EXAMINATIONS OF THE HONG KONG STATISTICAL SOCIETY

HIGHER CERTIFICATE IN STATISTICS, 2005

Paper III : Statistical Applications and Practice

Time Allowed: Three Hours

Candidates should answer **FIVE** *questions.*

All questions carry equal 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 9 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. An experiment was carried out to study the effect of two factors on the blood sugar level in rabbits. One factor was two types (I and II) of insulin and the other factor was the dose level ("low" or "high") at which each type of insulin was given. Four rabbits were used in the experiment, and each was given all four treatment combinations, in random order, separated by suitable intervals of time in the hope of avoiding carry-over effects from one treatment to the next.

Measurements of blood sugar level (mg/100cc) are given in the following table.

	Type of insulin			
	I		**II**	
	Dose level		*Dose level*	
	low	*high*	*low*	*high*
Rabbit A	61	47	76	63
Rabbit B	89	61	74	46
Rabbit C	86	63	79	58
Rabbit D	69	59	66	50
Treatment total	305	230	295	217

(i) Calculate the mean responses for the four treatment combinations, and illustrate these in a graph. Is there any suggestion of interaction between the two factors?

(4)

(ii) Copy and complete the following analysis of variance table, and subdivide the treatments into components each having one degree of freedom. Calculate also the standard error of a treatment mean.

Source	df	SS	MS	*F* ratio
Rabbit		297.19		
Treatment		1496.69		
Residual				
Total		2463.94		

(8)

(iii) Write a brief report on the results of the experiment.

(4)

(iv) Another experimenter suggests that it would save time if, instead, four different groups, each of four rabbits, were used in this type of experiment. Each treatment combination would be given to one of the groups, chosen at random. Discuss briefly the advantages and disadvantages of conducting an experiment in this way.

(4)

2. (i) An oil company wishes to test a new additive, which they think will decrease petrol consumption (and believe cannot increase it). An experiment is carried out in which miles per litre are recorded for different makes of the same type of car. Twelve cars are selected and divided randomly into two groups of six. Group A cars use petrol with the additive and Group B cars use petrol without the additive. The results are:

A	6.55	8.78	10.80	9.05	7.35	9.38
B	9.23	10.55	6.73	7.55	8.23	7.72

(a) Test whether there is evidence that the additive increases the mean mileage per litre. State any assumptions you make.

(6)

(b) Now consider the case where each sample consists of n cars, where n is large, yielding sample means $\bar{x}_A$ and $\bar{x}_B$. The company wishes to be 95% confident of detecting an increase in mean of 0.5 of a mile per litre when using the additive. Let the mean fuel consumption using petrol with additive be μ_A and the mean fuel consumption using petrol without additive be μ_B. Write down a probability statement that represents the company's objective.

Assuming that the data to hand give a reasonable estimate of the sampling variance of petrol consumption, calculate how large n must be to meet the company's objective.

(6)

(ii) The Director of Research and Development at the oil company decides that he is not satisfied with the design of the experiment in (i). He decides to conduct a new experiment in which the same six cars in Group A above are tested again using petrol without the additive. The results obtained are:

With additive	6.55	8.78	10.80	9.05	7.35	9.38
Without additive	6.15	7.73	10.34	8.16	7.27	8.02

(a) Explain why this type of experimental design is better than the one used in (i). Describe any other improvements you might make if another experiment was being planned.

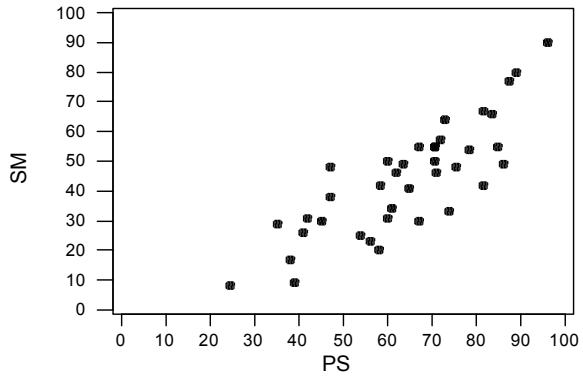
(3)

(b) Determine whether or not there is evidence that the additive increases mean miles per litre.

(5)

Turn over

3. A study into student progression on a mathematics degree looked at the performance of a random sample of 38 students in the year 2 Probability and Statistics (PS) examination and the year 3 examination in Statistical Modelling (SM). The scatter plot below shows the results.



