M2492006F1PV1

M249/J

Module Examination 2020

Practical Modern Statistics

Friday 12 June 2020 **10.00 am – 1.00 pm**

Time allowed: 3 hours

There are **four sections** in this examination, each worth 25% of the total mark.

In **each section** you should **attempt all questions**. You are advised to spend about 45 minutes on each section.

Include all your working, as some marks are awarded for this.

Write your answers in the spaces provided in this question paper in **pen**, though you may draw diagrams in pencil.

Crossed out work will not be marked.

You may request an answer book for further work, if needed.

Please fill in the grid below to show your Personal Identifier and Examination number, taken from your desk record.

Personal Identifier								
Examination No.								
Calculator Model								

For Examiner's use only:

Section	1	2	3	4	Total
Marks					

Section 1 (relates to Book 1 *Medical statistics*)

Questions 1 to 6

You should **attempt all questions**. ***This section is worth* 25%.**

Write your answers in the spaces provided.

Question 1

A study was undertaken to investigate whether an elevated level of an inflammatory marker is associated with an increased risk of coronary heart disease. Blood samples were taken from a sample of 18 225 men and their levels of the inflammatory marker were determined. Over the following three years, 266 of the men had a heart attack and their levels of the inflammatory markers were compared with those of a subset of the other men. That subset was chosen to match the men with a heart attack in terms of age and smoking status, with two men without a heart attack selected for each man who had suffered a heart attack.

(a) State whether this is a cohort or case-control study, and give a reason for your answer. [*2*]

(b) In the study, each man who had suffered a heart attack was matched for age and smoking status with two men who had not. What bias could this remove? Explain briefly how this bias might have arisen without the matching. [*2*]

Question 2

A potential side effect of a medicine is an upset stomach. In a trial, 56 volunteers received a new medicine and a further 85 volunteers received a placebo. The numbers having an upset stomach are shown in Table 1.

Table 1

	Upset Stomach		
	Yes	No	Total
New medicine	14	42	56
Placebo	2	83	85
Total	16	125	141

(a) Calculate the relative risk of an upset stomach for people taking the new medicine, relative to those taking the placebo. [1]

(b) Obtain a 95% confidence interval for the relative risk. [2]

(c) Summarize and interpret your results. [2]

Question 3

A study of association between obesity and cardiovascular disease (CVD) was undertaken. The data were split into two strata according to age (Stratum A: age < 50 years; Stratum B: age ≥ 50 years). The stratified data are shown in Table 2.

Table 2

A: Age under 50

Exposure category	CVD	No CVD
Obese	10	90
Not obese	35	465

B: Age 50 or more

Exposure category	CVD	No CVD
Obese	36	164
Not obese	25	175

(a) Obtain the stratum specific odds ratios for the two groups and the odds ratio for the aggregated data. [*2*]

(b) Calculate the Mantel–Haenszel odds ratio [*2*]

(c) Comment on the association between obesity and CVD and the role played by age in this association. [*2*]

Question 4

A case-control study was conducted to examine whether there is an association between dietary magnesium intake and risk of colorectal tumors. The intake level was divided into three ranges that each contained similar numbers of controls. The dose-specific odds ratios relative to individuals in the lowest range are shown in Table 3.

Table 3

Magnesium intake	Cases	Controls	Odds ratio
Under 316	270	236	1.00
316 to 358	267	236	0.90
Over 358	231	237	0.73

(a) Briefly describe the relationship between magnesium intake and colorectal tumors. [*1*]

(b) The chi-squared test for no linear trend gives the value 1.523 for the test statistic. What do you conclude, and why? [*2*]

Question 5

A randomized controlled trial is planned to investigate whether a new drug for treating arthritis is better than a standard treatment. The investigators are interested in the proportion of arthritis sufferers who have reduced symptoms after three weeks. The sample size for the trial was based on the following values at significance level $\alpha = 0.01$, power $\gamma = 0.85$ and design values $\pi_T = 0.45$, $\pi_C = 0.30$.

(a) What is the probability that the trial will fail to demonstrate a true effect? [*1*]

(b) What is the probability of improvement under the standard treatment assumed to be in this trial? [*1*]

(c) In the trial, half the patients will be allocated to the standard treatment and half to the new drug. What is the required sample size per group for this trial? [*3*]

Question 6

Several randomized controlled trials have been conducted to assess the effectiveness of propanolol (a beta blocker) at reducing mortality in patients who have suffered a heart attack. Eight studies with similar methodologies were combined in a meta-analysis and the resulting forest plot is shown in Figure 1. With each study an odds ratio of less than 1 indicates that propanolol reduced mortality.

