EXAMINATIONS OF THE HONG KONG STATISTICAL SOCIETY

ORDINARY CERTIFICATE IN STATISTICS, 2012

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 7 printed pages, **each printed on one side only**.
This front cover is page 1.
Question 1 starts on page 2.

There are 8 questions altogether in the paper.

1. The figures below are the hourly wage rates, in US dollars, for a random sample of workers in the US construction industry.

9.88	10.19	11.15	11.70	12.01	12.46
13.22	13.29	14.00	14.38	14.50	14.50
15.14	15.16	15.16	15.16	15.54	15.96
16.20	16.22	16.44	17.58	20.38	21.15

(i) Calculate the median and the inter-quartile range.

(4)

(ii) Draw a boxplot for these data.

(3)

2. A doctor has 8 patients with a particular medical condition. The doctor notes their ages in years as follows.

45 47 48 48 53 55 57 61

Find the mean of these figures and explain why it is likely to underestimate the mean age of these patients.

(3)

Give a better estimate of the mean age of these patients, explaining your reasoning.

(2)

3. Consider the following two cases. In one it is appropriate to calculate a coefficient of variation and in the other it is not.

(a) The mean temperature, in degrees Celsius, at noon on the 30 days of November 2011 in Manchester was 9.7 and the standard deviation was 3.3.

(b) The mean diameter, in millimetres, for a sample of 30 ball-bearings made on a machine was 8.810 and the standard deviation was 0.072.

Identify the case where it is not appropriate to calculate a coefficient of variation, giving a reason for your decision.

Calculate the coefficient of variation in the other case.

(4)

4. The table below shows data about the 'G8' group of industrialised countries for the year 2007. Gross Domestic Product (GDP) is a measure of the size of a country's economy. Carbon dioxide (CO_2) emissions are widely thought to be responsible for harmful climate change.

Country	*GDP in US $ billions*	*CO_2 emissions in millions of tonnes*
Canada	994	639
France	2 060	383
Germany	2 744	809
Italy	1 729	450
Japan	4 608	1 258
Russia	592	1 525
UK	2 155	587
USA	11 712	6 049

(i) Draw a scatter diagram of the data.

(4)

(ii) Calculate the ratio of CO_2 emissions to GDP for each country. Identify the country that has the highest figure for emissions in relation to its GDP and the country that has the lowest. Indicate the positions of the data points for these two countries on your scatter diagram.

(5)

(iii) Calculate Spearman's rank correlation coefficient for the data.

State, with a reason but without doing any calculations, whether you would expect the product-moment correlation coefficient to be less than, about the same as, or greater than Spearman's coefficient in this case.

(6)

(iv) Calculate the value of Spearman's rank correlation coefficient with the data point for the USA removed. Comment on your result.

(4)

5. A genetic condition is known to be present in a small proportion of the population. A test for this condition exists, but it does not detect the condition correctly in every case.

Let G be the event that a person has the genetic condition. Let T be the event that a person tests positive for the condition – that is, the test indicates that he or she has the genetic condition. Complements of these events are G', T' respectively.

You are given that $P(T\,|\,G) = 0.98$ and $P(T\,|\,G') = 0.01$.

(i) Write down the values of $P(T'\,|\,G)$ and $P(T'\,|\,G')$. What do the various probabilities indicate about the failure rates of the test?

(4)

You are now given that the genetic condition is present, on average, in 1 person per 400 of the population. A person is chosen at random and tested.

(ii) Draw a tree diagram and use it to find the probability that this person tests positive for the genetic condition.

(4)

(iii) Given that this person tests positive, find the probability that he or she actually has the genetic condition.

(3)

(iv) In practice, the probability that a person testing positive actually has the condition is likely to be higher than the answer to part (iii). Explain briefly why that is the case.

(2)

6. In a random sample taken in Brazil, 870 adults were assessed for their risk of developing a form of dementia. The adults were classified as low risk, medium risk or high risk. The data were analysed by sex and by age.

The contingency table for the data categorised by sex is as follows.

Sex	*Low risk*	*Medium risk*	*High risk*
Male	322	118	21
Female	327	45	37

(i) Calculate, for males and females separately, the percentages in each of the three risk categories.

Comment briefly on how risk appears to vary with sex.

(5)

The contingency table for the data categorised by age is given below, with some entries missing.

Age	*Low risk*	*Medium risk*	*High risk*	*Totals*
18–27	119	41	19	179
28–37	165	47		222
38–47		43	8	204
48–57	98		8	
58+	114	15	13	142
Totals	649			

(ii) Copy and complete the table.

Comment briefly on how the risk appears to vary with age.

(10)

(iii) State what further analysis of the data might be possible and why it might prove useful.

(2)

7. The data in the table below show quarterly house sales in a region of Scotland for 22 successive quarters. The table also shows the appropriate centred moving average.

(i) Calculate the missing values shown as *a*, *b* and *c*.

(6)

(ii) Without doing any calculations, describe the variation shown in the data across the four quarters of the year.

(2)

(iii) Calculate the average difference from the overall mean for each of the four quarters.

(5)

(iv) Draw a suitable graph and hence state what trend, if any, is evident in the data.

(6)

Quarter	*Sales*	*Sales*	*Sales*	*Sales*	*Moving average*
1	207				
2		223			
3			364		286.375
4				355	294.25
1	200				306.25
2		293			*a*
3			390		331.875
4				399	358
1	291				381.75
2		411			409.25
3			462		429
4				547	423.625
1	*b*				421.5
2		358			416.25
3			498		405.625
4				469	419.75
1	294				436.5
2		478			437.75
3			512		437.75
4				465	*c*
1	298				
2		337			

8. An investor bought shares in six investment trusts in 2005. Over the five years to 2010 he increased his number of shares in some trusts and decreased the number in others. The table shows the prices of the shares in 2005 and 2010 together with the numbers of shares held at those times. The share prices are in pence.

Investment trust	*2005 Price*	*2005 No. of shares*	*2010 Price*	*2010 No. of shares*
Japan	242.50	1200	162.25	600
Europe	530.50	650	836.00	850
USA	283.00	1000	368.50	1000
Ireland	582.75	500	665.50	200
China	69.50	1500	152.75	3000
Russia	99.25	1300	55.00	2000

(i) Calculate, for each trust, the simple price relative in 2010 taking 2005 as the base year. Find the simple average of these price relatives.

Calculate also the relative total value of the investor's shares in 2010 taking 2005 as the base year.

(7)

(ii) Calculate the Laspeyres index of share price relatives for these data.

(3)

(iii) The Fisher index of share price relatives for these data is 1.270. Use this figure to calculate the Paasche index.

(3)

(iv) Explain briefly how the three indices in parts (ii) and (iii) differ in the way they measure changes in the investor's share holding.

(3)