PRIFYSGOL ABERYSTWYTH UNIVERSITY

ADRAN MATHEMATEG / DEPARTMENT OF MATHEMATICS

ARHOLIADAU SEMESTER 2 / SEMESTER 2 EXAMINATIONS

MAI / MAY 2026

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 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.

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 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.

© 2026 Aberystwyth University

Section A

1. Classify the following variables as discrete or 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) Number of books borrowed from a library by each member in a month;

(b) Star ratings given to a restaurant (1 star / 2 stars / 3 stars / 4 stars / 5 stars);

(c) Types of payment method used for online purchases.

(d) Daily temperature (°C) recorded in different towns; [8 marks]

2. Biscuits and cakes occasionally appear in the Mathematics lecturers' staff tea-room; no-one knows where they come from. This strange phenomenon has occurred for several years, with a long term average rate of appearance of one batch of baked goods appearing every 10 working days.

In the last 50 working days, only two batches of baked goods have appeared. Test whether the rate of baked goods appearing has significantly decreased. In your answer, clearly state:

(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 appearances of baked goods in 50 working days when H_0 is true.

Use tables to evaluate the p-value and state your conclusion. [10 marks]

3. A wildlife charity wishes to investigate the relationship between the amount of wetland habitat restored and the number of bird sightings recorded at its reserves. To do this, 15 reserves were randomly selected and assigned one of five restoration targets with different sizes (in hectares): $x = 2$, 4, 8, 16, and 32. After one year, the corresponding number of bird sightings (in hundreds), y, was recorded. The summary statistics for the data are as follows:

$$\bar{x} = 12, \quad \bar{y} = 45, \quad S_{xx} = 1{,}280, \quad S_{xy} = 3{,}840, \quad S_{yy} = 18{,}000.$$

(a) Using the summary statistics provided, calculate the least squares regression line of y on x. [5 marks]

(b) Compute the value of R^2 and interpret its meaning in this context. [4 marks]

(c) Predict the expected number of bird sightings when the restoration target is 20 hectares. [2 marks]

Section A continues on next page

4. A theme park claims that visitors spend an average of 75 minutes queueing for its most popular rollercoaster, *"The Probability Plunge"*.

A group of second-year statistics students suspect that the true mean waiting time is longer during the summer holidays. To investigate, a random sample of 9 visitors was taken on a day in the summer holidays and their waiting times (in minutes) were recorded. The sample mean waiting time was 84 minutes and the sample standard deviation was 12 minutes.

Defining any notation you introduce and clearly stating any assumptions you make, conduct a hypothesis test to assess whether the average waiting time is longer than claimed. [10 marks]

5. A board game café advertises that players escape from its most difficult escape room 40% of the time.

A sceptical group of mathematics students believe the true success rate is lower. To investigate, they record the outcomes of 50 independent escape attempts. Only 11 groups successfully escape.

Let p denote the true probability that a group successfully escapes on a single attempt. Assume attempts are independent and that the probability of success remains constant. Conduct an appropriate hypothesis test to determine whether p is lower than claimed. [9 marks]

6. A game studio is investigating whether background sound affects players' level-completion time.

Players were randomly assigned to one of four sound conditions: silence, rainforest sounds, lo-fi beats, or heavy metal. In each group, 5 players attempted the same level, and their completion times (in seconds) were recorded.

A one-way ANOVA was conducted to compare the mean completion times between the four groups, and the following ANOVA table was obtained:

Source	SS	DF	MS	F-ratio	P
				4	0.026
Within groups					
Total (corr.)	280				

(a) Copy and complete the ANOVA table. [8 marks]

(b) State a suitable model underlying the analysis and carry out the usual hypothesis test at the 5% level, stating clearly and carefully your conclusions. [4 marks]

SECTION A CONTINUES ON NEXT PAGE

7. A streaming platform claims that 65% of viewers watch a newly released episode within 24 hours of its release. To estimate the true proportion of viewers who watch within 24 hours, a random sample of 120 viewers was taken. Of these, 78 reported that they watched the episode within 24 hours.

Let p denote the true proportion of all viewers who watch within 24 hours.

