PRIFYSGOL ABERYSTWYTH UNIVERSITY

ADRAN MATHEMATEG / DEPARTMENT OF MATHEMATICS

ARHOLIADAU SEMESTER 2 / SEMESTER 2 EXAMINATIONS

MAI – MEHEFIN / MAY – JUNE 2023

MA26620 - Applied Statistics

The questions on this paper are written in English.

Amser a ganiateir - 2 awr

- 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 FX83 neu FX85 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.

Ar ôl eistedd, gall myfyrwyr lenwi tudalen flaen y llyfryn atebion a'r papur presenoldeb.

Peidiwch ag agor y papur arholiad tan y dywedir wrthych am wneud hynny.

Time allowed - 2 hours

- 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 FX83 or FX85 calculators ONLY may be used.
- Statistical tables will be provided.
- Students may submit answers to this paper in either Welsh or English.

Once seated, students may complete the front cover of the answer book and the attendance slip.

Do not open the question paper until instructed to do so.

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) Annual household expenditure (£) on energy;

(d) Names given to babies born in Ceredigion in 2022. [8 marks]

2. A supermarket receives a delivery of apples which are wrapped as six-packs. A sample of 20 six-packs of apples is taken and is found to have mean weight of 700g (per six-pack), with a sample standard deviation of 30g. Clearly stating any assumptions made, construct a 95% confidence interval for the population mean weight of a six-pack of apples. [9 marks]

3. A child playing the board game *Snakes and Ladders* has doubts that the die being used is fair. His statistically-minded father records the number of times each number appears in the next 90 rolls. The results are as follows:

Score rolled	1	2	3	4	5	6
Frequency	21	18	10	15	17	9

Test whether the die is fair. [9 marks]

4. A company that runs a small fleet of ice-cream vans records their sales (y, in multiples of £1000) and maximum temperature (x, in °C) each day for eleven days. The data is summarised as follows:

$$\bar{x} = 20.0, \quad \bar{y} = 8.0, \quad S_{xx} = 175, \quad S_{xy} = 121, \quad S_{yy} = 106.$$

(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 and comment. [2 marks]

(c) Predict the ice cream sales for a day on which the maximum temperature is 22°C. [2 marks]

5. Over a long period of time, a large industrial company has typically recorded 4 notifiable workplace accidents per month. Test whether the rate of accidents has increased if 37 such accidents are recorded in a 6 month period. In your answer, state clearly:

(i) the meaning of any notation you introduce as well as any assumptions made;

(ii) which variable follows a Poisson distribution, defining its parameter;

(iii) the two hypotheses;

(iv) the distribution of the number of accidents in 6 months when H_0 is true. Use tables to evaluate the p-value and state your conclusion. [10 marks]

6. In each of five compost mixes, m 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**
				3	0.0281
Within composts			25		
Total (corr)	1425				

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

(b) Give the value of m. [2 marks]

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

7. A provider of vehicle breakdown cover claims that the average response time between receiving a call and arriving at a breakdown is 50 minutes.

Data concerning response time is collected from 25 randomly sampled breakdowns. These 25 breakdowns have an average response time of 60 minutes, with sample variance 361 minutes^2.

Clearly stating your hypotheses and any assumptions you make, test whether the expected time between the company receiving a call and arriving at a breakdown is greater than the 50 minutes claimed. [10 marks]

(Section B begins on following page)

Section B

8. A sports scientist claims that the resting heart rate of male undergraduate students who exercise regularly is more than five beats per minute lower than that of male undergraduate students who do not exercise regularly. 40 exercisers and 30 non-exercisers were selected at random and their resting heart rates were recorded. The mean and standard deviations are given below:

	Exercisers	**Non-exercisers**
Resting heart rate (bpm)	60	68
Standard deviation (bpm)	2.0	6.0

Note that due to the disparity between observed sample standard deviations for the two samples, an assumption of equal variances is not appropriate.

