THE ROYAL STATISTICAL SOCIETY

2004 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, 2004. Question 1

Part (i)

← N	I				II	III	IV	S →
DOOR	*ab*	(1)	*bc*	*b*				DOOR
	c	*abc*	*a*	*ac*				

There is likely to be a "climatic trend" from north to south, even in a glasshouse, increased by having doors at each end which will produce temperature changes when opened. Blocking in this direction, as shown, is therefore a good property of the design. The eight treatment combinations will be randomised in each block, independently of one another.

Part (ii)

(a) The remaining sums of squares are calculated as follows. We need the grand total, 304.0, and hence the "correction factor" $304.0^2/32 = 2888$.

$$\text{SS for blocks} = \frac{68.8^2}{8} + \frac{81.8^2}{8} + \frac{83.3^2}{8} + \frac{70.1^2}{8} - 2888 = 2909.6975 - 2888 = 21.6975 .$$

$$\text{SS for } A = \frac{(217.1 - 86.9)^2}{32} = 529.75125 .$$

$$\text{SS for } ABC = \frac{(155.8 - 148.2)^2}{32} = 1.80500 .$$

(We may check that the sums of squares for all seven main effects and interactions add up to the stated treatments total of 616.795.)

By subtraction, the residual SS = total SS – treatments SS – blocks SS = 36.7875.

Each main effect and interaction has 1 degree of freedom, giving 7 in all for the treatments, and the residual has 21.

Solution continued on next page

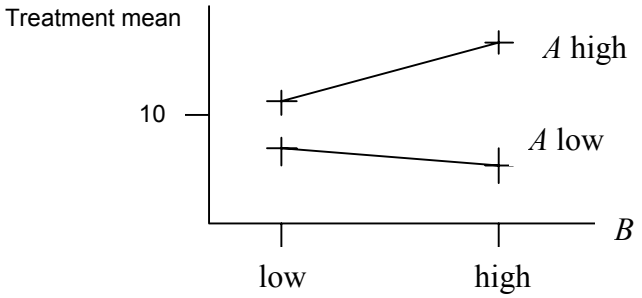
Hence:

SOURCE	DF	SS	MS	*F* value
Blocks	3	21.6975	7.2325	4.13 compare $F_{3,21}$
A	1	529.75125	529.75125	302.40 compare $F_{1,21}$
B	1	19.84500	19.84500	11.33 …
C	1	1.05125	1.05125	0.60 …
AB	1	62.16125	62.16125	35.48 …
AC	1	0.98000	0.98000	0.56 …
BC	1	1.20125	1.20125	0.69 …
ABC	1	1.80500	1.80500	1.03 …
Treatments	7	616.7950		
Residual	21	36.7875	1.7518	$= \hat{\sigma}^2$
TOTAL	31	675.2800		

(b) $F_{1,21}$ tests for the main effects and interactions (upper 5% point is 4.32) show that *A*, *B* and *AB* are significant. The blocks effect is also significant (upper 5% point of $F_{3,21}$ is 3.07).

To study the effects of *A* and *B* in the presence of an *AB* interaction, we need the table of *AB* means:

	A low	*A* high	[Treatments included]	
B low	6.04	11.39	(1), *c*	*a*, *ac*
B high	4.83	15.75	*b*, *bc*	*ab*, *abc*



(c) The decision to include blocking was wise. There is evidence that the blocks are not all the same as each other, and we see that the two end blocks performed less well than those in the centre.

Because the main effect of *C* was not significant and there were no interactions involving *C*, we may conclude that it does not matter which of the two experimental levels of *C* is used in practice. We may, of course, still wish to explore higher or lower levels in a later experiment.

Factors *A* and *B* interact, so their main effects should not be examined alone. Instead, we refer to the table of *AB* means. Clearly use of the high level of *A* has a beneficial effect, and this is increased by using the high level of *B*. On the other hand, *A* does not perform well at the low level and is even worse at this level if the high level of *B* is used.

