

EXAMINATIONS OF THE HONG KONG STATISTICAL SOCIETY



HIGHER CERTIFICATE IN STATISTICS, 2013

MODULE 3 : Basic statistical methods

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 ${}^n C_r$.

This examination paper consists of 8 printed pages.

This front cover is page 1.

Question 1 starts on page 2.

There are 4 questions altogether in the paper.

1. One criterion that students may apply when choosing a university for undergraduate studies is the number of scheduled contact hours per week. However, different disciplines have developed teaching patterns best suited to their subject material. Typically, scientific and engineering subjects may require more contact hours to allow for structured programmes of lectures and laboratory sessions. In contrast, arts and humanities subjects require fewer contact hours to foster independent, creative and original analysis.

The table below shows the sample mean numbers of contact hours for ten discipline categories at two universities, A and B. The values have been obtained from samples of students in each discipline at each university. It is of interest to investigate whether there is a significant overall difference between the mean numbers of contact hours at these two universities.

<i>Subject category</i>	<i>Mean number of hours at university A</i>	<i>Mean number of hours at university B</i>
Medical Studies	20.1	20.3
Technology and Engineering	19.2	21.4
Law	11.6	12.3
Business and Management Studies	12.2	13.1
Philosophical and Historical Studies	8.2	7.9
Classics and Linguistics	10.1	10.7
Mathematics and IT	14.9	16.1
Physical and Biological Sciences	16.0	17.8
Urban, Regional and Architectural Studies	16.2	15.8
Media and Social Studies	12.0	12.8
TOTAL	140.5	148.2

- (i) Perform a t test at the 5% significance level to examine whether the difference in contact hours between the universities has a mean of zero. State your null and alternative hypotheses clearly and report your conclusions. State any assumptions made in carrying out the t test. (12)
- (ii) Two more subject categories are added to the ten given above. The mean number of hours for Art and Design is 13.2 at university A and 10.7 at university B; the mean number of hours for Education is 13.5 at university A and 9.8 at university B. It is found that, with these additional subject categories, the assumptions required for a t test are violated. Carry out a Wilcoxon signed rank test on the full set of data on all twelve subject categories. Specify your null and alternative hypotheses, and report your conclusions. (8)

2. A city council is considering introducing a congestion charge for motorists travelling into or out of the city centre. The city is divided into ten administrative areas. In order to assess the popularity of such a measure, samples of residents from two of the administrative areas are asked whether or not they are in favour of the introduction of the congestion charge. The results are shown below.

	<i>In favour of the charge</i>	<i>Not in favour of the charge</i>
<i>Area 1</i>	61	95
<i>Area 2</i>	20	84

- (i) Perform a χ^2 test at the 5% significance level to investigate whether there is an association between the area of the city and the attitude to the proposed congestion charge. State your null hypothesis and report your conclusions. (9)
- (ii) Estimate the proportions who are not in favour of the proposed congestion charge for each of the two areas and calculate an approximate 95% confidence interval for the difference in these two proportions. (8)
- (iii) You could perform a hypothesis test to examine whether there is a difference in the proportions of those who are not in favour of the proposed congestion charge in the two areas. Without performing this test, outline briefly how its results would relate to your answers to parts (i) and (ii). (3)

3. The amount of a potentially toxic pollutant in the water of a river affects the edibility of mussels grown in its estuary. An environmental health officer has heard a report of a leak of this pollutant into the estuary and undertakes an investigation into how this has affected the mussel population. He takes a sample of ten mussels randomly from this population and measures the amount of the pollutant in parts per million (ppm) in each of them. These ten values are as follows.

39.5 38.6 44.9 36.4 45.6 46.6 36.1 32.3 35.0 35.5

- (i) Calculate a 99% confidence interval for the population mean. (6)

- (ii) Calculate a 95% confidence interval for the population variance. (5)

Official health guidelines state that mussels are safe to eat provided that the (population) mean level of pollutant does not exceed 36 ppm.

- (iii) Test, at the 1% significance level, whether or not the population mean level exceeds 36 ppm. State the null and alternative hypotheses and report your conclusions. (8)

- (iv) State briefly why the 99% confidence interval for the mean calculated in part (i) cannot be used directly to perform the required hypothesis test in part (iii). (1)

4. A manager wishes to compare the effect of two training methods on the time it takes workers to complete a particular task on a production line. Eighteen workers were each randomly assigned to one of the two training methods. At the end of the training, each worker was asked to complete the task and the time taken in minutes to do this was recorded.

Summary values calculated from the results are given below.

Training Method 1

Number of workers = 8

Sample mean = 35.22 minutes

Sample variance = 24.445 minutes²

Training Method 2

Number of workers = 10

Sample mean = 31.56 minutes

Sample variance = 20.027 minutes²

It is thought reasonable to assume that the times taken by workers to complete the task are Normally distributed.

- (i) Assuming that the underlying population variances are equal, construct a 95% confidence interval for $\mu_1 - \mu_2$, where μ_1 and μ_2 are the population mean times taken by workers to complete the task after completing training methods 1 and 2 respectively. Can we conclude that training method 2 results in faster completion times than training method 1? Briefly justify your answer. (8)

Now, instead, let σ_1^2 and σ_2^2 denote the population variances for the times taken to complete the task by workers trained using training methods 1 and 2 respectively. Using the sample variances above, the company statistician states that a 95% confidence interval for $\frac{\sigma_1^2}{\sigma_2^2}$ is (0.291, 5.883).

- (ii) Explain what the statistician's statement means and reproduce her calculations. (9)
- (iii) Write down a 95% confidence interval for the corresponding ratio of population standard deviations. (1)
- (iv) Is there enough evidence to conclude that workers following training method 1 show more variable completion times than those following training method 2? Give a reason for your answer. (2)

BLANK PAGE

BLANK PAGE

BLANK PAGE