PRIFYSGOL ABERYSTWYTH UNIVERSITY

ADRAN MATHEMATEG / DEPARTMENT OF MATHEMATICS

ARHOLIADAU SEMESTER 1 / SEMESTER 1 EXAMINATIONS

IONAWR / JANUARY 2026

MAM5120 – Statistics Concepts, Methods and Tools

The questions on this paper are written in English.

Amser a ganiateir - 3 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 ar lefel rhagoriaeth.
- Cyfrifianellau Casio FX-83 neu FX-85 YN UNIG a ganiateir.
- Mae modd i fyfyrwyr gyflwyno atebion i'r papur hwn naill ai yn y Gymraeg neu'r Saesneg.

Time allowed - 3 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 distinction level mark.
- Casio FX-83 or FX-85 calculators ONLY may be used.
- Students may submit answers to this paper in either Welsh or English.

© 2026 Aberystwyth University

The datasets for questions 4, 5, 6, 7, 8, 9, 10, 11 and 12 are available (as Q4.csv, Q5.csv etc.) in the Assignments tab of the Blackboard page for the module. I recommend that you create a project for the exam with a data folder and a scripts folder and download all of the datasets to the data folder. They can then be read in using, for example
`Q4 <- read.csv("data/Q4.csv")`.

Section A

1. Here we consider a biological test that offers a binary response (positive/negative). We assume each subject tests positive with a fixed probability, θ.

(a) What is the distribution of the number of positive tests in a sample of size 120?

(b) Suppose that out of 120 subjects, 43 test positive.

(i) Calculate the p-value for testing the null hypothesis, $H_0 : \theta = \frac{5}{12}$, against the alternative $H_A : \theta < \frac{5}{12}$.

(ii) What is the meaning of the p-value in part (b)(i)?

(iii) What is the conclusion of the statistical test?

[2 + 2 + 2 + 1 marks]

2. The number of parasitic wasp eggs found in each of 40 larvae are summarised in the following table:

number of eggs	frequency
0	8
1	10
2	14
3	6
4	2
5+	0

The number of eggs in a randomly selected larva may be modelled as a Poisson random variable with mean λ.

(a) If X is a random variable corresponding to the number of eggs in one larva (so $X \sim \text{Pois}(\lambda)$), what is the distribution of the total number of eggs found in 40 larvae?

(b) Carry out a hypothesis test to assess whether the data provide evidence that the rate parameter λ is greater than 1.5. Include your R code for the calculation.

[2 + 4 marks]

3. Suppose the weights of gannet eggs are known to follow a Normal distribution with mean 83g and variance 27.04g^2.

(a) Write down the distribution of the mean weight of 32 eggs.

(b) The weights of $N = 50$ eggs were measured. The summary statistics are as follows:

$$\bar{x} = 84.2\text{g} \qquad\qquad s^2 = 4.9^2\text{g}^2, \text{ i.e. } s = 4.9\text{g}.$$

(i) If

$$X \sim N(\mu, \sigma^2),$$

write down the distribution of

$$\frac{\bar{X} - \mu}{s/\sqrt{N}},$$

where

$$\bar{X} = \frac{1}{N}\sum_{i=1}^{N} X_i$$

is the random variable corresponding to the sample mean in a sample of size N and s^2 is the sample variance.

(ii) Calculate a 95% confidence interval for the mean using the summary statistics and the distribution of part (b)(i).

(iii) What is the interpretation of a 95% confidence interval?

[2 + 2 + 3 + 3 marks]

4. In an experiment to assess the effect of an enriched pig feed on growth, 6 pairs of piglets were used, with both members of each pair coming from the same litter, i.e. 6 litters in total. In each pair, one piglet was randomly assigned to a control group who received a standard diet, and the other received a supplemented diet. The increase in pig weight over two weeks was measured and the data are presented in the table below. (The data are in the file Q4.csv on Blackboard.)

pair	standard	supplement
1	0.4	0.7
2	0.6	0.6
3	0.6	0.8
4	0.5	0.4
5	0.3	0.7
6	0.4	0.5

(a) Clearly stating your null hypothesis, carry out an appropriate two-tailed test to compare the effectiveness of the two diets. Report the value of the test statistic, the distribution of the test statistic if the null hypothesis is true, the p-value and your conclusion.

(b) What assumptions about the distributions of weight gain in the two groups does the test of part (a) rely on?

(c) What diagnostic plot would help to check this assumption? (No need to do the plot.)

