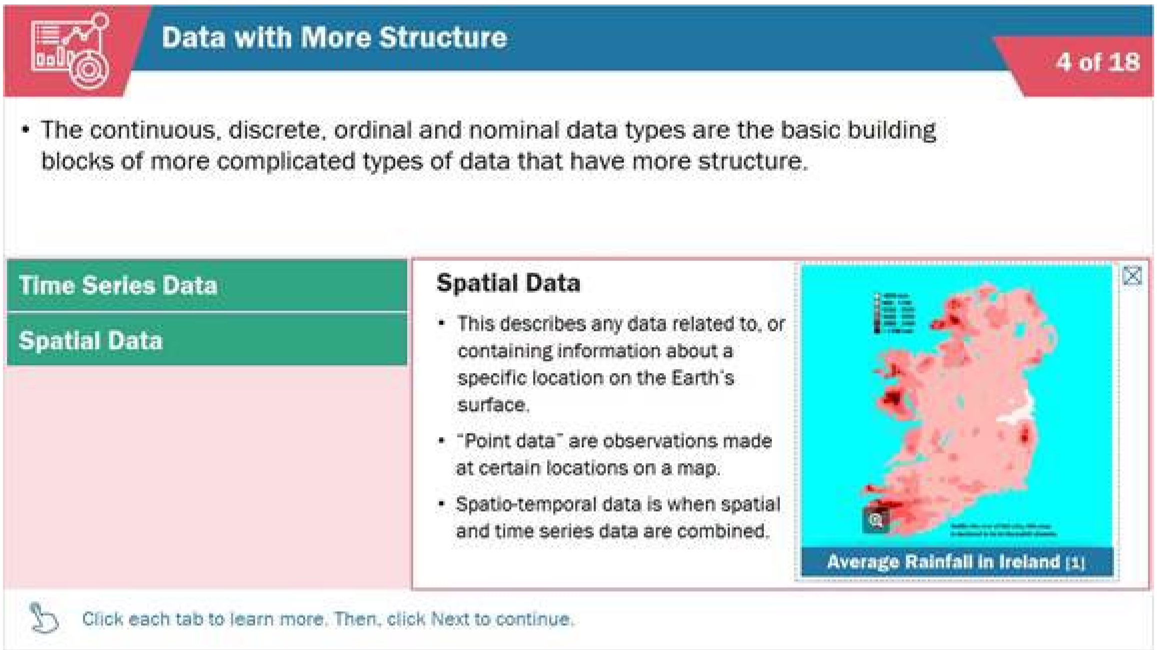
- Time series data are a set of observations where each one was observed at a specific time.
- Each observation:
  - Can be of one of the four types
  - Is structured by the observation time

Dublin Airport Hourly Temperature for Jan 1st-3rd, 1991[1]

Click each tab to learn more. Then, click Next to continue.

Time series data are a set of observations where each one was observed at a specific time. Each observation could be of one of the four types, but on top of that they are structured by the observation time, which also carries important information. Think of the data presented here on hourly temperature. Each temperature observation is continuous but the time ordering is important. If they were plotted from left to right in random order rather than chronologically; clearly a lot of information in the data would be lost.

Tab 2: Spatial Data



Data with More Structure

4 of 18

- The continuous, discrete, ordinal and nominal data types are the basic building blocks of more complicated types of data that have more structure.

Time Series Data

Spatial Data

Spatial Data

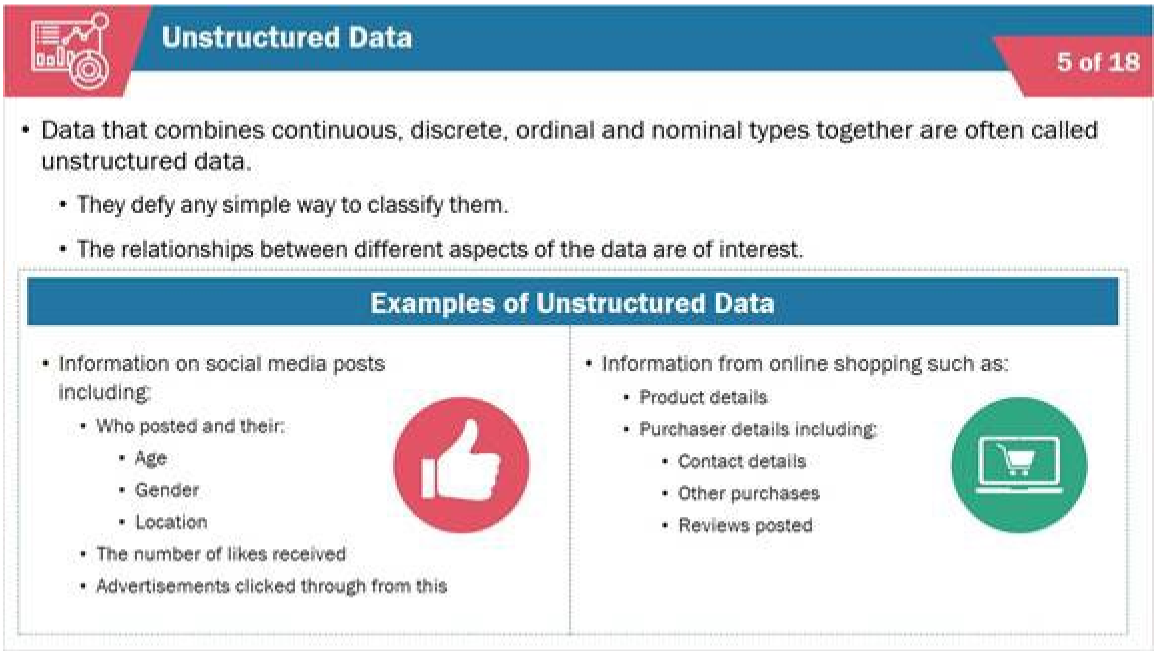
- This describes any data related to, or containing information about a specific location on the Earth's surface.
- "Point data" are observations made at certain locations on a map.
- Spatio-temporal data is when spatial and time series data are combined.

Average Rainfall in Ireland [1]

Click each tab to learn more. Then, click Next to continue.

Spatial data describes any data related to or containing information about a specific location on the Earth's surface or elsewhere. In the figure here, we have an example of image data, where each location in Ireland has a value associated with it. Other sorts of spatial data are 'point' data, where observations are only made at certain locations on a map. There is also spatio-temporal data, where spatial and time series data are combined. One can think about spatio-temporal data as like a video, with a sequence of images changing over time.

Slide 5: Unstructured Data



Unstructured Data

5 of 18

- Data that combines continuous, discrete, ordinal and nominal types together are often called unstructured data.
  - They defy any simple way to classify them.
  - The relationships between different aspects of the data are of interest.

Examples of Unstructured Data

- Information on social media posts including:
  - Who posted and their:
    - Age
    - Gender
    - Location
  - The number of likes received
  - Advertisements clicked through from this

- Information from online shopping such as:
  - Product details
  - Purchaser details including:
    - Contact details
    - Other purchases
    - Reviews posted

Moving beyond time and spatial data, in recent decades we have seen much data that combines continuous, discrete, ordinal and nominal types together. These data are often called unstructured, as they defy any simple way to classify them. Examples include: Social media posts along with information about the person who posted it such as age, gender, location, who else read it, how many likes it receives, which ads were clicked through from the post, etc.; or product details from an online shopping company, along with who bought it and their personal details, what else they bought, reviews of the product and so on.

These two examples reflect the fact that much of this sort of data has arisen out of the rise in importance of information technology over the last number of years. Often it is the relationships between different aspects of the data that are of interest (for example, relating the posts on a social media site that a user likes to what sorts of ads are most likely to interest them). These are complex questions and the subject of much work in statistics and computer science at the moment, and also well beyond the scope of this module.

Slide 6: Section: Summary Statistics

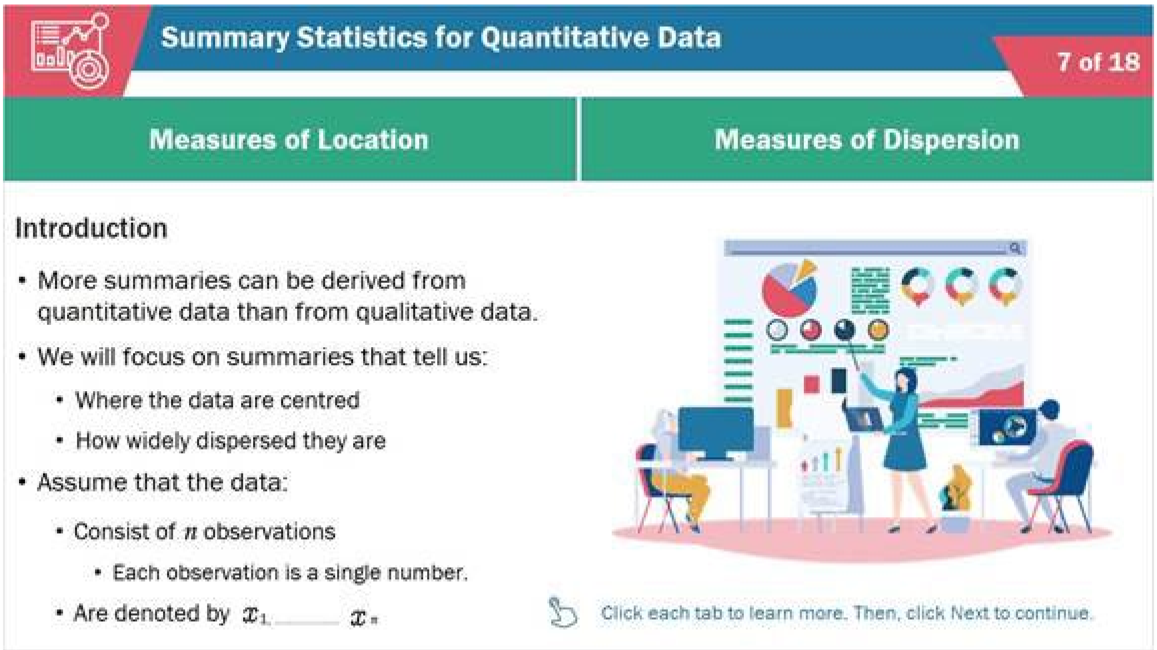
Summary Statistics

6 of 18

- Summary statistics are values derived from the data that help us to understand it better.

We are now going to look at summary statistics. Summary statistics are values that we derive from the data that help us to understand it better. We will look at summary statistics for both quantitative and qualitative data.

Slide 7: Summary Statistics for Quantitative Data



Summary Statistics for Quantitative Data

7 of 18

Measures of Location | Measures of Dispersion

Introduction

- More summaries can be derived from quantitative data than from qualitative data.
- We will focus on summaries that tell us:
  - Where the data are centred
  - How widely dispersed they are
- Assume that the data:
  - Consist of n observations
    - Each observation is a single number.
  - Are denoted by $x_1, \ldots, x_n$

Click each tab to learn more. Then, click Next to continue.

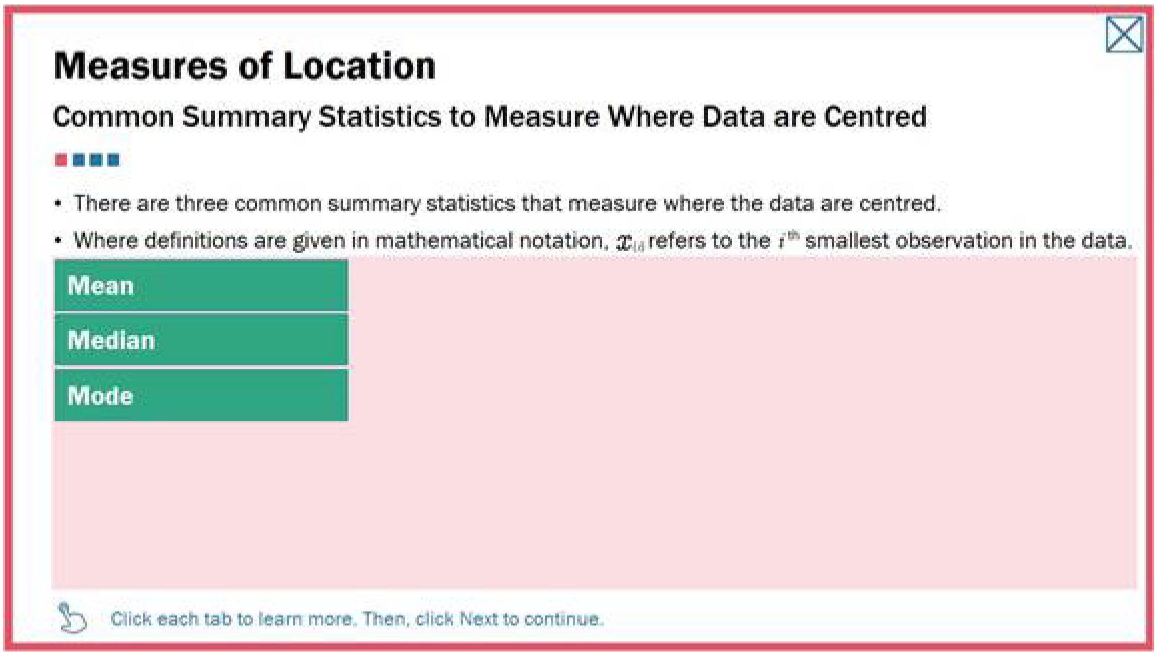
There are many more summaries that one can derive in the quantitative case than the qualitative case. In this presentation, we will focus on summaries that tell us something about where the data are centered and how widely dispersed they are. We will use some notation to define these summaries: we assume that our data consist of n observations, each observation is a single number and they are denoted x_1 up to x_n.

We are going to concentrate on two summaries for quantitative data:

- Measures of location; and
- Measures of dispersion

Click the tabs to learn about each one. When you are ready, click Next to continue.

Tab 1: Measures of Location



Measures of Location

Common Summary Statistics to Measure Where Data are Centred

- There are three common summary statistics that measure where the data are centred.
- Where definitions are given in mathematical notation, $x_{(i)}$ refers to the i^{th} smallest observation in the data.

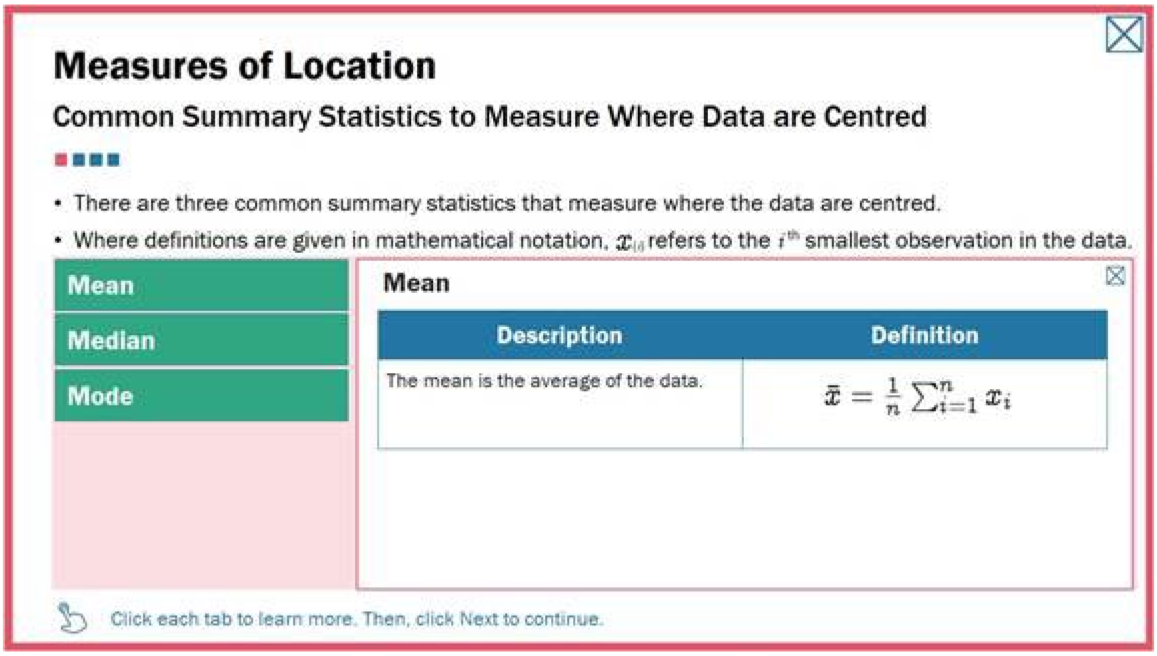
Click each tab to learn more. Then, click Next to continue.

Here we will look at the three most common summary statistics of quantitative data that measure where the data are centered, so-called measures of location, along with their definition. Where definitions are given in the mathematical notation, note that $x_{(i)}$ refers to the i^{th} smallest observation in the data.

Click the tabs to learn about each statistic. When you are ready click next to continue.

Tab 1: Mean: The mean is the average of the data, that is to say their sum divided by the number of observations, n.

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$



Measures of Location

Common Summary Statistics to Measure Where Data are Centred

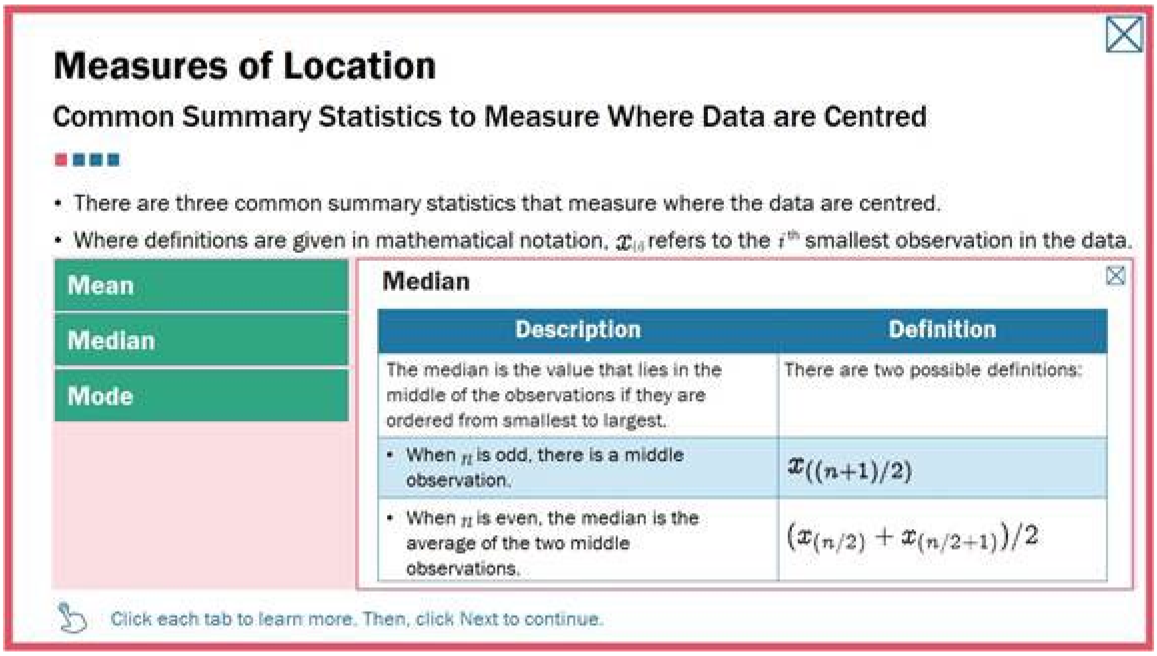
- There are three common summary statistics that measure where the data are centred.
- Where definitions are given in mathematical notation, $x_{(i)}$ refers to the i^{th} smallest observation in the data.

Mean

Description	Definition
The mean is the average of the data.	$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$

Click each tab to learn more. Then, click Next to continue.

