SECTION A: Compulsory questions

QUESTION 1

An exam paper provides students with a data set consisting of 24 times (in hours and minutes) and instructs them to illustrate it by means of a suitable diagram. Figures 1a and 1b below are two of the diagrams submitted.

(i) Criticise the two diagrams. **[4 marks]**

(ii) Assuming the numerical values in figure 1b are correct (with the stem in hours and leaves representing minutes), draw a cumulative frequency diagram. **[3 marks]**

(iii) Calculate the median and quartiles of the data set and show how they can be derived from your diagram. **[3 marks]**

[Total: 10 marks]

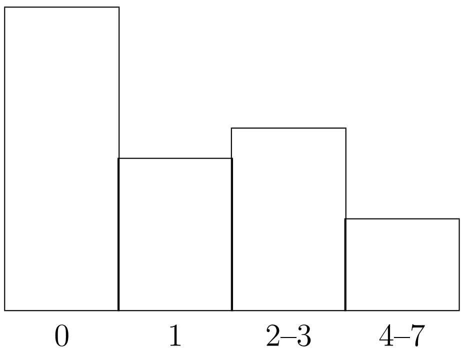


Figure 1a

7	32
4	21 48
3	55 35
2	22 38 10 44
1	52 07 21 17 39
0	37 11 26 49 11 18 02 09 29 04

Figure 1b

QUESTION 2

A is the event that I have at least one lecture during a day, B the event that I have at least one committee meeting, C the event that I have at least one meeting with a student and D the event that the first entry in my diary is for a lecture. If I have a committee meeting, it is always the first event in the diary for that day, and I never arrange meetings with students on days when I have committee meetings.

(i) Draw a Venn diagram to illustrate the situation. **[2 marks]**

(ii) Explain why $B \cap D \subset C \cap D \subset A \cap D$. **[2 marks]**

(iii) In a five-day week there are an average of 3 days when I have lectures, 3 days when I have meetings with students and 1.5 days when I have committee meetings. Suppose that A and B are independent, A and C are independent, and $\mathbb{P}(C|D) = \frac{1}{2}$.

(a) In a 10-week term, on how many days do I expect to have no appointments at all?

(b) Calculate the largest possible value for the probability that my day begins with a lecture and involves at least one meeting with a student. **[6 marks]**

[Total: 10 marks]

QUESTION 3

(i) I want to take a train to Edinburgh but all seats are booked. I am told that empty seats occur according to a Poisson process with rate $\lambda = 0.5$ per hour.

(a) Let T denote the time, in hours, until the first empty seat is available. Write down the density function of T and use it to calculate $\mathbb{P}(2 < T \leq 3)$.

(b) Use the Poisson distribution to calculate the probability that there are no empty seats in the first two hours but at least one in the third hour. **[5 marks]**

(ii) A function F is defined as

$$F(x) = \begin{cases} 0 & \text{if } x < 0 \\ \left(\frac{x}{1-x}\right)^2 & \text{if } 0 \leq x \leq c \\ 1 & \text{if } x > c \end{cases}$$

(a) Show that there is one value, c_0, of c for which F is the cumulative distribution function of a continuous random variable. Evaluate the density function f corresponding to F when $c = c_0$.

(b) Suppose X is a random variable with density function f. Calculate the expectation of $(1 - X)^3$. **[5 marks]**

[Total: 10 marks]

QUESTION 4

Andrew and Bob play the following game.

- Both players flip three fair coins and their score is given by the resulting number of heads.
- After seeing his score, but unaware of Bob's score, Andrew can decide to swap his score with that of Bob.

Andrew adopts the the following strategy: he will retain his score only if it is 2 or 3 and swap it if it is 0 or 1.
Denote by A and B Andrew's and Bob's final scores.

(i) Compute the probability of the following events: $\{A = 3 \cap B = 3\}$, $\{A = 2 \cap B = 1\}$, $\{A = 0 \cap B = 2\}$. **[5 marks]**

(ii) Compute the conditional expected value of Andrew's score given the number of heads in his coin flip. Deduce Andrew's expected score. **[3 marks]**

(iii) Andrew changes his strategy and decide to keep his score only if he sees 3 heads in his coin flip. Will his expected score increase? **[2 marks]**

[Total: 10 marks]

QUESTION 5

The efficiency of an assembly line is measured by the number X of functioning items before a faulty one is produced. The distribution of X is geometric with parameter p, with probability function

$$\Pr(X = k) = (1-p)^k p,\ k = 0, 1, 2, \ldots$$

where p is the probability that a given piece is faulty. For a geometric distribution, $\mathbb{E}(X) = \frac{1-p}{p}$ and $\mathrm{Var}(X) = \frac{1-p}{p^2}$.
Factory managers have recently introduced some changes in the assembly line that should result in greater efficiency. Prior to these changes, the value $p = 0.02$ was considered to be appropriate.
To assess whether the introduced changes are effective, a test of $H_0 : p = 0.02$ against $H_1 : p < 0.02$ is required.