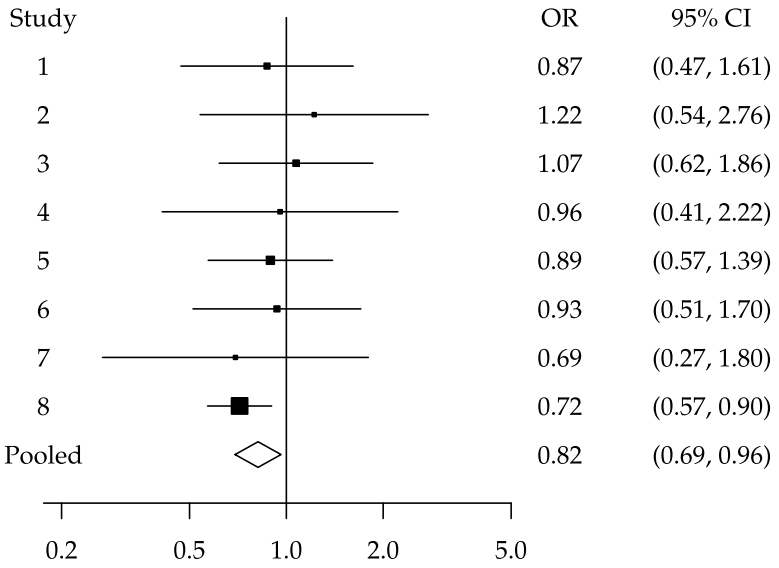


Figure 1

(a) Which study contributed most to the pooled odds ratio? Give a reason for your answer. [1]

(b) Interpret the pooled estimate of the odds ratio and its 95% confidence interval. [1]

For Examiner's use only:

Question No.	1	2	3	4	5	6	Total for Part 1
Mark							

Section 2 (relates to Book 2 *Time Series*)

Questions 7 to 14

You should **attempt all questions**. ***This section is worth* 25%.**

Write your answers in the spaces provided.

Question 7

Figure 2 shows the quarterly time series plot of the number of visits abroad by UK residents between 2015 and 2018.

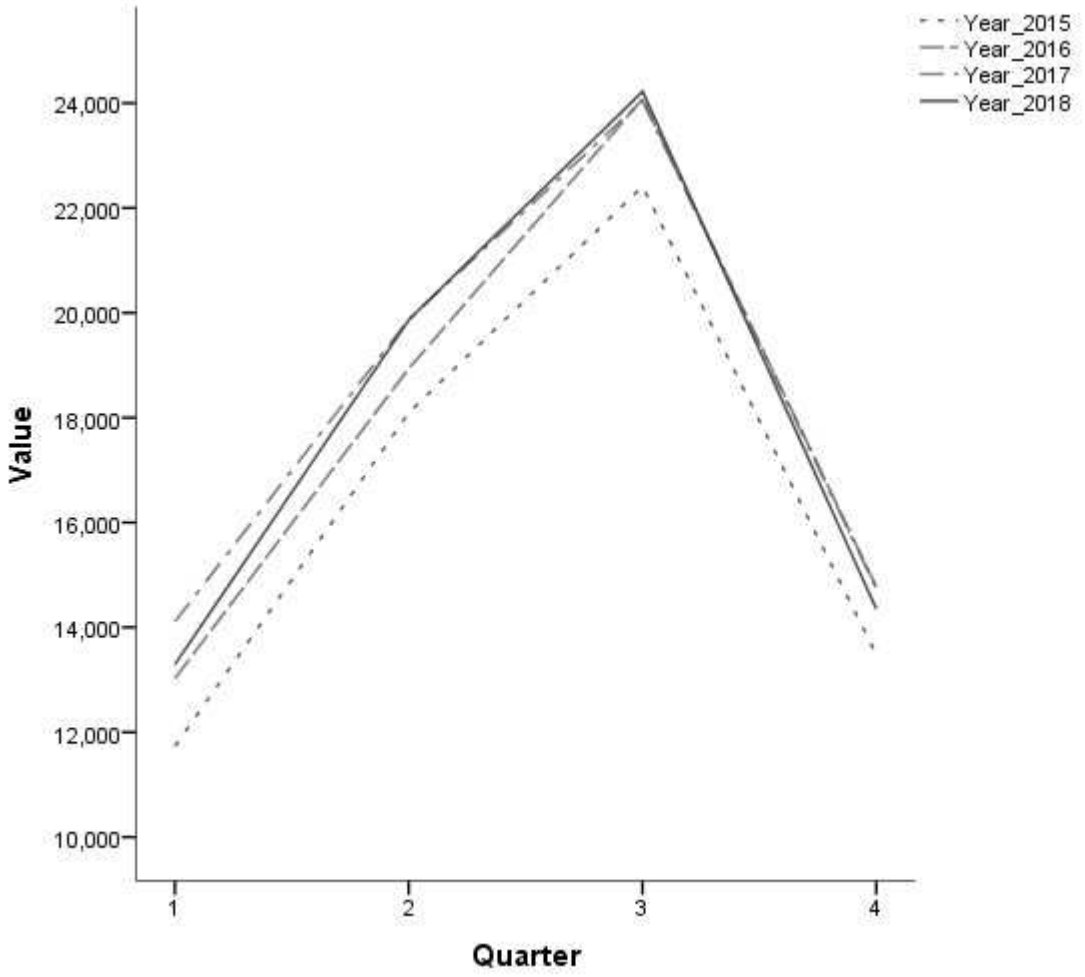


Figure 2

(a) Describe the seasonal variation in this time series. [*2*]

(b) Is there any evidence of an increasing or decreasing trend from year to year in the period 2015–2018? Explain your answer. [*2*]

Question 8

Figure 3 shows a time plot of the monthly UK consumer price index (CPI) inflation rate figures from January 2015 until December 2018.

