FAQ - The Data Science Course

General questions

1. I haven't received my gift at 50% completion. What should I do?

Write in the Q&A board and we will send it to you as soon as possible.

2. How can I get a certificate for completing the course?

Please refer to this article which thoroughly explains how to attain your certificate of completion!

https://support.udemy.com/hc/en-us/articles/229603868-Certificate-of-Completion

3. Why is my quiz/video not loading/rendering?

Unfortunately, such issues occur from time to time.

You can try and refresh the page several times. If this does not solve your problem, you can wait for a while or contact Udemy support at:

https://support.udemy.com/hc/en-us

4. How can I download all the course resources?

Well, read above!

If you have downloaded the PDF, refer to Section 1, Lecture 3, where we explain how to download all the materials for the course.

For your convenience, here is a link to the associated lecture. https://www.udemy.com/the-data-science-course-complete-data-science-bootcamp/learn/v4/t/lecture/12955006?start=1

5. What should I do after completing the course?

A note from your instructor:

I really prefer the more math-heavy books. Especially after completing this course, they will help you reinforce what you have learned and give you a better understanding of the mechanics of behind the formulas and the techniques used. In my opinion, introductory statistics books will overlap a lot with what we do in this course and will not prove that beneficial.

Light and fun:

An interesting read that may greatly aid you with your understanding is: 'A Field Guide to Lies and Statistics: A Neuroscientist on How to Make Sense of a Complex World' by Daniel Levitin. I found the book quite fun, while not technically heavy at all. It looks into different ways people misinterpret statistics (often on purpose).

If you are comfortable with Math:

I'd suggest the: 'Probability and Statistical Inference' by Nitis Mukhopadhyay. That must be one of the better readings on the topic (probability, distributions, confidence intervals, hypothesis testing, Bayesian statistics).

Then, for regressions (in a business / economics context), I would strongly suggest 'Econometric Analysis' by William Greene. That book is especially good, as it has several appendices, which include the linear algebra, probability, etc. that you may not know, or may have forgotten.

But these two books are really math heavy (probability, linear algebra, etc.).

If you want something more programming oriented:

Then one of the classics is 'Python Data Science Handbook' by Jake VanderPlas. Probably you know this one. It is especially good because it looks into both NumPy and pandas and how to manipulate data using them.

6. Can I download videos from the course?

Unfortunately, you cannot currently download this course on a desktop computer.

However, if you can use a mobile phone to watch this course, you can download the Udemy App. Then, you can download the lectures and watch our content offline.

Downloading Courses on iOS: https://support.udemy.com/hc/en-us/articles/229603928-Downloading-Courses-on-iOS

Downloading Courses on Android: https://support.udemy.com/hc/en-us/articles/115006973308-Downloading-Courses-on-Android

Lecture Specific Questions

1. When do we assume our data is from a population, rather than a sample?

Let's say, for example, in a biomedical field when a scientist is doing research using animals - some animals are given treatment A, and some animals are given treatment B, the rest are not given any treatment (control group). Each treatment group has 10 animals, so we have 30 in total.

In this scenario the population in question is **all animals.**

Why?

Because we want to find a cure for **all animals**, right? We don't want to find a cure for 30 of them. There's no point in doing that.

***

What about these three groups?

Now, we've got 3 **samples**, 10 each. In real life, by the way, we would usually draw a sample of 30 and then randomly divide it in 3. By all means these two scenarios are the same, but nobody is going to perform the sampling 3 times (10 animals each).

Once this is done, we have 3 groups of animals. Ideally, we want each of these 3 samples to be: random and representative.

***

Let's explore another example.

Imagine you are trying to cure all animals, and you've picked 10 dogs.

That would be highly uninformative.

What about horses, pigs, cows, cats, elephants? Given the same treatment, elephants will definitely react in a different way than dogs.

This points us to two potential problems:

1) Animals may be too different.

That's why treatments that work on rats don't always work on people and vice versa.

2) Given that there are lots and lots of different animals, a sample size of 30 (10) won't be enough (most animals would not be represented).

