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

Transcript & Notes: PART 5

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 5**

**Introduction Slide**

Hello and welcome to Part 5 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 part 5 we’ll look at hypothesis testing.

**Slide 2**

* Hypothesis testing is a statistical approach used to find the optimal answer to the questions we pose about the world around us.
* It uses available knowledge captured within data, to reach conclusions regarding the hypotheses we form. These conclusions are reached in a rigorous as opposed to an ad-hoc way.
* The hypothesis testing method is most useful, when undertaking experimental studies.
* Suppose we are asked to determine if a new medication works during an NHS trial.
* We form the hypotheses , the null hypothesis, and the alternative hypothesis . We then split some sample population into control and experimental groups.
* We can now use hypothesis testing to determine which of the hypotheses holds, over these groups.
* That is, we can determine statistically which hypothesis has the most evidence in it’s favour for each group.
* Here we are introducing the foundations of statistical inference, a topic central to both data science and machine learning.

**Slide 3**

* So far you’ve come across concepts from lots of different areas.
* You’ve learned about probability theory, the different types of data distribution (unimodal, bi-modal, multi-modal), the law of large numbers, and how to compute summary statistics over samples of data, and entire populations.
* We covered this material to help prepare you for the concepts I’ll very shortly introduce related to hypothesis testing. I’m sure you’re relieved that none of this time was wasted!
* So with that in mind, lets return to thinking about a distribution I’ve mentioned a few times during this course – the normal distribution.

**Slide 4**

* The normal curve is always a symmetric, unimodal, and bell-shaped.
* The shape of the curve is determined by two parameters.
  + The mean, 𝝁. This defines the location of the curve when plotted.
  + The standard deviation, 𝝈. This defines how spread out the curve is.
* Using this notation we can describe any normal curve via a pair, e.g. Here we can see this curve. We can then see how altering the parameters creates a different curve, with a different location and shape. A normal distribution that has a mean of zero and a standard deviation of 1 is special. It is known as the standard normal distribution.

**Slide 5**

* Please watch the video shown on the slide (https://youtu.be/mtbJbDwqWLE). It reviews the normal distribution in more detail.

**Slide 6**

* Suppose you’re given two normal distributions. These represent the test scores of a collection of students on two different tests.
* We then get scores for an individual student.
* They score 1800 on test 1, and 24 on test 2.
* We then collect details about the mean and standard deviation of the data for each test.
* The question is, did the student do better on test 1 or test 2?

**Slide 7**

* One way to answer this question, is to determine how many standard deviations from the mean each test result is.
* We’re assuming here the better result is the one further from the mean in the positive direction.
* We can use the 𝑍-score to determine how many standard deviations an observation 𝑥 falls above, or below, the mean.
* Let’s try that now. For test 1 we can compute the -score by substituting the variables we have, into the formula. This gives us 1800 minus 1500, divided by 300 (). When we work this out, we find that a test score of 1800 is exactly one standard deviation, or 1 sigma (1) from the mean. We can then do the same for test 2 as shown. Here we find that a test score of 24 is 0.6 from the mean result. From this we can conclude that the result for test 1 was better, given that it deviates further from the mean in the positive direction.

**Slide 8**

* Please watch the video shown on the slide (https://youtu.be/2tuBREK\_mgE). It reviews the -score in more detail.

**Slide 9**

* Suppose we wanted to know the percentile of the result for the student on test 1.
* This can be represented by the area below the score achieved by the student.
* The total area under the curve is equal to one. Think back to probability – there are lots of potential scores a student can get, but the probability of all those added together is 1. The probability of scoring a zero, plus the probability of scoring a 1 and so on, is equal to 1.

**Slide 10**

* So what percentile does the students’ performance correspond to?
* We’re looking for the probability that a score is less than 1800 in this case. In principle we could add up the probability of all scores below 1800 to get the answer.
* But this would be quite impractical to compute – there’s a lot of numbers to add up!
* Thankfully, there’s a trick we can use to easily compute the percentile, that applies to normally distributed data. This trick works by taking advantage of the characteristics of the normal distribution.

**Slide 11**

No matter how a normal distribution is defined in terms of it’s mean mu (), and standard deviation sigma (), the area under the curve is the same for any standard deviations. For example, for one standard deviation, irrespective of the shape of the normal distribution, approximately 68% of the data under the curve, falls within 1 standard deviation of the mean. It is also important to understand the relationship between the area under the curve and probability. The total area under the curve is equal to 1, and the probability of a random observation falling somewhere under the curve is also 1. So for example, an area under the curve equivalent to 0.682, falls within 1 standard deviation of the mean. This also tells us that 68.2% of the data falls in this region, and that there is a 68.2% chance that any random observation will fall into this range too! Now we can better appreciate this link between probability theory and the normal distribution. The area under the curve defined by a chosen deviation, can be interpreted as a probability.

