Problem Sheet 1

MATH1710 Probability and Statistics I

University of Leeds, 2023-24

*This is Problem Sheet 1, which covers material from Lectures 1 and 2 of the notes. You should work through all the questions on this problem sheet in advance of your tutorial in Week 2. Questions C1 and C2 are assessed questions, and are due in by **2pm on Monday 16 October**. I recommend spending about 4 hours on this problem sheet, plus 1 extra hour to neatly write up and submit your answers to the assessed questions.*

A: Short questions

*The first three questions are **short questions**, which are intended to be mostly not too difficult. Short questions usually follow directly from the material in the lectures. Here, you should clearly state your final answer, and give enough working-out (or a short written explanation) for it to be clear how you reached that answer. You can check your answers with the solutions-without-working at the bottom of this sheet; solutions-with-working will be available after Friday 13 October. If you get stuck on any of these questions, you might want to ask for guidance in your tutorial.*

A1. Consider again the "number of Skittles in each packet" data from Example 1.1.

$$59,\ 59,\ 59,\ 59,\ 60,\ 60,\ 60,\ 61,\ 62,\ 62,\ 62,\ 63,\ 63.$$

(a) Calculate the mean number of Skittles in each packet.

(b) Calculate the sample variance using the definitional formula.

(c) Calculate the sample variance using the computational formula.

(d) Out of (b) and (c), which calculation did you find easier, and why?

A2. Consider the following data sets of the age of elected politicians on a local council. (The "18–30" bin, for example, means from one's 18th birthday to the moment before one's 30th birthday, so lasts 12 years.)

Age (years)	Frequency	Relative frequency	Frequency density
18–30	1		
30–40	2		
40–45	4		
45–50	5		
50–60	6		
60–80	2		
Total	20	1	—

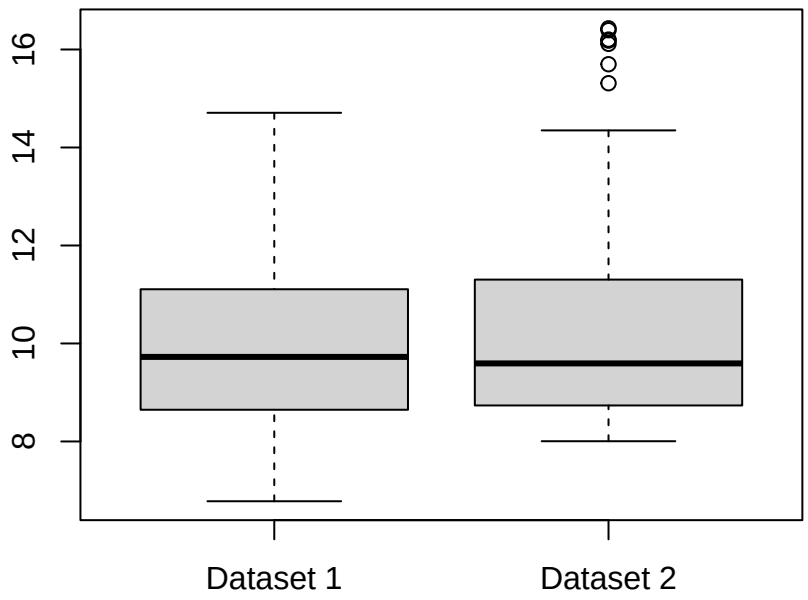
(a) Complete the table by filling in the relative frequency and frequency densities.

(b) What is the median age bin?

(c) What is the modal age bin?

(d) Calculate (the standard approximation of) the mean age of the politicians.

A3. Consider the two datasets illustrated by the boxplots below. Write down some differences between the two datasets.



B: Long questions

*The next four questions are **long questions**, which are intended to be harder. Long questions often require you to think originally for yourself, not just directly follow procedures from the notes. You may not be able to solve all of these questions, although you should make multiple attempts to do so. Here, your answers should be written in complete sentences, and you should carefully explain in words each step of your working. Your answers to these questions – not only their mathematical content, but also how to write good, clear solutions – are likely to be the main topic for discussion in your tutorial. Solutions will be available after Friday 13 October.*

B1. For each of the two datasets below, calculate the following summary statistics, or explain why it is not possible to do so: mode; median; mean; number of distinct outcomes; inter-quartile range; and sample variance.