***

So, let's simplify to 1 animal - dogs.

You get 30 dogs.

You test A on 10 of them, test B on 10 of them, and monitor the other 10.

These 30 dogs are a **sample** of the **population of all dogs.** They are not the population itself.

***

We are using samples for several important reasons:

1) Time efficiency -> it is much easier to draw a sample of 30 dogs, rather than ... all dogs (which is also impossible).

2) Monetary cost -> it is much **cheaper** to work with 30 dogs. Think about how many people must be involved to test a treatment on 5000 dogs. Where are you going to do this? Who is going to feed them, take care of them, walk them, etc.

3) Safety -> what if treatment A kills 9 out of 10 dogs? We don't want this to happen to thousands of dogs. That's why we **tested the treatment on only a sample** and not the whole population.

***

Let's think about a case where it may have been the population.

Imagine that there is a village which has an unknown disease which is killing the dogs there. We find that there are only 30 dogs that are contaminated.

We take them away from the others, so they don't spread the virus and then we decide to find a cure. We proceed as you indicated in the question.

This is the population of dogs that have this disease.

If we take 10 dogs out of the 30, it would be a **sample.**

All 3 samples taken together make up the **population.**

2. Is a sample random if, and only if, it is representative?

No, the statement: 'A sample is random IF AND ONLY IF the sample is representative?' is not correct.

Consider this. Here are the spoken languages in the USA

English 79.2%
Spanish 13.0%
Other Indo-European 3.7%
Asian and Pacific island 3.3%
Other 0.8%

If I contact 20 of my friends -> 16 English speaking, 2 Spanish speaking, 1 Asian speaking, 1 'Other Indo-European'. This is more or less a representative sample. However, it is definitely not random, because all of those 20 people were my friends. Thus, “representative” does not imply “random”.

You can claim that: a sample is not representative; therefore, it is not random.

That would be true, but the general idea is different - the biggest sampling issues come from the fact that people can take representative samples in a NON-random manner (intentionally to prove some wrong statement). See example with languages above.

If a sample is not representative, it may be because of lack of knowledge on the person sampling it, causing it to be non-random, too (unintentionally).

3. Does the weight of a person have a true 0?

What is meant is 'weight' as a variable, not 'weight of a person'. Sorry about that.

As you know the weight of nothingness is 0, so weight can be 0.

Weight of a person, *should* have the same properties as weight, right?

The fact that it is **not achievable** does not mean it is not a **true zero.** Unlike the temperature example where 0 degrees Celsius is a deceiving 0.

4. What is my null hypothesis?

In statistics, the null hypothesis is the statement **we are trying to reject**. We define it that way, so we can **reject** it with a statistical test.

Conventionally, we include the equality sign in the null hypothesis.

5. How can I create a histogram in Microsoft Excel?

For a thorough walkthrough for creating a histogram refer to the course notes for Descriptive Statistics (Section 10, Lecture 26 in the Course Content tab).

For your convenience here's a link to that lecture: https://www.udemy.com/the-data-science-course-complete-data-science-bootcamp/learn/v4/t/lecture/10764384

6. When do I use the one-tailed test instead of the two-tailed test?

We use the one-tailed test when the null-hypothesis is some *inequality*. In that case, only values on one side of the null can reject it.

For instance, if the null hypothesis states that "a" is greater than "b", then we can only reject if "a - b" is significantly lower than 0.

If instead, our null stated the two are equal, then we would have an *equation* that looks like "a - b = 0". We can reject this hypothesis if "a – b" is significantly *greater* **or** *significantly* lower than 0. In this case, we have two possible ways of rejecting the null (if it is significantly 1) lower or 2) higher than the hypothesized value), which is why we need a "two-tailed" test.