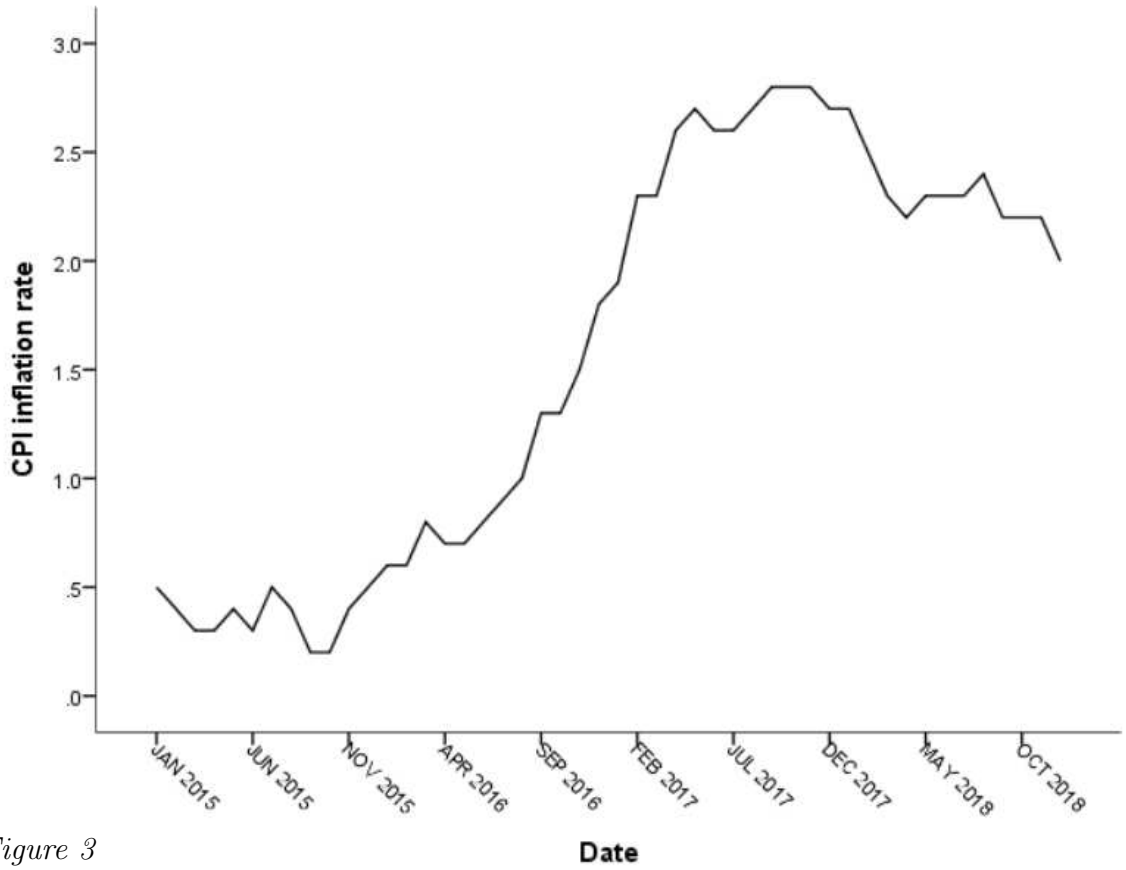


Figure 3

State which exponential smoothing model should be used for forecasting the CPI inflation rate: simple, Holt's or Holt-Winters. Explain why your chosen model is appropriate, and why each of the other two models is not appropriate. [*2*]

Question 9

Consider the quarterly seasonal time series given in Figure 4.

