THE ROYAL STATISTICAL SOCIETY

2003 EXAMINATIONS – SOLUTIONS

GRADUATE DIPLOMA

APPLIED STATISTICS

PAPER II

The Society provides these solutions to assist candidates preparing for the examinations in future years and for the information of any other persons using the examinations.

The solutions should NOT be seen as "model answers". Rather, they have been written out in considerable detail and are intended as learning aids.

Users of the solutions should always be aware that in many cases there are valid alternative methods. Also, in the many cases where discussion is called for, there may be other valid points that could be made.

While every care has been taken with the preparation of these solutions, the Society will not be responsible for any errors or omissions.

The Society will not enter into any correspondence in respect of these solutions.

Graduate Diploma, Applied Statistics, Paper II, 2003. Question 1

(i) Driver totals are: *A* 173, *B* 151, *C* 201, *D* 163.

"Correction factor" is $\frac{688^2}{16} = 29584$. Therefore total SS = 30042 – 29584 = 458.

SS for drivers = $\frac{173^2}{4} + \frac{151^2}{4} + \frac{201^2}{4} + \frac{163^2}{4} - 29584 = 29925 - 29584 = 341$.

SS for cars = $\frac{181^2}{4} + \frac{171^2}{4} + \frac{161^2}{4} + \frac{175^2}{4} - 29584 = 29637 - 29584 = 53$.

SS for roads = $\frac{182^2}{4} + \frac{174^2}{4} + \frac{164^2}{4} + \frac{168^2}{4} - 29584 = 29630 - 29584 = 46$.

Hence:

SOURCE	DF	SS	MS	*F* value
Cars	3	53	17.67	5.89 significant
Roads	3	46	15.33	5.11 significant
Drivers	3	341	113.67	37.89 very highly sig
Residual	6	18	3.00	$= \hat{\sigma}^2$
TOTAL	15	458		

[All *F* values are compared with $F_{3,6}$; upper 5% point is 4.76, upper 0.1% point is 23.7.]

There are differences between cars and between roads, both significant at the 5% level; these might not look very large differences, but the residual error variability, with which they are compared, is quite small. The difference between drivers is much greater – significant at the 0.1% level – with driver *C* having a relatively large value.

(ii) Combinations of all cars with all roads and all drivers would require 4×4×4 = 64 runs. The Latin square scheme, in 16 runs, allows orthogonal comparisons of the three factors, on the assumption that there are no interactions. A 4×4 square has only 6 degrees of freedom for residual, and often that would not be enough to give a reliable estimate of σ^2; here, however, the estimate is quite small, so a useful analysis has resulted. Using two squares would give ample degrees of freedom for *F* and *t* tests.

(iii) There are four "standard" 4×4 squares (letters in alphabetical order in first row and in first column), one of which must be chosen at random. The rows of this square are then permuted at random, as are the columns, to give a randomised design. The letters *A*, *B*, *C*, *D* are then allocated at random to the "treatments" (drivers). This gives a random choice from all possible 4×4 squares.

See next page for solution to (iv)

(iv) Note that *t* tests would show little difference among *A*, *B*, *D* but a significantly greater amount of wear when *C* is driving.

Contrasts:

	A	*B*	*C*	*D*	Value	Divisor	SS	*F* value
TOTAL	173	151	201	163				
Times of day	–1	1	–1	1	–60	16	225	75.00
Weekday/weekend	–1	–1	1	1	40	16	100	33.33
Interaction	1	–1	–1	1	–16	16	16	5.33

The *F* values are all compared with $F_{1,6}$; upper 5% point is 5.99, upper 1% point is 13.74, upper 0.1% point is 35.51. Thus the result for time of day is very highly significant, that for weekday/weekend is highly significant, and that for interaction is significant.

Morning times (*A*, *C*) give a great deal heavier wear; so do weekdays. But since *C* is different from the others, and *C* drove on weekday mornings, this may explain all of these results; we cannot give any firm conclusions.

