CA1_2025

September 17, 2025

Data Science - Coursework 1 (35%)

Short Style Data Science Questions - these will be answered and submitted during the Week 6 Data Science Workshop

Before starting:

- **Read the Instruction document in Learning Central > Assessment and Feedback > CA 1.**
- **Put away your phone.**
- Open your weekly jupyter notebooks and completed yourturn notebooks to use.

Plagiarism

Please take note of the University's policy on plagiarism, which is outlined in your student handbook.

Plagiarism is the act of passing off the words or ideas of others as if your own. Advice on avoiding plagiarism is given in the UG Student Handbook. There is also considerable help and advice on Learning Central and the University web site. Students need to be especially careful of plagiarism in computing tasks and you are advised not to share code through electronic means. Students working together during their weekly exercises and the coursework is great (and indeed encouraged) but you need to ensure that they you are not using each other's code or text.

Use of genAI

We will follow the University's traffic light system. This coursework is designated AMBER — you may use GenAI tools only for the computational aspect (e.g., coding cells). GenAI must not be used to write text, discussions, or interpretations. You may, however, use support tools such as dictionaries, thesauruses, or grammar/spell checkers (e.g., Grammarly), even if they are powered by GenAI. Any use of AI must be declared, as outlined in Learning Central > Assessment and Feedback.

QUESTION 1

A student is taking three modules, and the probability that they pass any individual module depends on the fraction of the weekly live sessions they attend f so that the probability of passing

the module is $0.82f$.

a.) If the student attends all the weekly live sessions, ($f = 1$), calculate the probability that they will

(i) pass all three modules,
(ii) fail one module and pass the other two
(iii) pass only one module,
(iv) pass no modules.

b.) Show that the sum of these four probabilities is 1.

c.) Calculate the four probabilities on the assumption that the student attends only half of the weekly live sessions.

d.) What fraction of weekly live sessions must the student attend to have a 50% chance of passing all three modules?

e.) Is this a realistic way to model the probability that the student will pass the modules?

[10 marks]

QUESTION 2

A group researching cancer have previously found that the genetic marker D3 is a useful indication that a person will develop the more aggressive form of melanoma skin cancer, in that D3 is present in 65% of the aggressive cases. However the test is expensive. A rival group claim that the marker M23 is more sensitive than D3, and works out considerably cheaper to test for. The rival research team manage to get DNA samples from 7 patients with the aggressive form of the disease, all of whom test positive for the genetic marker M23. Based on these results, is M23 a better marker for the disease than D3?

Explain your approach, give full mathematical working for your reasoning, include code cells for calculations and show labeled plots of the underlying functions. Briefly discuss your result.

[20 marks]

QUESTION 3

The state of Florida is thinking of relaxing its policy on sales of ice cream to allow ice cream to be sold in ince cream vans, since the police predict that this can reduce violence. After some extensive polling, they find that only 30% and 12% of Republican and Independent voters are, respectively, behind the change in the law, while 80% of the Democrat voters are in favour. You are visiting the state, and ask a Police Officer what they think of the idea. They say they're against the change to the law. What is the probability that they vote Democrat?

You may assume that voting in the Florida polls that year was split in the following way: 40% Democrat and 38% Republican. You can also assume that Independent covers everything that is not Republican or Democrat.

Explain your approach, give mathematical working for your reasoning, include code cells for calculations and briefly discuss your result.

[25 marks]

QUESTION 4

A computer chip manufacturer suspects that roughly half of its latest batch of CPUs contains a flaw. The accounts department are clearly concerned, and are trying to predict how the fault will affect the number of customers returning products. How many CPUs from the batch would they need to examine to know the probability that any given CPU is faulty to better than 2.5%?

Tip: think carefully about what you are trying to estimate here, you want the *error in your success probability* to be less than 2.5%. You can solve this by thinking about what probability distribution is required and looking at the mean and standard deviations.

Explain your approach, give mathematical working for your reasoning, include code cells for calculations and briefly discuss your result.

[20 marks]

Convert your .ipynb file to a PDF and upload in Assessment and Feedback > CA 1