IMPERIAL

Probability and Statistics

Coursework 2024

WE ARE
THE DEPARTMENT OF
BIOENGINEERING

CONTENTS

PREAMBLE

Introduction

The aim of this coursework is to get you to process some ‘real’ data and think about the results. The data are actually randomly generated using MATLAB (as we have seen in class for the Monte Carlo techniques), so you will each have different results.

The coursework involves 3 parts that will be marked

- Numerical answers and MCQ conclusions for your specific data set
- Written answers on your selection of tests and reflections about your observations
- Figures, representing and communicating your data as effectively as possible

An m-file of your code will also be uploaded, but not marked.

There are four parts to the coursework.

Questions 1-5

Numeric answers will be entered into Numbers_Template.xlsx.

- These include multiple choice, true/false and numeric answers, such as measures of location, test statistics, p-values, percentages etc.
- You will enter the values into the Numbers Template
- Instructions to enter data into the Numbers Template are given in purple.
- A MATLAB algorithm will mark these and add the correct answers and marks to your excel spreadsheet for each answer
- A number of significant figures or decimal places have been specified (3 decimal places, 1 decimal place, 3 significant figures or integers).
    - In many cases, this is more precise than is strictly appropriate for the data, but is so I can check that you have done the correct statistical process
    - The MATLAB command `round` can be used to round to a specified number of decimal places or significant figures.
    - The spreadsheet is deliberately formatted to display 6 decimal places to help you make sure you have rounded correctly. E.g. a p-value of 0.05129 to 3.s.f will be 0.0513, and on the spreadsheet will look like 0.051300. Don't worry about the trailing zeros.
    - Be careful reading values from the Command Window in MATLAB, which only shows four decimal places by default. Check in the workspace, or type `format long` into the Command Window to see more decimal places
    - To avoid rounding errors, only round your values when you put them into the spreadsheet
    - If you enter a value on excel that is less than 10^-6, it may appear as all zeros. If the information is correct, it will still be held in the cell. In excel, the bar under the ribbon shows the value in a given cell. As long as this is correct, the information is there and it will be ok.
    - If you do not properly enter your data to the specified precision, your answers will be numerically wrong and you will lose marks
- Questions 1, 2 and 4 will be possible after Lecture 7. Questions 3 and 5 will be possible after lecture 8.

Questions 6 & 7

- Text answers will be entered into Words and Figures_Template.docx
    - A word count is given for each question. Do not exceed the word count, or you will not get marks for that part of the question
    - Remember, there is a lot of bad statistical practice on the internet (like bar charts and over-interpretation of p-values) as well as various ways of doing things that may not align with how I have taught you in this course. So if you choose to use AI, remember that it may not have learned things the way I have taught you and it's answers may therefore not get marks. Nonetheless, it can sometimes be helpful for understanding. If you do choose to use AI and enter resulting text as your answer, make sure to reference it, per the departmental guidelines – failure to do so would be plagiarism.
    - Please leave in the font I have provided :)

Figures 1 & 2

- Figures and their corresponding captions will be inserted into Words and Figures_Template.docx
    - These are a chance to try out ways to visualise your data
    - Create figures in MATLAB, use export (manually or using `exportgraphics`) at 300 dpi or greater and enter into the template. Low quality figures will lose marks.
    - Marks will be given for completeness, clarity and style

MATLAB code

- <u>This will not be marked, but marks will be docked if it is not uploaded</u>
- This is so we can check for plagiarism if necessary, but also in some circumstances allows me to check where errors have occurred and potentially award working marks. This has historically made quite a difference in a few cases
- You may use in-built MATLAB functions or code your own. If the former, be careful to make sure inputs are as discussed in class
- For the Shapiro-Wilk test, download `swtest` from the Blackboard folder and place in your main MATLAB directory or your main working folder

Statistical points

- Use $\alpha = 0.05$ unless otherwise specified
- With your word answers, try to be precise and specific.