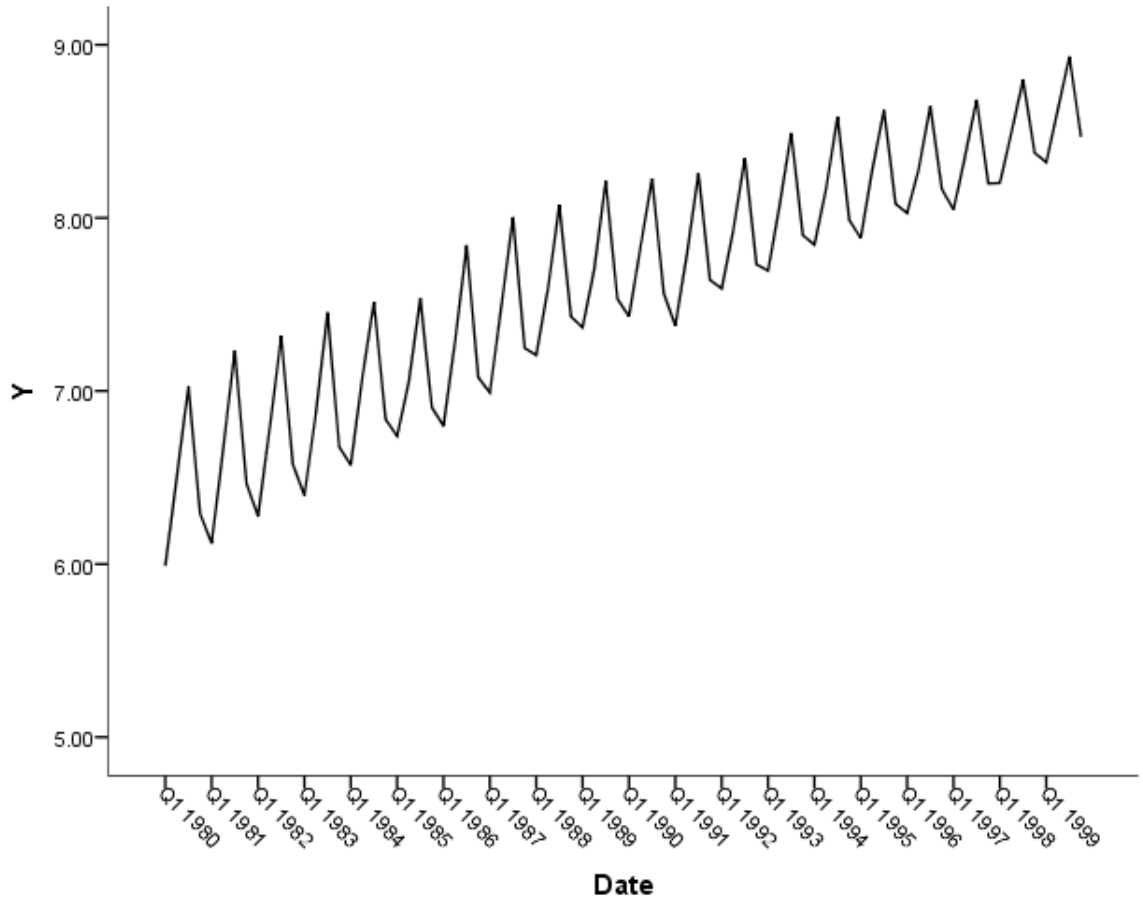


Figure 4

Which model is more appropriate for the time series in Figure 4, an additive decomposition model or a multiplicative decomposition model? Explain your answer. [*1*]

Question 10

The estimated seasonal factors for the monthly series of expenditure abroad (in £ millions) by UK residents from January 2012 to December 2017 are shown in Table 4.

Table 4

Month	Seasonal factor	Month	Seasonal factor
1	−452.34	7	730.26
2	−722.09	8	?
3	−468.19	9	370.99
4	−139.44	10	42.82
5	58.64	11	−285.49
6	243.62	12	−116.94

(a) Obtain the seasonal factor for August (Month 8). [*1*]

(b) The expenditure figure for August 2017 was £3014.00 millions. Calculate the seasonally adjusted expenditure figure for that month. [*1*]

Question 11

The optimal values of the smoothing parameters α, γ and δ obtained with the Holt–Winters exponential smoothing method applied to a monthly time series were $\alpha = 0.32$, $\gamma = 0.71$ and $\delta = 0.01$. Interpret these optimal parameters in terms of the weight given to recent observations. [*2*]

Question 12

A monthly time series has 36 observed values from January 2016 to December 2018. The first two values of the log transformed series are 7.98 in January 2016, and 7.93 in February 2016.

(a) Assuming a Holt's exponential smoothing method is used for the log transformed time series, suggest appropriate initial values for the level and the slope. [*2*]

(b) Table 5 shows the values of the SSE obtained for several pairs of values of the parameters α and γ, and the corresponding forecasts for January 2019.

Table 5

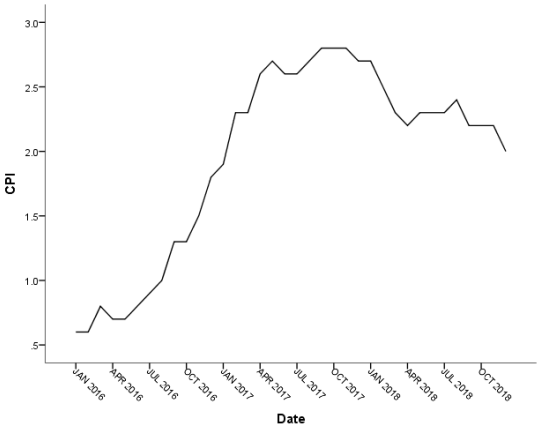
Model	α	γ	SSE	Forecast
A	0.39	0.001	2.609	7.83
B	0.39	0.005	2.605	7.81
C	0.40	0.001	2.601	7.79
D	0.40	0.005	2.602	7.80
E	0.41	0.001	2.676	7.82
F	0.41	0.005	2.763	7.85
G	0.50	0.01	2.802	7.73

