**EXAMINATIONS OF THE HONG KONG STATISTICAL SOCIETY**



**GRADUATE DIPLOMA, 2005**

**Applied Statistics II**

**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* $\begin{pmatrix} n \\ r \end{pmatrix}$ *is the same as* $^{n}C_{r}$.

1.　Four diets, A, B, C and D, were compared in a specially bred strain of guinea pigs. Four animals, of the same sex, from each of six litters (I to VI) were available. Litters were used as blocks to take out any genetic variation. The table below gives the increases in weight (in suitable units) over a fixed period; the animal from litter IV which should have received diet D showed signs of disease and was removed from the experiment before the final recordings.

| Litter | I | II | III | IV | V | VI | Total |
|---|---|---|---|---|---|---|---|
| Diet A | 24 | 34 | 41 | 27 | 36 | 32 | 194 |
| Diet B | 35 | 38 | 46 | 33 | 37 | 35 | 224 |
| Diet C | 40 | 44 | 54 | 38 | 46 | 40 | 262 |
| Diet D | 29 | 35 | 40 | $x$ | 38 | 34 | $176 + x$ |
| Total | 128 | 151 | 181 | $98 + x$ | 157 | 141 | $856 + x$ |

(i)　Explain briefly why a missing observation in a randomised block experiment makes analysis of the data less straightforward.

(2)

(ii)　A general formula which can be used to estimate a single missing value in treatment $i$ and block $j$ of a randomised block experiment is

$$\frac{tT_i' + bB_j' - G'}{(t-1)(b-1)} .$$

Define $t$, $b$, $T_i'$, $B_j'$ and $G'$ as used in this formula.

(2)

(iii)　Apply this formula to find an estimate of the missing value $x$ in this experiment. State what effect using this estimate will have on the rest of the analysis of these data. What assumptions about the lost observation must be made if this analysis is to be valid?

(5)

(iv)　Given that the residual (error) mean square in an analysis using the estimate of $x$ is 3.7111, find approximate standard errors for the differences between pairs of diet means. Using these, examine whether there are differences between the means of the four diets. State your conclusions.

You are now told that bran was included in the diets, and that the amount of added bran increased from diet A through diets B and C to diet D. Suggest any further analysis that might be useful.

(8)

(v)　Suppose now that the guinea pig receiving diet B in litter I died before the experiment was completed, so that the experiment had two missing values. Explain briefly, without further calculation, how the formula in part (ii) could be applied to estimate two missing values.

(3)

**Turn over**

2.  (i)   Explain what is meant by a complete factorial experimental design for a response variate which is influenced by $m$ factors A, B, C, … each at two levels. Describe briefly the way such a design can be divided into regular blocks of size $2^{m-1}$, illustrating your answer by reference to an example with three factors and two blocks.

[You may use the usual notation for treatment combinations in a 2-level factorial design without further explanation.]

(4)

(ii)  In an experiment to investigate the yield $y$ from a particular chemical process, the effect of four factors A, B, C, D was studied. Each factor had 2 levels and the 16 runs from all possible combinations of levels of the 4 factors were allocated to 4 blocks, each of 4 runs. There were two replicates of the 16 runs and the design (before randomisation) for each replicate is given below.

| Block 1 | (1) | *ab* | *acd* | *bcd* |
|---|---|---|---|---|
| Block 2 | *a* | *b* | *cd* | *abcd* |
| Block 3 | *c* | *ad* | *bd* | *abc* |
| Block 4 | *d* | *ac* | *bc* | *abd* |

(a)   Which interactions are confounded with blocks in this experiment?   (3)

(b)   Write down the sources of variation and associated degrees of freedom in an analysis of variance for the data $y$.   (3)

(c)   You are given that the residual (error) mean square from this analysis of variance is 9705 and that the remaining three- and four-factor interactions (those that were not confounded with blocks) are not significant at the 5% level.

The output **shown on the next page** lists the main effects and two-factor interactions calculated from the results, and the mean yields (in coded units) for each two-factor combination, where "−" and "+" indicate "low" and "high" levels of each factor.

