EXAMINATIONS OF THE HONG KONG STATISTICAL SOCIETY

HIGHER CERTIFICATE IN STATISTICS, 2008

Paper II : Statistical Methods

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. (i) Write down the standard model and assumptions for one-way analysis of variance.

(3)

(ii) A junior clinician carries out a small pilot trial on 12 patients to assess the effectiveness of an experimental drug to reduce blood pressure (BP). Doses at three different levels are given daily for a week, the dose levels being assigned to patients at random, and the reductions in BP (in millimetres of mercury) over this period are tabulated.

Dose level (mg)	*Reduction in BP*	*Dose level means*
4	10, 6, 8	8
10	12, 7, 8, 9	9
16	12, 14, 13, 12, 9	12

(a) Calculate the estimated residuals for a one-way analysis of variance model and comment briefly on how well the model assumptions appear to hold.

(3)

(b) Test the null hypothesis that the mean reduction in BP is the same for all dose levels against the alternative that the mean reduction for at least one dose level differs from the rest. Report your conclusion clearly. You are given that the sum of the observations is 120 and the sum of squares of the observations is 1272.

(11)

(iii) A medical statistician believes that, over the range of dose levels used, any possible effect of the drug should increase with dose. He therefore suggests that a simple linear regression of BP reduction on dose level should provide a better test of the effectiveness of the drug. Without carrying out further calculations, comment on this suggestion in the light of your analysis.

(3)

2. (a) (i) A random sample of paired data $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$ is drawn from the joint distribution of two random variables X and Y. For $i = 1, 2, \ldots, n$ let $r(x_i) = \text{rank}(x_i)$ and $r(y_i) = \text{rank}(y_i)$, and let $\bar{x}, \bar{y}, \bar{r}_x$ and $\bar{r}_y$ denote the respective mean values of the original and ranked data. What information do the quantities r and R given by the formulae

$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}},$$

$$R = \frac{\sum \left(r(x_i) - \bar{r}_x\right)\left(r(y_i) - \bar{r}_y\right)}{\sqrt{\sum \left(r(x_i) - \bar{r}_x\right)^2 \sum \left(r(y_i) - \bar{r}_y\right)^2}},$$

convey about the relationship between X and Y?

(5)

(ii) Draw x-y scatter plots to illustrate a situation in which both R and r are appropriate measures of the relationship between X and Y, and a situation in which neither R nor r is appropriate.

(3)

(b) The table below shows the population of the United Kingdom in millions (X), and the percentage aged over 65 (Y), over the period 1871–2001.

Year	1871	1901	1921	1931	1951	1961	1971	1981	1991	2001
y	4.98	4.61	6.07	7.37	15.78	17.15	19.07	20.48	20.94	20.04
x	26.1	36.9	42.8	44.8	48.8	51.3	54.0	54.7	55.4	56.4

(i) Draw a scatter diagram for x and y and comment briefly on the relationship between X and Y. A statistics student wishes to test at the 1% significance level the null hypothesis of zero correlation against the alternative that there is positive correlation between X and Y. Calculate the Pearson (product-moment) correlation coefficient for these data, perform the test and report your conclusion clearly.

[You are given that $\sum x = 471.2$, $\sum y = 136.49$, $\sum x^2 = 23053.04$, $\sum y^2 = 2303.5257$, $\sum xy = 6980.286$.]

(7)

(ii) Later in his course, the student learns about rank correlation. Calculate the rank correlation coefficient for these data and perform the test corresponding to that in part (b)(i). Which of these measures of association is more appropriate to use on these data?

(5)

3. (i) State the *Central Limit Theorem* (CLT) and explain how it is used in practice. (4)

The following data are an ordered random sample of the waiting times of 20 patients at a doctor's surgery.

0 1 1 1 2 2 2 3 3 4 4 5 6 7 8 9 11 13 16 22

[You are given that the sum and sum of squares of the data are 120 and 1350 respectively.]

(ii) Calculate the mean and variance of these waiting times, and use the CLT to calculate an approximate 90% confidence interval (CI) for the true mean waiting time, μ. (4)

(iii) Calculate the median and quartiles of the waiting time data and comment critically on the above use of the CLT. (6)

Suppose now that the waiting times follow an exponential distribution with unknown mean μ. You are given that, if the random variable $\overline{X}$ denotes the mean of a sample of size n from this distribution, then $2n\overline{X}/\mu$ follows a χ^2 distribution with $2n$ degrees of freedom.

(iv) Use this result and the table of the percentage points of the χ^2 distribution to obtain a 90% CI for μ using the above data. (3)

(v) Compare the CIs you have found in parts (ii) and (iv), and comment briefly. (3)

Turn over

4. In order to compare the effectiveness of two treatments, A and B, in abstracting water from different chemical compounds, each treatment was applied under controlled conditions to random 100-gram samples of 10 different compounds. The weights in grams of water removed were recorded.

Compound No.	1	2	3	4	5	6	7	8	9	10
Treatment A	13.31	17.59	4.14	17.27	10.12	3.70	5.02	7.92	8.76	11.66
Treatment B	13.50	17.92	4.70	17.09	10.48	4.05	4.93	7.87	8.84	12.11

You are given that the total weight of water extracted by treatment A was 99.49 grams and the total weight of water extracted by treatment B was 101.49 grams.

