THE ROYAL STATISTICAL SOCIETY

2001 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, 2001. Question 1

(a) (i) A randomised complete block design will be suitable. 'Treatments' are the recipes, 'blocks' are days, each block containing every treatment once (hence 'complete'), the order of recipes tested during the day being determined at random. It is assumed that there is no systematic variation from beginning to end of the day, so the units (bakes) in a day (block) will be handled in homogeneous conditions. There could be differences between days.

(ii) If there is systematic (trend) variation through the day this should be removed by a row-and-column design. For example:

		Columns = Days			
		1	2	3	4
Rows = Time of Day	I	D	C	A	B
	II	B	A	C	D
	III	C	B	D	A
	IV	A	D	B	C

A, B, C, D are the four recipes; A will be run first on day 3, second on day 2, third on day 4 and last on day 1, etc. Row differences take out time trend.

(iii) It may not be possible to run every recipe every day, so blocks are no longer complete. A balanced incomplete block allows comparisons between any two recipes to be made with the same precision. The following is such a design:

Day 1 ABC
Day 2 ABD
Day 3 ACD
Day 4 BCD

Each pair, (AB) etc, appears on two of the four days; that is $\lambda = 2$ in the usual notation.

NOTE. If time trend was also important, a Youden square could be used:

		DAY			
		1	2	3	4
TIME	I	A	B	C	D
	II	B	D	A	C
	III	C	A	D	B

Otherwise the order within each day would be randomised.

(b) (i) Shifts are 'blocks'; randomisation is within each of these. The treatments are a 2×3 factorial set for times and temperatures. With a completely randomised design, possible differences between shifts would not be removed, and would inflate residual 'error'.

(ii) $\sum y^2 = 114159,\ \ G = 1623,\ \ N = 24,\ \ \dfrac{G^2}{N} = 109755.375.$

$$\text{S.S. Time} = \frac{1}{8}\left(545^2 + 564^2 + 514^2\right) - \frac{G^2}{N} = 159.250.$$

$$\text{S.S. Temperature} = \frac{1}{2}\left(693^2 + 930^2\right) - \frac{G^2}{N} = 2340.375.$$

$$\text{S.S. Shifts} = \frac{1}{6}\left(420^2 + 382^2 + 416^2 + 405^2\right) - \frac{G^2}{N} = 145.458.$$

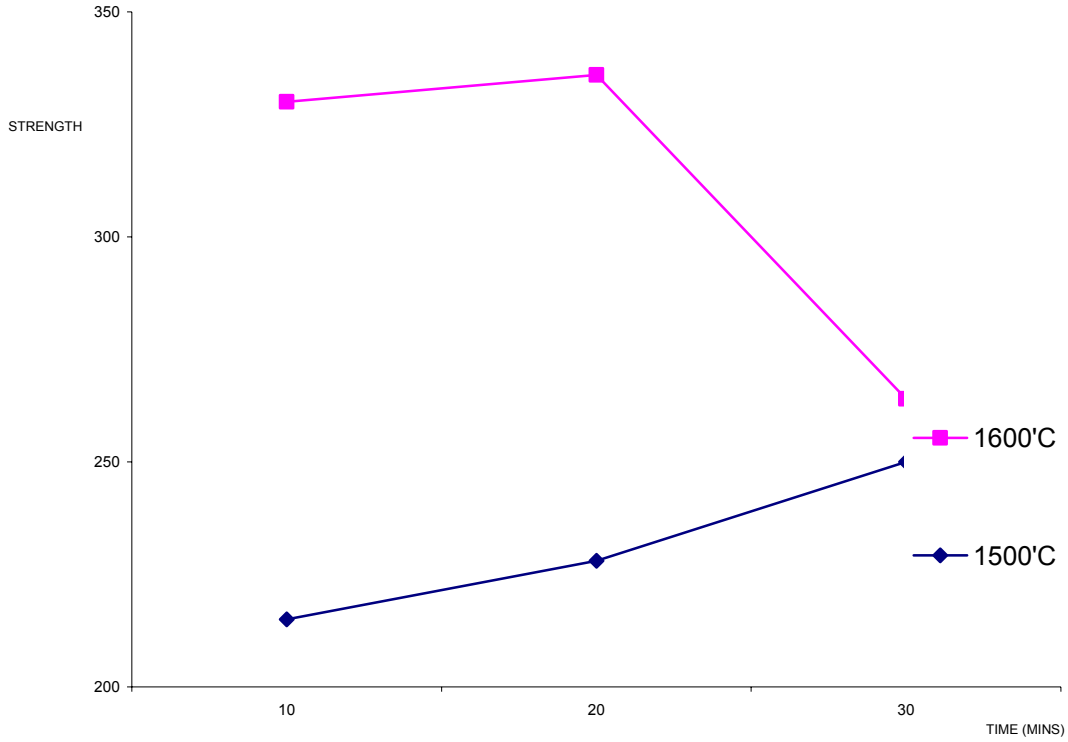
$$\text{S.S. All Treatments} = \frac{1}{4}\left(215^2 + ... + 264^2\right) - \frac{G^2}{N} = 3294.875.$$