Plots

- Colours are your choice, but try to choose clear colours (and ideally not default ones).
- For annotations, you can save a figure (in hi resolution), move into power point and export.
- Some functions that might be useful (you can find examples within the Lecture Code files)
    - `h=scatter(x,y,…)` - creates a scatter plot that is amenable to transparency
    - `h=plot(x,y,…)` - good for plotting lines. Use ‘linestyle’ and ’color’ for formatting
    - `h=fill(x,y,…` - for filled areas, requires all four corners of the filled area to be specified
    - `alpha(h,level)` - sets transparency level of object referenced by handle h
    - `xlabel/ylabel` - sets axis labels
    - `xticks/yticks` - sets the values of the ticks on the axes
    - `xticklabels/yticklabels` - sets the tick labels (e.g. if you want words instead of numbers)
    - `xtickformat/ytickformat` - sets the precision of the tick labels. E.g. if you had x-ticks at 0.1, 0.15, 0.2, 0.25 this is varying precision. The command `xtickformat(‘%.2f’)` would set the precision to 2 decimal places, so you would get 0.10, 0.15, 0.20, 0.25, which is much better!
    - `axis ([xl xu yl yu])` - sets the axis limits
    - `xlim([xl xu])` and `ylim([yl yu])`are alternatives to axis
    - `fontsize(h,size,units)` - sets fontsize for the object referenced by handle h

Captions

- Provide captions for your figures
    - You do not need to describe the interpretation of the results in the caption
    - Rather, succinctly clarify what each feature on the plot represents
    - Legends on the figures can also be used, but keep in mind the clarity of the plot.

Logistics

The logistics of this coursework, providing an individual data set and multiple types of feedback to ~350 students are a little awkward. Figure 1 explains the process. Please do try to follow the naming conventions.