(a) State the distribution of the number of viewers in the sample who watched within 24 hours. Also give any assumptions for this to be valid. [2 marks]

(b) State an unbiased estimator of p (denoted $\hat{p}$) and derive that its estimated standard error is given by

$$ESE(\hat{p}) = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}.$$

Calculate its numerical value. [4 marks]

(c) Obtain an approximate 95% confidence interval for p. [4 marks]

Section B

8. A Facebook group for fans of the TV show *Keeping Up with the Kardashians* recorded the number of new members per day, y, over a period of $n = 80$ days. The results were as follows:

y	Frequency
0	27
1	23
2	29
3	1

Test the hypothesis at the 5% significance level that the random variable Y, the number of new members per day, follows a Poisson distribution. Assume that daily observations are independent. [10 marks]

9. A nutrition researcher is investigating whether daily caffeine consumption differs between undergraduate and postgraduate students at a university. A total of 50 students are sampled, of whom 35 are undergraduates and 15 are postgraduates. The following table summarises the mean and standard deviation of daily caffeine intake (in milligrams):

	Undergraduates	**Postgraduates**
Mean caffeine intake (mg)	210.0	180.0
Standard deviation (mg)	40.0	35.0

Assume that daily caffeine intake in both groups is Normally distributed and that the population variances of the two groups are equal and denoted σ^2.

QUESTION CONTINUES ON NEXT PAGE

(a) Clearly stating any assumptions made, if X_i denotes the caffeine intake of the i-th undergraduate and Y_j denotes the caffeine intake of the j-th postgraduate, find the constant c (expressed in lowest terms) for which

$$\mathrm{Var}(\bar{X} - \bar{Y}) = c\sigma^2,$$

where $\bar{X}$ and $\bar{Y}$ denote the sample means for undergraduates and postgraduates respectively. [3 marks]

(b) Given that

$$S^2 = \frac{(n-1)S_1^2 + (m-1)S_2^2}{(n-1)+(m-1)}$$

denotes the pooled sample variance (you may use without proof that this is an unbiased estimate of σ^2), with $n = 35$ and $m = 15$, numerically evaluate the quantity $ESE(\bar{X} - \bar{Y})$. [2 marks]

(c) Hence construct a 95% confidence interval for the difference in mean daily caffeine intake between undergraduates and postgraduates. You may find the **Supplementary Table** (an extension of the t-tables) on page 7 of this exam paper useful. [5 marks]

10. A retailer is investigating daily sales (number sold per day) of three models of *Minions*-branded cuddly toys: Kevin, Stuart and Bob. In particular, they are interested in how sales are affected by:

- the location (either Student Union shop or Supermarket);
- the model (either Kevin, Stuart, or Bob).

On 18 days of similar footfall, sales were recorded, with 3 days observed for each location–model combination. The results are summarised in Table 1.

	Kevin	**Stuart**	**Bob**	*Row averages*
Student Union	30 34 38	20 24 28	22 26 30	
	Average 34.00	*Average 24.00*	*Average 26.00*	*Average 28.00*
Supermarket	22 25 28	13 16 19	12 15 18	
	Average 25.00	*Average 16.00*	*Average 15.00*	*Average* $56/3 \approx 18.67$
Column averages	*Average 29.50*	*Average 20.00*	*Average 20.50*	*Grand mean* $70/3 \approx 23.33$

Table 1: Daily sales (number sold) by location and Minion model.

The data are to be analysed using the following two-way ANOVA model:

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

where α_i and β_j denote the row (location) and column (model) effects, respectively, and γ_{ij} denotes the interaction between location and model.

