**IOC Topic 11b – Advanced Data Science**

Transcript & Notes: PART 3

Author: Dr. Robert Lyon

Contact: robert.lyon@edgehill.ac.uk ([www.scienceguyrob.com](http://www.scienceguyrob.com/))

Institution: Edge Hill University

Version: 1.0

**Topic 11b, Part 3**

**Introduction Slide**

Hello and welcome to Part 3 of Topic 11b, Advanced Data Science. During this topic I'll introduce what data science is, the basic principles underpinning data science, and some important data science tools that may be unfamiliar to you. My name is Dr. Robert Lyon, and I’ll be taking you through the learning material.

**Slide 1**

What material will we cover while studying this topic? Well, this topic aims to introduce…

* What data science is all about.
* Key concepts underpinning good data science – primarily the scientific method.
* Useful terminology that will help you navigate the world of data science.
* Important tools crucial for successful and reproducible data science – these are the tools provided by Statistics.
* Data collection & Experiment Design practices.
* Probability basics – very important for statistical inference.
* Data distributions that describe the characteristics of data.
* Hypothesis testing – a formal method for testing predictions.

The aim: to help you understand what it means to be a data scientist and to get you familiar with data science tools. In this part we’ll be looking at statistics, an essential tool for data science. We’ll also be covering data collection and how this relates to experiment design.

**Slide 2**

As data scientists we seek to answer questions using statistical methods and a collection of observations that we call data. These observations may be obtained through the entry of personal information into a website, the automated monitoring of user web activity or financial transaction records and so on. Data is at the core of everything we do in this field. Thus it pays to understand a little a about our data.

Data is a collection of observations described using variables. We use some simple notation to describe variables: . Here we have a dataset showing details of a number of U.K. rivers. There are three variables recorded in this data. The name of the river , the length of the river in kilometres , and the flow volume of the river in metres cubed per second .

Data shown on the slide from: <https://en.wikipedia.org/wiki/Longest_rivers_of_the_United_Kingdom>

**Slide 3**

We usually describe data in terms of population and sample data. A population is a complete dataset that contains all potential observations of an event or phenomena. A sample represents a subset of a population chosen in some way. An observation is an individual example from the population or from a sample. Population datasets are hard to come by as they are expensive to obtain and hard to store (due to their size). In such cases we can choose a sample of data from the population to tackle our data science research questions.

For example, if we wish to determine river quality in the U.K. a population dataset would include quality readings from every stream, brook, tributary and river in the country. This may be overkill. Instead we need only measure rivers of significant size, creating a sample. If the sample is chosen well, it will be representative of the river population as a whole. We can then use the sample to make estimates and inform environmental decision making.

Data shown on the slide from: <https://en.wikipedia.org/wiki/Longest_rivers_of_the_United_Kingdom>

**Slide 4**

There is some basic terminology we should be familiar with, when dealing with data (like that shown here) and variables. Variables may be numerical – which includes discrete variables, which are whole numbers such as number of customers or individual sales, or continuous variables which have decimal components such as temperature or weight.

Variables may also be categorical. Regular categorical variables describe categories such as human or animal, whilst ordinal categorical variables are categories which have a natural ordering. For example, the insurance coverage levels none, partial and comprehensive are ordinal variables.

**Slide 5**

Relationships may exist between the variables in a data set. You may need to verify the presence of a relationship, or prove that no such relationship exists. If variables are related or “associated”, then changing the value of one may impact the other. The variable that creates change is known as the independent variable. The variable being affected is known as the dependent variable. For example, an individuals caloric intake is a independent variable that affects their mass.

**Slide 6**

Relationships between variables can be represented graphically. Here we have some data. We can use a scatter plot to represent this data. A scatter plot allows two variables to be visualised using what we call a cartesian coordinates. In this coordinate system, we plot each observation in our dataset using coordinates. Thus, if we choose one variable from our dataset as the coordinate, and one as the coordinate, we get data ready for visualisation that looks like this. We can see how the data is plotted using the first data point as an example. Here we can see that the coordinate value is equal to 2.5, while the coordinate value is equal to 25. You can probably discern a trend in this data. As the value of the independent variable increases, the value of the dependent variable increases too. We call this a positive linear correlation.

**Slide 7**

In this example if the value of the independent variable decreases, the value of the dependent variable also decreases. We call this a negative linear correlation.

**Slide 8**

In some instances what we view as the independent variable, may have little to no apparent effect on the dependent variable. In such circumstances we say there is no discernible relationship of significance.