[6 + 2 + 1 marks]

5. Suppose an experiment was carried out to assess whether the level of heavy metal toxicity in fish populations had changed since 1995. In 1995, the mean level of heavy metals in tuna was found to be 132μg/kg.

In 2015, the levels of heavy metals in 15 fish samples were found to be

131.4	130.8	134.1	137.2	135.1
132.4	131.6	130.0	129.6	134.2
132.1	129.7	135.3	133.8	131.9

The data are in the file Q5.csv on the Blackboard page. Read in the data using

```
Q5 <- read.csv("data/Q5.csv")
```

and then if you type

```
Q5$heavy_metal_level
```

you can access the vector of numbers.

(a) Using 132 μg/kg as the true population mean for 1995, carry out a test at the 5% significance level to assess whether there has been a change in toxicity levels since 1995. Report your findings.

(b) Assuming the sample variance and sample size are held fixed, what is the critical region at the 1% level for the fish toxicity data? (In other words, for which values of the sample mean would we reject

$$H_0 : \mu_{2015} = 132$$

in favour of

$$H_A : \mu_{2015} \neq 132$$

at the 1% level, where μ_{2015} is the true population mean for 2015?)

[3 + 4 marks]

6. (a) Suppose that in a study to assess the effect of temperature on activity of an enzyme, the following data were collected. (The data are available in Q6.csv on Blackboard.)

temperature (°C)	enzyme activity (arbitrary units)
5	40
5	43
10	66
10	59
15	85
15	85
20	105
20	110
25	119
25	130

Regress enzyme activity on temperature and report your findings. Include the fitted model, an interpretation of the parameters in the model and an interpretation of the F statistic and associated p-value.

(b) Sketch and analyse appropriate diagnostic plots.

[6 + 4 marks]

7. In an agricultural experiment with a control treatment (standard fertiliser) (C) and three enhanced fertilisers (T1, T2 and T3), the following yield data were obtained. (They are available in Q7.csv on Blackboard.)

treatment	yield (kg/hectare)
C	2.05
C	2.61
C	2.52
C	2.40
T1	1.96
T1	2.07
T1	2.24
T1	2.38
T2	2.49
T2	3.07
T2	3.24
T2	3.51
T3	2.58
T3	3.12
T3	2.99
T3	3.09

(a) Conduct a one-way ANOVA and appropriate post hoc tests to establish which means are significantly different. Report your findings.

(b) Using the default treatment contrasts, interpret the coefficients in the summary table of the linear model.

[5 + 5 marks]

8. This question uses the `dplyr` package in R to carry out data manipulation. First type the following two lines of code to install and activate the `dplyr` library:

```
install.packages("dplyr")
```

```
library(dplyr)
```

You will need to use the verbs `filter`, `arrange`, `summarize`, `group_by` and `select` to answer this question.

The data set Q8.csv is available on the Blackboard site for this exam. Assuming you have the dataset in your `data` folder in your RStudio project, you can load in the data using the command `Q8 <- read.csv("data/Q8.csv")`. The dataset contains information about the fuel economy of 234 cars. The variable `cty` gives the city fuel economy (in miles per gallon) and the variable `hwy` gives the highway fuel economy (again in miles per gallon). Use the five verbs mentioned above to **write the R code** (i.e. give the R code as your answer to each question) to generate the following tables:

(a) A table with three columns, manufacturer, model and cty, showing only Volkswagen, Honda and Toyota cars and listed in decreasing order of cty (use `desc(cty)` in the relevant place in your code).

(b) A table with 2 rows, one for automatic cars and one for manual cars, and three columns, transmission, mean_hwy and sd_hwy, where mean_hwy gives the mean highway fuel economy for a particular transmission and sd_hwy gives the standard deviation of the highway fuel economy. NOTE: Make sure you use the transmission variable rather than the trans variable.

[6 + 5 marks]

Section B

9. In an experiment to investigate the effect of pH on the activity of an enzyme, the following data were collected. The data are available in Q9.csv on Blackboard.

pH	enzyme activity
7.2	172.5
3.9	233.6
10.6	45.4
10.1	75.4
6.0	187.8
6.5	183.8
9.3	87.7
6.8	185.4
9.7	66.6
9.0	103.6

(a) Describe the plot of enzyme activity against pH.

(b) Fit the following model:

$$\text{Model 1} : Y_i = \beta_0 + \beta_1 X_i + \epsilon_i.$$

What do you notice in a plot of residuals against fitted values for Model 1?

