EXAMINATIONS OF THE HONG KONG STATISTICAL SOCIETY

ORDINARY CERTIFICATE IN STATISTICS, 2015

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 study in Denmark investigated the relationship between the smoking habits of mothers and the length at birth of their babies. The mothers were divided into groups, each of size about 250, according to their level of smoking. The data were as follows.

Mean number of cigarettes smoked per day, x	0	3	7	15	23
Mean length (in cm) at birth of babies, y	52.0	51.8	50.8	50.6	50.4

(i)   Draw a scatter diagram of the data.

(4)

(ii)   Describe clearly and concisely what the scatter diagram indicates.

(3)

(iii) Discuss whether or not it would be appropriate to calculate a linear regression for y on x for these data.

(2)

2.　　An opinion poll carried out in the USA in 2010 asked those interviewed whether they thought spending on defence should be decreased, kept the same, or increased. Some of the results, classified by region of the country, are shown in Table 1. The percentages of people favouring decreasing defence spending, keeping it the same, and increasing it are shown for each region in Table 2.

	Northeast	Midwest	South	West	Totals
Decrease	30		51	56	169
Keep same		165	189		
Increase	93		229		554
Totals	266	355		296	

Table 1

	Northeast	Midwest	South	West
Decrease	11.3%	9.0%		18.9%
Keep same	53.8%	46.5%		56.1%
Increase	35.0%	44.5%		25.0%
Totals	100.0%	100.0%	100.0%	100.0%

Table 2

(i)　　Copy and complete the tables.

(5)

(ii)　　Compare attitudes to defence spending

(a)　　between those in the Northeast and those in the Midwest,

(b)　　between those in the South and those in the West.

(4)

3. The table below shows the numbers of deaths in road accidents in the years 2000, 2005 and 2010 for 33 countries.

	2000	2005	2010
Australia	1 817	1 627	1 366
Austria	976	768	552
Belgium	1 470	1 089	840
Bulgaria	1 012	957	775
Croatia	655	597	426
Cyprus	111	102	60
Czech Republic	1 486	1 286	802
Denmark	498	331	265
Estonia	204	169	78
Finland	396	379	270
France	8 079	5 318	3 992
Germany	7 503	5 361	3 651
Greece	2 037	1 658	1 281
Hungary	1 200	1 278	739
Irish Republic	415	399	212
Italy	7 061	5 818	3 998
Japan	10 403	7 931	5 745
Latvia	588	442	218
Lithuania	641	773	300
Luxembourg	76	45	32
Netherlands	1 082	750	640
New Zealand	462	405	375
Norway	341	224	210
Poland	6 294	5 444	3 907
Portugal	1 860	1 247	845
Romania	2 499	2 461	2 377
Slovakia	628	560	353
Slovenia	313	258	138
Spain	5 776	4 442	2 470
Sweden	591	423	266
Switzerland	592	409	327
United Kingdom	3 580	3 336	1 905
USA	41 945	43 443	32 788

(i) Describe the broad trend in numbers of road deaths in these countries from 2000 to 2010. Explain whether or not this trend is equally evident in the two periods 2000–2005 and 2005–2010.

(3)

(ii) Suggest two possible explanations for the fact that some countries do not fit in with the general trend. You should refer to specific countries in your answer.

(4)

(iii) State a useful way in which a country's road safety could be assessed using these data. Identify the best and worst performing countries.

(3)

(iv) The numbers of deaths in the USA are much greater than in the other countries for each of the 3 years. However, the USA is a much bigger country. Suggest two measures of road safety that take this into account.

(2)

4

4.   Eight patients with early signs of kidney malfunction were given a course of treatment intended to reduce their blood pressure. A blood pressure reading consists of a pair of measurements: the systolic or maximum pressure, and the diastolic or minimum pressure. In order to improve accuracy, a series of blood pressure readings are taken over a period of several hours; the readings are averaged to give the mean systolic pressure and mean diastolic pressure.

For each of the eight patients given the course of treatment, the table below shows
- u, the mean systolic pressure at the start of the treatment,
- v, the reduction in mean systolic pressure after 56 days,
- w, the mean diastolic pressure at the start of the treatment,
- x, the reduction in mean diastolic pressure after 56 days.