Graduate Diploma, Applied Statistics, Paper II, 2004. Question 2

Part (i)

In the usual notation, we have $v = 5$, $b = 10$ and $N = 30$, so $r = 6$ and $k = 3$. Thus $\lambda = r(k-1)/(v-1) = 3$.

The required design may be achieved by the following pattern of blocks, where each letter refers to one of the treatments:

Block 1	A B C
Block 2	A B D
Block 3	A B E
Block 4	A C E
Block 5	A D E
Block 6	A C D
Block 7	B C D
Block 8	B C E
Block 9	B D E
Block 10	C D E

Within each block, the order of the treatments allocated should be randomised, with a fresh randomisation for each block.

Part (ii)

(a) The batches are the blocks. This design has $v = 5$, $b = 10$ and $N = 20$, so $r = 4$ and $k = 2$. Thus $\lambda = r(k-1)/(v-1) = 1$.

The overall mean is $\bar{y} = 417/20 = 20.85$.

Let T_i represent the total for the ith treatment and $B^{(i)}$ the total of all batches in which treatment i appears. Also define $Q_i = kT_i - B^{(i)}$.

Treatment	kT_i	$B^{(i)}$	Q_i	$Q_i/(v\lambda)$	$\bar{y}' = \bar{y} + Q_i/(v\lambda)$
A	186	166	20	4.0	24.85
B	160	167	–7	–1.4	19.45
C	90	138	–48	–9.6	11.25
D	250	194	56	11.2	32.05
E	148	169	–21	–4.2	16.65
					Adjusted treatment means ↑

Solution continued on next page

(b) The given information shows that the total (corrected) sum of squares is 1352.55, and this will have 19 degrees of freedom. Also the residual after fitting batches and treatments is 171.50.

We need to find the *unadjusted* blocks (batches) sum of squares, and we can then find the *adjusted* treatments sum of squares.

The unadjusted batches sum of squares can be found from the given information as 1352.55 – 804.50 = 548.05. Alternatively, it can be calculated as $\Sigma B_j^2/2 - G^2/N$ where B_j is the total for batch j (e.g. 51 for batch 1) and G is the grand total (417). Alternatively again, it can be calculated directly as $(v-1)\Sigma Q_i^2/\{rvk(k-1)\}$.

Hence:

SOURCE	DF	SS	MS	F value
Batches (unadjusted)	9	548.05		
Treatments (adjusted)	4	633.00	158.250	5.54
Residual	6	171.50	28.583	$= \hat{\sigma}^2$
TOTAL	19	1352.55		

The F value of 5.54 is referred to $F_{4,6}$; upper 5% point is 4.53, upper 1% point is 9.15, so this is significant at the 5% level. There is some evidence of treatment differences.

(c) One way to investigate treatment differences is by a least significant difference analysis.

The variance of a difference between two (adjusted) treatment means is $2k\sigma^2/\lambda v$. We estimate σ^2 by 28.583 to get 2×2×28.583/5 = 22.8664. The square root of this is 4.782.

The 5%, 1% and 0.1% points of t_6 are 2.45, 3.71 and 5.96 respectively, so the least significant differences at these levels are 2.45×4.782 = 11.72, 3.71×4.782 = 17.74 and 5.96×4.782 = 28.50.

This suggests that *B*, *C* and *E* might all have the same effect with *D* having a greater effect. The status of *A* remains somewhat unclear. It appears, at the 5% level, to have a greater effect than *C*, but it cannot be distinguished from *B*, *D* or *E*.