Graduate Diploma, Applied Statistics, Paper II, 2003. Question 2

(i) Survival time distributions tend to be skew to the right, roughly speaking lognormal, and so the log transformation is likely to be a better basis for analysis than using untransformed data.

(ii) "Correction factor" is $\frac{4064^2}{36} = 458780.4444$.

Therefore total SS = 525276 – 458780.4444 = 66495.5556.

SS for gases main effect = $\frac{1177^2}{12} + \frac{1372^2}{12} + \frac{1515^2}{12} - 458780.4444 = 4797.7223$.

SS for cyanide main effect =

$$\frac{1557^2}{9} + \frac{1026^2}{9} + \frac{847^2}{9} + \frac{634^2}{9} - 458780.4444 = 51918.4444.$$

SS for treatments = 519688.6667–458780.4444 = 60908.2223 (and hence the SS for interaction is obtained by subtraction).

SOURCE	DF	SS	MS	*F* value	
Gas	2	4797.7223	2398.8612	10.30	Compare $F_{2,24}$
Cyanide	3	51918.4444	17306.1481	74.34	Compare $F_{3,24}$
Interaction	6	4192.0556	698.6759	3.00	Compare $F_{6,24}$
Treatments	11	60908.2223			
Residual	24	5587.3333	232.8056	$= \hat{\sigma}^2$	
TOTAL	35	66495.5556			

The upper 0.1% point of $F_{2,24}$ is 9.34; the main effect of gas is very highly significant.

The upper 0.1% point of $F_{3,24}$ is 9.55; the main effect of cyanide is very highly significant.

The upper 5% point of $F_{6,24}$ is 2.51; the interaction is significant.

There is an interaction, significant at the 5% level, as well as two very large main effects. See parts (iv) and (v) for continuation.

(iii)

Cyanide	(0.16)	(0.80)	(4)	(20)				
TOTAL	1557	1026	847	634	Value	Divisor	SS	*F* value
Linear	–3	–1	1	3	–2948	9×20	48281.6889	207.39
Quadratic	1	–1	–1	1	318	9×4	2809.0000	12.07
Cubic	–1	3	–3	1	–386	9×20	826.7556	3.56

[Note that the effective replication for each total is 9 (sum for G_1, G_2 and G_3).]

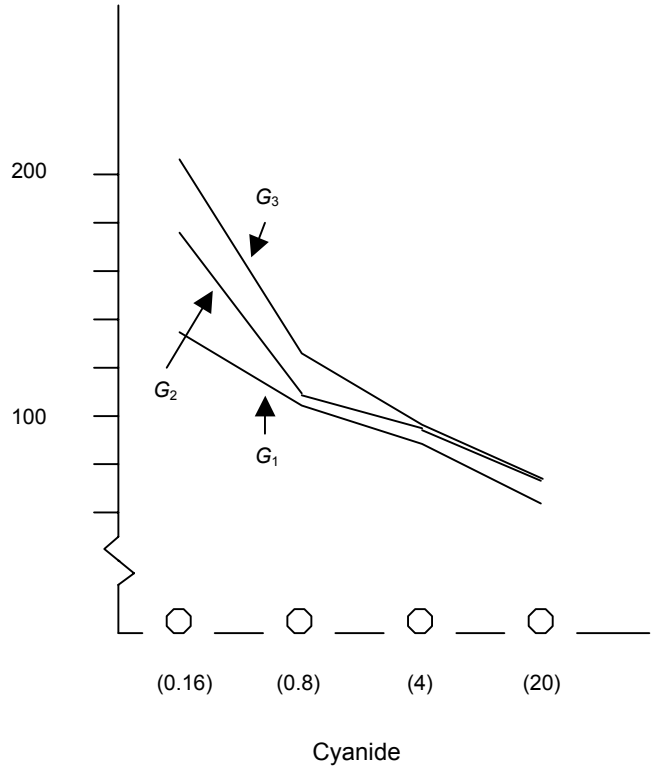
All *F* values are compared with $F_{1,24}$; upper 5% point is 4.26, upper 1% point is 7.82, upper 0.1% point is 14.03. So the linear effect is very highly significant, the quadratic effect is highly significant and the cubic effect is not significant.