The main effect of a factor is calculated as the difference between the means of the observations at high level and those at low level of the factor. Use the information given by the two-way tables of means to verify that the main effect of A is −77.50. Using a similar method, which you should explain carefully, verify that the "effect" for the BC interaction is 267.125.   (3)

(d)   Calculate the standard error of the factorial effects, and use this to carry out appropriate significance tests on the factorial effects.

It is desired to set the levels of the factors A, B, C, D to achieve as low a value of $y$ as possible. Carry out any further statistical tests that you consider appropriate to suggest levels of A, B, C, D which achieve this objective. State your conclusions clearly.   (7)

3

**Turn over**

**Computer output for question 2 part (ii)**

```
The main effect estimates and two-factor interactions


A    -77.500       AB    1.375       BC    267.125
B   -424.875       AC   41.750       BD  -177.500
C   -193.000       AD   55.375       CD    52.875
D    295.875




Two-way tables of means


          B -           B +
A -     736.500       310.250
A +     657.625       234.125



          C -           C +
A -     640.750       406.000
A +     521.500       370.250



          D -           D +
A -     403.125       643.625
A +     270.250       621.500



          C -           C +
B -     927.125       467.000
B +     235.125       309.250



          D -           D +
B -     460.375       933.750
B +     213.000       331.375



          D -           D +
C -     459.625       702.625
C +     213.750       562.500
```

3.  (i)   It is sometimes stated that two important principles in experimental design are *replication* and *randomisation*. Explain what these are and why they are important.

Suppose that in a drug trial the effect of a new drug is to be measured as the difference between the patient's state after treatment ("outcome") and his state before treatment ("baseline"). Give an example to show that testing whether the mean difference is zero might not always be a suitable way of determining whether the new drug is effective. How might the trial be modified so that the effectiveness of the new drug can properly be tested?

(8)

(ii)  Kidney failure patients are commonly treated on dialysis machines that filter toxic substances from the blood. Effective treatment depends on, among other things, duration of treatment and weight gain between dialysis treatments as a result of fluid build-up.

To study the effects of these two factors on the number of days hospitalised during the past year, a random sample of 10 patients was obtained from each of six groups of patients who had undergone treatment at a large dialysis facility. Treatment duration was categorised as short or long; weight gain between dialysis treatments was categorised as mild, moderate or severe.

The table below shows the number of days each patient was hospitalised in the previous 12 months.

|  | Short duration | | | | | | Long duration | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
|  | *Mild* | | *Moderate* | | *Severe* | | *Mild* | | *Moderate* | | *Severe* | |
|  | 0 | 2 | 2 | 4 | 15 | 16 | 0 | 2 | 5 | 1 | 10 | 15 |
|  | 2 | 0 | 4 | 3 | 10 | 7 | 1 | 7 | 3 | 3 | 8 | 4 |
|  | 1 | 5 | 7 | 1 | 8 | 30 | 1 | 4 | 2 | 6 | 12 | 9 |
|  | 3 | 6 | 12 | 5 | 5 | 3 | 0 | 0 | 0 | 7 | 3 | 6 |
|  | 0 | 8 | 15 | 20 | 25 | 27 | 4 | 3 | 1 | 9 | 7 | 1 |
| Total | 27 | | 73 | | 146 | | 22 | | 37 | | 75 | |

The sum of the squares of the 60 recorded observations is 5050.

(a)   Analyse the data to estimate effects attributable to treatment duration and weight gain, and their interaction. Present appropriate means and report briefly on the results.

(8)

(b)   Discuss any concerns you have about the validity of the analysis. Do you think a transformation might be appropriate for these data? Explain your reason, indicating why you cannot simply take logarithms, and suggest another transformation.

(4)

**Turn over**

4.    In a drilling operation, five factors A, B, C, D, E are thought to be important in influencing the volume of crude oil pumped.  The factors, and levels to be investigated in an experiment, are as follows.

    A :    rotational drill speed        60 rpm,  75 rpm
    B :    longitudinal velocity         50 fpm,  100 fpm
    C :    drill-pipe length             200 ft,  400 ft
    D :    drill-pipe diameter           3 ft,  6 ft
    E :    drilling angle                30 degrees,  60 degrees

It is known that factors A, B and C do not interact with each other, and also that factors C, D and E do not interact with each other.