**Slide 9**

Just because there appears to be a relationship between two variables, this does not necessarily mean that there is. In some cases, correlation does not imply causation. That is, just because the independent variable appears to affect change in the dependent variable, this does not mean it is responsible for the change. Let’s consider a famous example.

Here we have an annotated scatter plot. An independent variable describing chocolate consumption in kg per capita per year, is shown on the -axis. The dependent variable is the number of Nobel laureates per 10 million inhabitants which is shown on the -axis. The plot appears to show that countries consuming more chocolate produce the most Nobel prize winners. Whilst this plot appears to show a correlated relationship between these two variables, there isn’t one. There is no connection between them in reality – higher chocolate consumption probably doesn’t make a population smarter. Before we move on remember, variables can be associated or independent, but not both.

Flag Icons credit: Muharrem Şenyıl, Creative Commons License 3.0.

**Slide 10**

Before we go any further, let’s spend some more time thinking about variables, correlation and causation. When ready please watch the video (https://youtu.be/GtV-VYdNt\_g) on the slide.

**Slide 11**

* We sometimes encounter what we call confounding variables.
* These can be correlated with both the independent and dependent variables leading to spurious associations.
* Such variables are often unaccounted for (or not understood) when designing our experiments.
* We must try and ensure we aren’t being influenced by such variables in our work.
* We can mitigate the impact of confounding variables by controlling for them.
* This involves ensuring that these variables don’t change during experimentation.

When ready please watch the video (https://youtu.be/sKkCZqqZ3qE) shown on the slide. It discuses confounding variables in more detail.

**Slide 12**

* All of the variables considered so far, can belong to a particular class of variables that we call random variables.
* Random variables are numerical variables whose values are determined via the outcome of a random event or phenomena.
* Most variables we’ll encounter will be random. This includes things like how many website visits google will receive today, or the number of people viewing these slides at any given time.
* Random variables can be discrete or continuous.
  + Discrete random variables take on exact integer values (that’s whole numbers). An example would be the number of children or pets that you may have.
  + Continuous random variables on the other hand take on real values. An example would be height in centimetres or mass in kilograms.

Watch the video (https://youtu.be/dOr0NKyD31Q) on the slide to gain a better understanding of random variables. These will become more important as we progress.

**Slide 13**

When dealing with data consisting of random variables, there are three main types of study that we can undertake.

Sample studies involve estimating the values of a parameter in a population – this parameter could be the average salary of people in data science, or the average hours worked per week in the U.K. This study approach is very useful when the population is too large to analyse, i.e. there is too much data to collect or is too costly to collect.

Observational studies collect data in a way that does not directly interfere with how the data is created. For instance, we may collect data from a group of patients being treated for a specific condition. We could ask them to fill in a survey that allows them to describe how they feel. In such cases we are simply observing the results once the survey is finished. Observational studies come in two forms. A prospective study involves collecting and analysing information as events unfold. Retrospective studies analyse data after events have taken place, for example once the patients have left the medical trial.

Finally, there are Experimental studies. These are concerned with investigating the possibility of a causal connection between variables.

**Slide 14**

Please take a moment to watch the video (https://youtu.be/SaP1O0i1bdc) on the slide, which covers experimental studies in more detail.

**Slide 15**

* Let’s consider the main principles of experimental design, crucial for designing sound experimental studies. We begin with Controlling.
* When designing an experiment, we do our best to control for any differences between the experimental and control groups, which could skew our results.
* It can be difficult to control for everything – we often have imperfect knowledge and are sometimes unaware of variables that should be controlled for. For instance, if segmenting customers based on their personal details, we may forget to control for a subtle characteristic.
* Randomisation helps protect us from issues arising from variables we forget to control for.
* We must randomise the cases we choose from a population to account for any variables not controlled.

**Slide 16**

* We must design experiments that are replicable.
* If an experiment cannot be reproduced, we cannot validate the original results.
* Experiments become non-replicable when the data or tools used during an experiment are discarded.
* Good experiments come with logs, that describe what was done in sequence allowing for reproducibility.
* Sometimes we may know, or suspect, that variables other than the independent variable, affect the dependent variable.
* Under such circumstances we may group cases from the population based on this variable into blocks and then randomly select cases from each block to form the experimental and control groups.
* This strategy is known as blocking. For instance, if we are looking at the effect of income on credit risk, we might first split the data into low income and high-income blocks. We can then randomly assign half the examples from each block, to the control group and the other half to the experimental group. This strategy helps ensure that each group has an equal number of low income and high-income examples. We’ll encounter similar issues shortly when considering data sampling.