Tab 2: Median: The median is the value that lies in the middle of the observations if they are ordered from smallest to largest.



Measures of Location

Common Summary Statistics to Measure Where Data are Centred

- There are three common summary statistics that measure where the data are centred.
- Where definitions are given in mathematical notation, $x_{(i)}$ refers to the i^{th} smallest observation in the data.

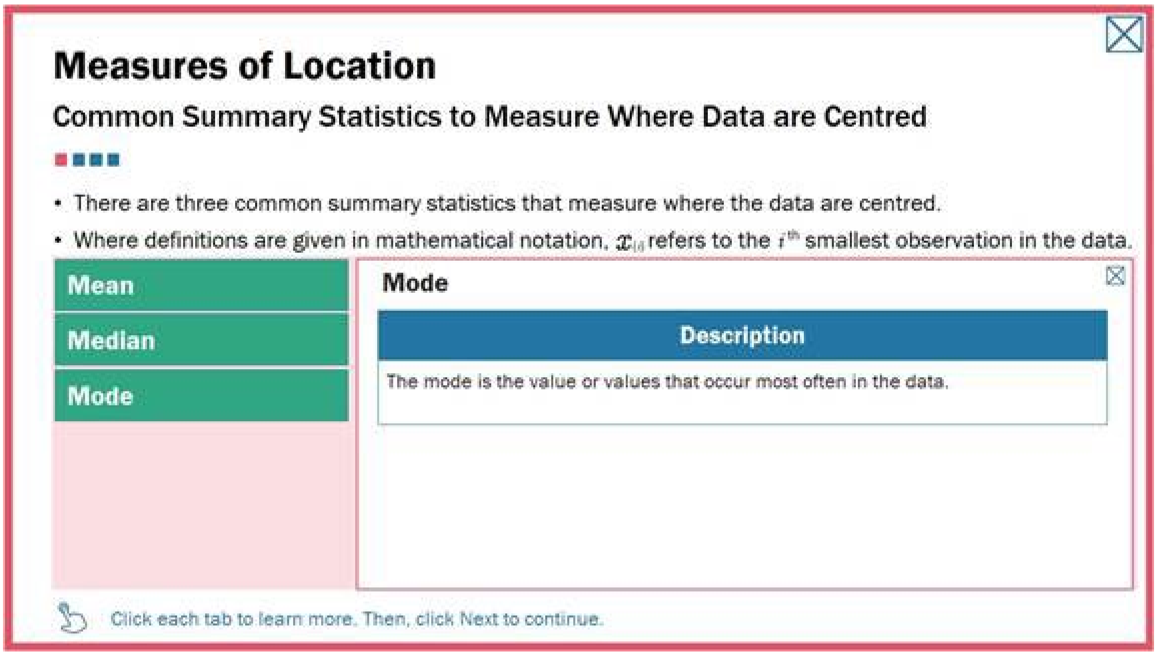
Median

Description	Definition
The median is the value that lies in the middle of the observations if they are ordered from smallest to largest.	There are two possible definitions:
• When n is odd, there is a middle observation.	$x_{((n+1)/2)}$
• When n is even, the median is the average of the two middle observations.	$(x_{(n/2)} + x_{(n/2+1)})/2$

Click each tab to learn more. Then, click Next to continue.

The median is calculated slightly differently depending on whether there are an odd or even number of observations. When n is odd, there is a middle observation (it is the (n+1)/2nd in the ordered list). However, when n is even, there are 2 middle observations (the n/2th and n/2 + 1th in the ordered list). In this case we take the average of the 2 middle observations as our median.

Tab 3: Mode: The mode is the value or values that occur most often in the data.



Measures of Location

Common Summary Statistics to Measure Where Data are Centred

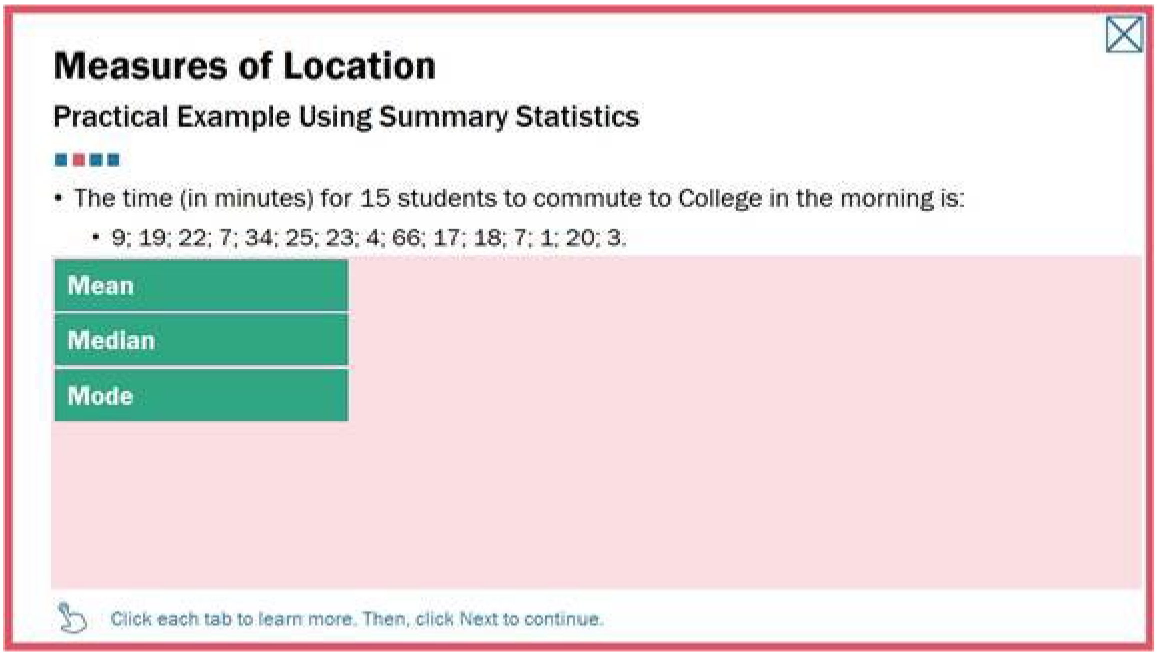
- There are three common summary statistics that measure where the data are centred.
- Where definitions are given in mathematical notation, $x_{(i)}$ refers to the i^{th} smallest observation in the data.

Mode

Description
The mode is the value or values that occur most often in the data.

Click each tab to learn more. Then, click Next to continue.

Tab 1.1: Practical Example Using Summary Statistics



Measures of Location

Practical Example Using Summary Statistics

- The time (in minutes) for 15 students to commute to College in the morning is:
  - 9; 19; 22; 7; 34; 25; 23; 4; 66; 17; 18; 7; 1; 20; 3.

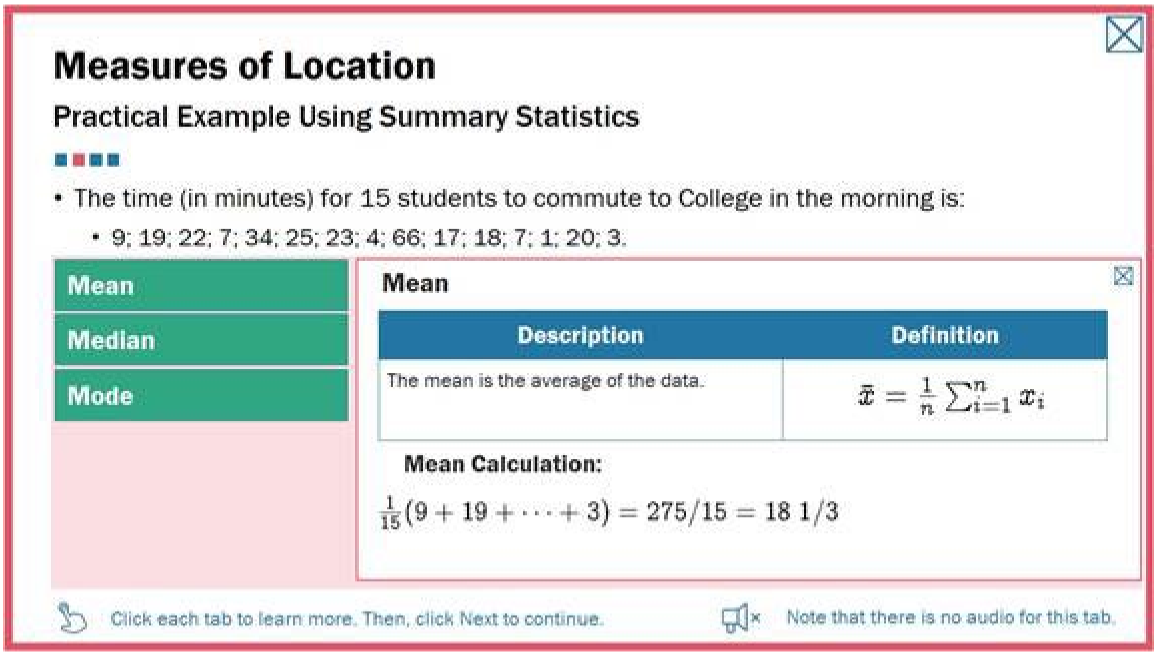
Click each tab to learn more. Then, click Next to continue.

Here is an example for you to work on. You are told that 15 students commute to college and you are given the time taken by each student to reach the college. Calculate the mean, median and mode for this data.

The travel times for each student are: 9; 19; 22; 7; 34; 25; 23; 4; 66; 17; 18; 7; 1; 20; 3.

When you reach your answer, click on the tabs to check your answers and confirm how the summary statistics are calculated.

Tab 1: Mean Calculation:



Measures of Location

Practical Example Using Summary Statistics

- The time (in minutes) for 15 students to commute to College in the morning is:
  - 9; 19; 22; 7; 34; 25; 23; 4; 66; 17; 18; 7; 1; 20; 3.

Mean

Description	Definition
The mean is the average of the data.	$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$

Mean Calculation:

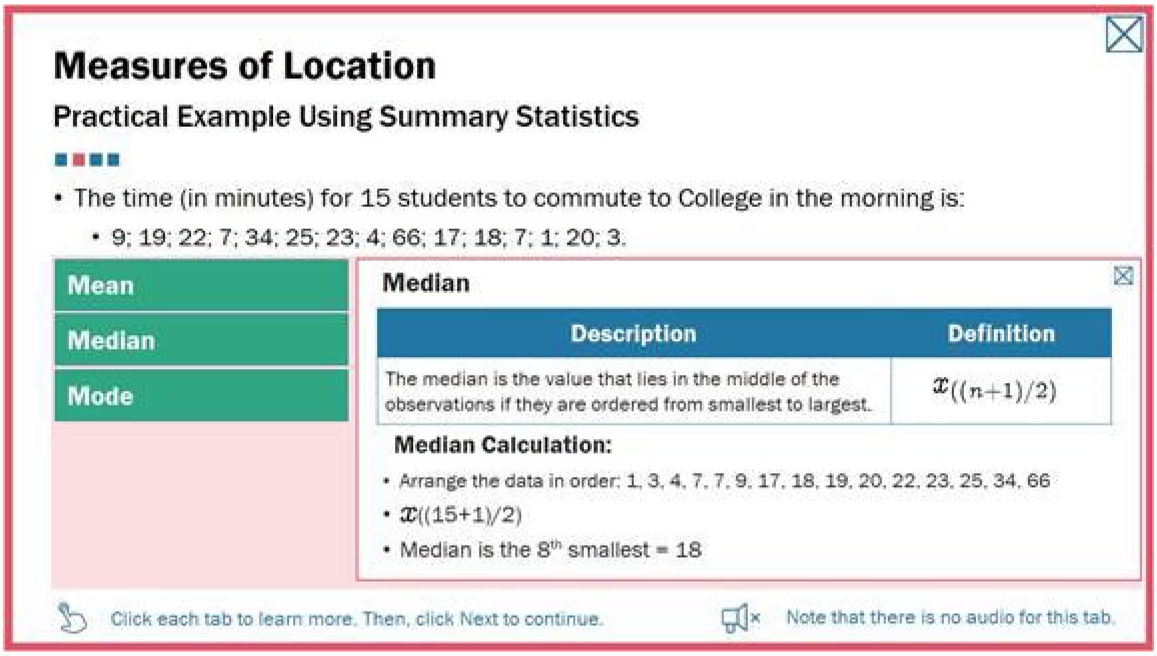
$\frac{1}{15}(9 + 19 + \cdots + 3) = 275/15 = 18\ 1/3$

Click each tab to learn more. Then, click Next to continue.

Note that there is no audio for this tab.

$$\frac{1}{15}(9+19+22+7+34+25+23+4+66+17+18+7+1+20+3) = \frac{275}{15}$$
$$= 18\frac{1}{3}$$

Tab 2: Median Calculation:



Measures of Location

Practical Example Using Summary Statistics

- The time (in minutes) for 15 students to commute to College in the morning is:
  - 9; 19; 22; 7; 34; 25; 23; 4; 66; 17; 18; 7; 1; 20; 3.

Median

Description	Definition
The median is the value that lies in the middle of the observations if they are ordered from smallest to largest.	$x_{((n+1)/2)}$

Median Calculation:

- Arrange the data in order: 1, 3, 4, 7, 7, 9, 17, 18, 19, 20, 22, 23, 25, 34, 66
- $x_{((15+1)/2)}$
- Median is the 8th smallest = 18

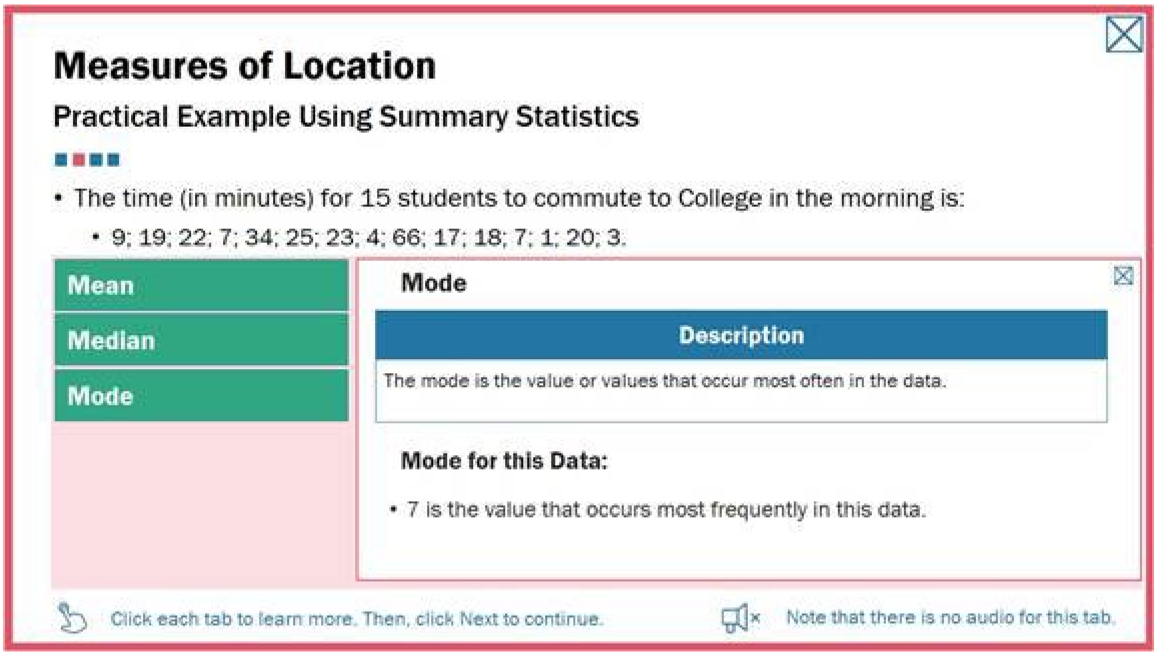
Click each tab to learn more. Then, click Next to continue.

Note that there is no audio for this tab.

Arrange the data in order: 1, 3, 4, 7, 7, 9, 17, 18, 19, 20, 22, 23, 25, 34, 66

$x\left(\frac{15+1}{2}\right)$= The 8th smallest value which is 18.

Tab 3: Mode:



Measures of Location

Practical Example Using Summary Statistics

- The time (in minutes) for 15 students to commute to College in the morning is:
  - 9; 19; 22; 7; 34; 25; 23; 4; 66; 17; 18; 7; 1; 20; 3.

Mode

Description
The mode is the value or values that occur most often in the data.

Mode for this Data:

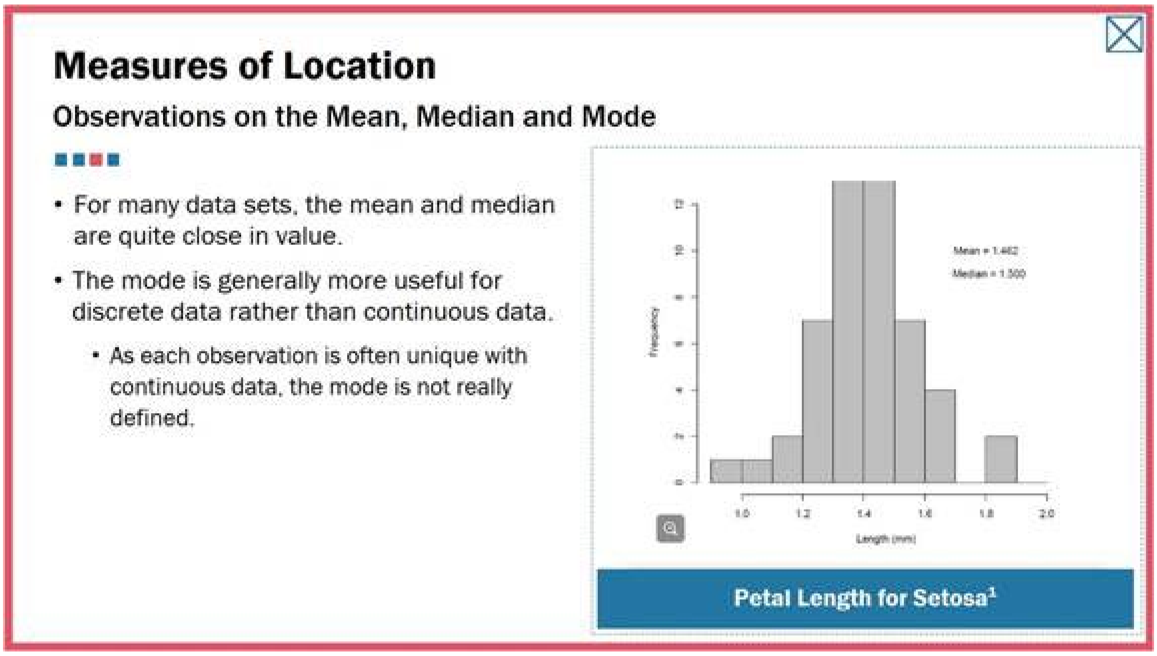
- 7 is the value that occurs most frequently in this data.

Click each tab to learn more. Then, click Next to continue.

Note that there is no audio for this tab.

7 is the value that occurs most frequently in this data.

Tab 1.2: Observations on the Mean, Median and Mode



Measures of Location

Observations on the Mean, Median and Mode

- For many data sets, the mean and median are quite close in value.
- The mode is generally more useful for discrete data rather than continuous data.
  - As each observation is often unique with continuous data, the mode is not really defined.