7. I can't get my head around Statistics – Practical Example: Inferential Statistics.

To make this as easy to follow as possible, we are going to examine one of the exercises from the course. Go to the exercises available in Section 14, Lecture 74 and examine task number 3: https://www.udemy.com/the-data-science-course-complete-data-science-bootcamp/learn/v4/t/lecture/10764532

We use the same logic as we did in the lecture on confidence intervals for difference of two means (independent samples, population variances assumed to be equal) -> Section 13, Lecture 70 (https://www.udemy.com/the-data-science-course-complete-data-science-bootcamp/learn/v4/t/lecture/10764516?start=180)

Back in Lecture 70, we got the result that the confidence interval for the difference between apple prices in NY and apple prices in LA **is positive.**

We made the conclusion that with 95% confidence, the difference of the means is between 0.47 and 0.92. In any case, **it is positive.** If the average apple prices in NY **MINUS** the average apple prices in LA is positive, then the apples in NY **are more expensive.**

Using the same logic, we construct the confidence intervals in the practical example.

1. If the confidence interval for the difference is strictly positive, then the first shop is outperforming the second.

2. If the confidence interval for the difference is strictly negative, then the second shop is outperforming the first.

3. If the confidence interval for the difference is from negative to positive (e.g. (-0.5, 0.5)), then we are not sure if one shop outperforms the other, so we cannot make a statement based on confidence intervals.

8. Why is the Central Limit Theorem important?

The main takeaway from the CLT is that no matter the underlying distribution, we know that the distribution of the **sample means** is normal (given a big enough sample).

This means that we can take 1 sample (a **SINGLE ONE!)** and say:

Imagine it was taken from the sampling distribution of the sample mean. It is normally distributed for sure, right? Because of the CLT.

So, when we are testing, we can use this insight and treat the sample mean as if it is coming from a normal distribution (because it is -> the sampling distribution of sample means). And since all our hypothesis tests depend on normality, this proves extremely useful.

9. How do I manually compute the P-value for T-statistics?

Unlike Z-scores, with T-statistics, we do not have a singular table of values, which matches T-scores with their associated P-values.

This comes from the fact that every T-statistic requires an associated degree of freedom. This means that if we want to have an associated P-value for any T-score with any degree of freedom, we would require a 3-dimentional “table”.

For a detailed discussion on the topic, please check this question: https://www.udemy.com/the-data-science-course-complete-data-science-bootcamp/learn/v4/questions/4896738

Thus, convention dictates that we have a single table, which includes numerous degrees of freedom, but only has data regarding the most commonly used P-values.

In short, the information you are looking for **is NOT in the t-table**, because it was intentionally cut out.

10. What does a p-value of 0.000 mean for a coefficient of a regressor?

In Hypothesis testing we always get a p-value associated with every single coefficient. The p-value is a measurement of variable significance. It indicates the likelihood of the term **NOT** being significantly different from 0.

A p-value of 0.000 suggests that the coefficient is significantly different from 0 (it has explanatory power).

11. When a model is non-significant, is it wrong?

A model that is non-significant is a model that has **no explanatory power whatsoever.**

For instance, say you are wondering if it is going to rain tomorrow or it is going to be sunny. There are ways to predict that and weather forecasting agencies are using them.

Now, I tell you: **here's my model**... I flip a coin. If it is heads it is going to be sunny, if it is tails, it is going to rain.

I may be correct sometimes, but my model is non-significant. Whenever I am correct, it is pure coincidence rather than 'the fruit of a good model'.

In essence, we prefer saying: non-significant, because especially in a situation with only 2 outcomes, the model will inevitably be correct sometimes. Thus, it doesn't make sense to call it: **wrong**. It is just not relevant or not significant.

12. I have all my Python notebooks and data files in the same directory but cannot load them. What should I do?

This is an issue that occurs due to your specific Anaconda installation. The easiest way to solve this is to reinstall Anaconda. However, we recommend that you use the **absolute path** of a file when loading the data.

Thus, you can write:

data = pd.read_csv('ABSOLUTE_PATH/real_esate_price_size.csv')

To me this looks like:

data = pd.read_csv('C:/Users/365/Desktop/The Data Science Course 2018 – All Resources/Part_4_Advanced_Statistical_Methods_(Machine_Learning)/S27_L142/real_estate_price_size.csv')

In your case, you can find that by opening the folder containing the files and copy-pasting the path.