**Slide 17**

There are a couple of more approaches to experiment design you may be familiar with.

* For instance, in the medical domain, researchers testing medications may keep patients in the dark about the treatment they are receiving.
* In this case the patients are said to be “blind”, thus this is a blind experiment.
* This helps researchers to avoid influencing patients simply by telling them about the medication they’re given – this helps avoid the placebo effect.
* In some cases, experiments are double-blind. Here the researchers don’t know about the treatment’s patients are receiving, thus any subconscious hints they may give off about the treatment will be avoided.
* These approaches are important for data science. For instance, suppose we’ve been asked to model user behaviour for an entertainment website. The aim is to find trends in viewing habits and how those correlate with time. If users are informed that their use will be monitored, they will likely change their behaviour. Though as you will find out later, users must be kept informed about what their data is used for.
* Experiments with no “blinding” are known as open trials.

**Slide 18**

* We’ve now learned about the different types of study we can undertake – sample, observational, and experimental studies. We’ve even looked at experimental studies in depth.
* All three types of study require the acquisition of data from some population.
* Data science questions, and research questions in general, are usually targeted toward a specific population.
* There could be very many examples in a population. Thus, it is usually too expensive to collect all the data from the entire population.
* Instead we collect data from an unbiased sample of the population. This represents a subset of examples called “cases”.
* Ideally, we aim to undertake data science investigations on large representative samples of data. We must often create such samples for ourselves.
* This involves selecting a sampling methodology which we can apply to the population. Sampling methodologies are important, as they can help us reduce bias in our chosen data samples.

**Slide 19**

* The most commonly used methodology is random sampling.
* Given a population this method simply chooses cases from the population at random, without taking any characteristics of the data into account. This is also known as random sampling without replacement. This means once a case has been randomly chosen, it won’t be chosen again.
* This type of sampling can be very effective.
* However, consider the following situation. Suppose you’re working for a credit card company. You’re asked to choose a default credit limit for new customers, based on a random sample of 100 cases taken from a population of customers. Sounds ok so far.
* Next, I tell you that the customer population is skewed toward those with incomes over £50,000. Over 90% of customers of this company earn above this amount. A random sample from the population is very likely to choose only high earning customers, since the population contains so many people in this category. We could end up with 100% of sampled customers cases being high earners. So simple random sampling can produce misleading samples and therefore mistakes – ultimately in this example, we may set too high a credit limit for new customers who burden themselves with unmanageable debt.

**Slide 20**

* We can think about this problem using basic probability. If 90% of the samples belong to customers earning over £50,000, that means there is a 90% (4 in 5) chance, that the first sample chosen will belong to the high earning group.
* Since samples are not replaced after being chosen (i.e. each case can be randomly chosen only once), this selection probability drops over time. Whilst for the less well-off group, the probability of selection increases over time.
* With each sample draw from the population, the probability of picking a case from the high earning group diminishes.
* If we take enough samples, eventually we’ll start to selecting cases from the less well-off group. But to get to this point we may need to set to a very high value!

**Slide 21**

* When randomly chosen samples become intrinsically biased, due to the composition of the population, we can apply a method called stratified sampling to try and get an unbiased selection of cases.
* Here were randomly sample a population as before.
* Yet this time we maintain the proportion of high and lower earners in the sample by randomly sampling each group or “strata”.
* The resulting sample is split so that the proportion of cases in each group, reflects the split in the population.
* Stratified sampling is useful, as it allows us to preserve population splits in our samples.
* In other words, it lets us preserve the true population distribution.

**Slide 22**

* Suppose we have a population of 100,000, and the population is split so that 90% of cases belong to Group 1, and 10% to Group 2. These are the resulting strata.
* We need to create a sample of the population of size .
* How many random samples should we make for group 1 and 2?
* For group 1 we need to randomly sample times from this strata.
* For group 2 we need to randomly sample times from this strata.
* This will give us a sample representative of the population.