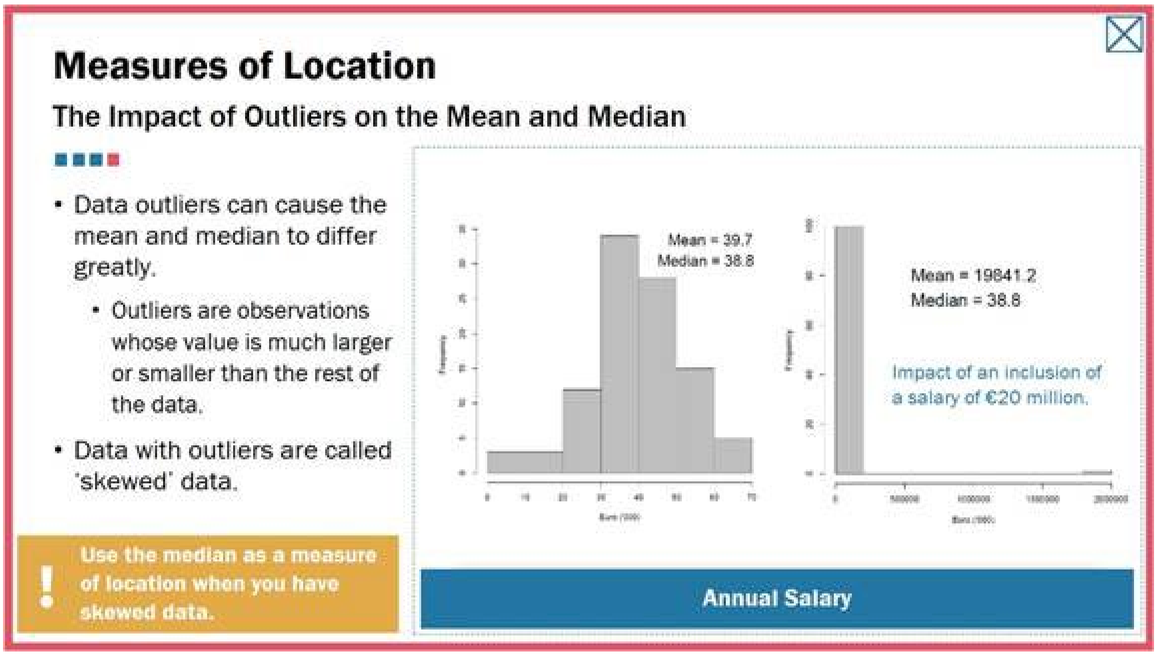
Petal Length for Setosa[1]

For many data sets, the mean and median are quite close in value. The figure on this slide shows an example where this is the case. It comes from a famous data set that measures properties of some different species of iris flowers. We'll use this data set again in this presentation. This plot is a histogram of the length of the petals of 50 specimens of a species called setosa. We'll define the histogram later but for now we note that it counts the number of observations in each of a set of intervals or bins. The mean and median of these lengths turn out to be very close together. The mean is

1.462 and the median is 1.500. Note that in this data set there are lots of values close to the mean and median, with a rather symmetric distribution of values about them.

One other thing to note is about the mode. It is generally more useful for discrete data, because with continuous data each observation is often unique and so the mode is not really defined.

Tab 1.3: The Impact of Outliers on the Mean and Median



Measures of Location

The Impact of Outliers on the Mean and Median

- Data outliers can cause the mean and median to differ greatly.
  - Outliers are observations whose value is much larger or smaller than the rest of the data.
- Data with outliers are called 'skewed' data.

! Use the median as a measure of location when you have skewed data.

Mean = 39.7
Median = 38.8

Mean = 19841.2
Median = 38.8

Impact of an inclusion of a salary of €20 million.

Annual Salary

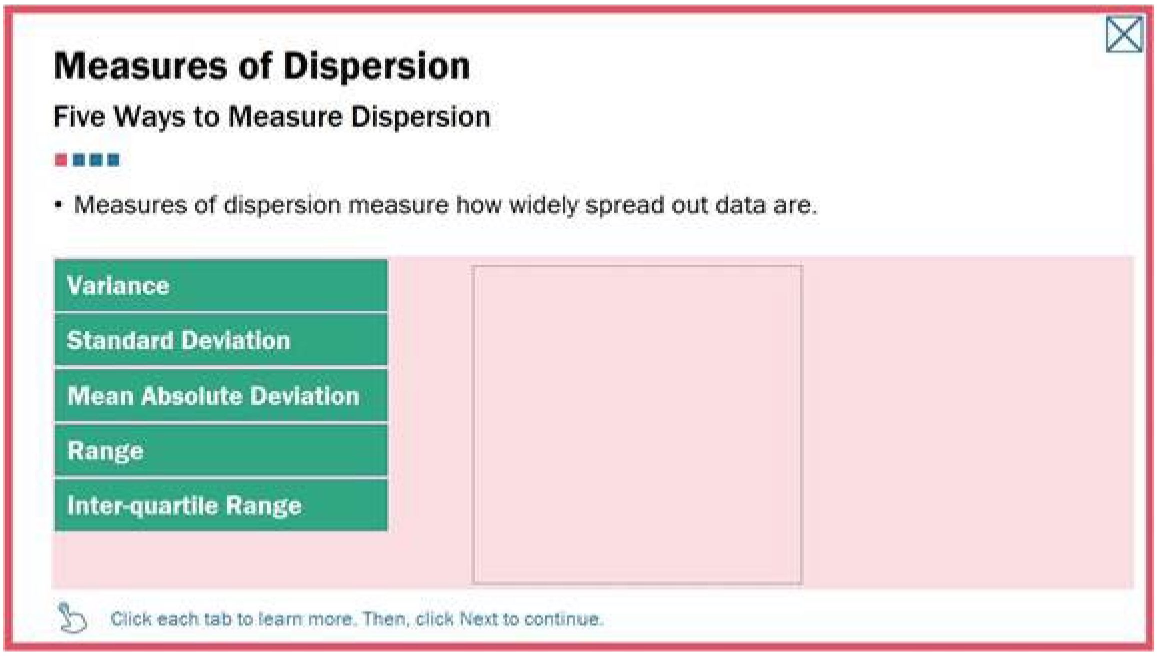
Finally, for measures of location, we ask the question as to when the mean and median can be quite different. One common way is when the data contain outliers, that is, observations whose values are much larger or smaller than the rest. In this slide we see an example where, initially, we have data where the mean and median are very close. The mean is 39.7 and the median is 38.8. Suppose there are 100 of us on this course and we look at our net income in the last year. Perhaps optimistically, we might get something like the data that we see in this figure where the mean and median are close, around €40,000. The highest of the 100 observations is around €70,000.

Now suppose that someone very wealthy enrols with an income that puts that person in the top 100 earners in the world (so a large income but also something that's not unrealistic and we expect to see in some individuals in most years).

Now what happens to the mean and the median? The mean is 19,841.2 and the median is 38.8. Well, as you can see in the second figure, the median barely moves; adding one more value to the data set, no matter how large, usually does not alter the value in the middle of the data very much at all. But the mean changes a lot and in fact it is about €20 million, many times larger than all of the observations except that of our very wealthy colleague on the course. The mean is not a good measure of the centre of the data in this case.

When we have data with outliers like this, the data are called skewed. In this case it is best practice to use the median as a measure of location.

Tab 2: **Measures of Dispersion**



Measures of Dispersion

Five Ways to Measure Dispersion

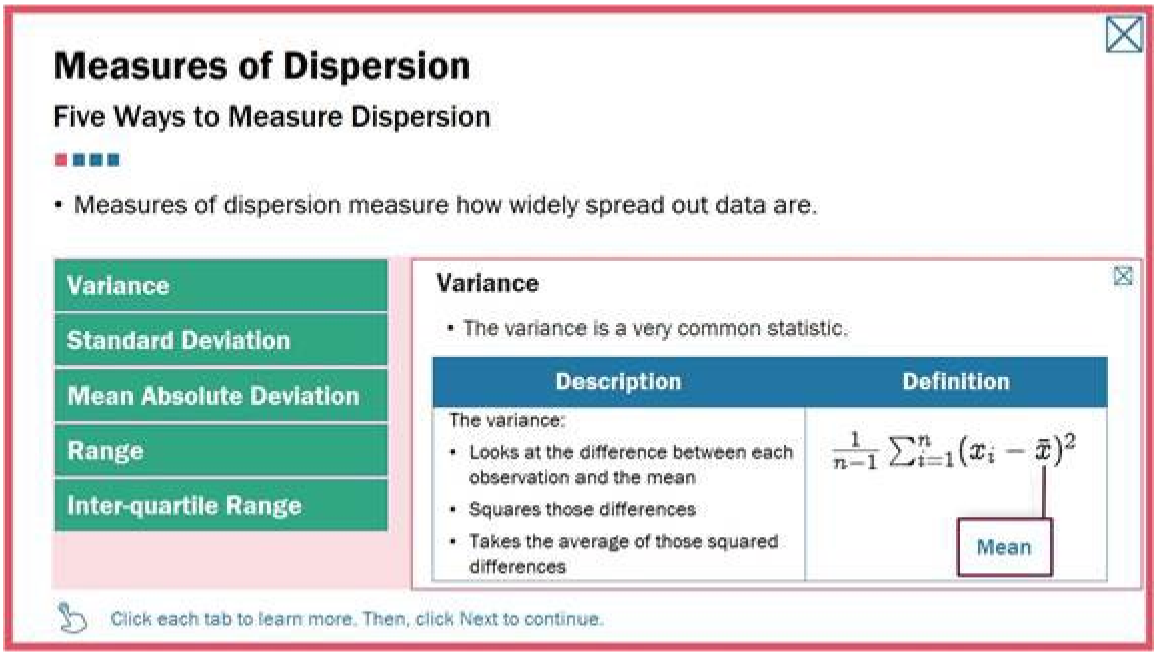
- Measures of dispersion measure how widely spread out data are.

Click each tab to learn more. Then, click Next to continue.

Now we move to summary statistics that measure how widely spread out the data are, so-called measures of dispersion. In this slide, we see five such measures.

Click the tabs to view each measure. When you are ready, click Next to continue.

Tab 1: Variance: The variance is a very common statistic and its definition needs some explanation.



Measures of Dispersion

Five Ways to Measure Dispersion

- Measures of dispersion measure how widely spread out data are.

Variance

- The variance is a very common statistic.

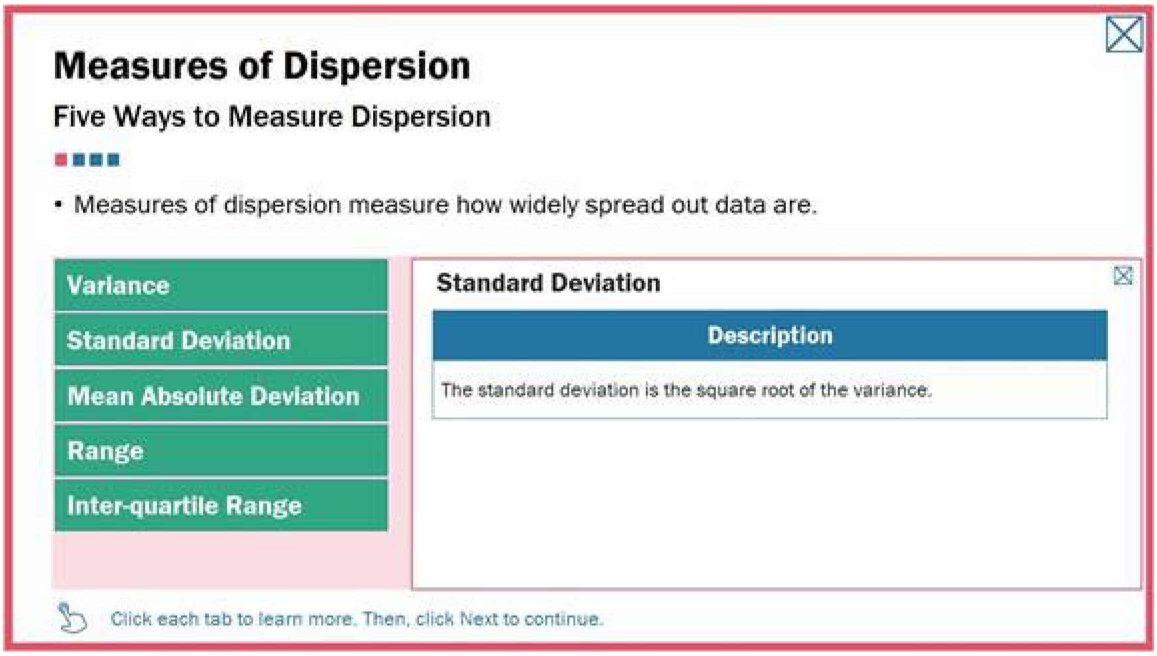
Description	Definition
The variance: • Looks at the difference between each observation and the mean • Squares those differences • Takes the average of those squared differences	$\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$ Mean

Click each tab to learn more. Then, click Next to continue.

Definition: $\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$

In the definition, first note that you can see some notation of an ‘x’ with a bar above it; this denotes the mean value of the data. The variance looks at the difference between each observation and the mean and squares those differences (so a positive and negative difference both contribute the same to the sum). It then takes the average of those squared differences (well, for technical reasons, it divides the sum by n-1 and not n) but if n is reasonably large, that difference doesn't matter very much.

Tab 2: Standard Deviation: The standard deviation is just the square root of the variance.



Measures of Dispersion

Five Ways to Measure Dispersion

- Measures of dispersion measure how widely spread out data are.

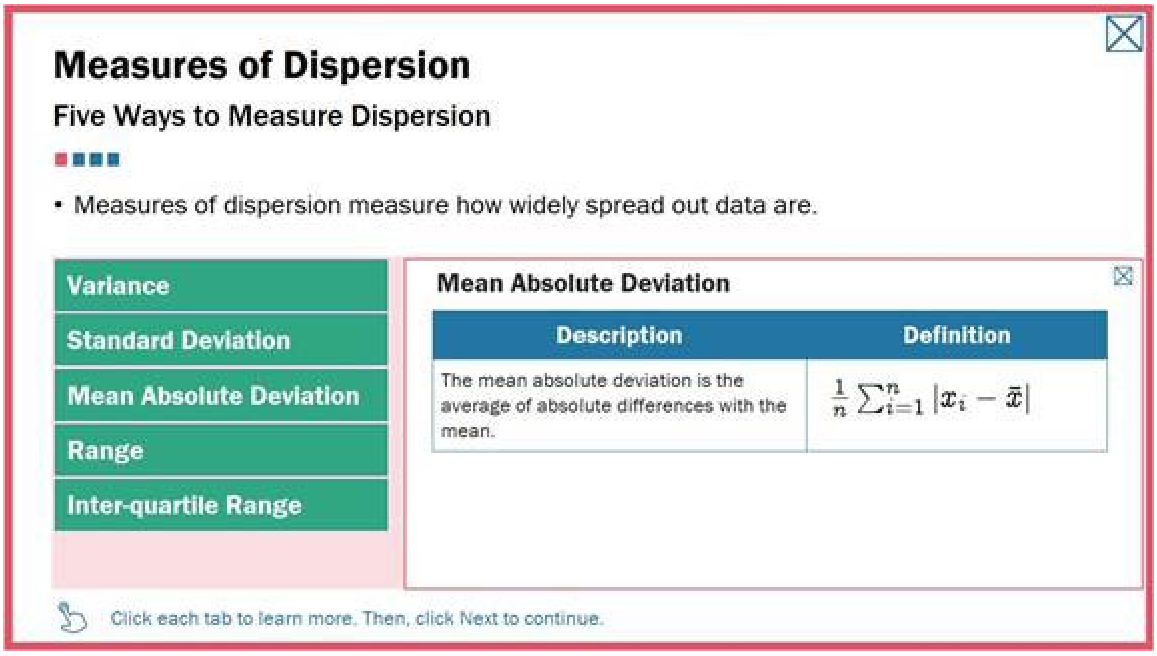
Standard Deviation

Description
The standard deviation is the square root of the variance.

Click each tab to learn more. Then, click Next to continue.

Tab 3: Mean Absolute Deviation: Similar to the variance, the mean absolute deviation is the average of absolute differences with the mean, as opposed to square differences.

Definition: $\frac{1}{n}\sum_{i-1}^{n}|x_i - \bar{x}|$



Measures of Dispersion

Five Ways to Measure Dispersion

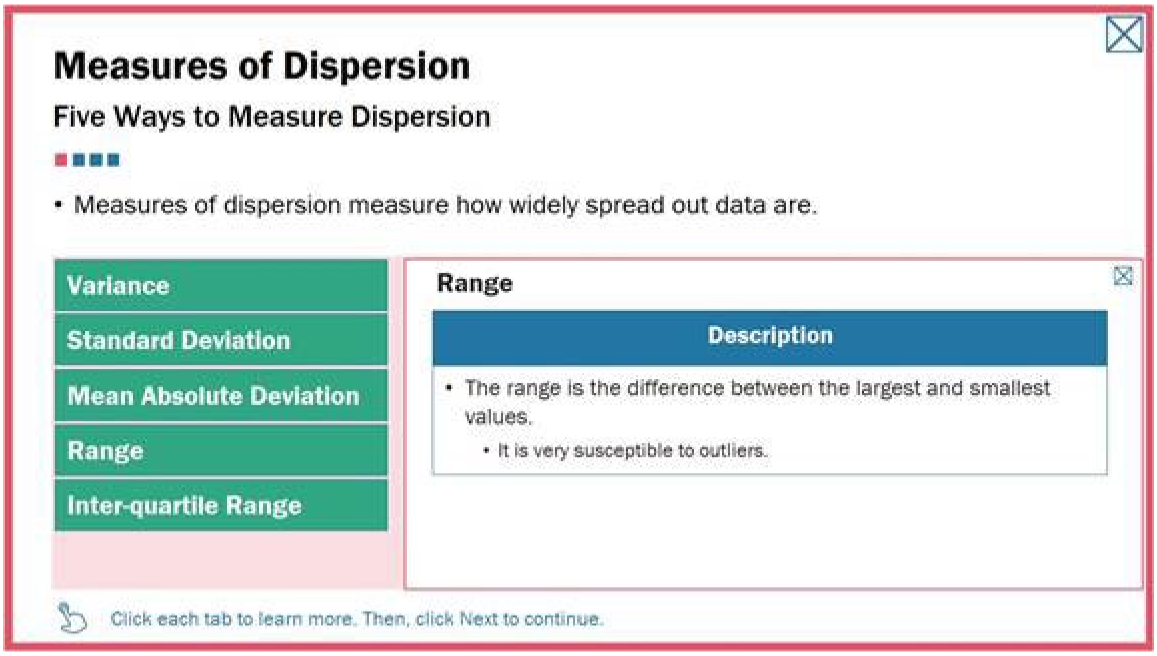
- Measures of dispersion measure how widely spread out data are.

Mean Absolute Deviation

Description	Definition
The mean absolute deviation is the average of absolute differences with the mean.	$\frac{1}{n}\sum_{i=1}^{n}\lvert x_i - \bar{x}\rvert$

Click each tab to learn more. Then, click Next to continue.

Tab 4: Range: The range is easy, just the difference between the largest and smallest values. Like the mean, the range is also very susceptible to outliers (think what happens to it with our salary example from the last slide).



Measures of Dispersion

Five Ways to Measure Dispersion

- Measures of dispersion measure how widely spread out data are.