(i) Identify the optimal combination of parameter values among those listed in Table 5. Briefly explain your choice. [*2*]

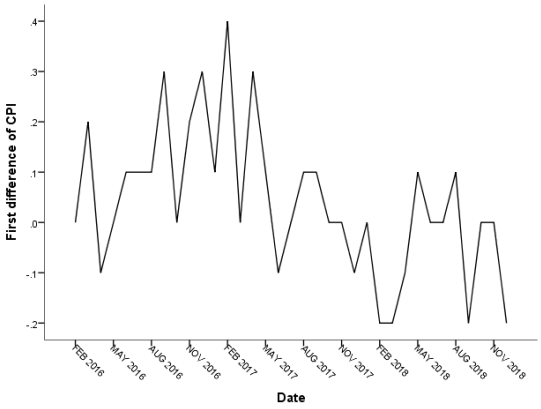
(ii) Obtain the forecasted value of the original series according to the optimal Holt's model for January 2019, rounded to two decimal places. [*2*]

Question 13

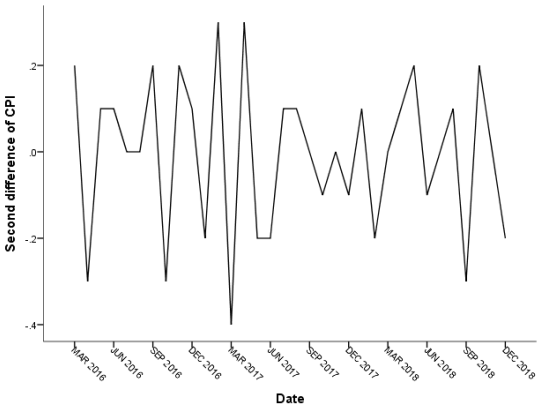
Figure 5(a) shows the time plot of the UK consumer price index (CPI) inflation rate from January 2016 to December 2018. Figures 5(b) and 5(c) show time plots of the first and second differences, respectively.



(a)



(b)



(c)

Figure 5

Identify the order of differencing required to obtain stationarity in mean and variance. Carefully explain your answer. [*2*]

Question 14

Figure 6 shows the correlogram (ACF) and the partial autocorrelogram (PACF) of a stationary time series at lags 1 to 20.

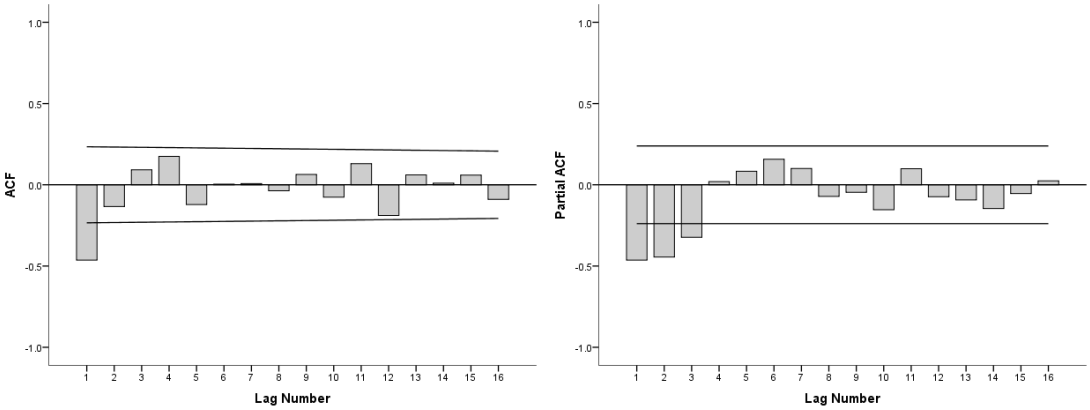


Figure 6

(a) Identify one plausible AR(p) and one plausible MA(q) model for this time series. Explain carefully the reasons for your choices. [*4*]

(b) Would a mixed ARMA model be a plausible model for this time series? Briefly explain your answer. [2]

For Examiner's use only:

Question No.	7	8	9	10	11	12	13	14	Total for Part 2
Mark									

Section 3 (relates to Book 3 *Multivariate analysis*)

Questions 15 to 20

You should **attempt all questions**. ***This section is worth* 25%.**

Write your answers in the spaces provided.

Question 15

Table 6 shows data on:
X_1: a baby's intake of breast milk (ml/24 hours)
X_2: the amount of milk substitute given to the baby (ml/24 hours)
X_3: the weight of the baby (kg)
X_4: the weight (kg) of the baby's mother
X_5: the height (cm) of the baby's mother.