Hence the Analysis of Variance:

ITEM	DF	SS	MS	
Shifts	3	145.458	48.486	$F < 1$
Time	2	159.250	79.625	$F \approx 1.2$
Temperature	1	2340.375	2340.375	
Time × Temperature	2	795.250	397.625	$F_{2,15} = 6.19$
	5	3294.875		
Residual	15	963.292	64.219	
TOTAL	23	4403.625		

Comparing 6.19 with $F_{2,15}$ gives a result that is significant at the 5% level, indeed nearly significant at the 1% level.

Shifts did not introduce extra variation. Times on average did not differ but the results of the interaction are to be interpreted, not the main effects. A graphical method using totals is sufficient:



At 1500^{o}C, there is a small, steady increase in strength with time; but at 1600^{o}C the strength at 10 and 20 minutes is much better than at 30 minutes. 1600^{o}C for not more than 20 minutes (nor less than 10 since we have no data) is the recommended combination.

Graduate Diploma, Applied Statistics, Paper II, 2001. Question 2

(i) A contrast is a generalisation of comparing two treatments. It is defined as a linear combination $\sum_{i=1}^{v} c_i T_i$ of totals of the v treatment responses, where $\sum c_i = 0$. If there is equal replication r, the variance of this contrast's value is $\frac{\sum c_i^2 \sigma^2}{r}$ expressed in units of the means, or $\sum r c_i^2 \sigma^2$ in totals. A second contrast $\sum_{i=1}^{v} d_i T_i$ is orthogonal to this if $\sum_{i=1}^{v} d_i = 0$ and $\sum_{i=1}^{v} c_i d_i = 0$.

If the total sum of squares for treatments is split into a set of v orthogonal contrasts, these will add to the total and will be mutually independent, so each can be tested as $F_{1,f}$ against the residual which has f df.

(ii) (a)

		C_1	C_2	S_3	S_6	S_{12}	A_3	A_6	A_{12}
Sulphur/no sulphur		3	3	−1	−1	−1	−1	−1	−1
Spring/Autumn		0	0	1	1	1	−1	−1	−1
Sulphur levels	L	0	0	−1	0	1	−1	0	1
	Q	0	0	1	−2	1	1	−2	1
Sulphur × time	L	0	0	−1	0	1	1	0	−1
	Q	0	0	−1	2	−1	1	−2	1
Controls		1	−1	0	0	0	0	0	0

[Note. C_1 and C_2 are the controls. S represents spring, A autumn. Subscripts 3, 6, 12 are the amounts of sulphur (the controls have no sulphur).]

