EXAMINATIONS OF THE HONG KONG STATISTICAL SOCIETY

ORDINARY CERTIFICATE IN STATISTICS, 2016

MODULE 2 : Analysis and presentation of data

Time allowed: Three hours

Candidates may attempt **all** *the questions.*

The number of marks allotted to each question or part-question is shown in brackets.

The total for the whole paper is 100.

A pass may be obtained by scoring at least 50 marks.

Graph paper and Official tables are provided.

Candidates may use calculators in accordance with the regulations published in the Society's "Guide to Examinations" (document Ex1).

This examination paper consists of 12 printed pages.
This front cover is page 1.
Question 1 starts on page 2.

There are 9 questions altogether in the paper.

1. A newspaper's website carried out a survey of customer satisfaction with UK banks in the following way. The website presented a list of 13 banks and invited people to vote for the bank that, in their experience, gave the best customer service. It was not possible for anyone to vote twice from the same computer. The banks were then rated according to the number of votes received.

   Identify three distinct weaknesses in this survey method. For each weakness identified, discuss briefly how it might be overcome if possible.

   (6)

2. A series of surveys in the United States, carried out from 1992 to 2000, asked random samples of men and women aged 75 to estimate whether or not they would survive to the age of 85. Ten years after each survey the actual numbers surviving in each cohort were determined. All the sample sizes were large. The table summarises the data.

	Men		*Women*	
Year of survey	*Percentage estimating they would survive to age 85*	*Percentage actually surviving to age 85*	*Percentage estimating they would survive to age 85*	*Percentage actually surviving to age 85*
1992	39.6	27.0	45.9	44.2
1994	38.9	27.8	43.1	46.6
1996	41.3	29.4	46.5	45.1
1998	39.2	30.2	46.2	45.2
2000	39.8	33.2	48.6	47.4

   (i) Describe what the data show about actual survival rates for men and women over time.

   (3)

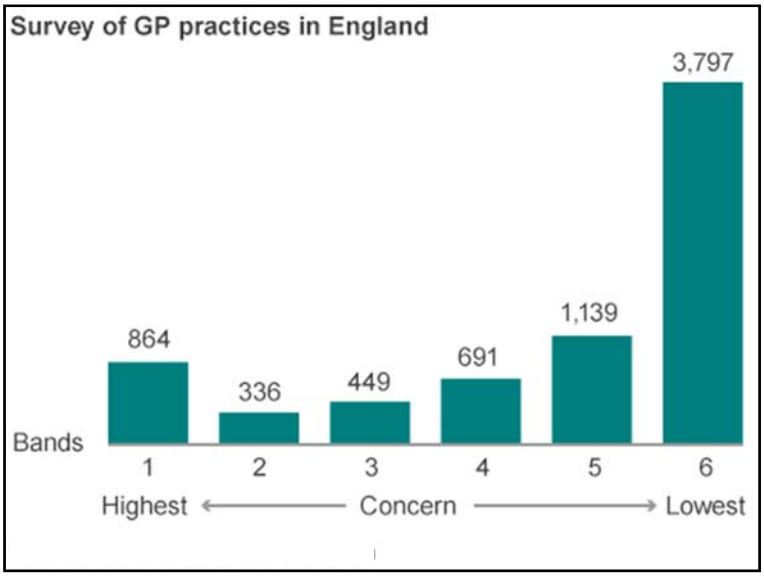
   (ii) Describe what the data show about estimated survival rates for men and women over time.

   (3)

   (iii) Compare the abilities of men and women to estimate their survival rates.

   (2)

3. A survey of family doctor practices (GP practices) in England was carried out by the Care Quality Commission in 2014. Each practice was given a score from 1 to 6, where 1 represents high cause for concern and 6 represents low cause for concern. The data are shown in the bar chart.



In writing about this story, a newspaper reporter calculated the mean score from these data.

(i) Find the value of the mean.

(2)

(ii) Discuss whether or not the mean is a useful measure for these data.

(3)

(iii) Imagine that you are reporting this story on a radio programme so that listeners cannot see the bar chart. Write a short paragraph summarising the most important features of the data for your listeners.

(4)

4. The table shows some recent data for the six largest Caribbean states by size of population. 'GDP per person' is a measure of the overall wealth of the country in relation to the size of its population.