**Slide 23**

* In some cases random and stratified sampling may not help.
* This is true when preserving the population distribution is unhelpful.
* This applies when we’re trying to target rarer groups in our populations.
* In these cases we can use weighted random sampling. This works by weighting the sampling so that it favours one strata over another.
* Suppose we have a population of 100,000, and the population is split so that 90% of cases belong to Group 1, and 10% to Group 2.
* We need to create a sample of the population of size .
* How many random samples should we make for group 1 and 2? We can use the simple formula , which is simply the number of samples multiplied by the weighting.
* For group 1 we set the weighting to . We need to randomly sample times from this strata.
* For group 2 we set the weighting to e need to randomly sample times from this strata.
* The weights must add up to 1.

**Slide 24**

* There are many types of sampling methods available.
* The best one to use depends on the question you’re trying to investigate.
* It is therefore up to you as the data scientist to choose an appropriate method.
* When a sample contains very few examples, any investigation we undertake can only yield what we call anecdotal evidence.
* This type of evidence may be true, yet can be very dangerous to use.
* I would strongly caution against using anecdotal data for anything other than providing general impressions.

**Slide 25**

* There are some real-world issues to consider when thinking about sampling.
* Suppose you conduct a customer survey to collect a data sample. If only 30% of customers respond, is that sample fair and representative of all customers? If the sample is not representative you may have encountered non-response bias. Your data has been skewed due to the non-response of many customers.
* Another common issue arises in this scenario, when certain subsets of customers are able to complete the survey because it’s simply easier for them to do so.
* This is known as a convenience sample. For instance, suppose you want to evaluate the accessibility of your company headquarters. You setup with a clipboard near the canteen, since you assume you’ll be able to nab lots of people there. For some it is convenient to dine in the canteen (they can afford it), whilst others cannot. Thus, bias has been inadvertently introduced into the data via convenience.
* Before moving on, please watch the video (https://youtu.be/be9e-Q-jC-0) shown on the slide. It will help you understand these issues a little better.

**Slide 26**

We’ve reached a checkpoint. Stop here and take a rest if needed. Let’s recap what we’ve introduced so far.

* Data sets.
* Populations vs. samples.
* Different types of variable.
* Scatter plots.
* Different forms of correlation.
* Experimental studies.
* Experimental design.
* Sampling methodologies.

That’s quite a lot! Take some time to digest that material, then return when you can. When ready, proceed to the next slide to learn how we apply statistics to data.

**Slide 27**

* Once we’ve collected sample data, we can start studying it.
* The first step almost always involves computing summary statistics that describe the data.
* Such statistics are incredibly useful as they can reveal broad trends, are easy to interpret, and easy to compute. Let’s illustrate this via an example.
* Suppose a company wants to know if targeted advertisements lead to an increase in the volume of sales (i.e. number of individual items sold) across a broad range of consumer types.
* They form a research question – does targeted advertising increase sales volume?
* To answer this, the company collects sample data from 450 website customers chosen at random across a range of demographic groups.
* The groups are equally split into two – an experimental group who will be exposed to targeted advertising and a control group. This data is summarised in the Table shown. Here we see the number of individual item purchases per customer, according to which group they belong to. There is so much data here that it can be hard to analyse in one go.

**Slide 28**

To compute a summary statistic for this data, we first summarise our data as shown. Here we can see that the experimental group consisted of 225 customers. Of those 95 made 1 or more purchases. The group as a whole made 140 purchases in total.

Can you compute the following summary statistic in this scenario – what proportion of customers made a purchase of one or more items overall? What is the proportion per group?

**Slide 29**

* In total 195 customers made 1 or more item purchases.
* The proportion of customers making a purchase overall is given by 195/450 = 0.433333 = ~43%.
* For the groups we have the following.
  + The control group proportion is given by 100/225= 0.444444 = ~44%.
  + Whilst for the experimental group the proportion is given by 95/225 = 0.4222222 = ~42%.
* Here we see that the control group had a higher ratio of customers making purchases.
* But we note that the experimental group did yield more sales overall.
* We can compute the average sales per customer that made 1 or more purchases (a spending customer), to determine if there is a difference.
* Try to compute that now.

**Slide 30**

* The average sales per customer who bought something in the control group, is given by 124 / 100 = 1.24 = ~ 1.2 sales per spending customer.
* For the experimental group we have 140 / 95 = 1.473684210526316 = ~1.5 sales per spending customer.
* Summary statistics provide some initial evidence suggesting that targeted advertising did increase the sales per customer. Summary statistics are useful, but be careful with them. They may not generalise well past the data you currently have.
* Consider if they would still apply with respect to a much larger customer sample.