Using the linear and quadratic components with coefficients (–1, 0, 1) and (1, –2, 1) has treated the levels as being on a logarithmic scale. For the interaction of one of these with time, simply change the relative signs of S and A (the main effects, with the same signs, being averages). It is doubtful whether the final contrast has any meaning, but it may be compared with residual as a check against unexpected variability.

(b) Using totals, these take the values (in the same order as above)

223, 14, 8, 50, 40, 118, –31

[Note. The signs may be either + or – depending on which way the contrast is written down; both are correct.]

Since we are not given the data, we cannot carry out a full ANOVA, and must use the means. The estimate $\hat{\sigma}^2 = 30$, given; and $r = 4$.

Residual will have 21 df, after removing blocks (3df) and treatments (7df); The 5%, 1% and 0.1% points of t_{21} are 2.080, 2.831 and 3.819 respectively.

Assuming that the data followed a Normal distribution with constant variance, the value of a contrast ÷ its standard error will follow t_{21}.

The total for each treatment has variance $r\sigma^2$, and so for a contrast using totals the variance is $r\sigma^2 \sum c_i^2$, i.e. the SE is $2\sigma\sqrt{\sum c_i^2}$ for this example, and for $\sigma^2 = 30$ this is $10.95445\sqrt{\sum c_i^2}$. The values of $\sum c_i^2$, in order, are 24, 6, 4, 12, 4, 12, 2. Hence

Contrast	1	2	3	4	5	6	7
t_{21} =	4.155	0.522	0.365	1.318	1.826	3.110	–2.001

The addition of sulphur has a highly significant benefit; and the sulphur × time interaction is curved, not linear. (A plot of the totals would help to show this.)

Graduate Diploma, Applied Statistics Paper II, 2001. Question 3

(i)

	A	*B*	*C*	*D*	*E*
Mean	394.75	413.00	86.75	37.75	35.25
SD (3df)	57.90	94.17	47.96	33.54	28.04

The mean and SD tend to increase together, though not very regularly. There may be need for a variance-stabilising transformation.

(ii) An analysis must assume constant variance, as well as Normality of distribution and model additivity and adequacy. Bartlett's test could be used to test the NH: $\sigma_A^2 = \sigma_B^2 = \sigma_C^2 = \sigma_D^2 = \sigma_E^2$; unfortunately it is not very sensitive and is affected by any non-Normality. Residuals after fitting 'blocks' and 'treatments' could be checked for evidence of Normality.

The correction term $\frac{G^2}{N} = \frac{3870^2}{20} = 748845.0$.

Blocks SS $= \frac{1}{5}\left(1088^2 + 982^2 + 1017^2 + 783^2\right) - \frac{G^2}{N} = 10244.2$.

Treatments SS $= \frac{1}{4}\left(1579^2 + \ldots + 141^2\right) - \frac{G^2}{N} = 597514.0$.

The Analysis of Variance is

ITEM	DF	SS	MS	
Blocks	3	10244.2	3414.73	$F \approx 1$ n.s.
Treatments	4	597514.0	149378.50	$F_{4,12} = 45.90$
Residual	12	39054.8	$3254.57 = \hat{\sigma}^2$	
TOTAL	19	646813.0		

Comparing 45.90 with $F_{4,12}$ gives a highly significant result.

(iii) The estimated variance of the difference between two treatments means is $\frac{2\hat{\sigma}^2}{4}$ which is 1627.285, standard error 40.34.

Means are:

A	*B*	*C*	*D*	*E*
394.75	413.00	86.75	37.75	35.25
(_____	_____)	(_____	_____	_____)
	∨		∨	
	n.s.		n.s.	

(tested as t_{12})

They clearly form two groups.

(iv) If the relation between mean and variance is roughly $\mu \propto \sigma^2$, then a square root transformation will stabilise variance for analysis. (This does not look very convincing, but may improve basic assumptions.) The four values of $\sqrt{y}$ for *A* are 20.928, 21.024, 17.861, 19.494 and their mean is 19.827, which transforms back to 393.1, very much as before.

