PRIFYSGOL ABERYSTWYTH UNIVERSITY

ADRAN MATHEMATEG / DEPARTMENT OF MATHEMATICS

ARHOLIADAU SEMESTER 2 / SEMESTER 2 EXAMINATIONS

MAI – MEHEFIN / MAY – JUNE 2025

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.

© 2025 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) Annual rainfall (mm) measured in different cities;

(b) Number of students enrolled in different university courses;

(c) Likert scale ratings of customer satisfaction (Strongly disagree / Disagree / Neutral / Agree / Strongly agree);

(d) Types of automobile brands. [8 marks]

2. A sports scientist is interested in whether a newly introduced training regimen has improved the average 5K running time. She collects a sample of 12 runners after completing the regimen and finds a sample mean finish time of 21 minutes with a sample standard deviation of 1.5 minutes.

Historical records indicate that the average 5K time before the training regimen was 22 minutes.

Clearly state your hypotheses and any assumptions you make, then test whether the new training regimen has significantly reduced the average 5K time. [10 marks]

3. A marketing firm wishes to investigate the relationship between advertising expenditure and monthly sales revenue for its regional offices. To do this, 15 offices were randomly selected and assigned one of five advertising budgets (in thousands of pounds): $x = 5$, 10, 20, 40, and 80. After one month, the corresponding sales revenue (in thousands of pounds), y, was recorded. The summary statistics for the data are as follows:

$$\bar{x} = 30, \quad \bar{y} = 150, \quad S_{xx} = 8{,}000, \quad S_{xy} = 40{,}000, \quad S_{yy} = 250{,}000.$$

(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 monthly sales revenue when the advertising expenditure is £$60{,}000$. [2 marks]

4. Over a long period of time, a local fire department typically receives 3 fire alarm activations per week. Test whether the rate of fire alarm activations has increased if 22 activations are observed in a 4-week period. 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 fire alarm activations in 4 weeks when H_0 is true.

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

5. Statistics lecturer Dr Kenobi believes that the probability of receiving a good haircut from his favourite Aberystwyth barbershop is 50%. Over the course of 20 haircuts, he experiences 13 good haircuts and 7 bad haircuts. Clearly state any assumptions you make, and conduct a suitable hypothesis test to determine whether the observed data provide evidence that the probability of a good haircut is greater than 50%. [10 marks]

6. A researcher is investigating the effect of coffee consumption on reaction times. Participants were assigned to one of four groups based on the number of cups of coffee consumed prior to testing (0, 1, 2, and 3 cups). In each group, 5 subjects had their reaction times (in milliseconds) recorded. An ANOVA was conducted to compare the mean reaction times between the groups, and the following ANOVA table was computed:

Source	**SS**	**DF**	**MS**	**F-ratio**	**P**
Between groups				4.5	0.020
Total (corr.)	944				

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

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

7. A customer suspects that a bag of fruity sweets is not evenly distributed among the five available colours: red, blue, green, yellow, and purple. The manufacturer claims that each color is equally likely to appear. To investigate, the customer randomly selects 120 sweets from a bag and records the following frequencies:

Colour	**Red**	**Blue**	**Green**	**Yellow**	**Purple**
Frequency	28	24	22	36	10

Test whether the observed frequencies provide evidence that the sweets are not evenly distributed. [8 marks]

SECTION B BEGINS OVERLEAF

Section B

8. A sports scientist is investigating the effect of player position on maximum vertical jump height (in centimeters) among basketball players. Out of 50 players, 40 are classified as guards and 10 as forwards. The following table summarizes the mean and standard deviation of jump heights for the two groups:

	Guards	**Forwards**
Mean jump height (cm)	70.0	65.0
Standard deviation (cm)	5.0	4.0

Assume that jump heights in both groups are Normally distributed and that the population variances of the two groups are equal.

(a) Clearly stating any assumptions made, if X_i denotes the vertical jump height of the i-th guard and Y_j denotes the vertical jump height of the j-th forward, show that

$$ESE(\bar{X} - \bar{Y}) = \sqrt{\frac{S^2}{8}},$$

where $\bar{X}$ and $\bar{Y}$ denote the sample means for guards and forwards respectively, and

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

denotes the pooled sample variance (you may use without proof that this is an unbiased estimate of σ^2), with $n = 40$ and $m = 10$. [5 marks]

(b) Hence construct a 95% confidence interval for the difference in mean jump heights between guards and forwards. You may find the **Supplementary Table** (an extension of the t-tables) on page 7 of this exam paper useful. [8 marks]

9. A national restaurant chain is investigating customer satisfaction ratings (on a scale from 0 to 100) for a new menu. In particular, they are interested in how ratings are affected by:

- the time of day at which the meal is served (either **Lunch** or **Dinner**);
- the type of service provided (either **Dine-in**, **Takeaway**, or **Delivery**).

In 18 branches of similar size, customer satisfaction ratings were recorded over a week. The results are summarized in Table 1 on the next page.

The data are to be analyzed 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 (time of day) and column (service type) effects, respectively, and γ_{ij} denotes the interaction between time of day and service.

QUESTION CONTINUES ON NEXT PAGE

	Dine-in	**Takeaway**	**Delivery**	*Row averages*
Lunch	75 80 78	70 72 74	65 68 67	
	Average 77.67	*Average 72.00*	*Average 66.67*	*Average 72.11*
Dinner	82 85 83	78 80 79	75 77 76	
	Average 83.33	*Average 79.00*	*Average 76.00*	*Average 79.44*
Column averages	*80.50*	*75.50*	*71.33*	*Average 75.44*

Table 1: Customer satisfaction ratings (0–100) by time of day and type of service.

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

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

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

The two-way ANOVA table for the model is:

```
                    Df   Sum Sq    Mean Sq  F value  Pr(>F)
Time of day          A   242.00    242.00   85.412   8.34e-07 ***
Service              B   252.78       E     44.608   2.78e-06 ***
Time of day:Service  C    10.33      5.17       F        0.203
Residuals            D    34.00      2.83
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

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

(e) Evaluate the quantity $Y_{\cdot 2 \cdot} - Y_{\cdot 3 \cdot}$ and explain what it is an estimate of. Show that its estimated standard error is approximately 2.04. [5 marks]

(f) Quoting in your answer the value from the table that informs your conclusion, would you say that the customer satisfaction rating:

(i) is the same for both times of day?

(ii) is the same for each of the three types of service?

(iii) changes from time of day to time of day in the same manner for each service? [3 marks]

Section B continues on next page

10. Using the data in Table 1 from **Question 9** for the Service factor only (i.e. the three treatment levels corresponding to Dine-in, Takeaway, and Delivery), assume that the overall treatment (Service) sum of squares ($SS_{SERVICE}$) is based on the column averages. Each treatment has $n = 6$ observations (3 from Lunch and 3 from Dinner), and the residual mean square error in this one-way setup is given as $MS_{ERROR} = 19.09$.

One contrast is defined as

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

where μ_j is the true mean customer satisfaction rating for the jth service (with $j = 1, 2, 3$ corresponding to Dine-in, Takeaway, and Delivery, respectively).

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

(b) Find a contrast representing the difference in satisfaction between Takeaway and Delivery of the form

$$\lambda_2 = c_1\mu_1 + c_2\mu_2 + c_3\mu_3.$$

Verify that it is orthogonal to λ_1. [3 marks]

(c) Using the provided column averages, 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. [4 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 Service 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. [7 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 8**.

	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.