Table 6

X_1 Milk intake (dl/24 hrs)	X_2 Supplement (dl/24 hrs)	X_3 Weight of infant (kg)	X_4 Weight of mother (kg)	X_5 Height of mother (cm)
842	250	5.002	65	173
844	0	5.128	48	158
841	40	5.445	62	160
770	0	5.667	63	165
965	60	5.106	55	162
644	240	5.196	58	170
629	0	5.526	56	153
979	30	5.928	78	175
843	0	5.263	57	170
805	230	6.578	57	168
648	555	5.588	58	173
764	0	4.613	58	171
491	0	4.360	49	162
873	60	5.882	54	163
771	315	5.618	68	179
839	370	6.032	56	175
932	130	6.030	62	168
678	0	4.727	59	172

Suppose that the data matrix based on Table 6 is denoted by $\mathbf{X}$.

(a) What are the values of x_{34} and x_{51}? [*1*]

(b) Should the data in Table 6 be standardized before carrying out a principal component analysis? Give two reasons why, or why not. [*2*]

(c) The mothers' heights in the data set have a mean of 167.61cm and a standard deviation of 6.84 cm. Calculate the standardized value of mother's height for the first observation in the data set. [*1*]

Question 16

The lower triangle of the sample correlation matrix from the data in Table 6 is:

$$\begin{array}{c} \\ X_2 \\ X_3 \\ X_4 \\ X_5 \end{array} \begin{array}{c} \begin{array}{cccc} X_1 & X_2 & X_3 & X_4 \end{array} \\ \begin{pmatrix} -0.101 & & & \\ 0.464 & 0.369 & & \\ 0.404 & 0.125 & 0.324 & \\ 0.143 & 0.562 & 0.140 & 0.570 \end{pmatrix} \end{array}$$

(a) Between which pair of variables does there appear to be the strongest linear relationship. Between which pair of variables does there appear to be the weakest relationship? [*1*]

(b) Use this correlation matrix to describe the relationship between the variables. [*2*]

Question 17

A principal component analysis was carried out on the standardized values of the variables presented in Table 6 of Question 15. The variance explained by each of the principal components (with one value missing) is given in Table 7, and the corresponding scree plot is given in Figure 7.

Table 7

Principal Component	Variance explained
1	2.230
2	1.290
3	
4	0.410
5	0.164

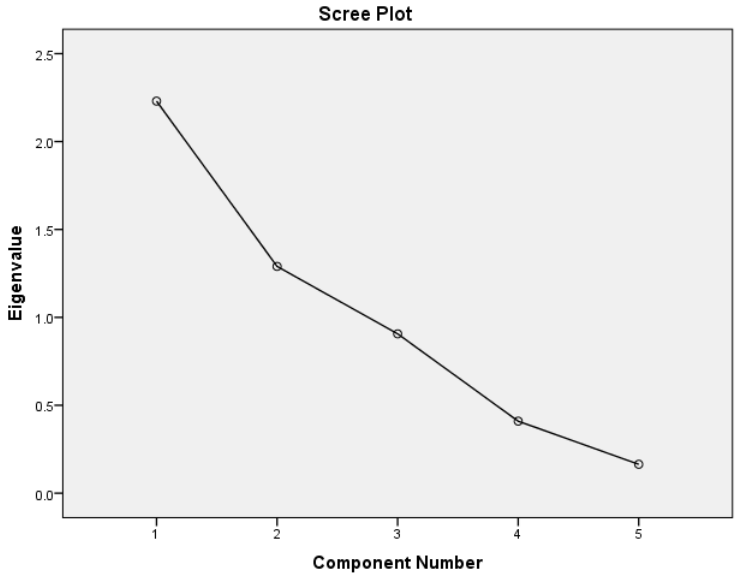


Figure 7

Use Table 6 and Figure 7 to answer the following questions.

(a) Calculate, to 3 decimal places, the variance explained by the third principal component. [*2*]

(b) Using Kaiser's criterion, how many components should be retained? Explain your answer. [*2*]

(c) Can you identify how many components should be retained using the scree plot. Explain your answer. [*2*]

(d) Calculate the percentage variance explained by the first two components. [*1*]

Question 18

Data were collected from 50 people on the value they placed on four types of precious stone. The data were standardised and then a principal component analysis carried out using SPSS. Table 8 gives the loadings for the first two principal components produced by SPSS.

Table 8

Variable	Component 1	Component 2
Diamond	0.451	−0.631
Ruby	0.578	−0.082
Sapphire	0.537	0.121
Emerald	0.417	0.762

Briefly interpret each of the components. [*4*]

Question 19

Three variables Y_1, Y_2 and Y_3 were recorded for a set of observations divided into four groups. The within-groups covariance matrix $\mathbf{W}$ and the between-groups covariance matrix $\mathbf{B}$ calculated using these data are as follows.

$$\mathbf{W} = \begin{pmatrix} 21.54 & 7.54 & 13.62 \\ 7.54 & 9.38 & 4.49 \\ 13.62 & 4.49 & 13.62 \end{pmatrix}$$

$$\mathbf{B} = \begin{pmatrix} 25.69 & -8.11 & 67.17 \\ -8.11 & 4.61 & -23.26 \\ 67.17 & -23.26 & 177.68 \end{pmatrix}$$

(a) Calculate the separations achieved by each of the three variables on its own, and hence identify the variable that achieves the best separation between the four groups. [*2*]

(b) Can the separation be improved substantially by using a linear combination of more than one variable? Explain your answer. [*2*]

Question 20

A data set consists of five variables recorded for 244 people whose work was classified as *personnel*, *mechanics* or *dispatchers*.