(v) $\hat{\sigma}^2$ is now 4.066, and $\sqrt{\frac{2\hat{\sigma}^2}{4}} = 1.426$. $t_{12,5\%} = 2.179$, so the least significant difference is 3.107.

A	*B*	*C*	*D*	*E*
19.8	20.2	9.1	5.8	5.6
(_____	_____)		(_____	_____)
	∨			∨
	n.s.			n.s.

There is now a suggestion that *C* is less effective (more poppies) than *D* and *E*: these are the two best treatments, while *A* and *B* are very poor.

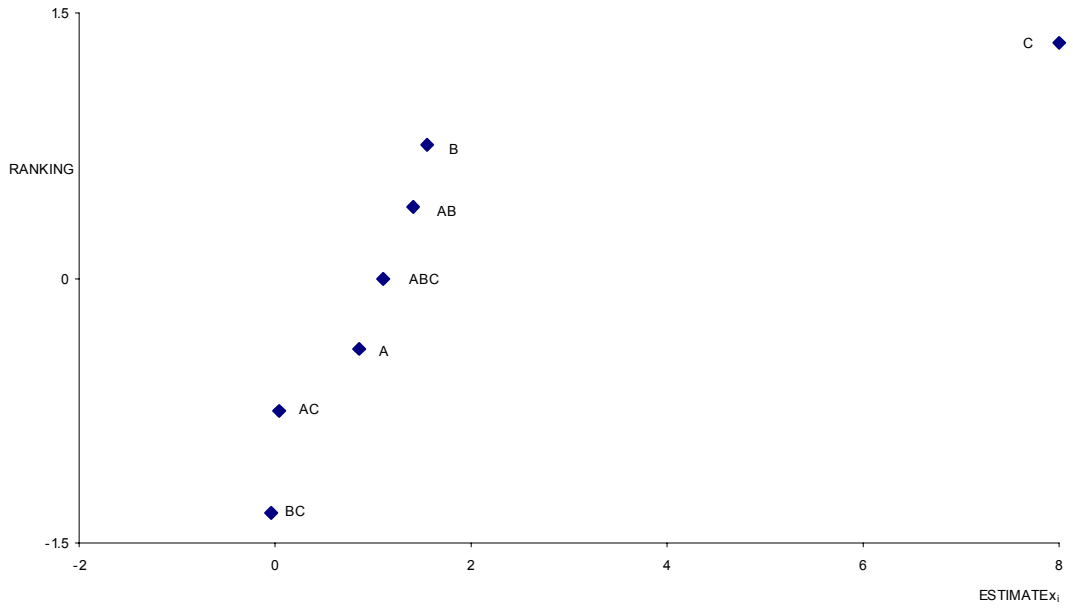
(vi) The only change is in the situation for *C*. In order to decide whether the basic assumptions are now better satisfied, plots of residuals should be checked for evidence for or against constancy of variance, Normality, and lack of relation between residual and size of fitted value. The transformation has improved the validity of the analysis, but it may be that it is not the best one.

Graduate Diploma, Applied Statistics, Paper II, 2001. Question 4

(a) In response surface analysis, when an experimental region has been identified whose centre (coded O) is thought to be near the maximum (or minimum) response, and k factors are being studied, a central composite design is suitable for fitting a quadratic model which allows the turning point to be located. It consists of a 2^k factorial (or fractional factorial), at points coded ±1 equidistant from O, together with several centre points (0, 0, …, 0) and a set of "axial" points ($\pm\alpha$, 0, …, 0), (0, $\pm\alpha$, 0, …, 0), …, (0, …, $\pm\alpha$). A central composite design can be built up from the first-order 2^k design by adding central and axial points. Blocking may be used to eliminate any changes in experimental conditions between the first-order design and later additions; it is possible to arrange for block parameters to be estimated independently of model parameters.