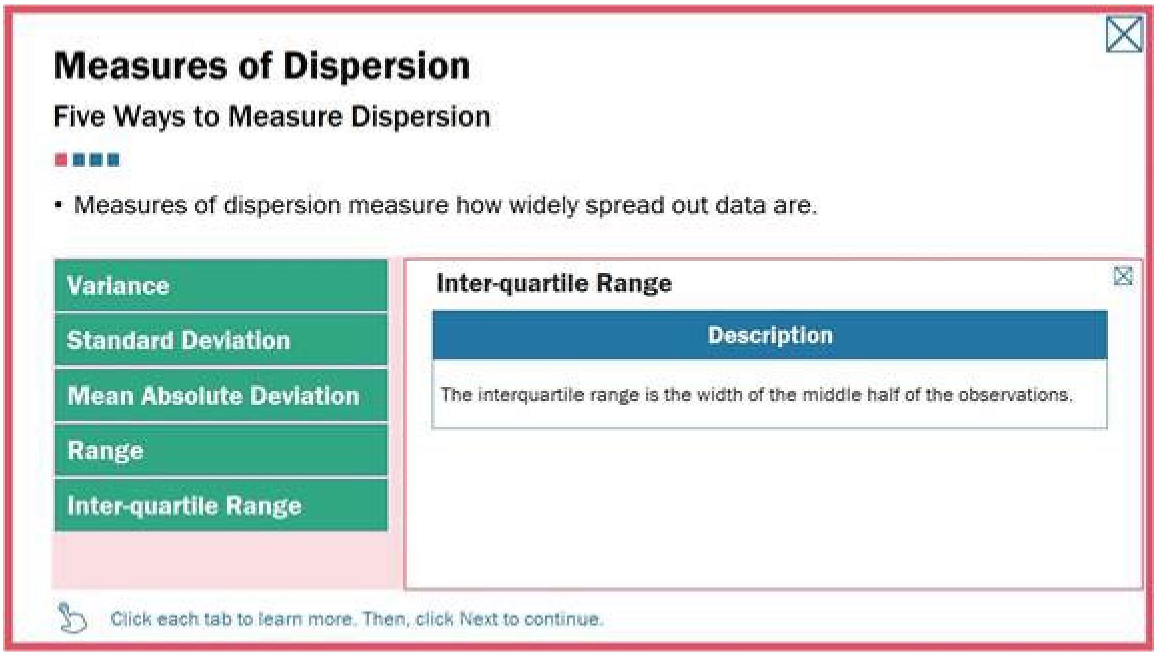
Range

Description

- The range is the difference between the largest and smallest values.
  - It is very susceptible to outliers.

Click each tab to learn more. Then, click Next to continue.

Tab 5: Inter-quartile Range: Because of the range's susceptibility to outliers, we use the inter-quartile range more often than the range. This is the width of the middle half of the observations, in other words, about one quarter of the observations have values smaller than the lower end of this interval and another quarter have values larger than the upper bound of the interval. This will be explained in more detail in the following slides.



Measures of Dispersion

Five Ways to Measure Dispersion

- Measures of dispersion measure how widely spread out data are.

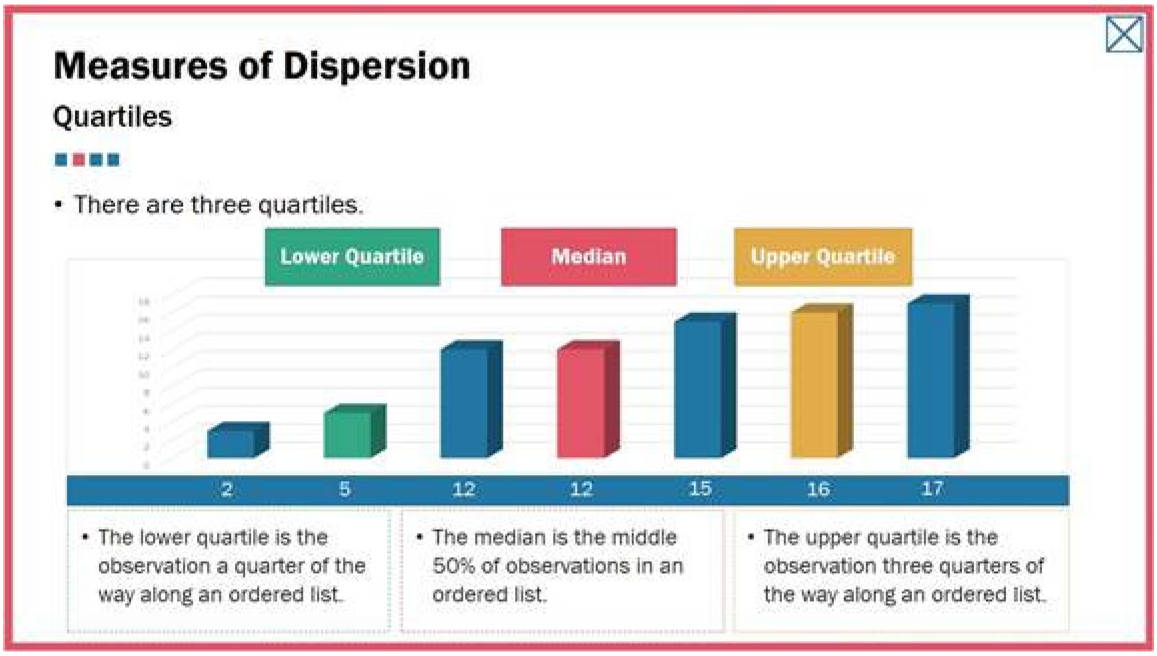
Inter-quartile Range

Description

The interquartile range is the width of the middle half of the observations.

Click each tab to learn more. Then, click Next to continue.

Tab 2.1: Quartiles



Measures of Dispersion

Quartiles

- There are three quartiles.

- The lower quartile is the observation a quarter of the way along an ordered list.
- The median is the middle 50% of observations in an ordered list.
- The upper quartile is the observation three quarters of the way along an ordered list.

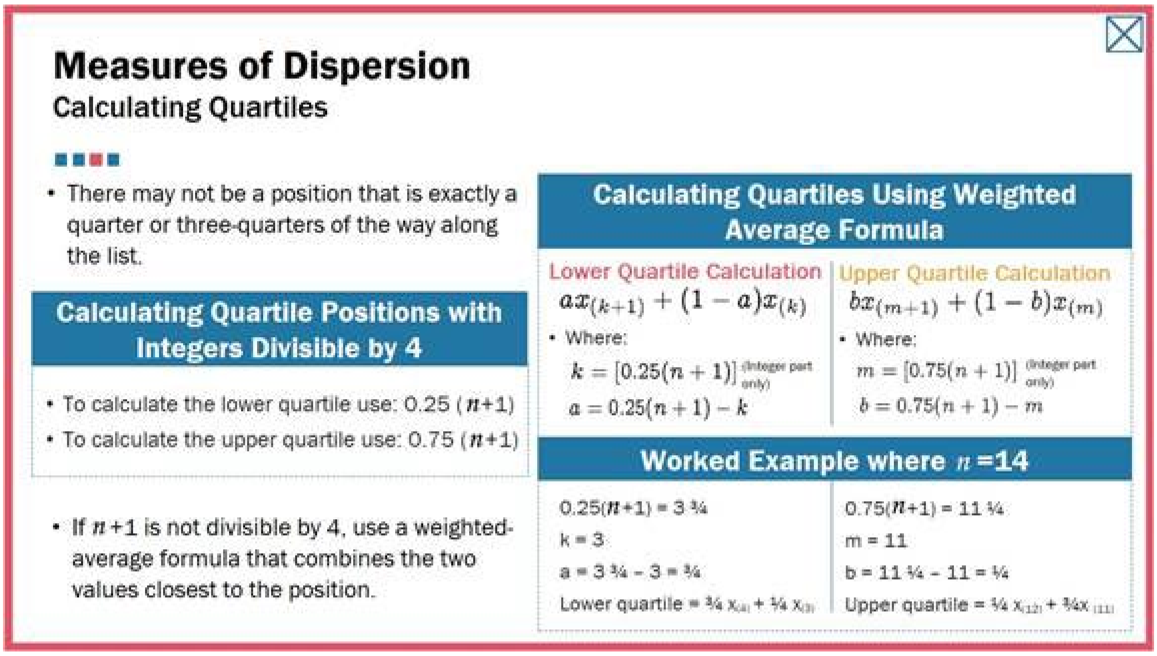
There are three quartiles:

The lower quartile is the observation that is ¼ of the way along the ordered list (going from smallest to largest)

The upper quartile is the observation that is ¾ of the way along the list.

And the median, that we've already met is the observation in the middle of the list of observations in order as we've already seen.

Tab 2.2: Calculating Quartiles



Measures of Dispersion

Calculating Quartiles

- There may not be a position that is exactly a quarter or three-quarters of the way along the list.

Calculating Quartile Positions with Integers Divisible by 4

- To calculate the lower quartile use: 0.25 (n+1)
- To calculate the upper quartile use: 0.75 (n+1)

- If n+1 is not divisible by 4, use a weighted-average formula that combines the two values closest to the position.

Calculating Quartiles Using Weighted Average Formula

Lower Quartile Calculation

$$ax_{(k+1)} + (1-a)x_{(k)}$$

- Where:

$k = [0.25(n+1)]$ (Integer part only)

$a = 0.25(n+1) - k$

Upper Quartile Calculation

$$bx_{(m+1)} + (1-b)x_{(m)}$$

- Where:

$m = [0.75(n+1)]$ (Integer part only)

$b = 0.75(n+1) - m$

Worked Example where n =14

0.25(n+1) = 3 ¾

k = 3

a = 3 ¾ - 3 = ¾

Lower quartile = ¾ $x_{(4)}$ + ¼ $x_{(3)}$

0.75(n+1) = 11 ¼

m = 11

b = 11 ¼ - 11 = ¼

Upper quartile = ¼ $x_{(12)}$ + ¾$x_{(11)}$

Like the median, there may not be a position that is exactly ¼ or ¾ of the way along the list. There is when n+1 is divisible by 4 e.g. a list of length 7 or 23. In this case, the position of the quartiles are ¼ (n+1) and ¾ (n+1) of the way along the list. When this is not the case, we use a formula that combines the 2 values closest to this position.

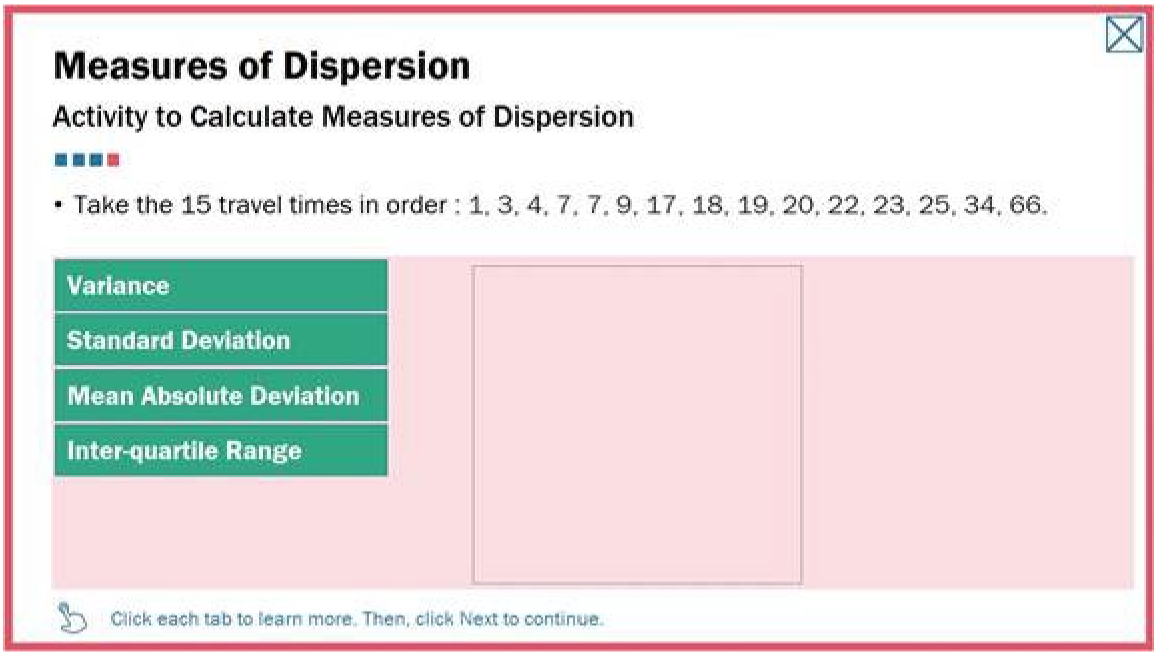
There are several different ways in which they can be combined, leading to potentially different values for the quartiles. Here we use a formula that is a little bit more complicated than that for the median but is one of the most common in use.

It involves using a weighted average of the two observations on either side in the ordered list.

This formula relies on calculating values that we denote by a and b. For example, if n is 14 then n+1 is 15 and is not divisible by 4, so we must use the formula.

To calculate the lower quartile, 0.25(n+1) is 15 divided by 4, which is 3 ¾. The integer part of this, denoted by k in the formula, is therefore 3. Then a is just the result of the division, 3 ¾, minus the integer part, 3, which gives a = ¾. The similar calculation to calculate the upper quartile gives m = 11 and b = ¼. So the lower quartile is ¾ times the (k+1)th smallest value (or 4th smallest) + ¼ times the kth smallest (or 3rd smallest). The upper quartile is ¼ times the twelfth smallest plus ¾ times the eleventh smallest value.

Tab 2.3: Activity to Calculate Measures of Dispersion



Measures of Dispersion

Activity to Calculate Measures of Dispersion

- Take the 15 travel times in order : 1, 3, 4, 7, 7, 9, 17, 18, 19, 20, 22, 23, 25, 34, 66.

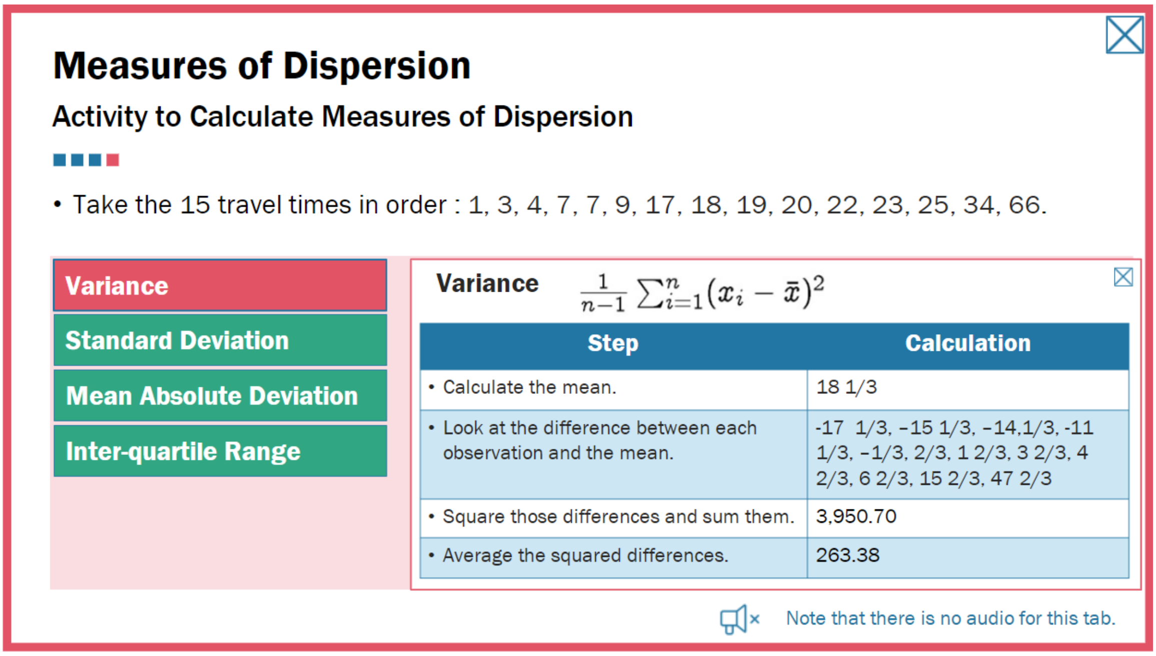
Click each tab to learn more. Then, click Next to continue.

Take some time to calculate the variance, standard deviation, mean absolute deviation and inter-quartile range for the 15 travel times we looked at in our earlier example. Ensure you work through the calculations to see that they match the definitions given earlier.

The 15 travel times are as follows: 1, 3, 4, 7, 7, 9, 17, 18, 19, 20, 22, 23, 25, 34, 66.

When you are ready, click the tabs to reveal the answers.

Tab 1: Variance Calculation:



Measures of Dispersion

Activity to Calculate Measures of Dispersion

- Take the 15 travel times in order : 1, 3, 4, 7, 7, 9, 17, 18, 19, 20, 22, 23, 25, 34, 66.

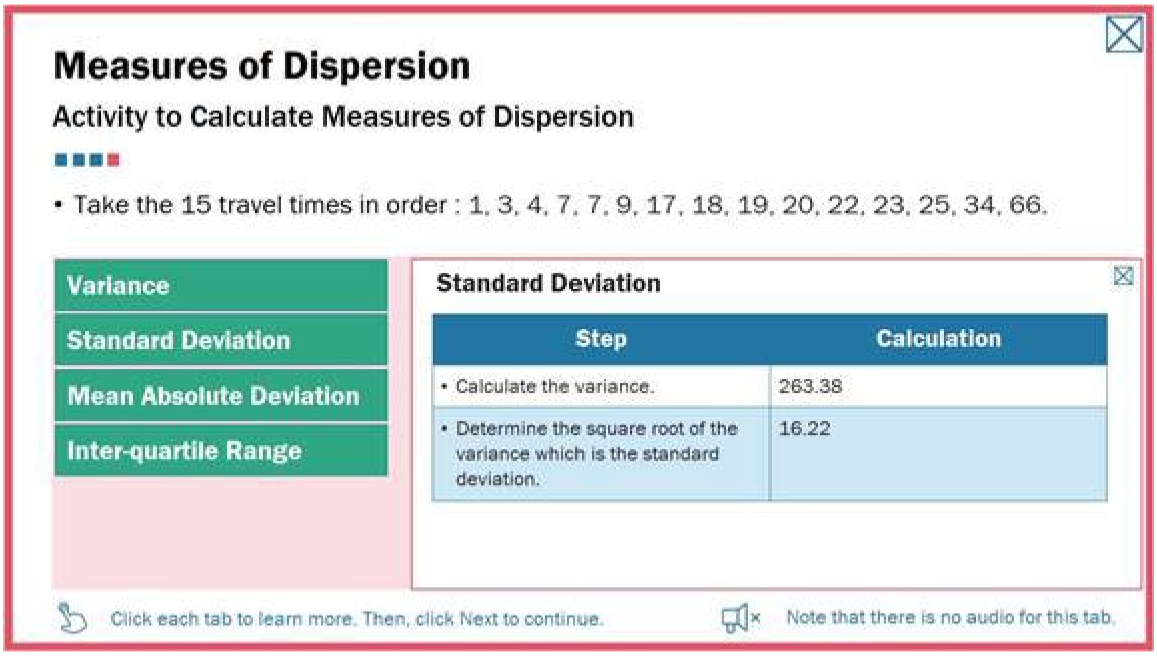
Variance $\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2$

Step	Calculation
• Calculate the mean.	18 1/3
• Look at the difference between each observation and the mean.	-17 1/3, –15 1/3, –14,1/3, -11 1/3, –1/3, 2/3, 1 2/3, 3 2/3, 4 2/3, 6 2/3, 15 2/3, 47 2/3
• Square those differences and sum them.	3,950.70
• Average the squared differences.	263.38

Note that there is no audio for this tab.

Step	Calculation
Calculate the mean.	$18\frac{1}{3}$
Look at the difference between each observation and the mean.	$-17\frac{1}{3}, -15\frac{1}{3}, -14\frac{1}{3}, -11\frac{1}{3}, -11\frac{1}{3}, -9\frac{1}{3}, -1\frac{1}{3}, -\frac{1}{3}, \frac{2}{3}, 1\frac{2}{3}, 3\frac{2}{3}, 4\frac{2}{3}, 6\frac{2}{3}, 15\frac{2}{3}, 47\frac{2}{3}$
Square those differences and sum them.	3,950.70
Average the squared differences.	263.38

Tab 2: Standard Deviation Calculation:



Measures of Dispersion

Activity to Calculate Measures of Dispersion

- Take the 15 travel times in order : 1, 3, 4, 7, 7, 9, 17, 18, 19, 20, 22, 23, 25, 34, 66.