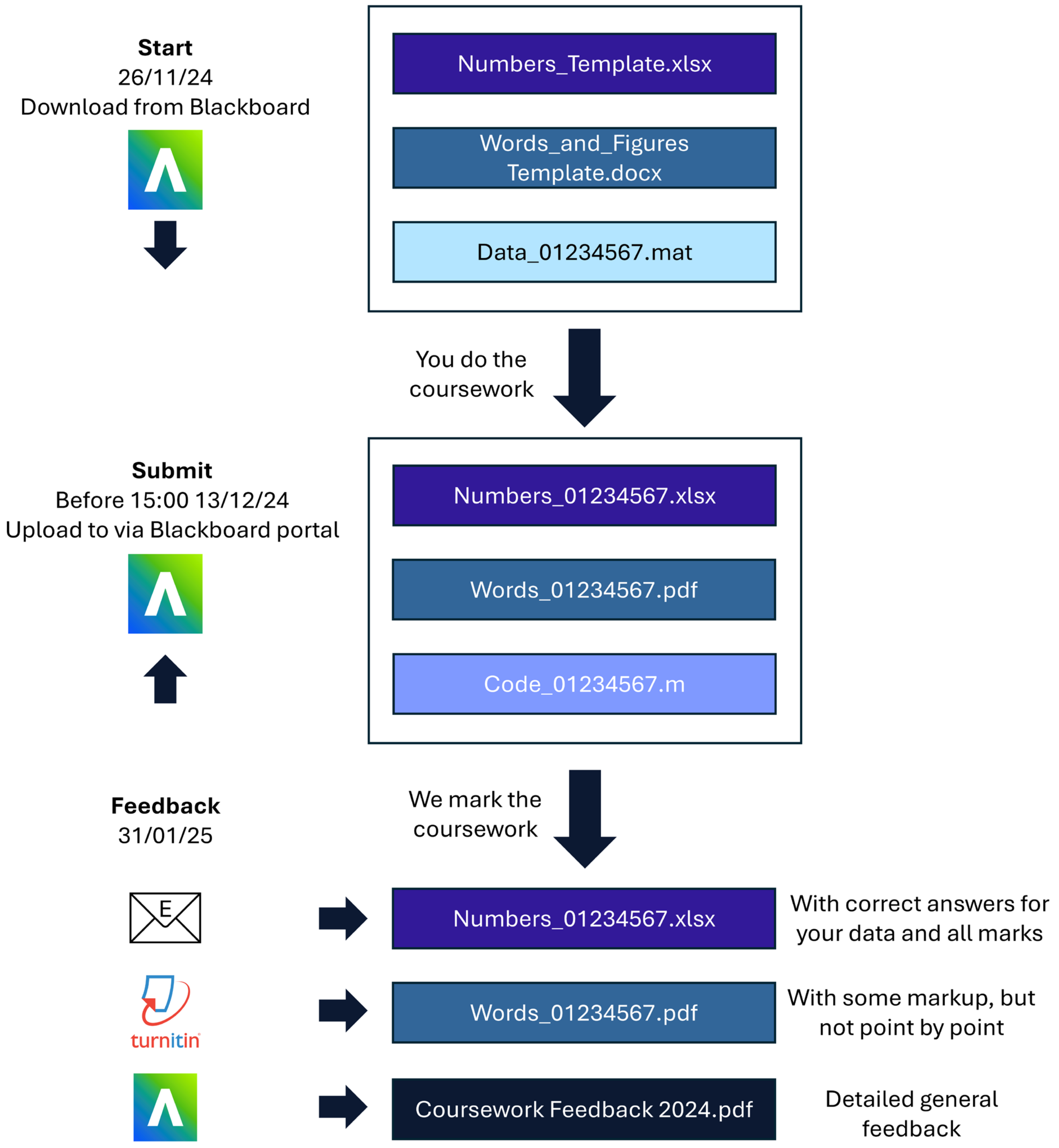


Figure 1 – Submission and Feedback details.

Each of you has your own .mat file that contains your data, against which your submission will be marked.

Start – download from Blackboard

1. the data with your CID number from the Sharepoint folder linked on Blackboard. For example, if your CID number was 1234567, then your data will be in Data_1234567.mat.
2. Words and Figures_Template.docx
3. Numbers_Template.xlsx

Submit – Upload via Blackboard portal. When you have finished doing the coursework

1. Save your Words and Figures file as pdf with your CID in the form: 'Words_[your CID].pdf', e.g. Words_1234567.pdf
2. Save your Numbers Template file with the name 'Numbers_[YourCID].xlsx'

e.g. Numbers_1234567.xlsx

3. Save your MATLAB code with the name 'Code_[your CID].m', e.g. Code_1234567 (note not '.mat' which is a data file, but '.m')
4. **To submit** your coursework, submit all three files (.pdf, .xlsx, .m) via the Blackboard Assignments portal.
5. Remember to use the Ed board with the tag 'Coursework' if you have questions between lectures and study groups

Feedback

1. You will receive by email your submitted answers, along with the correct answers. Where common errors have been made, I can identify these and have given partial marks and a comment.
2. Your Words and Figures will be marked up but without explicit feedback. This should give you a general idea of mark distribution, but will not give you explicit details on every question.
3. The Coursework Feedback document will be shared on Blackboard after we release the marks. Use this, along with the other two documents to evaluate if there are places where you may wish to improve your understanding.

BACKGROUND

In this coursework, we will investigate the question of how different sources of nitrogen affect the growth of bacteria. The purpose of this is to optimise the output of protein-generating bacteria.

Bacterial growth can be measured by quantifying the optical density (OD), which is related to the amount of light blocked by the bacteria present. The measurement principle is that when there are more bacteria, less light will pass through and the OD will be higher. We start with a baseline number of bacteria, which should yield an optical density of 0.35. We then measure it again 180 minutes after the start of the experiment and use this to estimate the rate of growth (a higher OD means more bacteria)

The two nitrogenous sources of interest are arginine (A) and urea (U). To investigate which is more effective at promoting bacterial growth, you will analyse the data from a number of different experimental paradigms.

***Disclaimer**: although this question is inspired by the optimisation described above, priority has been given to evaluating your statistical skills rather than being reliable observations of the true behaviour of bacteria.*

NUMERIC QUESTIONS

To be completed in the Numbers Template

Question 1

In the first round of experiments, you want to investigate whether the test is working (using arginine), i.e. whether arginine is increasing the OD compared to the baseline value of 0.35. If it is not greater, but less, you wouldn't care, as either way would require you to troubleshoot.

You acquired OD values from N=16 observations using arginine (`A1`).