**Slide 31**

* The mean is a summary statistic that allows you to compute the average of a data set.
* There are actually two forms of the mean that we consider as data scientists.
* First there is the population mean, mu (), which is the average for the entire population. We can define mu as the sum of the obervations dividied by the total number of observations (). This notation may be confusing at first. If we have the population dataset shown in the table, then is the number of cases. While the sigma notation () tells us to add up all the variable values for example 1 to 5. Here we can see how this computation proceeds.
* The population mean is sometimes described with slightly different notation. This is written as pronounced “e of x”. This can be understood to refer to the expected value of If you think about it for a moment, this notation does make sense. The most likely value we can expect for a variable , is its mean (average) value. This is true so long as is a random variable. That is, a variable which assumes values randomly.

**Slide 32**

* Some encounter the upper-case sigma symbol and become uncomfortable with the notation. Questions such as what does it mean, or why use that symbol, regularly crop up.
* I want you all to understand the notation hence I take a moment to elaborate. It’s best to simply view the mathematics you encounter (if unfamiliar with it) as a new language. It may be unclear if you’re the grammar and syntax is completely new to you.
* Sigma simply means summation – which involves adding one or more numbers or variables together. Here we can see this notation used to add up the first three natural numbers (1, 2, 3).
* Sigma notation is often extended to make it more useful. This involves applying a range on the summation. The letter is used to denote the lower limit, and the letter the upper limit. So if we specify these variables we can restrict the summation.
* We often use Sigma to add up the values of variables – of which there are usually many. If our variables are not indexed, there is usually only one variable and we don’t need to use the summation symbol. If our variables are indexed, we can use the lower and upper limit to restrict which variable values we include in our summation. For example, suppose we are given which contains the values 4, 5 and 6. Then , and . So the sum of these variables is now easy to compute. If this notation is new to you, take a moment to digest it.

**Slide 33**

* Now back to the mean. There are two types. First there is the population mean. Second there is the sample mean.
* This is the mean of a data sample, which is taken from a population.
* The formula for the sample mean is almost identical to the population mean. Except it uses a different value for the parameter . Here little describes the number of examples in the sample – not the whole population.
* We normally denote the sample mean as x-bar ().
* The mean is an important metric, as it allows us to estimate the central value of a dataset. This is an important concept to understand.

**Slide 34**

* The mean helps us describe the centre of a dataset.
* However, it is also important to understand how variable or spread the data is – especially because this may help us determine if outliers are impacting our summary statistics.
* The distance of an observation from the population or sample mean, is called it’s deviation.
* We use two variability measures in statistics to measure this deviation – the variance and the standard deviation.

**Slide 35**

* The population variance is denoted by lower case Sigma squared ().
* It is defined by the following formula (which you don’t need to remember, no memory tests here): .
* While the sample variance is given by this formula : .
* The correct formula to use depends on whether or not your dealing with a population or a sample – and that’s for you to determine.
* How to interpret the variance? Well datasets which are less varied, deviate less from the mean value, which is at the centre of the data distribution. Datasets which are more varied deviate more from the mean, and thus have a higher variance.

**Slide 36**

* The population standard deviation denoted by lower case sigma (), is given by the formula: .
* While the sample standard deviation is denoted by , and is given by the formula: .
* Note the relationship between the variance and the standard deviation. The standard deviation, both for a population and a sample, is equal to the square root of the variance.

**Slide 37**

* What does deviation from the mean translate into in practice? We can visualise this to provide a more intuitive understanding of the standard deviation and variance.
* Suppose we have some data describing the values obtained by a random variable. It doesn’t matter what this variable represents, but it could be height, mass, age or anything else. The plot shows probability density on the -axis, which describes the likelihood of the random variable taking on an arbitrary value.
* We can see that the most likely value with the highest probability density is the mean value.
* The -axis shows the deviation from the mean value expressed in terms of sigma () units. 1 sigma represents 1 standard deviation from the mean.
* If the data is spread out, then the curve will spread over a wider range of values and thus exhibit a higher deviation from the mean. If the data is dense, it will exhibit a lower deviation from the mean.
* We have a name for data that produces a bell-shaped curve when plotted. We say that the data is normally distributed, or that the data is Gaussian distributed (named for Carl Friedrich Gauss who first defined this distribution).

**Slide 38**

* For data that looks like this, there are some characteristics to be aware of. In general, approximately 68% of observations will assume values within 1 standard deviation of the mean.
* Approximately 95% of observations will assume values within 2 standard deviations of the mean.
* Finally, approximately 99% of observations will assumes values within 3 standard deviations of the mean.
* Why is this important? Well, we can use the concept of deviation to determine how unusual an individual observation is. We can simply determine how far it deviates from the mean. Values that are 3 or more standard deviations from the mean are considered unusual.

