The Open University

M246/W

Second Level Course Examination 1998

Elements of Statistics

Friday, 23 October, 1998 10.00 am – 1.00 pm

Time allowed: 3 hours

This examination consists of 33 questions of varying length: the mark allocation for each question is shown on the paper. The marks add up to 120, so you do not need to answer all the questions to obtain full marks, but you should answer as many questions as you can.

The examination paper is divided into five parts, all worth roughly the same number of marks. You are advised to spend about 35 minutes on each part. Note the division into parts has been done only to help you in allocating your time in the examination.

At the end of the examination:
Check that you have completed the grid below, and have written your personal identifier and examination number on any supplementary answer book used.
Failure to do so will mean that your work cannot be identified.

Examination No.								
Personal Identifier								

If you have used a supplementary answer book, attach this examination paper to the *front* of it, using the paper fastener provided.

For Examiner's use only:

I	II	III	IV	V	Total

PART I

Questions 1 to 2 carry a total of 23 marks.

Question 1

The following tables are taken from the Annual Review of Government Funded Research and Development 1990. They show Government expenditure (in £million) on Research and Development in broad groups of manufacturing industry over the 10 year period from 1978–1988. Table 1 gives the raw figures, Table 2 gives the figures adjusted to reflect price changes over the period.

Table 1: Unadjusted expenditure on R & D

Year	1978	1981	1983	1985	1986	1987	1988
Chemicals	394.1	617.4	735.0	941.9	1037.9	1303.0	1573.5
Mechanical engineering	174.1	234.0	249.6	262.6	268.4	285.6	301.0
Electronics	656.6	1235.3	1473.9	1758.6	1949.7	1854.7	2063.6
Electrical engineering	100.8	120.8	117.7	125.6	152.5	142.2	147.0
Motor vehicles	129.7	180.4	239.5	371.6	394.2	450.5	468.3
Aerospace	424.6	762.9	720.0	818.0	829.6	870.9	814.0
Other manufactured products	332.6	361.1	334.3	394.9	438.1	462.8	504.8
Total	2212.5	3511.9	3870.0	4673.2	5070.4	5369.7	5872.2

Table 2: Adjusted expenditure on R & D

Year	1978	1981	1983	1985	1986	1987	1988
Chemicals	752.7	772.5	812.9	941.9	1002.0	1200.9	1362.3
Mechanical engineering	332.5	292.8	276.0	262.6	259.1	263.2	260.6
Electronics	1254.0	1545.6	1630.1	1758.6	1882.3	1709.4	1786.7
Electrical engineering	192.5	151.1	130.2	125.6	147.2	131.1	127.3
Motor vehicles	247.7	225.7	264.9	371.6	380.6	415.2	405.5
Aerospace	810.9	954.6	796.3	818.0	800.9	802.7	704.8
Other manufactured products	635.2	451.8	369.7	394.9	423.0	426.5	427.5
Total	4225.5	4394.1	4280.1	4673.2	4895.1	4949.0	5074.7

(a) The figures in Table 2 have been adjusted to correspond to prices in a particular year. Which year has been used?

The adjustment has been done by means of multiplicative weighting factors: for example multiplication of each unadjusted expenditure in the 1978 column of Table 1 by a factor of 1.910 gives the figures in the 1978 column of Table 2. Calculate the weighting factors for other years using the figures for the *Chemicals* category. Look specifically at the category *Other manufactured products* for the year 1988: is there any reason to suspect the figures quoted? [5]

(b) Draw a rough diagram to illustrate the pattern of change in weight factors over the period 1978–88. Comment briefly on what your diagram shows in terms of the value of the pound. What is the link between the weights and price changes over the period? Draw a graph which shows directly how prices have changed and comment on how prices have changed. [5]

(c) Draw a rough diagram to illustrate the patterns of funding in the areas which could broadly be described as "engineering related" (*i.e.* mechanical engineering, electrical engineering, motor vehicles and aerospace). Comment on the comparisons between patterns. [4]

(d) The table below contains the unadjusted figures for Government expenditure on research and development for non-manufactured products for the period 1978–88.

Table 3: Unadjusted expenditure on R & D

Year	1978	1981	1983	1985	1986	1987	1988
Non-manufactured products	111.7	280.8	293.5	448.4	880.3	963.6	988.6

Adjust these figures so that they are comparable across the period 1978–88. Comment upon the pattern of expenditure on research and development for non-manufactured products, paying particular attention to their comparison with Government expenditure on research and development in the manufacturing industries. [6]

Question 2

Find the sample mean, $\bar{x}$, and the sample variance, s^2, for the following data.