You can assume for this question that the data are sampled from normal distributions.

a) Is this statement true or false? Enter 1 (true) or 0 (false) into the Numbers Template.
"As the sample size is 16, it is appropriate to use a Z-test for this analysis"

b) Which of these statements is the most accurate? Enter the appropriate number into the Numbers Template.
1. "As we only care about an increase in the OD, a lower tailed test is appropriate"
2. "As we only care about an increase in the OD, an upper tailed test is appropriate"
3. "As we only care about an increase in the OD, a two-tailed test is appropriate"

c) Select the appropriate statistical test, calculate the test statistic and p-value. Enter the test statistic (to 3 decimal places) and p-value (to 3 significant figures) into the Numbers Template.

d) Would you reject the null hypothesis? Enter 0 (do not reject) or 1 (reject) into the Numbers Template.

Question 2

Depending on your data from Part 1, you were either satisfied that the test was working, or were not satisfied, and realised that it was an incorrect illumination setting, which you fixed. You now proceed to do an experiment with N=16 observations for urea (`U2`) and N=16 observations for arginine (`A2`). You want to know if one of the nitrogen sources is better than the other.

You can assume for this question that the data are sampled from normal distributions.

a) Find the difference between the sample means (Urea minus Arginine), $\bar{Y}$, the number of degrees of freedom on this value, and the corresponding upper and lower limits of its 95% confidence interval. Enter these values to 3 decimal places into the Numbers Template.

b) Select the appropriate statistical test, calculate the test statistic and p-value. Enter the test statistic (to 3 decimal places) and p-value (to 3 significant figures, but do not worry about trailing zeros) into the Numbers Template.

c) Would you reject the null hypothesis? Enter 0 (do not reject) or 1 (reject) into the Numbers Template.

Question 3

Your supervisor asks you to take a look at some data a previous student in the lab had carried out. They used a similar approach and also have N_U=12 values for urea (`U3`) and N_A=11 arginine (`A3`), as one observation failed. You still want to know if one of the nitrogen sources is better than the other.

Do not assume anything about the underlying population distributions for this question.

a) In order to identify the proper statistical test, we must evaluate whether the data can reasonably be assumed to have been sampled from a normal distribution. Carry out Shapiro-Wilk tests on these data. Enter the *p*-values to 3 significant figures into the Numbers Template

b) Is this statement true or false? Enter 1 (true) or 0 (false) into the Numbers Template.

"If the null hypothesis of an underlying normal distribution is rejected for either sample, then a t-test is not appropriate"

c) Check your data set for outliers, identifying them as data points that are more than 3 median absolute deviations away from the median (using k=1.4826). Enter the number of outliers for each data set as an integer into the Numbers Template.

d) Remove any outliers from your data and re-run the Shapiro-Wilk test for both data sets. Enter the *p*-values to 3 significant figures into the Numbers Template.

e) Select the appropriate statistical test, calculate the test statistic and p-value. Enter the test statistic (as an integer) and p-value (to 3 significant figures) into the Numbers Template. Make sure the test statistic is the one we use in this course, not just what MATLAB outputs.

f) Would you reject the null hypothesis? Enter 0 (do not reject) or 1 (reject) into the Numbers Template.

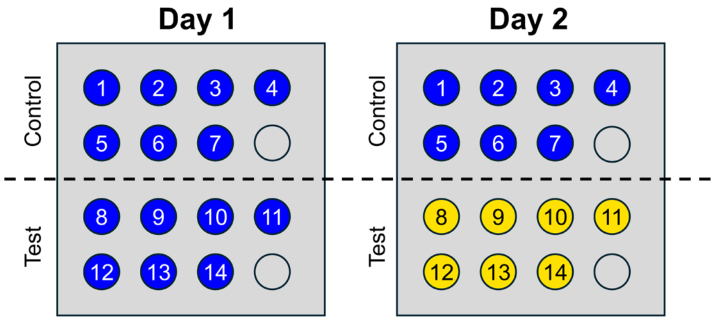
Question 4

You have read in a reputable journal that if you use arginine on day one

1) adding additional arginine the second day has no effect
2) adding urea on the second day will further increase the amount of bacteria.