	GDP per person (thousands of US $)	*Number of doctors per 1000 people*
Cuba	6.1	6.40
Haiti	1.7	0.25
Dominican Republic	5.8	1.88
Puerto Rico	28.5	1.75
Jamaica	5.3	0.85
Trinidad and Tobago	18.4	1.18

(i) Draw a scatter diagram for these data.

(3)

(ii) Find the value of Spearman's rank correlation coefficient for these data.

(3)

You are now **given** that the product moment correlation coefficient for these data is –0.075.

(iii) Comment briefly on the difference in the values of the two correlation coefficients.

(2)

(iv) Discuss briefly the relationship between GDP and numbers of doctors for these six countries. You should refer to your diagram and to the correlation coefficients in your answer.

(4)

5. In the UK, letters can be posted by two services, called ‘first class’ and ‘second class’. First class post is intended to be faster.

A letter posted first class on a Monday has the following probabilities of arriving on Tuesday to Friday of the same week.

Tuesday	Wednesday	Thursday	Friday
0.6	0.2	0.1	0.1

You should assume that the arrival days of letters are independent of one another.

(i) I post three letters first class on Monday. Find the probability that

(a) they all arrive on Tuesday,

(b) none of them arrive on Tuesday,

(c) one arrives on Wednesday, one on Thursday and one on Friday.

(7)

A letter posted second class on a Monday has the following probabilities of arriving on Tuesday to Friday of the same week.

Tuesday	Wednesday	Thursday	Friday
0.1	0.2	0.3	0.4

Again, you should assume that the arrival days of letters are independent of one another.

(ii) I post one letter first class and one letter second class on Monday. Find the probability that

(a) the two letters arrive on the same day,

(b) the second class letter arrives before the first class letter.

(5)

6. In a survey conducted in the UK in 2013, a random sample of 2000 people were asked to state their main source of news. The answers available were: television (T), the internet (I), newspapers (N), radio (R). The following table shows the percentages of people in various age groups giving each answer.

	Age			
	18–29	30–49	50–64	65+
T	50	50	58	68
I	27	28	18	6
N	7	6	8	18
R	3	7	7	4

(i) The columns of figures do not sum to 100. Explain why some people might not answer T, I, N or R.

(3)

(ii) Find, for the 18–29 age group, the percentage figures for T, I, N and R when corrected to add to 100. (You do not need to do this for other age groups.)

(3)

(iii) Describe briefly what the data indicate about the main source of news for people in different age groups.

(4)

7. A recently published dictionary of English contained definitions for 118 611 words. These words ranged in length from 1 letter to 25 letters. The data on word length are summarised in the table.

Number of letters, x	Number of words, f	Cumulative number of words
1	3	3
2	93	96
3	754	850
4	3027	3877
5	6110	9987
6	10083	20070
7	14424	34494
8	16624	51118
9	16551	67669
10	14888	82557
11	12008	94565
12	8873	103438
13	6113	109551
14	3820	113371
15	2323	115694
16	1235	116929
17	707	117636
18	413	118049
19	245	118294
20	135	118429
21	84	118513
22	50	118563
23	23	118586
24	16	118602
25	9	118611

You are given that $\Sigma xf = 1\,094\,334$ and that $\Sigma x^2 f = 11\,091\,638$.

(i) Find the mean and standard deviation of the number of letters per word.

(3)

(ii) Construct a box and whisker plot (boxplot) for the data.

(4)

(iii) Find the mode of the data.

Discuss briefly how useful the mean, median and mode are as indicators of central tendency for these data.

(3)

(iv) An investigation of the lengths of words in a newspaper found that the mean number of letters per word was 4.9. Discuss these findings in the light of the data from the dictionary.

(2)

8. The data in this question relate to the size and value of the housing stock in the regions of England in 2001 and 2011. In the table below,

P_{2001} denotes a price index for housing in 2001,

P_{2011} denotes the corresponding price index for housing in 2011,

Q_{2001} denotes the number of housing units, in thousands, in 2001,

Q_{2011} denotes the number of housing units, in thousands, in 2011.

Region	P_{2001}	Q_{2001}	P_{2011}	Q_{2011}
North	234.0	1115	483.1	1164
Yorks & Humber	257.5	2155	513.2	2294
North West	255.7	2945	485.0	3111
East Midlands	287.3	1797	517.9	1961
West Midlands	301.5	2225	530.9	2358
East	322.6	2308	544.1	2520
South West	327.8	2181	547.3	2403
South East	354.7	3392	553.1	3683
London	428.3	3090	659.6	3318
Totals		21208		22812

$\Sigma P_{2001}Q_{2001} = 6\,742\,056.6$
$\Sigma P_{2001}Q_{2011} = 7\,261\,010.9$
$\Sigma P_{2011}Q_{2001} = 11\,548\,569.6$
$\Sigma P_{2011}Q_{2011} = 12\,427\,822.3$

(i) Calculate a simple price relative for houses in the North region in 2011, taking 2001 as the base year.