(i) Based on a single observation X, it is suggested to use the following criterion of rejection of the null hypothesis: reject H_0 if $X > c$ for some threshold c. Explain why this is reasonable. **[1 mark]**

(ii) For $c = 150$, compute the type I error probability and the type II error probability if $p = 0.01$. (the following formula may be used: $1 + \alpha + \alpha^2 + \ldots + \alpha^k = \frac{1-\alpha^{k+1}}{1-\alpha}$ for $\alpha \neq 1$) **[4 marks]**

(iii) A sample $X_1, \ldots, X_{40}$ is observed. Suggest a possible test statistic and, using a suitable approximation, compute the critical region that guarantees a 5% type I error. Compute the p value if $\sum X_i = 2400$. What is your conclusion? **[5 marks]**

[Total: 10 marks]

QUESTION 6

An insurer suspects that claim estimates received from garage XYZ are too high. To detect a possible fraud, a comparison with claim estimates received from the more reliable garage AAA, for similar cars and accidents, is made. The results are presented in the table below.

garage	number of claims	sample mean	sample variance
XYZ	40	£1850.47	£^{2} 46643.68
AAA	50	£1496.19	£^{2} 40367.96

It is assumed that claim estimates received from garages XYZ and AAA are distributed normally with mean and variance μ_X, σ_X^2 and μ_Y, σ_Y^2 respectively.

(i) Show that it is plausible to assume that the variance of claim estimates received from the two garages is the same. (Use $F_{39,49,2.5\%} = 1.8082$ and $F_{49,39,2.5\%} = 1.8463$) **[3 marks]**

(ii) Develop a test to decide at the 5% level if garage XYZ is actually defrauding the insurer. **[5 marks]**

(iii) The insurer wants further evidence and compares the data obtained from garage XYZ with the historical industry claim average for a very large number of similar cars and accidents, which is £1600. What is the conclusion of the appropriate test in this case? **[2 marks]**

[Total: 10 marks]

SECTION B: Optional questions

QUESTION 7

(i) Give a definition of a *partition*. **[2 marks]**

(ii) Provide a proof of Bayes' Theorem, which states that, if $\{A_1, A_2, \ldots, A_n\}$ is a partition of the state space Ω, then

$$\mathbb{P}(A_i|B) = \frac{\mathbb{P}(B|A_i)\,\mathbb{P}(A_i)}{\sum_{j=1}^{n} \mathbb{P}(B|A_j)\,\mathbb{P}(A_j)}.$$

[4 marks]

(iii) In a television game show a contestant has to pick one of n identical doors, where $n \geq 3$. There is a prize worth £50,000 located behind one of the doors. Whichever door the contestant picks, the game show host points to a different door and truthfully says "There is no prize behind this door". The contestant can then choose whether to stay with the original door or to change to a different door.

(a) Draw a tree diagram to show the situation if the contestant decides to stay with the original door, and another tree diagram for the case where the contestant decides to choose a different door instead.

(b) Which situation has the greatest probability that the contestant wins the prize? **[6 marks]**

(iv) In a revised version of the game show there are $3a + 1$ doors facing the contestant, where $a \geq 2$. Behind one door is a £50,000 prize, behind a doors are prizes of £10,000, and behind the remaining $2a$ doors there are no prizes. When the contestant picks a door, the game show host points to a different door and says "There is a £10,000 prize behind this door". The contestant can choose to stay with the original door, change to the door the game show host is pointing out, or change to a different door.

(a) Derive an expression involving a for the expectation of the prize behind the selected door (A) if the contestant stays with the original door, (B) if the contestant chooses the door indicated by the game show host, or (C) if the contestant picks a different door.

(b) Show that, for every value of $a \geq 2$, there is a strategy which is better than just staying with the door originally chosen. **[8 marks]**

[Total: 20 marks]

QUESTION 8

(i) The amount spent in a large supermarket by shoppers with large shopping trolleys may be assumed normally distributed with a mean value of £65.50 and standard deviation £30.

(a) What proportion of shoppers with large trolleys spend £100 or more?

(b) 8 people are seen entering the supermarket with large trolleys. What is the probability that fewer than half of them will spend £100 or more?

(c) Calculate, to the nearest 10p, the sum v such that only $\frac{1}{2}\%$ of shoppers with large trolleys spend more than v.

(d) Use a suitable approximation to evaluate the probability that at least three shoppers out of the first 300 who enter the supermarket with large trolleys will spend more than v.

(e) 2% of all the money spent by the shoppers in the supermarket is retained as profit by the supermarket's owners. If 12,000 shoppers visit the supermarket in one day, calculate the expectation and standard deviation of the owners' profit for the day.

[13 marks]

(ii) Phoebe is writing Christmas cards to her 100 friends. Half her friends live the in UK, so she needs a 62p stamp on the envelope, a fifth of her friends live in other EU countries, and she needs an 85p stamp for these, and the rest live outside the EU, so the stamp required is 105p.

(a) A visitor to Phoebe's room picks up a Christmas card at random. Calculate the expectation, μ_X, and the variance, σ_X^2, of X, the cost of the stamp required to post the card.