See next page for solution to (iv) and (v)

(iv) The means are

	Cyanide concentration			
	0.16	*0.80*	*4*	*20*
G_1	134.7	105.7	89.0	63.0
G_2	177.7	109.7	96.0	74.0
G_3	206.7	126.7	97.3	74.3

Mean survival times



(v) Survival times for given cyanide concentration are always longer for G_3 than for G_2, and for G_2 than for G_1. There is a sharp drop from (0.16) to (0.8) and a less steep drop afterwards.

G_1 is almost linear; the others have a quadratic tendency to begin with. This is the main reason for the gas/cyanide interaction.

G_1 and the highest cyanide level are the best for future use.

Graduate Diploma, Applied Statistics, Paper II, 2003. Question 3

Part (i)

(a) $y_{ijk} = \mu + \tau_i + \beta_j + \varepsilon_{ijk}$

for $i = 0$ to v ($i = 0$ refers to treatment S)

$j = 1$ to b

$$k = \begin{cases} 1 & \text{if } i \neq 0 \\ 1 \text{ to } c & \text{if } i = 0 \end{cases}.$$

μ is the population overall grand mean. τ_i is the population mean effect due to treatment i, β_j is the population mean effect due to being in block j. ε_{ijk} are the residual error terms, assumed mutually independent N(0, σ^2) random variables.

It is necessary to have $\sum_{j=1}^{b} \beta_j = 0$ and $c\tau_0 + \sum_{i=1}^{v} \tau_i = 0$ $\left(\text{i.e. } \sum r_i \tau_i = 0\right)$, where τ_0 refers to S.

(b) Minimise $\Omega = \sum \varepsilon_{ijk}^{\ 2} = \sum_{i,j,k} \left(y_{ijk} - \mu - \tau_i - \beta_j\right)^2$ to obtain least squares estimators. Estimators of μ and β_j can be found, but in fact will not enter the comparison of treatment effects due to the randomised block structure. To find the estimator of τ_i, consider $\frac{\partial \Omega}{\partial \tau_i} = -2\sum_{j,k} \left(y_{ijk} - \mu - \tau_i - \beta_j\right)$ and set this equal to zero. (The result can readily be confirmed to be a minimum by considering second derivatives.) Noting that $\sum_j \beta_j = 0$, and for convenience writing T_i for the total for treatment i, we get that

for $i = 0$, $T_0 = bc\left(\hat{\mu} + \hat{\tau}_0\right)$,

for $i = 1$ to v, $T_i = b\left(\hat{\mu} + \hat{\tau}_i\right)$.

Thus

$$\hat{\tau}_i - \hat{\tau}_0 = \frac{T_i}{b} - \frac{T_0}{bc}.$$

(c) Immediately, we have $\text{Var}\left(\hat{\tau}_i - \hat{\tau}_0\right) = \sigma^2 \left(\frac{1}{b} + \frac{1}{bc}\right)$.

See next page for solution to (ii)

Part (ii)

(a) "Correction factor" is $\frac{306^2}{24} = 3901.5$. Therefore total SS = 5076 – 3901.5 = 1174.5.

SS for blocks = $\frac{100^2}{6} + \frac{88^2}{6} + \frac{67^2}{6} + \frac{51^2}{6} - 3901.5 = 237.5$.

SS for treatments = $\frac{162^2}{8} + \frac{42^2}{4} + \frac{50^2}{4} + \frac{25^2}{4} + \frac{27^2}{4} - 3901.5 = 783.5$.

Hence:

SOURCE	DF	SS	MS	[*F* tests not required]
Blocks	3	237.5	79.17	
Treatments	4	783.5	195.88	
Residual	16	153.5	9.594	$= \hat{\sigma}^2$
TOTAL	23	1174.5		

The variance of the difference between sample means for $S - A$ (or $S - B$ etc) is $\sigma^2\left(\frac{1}{4}+\frac{1}{8}\right) = 3\sigma^2/8$, estimated by $3\times 9.594/8 = 3.598$; thus the standard error is $\sqrt{3.598}$ = 1.90. A 95% confidence interval for one of these differences is therefore given by $\bar{y}_S - \bar{y}_A \pm (2.120)(1.90) = \bar{y}_S - \bar{y}_A \pm 4.03$, where 2.120 is the double-tailed 5% point of t_{16}.