(2)

(ii) Calculate the percentage increase, from 2001 to 2011, in

(a) the total number of housing units in England,

(b) the total value of the housing stock in England.

(4)

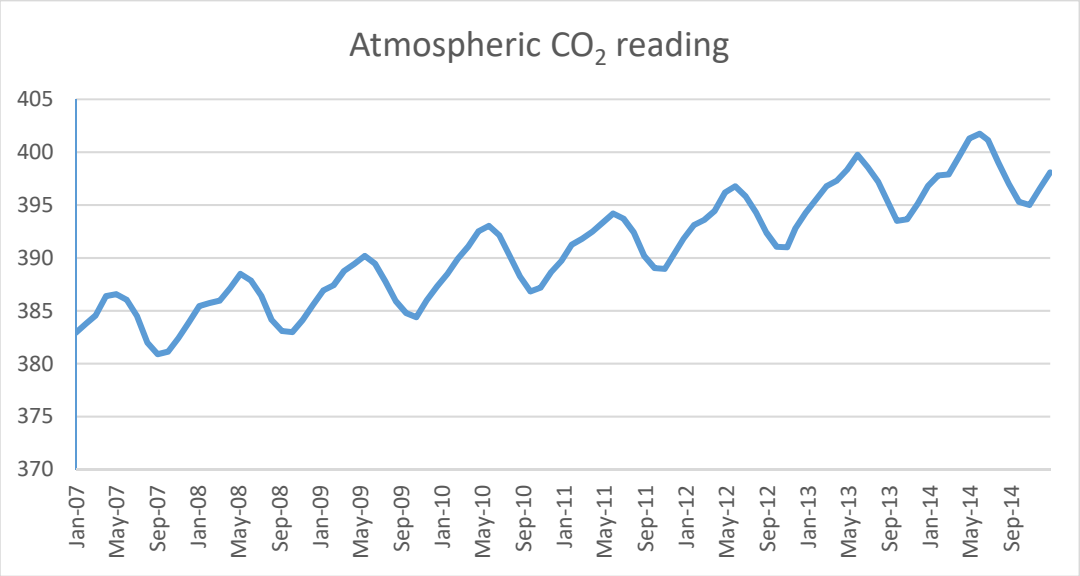
(iii) Calculate the average annual rate of house price inflation in London from 2001 to 2011.

(3)

(iv) Calculate the Laspeyres price index for the data.

(2)

9. The graph below shows readings of atmospheric CO_2 levels (in parts per million) at a monitoring station in Hawaii, taken over 8 years from January 2007 to December 2014. Note that the vertical axis does not start at zero.



(i) Describe the trend and variation in these data. State, with a reason, whether it would be more appropriate to use an additive or multiplicative model to analyse the seasonal variation. (6)

(ii) A climate change blogger drew the graph of these data so that the origin on the vertical axis was zero. Explain briefly what the graph would now look like.

The blogger argued that the total change in CO_2 readings over the eight years was small in relation to the average reading of about 390, and went on to say that worries about climate change may be exaggerated. Discuss briefly the validity of this argument. (4)

The mean annual readings of CO_2 levels for the eight years are given in the table, with summary data below. (The y readings have been rounded to 1 decimal place; the summary statistics have been calculated from the raw data.)

Year after 2000, t	7	8	9	10	11	12	13	14
Reading, y	383.8	385.6	387.4	389.8	391.6	393.8	396.5	398.4

$\Sigma t = 84$ $\Sigma y = 3126.86$ $\Sigma t^2 = 924$ $\Sigma y^2 = 1\,222\,346$ $\Sigma ty = 32\,920.98$

(iii) Calculate the least squares regression line for y on t.

Calculate also the product moment correlation coefficient for the data. (6)

(iv) Use your regression line to predict the mean annual reading for 2016 and also to estimate the year in which the mean annual reading is predicted to exceed 500 for the first time.

Comment on the validity of these two estimates. (4)

BLANK PAGE

BLANK PAGE

BLANK PAGE