A regression model is proposed to describe the result on the SM course, y, as a linear function of the result achieved on the PS course, x.

The model is given as $y_i = \alpha + \beta x_i + e_i$, where the $\{e_i\}$ are independent and distributed as N(0, σ^2).

(i) Find least squares estimates, $\hat{\alpha}$ and $\hat{\beta}$, of α and β, given the following summary information.

$$\sum_{i=1}^{38} x_i = 2436 \quad \sum_{i=1}^{38} y_i = 1670 \quad \sum_{i=1}^{38} x_i y_i = 116888 \quad \sum_{i=1}^{38} x_i^2 = 166991 \quad \sum_{i=1}^{38} y_i^2 = 86402$$

(5)

(ii) Interpret the fitted model from a practical point of view and comment on the fit of the model.

(4)

(iii) Calculate an estimate for σ^2 and calculate the estimated variances of $\hat{\alpha}$ and $\hat{\beta}$.

(6)

(iv) Give a 95% confidence interval estimate for the mean mark in the SM examination for students who score 80 on the PS examination. Comment on your answer.

(5)

Turn over

4. (i) Describe four types of component into which a time series can be decomposed. Define an additive model for a time series with these components and demonstrate how a multiplicative model might be converted to an additive model.

(4)

(ii) The half-yearly profits of a company (in £million) are shown below for the past 10 years. These 20 figures have been inflation adjusted to present them at today's values.

Year	*Feb–Jul*	*Aug–Jan*	*MA1*	*MA2*
1	1.6	1.2		1.45
2	1.8	1.4	1.55	1.625
3	1.9	1.6	1.7	1.775
4	2.0	1.7	1.825	1.875
5	2.1	1.7	1.9	1.9
6	2.1	1.8	1.925	1.975
7	2.2	1.8	2	2
8	2.2	1.8	2	2.025
9	2.3	1.9	2.075	2.125
10	2.4	2.0	2.175	

(a) A plot of the data is shown below. Describe its main features.

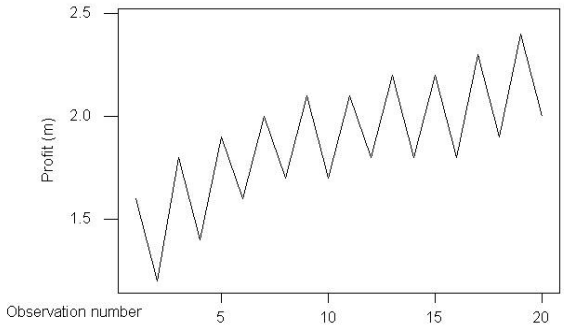
(3)

(b) An additive model is assumed, and the table above shows the half-yearly moving average values. Identify from these values the form of the moving average used. Comment on the suitability of the moving average chosen. Suggest an alternative moving average that might be used and contrast it with the one chosen.

(4)

(c) Using the moving average given in the table, obtain the detrended series and from this estimate the seasonal components. Hence calculate the irregular components and plot them against time. Comment on the fit of the model.

(9)



Turn over

5. (i) From a very large batch of mass produced widgets, a simple random sample of 20 is selected and each widget in the sample is inspected to see whether or not it meets the quality requirement. The batch is accepted if the sample contains 0 or 1 defective widgets.

In a simple random sample, every widget has the same probability of being selected for the sample. Calculate the exact probability of accepting a batch if it contains a proportion p of defectives, for $p = 0.01, 0.05, 0.1$.

(4)

(ii) The above scheme is now modified in that, if the sample contains more than 2 defectives the batch is still rejected, but if the sample of 20 contains just 2 defectives, a further random sample of size 20 is selected, which may be assumed independent of the first sample. If the second sample contains no defectives then the batch is accepted, otherwise the batch is rejected.

List the possible numbers of defectives in the two samples which lead to acceptance of the batch, and hence calculate the probability of accepting a batch, for $p = 0.01, 0.05, 0.1$.

(7)

(iii) Rejected batches are subject to 100% inspection and all defectives are removed. If the batch size is 1000, calculate the expected total number of items inspected in the two sampling schemes in (i) and (ii), for each of the values of p given. Comment briefly on your results.

(9)

6. (i) What do you understand by the term *simple random sampling*? Describe conditions under which it may not be a suitable sampling procedure or where it would be desirable to combine it with some other sampling method.

(4)

(ii) Choose three different types of *non-sampling error*, and briefly describe circumstances in surveys that give rise to these errors.

(7)

(iii) Outline the main disadvantages of telephone surveys.