Graduate Diploma, Applied Statistics, Paper II, 2004. Question 3

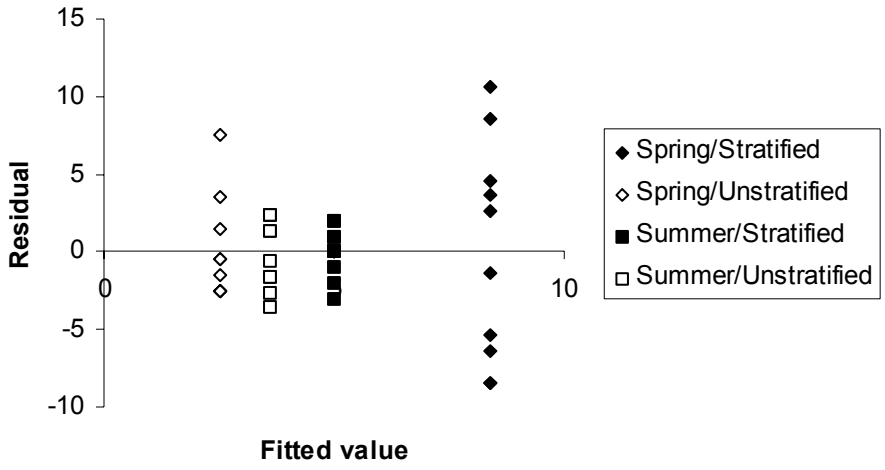
[solution continues on next page]

(i) In a one-way analysis of variance, the residual for any plot is the difference between the observed value and the fitted value, which is simply the mean for that treatment. For example, the mean for Spring/Stratified is 8.4, so that the residual for observation '12' is 12 – 8.4 = 3.6. The sum of the residuals for this treatment, and for each of the other treatments, will of course be 0.

Before carrying out further analysis, note that each data item should actually have an underlying binomial distribution with $n = 20$. There are a number of extreme values, near to 0 or 20. The ranges of the data for the four treatments are Spring/Stratified 0 to 19, Spring/Unstratified 0 to 10, Summer/Stratified 2 to 7, Summer/Unstratified 0 to 6. All this suggests that the required assumptions of underlying Normality and equal variances for the four treatments seem unlikely to be the met. An angular transformation may be necessary to stabilise variance.

The residuals can be plotted against the fitted values. The plot is shown below. The pattern of the scatter should be the same for each treatment. It does not appear to be so. For example, spring storage gives more variable results than summer storage.

[**Note.** There are coincident points for each fitted value. The full list of residuals is given in the question.]



A Normal probability plot would be possible if a computer is available. (It gives some evidence of non-Normality due to the extreme values; these are also noticeable in the plot above, though not in a way that suggests lack of symmetry.)

[Candidates might mention Bartlett's test for variance homogeneity; but it should be emphasised that it is sensitive to non-Normality and so in this case is unlikely to be useful.]

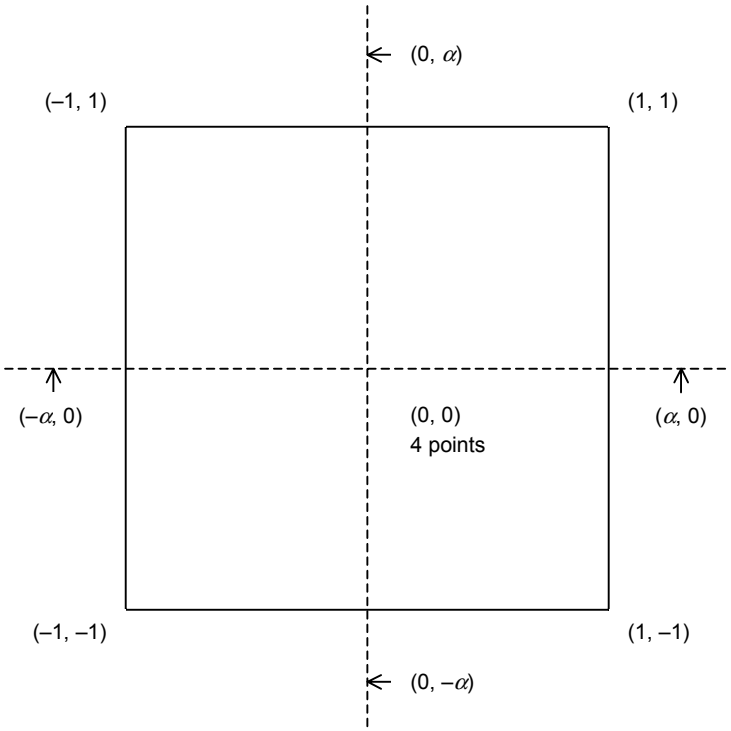
Since the layout is effectively "completely randomised", there are no classifications other than treatments that can be studied.