**Slide 39**

Now that we’ve taken some time to think about deviation and spread, you’re ready to tackle the video (https://youtu.be/MRqtXL2WX2M) shown on the slide. Please watch when ready.

**Slide 40**

We covered some summary statistics so far. This includes the mean, variance and standard deviation. The following video (https://youtu.be/dn6lHbg5Vs0) will review summary statistics as a tool of exploratory data analysis. Please watch when ready.

**Slide 41**

* We’ve heard about the population mean and the sample mean.
* But in the real world it can be unusual for us to have access to the population mean.
* Consider an example. Suppose Amazon want to analyse and predict likely sales per consumer over the next 12 months.
* Amazon has data describing existing customers it could use for this task. However, it doesn’t have data for potential customers, i.e. those who’ve not yet shopped with Amazon. In this instance only sample data is available. Therefore, the population mean cannot be computed – not unless they investigate every potential customer on the planet!
* In this instance the sample mean is still very useful. It can be used to estimate the population mean. The estimate gets more and more accurate as the sample set increases in size. This is due to something called the law of large numbers. Due to this law, the sample mean approaches the population mean, as approaches infinity (). Watch the video (https://youtu.be/VpuN8vCQ--M) on the slide to learn more about this concept.

**Slide 42**

* The law of large numbers tells us we can use the sample mean to reasonably estimate the population mean if we have enough samples, which is clearly very useful.
* However, the population mean and the sample mean can be skewed. Thus, their ability to help summarise data can be diminished.
* This happens because the mean is not robust to outliers.
* Outliers are “extreme” data points that skew the average providing a misleading impression of the central point of the data. Outliers drag the mean toward them.
* Consider the data shown in the table. It describes self-reported salaries for credit card customers in London.
* We aim to estimate the optimal amount of credit to offer a customer. At first, we use the sample mean salary to guide us.
* In this example the mean salary is £39,000. This is much higher than all but one salary in the data.
* In such cases we can use different measures of centrality, to estimate the midpoint of the data. This is important as we can see how the mean has been dragged toward the outliers in the data.

**Slide 43**

* One way to overcome the influence of outliers on our analyses, is to use more robust statistics.
* For instance, we could use the median to estimate the midpoint.
* The median is the central value in the data when ordered – in this case, the value in row 3 (£24,000). If there are an even number of observations, then the median is the central two data points divided by two.
* We could also use the mode – this is simply the most common value found in the data.
* In this case the mode is £24,000.

**Slide 44**

* Data can be confusing however.
* Consider these three fictitious datasets which are summarised by way of a histogram.
* Believe it or not, these data sets have the same mean () and the same standard deviation ().
* This is why it becomes important to visualise our data, to better interpret the data we have.

**Slide 45**

* In the previous slide we introduced the histogram.
* This is a type of data visualisation tool, that describes the frequencies of observations / outcomes in data.
* Outcomes are first grouped in to bins covering ranges. Here we can see the first bin covers values in the range 0 to 1. The last bin covers values in the range 7 to 8.
* We count the frequencies of values falling in to these ranges, like those summarised in the table shown.
* Finally, we plot bars where bar height is equal to the frequency of examples falling in the bin range.
* The histogram is a simple but elegant data visualisation tool.

**Slide 46**

* There is some terminology associated with the characteristics of the data we analyse.
* Sometimes data is described as long tailed. This means the data has long tails which trail off from the distribution, possibly infinitely.
* Data can be left-skewed, right-skewed, or symmetric with respect to the data density.
* Data sets can be unimodal, bi-modal, or multi-modal – which roughly translates to a description of the number of peaks in the data.

**Slide 47**

We’ve reached another checkpoint. Let’s recap what we’ve introduced so far.

* Data sets
* Populations vs. samples
* Different types of variable
* Scatter plots.
* Different forms of correlation.
* Experimental studies.
* Experimental design
* Sampling methodologies.

While in the second half of this material you learned about,

* Summary statistics (mean, variance, standard deviation, mode and median).
* Why such statistics are important.
* Sample versus population statistics.
* The law of large numbers.
* Outliers and their impact on summary statistics.
* Histograms.
* Terms used to describe dataset characteristics.

This puts you in a great place to tackle our next topics - probability and data distributions.