Standard Deviation

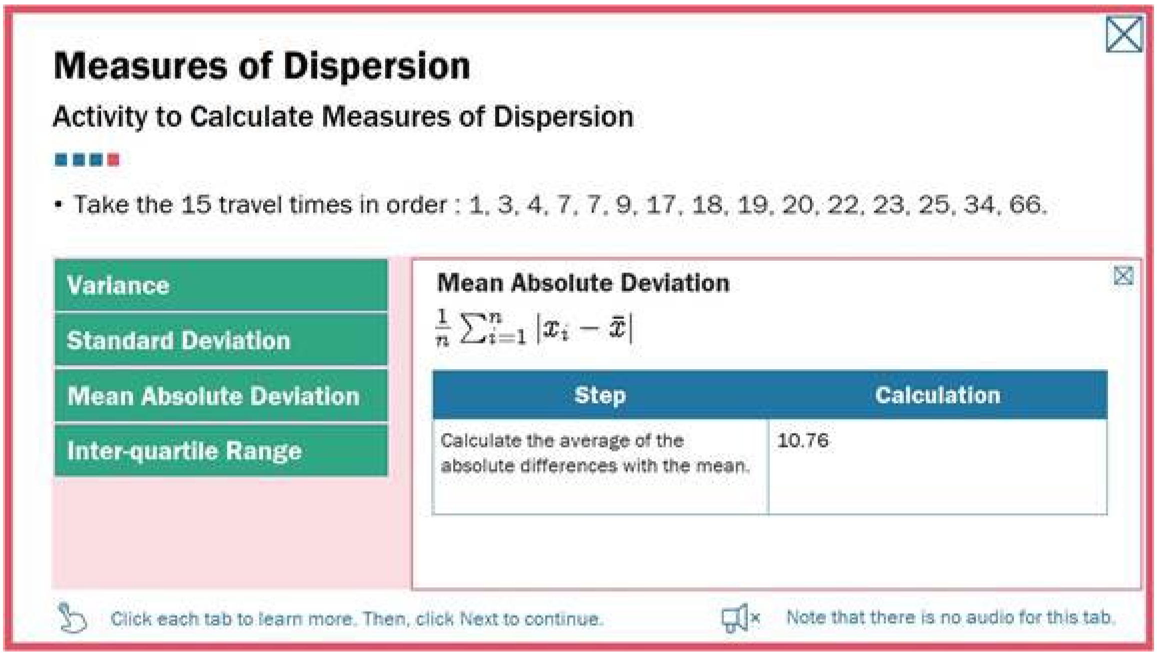
Step	Calculation
• Calculate the variance.	263.38
• Determine the square root of the variance which is the standard deviation.	16.22

Click each tab to learn more. Then, click Next to continue.

Note that there is no audio for this tab.

Step	Calculation
Calculate the variance.	263.37
Determine the square root of the variance, which is the standard deviation.	16.22

Tab 3: Mean Absolute Deviation



Measures of Dispersion

Activity to Calculate Measures of Dispersion

- Take the 15 travel times in order : 1, 3, 4, 7, 7, 9, 17, 18, 19, 20, 22, 23, 25, 34, 66.

Mean Absolute Deviation

$\frac{1}{n}\sum_{i=1}^{n}|x_i - \bar{x}|$

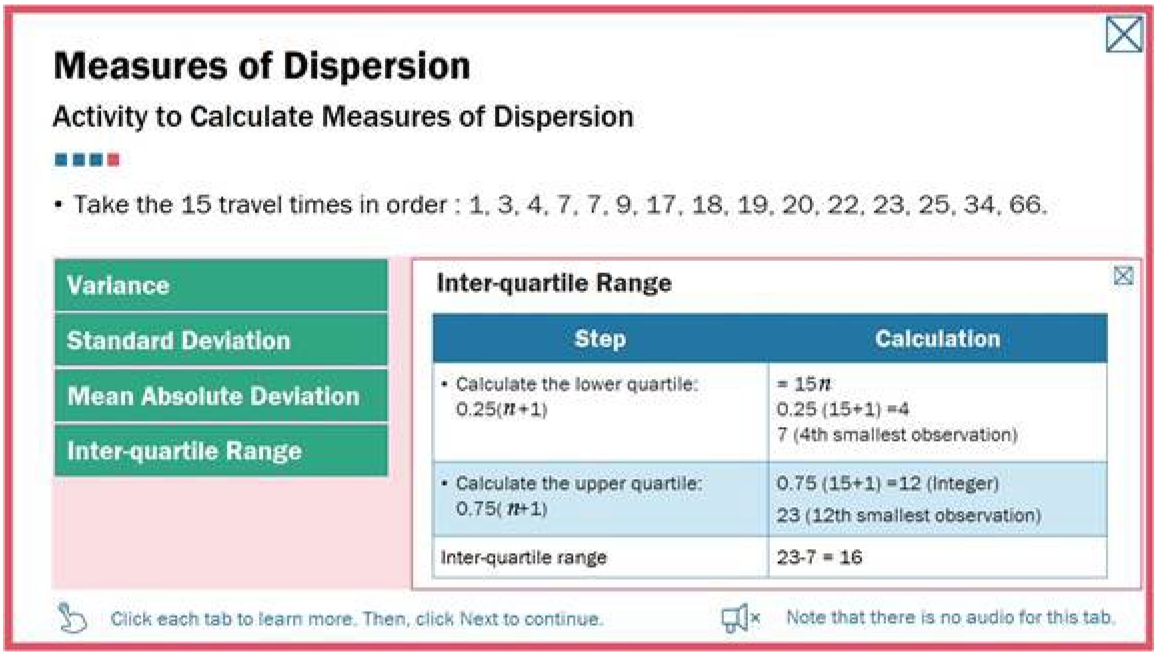
Step	Calculation
Calculate the average of the absolute differences with the mean.	10.76

Click each tab to learn more. Then, click Next to continue.

Note that there is no audio for this tab.

Step	Calculation
Calculate the average of the absolute differences with the mean.	10.76

Tab 4: Inter-quartile Range Calculation:



Measures of Dispersion

Activity to Calculate Measures of Dispersion

- Take the 15 travel times in order : 1, 3, 4, 7, 7, 9, 17, 18, 19, 20, 22, 23, 25, 34, 66.

Inter-quartile Range

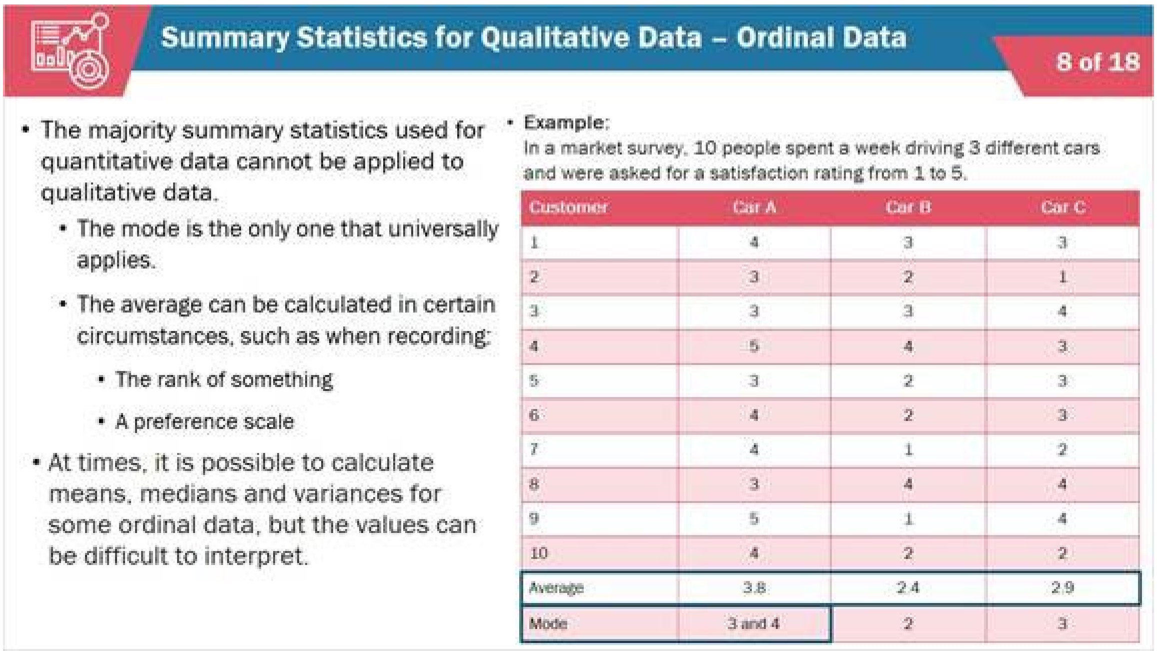
Step	Calculation
• Calculate the lower quartile: 0.25(n+1)	= 15n 0.25 (15+1) =4 7 (4th smallest observation)
• Calculate the upper quartile: 0.75(n+1)	0.75 (15+1) =12 (Integer) 23 (12th smallest observation)
Inter-quartile range	23-7 = 16

Click each tab to learn more. Then, click Next to continue.

Note that there is no audio for this tab.

Step	Calculation
Calculate the lower quartile: 0.25(n+1)	15=n 0.25(15+1)=4 Lower quartile = 7 (4th smallest observation)
Calculate the upper quartile: 0.75(n+1)	0.75(15+1)=12(Integer) Upper quartile = 23 (12th smallest observation)
Inter-quartile Range	23 - 7 = 16

Slide 8: Summary Statistics for Qualitative Data – Ordinal Data



Summary Statistics for Qualitative Data – Ordinal Data

8 of 18

- The majority summary statistics used for quantitative data cannot be applied to qualitative data.
  - The mode is the only one that universally applies.
  - The average can be calculated in certain circumstances, such as when recording:
    - The rank of something
    - A preference scale
- At times, it is possible to calculate means, medians and variances for some ordinal data, but the values can be difficult to interpret.

- **Example:**
  In a market survey, 10 people spent a week driving 3 different cars and were asked for a satisfaction rating from 1 to 5.

Customer	Car A	Car B	Car C
1	4	3	3
2	3	2	1
3	3	3	4
4	5	4	3
5	3	2	3
6	4	2	3
7	4	1	2
8	3	4	4
9	5	1	4
10	4	2	2
Average	3.8	2.4	2.9
Mode	3 and 4	2	3

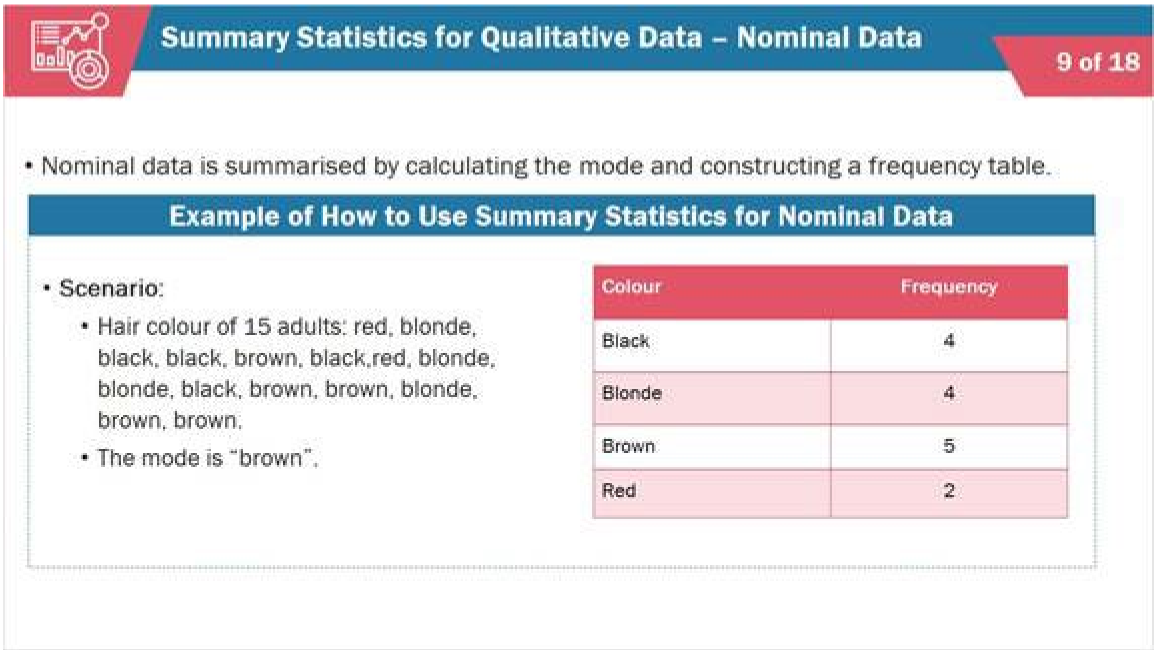
When it comes to qualitative data, unfortunately most of the summary statistics that we looked at do not make sense as they relied on the data being numbers. Indeed, the mode is the only one that we could universally apply. In some cases of ordinal data, where it is recording the rank of something (e.g. a list of preferences of candidates in an election) or something like a preference scale (e.g. a customer survey where satisfaction with a product or service is given on a scale from 1 to 5), it is possible to calculate the average. In this example you can see data from 10 customers who evaluated 3 cars and rated them on a scale from 1 to 5.

Customer	Car A	Car B	Car C
1	4	3	3
2	3	2	1
3	3	3	4
4	5	4	3
5	3	2	3
6	4	2	3
7	4	1	2
8	3	4	4
9	5	1	4
10	4	2	2
Average	3.8	2.4	2.9
Mode	3 and 4	2	3

Here the average can be calculated and is useful to compare the satisfaction rating of the 3 cars. Note that for Car A the mode is not unique as both the ratings of 3 and 4 have the largest number of observations for that car.

At times, it is also possible to calculate means, medians and even variances, but the values can be difficult to interpret.

Slide 9: Summary Statistics for Qualitative Data – Nominal Data



Summary Statistics for Qualitative Data – Nominal Data

9 of 18

- Nominal data is summarised by calculating the mode and constructing a frequency table.

Example of How to Use Summary Statistics for Nominal Data

- Scenario:
  - Hair colour of 15 adults: red, blonde, black, black, brown, black,red, blonde, blonde, black, brown, brown, blonde, brown, brown.
  - The mode is "brown".

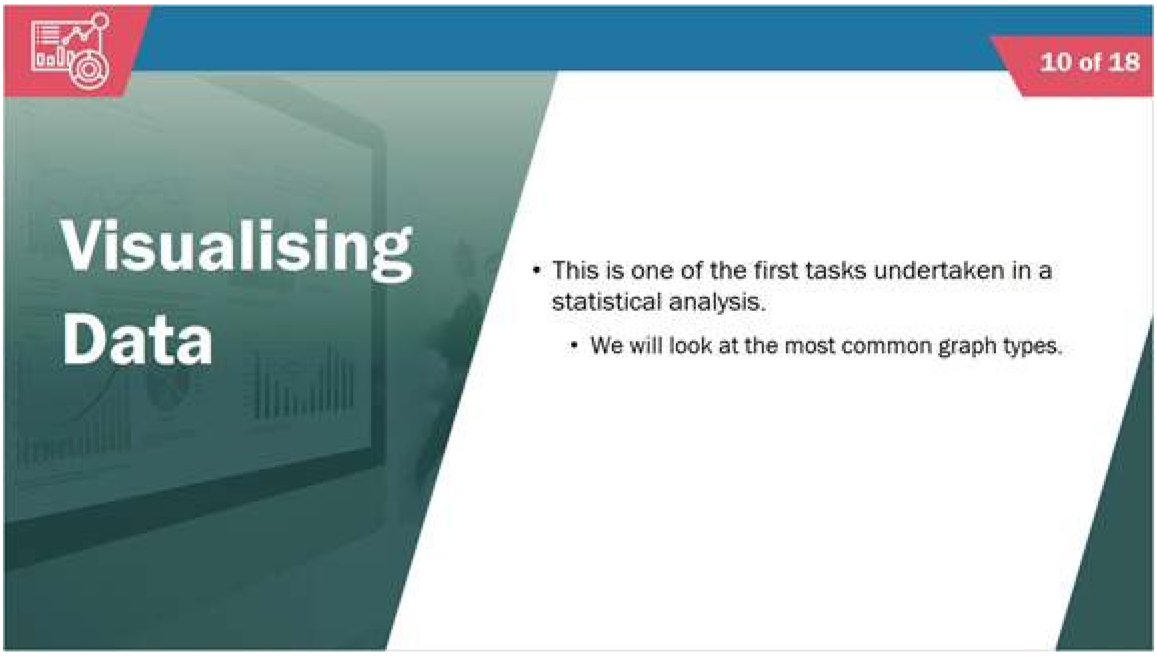
Colour	Frequency
Black	4
Blonde	4
Brown	5
Red	2

For nominal data, usually all that one can do by way of summarising the data is to calculate the mode and construct a frequency table that lists the number of observations of each different name in the data. Here is a simple example of a frequency table showing data gathered about the hair colour of fifteen people.

Colour	Frequency
Black	4
Blonde	4
Brown	5
Red	2

The mode is brown.

Slide 10: Section: Visualising Data



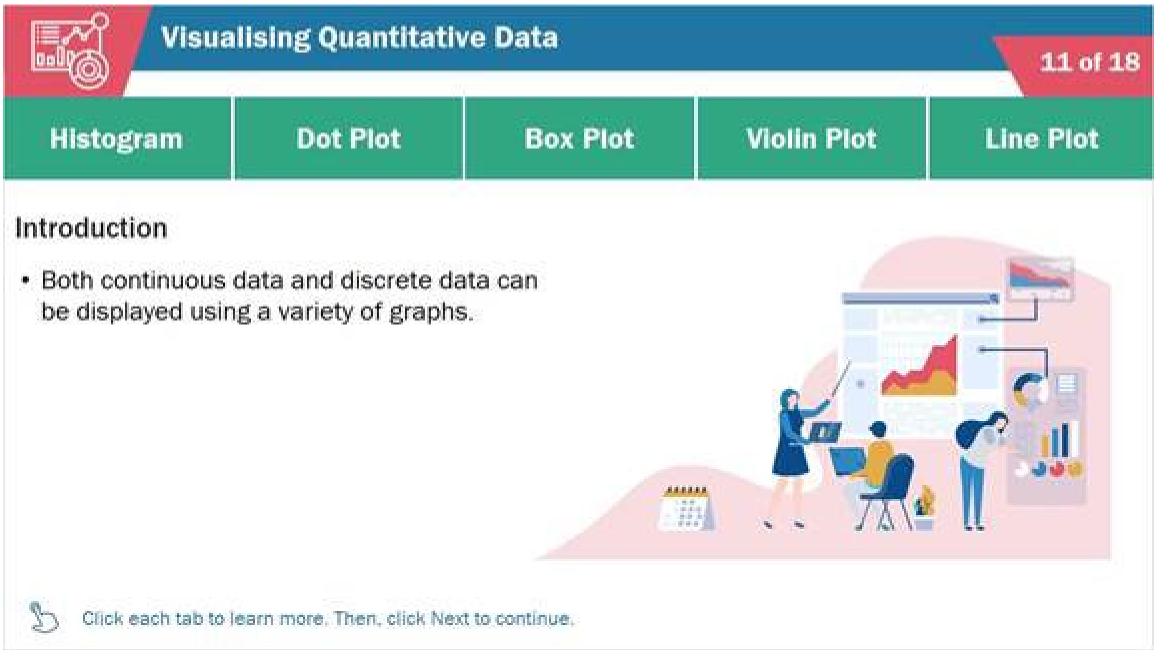
10 of 18

Visualising Data

- This is one of the first tasks undertaken in a statistical analysis.
  - We will look at the most common graph types.

