1872 PRIFYSGOL ABERYSTWYTH UNIVERSITY

ADRAN MATHEMATEG / DEPARTMENT OF MATHEMATICS

ARHOLIADAU SEMESTER 2 / SEMESTER 2 EXAMINATIONS

MAI – MEHEFIN / MAY – JUNE 2024

MA26620 – Applied Statistics

The questions on this paper are written in English.

Amser a ganiateir - 2 awr **Time allowed - 2 hours**

- Arholiad llyfr agored: caniateir hyd at bum taflen o bapur A4 sy'n cynnwys nodiadau ysgrifenedig.
- Gellir rhoi cynnig ar bob cwestiwn.
- Rhoddir mwy o ystyriaeth i berfformiad yn rhan B wrth bennu marc dosbarth cyntaf.
- Cyfrifianellau Casio FX-83 neu FX-85 YN UNIG a ganiateir.
- Darperir tablau ystadegol.
- Mae modd i fyfyrwyr gyflwyno atebion i'r papur hwn naill ai yn y Gymraeg neu'r Saesneg.

- Open book examination: up to five sheets of A4 paper containing handwritten notes permitted.
- All questions may be attempted.
- Performance in section B will be given greater consideration in assigning a first class mark.
- Casio FX-83 or FX-85 calculators ONLY may be used.
- Statistical tables will be provided.
- Students may submit answers to this paper in either Welsh or English.

Section A

1. Classify the following variables as discrete, continuous or effectively continuous. For discrete variables, state whether they are of ordinal or nominal type. For continuous variables, state whether they are of ratio or interval type.

(a) Highest daytime temperatures (°C) of locations in Wales;

(b) Voters' orders of preference for candidates standing in an election;

(c) Total daily sales (£) in a supermarket;

(d) Names of species of sea slugs. [8 marks]

2. A mathematics lecturer uses software to display the screen of his tablet on the large screen at the front of the lecture theatre. Frustratingly, the software sometimes freezes and needs to be restarted. Over a long period of time, the software needs to be restarted an average of three times in every two hours of lecturing.

After a software update, the lecturer suspects the problem may have worsened. Test whether the rate of software freezing has significantly increased if 16 restarts are required in eight hours of lectures. In your answer, state clearly:

- the meaning of any notation you introduce as well as any assumptions made;
- which variable follows a Poisson distribution, defining its parameter;
- the two hypotheses;
- the distribution of the number of restarts required in eight hours of lectures when H_0 is true. Use tables to evaluate the p-value and state your conclusion. [10 marks]

3. In each of four compost mixes labelled A, B, C, and D, an equal number of sunflower seeds were planted. The heights of the resulting plants in mm were recorded after six weeks and the following ANOVA table computed:

Source	**SS**	**DF**	**MS**	**F-ratio**	**P**
Between composts				5	0.00589
Total (corr.)	141	35			

(a) How many sunflower seeds were planted in compost mix A? [2 marks]

(b) Copy and complete the table. [7 marks]

(c) State a model underlying the analysis and carry out a one-way ANOVA hypothesis test, stating clearly your conclusions. [5 marks]

4. A *Thomas the Tank Engine* themed advent calendar has 25 doors. Behind each is a small figurine of either a locomotive or a carriage. The placement of the figurines behind the doors is random, though the manufacturer claims that on average, two thirds of figurines are locomotives and one third are carriages.

Two children, both more fond of locomotives than carriages, are disappointed that between their two advent calendars, they only have 25 locomotives (and 25 carriages). Clearly stating any assumptions, conduct a hypothesis test to assess the strength of evidence regarding whether the true ratio of locomotives is lower than claimed.[10 marks]

5. A phone manufacturer is investigating how long it takes for their latest model to boot up. A sample of 15 phones is randomly selected, and the boot-up time for each is measured. The sample mean boot-up time is 18 seconds, and the sample standard deviation is 4 seconds.

Clearly stating any assumptions made, construct a 95% confidence interval for the true average boot-up time for the latest model. [8 marks]

6. A city's police force investigates the relationship between population density (X, thousands of people per square mile) and crime rate (Y, number of crimes per 100,000 people). The data from 50 districts is summarised below:

$$\sum_{i=1}^{50} x_i = 1,500; \quad \sum_{i=1}^{50} y_i = 2,850; \quad S_{xx} = 30; \quad S_{yy} = 50; \quad S_{xy} = 32.$$

(a) From the summary statistics given, calculate the least squares regression line of y on x. [6 marks]

(b) Calculate the value of R^2. What precisely does this number tell you? [3 marks]

(c) Predict the crime rate in a region whose population density is 25,000 people per square mile. [2 marks]

(d) What is the expected change in crime rate if the population density increases by 3,000 people per square mile? [2 marks]

7. Using the one-way ANOVA notation seen in lectures (where Y_{ij} denotes the j-th observation in group i in an experimental setup where there are m observations in each of k groups, and • denotes averaging over the respective subscript), prove that

$$\sum_{i=1}^{k}\sum_{j=1}^{m}(Y_{ij} - Y_{\bullet\bullet})^2 = \sum_{i=1}^{k}\sum_{j=1}^{m}(Y_{ij} - Y_{i\bullet})^2 + m\sum_{i=1}^{k}(Y_{i\bullet} - Y_{\bullet\bullet})^2.$$

[7 marks]

Section B begins on next page.