(3)

(iv) A poll was conducted on the support given by the public to a new government health policy initiative. Of 1015 people surveyed, 853 expressed support. One year later, a similar poll of whether support was still as high yielded 780 out of 1005 people in favour. Determine whether there is any evidence of a decrease in the proportion of people supporting the policy.

(6)

7. The amount (£) paid out by an insurance company on claims arising from a certain group of policies is considered to follow an exponential distribution with mean μ.

(i) For a random sample of n claims $x_1, x_2, \ldots, x_n$, write down the likelihood function in terms of μ and hence find the maximum likelihood estimator of μ.

(5)

(ii) Over the course of a year, 96 claims are received and the amounts (x, in £) paid out by the insurance company are shown below, grouped into categories. The sample mean claim size is £2989.8. Calculate the missing entries in the column of expected numbers of claims based on the exponential model, and carry out an appropriate statistical test to determine whether or not an exponential model is suitable for these data.

Interval	*Number of claims*	*Expected number of claims*
$0 \le x < 250$	11	7.70
$250 \le x < 500$	16	7.08
$500 \le x < 1000$	16	
$1000 \le x < 1500$	10	10.58
$1500 \le x < 2000$	10	8.95
$2000 \le x < 3000$	11	
$3000 \le x < 4000$	7	10.01
$4000 \le x < 6000$	5	12.29
$6000 \le x < 8000$	4	
$8000 \le x < 10000$	2	
$10000 \le x < 20000$	2	3.27
$20000 \le x < 40000$	1	0.12
$40000 \le x < 60000$	1	0.00

(10)

(iii) Use the fitted exponential model to estimate the probability that the amount paid out on a claim exceeds £20 000. Comment on your answer in relation to the data.

(3)

(iv) Use your answers to (ii) and (iii) to discuss the characteristics of an improved model for claim payments on this group of policies.

(2)

8. An investigation was carried out by a flour manufacturer into the production line variations of a certain baking process. Squares of puff pastry were baked in a tray containing 7 squares across the tray and 10 squares the length of the tray.

Data were collected on the size of the finished product, just after coming out of the oven. For each "square", the width, length, and height (mm) were measured and also, on the basis of these measurements, the approximate volume (mm^3) was calculated. The data are summarised below.

Present your conclusions about the baking process in as informative a way as possible, including suitable plots, but avoiding formal statistical material (such as significance tests or confidence intervals). You should consider questions such as whether there are detectable effects due to the position on the tray of the puff pastry square, possible relationships between the dimensions of the squares and any implications for the uses to which the pastry squares might be put. Highlight any additional information about the raw data that might be useful.

(20)

Summary data classified by pos-w, i.e. position across the width of the tray

Variable length, N = 10

pos-w	Mean(l)	SD(l)	Med(l)	Min(l)	Max(l)
1	86.4	1.42984	86.5	85	89
2	87.3	1.25167	87.0	86	90
3	87.5	1.43372	88.0	84	89
4	86.7	1.70294	86.5	84	90
5	86.2	1.81353	86.0	83	89
6	84.7	1.63639	84.5	83	88
7	84.3	1.49443	84.0	82	87

Variable width, N = 10

pos-w	Mean(w)	SD(w)	Med(w)	Min(w)	Max(w)
1	77.1	1.91195	77.5	72	79
2	78.1	1.44914	77.5	77	81
3	77.0	2.78887	76.5	74	83
4	76.1	1.66333	76.0	73	79
5	76.8	0.91894	77.0	76	79
6	78.3	1.41814	78.0	77	81
7	79.1	1.66333	79.5	76	81

Variable height, N = 10

pos-w	Mean(h)	SD(h)	Med(h)	Min(h)	Max(h)
1	27.8	2.44040	27.5	24	31
2	27.9	2.84605	28.0	24	32
3	29.3	2.75076	28.0	27	34
4	29.2	2.44040	30.0	25	32
5	28.4	2.01108	29.0	24	31
6	30.6	2.50333	30.0	27	36
7	31.5	3.89444	31.5	27	38

Variable volume, N = 10

pos-w	Mean(v)	SD(v)	Med(v)	Min(v)	Max(v)
1	185044	15396.0	183344	159120	207669
2	190217	19868.5	187910	167475	216832
3	196930	13628.0	193604	178524	221408
4	192254	11153.5	194940	174174	204972
5	187874	11941.2	189535	162336	200970
6	202796	15521.1	201192	181305	238392
7	209629	22467.7	205096	183222	242609