(b) Once Phoebe has finished writing her cards and putting them in envelopes, she thinks to herself, "The total amount I need to cover postage has a mean of $100\mu_X$ and variance $100\sigma_X^2$." What mistake is she making? What are the correct figures for the expectation and variance of the total postage?

[7 marks]

[Total: 20 marks]

QUESTION 9

(i) (a) Define what is meant by a *simple random sample.*

(b) Consider the following series of yearly minimum temperatures in Celsius recorded in London over 20 years

$$-0.52,\ 3.02,\ 6.00,\ 2.50,\ 2.97,\ 4.68,\ 8.00,\ 0.31,\ -2.93,\ 2.46,$$

$$4.71,\ 3.03,\ 7.60,\ 1.68,\ -5.92,\ 1.02,\ 2.47,\ 0.96,\ 6.13,\ -3.19$$

Discuss whether this can be considered a simple random sample.

(c) It is assumed that these observations come from a normal distribution with mean μ and variance σ^2. Denoting by x_i the minimum temperature in year i, it is known that $\sum_i x_i = 44.98$ and $\sum x_i^2 = 343.91$. Estimate μ and σ^2 and construct 99% confidence intervals for these parameters.

[8 marks]

(ii) An insurer sells 2 insurance contracts providing protection against events related to the minimum temperatures recorded in Celsius in London and Edinburgh in a given year, denoted respectively by X and Y. The model the insurer uses assumes that X and Y are normally distributed with

$$\mu_X = -1,\ \sigma_X^2 = 16,\ \mu_Y = -5,\ \sigma_Y^2 = 36,\ \rho_{X,Y} = \rho.$$

The contracts make payments based on an index computed as follows

$$Z = W - 1.3V,$$

where W and V are minimum temperatures in London and Edinburgh computed in Fahrenheit. The relation between Celsius and Fahrenheit is

$$\text{Fahrenheit} = 32 + 1.8 \times \text{Celsius}.$$

(a) Compute the distribution of Z in terms of ρ. What is the possible range of values for $\mathrm{Var}(Z)$? Is there a value of ρ such that Z and W are independent?

(b) Assume in the rest of the question that $\rho = 0.7$.

The first contract pays £1 million if the index Z is greater than 10, and nothing otherwise.

The second contract pays £1 million if the index Z is greater than 10 but less than or equal to 15, £2 millions if the index is greater than 15 and pays nothing otherwise.

Find the joint distribution of the random payments under the two contracts .

(c) What is the expected payment on the 2nd contract conditional on the payment on the 1st contract?

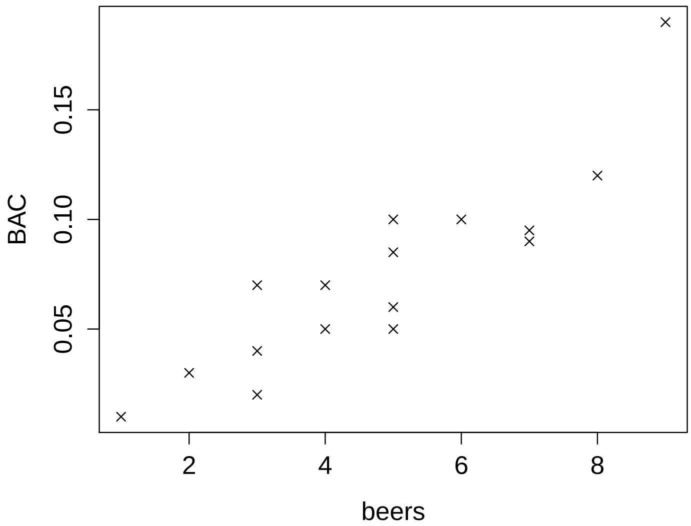
[12 marks]

[Total: 20 marks]

QUESTION 10

(i) State the assumptions underlying a linear regression model, derive and solve the normal equations. **[8 marks]**

(ii) In order to assess the effect of alcohol on people, each student in a group of 16 is assigned a random number of cans of beer. Thirty minutes after drinking, their BAC (blood alcohol content, an indicator of the percentage of alcohol in the blood) is recorded. The results are presented in the following graph.



Denoting x_i the number of cans assigned to student i and y_i the corresponding BAC, it is given that $\sum x_i = 77$, $\sum x_i^2 = 443$, $\sum y_i = 1.18$, $\sum y_i^2 = 0.11625$, $\sum x_i y_i = 6.98$.

(a) A linear model of BAC against number of beer cans is considered. Calculate the estimates of the intercept and slope.

(b) Compute the variance of the estimator of the slope and check at the 1% level of significance whether there is evidence that the number of beers has an effect on the blood alcohol content.

(c) Compute the coefficient of determination R^2.

(d) A student thinks he can drive legally 30 minutes after he drinks 5 beers in a country where a driver's blood alcohol content should not be higher that 0.08. Predict the student's BAC after this number of beers and compute a 95% prediction interval for the value which would be observed if the measurement were carried out. Comment on the result.

(e) The students in the sample differ by sex, weight and drinking habit. Which of the assumptions stated in (i) are most affected by this heterogeneity?

[12 marks]

[Total: 20 marks]