**Slide 12**

We just found that we can interpret the area under the curve as a probability. Now consider this - instead of measuring the data between symmetric standard deviations, say from minus 1 to plus 1, we could measure between any two standard deviations under the curve. If we do this, we’ll be able to determine the area under the curve between these two points. Now If I say that the top plot represents 0% of the data, and the bottom represents 100%, perhaps it will become obvious how we can compute percentiles. Any points chosen between these extremes describe a proportion of the data corresponding to a percentile. Suppose we choose . This corresponds to zero deviation from the mean, thus it points to the mean value. We know that 50% of the data sits below this point from equal to negative infinity to equals zero. Thus corresponds to the 50th percentile. There is a 50% chance that a random data point chosen from the distribution, will fall into this range.

**Slide 13**

We’ve seen how a -value can be interpreted as a percentile by finding the area under the curve up to the -value, then turning that value into a probability. This approach works for any normal distribution. Now, because interpreting -values is something we need to do often, and given that -values correspond to the same percentiles for all normal distributions, some kind people have already computed all potential -values for us. These values are stored in tables. We call these normal probability tables.

This is a table detailing the percentile an observation falls into, for a specific -score. Suppose we have a -score, , which equals zero (). This -score exhibits no deviation from the mean whatsoever. This is in the exact middle of the distribution, thus it refers to the score 1500.

So corresponds to the 50th percentile in the data. This tells us that 50% of the data sits below a score of 1500. In the normal probability table this is the value we find for a -score equal to zero. The table is very big in reality, so I don’t show it all here.

**Slide 14**

We can now use the normal probability table to answer our initial question. If we look for the row that has equal to one, and find the appropriate column, we obtain an answer of 0.8413. This is 84.13%, or approximately 84%. So the student is in the 84th percentile for test 1. This means they did very well indeed. Only 16% of students did any better.

**Slide 15**

* There are actually two normal probability tables. One for when is negative, and one for when is positive. We used the positive table in the previous slide.
* You don’t need to worry about remembering normal probability tables.
* We can create them in code and extract values from them automatically.
* What matters is that you understand that:
  + normal probability tables exist.
  + that they can be used to determine what percentile an observation is in.
  + and finally, that you must usually compute the -score to make use of them.

**Slide 16**

* Sometimes we may not be looking for simple percentiles for our data.
* We may wish to know what proportion of our data sits between two specific positions.
* We can use the concepts we’ve already learned to answer such questions.
* We can do this by first calculating percentiles and then subtracting them from 1.
* Once we determine the reminder, we can use this in further calculations. For example, suppose we want to know the proportion of data contained within this region. We can calculate this by first determining how much data is under the curve on the left-hand side of the region. Then we can determine the area under the curve on the right-hand side. Now we can just do some simple math to calculate the area in the middle.

**Slide 17**

* When we collect data, it usually represents a sample of data from a much larger population.
* When we compute summary statistics over the sample, how accurate are they? After all, they are just estimates.
* The sample mean you compute for any data sample, won’t be exactly equal to the population mean. The sample mean might vary from the true population mean quite a lot, especially if the sample is small. We can try and quantify the uncertainty in our estimates via a measure called the standard error.
* The standard deviation associated with an estimate is called the Standard error of an estimate.
* The standard error for the sample mean is an important statistic. It gives us an indication of the uncertainty we have in our sample mean.
* The standard error for the sample mean is given by the formula shown on the slide, where sigma () is the standard deviation, and is the number of observations in the sample.

**Slide 18**

* The sample mean obtained for a collection of observations, represents an estimate of the true population mean ().
* If we were to take another random sampling from the population, and then recompute the sample mean, we’ll get a slightly different mean estimate.
* If we were to take many random samples from the population, and compute the sample mean for each of them, we would obtain a distribution for the sample mean, as shown on the slide. It is important to note that this distribution is normally distributed.
* The average of this distribution, is going to be very close to the true population mean. Due to the law of large numbers we know that if we take infinite samples, the mean of this distribution would equal the population mean! But we can’t be expected to create such a distribution each time we want to estimate the sample mean. It is clearly impractical to take lots of samples, and compute the sample mean for each of them!
* But if we can’t compute lots of sample means, how can we be sure that a single estimate is useful? How confident can we be in our sample mean estimate?