QUESTION CONTINUES ON NEXT PAGE

(a) List the usual two-way ANOVA restrictions on the parameters in this model. [3 marks]

(b) Give the least squares estimates of μ, β_2 (the effect for Stuart), and γ_{22} (the interaction effect for Supermarket and Stuart). [3 marks]

The two-way ANOVA table for the model is:

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Location	A	392.00	392.00	31.360	1.16e-04 ***
Model	B	343.00	E	13.720	7.93e-04 ***
Location:Model	C	7.00	3.50	F	0.761
Residuals	D	150.00	12.50		

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(c) Give the missing values A, B, C, D, E, and F from the ANOVA table. [3 marks]

(d) Evaluate the quantity $Y_{\bullet 2\bullet} - Y_{\bullet 3\bullet}$ and explain what it is an estimate of. Show that its estimated standard error is approximately 2.04. [4 marks]

(e) Quoting in your answer the value from the table that informs your conclusion, would you say that daily sales:

(i) are the same for both locations?

(ii) are the same for each of the three Minion models?

(iii) change from location to location in the same manner for each model? [3 marks]

11. Using the data in Table 1 from **Question 10** for the Model factor only (i.e. the three treatment levels corresponding to Kevin, Stuart, and Bob), assume that the overall treatment (Model) sum of squares (SS_{MODEL}) is based on the column averages. Each treatment has $n = 6$ observations (3 from Student Union shop and 3 from Supermarket), and the residual mean square error in this one-way setup is given as $MS_{ERROR} = 36.6$.

One contrast is defined as

$$\lambda_1 = \mu_1 - \frac{1}{2}(\mu_2 + \mu_3),$$

where μ_j is the number of sales for the jth model (with $j = 1, 2, 3$ corresponding to Kevin, Stuart, and Bob, respectively).

(a) Give an interpretation of what λ_1 represents. [2 marks]

(b) Write down a contrast representing how many more Bob toys sell than Stuart. Verify that it is orthogonal to λ_1. [3 marks]

(c) Compute unbiased estimates of λ_1 and λ_2. [2 marks]

(d) Calculate the sum of squares associated with each of the contrasts λ_1 and λ_2. [2 marks]

(e) Construct a modified one-way ANOVA table which has a row for each of λ_1 and λ_2 in place of the Model factor. Your table should include degrees of freedom, sum of squares, mean squares, F values, and an indication of whether the p-values are less than 0.05 or greater than 0.05, quoting the critical value from the Statistical Tables that informs these p-values. [5 marks]

QUESTIONS END

SUPPLEMENTARY TABLE ON NEXT PAGE

Supplementary Table

The following table, an extension of the t-distribution table given in the Statistical Tables, may be useful when answering **Question 9**.

	Tail area probability										
df	25%	20%	15%	10%	5%	2.5%	1%	0.5%	0.25%	0.1%	0.05%
40	0.6807	0.8507	1.0500	1.3031	1.6839	2.0211	2.4233	2.7045	2.9712	3.3069	3.5510
41	0.6805	0.8505	1.0497	1.3025	1.6829	2.0195	2.4208	2.7012	2.9670	3.3013	3.5442
42	0.6804	0.8503	1.0494	1.3020	1.6820	2.0181	2.4185	2.6981	2.9630	3.2960	3.5377
43	0.6802	0.8501	1.0491	1.3016	1.6811	2.0167	2.4163	2.6951	2.9592	3.2909	3.5316
44	0.6801	0.8499	1.0488	1.3011	1.6802	2.0154	2.4141	2.6923	2.9555	3.2861	3.5258
45	0.6800	0.8497	1.0485	1.3006	1.6794	2.0141	2.4121	2.6896	2.9521	3.2815	3.5203
46	0.6799	0.8495	1.0483	1.3002	1.6787	2.0129	2.4102	2.6870	2.9488	3.2771	3.5150
47	0.6797	0.8493	1.0480	1.2998	1.6779	2.0117	2.4083	2.6846	2.9456	3.2729	3.5099
48	0.6796	0.8492	1.0478	1.2994	1.6772	2.0106	2.4066	2.6822	2.9426	3.2689	3.5051
49	0.6795	0.8490	1.0475	1.2991	1.6766	2.0096	2.4049	2.6800	2.9397	3.2651	3.5004
50	0.6794	0.8489	1.0473	1.2987	1.6759	2.0086	2.4033	2.6778	2.9370	3.2614	3.4960

Table 2: t-distribution critical values (one-tailed) for selected tail area probabilities for 40 to 50 degrees of freedom.