(ii) The assumptions are that the (true) residuals (experimental errors) are i.i.d. (independent identically distributed) N(0, σ^2) and that there is no systematic variation except treatments.

The required discussion is included in the solution to part (i) above. An angular transformation is suggested, but even then we may not have good results because of the different patterns of scatter within the treatments. The usual F and t tests might well be unreliable. Individual comparisons between treatments could be made, not using the overall estimate of variance from the ANOVA; or a non-parametric comparison could be made.

Graduate Diploma, Applied Statistics, Paper II, 2004. Question 4

Part (i)



I : The factorial points are the major contributors to estimating linear terms and the interaction. This part of the design is variance-optimal for this purpose and is the only source of information on the interaction.

II : The centre points contribute towards estimating quadratic effects. They also provide an internal estimate of residual error ("pure error"). A lack-of-fit test for a model is thus possible.

III : The axial points are the major contributors to estimating quadratic terms.

Solution continued on next page

Part (ii)

In a rotatable design, the variance of the fitted response $\hat{Y}$ is the same at any point which is the same distance from O (the centre of the design). Unless it is known that a particular direction is important for study, this property is good. (Designs for fitting non-polynomial models may require a different approach.)

Rotatability requires that α^4 equals the number of factorial points, i.e. 4 here. So we need $\alpha = \sqrt{2}$.

An orthogonal design is one for which $\mathbf{X}^T\mathbf{X}$ is diagonal apart from non-zero entries in the positions that correspond to the product of the constant and quadratic terms, where $\mathbf{X}$ is the design matrix. This means that the covariances of the estimated coefficients are zero, i.e. they are estimated independently given the assumption of underlying Normality, except for the covariances between the estimated constant and quadratic parameters which cannot be arranged to be zero.

Here, for fitting $Y = a + b_1x_1 + b_2x_2 + b_{12}x_1x_2 + b_{11}x_1^2 + b_{22}x_2^2$, $\mathbf{X}^T\mathbf{X}$ is

$$\begin{bmatrix} 12 & 0 & 0 & 0 & 8 & 8 \\ 0 & 8 & 0 & 0 & 0 & 0 \\ 0 & 0 & 8 & 0 & 0 & 0 \\ 0 & 0 & 0 & 4 & 0 & 0 \\ 8 & 0 & 0 & 0 & 12 & 4 \\ 8 & 0 & 0 & 0 & 4 & 12 \end{bmatrix}$$

As noted above, the first row and column of this matrix relate to the constant term a in the model for Y; non-zero entries in this row and column do not make the design non-orthogonal. But when a design is orthogonal, all off-diagonal entries in other rows and columns of the $\mathbf{X}^T\mathbf{X}$ matrix must be zero. This is not the case, so the design is not orthogonal.

Solution continued on next page

Part (iii)

(a) The total (corrected) sum of squares is 1768.917, with 11 degrees of freedom. The residual sum of squares is 35.344.

Quadratic terms account for 614.260 – 35.344 = 578.916, with 2 degrees of freedom.

The interaction term accounts for 854.510 – 614.260 = 240.250, with 1 degree of freedom.

Finally, the difference 1768.917 – 854.510 = 914.407 is due to the linear terms, with 2 degrees of freedom.

Hence:

SOURCE	DF	SS	MS	*F* value
Linear	2	914.407	457.20	77.6
Interaction	1	240.250	240.25	40.8
Quadratic	2	578.916	289.46	49.1
	5	1733.573		
Residual	6	35.344	5.891	
TOTAL	19	1786.917		

The F values are referred to $F_{2,6}$, $F_{1,6}$ and $F_{2,6}$ respectively. All are very highly significant (upper 0.1% point of $F_{2,6}$ is 27.00 and of $F_{1,6}$ is 35.51.