(a) What is the maximum number of discriminant functions that are useful for discriminating between the three classifications using the five variables? [*1*]

(b) Stacked histograms of the first discriminant function are shown in Figure 8 (top: *personnel*, middle: *mechanics*, bottom: *dispatchers*).

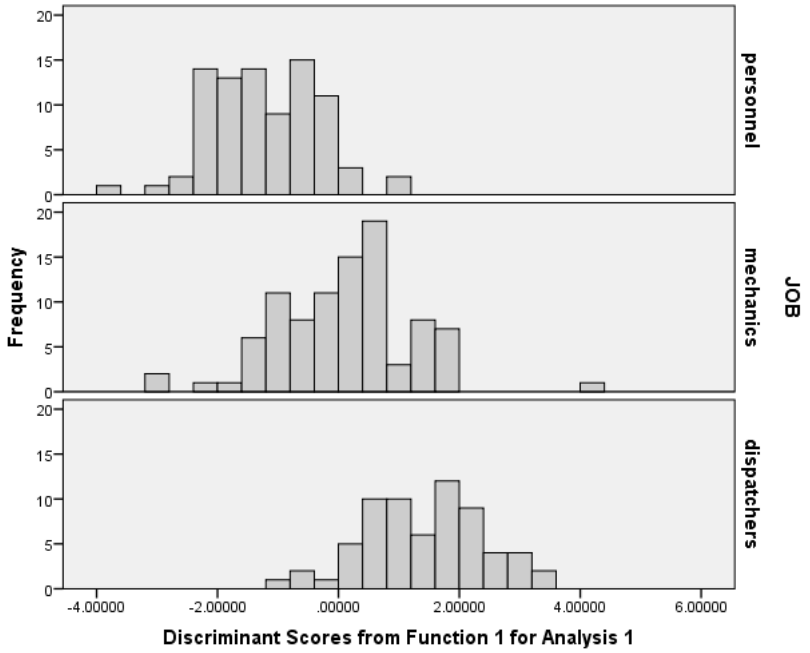


Figure 8

Briefly comment on the usefulness of this discriminant function for distinguishing between the three types of work. [*2*]

For Examiner's use only:

Question No.	15	16	17	18	19	20	Total for Part 3
Mark							

Section 4 (relates to Book 4 *Bayesian statistics*)

Questions 21 to 27

You should **attempt all questions**. ***This section is worth* 25%.**

Write your answers in the spaces provided.

Question 21

Three different machines were used for producing a large lot of similar manufactured items. Twenty percent of the items were produced by machine A, 30% by machine B and 50% by machine C. Suppose that 3% of items produced by machine A are defective, 4% produced by machine B are defective and 2% produced by machine C are defective.

(a) What proportion of items in the complete lot are defective? [*1*]

Suppose an item is randomly selected from the complete lot.

(b) What is the probability it was produced by machine A and is a defective item? [*1*]

(c) If the item is defective, what is the probability it was produced by machine A? [*2*]

Question 22

Figure 9 contains plots of four prior distributions.