Patient number	Mean systolic, u	Systolic reduction, v	Mean diastolic, w	Diastolic reduction, x
1	138	10	80	2
2	136	26	70	8
3	128	2	78	4
4	128	8	78	10
5	134	16	72	0
6	137	20	63	5
7	145	20	66	8
8	120	6	63	1
Mean	133.25		71.25	4.75
Variance	59.0714		47.6429	13.3571

(i)   Calculate the mean and variance of the values of v.

(3)

(ii) Describe the apparent effects of the treatment on the blood pressure readings for these patients.

(3)

The table below shows product moment correlation coefficients between the variables.

	u	v	w	x
u	1.0000			
v	0.6950	1.0000		
w	−0.0417	−0.4634	1.0000	
x	0.2924	0.3264		1.0000

(iii) Calculate the product moment correlation coefficient between w and x.

(3)

(iv) Explain briefly what the correlation coefficients in the three shaded cells indicate.

(3)

5. According to the National Health and Nutrition Examination Survey in the USA, prevalence of myopia (short sightedness) varies by racial group as follows.

Racial group	Percentage having myopia
White	26.4
Black	14.5
Hispanic	18.3
Other	20.7
All	23.9

Suppose that one person is chosen at random from each of the four racial groups. Find the probability that

(i) all 4 have myopia,

(2)

(ii) none of them has myopia.

(3)

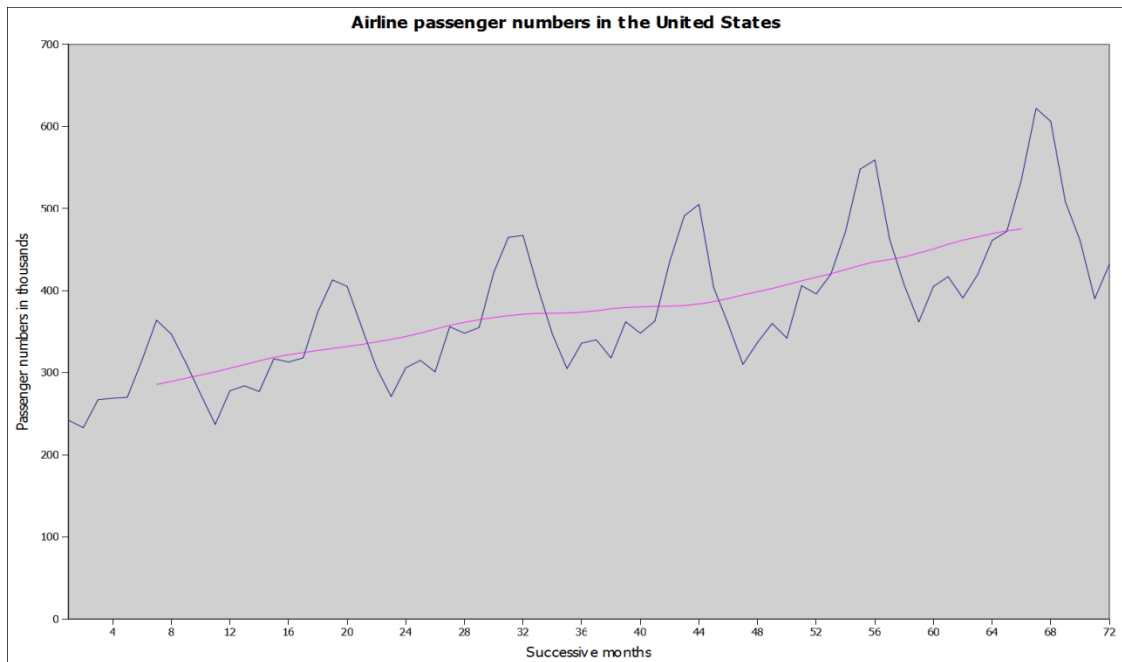Now suppose that people are chosen at random from the whole population.

(iii) Find the probability that, if 3 people are chosen, at least 1 has myopia.

(3)

(iv) Find the smallest number of people that should be chosen in order to be 95% certain that at least 1 has myopia.

(3)

6.   The table below shows the prices, P, and quantities grown, Q, of the cereals wheat, barley and oats in the UK in the years 2002 and 2012.