Once you copy-paste this path, you should **CHANGE ALL SLASHES from backward to forward.** This means that if your path is C:\Users\YOUR_COMPUTER_NAME\Desktop\... , you should make it look like: C:/Users/YOUR_COMPUTER_NAME/Desktop/…

Note that, due to the standards for referencing paths, in Python, instead of a backward slash (\ ), you should use a forward slash (/).

13. I cannot install TensorFlow.

You can check this lecture: https://www.udemy.com/the-data-science-course-complete-data-science-bootcamp/learn/v4/t/lecture/11258934

Or read further.

There is a compatibility issue between TF and Python 3.7. Until TF 2.0 is released, it may take you more time and effort to install TensorFlow. If you are experiencing an issue, here is how to downgrade to Python 3.6 so you can use it:

First, open Anaconda prompt as an administrator.

Then create a new environment and install python 3.6 there.

conda create --name py36 python=3.6

Activate (enter) the new environment you just created

conda activate py36

Then install tensorflow in that environment

conda install tensorflow

Since you may not have some of the packages that you had previously installed, please install them (I believe matplotlib is the only one of them):

conda install matplotlib

You will now have that environment; however, we must make it accessible in Jupyter. One of the easier ways is through the dedicated package, so you'll need to install it.

First install the ipykernel (in case you don't have it)

conda install ipykernel

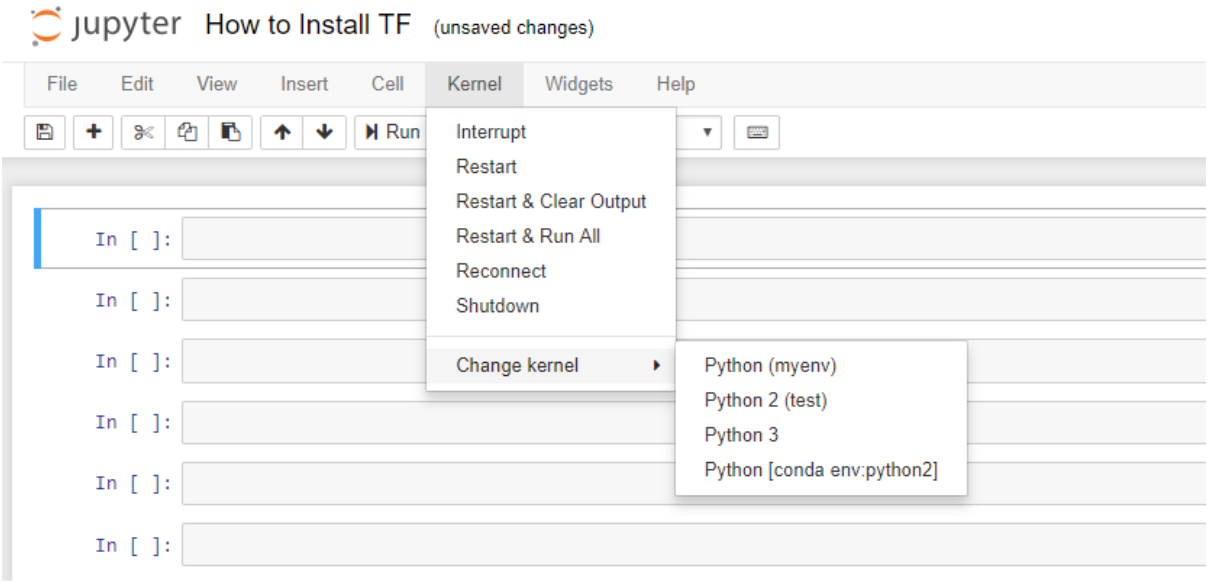
Then install the dedicated package

conda install nb_conda_kernels

After this is done, restart close Anaconda prompt and Jupyter. Then start Jupyter again. You should be able to choose your kernel. Needless to say, select the one in which you've installed TensorFlow.

Kernel -> Change kernel -> *One with TF*

Here's a screenshot of my view:



Hope this helps!

Best,
The 365 Team