(a) Shirt sizes for the $n = 16$ members of a university football squad:

Colour	Xtra Small	Small	Medium	Large	Xtra Large
Number of shirts	0	1	6	4	5

(b) Six packets of Skittles are opened together, a total of $n = 361$ sweets. The colours of these sweets is recorded as follows:

Colour	Red	Orange	Yellow	Green	Purple
Number of Skittles	67	71	87	74	62

B2. A summary statistic is informally said to be "robust" if it typically doesn't change much if a small number of outliers are introduced to a large dataset, or "sensitive" if it often changes a lot when a small number of outliers are introduced. Briefly discuss the robustness or sensitivity of the following summary statistics: **(a)** mode; **(b)** median; **(c)** mean; **(d)** number of distinct outcomes; **(e)** inter-quartile range; and **(f)** sample variance.

B3. Let $\mathbf{a} = (a_1, a_2, \dots a_n)$ and $\mathbf{b} = (b_1, b_2, \dots, b_n)$ be two real-valued vectors of the same length. Then the *Cauchy–Schwarz inequality* says that

$$\left(\sum_{i=1}^{n} a_i b_i\right)^2 \leq \left(\sum_{i=1}^{n} a_i^2\right)\left(\sum_{i=1}^{n} b_i^2\right).$$

(a) By making a clever choice of (a_i) and (b_i) in the Cauchy–Schwarz inequality, show that $s_{xy}^2 \leq s_x^2 s_y^2$.

(b) Hence, show that the correlation r_{xy} satisfies $-1 \leq r_{xy} \leq 1$.

B4. A researcher wishes to study the effect of mental health on academic achievement. The researcher will collect data on the mental health of a cohort of students by asking them to fill in a questionnaire, and will measure academic achievement via the students' scores on their university exams. Discuss some of the ethical issues associated with the collection, storage, and analysis of this data, and with the publication of the results of the analysis. Are there ways to mitigate these issues?

(It's not necessary to write an essay for this question – a few short bulletpoints will suffice. There may be an opportunity to discuss these issues in more detail in your tutorial.)

C: Assessed questions

*The last two questions are **assessed questions**. This means you will submit your answers, and your answers will be marked by your tutor. These two questions count for 3% of your final mark for this module. If you get stuck, your tutor may be willing to give you a small hint in your tutorial.*

*The deadline for submitting your solutions is **2pm on Monday 16 October** at the beginning of Week 3. Submission will be via Gradescope, which you can access via Minerva or on the Gradescope mobile app. You should submit your answers as a single PDF file. Most students choose to hand-write their work on paper, then scan-and-submit it to using the Gradescope app on their phone. Your work will be marked by your tutor and returned on Monday 23 October, when solutions will also be made available.*

Question C1 is a "short question", where brief explanations or working are sufficient; Question C2 is a "long question", where the marks are not only for mathematical accuracy but also for the clarity and completeness of your explanations.

You should not collaborate with others on the assessed questions: your answers must represent solely your own work. The University's rules on academic integrity – and the related punishments for violating them – apply to your work on the assessed questions.

C1. The monthly average exchange rate for US dollars into British pounds over a 12-month period was:

$$1.306,\ 1.301,\ 1.290,\ 1.266,\ 1.268,\ 1.302,$$
$$1.317,\ 1.304,\ 1.284,\ 1.268,\ 1.247,\ 1.215.$$

(a) Calculate the median for this data.

(b) Calculate the mean for this data.

(c) Calculate the sample variance for this data.

(d) Is the mode an appropriate summary statistic for this sort of data? Why/why not?

C2. **(a)** Suppose that a dataset $\mathbf{x} = (x_1, x_2, \ldots, x_n)$ (with $n \geq 2$) has sample variance $s_x^2 = 0$. Show that all the datapoints are in fact equal.

(b) Prove the following computational formula for the sample covariance:

$$s_{xy} = \frac{1}{n-1}\left(\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}\right).$$