Section B

8. In the first week of December, a national supermarket chain wishes to investigate sales of wall calendars for the year ahead. In particular, they are interested in how sales are affected by:

- the location of a small calendar display stand within the shop (either adjacent to the checkouts, or near the pharmacy section);
- the title of the calendar on the display stand (either '*Cute Kittens*', '*Supercars*', or '*Cliff Richard*').

Each of 18 stores of similar size reports their sale numbers across the week. The results are as follows:

	Kittens	**Supercars**	**Cliff Richard**	*Row averages*
Checkout	36	34	16	
	30	28	16	
	33	28	19	
	Average 33	*Average 30*	*Average 17*	*Average 26.667*
Pharmacy	16	20	18	
	22	18	15	
	28	19	21	
	Average 22	*Average 19*	*Average 18*	*Average 19.667*
Column averages	*27.5*	*24.5*	*17.5*	*23.167*

The data are to be analysed using the following model, in which α_i and β_i respectively denote the row and column effects, while γ_{ij} denote interaction terms:

$$\mathbb{E}[Y_{ijk}] = \mu + \alpha_i + \beta_j + \gamma_{ij}.$$

(a) List the usual two-way ANOVA restrictions on the parameters in this model. [2 marks]

(b) Give the values of Y_{123} and $Y_{\bullet 2 \bullet}$. [2 marks]

(c) Give the least squares estimates of μ, β_2 and γ_{22}. [3 marks]

The two-way ANOVA table for the model is:

```
               Df  Sum Sq  Mean Sq  F value    Pr(>F)
Location        A   220.5   220.50   18.900  0.000949 ***
Title           B   316.0        E   13.543  0.000837 ***
Location:Title  C   144.0    72.00        F  0.014350 *
Residuals       D   140.0    11.67
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

(d) Give the missing values A, B, C, D, E, and F. [3 marks]

(e) Evaluate the quantity $Y_{\bullet 1 \bullet} - Y_{\bullet 3 \bullet}$ and explain what it is an estimate of. Show that its estimated standard error is approximately 1.97. [5 marks]

Question 8 continues on next page

(f) Indicating in your answer which value from the table informs your conclusion, would you say that the number of calendars sold:

(i) is the same for both locations?

(ii) is the same for each of the three titles?

(iii) changes from location to location in the same manner for each of the three calendars? [3 marks]

9. An aspiring YouTuber wonders whether the number of views on a video she released a year ago might be well-modelled by a Poisson distribution, since she perceives them to occur at a fairly constant (albeit low) rate. She makes a note of how many views there are each day for 90 days. The results are as shown in the following table:

Number of views	Frequency
0	30
1	30
2	20
3	3
4	6
5	1

Conduct a chi-squared test to evaluate whether a Poisson distribution is a good fit to the observations. [12 marks]

10. For the data given in Question **8**, suppose the location variable is disregarded, so the data are now viewed as three groups (i.e. as a one-way setup), each of six observations. Denote the group means μ_{Kittens}, μ_{Cars}, and μ_{Cliff}.

(a) Construct two contrasts between group means representing:

(i) How many fewer copies of the '*Cliff Richard*' calendar sell than the average of the other two calendars;

(ii) How many more copies of the '*Cute Kittens*' calendar sell than '*Supercars*'.

Show that these contrasts are orthogonal and estimate their values. [6 marks]

(b) Calculate the sum of squares associated with each contrast. [2 marks]

(c) An R command and its output is given below:

```
> summary(aov(Sales ~ Title))
            Df Sum Sq Mean Sq F value Pr(>F)
Title        2  316.0  158.00   4.698 0.0261 *
Residuals   15  504.5   33.63
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Give a modified version of this one-way ANOVA table that replaces the row which begins '`Title`' with two rows, one for each contrast that you have constructed. For the column headed '`Pr(>F)`', state only whether the value is greater than or less than 0.05. Which critical value from the statistical tables informed this conclusion? What do you conclude from the modified table? [5 marks]

Section B concludes on next page

11. An advertising agency asserts that 12% of visitors to websites featuring their advertisements click on their banner to learn more about the promoted product. A client discovers that 24 out of 356 visitors clicked on the advertisement for their product.

Clearly stating any assumptions and hypotheses, conduct a hypothesis test to assess whether the rate of click-throughs is less than that claimed by the agency. In doing so, you should use an appropriate approximation with a continuity correction. [7 marks]

END OF PAPER