(b) (i) $$ABC = \frac{(abc + a + b + c) - (ab + bc + ac + (1))}{4} = \frac{64.8 - 60.4}{4} = 1.10 .$$

(ii) On the hypothesis that no factors have any effect, the estimates of main effects and interactions will be $\mathrm{N}\left(0, \frac{4\sigma^2}{n}\right)$, giving an approximately straight line. Points away from the line indicate real effects.



C is clearly a substantial effect. AC may appear to be off the line which fits the others, but not enough to warrant further study. It is generally those off the line at either end which need examination.

(iii) $y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + \varepsilon_i$, $\varepsilon_i \sim \mathrm{N}\left(0, \sigma^2\right)$, i.i.d.

$\hat{\beta}_0 = \text{grand mean} = \frac{1}{8}(125.2) = 15.65.$

$\hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_3$, with coding ± 1 for levels as here, are one-half of the main *A*, *B*, *C* effects. Hence the model is $y_i = 15.65 + 0.425x_1 + 0.775x_2 + 4.0x_3$. It does not fit *ac* or *bc* very well, nor *abc*, so some further terms may be needed.

(iv) If $\bar{y}_f$ and $\bar{y}_c$ are the means of n_f and n_c points in the factorial and centre parts of a design, than a SS for curvature is $n_f n_c \frac{\left(\bar{y}_f - \bar{y}_c\right)^2}{\left(n_f + n_c\right)}$.

$\bar{y}_f = 15.65$, as above. With $n_c = 4$, we can estimate σ^2 as these values' variance.

$\hat{y}_c = 15.825$ and $\hat{\sigma}^2 = 0.1092$ with 3 d.f.

SS curvature $= \frac{8 \times 4}{12}(15.65 - 15.825)^2 = 0.0817, < \hat{\sigma}^2$, so no indication of a pure quadratic effect.

For lack of fit, the SS for interaction is $2\left(1.4^2 + 0.05^2 + \left(-0.05\right)^2 + 1.1^2\right) = 6.35$.

$4\hat{\sigma}^2 = 0.4368$, and $F_{4,3} = \frac{6.35}{0.4368} = 14.54$, significant at the 1% level, giving evidence of lack of fit.

(v) There are 12 runs to be made, the 2^3 factorial plus 4 centre points. Confound the *ABC* interaction between days, so that *a*, *b*, *c*, *abc* and 2 centre points occur on Day 1 (or 2 – choice at random) and (1), *ab*, *ac*, *bc* and 2 centre points on the other day. The *ABC* interaction is the least likely to be important, and the centre points will help to assess possible block (day) effects.

(vi) A further block with the axial points $\begin{pmatrix} \alpha & 0 & 0 \\ 0 & \alpha & 0 \\ 0 & 0 & \alpha \\ -\alpha & 0 & 0 \\ 0 & -\alpha & 0 \\ 0 & 0 & -\alpha \end{pmatrix}$ may be used, if no day effect is evident.

It will be rotatable if $\alpha = 2^{\frac{3}{4}} = 1.682$.

Graduate Diploma, Applied Statistics, Paper II, 2001. Question 5

Credit is always given for good examples, especially from personal experience.

(a) (i) Face-to-face interviews will always be appropriate in quota sampling, and will allow interviewers to decide in advance whether to approach a particular person or not, to fill their quota of age-groups etc. They are also desirable when questions may need some explanation, or in situations where communication is difficult, literacy is low, or a pilot survey is being tested out. In these cases the danger of questions not being understood, or of people refusing to respond, can be reduced.

(ii) Telephone interviews may be suitable if people are willing to give time to answering questions which appear to have little relevance to them, and the problems of obtaining a sample that really represents a population are less important. They can also be used for specialised enquiries of a small, well-defined population, e.g. a particular area of industry or business, when there may be someone in an office who can give the information required, provided they know the agency carrying out the survey and do not have sensitive commercial information that they are not willing to give. A weakness is that refusal to respond is easy, also that different parts of a population will not all be easily available at the same time of day.