**Slide 19**

* We can apply what we call “confidence intervals” to our estimates, to help quantify our confidence level.
* A confidence interval contains the plausible range of values for an estimated parameter, when taking into account uncertainty in the estimate, using the standard error.
* For example, suppose we have an estimate for some normally distributed parameter equal to 10.
* Suppose we also know the estimated parameter has a standard error of 1.
* This means it can plausibly deviate by plus or minus 1.
* We can model this by creating an interval that takes into account this deviation.
* The plausible range is given by the parameter plus 1, and minus 1. This interval now represents our uncertainty - we know the parameter likely falls somewhere in this range. We just don’t know for certain exactly where it falls.
* This is a confidence interval. We can interpret this interval graphically. As the standard error is roughly equal to 1 standard deviation from the mean, we can see that this interval covers 68% of the area under the curve.

**Slide 20**

* We can now formalise this understanding. We can construct a 95% confidence interval over the parameter we wish to estimate, in this case the sample mean, via the following simple formula: estimate The symbol that looks like a plus with a line underneath it (), is called the plus-minus sign. This tells us that there are two outputs from this equation. This can be explained using a simple diagram as shown. I include a description of the terms, to make things clearer. Next, we’ll look at a practical example to see how this works.

**Slide 21**

* Suppose we have some sample data.
* The data has a sample mean of 95.61.
* The standard error for the mean is 1.58.
* We can compute the 2 sigma confidence interval for this example, by plugging in the values for the sample mean and standard error. This gives us the following results. Here’s an expanded version of the computation. It shows how the values were obtained.
* We can see how these values correspond to regions on the plot.

**Slide 22**

* Please watch the video shown on the slide (https://youtu.be/ulBG4SOHKS0). It reviews confidence intervals, and the standard error, in more detail.

**Slide 23**

* What I showed you in the last example wasn’t exactly right. I showed you how to compute an approximate 95% confidence interval using this formula. Doing this was convenient, because the value 2 we used, corresponds to exactly 2 sigma! An exact 95% confidence interval can be computed with the updated formula shown. Here we’ve replaced the value 2 for 1.96. We have to update the formula, as I’ve been using rounded numbers for simplicity in my earlier slides. Yet in reality, 95% of all data falls with 1.96 deviations from the mean of a normal distribution.
* Perhaps this confidence level isn’t good enough for you – well you can compute a 99% confidence interval using the formula shown.
* These intervals will apply to normal data only.

**Slide 24**

All the material we’ve covered so far has been building up to this – starting to test hypotheses.

* We can start testing competing hypotheses using confidence intervals. Suppose we have a dataset describing the finishing times of runners in a race.
* We want to determine if the runners finished in a faster time this year, compared to last year.
* We form two competing hypotheses for this data. The null hypothesis is that there is no difference in average finishing times. The alternative hypothesis, is that the average time was different this year compared to last.
* The average time for last year’s run was 93.29 minutes (93 minutes and 17 seconds). We can reframe our hypotheses given this data. The null hypothesis now says, that the mean time for this years run, is equal to the mean recorded last year. While the alternative hypothesis now says that the mean time for this years run, is not equal to the mean recorded last year – something has changed.

**Slide 25**

* We can test these hypotheses using confidence intervals.
* To do this, we first we collect a sample of data from this years runners. We select runners at random, and compute the average time. We obtain a sample mean of 95.61 minutes.
* We compute the sample standard deviation and find it to be 31.2. The finishing time deviates by up to approximately 31 minutes between runners.
* We now compute the standard error for the data. We find this to be 3.12.
* Next, we try to build a 95% confidence interval around our sample mean. We first compute the value we will subtract and add, to our mean value, to build the lower and upper limits for the interval.
* Then we compute the intervals as shown. We can see here that the value for the null hypothesis, 93.29, falls within our confidence interval for the sample mean.
* Because the mean of last years run, falls within the range of plausible mean values for this years data, we cannot say that the null hypothesis is implausible.
* Thus, we fail to reject the null hypothesis – given the sample we have, there is no evidence to claim this years runners were faster or slower than last years!

**Slide 26**

* In our last example, we decided to accept the null hypothesis, that there was no difference in the average run time.
* There is a chance, this decision was wrong – even if it was the best decision with the evidence available.
* In general, for any hypothesis test, there are four potential test outcomes, summarised by this table.
* If the null hypothesis is true, and we do not reject it, we’ve made the correct decision. Likewise, if the alternative hypothesis is true, and we accept it, we’ve similarly made the correct decision.
* If we rejected the null hypothesis and accepted the alternative hypothesis when we shouldn’t have, we’ve made a Type I (type one) error.
* While if we reject the alternative hypothesis when we shouldn’t, we’ve made a Type II (type two) error.
* When running hypotheses tests, we aim to minimise the errors we make. Confidence intervals are great, but alone they don’t really help us achieve that. Instead we try to use significance levels to determine how significant a result is, before making a decision.