1 3 3 5 9 9 [3]

PART II

Questions 9 to 11 carry a total of 24 marks.

Question 3

Find the five-figure summary and calculate the interquartile range for the following sorted data sample.

2.6	3.1	3.5	4.5	4.6	4.8	5.0	5.2
5.6	6.1	6.4	7.0	7.1	7.1	7.5	7.7
8.0	8.1	8.2	8.4	8.6	9.3	9.5	9.9
10.4	10.5	10.7	10.8	11.2	11.3	11.4	11.6
11.6	11.7	11.9	12.0	12.6	13.4	15.7	19.2

[4]

Question 4

What are the two main advantages of bar charts over pie charts? [2]

Question 5

What are the properties of a probability density function? [2]

Question 6

If the random variable X is Triangular(3), calculate the probability $P(X \geq 1)$. [2]

Question 7

Calculate the mean and variance of the discrete probability distribution given below.

x	0	1	2	3
$p(x)$	0.375	0.375	0.125	0.125

[3]

Question 8

The joint probability mass function $p_{X,Y}(x, y)$ of the random variables X and Y is given in the following table.

		x		
		0	1	2
y	−1	0.025	0.05	0.025
	1	0.225	0.45	0.225

Are X and Y independent random variables? What leads you to your answer? [4]

Question 9

The probability of NOT winning any money if you enter a single set of numbers in the British National Lottery is 0.9814 (to four decimal places). The random variable X is the number of draws up to and including the first win of someone who enters a single set of numbers in each draw. What probability distribution is an appropriate model for X? Find the mean and variance of X. [3]

Question 10

The c.d.f. of the exponential distribution with parameter λ is given by

$$F(x) = P(X \leq x) = 1 - e^{-\lambda x}, \quad x \geq 0,$$

where $\lambda > 0$. Calculate the upper quartile of the exponential distribution with parameter $\lambda = 2$. [3]

Question 11

What probability distribution could reasonably be used to model the total number of faulty silicon chips in a sample of 200, where the probability that a chip is faulty is 0.002 independently of the other chips? [1]

PART III

Questions 12 to 18 carry a total of 23 marks.

Question 12

Cars arrive at a petrol station at a rate of two per minute. You may assume that the car arrivals are independent of one another and occur randomly during the day.

Give the distribution of each of the random variables defined below:

(a) X: the number of arrivals between 6.00 pm and 6.45 pm;

(b) Y: the time interval in minutes between successive arrivals. [3]

Question 13

The independent random variables X and Y have exponential distributions with means of 3 and 1, respectively. What are the mean and the variance of their difference, $X - Y$? [3]

Question 14

Suggest from your own experience a continuous random variable in which the variation observed is such that it would NOT be well-modelled by a normal distribution, and say why. [3]

Question 15

Suppose that the variation in heights of eleven-year-old girls in Scotland can be adequately modelled by a normal distribution with mean 54.9 inches and variance 12.25. Find the proportion of eleven-year-old Scottish girls who are at least 56 inches tall. [4]

Question 16

Using results from your *Handbook* it is possible to write down the mean, μ_X, and the variance, σ_X^2, for the number of rolls, X, of a fair six-sided die up to and including the first six, as $\mu_X = 6$ and $\sigma_X^2 = 30$.

(a) Show that the total number of rolls, T, of a fair six-sided die up to and including the 25th six has mean $\mu_T = 150$ and variance $\sigma_T^2 = 750$.

(b) Use normal tables to find an approximation to the probability that the number of rolls up to and including the 25th six is less than 100. [5]

