PRIFYSGOL ABERYSTWYTH UNIVERSITY

ADRAN MATHEMATEG / DEPARTMENT OF MATHEMATICS

ARHOLIADAU SEMESTER 2 / SEMESTER 2 EXAMINATIONS

MAI / MAY 2020

MA26620 - Applied Statistics

The questions on this paper are written in English.

If you have questions about the paper during the exam, contact the module co-ordinator, Dr Adam Vellender, on asv2@aber.ac.uk.

You should write out solutions to the paper and upload them to Blackboard as a single PDF file.

Amser a ganiateir - 3 awr
Mae'n rhaid cyflwyno eich atebion erbyn 12:30 (amser y DU).

- Gellir rhoi cynnig ar bob cwestiwn.
- Rhoddir mwy o ystyriaeth i berfformiad yn rhan B wrth bennu marc dosbarth cyntaf.
- Mae modd i fyfyrwyr gyflwyno atebion i'r papur hwn naill ai yn y Gymraeg neu'r Saesneg.
- Mae tablau ystadegol ar gael ar Blackboard.

Time allowed - 3 hours
Submission must be completed by 12:30 (UK time).

- All questions may be attempted.
- Performance in section B will be given greater consideration in assigning a first class mark.
- Students may submit answers to this paper in either Welsh or English.
- Statistical tables are available on Blackboard.

Section A

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

(a) Heights of trees (m) in Penglais Woods;

(b) Weekly spending (£) of families on fast food;

(c) Military ranks;

(d) Names of rabbit breeds. [8 marks]

2. An ornithologist is interested in the effect of garden feeding upon the weight of European goldfinches, *Carduelis carduelis*. She records the weight of 12 goldfinches found in her garden and calculates a sample mean of $\bar{x} = 15.2$ grams, with sample standard deviation $S = 0.8$ grams.

A large 1960s study conducted before the rise in popularity of feeding garden birds found the average weight of goldfinches at that time to be 14.5 grams.

Clearly stating your hypotheses and any assumptions you make, test whether modern-day goldfinches are significantly heavier than they were in the 1960s. [10 marks]

3. A pharmaceutical firm would like to obtain information on the relationship between the dose level of a drug product and its potency. To do this, a sample containing virus particles was inserted into each of 15 test tubes. They were then incubated for 5 days at 30°C. Three test tubes were randomly assigned to each of the five different dose levels investigated ($x = 2$, 4, 8, 16 and 32mg). A measure of protective strength of the product against the virus culture, y, was recorded.

The data is summarised as follows:

$$\bar{x} = 12.4, \quad \bar{y} = 15.8, \quad S_{xx} = 1785.6, \quad S_{xy} = 3966, \quad S_{yy} = 9200.$$

(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. [3 marks]

(c) Predict the value of the measure of protective strength of the product against the virus culture at a dose level of 20mg. [2 marks]

4. A mother observes her toddler picking up a spoon 20 times while he is learning to use cutlery. She suspects the toddler has a slight preference for using his right hand to pick up a spoon, and observes 13 right-handed pick-ups and 7 left-handed pick-ups. Clearly stating any assumptions, conduct a suitable hypothesis test to determine the strength of evidence regarding whether the toddler has a preference. [9 marks]

5. For quality control purposes, the weights (measured in grams) of 14 packets of chocolate digestive biscuits are measured at a chocolate factory. Their sample mean weight is 270g, with a sample standard deviation of 2g. Clearly stating any assumptions made, construct a 95% confidence interval for the population mean weight of a packet of chocolate digestive bisuits. [9 marks]

6. As part of vehicle crash safety testing, five cars of each of four new types of car were driven into a wall at 30mph. The pressure applied to the head of the driver (crash test dummy) was recorded for each collision and the following ANOVA table computed:

Source	SS	DF	MS	F-ratio	P
Between models				4	0.027
Total (corr)	2800				

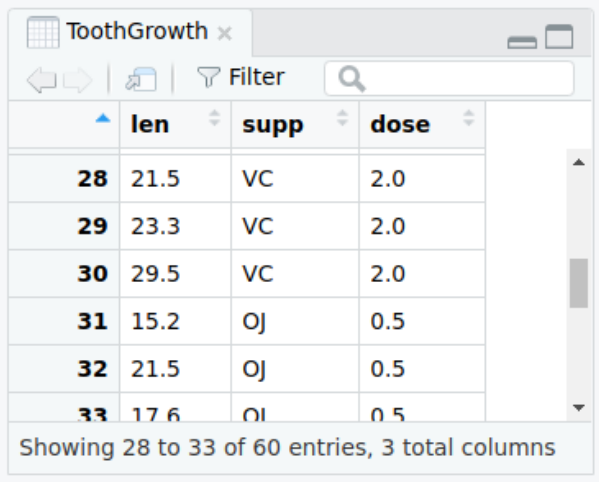
(a) Copy and complete the table. [8 marks]

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

7. This question concerns commands in the statistical package R/RStudio.

An experiment was conducted studying the effect of vitamin C on tooth growth in 60 guinea pigs. Each animal received one of three dose levels of vitamin C (0.5, 1, and 2 mg/day) by one of two delivery methods, (orange juice, coded as OJ, or chemically produced vitamin C, coded as VC). The response is the length of odontoblasts (cells responsible for tooth growth) in 60 guinea pigs.

The resulting data was imported into RStudio as a dataframe named `ToothGrowth` and then attached. A screenshot of some of the rows of data is given below.


ToothGrowth

Filter

	len	supp	dose
28	21.5	VC	2.0
29	23.3	VC	2.0
30	29.5	VC	2.0
31	15.2	OJ	0.5
32	21.5	OJ	0.5
33	17.6	OJ	0.5

Showing 28 to 33 of 60 entries, 3 total columns

(a) Write down an R command that would return the mean odontoblast length for the whole dataset. [1 mark]

(b) Write down an R command that would return a multiple boxplot of odonotoblast length, with a box for each of the two delivery methods. You are not required to add titles or edit axis labels. [2 marks]

(c) Briefly explain the effect of running the command
`tapply(len,dose,median)`. [2 marks]

(d) Briefly explain the effect of running the command
`groupA <- subset(ToothGrowth, dose > 1 & supp == "OJ")`. [3 marks]

(e) Write down an R command that would return a two-way ANOVA table for this data. [2 marks]

Section B

8. Over a long period of time, a hotel has typically experienced 10 booking cancellations per month. Test whether the rate of cancellations has increased if 48 cancellations are observed in a 4 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 cancellations in 4 months when H_0 is true.

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

9. In an experiment into the effect of diet on longevity, 109 rats were fed a regulated diet while 81 were given unlimited access to food.

Let X_i for $i = 1, \ldots, 109$, denote the lifetime in days of the rats with unlimited food, and Y_i for $i = 1, \ldots, 81$, denote the lifetime in days of the rats on a regulated diet. Further, suppose $X_i \sim N(\mu_X, \sigma_X^2)$ independently, and $Y_i \sim N(\mu_Y, \sigma_Y^2)$ independently.

The average lifetime and standard deviation of the rats in each group are given below.

	Unlimited	**Regulated**
Mean lifetime (days)	968	684
Standard deviation (days)	285.6	134.1

An F-test suggests that an assumption of equal population variances would be inappropriate for this data.

(a) By considering the variance of $\bar{X} - \bar{Y}$, show that

$$ESE(\bar{X} - \bar{Y}) = \sqrt{\frac{S_X^2}{109} + \frac{S_Y^2}{81}},$$

where $\bar{X}$ and $\bar{Y}$ respectively denote the sample means of the X_i and Y_i, while S_X^2 denotes the sample variance of the X_i, and S_Y^2 denotes the sample variance of the Y_i. [5 marks]

(b) Hence conduct a hypothesis test to determine whether rats with access to unlimited food live longer than those on the experiment's regulated diet.