(i) Noting any assumptions made and taking into account the paired nature of the data, carry out a two-tailed parametric test of the null hypothesis that treatments A and B are equally effective on average.

(10)

(ii) Carry out a non-parametric test corresponding to the parametric test in part (i), noting any assumptions needed for this non-parametric test.

(9)

[In parts (i) and (ii) the null and alternative hypotheses should be clearly stated in terms of population quantities.]

(iii) Comment briefly on your results.

(1)

5. In a quality control procedure, a random sample of 10 items is drawn from a large batch of items and tested for defectives. If there are no defective items the batch is accepted, and if there are two or more defective items the batch is rejected. If there is one defective item, a second random sample of 10 items is taken: if there are no defectives in this second sample the batch is accepted, but if there are any defectives in the second sample the batch is rejected.

(i) Show that the overall probability of accepting the batch is

$$P(\text{accept batch} \mid p) = (1-p)^{10}\left[1+10p(1-p)^{9}\right],$$

where p is the batch proportion defective.

(3)

(ii) Calculate

(a) $P(\text{accept batch} \mid p = 0.05)$,

(b) $P(\text{reject batch} \mid p = 0.15)$.

Sketch the operating characteristic curve.

(11)

(iii) The procedure described above is implemented for 100 batches, in all of which 1% of items are defective. Use a suitable approximation to calculate the probability that more than two batches are rejected.

(6)

6. The mean reaction times in tenths of a second, X, of two groups of subjects to a beep stimulus are given below. The subjects in the first group were new to experiments of this type, but those in the second group had taken part in similar experiments before. Assume that the groups of subjects are random samples from large populations of "new" and "experienced" subjects.

New subjects	1.6	3.3	3.0	2.0	2.1	2.3	2.5		$\sum x = 16.8$	$\sum x^2 = 42.40$
Experienced subjects	2.6	2.4	2.2	1.5	1.7	1.9	2.7	1.8	$\sum x = 16.8$	$\sum x^2 = 36.64$

(i) Calculate the mean and variance of each sample. Assuming that in each population reaction times are Normally distributed, test for the equality of variances of the populations represented by these samples against a two-sided alternative.

(5)

(ii) Now also assuming that the population variances are equal, carry out a *t* test of the null hypothesis that there is no difference in the mean times taken by new and experienced subjects, against the alternative hypothesis that the mean time for new subjects is longer.

(6)

(iii) Assume now that the distributions of reaction times in the two populations may have different medians, and may not be Normal. Carry out a non-parametric test of the null hypothesis that the median times of the two distributions are equal, against the alternative that the median time for new subjects is longer than that for experienced subjects.

(6)

(iv) Comment briefly on the results of the tests in parts (ii) and (iii).

(3)

7. The table below shows the frequency distribution of the numbers of attempts required by 200 candidates at a driving school to pass the driving test.

Number of attempts	1	2	3	4	5	6 or more
Number of candidates	70	46	24	18	14	28

It is thought that the number of attempts, X say, that a candidate requires to pass follows a form of distribution such that

$$P(X=x)=q^{x-1}p, \quad x=1,2,3,\ldots,$$

where $0<p<1$ and $q=1-p$.

(i) Use the χ^2 goodness-of-fit test to test the hypothesis that the tabulated data constitute a random sample from a distribution of the given form with $p = 0.4$.

(9)

(ii) Test the hypothesis that the above data follow the given form of distribution with p unspecified. To perform this test p must be estimated from the data and for this purpose you are given that an appropriate estimate of p based on the tabulated data is 1/3.

(8)

(iii) Comment briefly on your results.

(3)

Turn over

8. An experiment is carried out to assess the mean effects of factors F1 (at r levels) and F2 (at c levels) on a quantity of interest. One observation of the outcome is taken at each of the rc factor level combinations.

(i) Identify a suitable analysis of variance procedure to test for differences between the mean effects of levels of either factor. You should assume that raw experimental material is allocated to all factor combinations at random, and that the errors in the data are realisations of independent, zero-mean, constant-variance Normally distributed random variables. Write down and interpret the formal model for your chosen procedure, and show (but do not derive) the partitioning of the total sum of squares into contributions arising from factor F1, factor F2 and error.

(6)

(ii) The table below shows the units of material produced by machines A, B and C when each is operated by 4 technicians, W, X, Y and Z, in each case working for a standard shift.

		Technician				
		W	*X*	*Y*	*Z*	*Row Total*
Machine	*A*	200	186	192	145	723
	B	217	162	186	171	736
	C	190	162	149	150	651
	Column Total	607	510	527	466	2110

Perform an appropriate analysis of variance as in part (i), testing at the 5% significance level, and report your conclusions clearly. You are given that the sum of squares of the observations is 376 700.

(9)

(iii) What is meant by saying that the effects of machines and technicians are assumed to be *additive*? How is the analysis above compromised if this assumption does not hold?

(5)