

EXAMINATIONS OF THE HONG KONG STATISTICAL SOCIETY



HIGHER CERTIFICATE IN STATISTICS, 2008

(Modular format)

MODULE 1 : Data collection and interpretation

Time allowed: One and a half hours

*Candidates should answer **THREE** questions.*

Each question carries 20 marks.

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

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).

The notation \log denotes logarithm to base e .

Logarithms to any other base are explicitly identified, e.g. \log_{10} .

Note also that $\binom{n}{r}$ is the same as nC_r .

This examination paper consists of 4 printed pages **each printed on one side only**.

This front cover is page 1.

Question 1 starts on page 2.

There are 4 questions altogether in the paper.

1. (i) No sampling frame is perfect. Explain why this is so, illustrating your answer with examples. As far as possible these should be of a frame or frames with which you are familiar. (8)
- (ii) A researcher has access to a sampling frame of addresses and a sampling frame of school children in a particular area, and wishes to select a sample of households that have children of school age.
- (a) Discuss how each frame might be adapted so that a suitable sample may be drawn from it. In your discussion you should point out any imperfections each frame has for this purpose. (9)
- (b) Explain, with reasons, which of these two frames you think would be the better one for the researcher to use. (3)
2. A research organisation is planning a survey to investigate patients' perceptions of post-operative care in hospitals. A list of hospitals is available for each of the ten regions in the country, and on this list it is indicated whether or not the hospital is a teaching hospital. Samples of patients who leave selected hospitals in a specified week following an operation are to be chosen. These patients will be mailed a self-completion questionnaire two weeks after they have left hospital.
- (i) Suggest a two-stage sampling design that might be used to choose a sample of hospitals. Discuss the benefits and drawbacks in this application of the scheme you have suggested. (9)
- (ii) Due to the confidentiality of patient records, samples of patients are to be selected from administrative records by hospital staff. What information do you think the research organisation might need about the format and structure of the records in order to devise a scheme for the selection of patients? (5)
- (iii) Draft six questions on post-operative care suitable for inclusion in the questionnaire to be sent to patients. These questions should be in closed form. (6)

3. (i) The table below shows, for women and for men in Great Britain in two different years, the percentages with total individual incomes in quintile groups defined by the total individual incomes of all adults in that year. The quintiles divide the distribution of incomes into five equal parts, so that the lowest 20% of total incomes fall in the first quintile, the next 20% in the second quintile, and so on.

The means of the total individual incomes were £227 per week for all women and £408 for all men in 2003/4. The corresponding means of the 1996/7 incomes expressed in terms of 2003/4 prices were £177 and £347.

	Quintile groups					Population ('000s)
	Bottom	2nd	3rd	4th	Top	
1996/7	<i>Percentages in group</i>					
<i>All women</i>	28	25	21	15	10	22289
<i>All men</i>	11	14	19	25	31	20591
2003/4	<i>Percentages in group</i>					
<i>All women</i>	27	25	21	16	11	22913
<i>All men</i>	12	15	19	24	30	21399

Write a report on the main points shown in this table, paying particular attention to any income inequalities between women and men, and any changes over time. Your report should be written in such a way that it can be understood by non-technical readers.

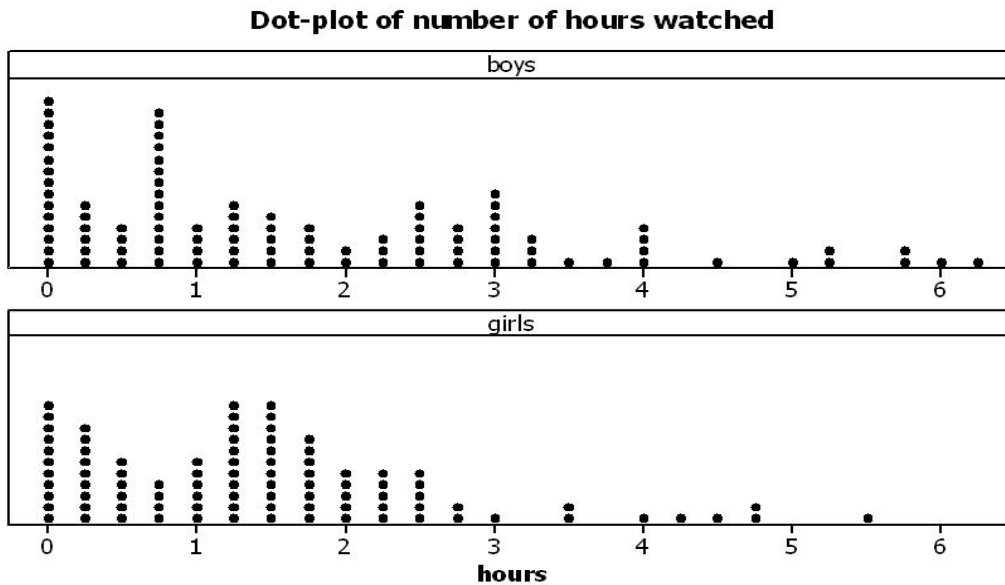
(13)

- (ii) Discuss the main problems in obtaining reliable information about people's annual incomes.

(7)

4. (i) A health-related survey of school children has been undertaken. One of the questions was "How long did you spend watching television after school yesterday? Give your answer to the nearest quarter of an hour". Some computer output relating to the responses that were given to this question by a random sample of boys and girls aged 15–16 is given below.

No. of hours	N	Mean	SE Mean	StDev	Minimum	Q1	Median	Q3	Maximum
Boys	97	1.799	0.164	1.613	0.000	0.500	1.250	2.875	6.250
Girls	92	1.473	0.126	1.205	0.000	0.500	1.250	2.000	5.500



- (a) The arithmetic mean, the median and the mode are three common measures of location of a distribution. Write sentences to explain what they are to someone who is not familiar with these terms. You should incorporate values from the output into your explanations. (7)
- (b) What advantages do you think dot-plots have over histograms in this example? (3)
- (c) Use the information given in the output to draw side-by-side box and whisker plots and comment briefly on the main features that they show. (6)
- (ii) What advice would you give a researcher who wants to conduct a survey of school children about their use of drugs? (4)