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

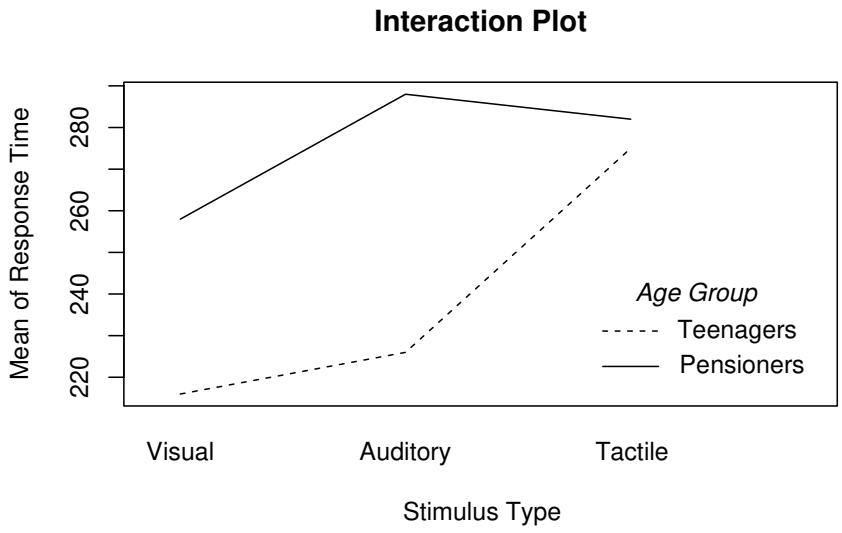
$$\nu = \frac{(f_X + f_Y)^2}{\frac{1}{n-1} f_X^2 + \frac{1}{m-1} f_Y^2}.$$

Here, $f_X = S_X^2/n$ and $f_Y = S_Y^2/m$ (n and m being the sample sizes of X and Y respectively). If the result isn't an integer, rounding down leads to a better approximation.

[10 marks]

10. A psychologist is interested in reaction times of people of different ages to stimuli of visual, auditory and tactile types. She takes 15 teenagers and 15 pensioners and splits each age group into three equally sized subgroups, each of which receive either a visual, auditory or tactile stimulus and are asked to press a button as soon after the stimulus as they can. She records the time taken (in milliseconds, ms) to press the button. The results, along with some R output, are given below:

	Visual	**Auditory**	**Tactile**	*Row averages*
Teenagers	200 220 210 220 230	220 240 180 240 250	280 290 290 305 210	
	Average 216	*Average 226*	*Average 275*	*Average 239*
Pensioners	250 260 300 260 220	200 260 400 270 310	270 240 300 300 300	
	Average 258	*Average 288*	*Average 282*	*Average 276*
Column averages	*237*	*257*	*278.5*	*257.5*




```
> summary(aov(responseTime~ageGroup*stimulusType))
                     Df Sum Sq Mean Sq F value Pr(>F)
ageGroup              1  10268   10268   6.610 0.0168 *
stimulusType          2   8615    4307   2.773 0.0825 .
ageGroup:stimulusType 2   3875    1938   1.247 0.3052
Residuals            24  37280    1553
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

[*Question continues overleaf*]

The model used to analyse the data is

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

where the α's refer to the age group effects and the β's refer to the stimulus type effects.

(a) Give the ranges of the subscripts i, j, k and note the constraints imposed on the parameters in this model. [4 marks]

(b) In the standard notation, give the values of $Y_{22\bullet}$ and $Y_{\bullet 2\bullet}$. [2 marks]

(c) Quote unbiased estimates of μ, β_2 and γ_{21}. [4 marks]

(d) Are there significant differences between the two age groups? How about between the three stimulus types? In your answer, quote which parts of the output lead you to your conclusions. [4 marks]

(e) Is there a significant interaction between the age of participant and the type of stimulus? Give reasons for your answer. What does this mean? What specifically does the interaction plot tell you? [5 marks]

(f) Estimate the contrast $\lambda = \frac{1}{2}\beta_1 + \frac{1}{2}\beta_2 - \beta_3$ and calculate its associated sum of squares. Explain what this contrast represents. [6 marks]

Hint: The sum of squares associated with a contrast, in notation consistent with that used throughout the lecture course, is $\dfrac{m\hat{\lambda}^2}{\sum c_i^2}$.