(iii) When a reasonably short, clear and simple set of questions is to be asked, especially on a regular basis from a population used to dealing with them, postal questionnaires should achieve a fairly high response rate. Professional bodies, industry and commerce, educational organisations can often use these successfully, with at most a small number of reminders for non-response. It may be possible to visit some non-respondents later. General surveys containing a large number and assortment of questions will not usually gain a high percentage of responses.

(b) Sensitive questions asked directly will often not be answered, at least not truthfully. Surveys including such questions will be very unreliable and very likely biased.

Randomised response allows the respondent to keep their own situation unknown to the interviewer. By a suitable confidential randomisation method (such as selecting a sealed envelope from a box), a proportion P of a population can be presented with a statement "I am HIV positive" to which they reply 'Yes' or 'No', and the remaining $(1 - P)$ have the opposite, to which they also reply 'Yes' or 'No'. The interviewer records Yes/No without knowing which statement a person has received. By reducing failure rates to gain responses, this method can be more precise and less biased.

The alternative statement, instead of being the opposite of the first, can be something with known probability in the population, such as month of birth [e.g. 'I was born in May' (Yes/No)].

(c) Income is a sensitive issue. Refusal or untrue answers are likely. The question can only be answered accurately by people whose income is paid regularly, on some form of salary scale, not depending on casual earnings, bonuses, tips etc.

Salary scales are usually quoted as an annual, not weekly, figure.

The survey only includes married women, not cohabiting "partners".

A possible series of questions might be:

1. Are you married or living with your partner?
   Married Partner No

2. Is your husband or partner in paid employment?
   Full time Part time No

3. [Here give a list of classifications of work, e.g. Professional, Manual, … .]

4. Do you know how much your husband or partner earns?
   If so, tick the box which indicates his total annual income:
   Less than £5000
   ……[go up in £5000 steps to, say, £30000, then in £10000 steps, and finish with the following]
   £50000 to £59999
   £60000 and above.

   (Make the classes distinct, not £10000 – £15000, £15000 – £20000 etc.)

Graduate Diploma, Applied Statistics, Paper II, 2001. Question 6

(i) (a) $\sum y_i = (200\times 0)+(240\times 1)+(50\times 2)+(10\times 3) = 370$.

$\sum y_i^2 = (200\times 0)+(240\times 1)+(50\times 4)+(10\times 9) = 530$.

$$s_Y^2 = \frac{1}{499}\left(530 - \frac{370^2}{500}\right) = \frac{256.2}{499} = 0.51343.$$

Estimate $\bar{y} = \frac{370}{500} = 0.74$. Approximate 95% limits are $\bar{y} \pm 1.96\sqrt{\frac{(1-f)s_Y^2}{n}}$.

$$\sqrt{\frac{(1-f)s_Y^2}{n}} = \sqrt{\frac{0.95\times 0.51343}{500}} = \sqrt{0.0009755} = 0.0312.$$

Limits are 0.74 ± 0.0612, i.e. 0.6788 to 0.8012, so for the whole population multiply by 10000 to give 6790 to 8010 approximately. The point estimate is $20 \times 370 = 7400$.

(b) $$\hat{R} = \frac{\bar{y}}{\bar{x}} = \frac{\sum y_i}{\sum x_i} = \frac{370}{\sum x_i}.$$

Now, $\sum x_i = (1\times 130)+(2\times 300)+(3\times 45)+(4\times 20)+(5\times 5) = 970.$

Hence the point estimate is 0.381.