The differences are

$S - A$: 9.75 $S - B$: 7.75 $S - C$: 14.00 $S - D$: 13.50.

Thus the intervals are

$S - A$: (5.72, 13.78)
$S - B$: (3.72, 11.78)
$S - C$: (9.97, 18.03)
$S - D$: (9.47, 17.53).

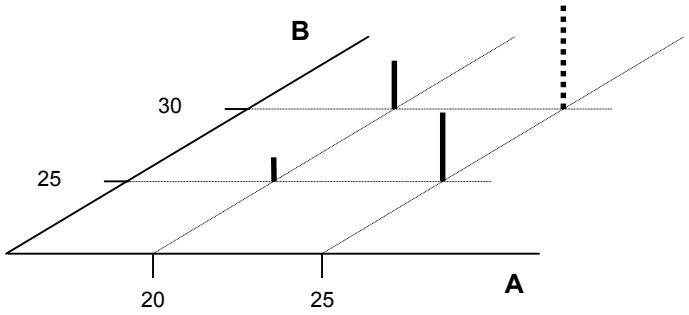
(b) The intervals obtained in (ii)(a) appear to show that all treatments *A*, *B*, *C*, *D* are better than *S*, since no interval contains zero. But the four calculations are not independent of one another, and we have a "multiple comparisons" problem. Another concern is that small percentages, as many of these are, should perhaps be given an arc-sine transformation; however, it is unlikely that inferences would be much different if this were done, on these particular figures. Perhaps homogeneity of variance should be checked using residuals.

Graduate Diploma, Applied Statistics, Paper II, 2003. Question 4

Part (i)

(a) The plant manager has not allowed for the possibility of interaction between the two factors. He will have observed the yield for "*A* 20 *B* 25", for "*A* 20 *B* 30" and for "*A* 25 *B* 25", but not for "*A* 25 *B* 30".

The diagram below indicates, by the vertical bars, what the yields might be at the three observed treatment combinations. It appears that the yield is increasing as *A* increases and as *B* increases. If there is no interaction, the yield at the unobserved combination "*A* 25 *B* 30" would be as indicated by the dashed vertical bar. But if there is interaction, the yield there would be either higher or lower. Unless this combination is actually used in the experiment, we do not know what the yield there is and so we cannot tell whether or not there is any interaction. Putting this another way, we need the observations at all four treatment combinations in order to discover the shape of a response surface in the experimental region *A* 20–25, *B* 25–30.



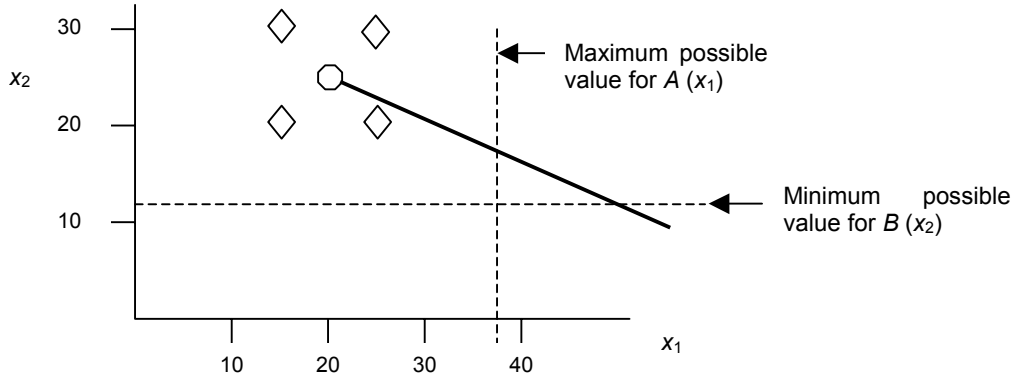
(b) If the current operating settings are near-optimal, exploring a region around them is quite likely to find an optimum, or at least to show in which direction each factor should be changed so as to get nearer the optimum. The replication provides adequate degrees of freedom to test the fit of a response surface model and obtain a good estimate of natural residual variation. But a 2-level experiment only allows an interaction term (x_1x_2) to be included in a model, and not curved terms (x_1^2, x_2^2), where x_1 and x_2 are the levels of *A* and *B*. Some centre points would be needed. If 16 runs cannot be completed under uniform conditions, some blocking may be required.