	2002		2012	
	P (£/tonne)	Q (10^3 tonnes)	P (£/tonne)	Q (10^3 tonnes)
Wheat	65.4	3615	168.7	2421
Barley	63.8	807.6	166.0	263.2
Oats	53.3	41.20	183.4	10.91

(i)   Calculate simple price relatives for the three cereals in 2012, taking 2002 as the base year.

(3)

(ii)   Find, for each cereal, the average annual rate of inflation for the 10 years from 2002 to 2012.

(3)

(iii) Calculate an index for the total value of these three cereals in 2012, taking 2002 as the base year.

(3)

(iv)   Find the Paasche price index for the data. Explain briefly the underlying principle of the Paasche price index.

(3)

7.   The graph below shows monthly airline passenger numbers, in thousands, in the United States over a period of six years in the 1950s. The 12-point centred moving average is also shown.

(i)   Describe the trend and the seasonal variation shown by the data.

(3)

The raw data and corresponding moving averages begin as shown in Table 1.

Month	Passenger numbers (thousands)	Moving average
Jan	242	
Feb	233	
Mar	267	
Apr	269	
May	270	
Jun	315	
Jul	364	285.75
Aug	347	a
Sep	312	293.25
Oct	274	297.17
Nov	237	301.00
Dec	278	305.46
Jan	284	309.96
Feb	277	314.42

Table 1

Question continued on the next page

8

(ii)   Explain, with reference to the graph, why it appears that a multiplicative model would be preferable to an additive model for the seasonal variation in these data.

(2)

(iii)   State how the moving average for July, 285.75, is calculated.

Calculate the moving average for August, shown as *a* in the table.

(3)

The raw data and moving averages for July in five successive years are as shown in Table 2.

Passenger numbers (thousands)	Moving average
364	285.75
413	329.54
465	369.46
491	381.83
548	430.71

Table 2

(iv)   Show, by considering the figures in Table 2, that a multiplicative model is preferable to an additive model in calculating the seasonal effect for July.

(4)

8.   A report published by the International Telecommunications Union contained the following information on internet use. Tables 1 and 2 show, for the years 2005, 2010 and 2013, the percentages of people in various categories who use the internet.

Year	2005	2010	2013
World population (billions)	6.5	6.9	7.1
Using the internet	16%	30%	39%
Users in the developing world	8%	21%	31%
Users in the developed world	51%	67%	77%

Table 1: Percentages of people using the internet over time

	Year	2005	2010	2013
	Africa	2%	10%	16%
	Americas	36%	49%	61%
Region	Arab States	8%	26%	38%
	Asia and Pacific	9%	23%	32%
	CIS	10%	34%	52%
	Europe	46%	67%	75%

Table 2: Percentages of people using the internet in different regions

NB: CIS is the Commonwealth of Independent States, consisting of Russia and some former members of the Soviet Union.

(i)   Calculate the number of internet users in the world in 2013.

(1)

(ii)   About 82% of the world's population in 2013 were from developing countries. Show that the figures for internet use are consistent with this.

(3)

(iii)   Use the data in Table 2 to show, in a suitable graphical form, internet use in 2013 by region.

(3)

(iv)   Represent the data in Table 2 in a suitable graphical form to allow comparisons to be made of rates of growth of internet use in different regions.

(3)

(v)   Comment on the usefulness or otherwise of the regional groupings used in Table 2.

(2)

9. The data below are the heights, rounded to the nearest centimetre, of a group of 45 adult male patients being treated at a clinic for obesity.

154	172	183	180	171	170	170	171	178
180	178	162	177	184	178	179	170	176
178	167	183	171	176	177	172	175	160
170	176	128	171	164	172	180	169	168
170	168	172	171	166	170	165	176	167

The following five summary statistics have been calculated.

Lower quartile 169   Median 171   Upper quartile 177

Mean 171.4   Standard deviation 9.1

(i) Identify, with a reason, an outlier in the data.

(2)

(ii) Explain, in general terms, the different ways in which an outlier may be treated.

(2)

(iii) State, for this particular case, two plausible explanations for the outlier.

(2)

(iv) Assume that, in this particular case, the outlier is removed from the data set. State the effect on each of the five summary statistics. (*Note: you are not required to do any calculations.*)

(5)

BLANK PAGE