(i)    Produce a ¼ replicate of a $2^5$ factorial design suitable for fitting a first order response surface.

(7)

(ii)   Explain how you might modify this design so that you could test for lack of fit of the first order response surface.  Outline how the $F$ test statistic would be obtained.

(5)

(iii)  Describe briefly how and when the design you have constructed might be sensibly augmented to form a design suitable for fitting a second order response surface.

(4)

(iv)   Comment on the suitability of $2^{5-2}$ fractional factorial designs as first order designs, illustrating your answer by reference to your experimental designs in parts (i) and (ii).

(4)

**Turn over**

5.  (i)   Two methods sometimes used in sampling are (1) convenience sampling, choosing people available when sampling is in progress, and (2) volunteer sampling, using people who have offered to take part.

    (a)   Discuss the advantages and disadvantages of *simple random* sampling versus *convenience* sampling for selecting a sample for a survey.

    (b)   Explain how *convenience* sampling and *volunteer* sampling differ in their construction. Give an example of volunteer sampling, indicating briefly typical problems likely to be associated with using this method of sampling.

    (6)

    (ii)  A farmer wishes to estimate the total weight of fruit to be produced in a field of zucchini (squash), by sampling plants just prior to harvest. The field consists of 20 rows with 400 plants per row. The total weight (kg) of fruit will be recorded for each plant sampled.

    (a)   The farmer is wondering whether to use simple random sampling or systematic sampling. Give reasons why, for this survey, systematic sampling might be preferred. What might be its drawbacks?

    (2)

    (b)   The farmer aims to estimate the total weight of fruit produced by the field to within 2000 kg, with 95% probability. Previous work shows that the standard deviation of the total weight of fruit on a single plant can be taken to be 2 kg. If the farmer were to use simple random sampling, show that his aim would be achieved with a sample of size about 240.

    (3)

    (c)   The farmer decides to use systematic sampling. Would you consider a 1-in-33 systematic sample to be an appropriate sampling scheme? Explain clearly how you would carry out this sampling scheme, and comment on whether it is reasonable to use the simple random sampling formula you used in part (ii)(b) when calculating the sample size.

    (3)

    (iii) An organisation wishes to commission a survey, taking a simple random sample of $n$ schools from a population of $N$ schools, in order to estimate two quantities: the proportion $P_a$ of schools in the population that have a playing field, and the proportion $P_b$ of students in the population that attend a school with a playing field.

    Explain how you would estimate each of the proportions $P_a$ and $P_b$. Discuss briefly the properties of each estimator.

    (6)

7

6. (i) In the context of stratified sampling, explain what is meant by a stratum. Give three general conditions under which you might decide to use a stratified random sample.

(6)

(ii) In stratified random sampling, the sampling variance of the estimated population mean is minimised for a fixed total sample size, $n$, if

$$n_h = n \frac{N_h S_h}{\sum N_h S_h}.$$

Define the terms $n_h$, $N_h$ and $S_h$ as used in this result.

(2)

(iii) A survey into the cost of house repairs in a certain large district is being planned. The budget for the survey is US$20,000. Two possible designs have been suggested.

Design 1 : a simple random sample at a cost of $4,000 plus $10 per house sampled.

Design 2 : a stratified random sample using 5 classes of house (I to V) as the strata at a cost of $10,000 plus $10 per house sampled.

Rough estimates of the means and standard deviations of the cost of house repairs in each stratum are as follows.

| | Class of house | | | | |
|---|---|---|---|---|---|
| | I | II | III | IV | V |
| Percentage of houses in class | 10 | 10 | 20 | 30 | 30 |
| Mean cost of repairs ($) | 1350 | 1100 | 850 | 600 | 400 |
| Standard deviation of cost ($) | 600 | 400 | 300 | 200 | 160 |

For each design, calculate an estimate of the variance of the estimated mean cost of house repairs for this district. Giving your reasons, state which design you would recommend.