$\sum x_i^2 = (1\times 130)+(4\times 300)+(9\times 45)+(16\times 20)+(25\times 5) = 2180$.

$$s_X^2 = \frac{1}{499}\left(2180 - \frac{970^2}{500}\right) = \frac{298.2}{499} = 0.59760.$$

$$s_{XY} = \frac{1}{499}\left(795 - \frac{970\times 370}{500}\right) = \frac{77.2}{499} = 0.15471$$

$$\mathrm{Var}\left(\hat{R}\right) = \frac{0.95\left(\left\{0.381^2\times 0.5976\right\} - 2\times 0.381\times 0.15471 + 0.51343\right)}{500\times\left(\frac{970}{500}\right)^2} = \frac{0.45817}{500\times 1.94^2}$$

$$= 0.00024347,$$

so $\mathrm{SE}\left(\hat{R}\right) = 0.0156$.

Approximate 95% limits to the ratio are $0.381 \pm 1.96\times 0.0156$, i.e. 0.381 ± 0.031 or 0.350 to 0.412.

(c) Households have one or more cars per adult if $y_i \geq x_i$. The relevant groups (x, y) are (1, 1), (1, 2), (1, 3), (2, 2), (2, 3) and (3, 3); total 109.

Hence $\hat{p} = \dfrac{109}{500} = 0.218$.

95% limits to p are $\hat{p} \pm 1.96\sqrt{\dfrac{\hat{p}(1-\hat{p})}{n}} = 0.218 \pm 1.96\sqrt{\dfrac{0.218 \times 0.782}{500}}$

$= 0.218 \pm 0.036$, i.e. (0.182, 0.254).

(ii) $\hat{\bar{Y}} = \hat{R}\bar{X} = 0.381 \times 1.8 = 0.6858$, hence $\hat{Y} \approx 6860$.

$\text{Var}\left(\hat{\bar{Y}}\right) = \bar{X}^2 \text{Var}\left(\hat{R}\right) = (1.8)^2 \times 0.00024347 = 0.0007888$, and SE = 0.0281.

95% limits now are $0.6858 \pm 1.96 \times 0.0281$ or 0.6858 ± 0.0550,

i.e. 0.6308 to 0.7408.

$\therefore$ revised limits for $\hat{Y}$ are 6310 to 7410.

The estimate of $\bar{x}$ was higher than the actual value, which affected both the estimated mean and variance.

Graduate Diploma, Applied Statistics, Paper II, 2001. Question 7

(i) In stratified sampling, a population is divided into groups (strata) and the strata each have a simple random sample taken from them. If proportional allocation is used, the fraction of the stratum population that is sampled is the same in every stratum, i.e. $\frac{n_i}{N_i}$ is the same for all *i*. So it is here, equal to $\frac{1}{6}$.

(ii) Simple random samples within strata yield unbiased estimates of means, $\bar{y}_i$. Weighting these by stratum sizes gives the overall estimate $\bar{y}_{st} = \frac{1}{N}\sum_{i=1}^{M} N_i \bar{y}_i$ ($M = 4$, the number of strata).

$\mathrm{Var}(\bar{y}_{st}) = \frac{1}{N^2}\sum_{i=1}^{M} N_i^2 \mathrm{Var}(\bar{y}_i) = \frac{1}{N^2}\sum_{i=1}^{M} N_i^2 \frac{S_i^2(1-f_i)}{n_i}$; *SE* is square root of this.
(This simplifies when proportional allocation is used.)

(iii) $N = 120$.

$$\bar{y}_{st} = \frac{1}{120}\{(24\times 99.3)+(36\times 100.0)+(30\times 98.0)+(30\times 100.0)\} = \frac{11923.2}{120} = 99.36.$$

$$\mathrm{Var}(\bar{y}_{st})$$
$$= \frac{1}{120^2}\left[\left\{\frac{24^2\times 9^2}{4}\times\frac{5}{6}\right\}+\left\{\frac{36^2\times 7.46^2}{6}\times\frac{5}{6}\right\}+\left\{\frac{30^2\times 6.28^2}{5}\times\frac{5}{6}\right\}+\left\{\frac{30^2\times 10.61^2}{5}\times\frac{5}{6}\right\}\right]$$

= 2.9541, and hence SE = 1.719.