(b) The "pure error" is the sum of squares between the four centre points, i.e. $76^2 + 79^2 + 83^2 + 81^2 - \frac{319^2}{4} = 26.75$. This has 3 degrees of freedom.

Hence the lack of fit sum of squares is 35.344 – 26.75 = 8.594, also with 3 (= 6 – 3) degrees of freedom.

We compare the lack of fit with the pure error by the usual F test: $\frac{8.594/3}{26.75/3} = 0.321$, which we refer to $F_{3,3}$. This is obviously not significant; there is no suggestion of lack of fit.

Returning to the F tests in part (a), we can confidently say that there is extremely strong evidence of linear, interaction and quadratic effects.

(c) Given adequate computer graphics, the best follow-up is to obtain contour plots of yield as a function of temperature and concentration near the maximum (assuming there is one in the experimental region). The position of the maximum can be estimated, together with rates of change from it in various directions.

Graduate Diploma, Applied Statistics, Paper II, 2004. Question 5

[solution continues on next page]

(i) As is clear from the data, the six strata split into three with fairly low density of caribou and three with much higher density. There are also some variations in the values of s_h between the six strata. Stratified sampling ensures that all these six strata will be represented adequately, and that an estimate of the total number of animals will have a smaller standard deviation than for simple random sampling.

(ii) The estimated total is $\hat{Y}_{st} = N\overline{y}_{st} = \sum_{h=1}^{L} N_h \overline{y}_h$ (where there are L strata), so

$$\hat{Y}_{st} = (400\times 24.1) + (40\times 25.6) + (100\times 267.6) + (40\times 179.0) + (70\times 293.7) + (120\times 33.2) \;=\; 69127 .$$

The estimated variance of $\hat{Y}_{st}$ is given by

$$\sum_{h=1}^{L} N_h (N_h - n_h)\frac{s_h^{\,2}}{n_h} \;=\; 400\times(400-98)\times\frac{74.7^2}{98} \;+\; \ldots \;=\; 84123268.3 \;,$$

so the estimated standard error is 9171.9.

(iii) N_h is the true number in stratum h. S_h is the true standard deviation in stratum h. w_h is n_h/n, the proportion of the whole sample that comes from stratum h. V is the value specified for Var($\hat{Y}_{st}$).

(iv) Optimal allocation minimises the variance Var($\hat{Y}_{st}$) (equivalently, Var($\overline{y}_{st}$)) for fixed total sample size n.

As well as allocating more sampling to strata with larger population sizes, it allocates more to those with larger standard deviations, so the precision is comparable with those having lower variability. In the present survey, there are wide variations among the stratum sizes and standard deviations; proportional allocation with the same sample size as optimal allocation is likely to lead to a considerably larger value of Var($\hat{Y}_{st}$).

(v) We use the formula quoted in part (iii) of the question, taking the estimates s_h from the preliminary aerial survey as though they were the true values S_h.

Optimal allocation with constant cost of sampling any unit has w_i (= n_i/n) given by $w_i = \dfrac{N_i S_i}{\sum_{h=1}^{L} N_h S_h}$. We have $\Sigma N_h S_h$ = 133903, using the preliminary survey values.

Further, we see that $\frac{N_h^2 S_h^2}{w_h} = (N_h S_h)\left(\sum_{h=1}^{L} N_h S_h\right)$, so that $\sum \frac{N_h^2 S_h^2}{w_h} = \left(\sum N_h S_h\right)^2$.

Also we have $\Sigma N_h S_h^2 = 47882186$ (this appears in the denominator of the formula).

Finally, we need V. The criterion of $d = 8000$ with (one-sided) tail probability 0.025 gives $V = (8000/1.96)^2$.

$$\therefore \quad n = \frac{(133903)^2}{\left(\frac{8000}{1.96}\right)^2 + 47882186} = 277.804 \ .$$

So we take $n = 278$. The allocation in each stratum is then given by

$$n_i = 278 w_i = 278 \frac{N_i S_i}{\sum N_h S_h} \ ,$$

which gives $n_1 = 62.03$, $n_2 = 5.29$, $n_3 = 122.39$, $n_4 = 12.54$, $n_5 = 51.08$, $n_6 = 24.66$.

