CM2105 - Frequently Asked Questions

[Q&A] Week 1 – Code – Remove index name in Pandas

Q: del score.index.name does not work with my Pandas.

A: This might be due to the installed version of Pandas, which does not support his operation. You should be able to find alternative ways to achieve the required output. See the link below:

https://stackoverflow.com/questions/29765548/remove-index-name-in-pandas

Important Note: In the future labs or coursework, as long as you can work out the required output, you are free to use any methods necessary (within the required libraries specified for the assignment) in your programme. The Python programming skills are not the learning outcomes of the module, so will NOT be assessed. This means, for example, the efficiency of code will not be assessed. For the coursework, as long as your submitted code is executable and error free (with your own installation of Jupyter Notebook and versions of libraries), marks will be awarded based on the results/output of the code.

[Q&A] Distribution of sample means

Q: When plotting sample means, how do you plot is using the normal distribution? And when gathering data, I imagine that if you want to get another mean, you would get a new sample from the population?

A: When plotting the distribution of sample means, the following assumptions are made: (1) the distribution is normal (bell curve); (2) the mean (**μ**) of the distribution is assumed to be the mean of the current sample; (3) the standard deviation (**σ**) of the distribution is assumed to be equal to the standard deviation of the current sample divided by the square root of the sample size.

Once you have worked out **μ and σ**, your bell curve can be exactly determined by a mathematical formula. But many software packages can help plot a normal distribution, try this:

https://homepage.divms.uiowa.edu/~mbognar/applets/normal.html

The answer to the second question: if you were to get a new sample from the population, you would get a new mean. So, the mean is essentially a variable, and the confidence interval reflects such variation of all possible means (of samples) you would get from the population.

[Q&A] Standard deviation – Python code of Labs and Coursework

Q: While I'm working out the standard deviation, I found that using the std() function in numpy and pandas library have returned different results. Which one should we use in our Labs and Coursework or which one do you prefer?

A: As you can find in the handout "Descriptive statistics" (available on Learning Central), there are two different formulas for variance/standard deviation, depending on whether the data is being considered a population of its own, or the data is a sample representing a larger population. There are many online resources to explain this (https://www.khanacademy.org/math/statistics-probability/summarizing-quantitative-data/variance-standard-deviation-sample/a/population-and-sample-standard-deviation-review).

In Jupyter programme, the Numpy function calculates the population standard deviation by default (N in the denominator); and Pandas calculates the sample standard deviation by default (N-1 in the denominator).

For example, given a DataFrame (df) of your data:

Numpy:
Population standard deviation: std(df)
Sample standard deviation: std(df, ddof=1)

Pandas:
Population standard deviation: df.std(ddof=0)
Sample standard deviation: df.std()

For the purpose of consistency, I suggest that everybody uses "Population standard deviation" in the Labs and Coursework whenever standard deviation is used.

[Q&A] Coursework

1. **[Q]** As you have shown in the sample table 1 and 2, there are only 6 rows displayed. I wanted to ask whether this should be followed exactly, by using .head(6), or are we supposed to just display the table as it is.

[A] "Sample output" only shows a part of required output. Display the returned table as it is.

2. **[Q]** Indicated in the Pro-Forma, we are supposed to out put the mean, min value, max value and standard deviation rounded to two decimal places. I wanted to ask if we have to force the program to out two zeros at the end of a whole number or is the default output fine? e.g. 100 - rounded would give us 100.0, is this fine or do we need, 100.00.

[A] In this case, both 100.0 and 100.00 are correct.

3. **[Q]** For the Confidence Interval I wanted to ask whether this is supposed to be done manually as done in the lab i.e. err=1.96*std(dp1)/sqrt(len(dp1)) or are we allowed to use a prebuilt function?

[A] Yes, any functions within the specified libraries are allowed for the coursework. The "CW_your student number.ipynb" file specifies the libraries allowed for the coursework.

To make this clear, the allowed libraries: matplotlib and numpy (imported by %pylab), pandas, scipy, and statsmodels.api.

4. **[Q]** Further, I wanted to ask, for the graphing in question 4 are we supposed to use a specific library or is it allowed for us to do it what ever way we please. If there is a specific library you would like us to use, could you please name it.