(c) Fit the following two models:

$$\begin{aligned} \text{Model 2} : Y_i &= \beta_0 + \beta_1 X_i + \beta_2 X_i^2 + \epsilon_i \\ \text{Model 3} : Y_i &= \beta_0 + \beta_3 X_i^2 + \epsilon_i. \end{aligned}$$

(For example for Model 2 use `model2<-lm(enzyme_activity~pH+I(pH^2))`.)
Using either the `anova()` function or the `stepAIC()` function in the MASS library, identify the best fitting model of Models 1, 2 and 3. Explain your reasoning in selecting the best model.

[3 + 2 + 5 marks]

10. This question uses the following data. (The data are available in Q10.csv on Blackboard.)

X	Y	X	Y
1	4.0	5	9.4
1	4.7	5	14.3
1	5.3	5	13.8
2	4.7	6	13.2
2	7.4	6	9.3
2	7.1	6	11.1
3	9.3	7	25.6
3	12.3	7	18.4
3	5.3	7	13.0
4	16.1	8	11.5
4	8.0	8	17.4
4	6.5	8	5.7

(a) In a plot of Y against X what assumption of linear modelling is clearly violated? How can you transform the data to improve the fit of the assumptions of a linear model?

(b) What is the model for regressing the transformed variable on the explanatory variable?

(c) Fit and interpret the linear model in part (b), reporting the fitted model and the interpretation of the X coefficient in the transformed model.

[2 + 2 + 3 marks]

11. Consider the following data table. (The data are available in Q11.csv on Blackboard.)

X_1	X_2	X_3	Y
2.6	4.5	0.8	29.1
1.1	5.3	0.5	27.9
1.5	4.0	0.1	30.7
2.5	4.1	-1.1	41.9
1.5	4.1	-1.3	34.9
2.7	4.8	-1.3	40.8
2.5	3.9	1.6	29.4
1.3	4.1	0.4	26.3
2.8	4.1	-1.0	39.6
2.2	3.6	2.0	26.0
2.2	4.7	1.6	26.8
2.7	3.7	2.2	22.2

Fit and interpret a linear model with Y as the response variable and X_1, X_2 and X_3 as the covariates. Discuss in particular the interpretation of the X_1 coefficient in the summary table.

[6 marks]

12. In an experiment to compare early flowering (WT = wild type) *Arabidopsis thaliana* plants with a late flowering mutant (mut), the growth of the plants (in mm) over a six-week period were measured under both short days (SD = 8 hours daylight, 16 hours darkness) and long days (LD = 16 hours daylight, 8 hours darkness). There were three plants per experimental condition (WT/LD, WT/SD, mut/LD, mut/SD). The following data were collected. (The data are available in Q12.csv on Blackboard.)

day length	ecotype	growth (mm)
LD	WT	5.8
LD	WT	6.4
LD	WT	6.0
LD	mut	3.2
LD	mut	3.6
LD	mut	3.1
SD	WT	4.9
SD	WT	4.6
SD	WT	4.4
SD	mut	3.7
SD	mut	4.0
SD	mut	3.8

(a) Identify the factors and levels in the experiment

(b) Fit the model

$$Y_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \epsilon_{ijk},$$

where α_i is the main effect for day length, β_j is the main effect for ecotype and $(\alpha\beta)_{ij}$ is the interaction effect. Present and interpret the ANOVA table, quoting the F-value, the p-value and a clear statement of your conclusion. Include an interpretation of the intercept and the day_lengthSD coefficients in the linear model.

(c) Describe the interaction plot that has separate lines for the different ecotypes. How does this relate to the ANOVA table of part (b)?

[2 + 8 + 4 marks]

13. Suppose that six coins are tossed and the number of coins showing a head is recorded. In an experiment involving 400 tosses of six coins, the following data were recorded.

no. of heads	frequency
0	12
1	38
2	110
3	120
4	85
5	30
6	5

(a) Estimate the probability of a head using the maximum likelihood estimator (MLE) for the probability of a head (total number of heads over total number of coins).

(b) Using the MLE estimated above to calculate the expected values, carry out a chi-squared goodness of fit test to establish whether the data could come from a binomial distribution. Explain your reasoning carefully and include any R commands you use in your answer. (NOTE: I have written the question so none of the expected values are less than 5, so there is no need to pool any of the bins.)

[3 + 9 marks]