You can assume for the purpose of this sub-question that this is true and there are no outliers

You decide to invite the class to help you out with an experiment to see if you can reproduce the result. The figure below visualises the experimental design.



Each student has a plate with 14 filled wells in it. On day one, each student treats all 14 wells with arginine and measures the OD. On day two, wells 1-7 are treated with arginine again (control), while wells 8-14 are treated with urea (test).

This data set will allow us to investigate the concepts of statistical power and statistical difference. The data for this question can be found in arrays `C1`, `C2`, `T1` and `T2` (C1=control, day 1 etc.). Note that for each variable there are 350 columns, with each column representing one student.

You can assume for this question that the data are sampled from normal distributions.

a) Is this statement true or false? Enter 1 (true) or 0 (false) into the Numbers Template.

"As the measurements on day two are on the same wells as day one, the data can be considered to be paired"

b) Control Data: for each student's data, use an appropriate statistical test to investigate whether there is a statistical difference between the control samples on days one and two. How many of the *p*-values would lead you to reject the null hypothesis? Enter the percentage (1 d.p.) of rejections of the null hypothesis into the Numbers Template.

c) Test Data: for each student's data, use an appropriate statistical test to investigate whether there is a statistical difference between the test samples on days one and two. How many of the *p*-values would lead you not to reject the null hypothesis? Enter the percentage (1 d.p.) of non-rejections into the Numbers Template.

d) Test Data: using MATLAB calculate the expected statistical power of a given future test with 8 wells. To find appropriate estimates of μ_1 and s, use all the well-by-well differences from the test data (note `X(:)` converts X into an n-by-1 array). Enter the expected power as a percentage (1 d.p.) into the Numbers Template.

e) Using your estimated of μ_1 and s from part (d) calculate how many more wells per sample you might need to obtain a power of 90% with α=0.01 than you would with α=0.05? Enter the number into the Numbers Template.

Question 5

Finally, you wish to evaluate how effective an antibiotic is at different concentrations. Concentrations of 0, 2, 4, 6, 8 and 10 μg/ml are made up and optical density is measured 3 times for each concentration after using Urea to enhance growth for 2 days. The data can be found in `Conc` and `OD`.

a) Calculate the sample correlation coefficient and evaluate whether it is statistically different from zero. Enter the *p*-value and to 3 significant figures into the Numbers Template

b) Calculate the coefficient of determination. Round to 3 decimal places and enter into the Numbers Template

c) Carry out a regression analysis on the data, calculating the slope and the 95% confidence interval on the slope. Enter these values to 3 decimal places into the Numbers Template.

d) Calculate the prediction bands and use them to estimate the range of expected values (prediction interval) of OD at a concentration of 2 μg/ml of the antibiotic. Enter the upper and lower bounds of this range into the Numbers Template.

TEXT QUESTIONS

To be completed in the Words and Figures template

Question 6 – test selection

In the numerical questions, you selected various statistical tests. For each, explain why you selected the one you did and what the null hypothesis is in words. Be as precise as possible.

a) For Question 1c (< 60 words).
b) For Question 2b (< 60 words).
c) For Question 3e (< 60 words).
d) For Questions 4b and 4c (< 60 words).

Question 7 - reflections

a) In Question 2, you carried out a null hypothesis test, from which you may or may not have rejected the null hypothesis. Irrespective of your personal result, what could you have concluded from the data if the null hypothesis was not rejected. Be as precise and comprehensive as possible (< 60 words)?
b) Reflect on your answers to 4b. How do they compare to what you expected? (< 100 words).
c) Reflect on your answers to 4c and 4d. How do they compare to what you expected? (< 100 words).
d) Reflect on your answers to Question 5. What do the statistics you calculated tell you about the real-world application? (< 100 words).

FIGURES

See preamble for marking criteria for plots and captions.

Figure 1

Plot the data for Question 2 in a way that you think most clearly communicates the data and the statistical analysis you carried out.

Figure 2

Plot the *p*-values from Questions 4b and 4c on a single graph. Use a log-scale for the p-values. Add annotations, colours etc. to tell the story of your data.

Figure 3

Plot your data for Question 5 and visualise the outcomes of your regression analysis. You may wish to include visualisation of the values calculated in part d.