Part (ii)

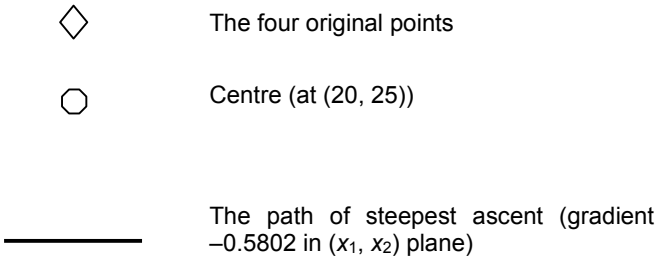
(a) Three parameters have been estimated, i.e. a, b_1 and b_2 in $y = a + b_1x_1 + b_2x_2$, and so there are 13 degrees of freedom for residual. One of these represents interaction (since no x_1x_2 term is in the model) and the others are "pure error". The single degree of freedom sum of squares for interaction can be used as a test of fit of the model, by comparing it with the residual mean square from the 12 "pure error" degrees of freedom.

See next page for solution to (ii)(b), (c), (d)

(b) and (c) Steepest ascent is the path where –0.5875 units are moved in the x_2 direction for every 1.0125 units in x_1. Thus the path has gradient –0.5875/1.1025 = –0.5802 in the (x_1, x_2) plane. The path passes through (0, 0), where this represents the coding of the centre, or current operating, point (20, 25). It is shown on the graph. As a "practical" path, it obviously cannot go to the right of the line representing the maximum possible value for x_1 (or below the line representing the minimum possible value for x_2).



KEY:



We are not told the responses. Those at (25,20) would have been interesting because the path towards the maximum passes nearby.

Useful settings for a follow-up experiment might be x_1 = 25, 30, 35 for which the corresponding x_2 values on the path of steepest ascent are 22.1, 19.2, 16.3. Assuming there was no departure from linearity, these three points could be used in a second (replicated) experiment to fit a linear model. (If the first experiment had suggested non-linearity, a less simple design would have been needed.)

(d) It would have been useful to locate more than one point at the centre. Besides improving the original design, this would have allowed a t test of the null hypothesis "y = 145" to be carried out. About five points could be used at the centre.

Graduate Diploma, Applied Statistics, Paper II, 2003. Question 5

[solution continues on next page]

(i) In random sampling from a population, units are selected by a probability mechanism. Simple random sampling from a finite population gives every item the same probability of selection, but in less simple methods the probabilities need not be the same. For example, in stratified sampling a random sample is taken from each stratum, but the strata are usually of different sizes. Other methods include cluster sampling and multi-stage sampling, in which primary units (for example geographical units such as villages) are selected at random from all those available and these units are either studied completely or subsampled.

Exact estimation methods for means, totals or proportions can be developed for a method of sampling that is based on probability rules. However, these sampling methods require setting up carefully and this can be very time-consuming and expensive.

Non-random sampling methods are usually much quicker, particularly quota sampling in which interviewers are typically sent to central points, such as shopping areas, and given a quota of people to be interviewed. These are specified by characteristics, such as age-group or sex or voting intentions, which can be discovered by a few simple questions so that the specified number (quota) in each sub-group of the population can be obtained. There is no restriction on which actual individuals in each sub-group shall be interviewed, and the easiest to obtain (the most co-operative) will usually be included in the sample. Bias often results from this, and usually also the population to be found in the shopping area (if that is indeed the situation) at the time of the survey is not representative of the whole population of the town. Analysis has to use the methods based on probability because no others are available.