Question 17

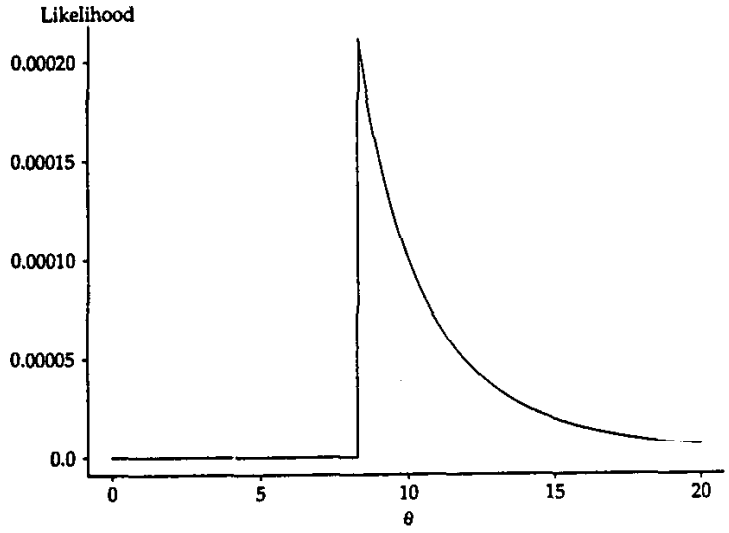
The following sample of size three was drawn from a population with a geometric distribution with parameter p.

3 1 9

Write down the likelihood of p for this sample. [3]

Question 18

The likelihood of θ for a particular sample of size four from a continuous uniform distribution $U(0, \theta)$ is shown below.



Mark the diagram to indicate the position of the maximum likelihood estimate of θ. Write down the approximate value of the maximum likelihood estimate of θ based on this diagram. [2]

PART IV

Questions 19 to 26 carry a total of 25 marks.

Question 19

Given that it is possible to derive confidence intervals based on just one observation, give two reasons why, in practice, samples of sizes much larger than one are normally taken to obtain confidence intervals. [2]

Question 20

A dataset consisting of 2608 counts of the numbers of particles given off in certain short time intervals by the radioactive decay of polonium has a sample mean of 3.87 and a sample variance of 3.70. Assuming that the Poisson distribution is a good model for these data, calculate an approximate 95% confidence interval for the mean count. [3]

Question 21

What information given in Question 20 supports the Poisson assumption made there? Nevertheless, one investigator of these data has suggested, on other grounds, that a Poisson assumption may not be appropriate. Obtain an approximate 95% confidence interval for the mean count for the polonium data of Question 20 if the Poisson distribution cannot be assumed. [3]

Question 22

The birth weights (in grams) of 15 babies with a certain condition were recorded. On the assumption that the resulting data come from a normal distribution, a 90% confidence interval for the mean birth weight of such babies is $(2898, 3501)$. Using this information, perform a test of the hypothesis that the population mean is 2800g. What can you say about the amount of evidence against the null hypothesis? [4]

Question 23

State the assumptions of the paired t-test for testing that a mean difference is zero. [2]

Question 24

The following is an extract from an SSC logfile.

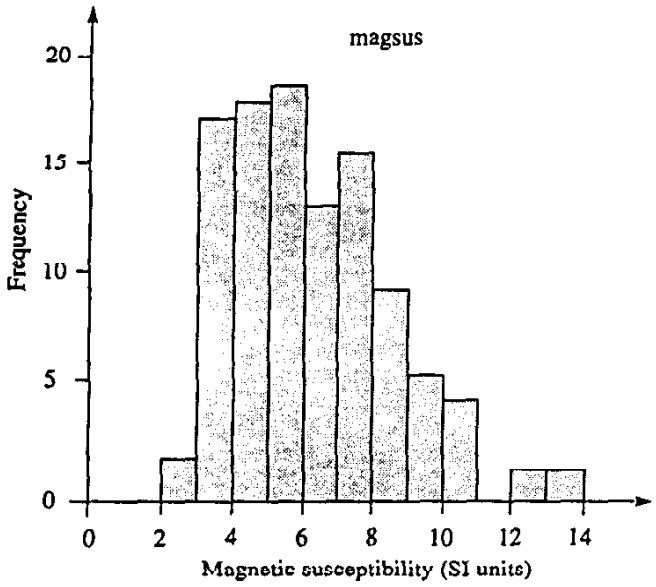
```
mean((control,stress))
  30.59      27.78
vare((control,stress))
  4.552    2.98
t2test(control,stress)
    t =  3.699     df = 24
    SP (obtained direction) = 0.0005622
    SP (opposite direction) = 0.0005622
    SP (total) = 0.001124
```

Describe the steps of the statistical procedure being carried out and state the conclusion that you would draw.

What is the sum of the sizes of the two samples involved? [5]

Question 25

The histogram below is based on a dataset comprising 101 measurements on the magnetic susceptibility of certain rocks of archaeological interest. What is the major feature of the appearance of this histogram that indicates that the normal distribution is not an appropriate model for the data as they stand? Suppose that you still hoped to be able to use a normal model in conjunction with analysing these data. Briefly describe how you might be able to do this.