**Slide 27**

* Confidence intervals are simplistic when it comes to hypothesis testing.
* Suppose we use a 95% confidence interval for some sample mean data, where the null hypothesis is accepted if the sample mean falls within 1 standard deviation of the mean.
* Sometimes the evidence against the null hypothesis may be overwhelming, like here. In this example the sample mean sits far away from the mean value represented by the null hypothesis.
* But sometimes we may be on the cusp of rejecting the null hypothesis, but don’t quite have enough evidence to reject it as we see here.
* In these situations it’s helpful to be able to quantify our confidence in the decisions we make. We can do this using a tool called, the P-value.

**Slide 28**

Please watch the video shown on the slide (https://youtu.be/a\_l991xUAOU). It reviews decision errors in more detail.

**Slide 29**

* P-values allow us to test the strength of the evidence against the null hypothesis.
* The P-value is a conditional probability – it is the probability of observing data at least as favourable to the alternative hypothesis as our current dataset is, if the null hypothesis is true.
* It may help to think of this description via a tree diagram. We can see here that the p-value is simply assessing the probability of seeing data this favourable to the alternative hypothesis, given that the null hypothesis is true.
* We usually use a summary statistic such as the sample mean to help compute a P-value.

**Slide 30**

* It’s best to illustrate how we can use P-values via an example. Then you’ll be able to apply the method for yourself by following the steps provided.
* A national sleep study suggests students sleep on average 7 hours per night.
* You’re a data scientist at a local education authority, and you are asked to determine if students in your area are similar.
* You collect data from a student sample (), and find that students in your area are sleeping on average over seven hours.
* You want to verify that your students are indeed different from the national sample.
* You form two hypotheses.
  + The null hypothesis, , which is that there is no real difference between the students in the national survey and your area.
  + The alternative hypothesis, , which is that there is a real difference - your students get more than 7 hours sleep on average.

**Slide 31**

* We begin by considering the data collected during the national sleep study. This data is normally distributed, with a mean of 7 hours as shown.
* To test our hypotheses, we select students at random from our local student cohort, and compute the average time. We obtain a sample mean of 7.42 hours.
* We compute the sample standard deviation and find it to be 1.75 hours. Remember, that’s the deviation between students in our local cohort.
* So how to interpret what we have so far? Well, we can look at the national data and ask ourselves – how different is the sample mean we computed, from the mean of the national trial?
* We may ask, what is the probability of seeing a sample mean this extreme, in the national trial data? To answer this, we can use the probabilities obtainable using the -score.
* We compute the -score as follows. This describes the area under the curve, below a value of 7.42.

**Slide 32**

* We take the -score, and lookup the area under the curve using the positive normal probability table. We can see the relevant part of the table here.
* We first find the row that contains 2.4, then the column that contains the value corresponding to the second decimal place in . This takes us to the 0.07 column. We find a probability value here of 0.9932 or 99.32%.

**Slide 33**

* We now know the probability of obtaining a sample mean less than 7.42 – it’s approximately 93%.
* This appears to suggest that it is pretty unlikely to obtain our sample mean, if the null hypothesis is true.
* Now think back to the P–value. The P-value is the probability of observing data favourable to the alternative hypothesis, if the null hypothesis is true.
* In other words, if the mean sleep time is indeed 7 hours as the national study suggests, then what is the probability of obtaining a sample mean of 7.42 or higher?
* We compute the P-value to quantify this via simple subtraction. The probability of observing a sample mean of 7.42 or higher, if the null hypothesis is true, is just 0.007 which is 0.7%. That is very low! What this is saying, is that it is highly unlikely that we’d obtain this sample mean if the null hypothesis is true. This is evidence in favour of the alternative hypothesis. But can we reject the null hypothesis?
* Before we can decide, we must set a significance level. If the P-value falls below this level, then we consider the result to be significant. Otherwise it is not significant.
* Here we set our significance level alpha () to 0.05, which corresponds to a 95% confidence level. Given that our P-value is below our significance level, we reject the null hypothesis in favour of the alternative hypothesis here.