Now we move to the final topic for this session, that of visualising data. Along with calculating some summary statistics, this is one of the first tasks that are undertaken in a statistical analysis. Here we will look at the most common graph types, although there are many others that have been proposed. We'll look at quantitative data first, and then look at what one can do with qualitative data.

Slide 11: Visualising Quantitative Data



Visualising Quantitative Data

11 of 18

Histogram | Dot Plot | Box Plot | Violin Plot | Line Plot

Introduction

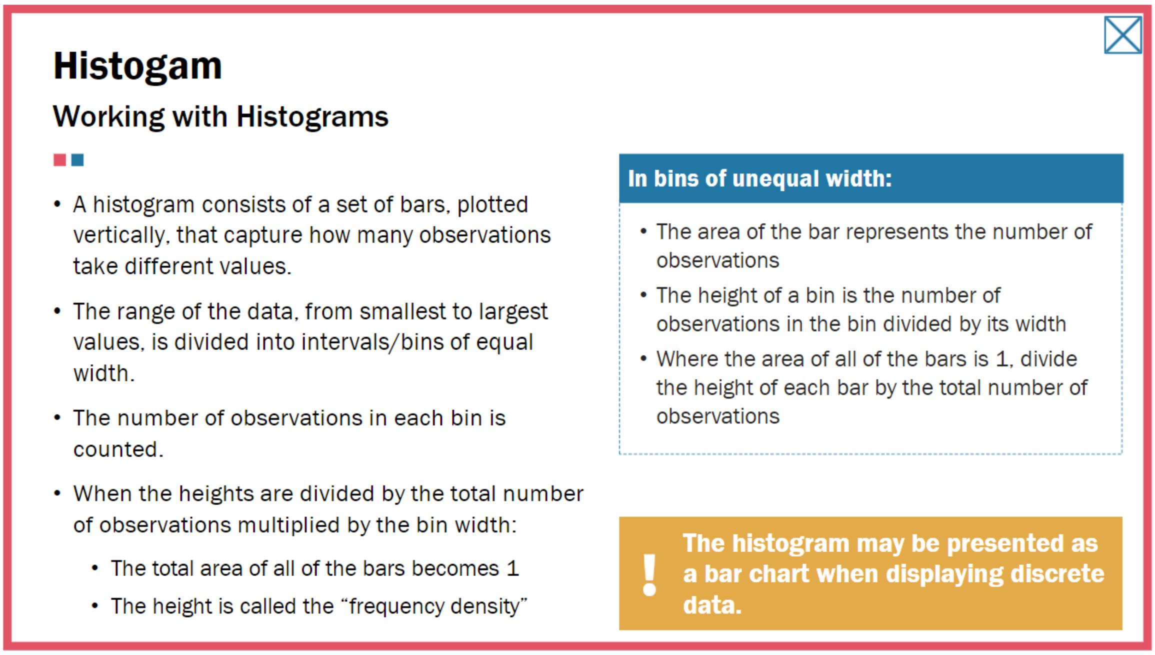
- Both continuous data and discrete data can be displayed using a variety of graphs.

Click each tab to learn more. Then, click Next to continue.

Both continuous data and discrete data can be displayed using variety of graphs. We are going to look at the following graphs in more detail: histogram, dot plot, box plot, violin plot and line plot.

Click the tabs to learn more. When you are ready, click Next to continue.

Tab 1: Histogram



Histogam

Working with Histograms

- A histogram consists of a set of bars, plotted vertically, that capture how many observations take different values.
- The range of the data, from smallest to largest values, is divided into intervals/bins of equal width.
- The number of observations in each bin is counted.
- When the heights are divided by the total number of observations multiplied by the bin width:
  - The total area of all of the bars becomes 1
  - The height is called the "frequency density"

In bins of unequal width:

- The area of the bar represents the number of observations
- The height of a bin is the number of observations in the bin divided by its width
- Where the area of all of the bars is 1, divide the height of each bar by the total number of observations

! The histogram may be presented as a bar chart when displaying discrete data.

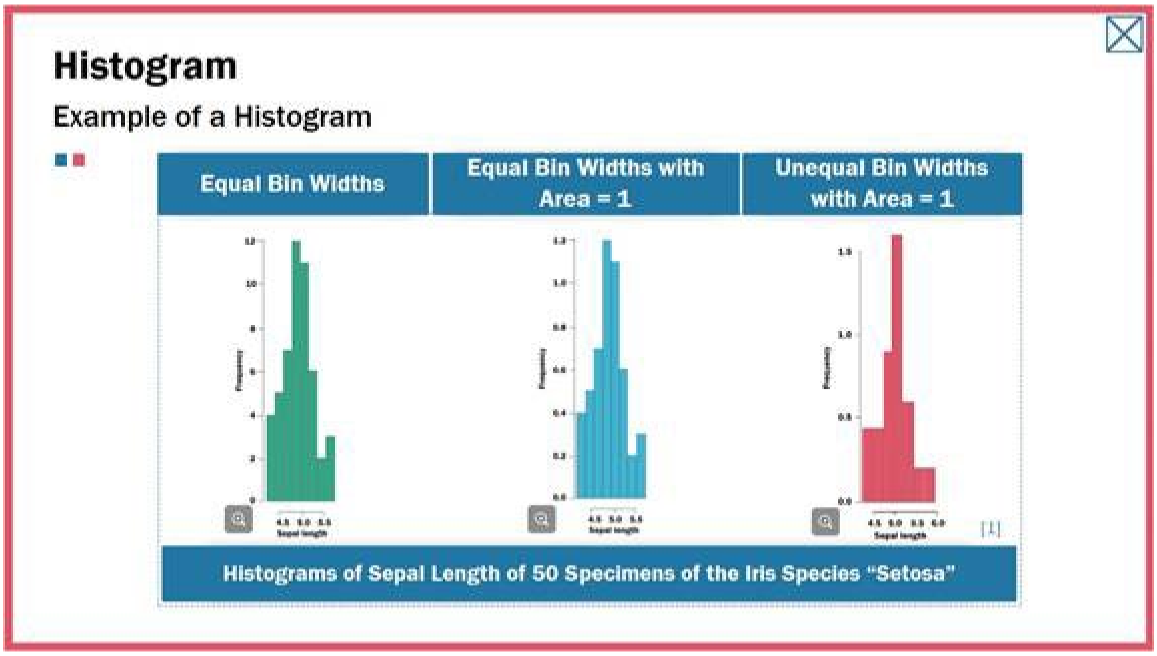
We saw the histogram earlier in the presentation. It consists of a set of bars, usually plotted vertically, that capture how many observations take different values. In its simplest form, the range of the data (from smallest to largest values) is divided into intervals, or bins, of equal width and the number of observations in each bin is counted. These frequencies are the heights of each bar.

Sometimes these heights are divided by the total number of observations multiplied by the bin width. In this case, the total area of all of the bars is one and the height is called the 'frequency density'.

It is also possible to divide up the data range into bins of unequal width. When this is done, it is not the height of the bar that represents the number of observations but its area, so the height of a bin is the number of observations in the bin divided by its width. Further, we can also create a histogram where the area of all of the bars is one by additionally dividing the height of each bar by the total number of observations. In fact, this is the way that an unequal bin width histogram is usually displayed.

Note that the histogram may be in the special form of a bar chart when displaying discrete data. In a bar chart, each separate discrete value in the range of the data has its own bar which is the number of observations with that value.

Tab 1.1: Examples of Histograms



Histogram

Example of a Histogram

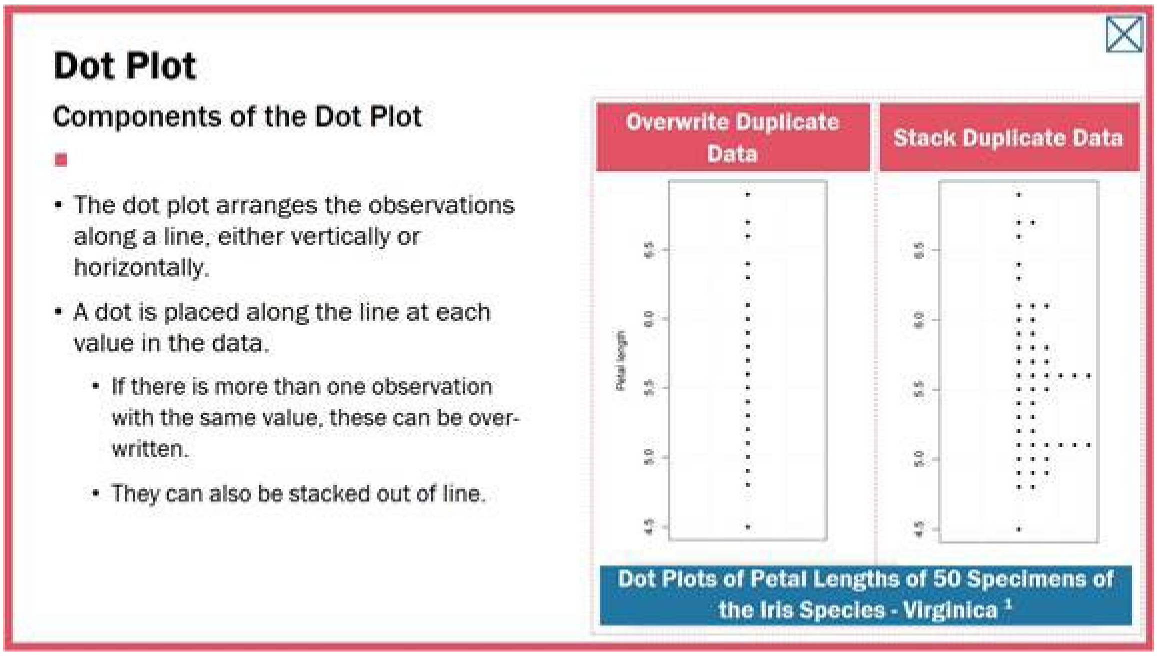
Equal Bin Widths | Equal Bin Widths with Area = 1 | Unequal Bin Widths with Area = 1

[1]

Histograms of Sepal Length of 50 Specimens of the Iris Species "Setosa"

On this slide you see an example of three types of histogram, applied to the same data set. In the first histogram, the bins are of equal width, the second displays bins of equal width with an area equal to one and the third displays bins of unequal width with an area of one.

Tab 2: Dot Plot



Dot Plot

Components of the Dot Plot

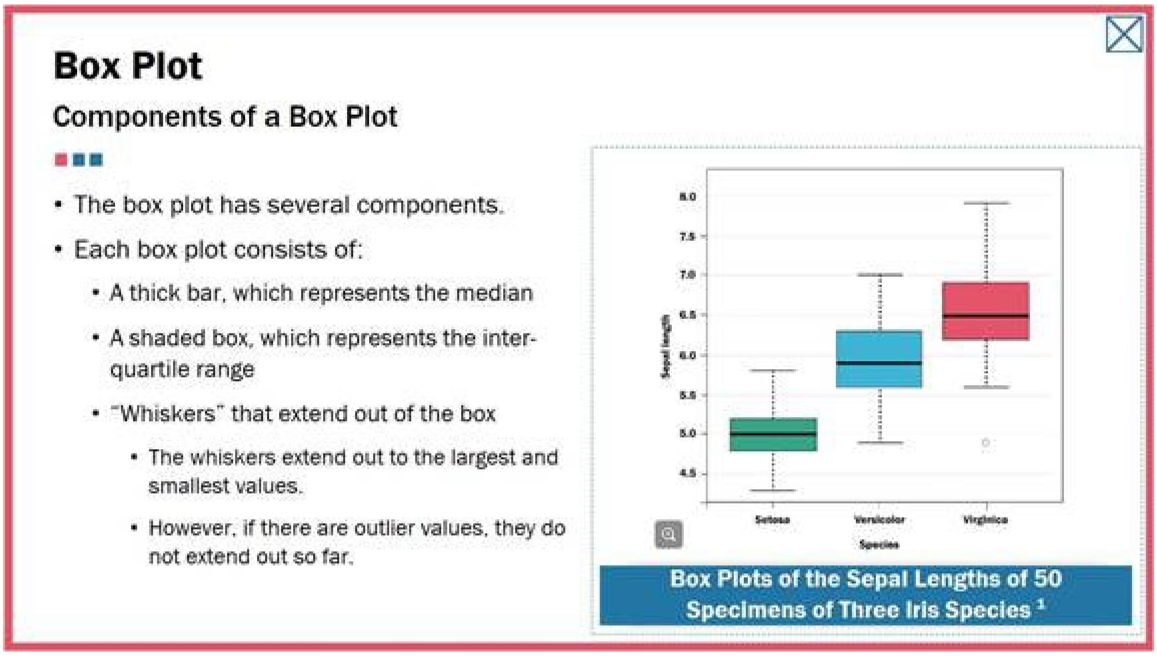
- The dot plot arranges the observations along a line, either vertically or horizontally.
- A dot is placed along the line at each value in the data.
  - If there is more than one observation with the same value, these can be overwritten.
  - They can also be stacked out of line.

Overwrite Duplicate Data | Stack Duplicate Data

Dot Plots of Petal Lengths of 50 Specimens of the Iris Species - Virginica [1]

The dot plot arranges the observations along a line, either vertically or horizontally. A dot is placed along the line at each value in the data. In the case of more than one observation with the same value then these are overwritten and the information about how many observations take a particular value is lost. However, these can be stacked out of the line, as shown here in some more observations from the iris data set, to recover this information.

Tab 3: Box Plot



Box Plot

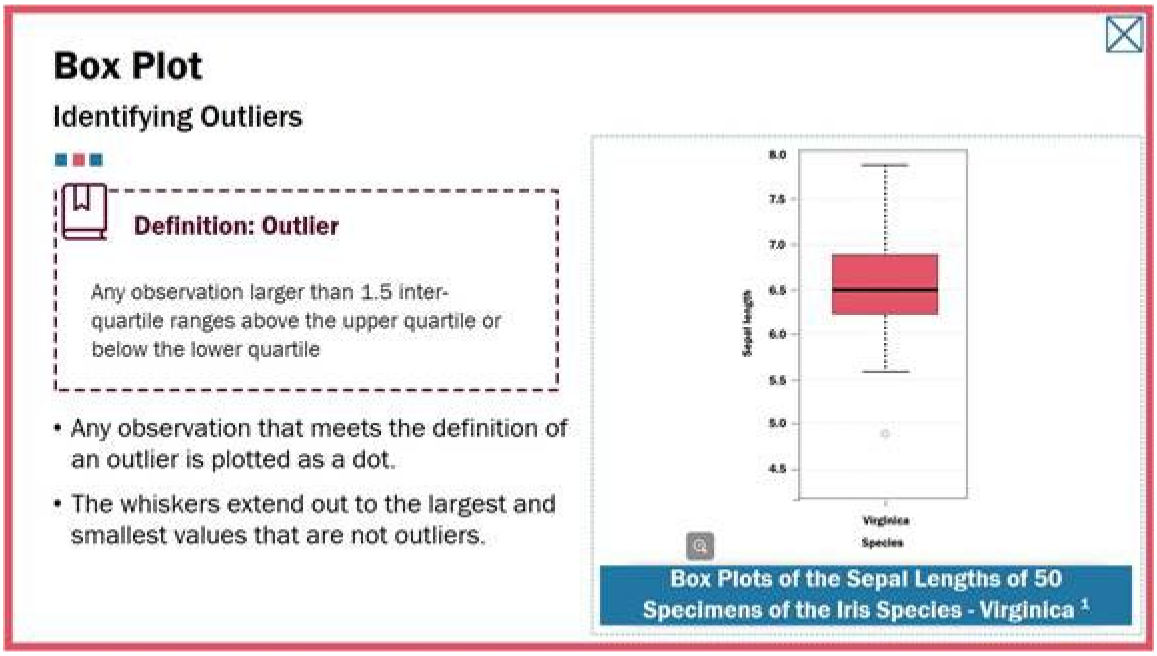
Components of a Box Plot

- The box plot has several components.
- Each box plot consists of:
  - A thick bar, which represents the median
  - A shaded box, which represents the inter-quartile range
  - "Whiskers" that extend out of the box
    - The whiskers extend out to the largest and smallest values.
    - However, if there are outlier values, they do not extend out so far.

Box Plots of the Sepal Lengths of 50 Specimens of Three Iris Species [1]

The box plot has several components to it. Look at the figure on this slide where there are three box plots. Each represent the length of the sepal of 50 specimens of a different iris species. You'll see that each box plot consists of a thick bar (this is the median), a shaded box (that extends from the lower to upper quartiles of the data, so represents the inter-quartile range), and then lines that extend out of this box. These lines are called whiskers and do one of two things. The simplest case is that they extend out to the largest and smallest values in the data. However, if there are outlier values, they do not extend out so far.

Tab 3.1: Identifying Outliers



Box Plot

Identifying Outliers

Definition: Outlier

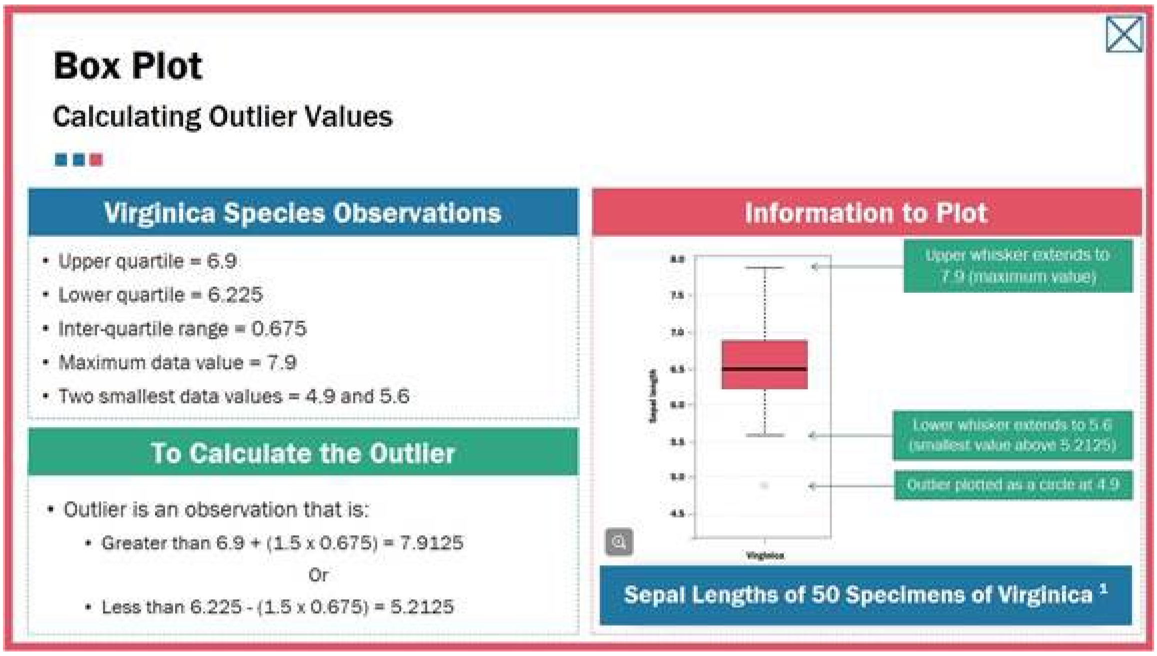
Any observation larger than 1.5 inter-quartile ranges above the upper quartile or below the lower quartile

- Any observation that meets the definition of an outlier is plotted as a dot.
- The whiskers extend out to the largest and smallest values that are not outliers.