[4]

Question 26

The Wilcoxon signed rank test is usually preferred to the sign test. Give two reasons for this (one reason may be a consequence of the other). [2]

PART V

Questions 27 to 30 carry a total of 25 marks.

Question 27

A dataset consists of the age at death (in years) and the length of the lifeline on their left hand (measured *post mortem*, in cm) for each of 50 individuals. The question of interest is whether the length of the lifeline is useful in predicting age at death. Restate this question in statistical terms. A 90% confidence interval for the slope of the regression line of age at death against length of lifeline is $(-4.05, 1.31)$. Use this information to give an answer to the question of interest. [3]

Question 28

Which is longer, a 95% confidence interval for the regression mean at x or a 95% prediction interval for the response at x? Give a reason for your answer. [2]

Question 29

Draw rough residual plots which reflect:

(a) a systematic discrepancy between the fitted and actual regression curves;

(b) a response variance increasing with the explanatory variable. [3]

Question 30

(a) The following extract is taken from the October 1997 issue of the *NFU* (National Farmers Union) *Newsletter*:

> "Out of a total of 1,911 tested [in 1997], 481 badgers were positive for TB. In 1995 out of a total of 1,509 examined, 424 were TB positive."

In the light of this, critically discuss the headline the extract appeared under:

> "TB Increase in Badgers".

(b) Describe the following SSC output, and what it contributes to the discussion above.

```
fisher(424,1085,481,1430)
    SP (obtained direction) = 0.0296
    SP (opposite direction) = 0.02624
    SP (total)              = 0.05583
```

[7]

Question 31

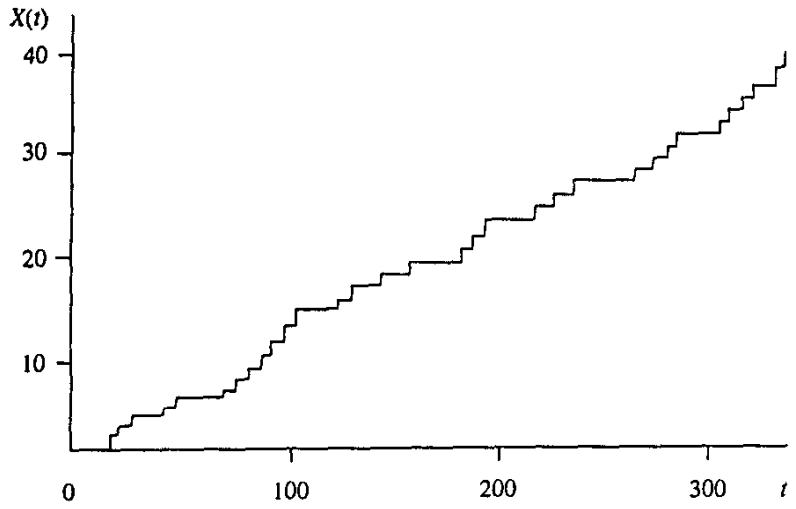
Draw rough scatterplots to illustrate:

(a) a sample of bivariate data with a very high Spearman correlation but not such a high Pearson correlation,

(b) a sample of bivariate data for which the Pearson correlation is close to zero but the variables have a strong dependence. [3]

Question 32

The figure below plots $X(t)$ against time, t.



Which of the following statements is true?

A: The process takes discrete values.

B: The process takes continuous values.

C: The process takes place in discrete time.

D: The process takes place in continuous time.

Give one type of random process which could result in a plot like this. [3]

Question 33

In a mood assessment exercise, patients in a psychiatric clinic were categorized daily at 5pm as 'happy today' or 'not happy today'. Over a 49-day period of observation, a patient was recorded as happy on 20 days and not happy for the remaining 29, with a matrix of transition frequencies given by

$$\begin{array}{r} \text{Not happy} \\ \text{Happy} \end{array} \begin{bmatrix} 21 & 8 \\ 8 & 11 \end{bmatrix}.$$

(a) Assuming a Markov chain model for the mood swings for this patient, use these data to estimate the mood transition probabilities.

(b) Over the period, the number of runs expected in a Bernoulli model is 24.67. The observed number of runs was 17, giving an SP of 0.025. What conclusions can you draw from these results? [4]