Systematic sampling is done from a population whose members are listed in some standard order (such as alphabetical). It consists of choosing a random starting point at the beginning of the list followed by a regular selection of every kth item, where $k = N/n$ = (population size)/(sample size). Systematic sampling (with random starting point) is much quicker and simpler than pure random sampling. There may be refusals, as in any method of choosing individuals, but this is so in random sampling also. Provided enough is known about possible regular trends in the list used, this method does have a reasonable theoretical base. If there are no trends, a systematic sample might behave as if it were a simple random sample, though strictly speaking it is not. Sometimes the methods for cluster samples can be used for analysis, if there are no trends.

(ii) If n members are selected at random from N, without replacement, the population variance (defined as $\frac{1}{N-1}\sum_{i=1}^{N}\left(X_i - \text{population mean}\right)^2$) is estimated by $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2$.

For the case of a binary variable, each x_i will be either 0 or 1 according as the characteristic being studied is absent or present. Suppose we take a sample of size n and find r individuals with the characteristic, so that r/n is the sample proportion with the characteristic. Then we will have

$$\sum x_i = r.1 + (n-r).0 = r \quad \text{and} \quad \sum x_i^2 = r.1^2 + (n-r).0^2 = r\,.$$

Therefore

$$s^2 = \frac{1}{n-1}\left\{\sum x_i^2 - \frac{\left(\sum x_i\right)^2}{n}\right\} = \frac{1}{n-1}\left(r - \frac{r^2}{n}\right) = \frac{r}{n-1}\left(1 - \frac{r}{n}\right),$$

and now writing $p = r/n$ we have $s^2 = np(1-p)/(n-1)$, as required.

A 95% confidence interval for the population mean is $\bar{x} \pm 1.96s/\sqrt{n}$, assuming n is fairly large. Hence $1.96s/\sqrt{n} \le 1.5$, giving $\frac{1.96}{1.5} \le \sqrt{\frac{n}{168.33}}$, from which we obtain $n \ge 168.33 \times 1.7074 = 287.4$.

Similarly, a 95% confidence interval for the population proportion is $\hat{p} \pm 1.96\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$, so we require $1.96\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \le 0.04$, and from this we obtain $1.96\sqrt{\frac{0.36 \times 0.64}{n}} \le 0.04$ so that $n \ge \frac{(1.96)^2 \times 0.36 \times 0.64}{(0.04)^2} = 553.19$.

Thus we need n at least 554.

(iii) If a population (e.g. a geographical region) can be split into clusters (e.g. towns, villages), sampling can be based on these clusters. Either a random sample of clusters is chosen and these are studied completely, which is "one-stage", or a sub-sample of units may be taken at random for study from each chosen cluster, which is "two-stage". The sample of clusters could be simple random, stratified random or systematic with random starting point.

Stratified sampling splits a population into various groups, according to some specified characteristic such as urban or rural areas, which are expected to be relatively homogeneous within themselves – which clusters might not be. Stratified sampling requires a complete listing of the whole population, whereas cluster sampling only requires that for the chosen clusters (and of course an initial list of clusters). Cluster sampling is often used for administrative convenience, in limiting the area that is to be covered, and in reducing costs, while stratified sampling aims to give a precise estimate of the population parameters through careful choice of homogeneous strata; cluster sampling might not give any better precision than simple random sampling.

In the UK, the Family Expenditure Survey stratifies into quite large geographical areas (by postcode) and uses cluster sampling to locate different communities within the areas.

Graduate Diploma, Applied Statistics, Paper II, 2003. Question 6

Part (i)

(a) $\mathrm{Cov}\left(\hat{R},\bar{x}\right) = E\left(\hat{R}\bar{x}\right) - E\left(\hat{R}\right)E\left(\bar{x}\right) = E\left(\bar{y}\right) - E\left(\hat{R}\right)E\left(\bar{x}\right) = \bar{Y} - \bar{X}E\left(\hat{R}\right).$

This gives $E\left(\hat{R}\right) = -\frac{1}{\bar{X}}\mathrm{Cov}\left(\hat{R},\bar{x}\right) + \frac{\bar{Y}}{\bar{X}}$, i.e. $E\left(\hat{R}\right) - R = -\frac{\mathrm{Cov}\left(\hat{R},\bar{x}\right)}{\bar{X}}$.