(12)

8

7.    In a recent opinion poll, a simple random sample of 1600 voters in a town was selected from the electoral roll (list of electors) and interviewed, in order to estimate support in the town for the candidates from two parties, A and B, in a forthcoming parliamentary election.   The electoral roll lists 160,000 persons, divided geographically into seven wards (areas) that have distinct characteristics in terms of economic well-being.

(i)     State the advantages and disadvantages of using *simple random* sampling to select this sample of 1600 voters from the electoral roll.   Suggest another sampling method that could be used with advantage to select the voters. Explain what benefits your method has over simple random sampling.

It is nearly a year since the electoral roll was compiled.   Explain what consequences this may have for the survey.

(7)

(ii)    Voters were asked to recall the party they voted for at the last general election, and to say which party they would support if there were a general election tomorrow.  The results are shown below.

| | Opinion poll ($n = 1600$) | |
|---|---|---|
| | *Party A* | *Party B* |
| *Current voting intention* | 720  (45%) | 880  (55%) |
| *Recalled voting behaviour* | 560  (35%) | 1040  (65%) |

The correlation is 0.60 between current voting intention and recalled voting behaviour (if a vote for A is scored as 1 and a vote for B is scored as 0).

You may ignore the effect of any changes that may have taken place in the composition of the electorate since the last election.

(a)    Using information in the recent opinion poll, give an approximate 95% confidence interval for *p,* the current proportion of party A supporters.

(4)

(b)    In the last parliamentary election, 30% actually voted for party A. Explain how one could allow for the sample having too large a proportion of supporters of A by using a regression estimator.  Give an approximate 95% confidence interval for *p,* based on the regression estimate.

(7)

(c)    In practice, a proportion of those people who have changed their vote will state incorrectly that they had previously voted for the party to which they have now switched.  Explain briefly how this would affect the regression estimator.

(2)

[You may assume $\mathrm{Var}\left(\hat{p}_{LR}\right)=\left(1-f\right)\left(1-r^2\right)s_y^2/n$ , using standard notation.]

**Turn over**

8. (a) Explain clearly the differences between infant, neonatal, perinatal and maternal mortality rates and show how they are calculated.

(6)

(b) The age-sex structure of the population of *U*, a developing country, is given (in thousands) below. Draw an age pyramid to illustrate these data.

| Age | Males | Females |
|---|---|---|
| 0 | 35.7 | 34.8 |
| 1 – 4 | 143.9 | 140.0 |
| 5 – 14 | 328.1 | 320.6 |
| 15 – 24 | 202.4 | 216.2 |
| 25 – 34 | 120.7 | 142.2 |
| 35 – 44 | 114.7 | 123.3 |
| 45 – 54 | 93.6 | 87.2 |
| 55 – 64 | 63.5 | 60.4 |
| 65 – 74 | 50.2 | 47.6 |
| 75 – 84 | 9.0 | 12.8 |
| 85 and over | 0.9 | 1.7 |

(6)

(c) (i) The sex-age-specific death rates (per thousand) for *U* and for the population of *D*, a developed country, are given below. Explain clearly how these rates have been calculated.

| | Country D | | Country U | |
|---|---|---|---|---|
| Age | Males | Females | Males | Females |
| 0 | 33.2 | 25.5 | 54.2 | 41.1 |
| 1 – 4 | 1.2 | 1.0 | 3.3 | 3.5 |
| 5 – 14 | 0.6 | 0.4 | 0.9 | 0.6 |
| 15 – 24 | 1.5 | 0.6 | 1.3 | 0.9 |
| 25 – 34 | 1.9 | 1.1 | 2.7 | 1.7 |
| 35 – 44 | 3.7 | 2.2 | 4.1 | 3.1 |
| 45 – 54 | 9.7 | 5.1 | 7.3 | 5.0 |
| 55 – 64 | 22.9 | 11.8 | 15.8 | 9.9 |
| 65 – 74 | 51.6 | 30.7 | 34.5 | 24.5 |
| 75 – 84 | 101.3 | 75.1 | 69.7 | 55.4 |
| 85 and over | 202.6 | 202.5 | 198.5 | 161.8 |

(4)

(ii) Outline the similarities and differences in mortality levels between the two countries, and between males and females.

(4)