Box Plots of the Sepal Lengths of 50 Specimens of the Iris Species - Virginica [1]

There are several definitions of what an outlier value in data is but for box plots, the most common makes use of the inter-quartile range. The outlier is defined as any observation larger than 1.5 inter-quartile ranges above the upper quartile or below the lower quartile. We can see an outlier observation in the Virginica box plot in our example. Any observation that meets the definition of an outlier is plotted as a dot, and the whiskers extend out to the largest and smallest values that are not outliers.

Tab 3.2: Calculating Outlier Values



Box Plot

Calculating Outlier Values

Virginica Species Observations

- Upper quartile = 6.9
- Lower quartile = 6.225
- Inter-quartile range = 0.675
- Maximum data value = 7.9
- Two smallest data values = 4.9 and 5.6

To Calculate the Outlier

- Outlier is an observation that is:
  - Greater than 6.9 + (1.5 x 0.675) = 7.9125

    Or
  - Less than 6.225 - (1.5 x 0.675) = 5.2125

Information to Plot

Upper whisker extends to 7.9 (maximum value)

Lower whisker extends to 5.6 (smallest value above 5.2125)

Outlier plotted as a circle at 4.9

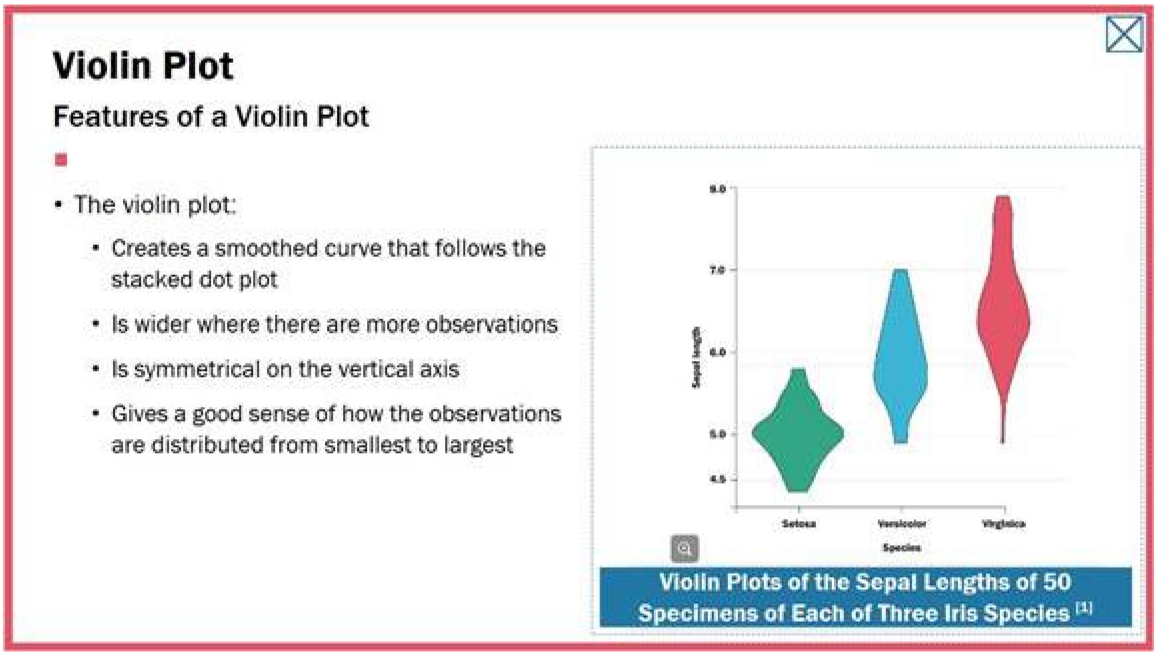
Sepal Lengths of 50 Specimens of Virginica [1]

Here are the calculations that show how an outlier is discovered in the data set for the Virginica species of iris and what that means for the box plot. So, the upper quartile is 6.9 and the lower quartile is 6.225 from which we can calculate the interquartile range as 0.675. The maximum data value is 7.9 and the two smallest values in the data are 4.9 and 5.6.

To calculate the outlier values, we need to work out the observations greater than 1.5 inter-quartile ranges above the upper quartile (which in this case is 7.9125) and below the lower quartile (which is 5.2125).

As you can see, there is no observation greater than 7.9125 so the upper whisker extends out to the maximum value of 7.9. If we look at the two smallest values in the data 4.9 and 5.6, one value is less than 5.2125 meaning the lower whisker extends to smallest observation above 5.2125 (which is 5.6) and the outlier is plotted as a circle (at 4.9).

Tab 4: Violin Plot



Violin Plot

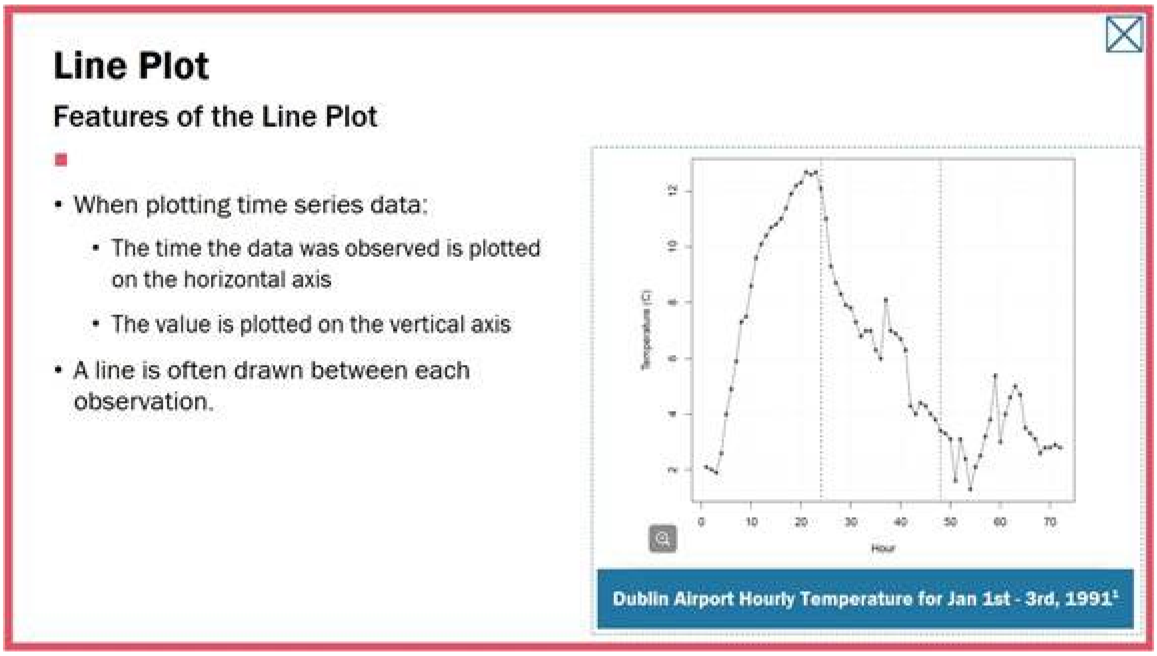
Features of a Violin Plot

- The violin plot:
  - Creates a smoothed curve that follows the stacked dot plot
  - Is wider where there are more observations
  - Is symmetrical on the vertical axis
  - Gives a good sense of how the observations are distributed from smallest to largest

Violin Plots of the Sepal Lengths of 50 Specimens of Each of Three Iris Species [1]

The violin plot has some resemblance to the stacked dot plot. While we do not discuss the details of how it is constructed, the idea is that the violin plot creates a smoothed curve that follows the stacked dot plot, and so is wider where there are more observations. It is also made to be symmetric on the vertical axis. It gives a good sense of how the observations are distributed from smallest to largest.

Tab 5: Line Plot



Line Plot

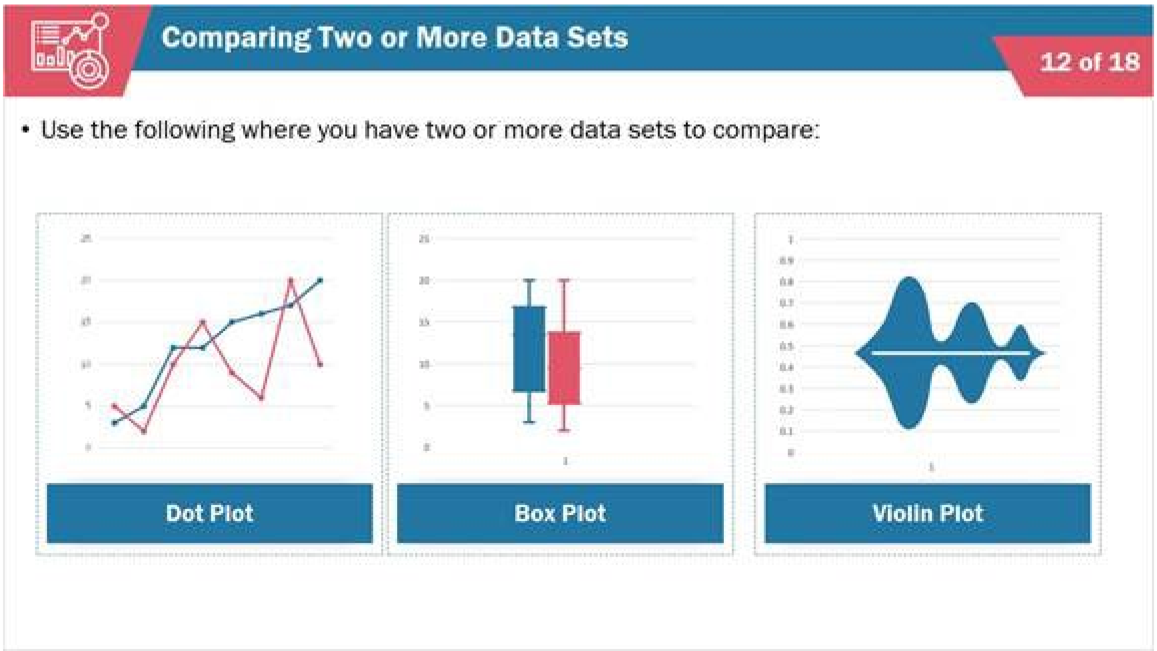
Features of the Line Plot

- When plotting time series data:
  - The time the data was observed is plotted on the horizontal axis
  - The value is plotted on the vertical axis
- A line is often drawn between each observation.

Dublin Airport Hourly Temperature for Jan 1st - 3rd, 1991[1]

Time series data are usually plotted with the time that the data was observed on the horizontal axis and the value on the vertical axis. Often a line is drawn between each observation.

Slide 12: Comparing Two or More Data Sets



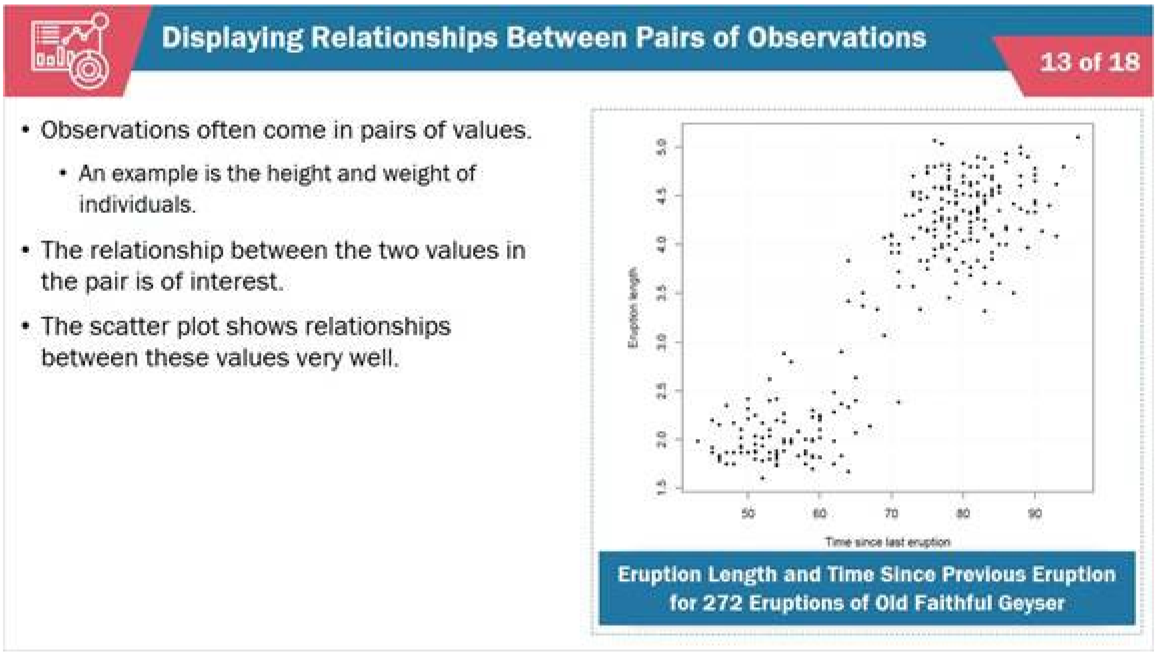
Comparing Two or More Data Sets

12 of 18

- Use the following where you have two or more data sets to compare:

Where you have two or more data sets and want to compare them, the dot, box and violin plots are all useful, as we have seen in some of the examples that we've looked at.

Slide 13: Displaying Relationships Between Pairs of Observations



Displaying Relationships Between Pairs of Observations

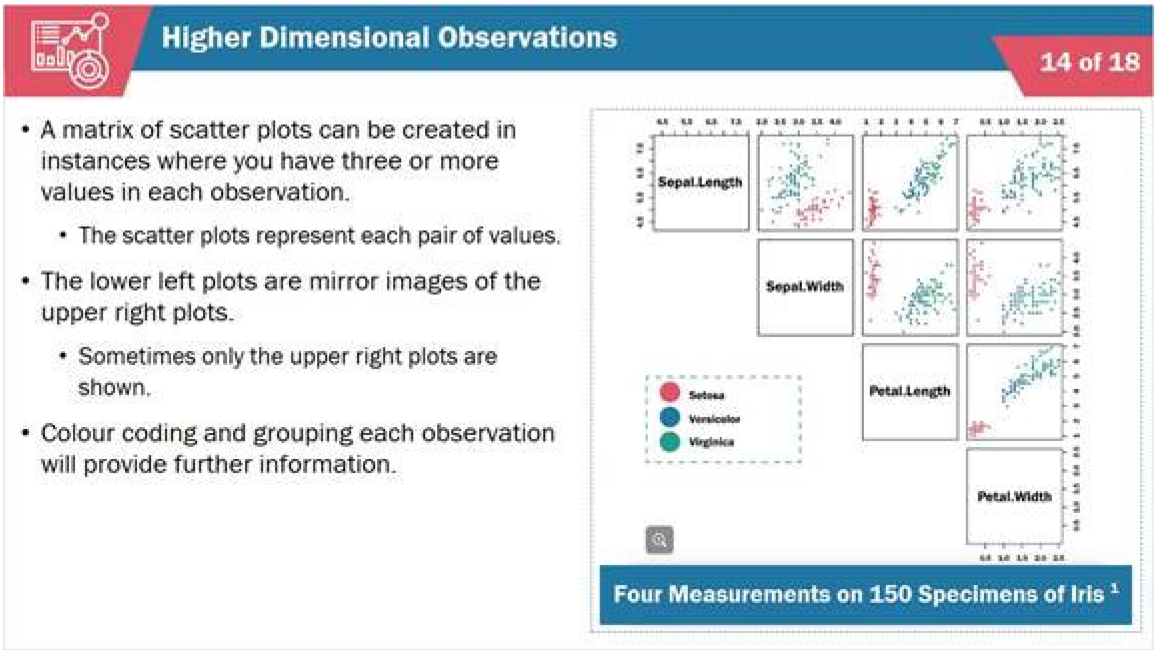
13 of 18

- Observations often come in pairs of values.
  - An example is the height and weight of individuals.
- The relationship between the two values in the pair is of interest.
- The scatter plot shows relationships between these values very well.

Eruption Length and Time Since Previous Eruption for 272 Eruptions of Old Faithful Geyser

Often, we have observations that come in pairs of values, for example the height and weight of individuals. Often, it's the relationship between the two values in the pair that is of interest. The scatter plot shows these relationships very well. On this slide we see an example to do with the length of eruptions of the Old Faithful geyser in Yellowstone National Park, plotted with the time since the previous eruption. This scatter plot illustrates two important properties of the data: first that longer eruption times are associated with a longer wait since the last eruption, and that there appear to be two groups of observations, one with smaller eruption lengths and waiting times, and the other group with larger values of both.

Slide 14: Higher Dimensional Observations



Higher Dimensional Observations

14 of 18

- A matrix of scatter plots can be created in instances where you have three or more values in each observation.
  - The scatter plots represent each pair of values.
- The lower left plots are mirror images of the upper right plots.
  - Sometimes only the upper right plots are shown.
- Colour coding and grouping each observation will provide further information.

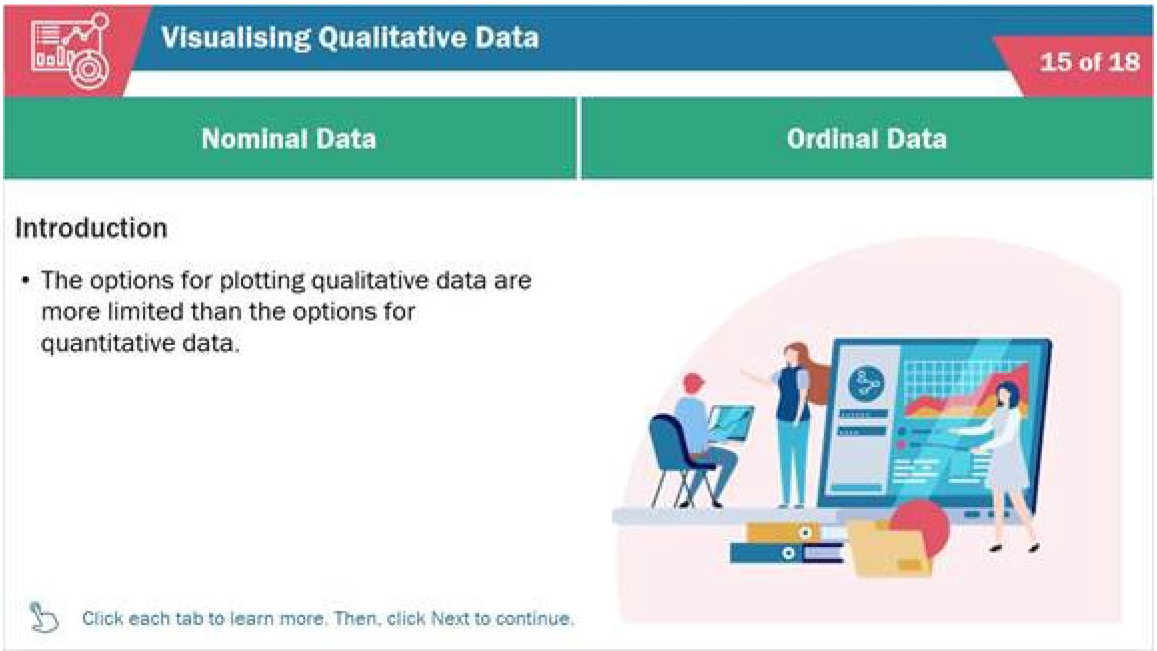
Four Measurements on 150 Specimens of Iris [1]

Where you have three or more values in each observation, a matrix of scatter plots can be created that are the scatter plots for each pair of values. Here we have four measurements on 150 specimens of iris and we see each possible scatter plot (there are six distinct pairs of four measurements if we ignore the order in which they appear).

The upper right scatter plots and lower left scatter plots are just mirror images of each other (with the variables taking different axes) and so sometimes just the upper right ones are shown, as here.

Recall that these 150 iris specimens were of three different species. Here we have colour coded each observation by species, which brings out further information in the data about their relative sizes and the relationships between these measurements.