(b) $f = \frac{n}{N}$ and $\hat{R} = \frac{\bar{y}}{\bar{x}}$, so that $\hat{R}\bar{x} = \bar{y}$ or $\bar{y} - \hat{R}\bar{x} = 0$.

Hence the estimator of $\mathrm{Var}\left(\hat{R}\right)$ given in the question is

$$\frac{1-f}{n\bar{x}^2}\cdot\frac{1}{n-1}\sum\left\{\left(y_i-\bar{y}\right)-\hat{R}\left(x_i-\bar{x}\right)\right\}^2$$

$$=\frac{1-f}{n\bar{x}^2}\cdot\frac{1}{n-1}\left\{\sum\left(y_i-\bar{y}\right)^2-2\hat{R}\sum\left(y_i-\bar{y}\right)\left(x_i-\bar{x}\right)+\hat{R}^2\sum\left(x_i-\bar{x}\right)^2\right\}$$

$$=\frac{1-f}{n\bar{x}^2}\left(s_Y{}^2-2\hat{R}\hat{\rho}s_Xs_Y+\hat{R}^2s_X{}^2\right)$$

in which $s_Y{}^2$, $s_X{}^2$ are the estimated variances of Y and X, and $\hat{\rho}$ is the estimated correlation coefficient for X and Y.

Part (ii)

The ratio method works well when Y is proportional to X, with the relation passing through the origin. It will <u>not</u> be better than a simple random sample when ρ is less than 0 or when the relation does not pass through the origin (in which case a regression estimator is required instead).

See next page for solution to (iii)

Part (iii)

(a) The sugar content of an individual fruit should be roughly proportional to its weight, in fruit from the same source and batch.

(b) Since we are not told *N*, the total number of oranges, a ratio estimator is used rather than regression. Counting the whole batch would take a very long time for what might be a very small improvement in precision.

$$\sum x = 1975, \quad \sum y = 110.9. \qquad X_T = 820.$$

$$\hat{R} = \frac{\overline{y}}{\overline{x}} = \frac{\Sigma y}{\Sigma x} = 0.05615. \quad \therefore \hat{Y}_T = \hat{R} X_T = 46.045 \text{ (kg)}.$$

We have $\text{Var}\left(\hat{Y}_T\right) = X_T{}^2 \text{Var}\left(\hat{R}\right)$.

Also, on neglecting *f* which will be very small (as *n* is only 10), we have that the value of the estimator of $\text{Var}\left(\hat{R}\right)$ is

$$\frac{1}{10\overline{x}^2} \cdot \frac{1}{9} \sum \left(y_i{}^2 - 2\hat{R} x_i y_i + \hat{R}^2 x_i{}^2 \right)$$

$$= \frac{1}{90} \cdot \frac{1}{(197.5)^2} \left(1268.69 \; - \; 2 \times 0.05615 \times 22194.8 \; + \; (0.05615)^2 \times 392389 \right)$$

$$= \frac{1}{351056.25} (1268.69 - 2492.476 + 1237.133) \; = \; \frac{13.34687}{351056.25}$$

which on multiplying by $(820)^2$ gives that the value of the estimator of $\text{Var}\left(\hat{T}\right)$ is 25.5641, i.e. the standard error is 5.056.

(c) The half-width of the interval, $ts/\sqrt{n}$, is to be less than 2. Thus $s/\sqrt{n} < 1$ and 25 oranges will achieve this approximately.

Graduate Diploma, Applied Statistics, Paper II, 2003. Question 7

[solution to (b) is on next page]

(a) The main points to be mentioned are

- questions need to be in logical order, brief, clear, not worded to point towards any particular answer, nor to annoy respondents
- questions should be relevant to the purposes of the enquiry or designed to provide valuable supplementary information, answerable by ticking a box or giving a fairly brief reply
- questions should be understandable in the local language or by a simple explanation provided for an enumerator
- questions should be laid out to allow easy extraction of data from the responses for analysis
- the questionnaire should have a neat and attractive appearance.