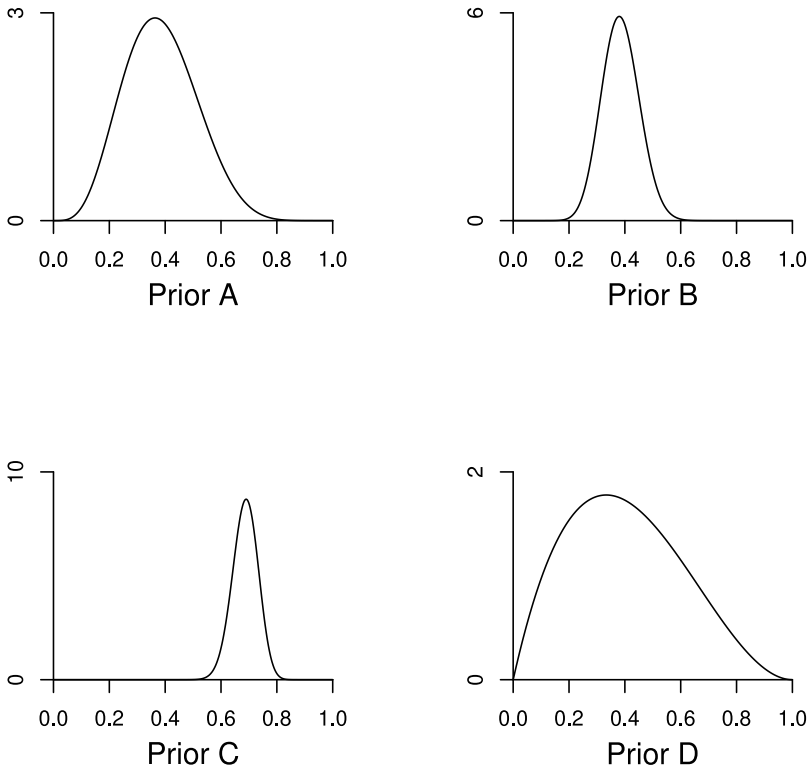


Figure 9

Place these prior distributions in order from weakest to strongest.
Justify your answer. [*2*]

Question 23

Pat is about to take up a new job in which she will sometimes have to go on emergency calls. Suppose that the number of calls she will have to answer in a week follows a Poisson distribution and let θ be the average number per week.

(a) Pat decides to represent her opinion about θ by a gamma prior distribution, Gamma(a, b). Give one reason why a gamma prior is an appropriate choice. [*1*]

(b) Pat chooses $a = 4$ for the first parameter of the gamma prior for θ. What value should she choose for b if she expects to make 1.5 calls per week, on average? [*1*]

(c) In her first five weeks, Pat handles 2, 1, 3, 2 and 4 calls. State the posterior for θ, given the data and Pat's prior. [*1*]

(d) Discuss how the numbers of calls in the first five weeks have changed Pat's beliefs. [*2*]

Question 24

Suppose that data X can be modelled by the normal distribution $N(\mu, 30)$. Jim's prior for μ has median 100. The interval $(97, 103)$ gives the central 50% of his prior.

(a) Calculate the values of the parameters a and b of the Normal prior $N(a, b)$ that best matches Jim's beliefs about μ. [*3*]

(b) Three observations on X are made and their sample mean is 102.0. Using the prior you specified in part (a), calculate the posterior for μ. [*3*]

(c) Janet's prior for μ has mean 101, and its lower and upper quartiles are 99 and 110. Give one reason why Janet's beliefs cannot adequately be represented by a normal prior. [*1*]

Question 25

Write down the model corresponding to the following WinBUGS model definition.

```
model
{
    x ~ dbin (p,n)
    p ~ dbeta (a,b)
    n  <-  12
    a  <-  5
    b  <-  8
}
```

[2]

Question 26

Samples from the posterior distributions of parameters α and β were obtained using WinBUGS. Using these samples, the following numerical summaries were produced.

Table 9

node	mean	sd	MC error	2.5%	median	97.5%	start	sample
alpha	-60.8	5.137	0.2208	-70.6	-60.87	-50.92	1	1000
beta	34.31	2.888	0.1252	28.72	34.35	39.93	1	1000

(a) Write down an estimate of the posterior mean and 95% equal-tailed credible interval for β. [*1*]

(b) Describe, in general terms, what would happen to the values of sd and MC error if the number of iterations were to increase substantially. Explain your answer. [*2*]

Question 27

Using WinBUGS, a chain of 1000 iterations was run so as to sample from the posterior for a parameter β. The resulting trace plot for β is given in Figure 10.

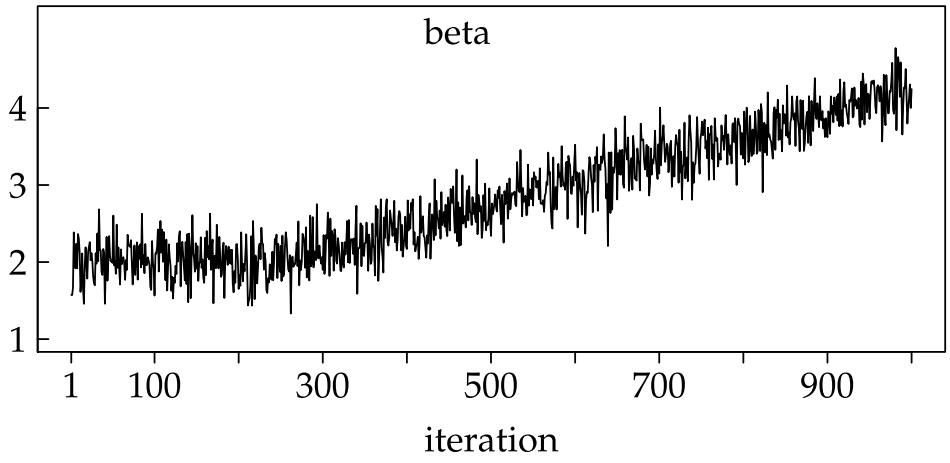


Figure 10

Do you think the chain converged? Justify your answer. [*2*]

For Examiner's use only:

Question No.	21	22	23	24	25	26	27	Total for Part 3
Mark								

[END OF QUESTION PAPER]