Slide 15: Visualising Qualitative Data



Visualising Qualitative Data

15 of 18

Nominal Data | Ordinal Data

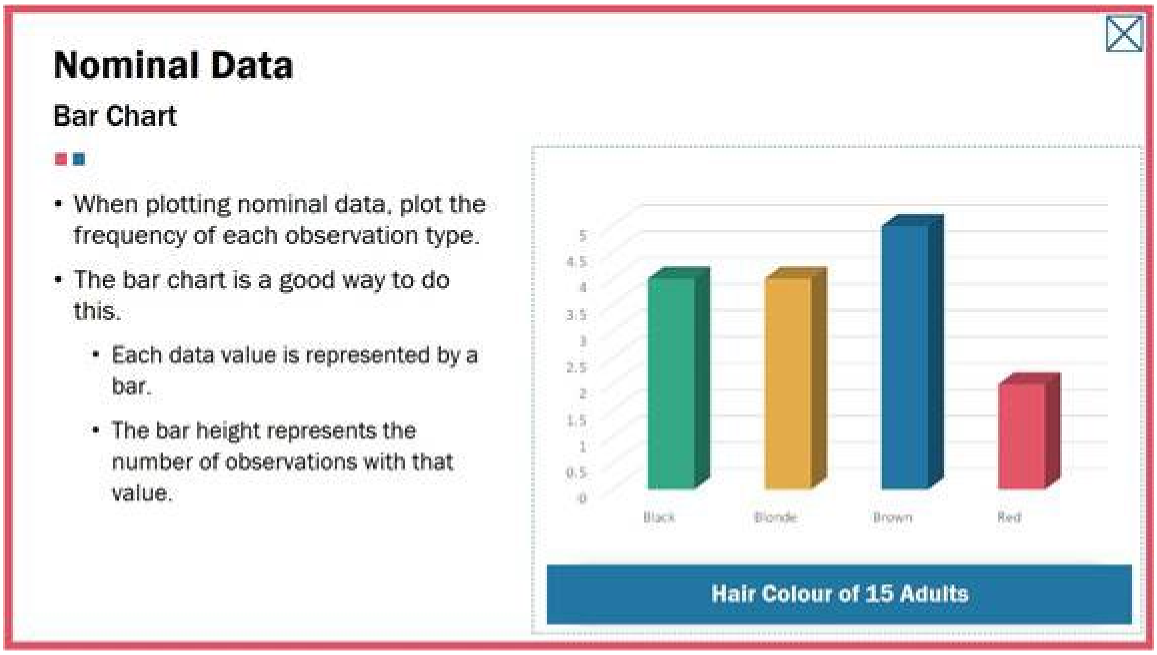
Introduction

- The options for plotting qualitative data are more limited than the options for quantitative data.

Click each tab to learn more. Then, click Next to continue.

Now we turn to qualitative data. The options for plotting are more limited than the quantitative case. Click the tabs to learn about visualising nominal and ordinal data. When you are ready, click next to continue.

Tab 1: Nominal Data



Nominal Data

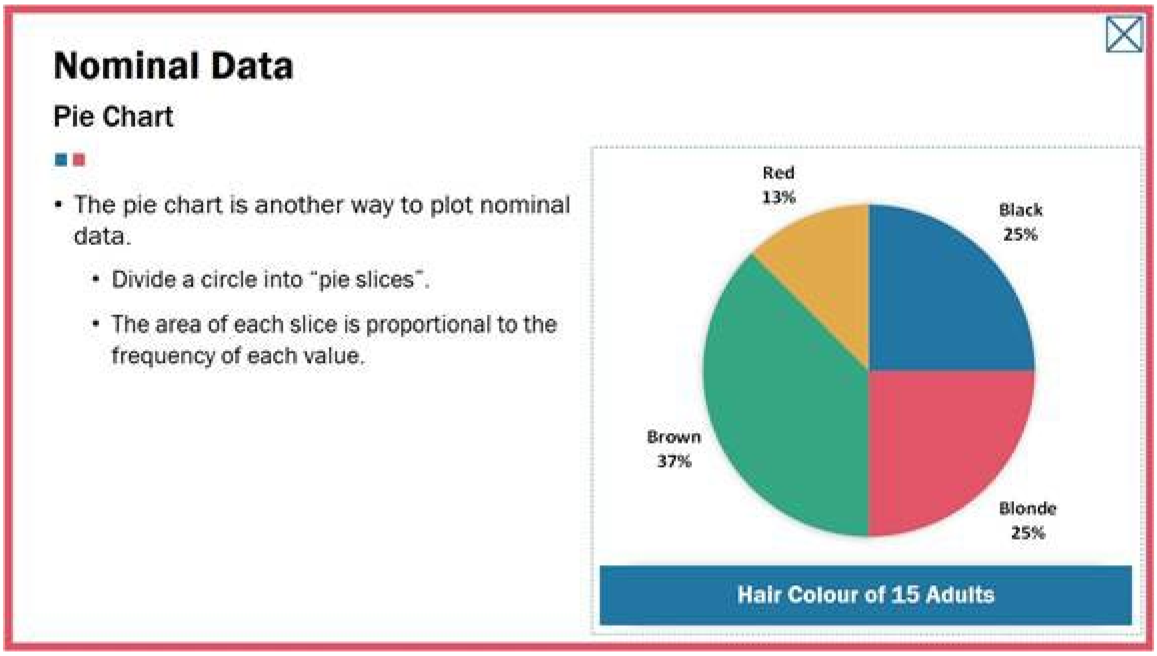
Bar Chart

- When plotting nominal data, plot the frequency of each observation type.
- The bar chart is a good way to do this.
  - Each data value is represented by a bar.
  - The bar height represents the number of observations with that value.

Hair Colour of 15 Adults

For nominal data, all that one can do usually is to plot the frequency of each observation type. The bar chart is a good way to do this, with one bar for each value in the data and the bar height being the number of observations with that value.

Tab 1.1: The Pie Chart



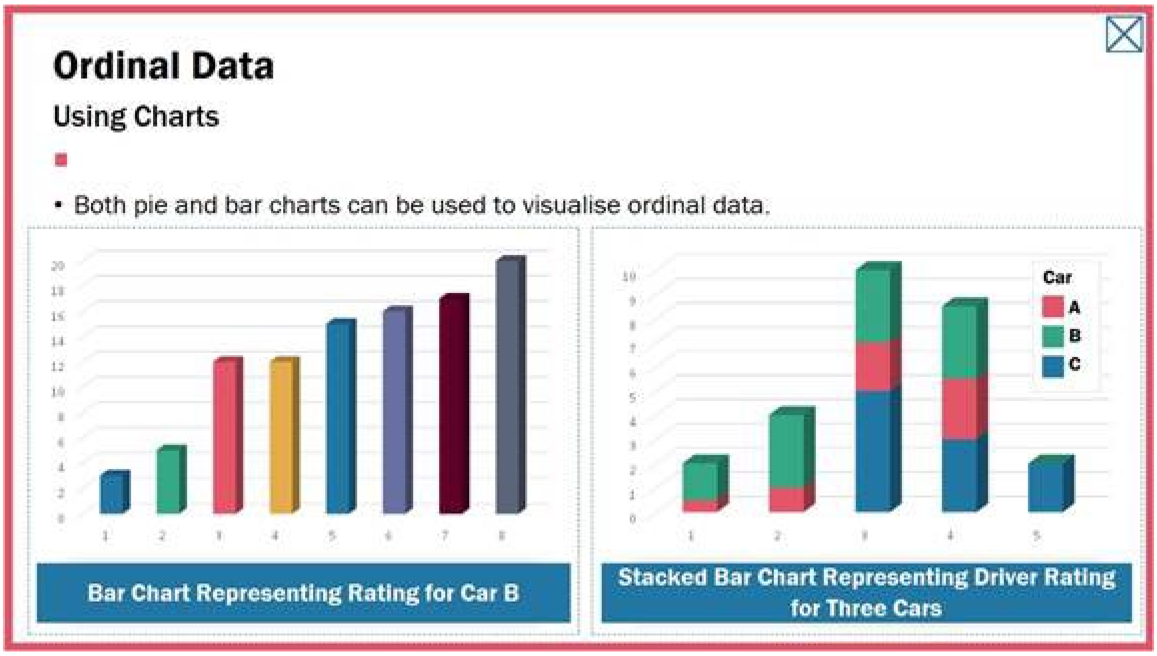
Nominal Data

Pie Chart

- The pie chart is another way to plot nominal data.
  - Divide a circle into "pie slices".
  - The area of each slice is proportional to the frequency of each value.

The pie chart is another way to illustrate nominal data. A circle is divided up into slices whose area is proportional to the frequency of each value. You can add additional information such as the proportion of observations for each name - seen in this example as a percentage.

Tab 2: Ordinal Data



Ordinal Data

Using Charts

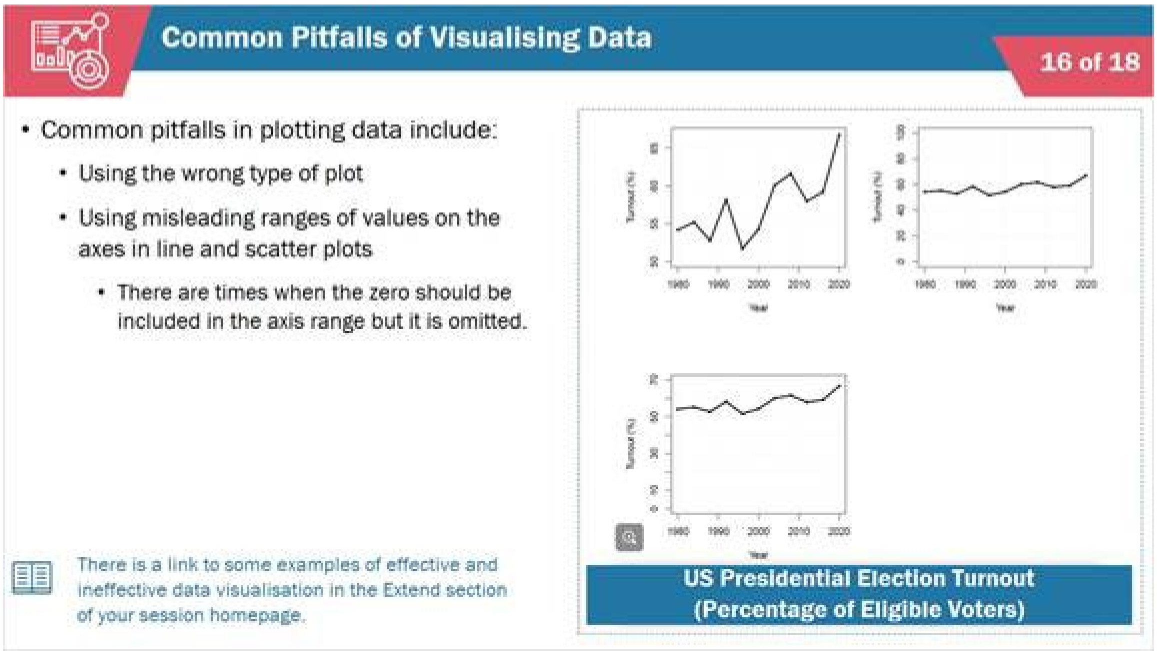
- Both pie and bar charts can be used to visualise ordinal data.

20
18
16
14
12
10
8
6
4
2
0
1
2
3
4
5
6
7
8
Bar Chart Representing Rating for Car B
10
9
8
7
6
5
4
3
2
1
0
1
2
3
4
5
Car
A
B
C
Stacked Bar Chart Representing Driver Rating for Three Cars

Ordinal data can also be visualised with a bar or pie chart. Here the data for the satisfaction rating of Car B in our ordinal data example is visualised as a bar chart.

We can take all three cars in that example and create a stacked bar chart, where the frequency of each satisfaction rating are stacked on top of each other, usually in different colours. This shows well how the different cars were rated by all of the customers, in particular that Car A got more higher ratings than the other cars.

Slide 16: Common Pitfalls of Visualising Data



Common Pitfalls of Visualising Data

16 of 18

- Common pitfalls in plotting data include:
  - Using the wrong type of plot
  - Using misleading ranges of values on the axes in line and scatter plots
    - There are times when the zero should be included in the axis range but it is omitted.

There is a link to some examples of effective and ineffective data visualisation in the Extend section of your session homepage.

US Presidential Election Turnout (Percentage of Eligible Voters)

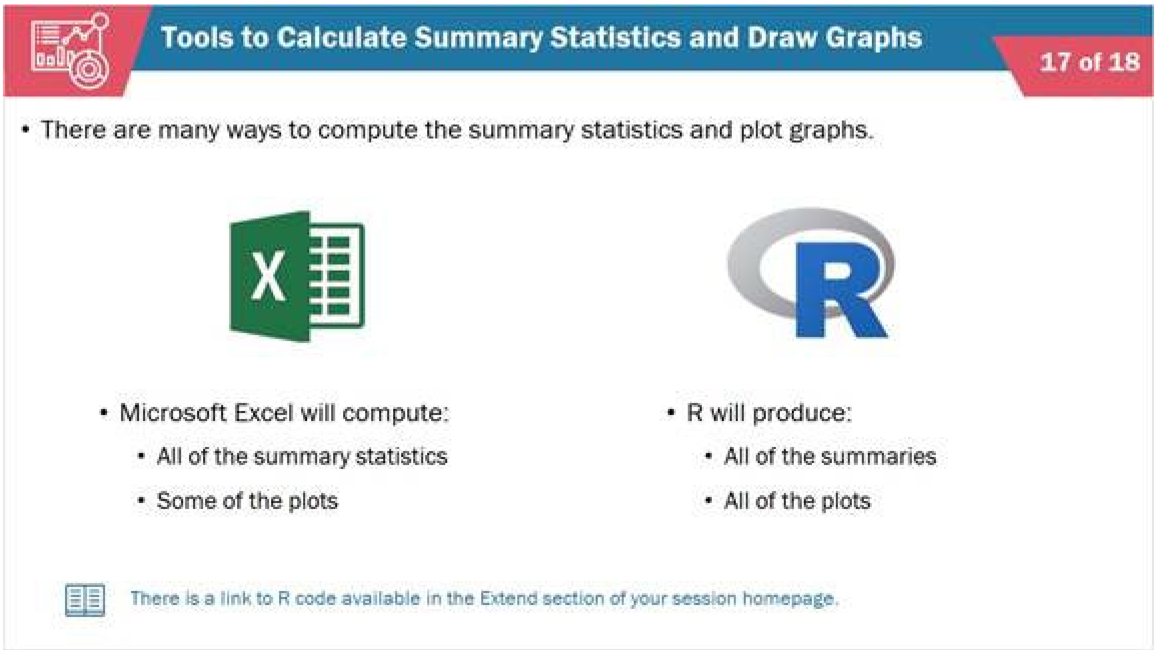
There are many pitfalls in plotting data. These include:

- Using the wrong type of plot – for example a pie chart for quantitative data
- Using misleading ranges of values on the axes in line and scatter plots. Most commonly this is when an axis range should include zero but does not.

Here’s an example of using a misleading range of values in a line plot of some time series data. With these data concerning turnout in US presidential elections, this plot can be misleading as it looks like turnout has increased by a factor of 3 or 4 from 1980. The vertical axis range starts at 50% turnout. More reasonable is as shown in this second graph, to plot this axis with a lower value of 0%, the lowest possible in this case. However, that leaves the plot looking rather flat and does not bring out the trend in the data, which is broadly that turnout has increased. This third plot is some sort of compromise that tries to bring this trend out more by reducing the range somewhat.

There are links to several resources that show some of the many examples of bad data visualisation, as well as some examples of really effective visualisations of data sets that are far larger and more complex than those we look at here available in the Extend section of the session homepage.

Slide 17: Tools to Calculate Summary Statistics and Draw Graphs



Tools to Calculate Summary Statistics and Draw Graphs

17 of 18

- There are many ways to compute the summary statistics and plot graphs.

- Microsoft Excel will compute:
  - All of the summary statistics
  - Some of the plots

- R will produce:
  - All of the summaries
  - All of the plots

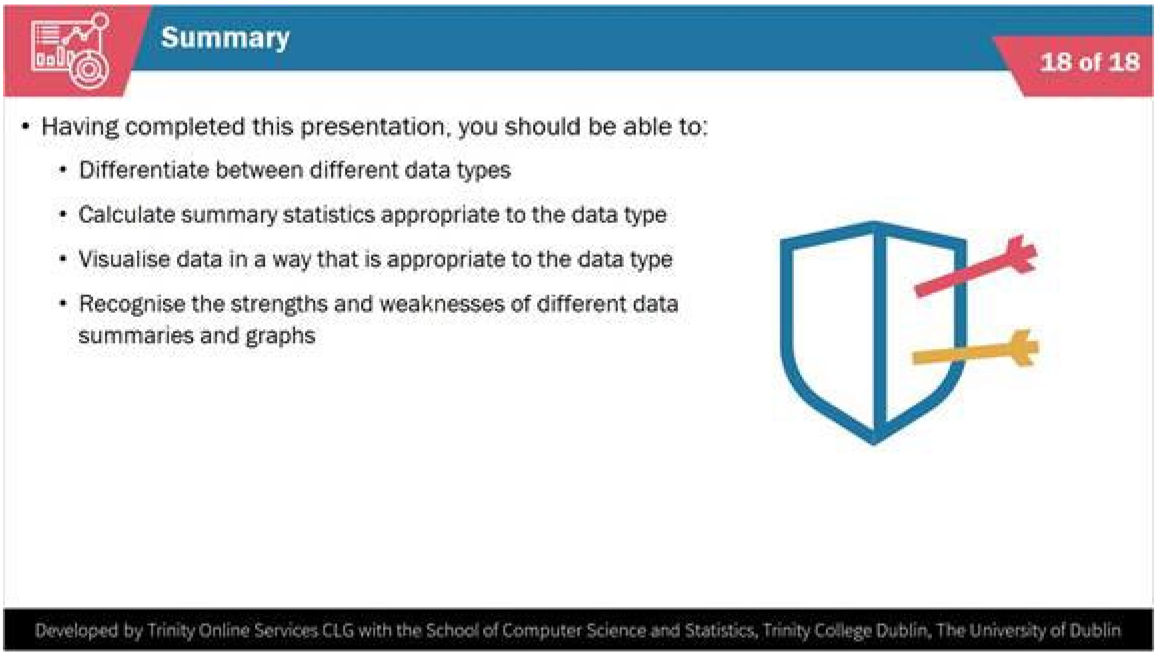
There is a link to R code available in the Extend section of your session homepage.

To conclude this session, we should say that there are many ways to compute the summary statistics and plot the sorts of graphs that have been discussed here. Excel will compute the summary statistics that have been defined, and also some of the plots. The R programming language for statistics, that you will be learning about in this course, will do all of the summaries and plots. Indeed, all of the plots in this presentation were created in R. The code to do it is available in the Extend section of the session homepage. Once you have learnt some of the basics about R, I recommend looking at the code to see how you can create these sorts of plots as well.

Images(s):

R-project (2016). *R Logo [Illustration].* From https://www.r-project.org/logo/Rlogo.svg under CC-BY-SA 4.0 license

Slide 18: Summary



Summary

18 of 18

- Having completed this presentation, you should be able to:
  - Differentiate between different data types
  - Calculate summary statistics appropriate to the data type
  - Visualise data in a way that is appropriate to the data type
  - Recognise the strengths and weaknesses of different data summaries and graphs

Developed by Trinity Online Services CLG with the School of Computer Science and Statistics, Trinity College Dublin, The University of Dublin

Having completed this presentation, you should be able to:

- Differentiate between different data types
- Calculate summary statistics appropriate to the data type
- Visualise data in a way that is appropriate to the data type
- Recognise the strengths and weaknesses of different data summaries and graphs