Approximate 95% limits for the true mean are $99.36 \pm 1.96\times 1.719$
$= 99.36 \pm 3.37$ or 95.99 to 102.73.

(iv) $\mathrm{Var}(\bar{y})$ by simple random sample $= \frac{5}{6}\times\frac{7.75^2}{20} = 2.5026$.

Limits now are $99.36 \pm 1.96\sqrt{2.5026} = 99.36 \pm 3.10$ or 96.26 to 102.46.

(v) Ratio of variances = efficiency = $\frac{\text{Var}(\bar{y})}{\text{Var}(\bar{y}_{st})} = \frac{2.5026}{2.9541} = 0.847$ (or 84.7%).

Stratification may not have been well chosen, since within the same chain the sales can vary greatly. Size of store, reflecting size of turnover, may have been a better choice.

(vi) Strata should be internally homogeneous. Construction can be on the basis of past records of the variable being studied, or of something closely correlated to it. Any major variation should be between strata, not within.

Graduate Diploma, Applied Statistics, Paper II, 2001. Question 8

(a) The life table describes the survival pattern of a group of individuals throughout life to the age-specific death rates currently observed in a particular community. It is a convenient summary of current mortality rather than a description of the actual mortality experience of any group.

A current life table summarises current mortality and may be used as an alternative to standardisation for comparing mortality patterns in different communities.

A cohort life table describes the actual survival experience of a group or cohort of individuals born about the same time.

(b)

Age	$_{10}q_x$	l_x	$_{10}d_x$	$_{10}L_x$	T_x	e_x
0	0.250	1000	250	8750	54390	54.39
10	0.024	750	18	7410	45640	60.85
20	0.040	732	30	7170	38230	52.23
30	0.051	702	36	6840	31060	44.25
40	0.062	666	42	6450	24220	36.37
50	0.091	624	57	5955	17770	28.48
60	0.172	567	98	5180	11815	20.84
70	0.335	469	157	3905	6635	14.15
80	0.624	312	195	2145	2730	8.75
90	1.000	117	117	585	585	5.00
100		0	0	0	0	0

l_x = number attaining age x (of each year's cohort)

$_{10}d_x$ = number dying within 10 years of attaining age x (= $l_x \, {}_{10}q_x$)

$_{10}L_x$ = number living between ages x and $x + 10$ $\left(= 10 \times \frac{\{l_x + l_{x+10}\}}{2} \right)$

T_x = number of persons aged x or greater in a life table $\left(= {}_{10}L_x + {}_{10}L_{x+10} + ... \right)$

e_x = average future lifetimes of persons aged x $\left(= \frac{T_x}{l_x} \right)$.

(i) Age distribution = $100\dfrac{10L_x}{54390}$.

Age	%		Age	%
0 –	16.09		60 –	9.52
10 –	13.62		70 –	7.18
20 –	13.18		80 –	3.94
30 –	12.58		90 –	1.08
40 –	11.86		100 –	0
50 –	10.95			

(ii) Expected age at death, when present age is x, is $x+\dfrac{T_x}{l_x}=x+l_x$; so for age 20 this is 72.23.

(iii) Life expectancy is (ii) at age 0: $0+\dfrac{54390}{1000}=54.39$.

(iv) Uniform death rate within each 10-year group; unlikely in later life. Same death rates in both sexes, again unlikely. Epidemics may make a slight difference; probabilities presumably based on a large amount of data so this is accounted for to some extent.

(c) Population *B* started with a cohort $(1+0.01)^{x+5}$ times the number in the cohort of Population *A*. A suitable form of the age distribution of *B* that would permit comparison with that for *A* is $100\times\dfrac{(1+0.01)^{-(x+5)}\,{}_{10}L_x}{\sum_i (1+0.01)^{-(x_i+5)}\,{}_{10}L_{x_i}}$.

The increasing birth rate in *B* will lead to higher proportions in the lower age groups as compared with *A*.