**Slide 34**

* What we’ve just done, is perform a one-sided hypothesis test. We were only computing the probability of seeing a sample mean greater than 7 hours, so we only considered one tail of the probability distribution.
* But if our hypotheses are defined differently, we may need to perform a two-sided test. Suppose we change the alternative hypothesis, so that it no long states that the sample mean is greater 7, but rather that it is not equal to 7.
* Suppose we take a different sample from our local student population. We select students at random, from our local student cohort, and compute the average time. We obtain a sample mean of 6.83 hours.
* We compute the sample standard deviation and find it to be 1.8 hours. Remember, that’s the deviation between students in our local cohort.
* We now ask, what is the probability of seeing a sample mean less than 6.83 in the national trial data? To answer this, we again use the probabilities obtainable using the -score.
* We compute the -score as follows. This describes the area under the curve, below a value of 6.83. This has allowed us to compute the probability for the left-hand side of the distribution. This probability is approximately 21%.

**Slide 35**

* So there’s an almost a 21% chance of getting an “extreme” sample mean less than 6.83.
* But we aren’t just looking for the probability of a sample mean less than 6.83. We’re looking for the probability of a value not being equal to 7. So we must consider the likelihood of getting a mean value larger than 7 too.
* As the normal distribution is symmetric, we’ve already computed the chance of obtaining a value equally “extreme” but on the right-hand side of the distribution.
* We know that this right-hand tail also covers 20.9% of the distribution as shown.
* We compute the P-value in this instance by adding up these two probabilities.
* This gives us a p-value of 0.4180 or 41.8%.

**Slide 36**

* Now we set our significance level alpha () to 0.05, which corresponds to a 95% confidence level.
* Given that our P-value is much greater than our significance level, we reject the alternative hypothesis in favour of the null hypothesis here.
* We can see that this makes sense, given that 58.2% of the data is at least as extreme as our sample mean. So our sample mean is not significantly different from the national trial mean of 7. Hence, we have to reject the alternative hypothesis.

**Slide 37**

* Some of what we’ve covered here may not make sense yet. That’s ok, because nobody becomes a hypothesis testing expert over night!
* What matters is that you appreciate what's happening and why.
  1. That we form hypotheses to answer questions about our data.
  2. We collect data samples to test them.
  3. We compute summary statistics over the data sample, such as the sample mean and sample standard deviation.
  4. We compute the -score and use this along with normal probability tables to determine the area under the curve.
  5. We use these areas to represent probabilities as p-values, and evaluate them with respect to some significance level, alpha().
* A little practice will help make these ideas clearer.

**Slide 38**

* Please watch the video shown on the slide (https://youtu.be/-FtlH4svqx4). It reviews hypothesis testing with p-values, in more detail.

**Slide 39**

* Please watch the video shown on the slide (https://youtu.be/mvye6X\_0upA). It reviews one and two tail hypothesis testing, in more detail.

**Slide 40**

* Ok, so you’ve sat through a large number of slides, and covered a lot of material.
* You probably don’t realise, but you’ve picked up a great deal of fundamental statistical knowledge that’s crucial to data science.
* To illustrate this, I’ve created some activities for you to try – they’ll let you apply the knowledge you’ve picked up.
* The activities can be found in a Google Colab notebook. It contains code, hints, tips, plus links to videos and tutorials that you’ll find useful. It will guide you through applying the concepts you’ve learned so far, and show you why probability theory and hypothesis testing is useful!

The notebook includes instructions on how to use it, so please read through those carefully. The main thing to remember is that you can only access the notebook via the Google chrome web browser.

The link to the resource is shown on the slide.

https://colab.research.google.com/drive/1sq5txv7MQy4uXxBKVjI1E9DaI6sHxpmA

**Slide 41**

There are lots of useful resources out there, that will help you build your knowledge in the areas we’ve covered.

* I strongly recommend looking at a book called Open Intro Statistics (<https://www.openintro.org/stat/textbook.php?stat_book=os>). This is an entirely free introductory text that will teach you all the statistics you need to get started in data science.
* I can also recommend the book Data Science from Scratch: First Principles with Python 2nd Edition.
* Although I haven’t read the following books, these may be of use.

I also recommend a couple of other resources, that you may find interesting.

Tool

* Kaggle (<https://www.kaggle.com/>) – an online platform where you can tackle data science challenges.
* Toward data science – a website where data science practitioners share ideas, tutorials and advice <https://towardsdatascience.com/>.

**Slide 42**

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

* Normal distributions.
* The -score.
* Probability tables.
* Standard error.
* Confidence intervals.
* And finally, Hypothesis testing.

From here you can pursue the activities I’ve provided in Google Colab, or watch the next set of slides which cover the ethics of data science. It’s entirely up to you.

Although this isn’t the last set of slides in the course, it happens to be the last set I’ve recorded. So I take this opportunity to wish you all the best in your future tech careers.