(i) "Please indicate how much is your income each month"

People might well think this is an invasion of privacy – and are unlikely to know exactly, because of tax (etc) deductions. It would be better to provide boxes covering *ranges of income*, to be ticked.

(ii) A: "Would you support an increase in national insurance contributions to be spent on health and education?"

B: "Would you support an increase in national insurance contributions?"

If A is asked first, the answer to it is more likely to be "yes" than if B is asked first followed by various possible purposes for the money.

(iii) "How much alcohol do you drink in an average week?"

OPEN: Please specify _____

CLOSED: Wine [then provide a number of boxes to tick, from 0 upwards, specifying number of glasses]
Spirits [similar]
Beer [this time give measures in pints/half-pints]
Other [this is to cover cider, perry and other alcoholic drinks which should be listed]

(iv) "Do you favour Britain entering the single European Currency?"

Tick-boxes such as
Yes, immediately
Yes, when conditions are right
(etc)
No, never

This is a question with a balanced set of alternatives.

"Do you favour or oppose Britain entering the single European Currency?" is not that; in fact it is not a good question at all.

(b) Direct questions on sensitive and highly personal matters such as teenage use of drugs may lead to refusal to answer or to people giving incorrect answers. This non-response error will most likely bias the estimated proportion of users, and also make the answers more variable than is really the case in the population. Non-responders often tend towards being users rather than non-users. It is unwise to treat the replies of the responders as giving a fully accurate picture.

Demographic and social factors, and responses to other questions on attitudes, may vary between users and non-users. Missing responses can sometimes by satisfactorily imputed by matching non-response people, for these factors and attitudes, with those who have replied.

The usual reasons for non-response in all surveys also apply, such as non-availability at the time an interviewer visits; and these might well not be the same for users as for non-users. By mail or telephone, failure to reply or refusal to participate are common reasons for bias. Reminders by mail sometimes help. A follow-up telephone call could be attempted if there was no reply to a first call, though not if an outright refusal was made.

Graduate Diploma, Applied Statistics, Paper II, 2003. Question 8

(a) Fertility relates to the number of live births a woman has had. (It is thus, in this sense, the "opposite" of childlessness.)

A period analysis considers the births occurring in a specified period of time, usually one year.

A cohort analysis considers all births occurring to a specific group of women, usually to all those born in a particular year or all those married in a particular year.

(b) (i) The crude fetal death rate per 1000 births

$$= \frac{\text{number of fetal deaths}}{\text{total number of births}} \times 1000 .$$

A: $\frac{308}{41309} \times 1000 = 7.456$. B: $\frac{415}{60710} \times 1000 = 6.836$. C: $\frac{209}{33217} \times 1000 = 6.292$

(ii)

Age of mother	Health district		
	A	B	C
< 20	8.331	9.370	7.216
20 – 24	6.951	6.334	5.226
25 – 29	6.913	6.212	5.645
30 – 34	7.385	6.835	6.173
35 +	12.034	8.545	11.985

(iii) The age-adjusted fetal death rate per 1000 births using A as standard

$$= \sum_{\substack{\text{ages} \\ \text{of} \\ \text{mother}}} \frac{\text{total births in A} \times \text{age-specific death rate}}{\text{overall total births in A (i.e. 41309)}} .$$

A: this will be equal to the crude fetal death rate, i.e. 7.456

B: $\frac{280886.541}{41309} = 6.800$

C: $\frac{253837.439}{41309} = 6.145$

District A has the highest rates for all ages of mother except <20. The age-adjusted rate for B is similar to its crude rate because the proportions of total births in the age-groups are similar to A's (the standard). The age-adjusted rate for C is lower than the crude rate because A (the standard) has a smaller proportion of total births to mothers of 35+, for whom the age-specific rate is comparatively high.