[A] The "CW_your student number.ipynb" file specifies the libraries allowed for the coursework.

5. **[Q]** In addition to all this, are we allowed to make more imports to make our program better or do we have to only use the ones already given.

[A] Yes, you can make more imports, as long as they are from the specified libraries allowed for the coursework. The "CW_your student number.ipynb" file specifies the libraries allowed for the coursework.

6. **[Q]** Finally, is it possible if you could provide us with the sample outputs for the other questions as well like you did for 1 and 2, in order to make the requirements clearer and more easy to understand.

[A] There are only sample outputs for cell1 and cell2. I am happy to clarify the stated requirements if there are any further queries.

[Q&A] Coursework

1.Q: For the question "print the 95% confidence interval of academic reputation " mean , is it the confidence level of what exactly ? Or does it simply mean respective quantiles?

A: As stated in the question, this refers to the **95% confidence interval** of the "**data contained in the variable called "Academic Reputation" for the UK universities in the world's top 200**".

2. Q: For part 2 , what package or libraries should we use " image quality prediction" stats models" scimitar-learn?

A: **The "CW_your student number.ipynb" file (see cell1) specifies the libraries allowed for the coursework. To make this clear, ONLY the following libraries are allowed: matplotlib and numpy (imported by %pylab), pandas, scipy, and statsmodels.api.**
NOTE: teaching is on-going, the image data processing will be covered in the next two weeks.

[Q&A] Coursework

Q: In the regards to coursework of CM2105 (Data Processing and Visualisation) in Q1 part 4 it asks to compare the 'Citations per Faculty' of universities in the United States, United Kingdom, Canada, Australia, Netherlands and Germany. Do we need to get the mean of 'citations per faculty' for each country or do we need to use all the values of 'citations per faculty' for each country? If possible, Is there a sample output we can receive for the graph?

A: Please be aware this is a piece of assessed coursework, so we are not in a position to provide feedback on your attempts or methods before the submission deadline. But, we are happy to help clarify the stated requirements.

Note: The published/moderated marking scheme only allows sample output for cell1 and cell2.

[Coursework] Read and display grayscale images

Use functions (links) from matplotlib:
imread and imshow
Sample code:

```
# read and display a grayscale image
%pylab inline
img=imread('m1.png')*255
figure(figsize(16, 4))
subplot(1,3,1)
imshow(img,cmap='gray')
title('Query image')
# get information of image size
row = size(img, 0)
col = size(img, 1)
px=row*col
```

[Q&A] Coursework

Q: on question 5 of the coursework it says to use a chosen test to show the difference between the two countries, I want to use a histogram to do so, but by 'tests' does that mean I can't use a histogram, I have to use something like t-test for example.

A: The question assesses the understanding of the tests for statistical significance (i.e., for comparing means).

[Q&A] Coursework

Q: While plotting 95% Confidence interval error bars can we use yerr = 2*std? Or 2*standard deviation to obtain 95% Confidence Interval

A: Please be aware this is a piece of assessed coursework, so we are not in a position to provide feedback on your attempts or methods before the submission deadline. But, we are happy to help clarify the stated requirements.

[Q&A] Coursework

Q1: For Q6, it does not say which two photos we have to derive the correlation for. For example, m1 and m5. Do you mean we have to derive the correlation of each image?

A1: The question states the requirement: correlation between the **two variables** “average pixel intensity” and “image quality”, for the **given dataset** “model.zip” of **ten images**.

Q2: I have a question regarding Q1.Part 2. In question 7 we are asked to print the mean squared error (MSE) of the model in the "model" dataset. In lecture 10: Regression - applications, we are shown how to calculate the mean absolute error (MAE) , would it be correct to assume this method is also the correct way to get the mean squared error (MSE)?

A2: In terms of evaluating a model’s performance, the principle is the same: analysing the errors the model is making when predicting the target variable. There are few popular error measures, including MAE and MSE. Their mathematical formulations are slightly different:

MAE: https://en.wikipedia.org/wiki/Mean_absolute_error

MSE: https://en.wikipedia.org/wiki/Mean_squared_error