(a) If X_i denotes the resting heart rate of the i-th regularly-exercising participant and Y_i denotes the resting heart rate of the j-th participant who does not exercise regularly, show that

$$ESE(\bar{X} - \bar{Y}) = \sqrt{\frac{S_X^2}{40} + \frac{S_Y^2}{30}}$$

and hence find the numerical value of $ESE(\bar{X} - \bar{Y})$. [5 marks]

(b) Hence perform a two-sample t-test to assess the strength of evidence supporting the sports scientist's claim.

Hint: you may find it useful to recall that to a good approximation, the T-statistic in the case of unequal variances is distributed as $t_{[\nu]}$, where

$$\nu = \frac{(f_X + f_Y)^2}{\frac{1}{n_X - 1} f_X^2 + \frac{1}{n_Y - 1} f_Y^2},$$

where $f_X = S_X^2/n_X$, $f_Y = S_Y^2/n_Y$, and the two sample sizes are denoted n_X and n_Y. If the result isn't an integer, rounding down leads to a better approximation.

[11 marks]

9. A manufacturer produces a certain type of light bulb that is supposed to have a failure rate of no more than 4%. To test whether the actual failure rate is within the acceptable range, a sample of 450 light bulbs is selected at random, and it is found that 25 of them have failed. Using a suitable approximation, conduct a suitable hypothesis test to determine if there is sufficient evidence to conclude that the actual failure rate is higher than 4%. [10 marks]

10. Eight students at each of four schools (coded A, B, C, D) are studying towards their *Further Mathematics* A-level qualification. They each sit a test which assesses their ability in early calculus, marked out of 100. A one-way ANOVA analysis is conducted on the result; an R command and its output is shown below:

```
> summary(aov(score~school))

            Df Sum Sq Mean Sq F value  Pr(>F)
school       3   1374   457.9   4.667 0.00911 **
Residuals   28   2747    98.1
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

> tapply(score,school,mean)
     A      B      C      D
73.000 73.250 58.375 62.250
```

In recent inspections, schools A and B were rated ‘outstanding’, while schools C and D were rated ‘good’.

(a) Write down three contrasts, λ_1, λ_2 and λ_3, respectively quantifying the difference in mean test scores between:

(i) the two ‘outstanding’ schools;

(ii) the two ‘good’ schools;

(iii) ‘outstanding’ schools and ‘good’ schools.

Show that the three contrasts are mutually orthogonal and give an unbiased estimate of each. [8 marks]

(b) Calculate the sum of squares associated with each contrast and state precisely how these relate to a figure in the R output above.

You may find it useful to recall that, using the notation used in lectures,

$L(\hat{\lambda}) = \left(m\hat{\lambda}^2\right) / \left(\sum_{i=1}^{k} c_i^2\right)$. [4 marks]

(c) Copy and complete the following table and state what conclusions can be drawn from it. In the column headed ‘p-value’, use the supplementary data given at the end of this exam booklet to give the smallest possible range of probabilities within which each p-value falls.

Source	SS	DF	MS	F	p-value
λ_1					
λ_2					
λ_3					
Residuals					

[12 marks]

End of paper.

(Supplementary data overleaf.)

Supplementary data The following table shows $p = P(F_{1,28} > c)$ for a variety of values of p and may be useful for question 10.

p	c
0.0001	20.52519
0.0005	15.48468
0.001	13.49759
0.005	9.28377
0.01	7.63562
0.02	6.08678
0.03	5.22753
0.04	4.63944
0.05	4.19597
0.10	2.89385
0.15	2.19082
0.20	1.72273
0.25	1.37997
0.30	1.11511
0.35	0.90342
0.40	0.73042
0.45	0.58698
0.50	0.46697
0.55	0.36612
0.60	0.28135
0.65	0.21039
0.70	0.15155
0.75	0.10355
0.80	0.06542
0.85	0.03643
0.90	0.01608
0.95	0.00400