However, the total size of stratum 3, N_3, is only 100; so we must take $n_3 = 100$.

The remaining 178 are then allocated in the same ratios as before, by multiplying each by 178/155.61. This gives $n_1 = 70.96$, $n_2 = 6.05$, $n_4 = 14.34$, $n_5 = 58.43$, $n_6 = 28.21$.

Finally,

$$n_1 = 71, \ n_2 = 6, \ n_3 = 100, \ n_4 = 14, \ n_5 = 58, \ n_6 = 28.$$

With this allocation, the estimated variance of $\hat{Y}_{st}$ is given by

$$\sum_{h=1}^{L} N_h (N_h - n_h) \frac{s_h^2}{n_h} = 400 \times (400 - 71) \times \frac{74.7^2}{71} + \ ... \ = 18610118.04$$

(note there is a ZERO contribution to the sum from stratum 3, where we have a 100% sample), so the estimated standard error is 4313.9.

Graduate Diploma, Applied Statistics, Paper II, 2004. Question 6

Part (a)

(i) We have 50 strata (geographical regions) covering the UK. Each stratum is divided into postcode sectors, some of which are selected (with probability proportional to size) to be sampled. In the chosen sectors, simple random sampling of households is carried out. This procedure is cluster sampling: the postcode sectors form the clusters, as they are either sampled or not used at all.

In Stage 1, the sampling units are the postcode sectors and the sampling plan is probability proportional to size (PPS).

In Stage 2, the sampling units are the households and the sampling is simple random.

PPS sampling is useful when clusters of population units vary in size. With simple random sampling, units in a large cluster have lower probability of being sampled than those in a small cluster. This may lead to estimates of means or totals which are biased and have large variances. PPS is an easier alternative to "weighted" sampling, for the purpose of giving every element in the target population the same probability of being in the sample. (If clusters are not very different in size, the extra effort in sampling by PPS may not be worthwhile.) PPS sampling can reduce the cost of obtaining estimates with specified precision.

(ii) Street directories may be out of date. Some buildings may no longer exist; others may be empty or in non-residential use, or may contain several households. Households moving into the area may not be located if the whole area is imperfectly covered by the directory.

In the UK, the electoral register (list of voters) is likely to be much more up to date (though it will not be perfect), and it does allow households to be identified by name. Similar lists exist in many countries.

Solution continued on next page

Part (b)

Section 1 : Question 2 needs more boxes, for widowed, divorced/separated, living with partner.

Question 3 could be more specific and ask how many adults and how many children are living at home.

Question 4 needs to say annual combined household income, and it should be made clear where (e.g.) £10000 is to be entered by having "£5000 – £9999", "£10000 – £19999", etc.

Section 2 : Question 2 should have an instruction to put a mark in all relevant boxes.

Question 3 should say per week, to avoid relying on memory/ guesswork for longer periods. Also, the present question 3 could possibly be called "main shopping", with another question for "top-up shopping" with categories (say) "Under £10", "£10 – £19.99", "£20 plus".

Question 4 might refer to the previous month only (again to avoid relying on memory/guesswork) and should give some numbers such as "More than 5 times", "3 to 5 times", "Once or twice" and "Never (or hardly ever)".

Question 5 should say per week and perhaps be explicitly restricted to the last week. It should include a box for 0 (zero) and possibly one for "More than 4".

Graduate Diploma, Applied Statistics, Paper II, 2004. Question 7

[solution continues on next two pages]

Part (i)

Possible answers include:

when there are obvious clusters in the target population;

when there is no reliable list of the whole population and it is expensive to construct one;

when there are similar clusters such as villages which will give similar results so that not all need be covered;

when cost is to be kept within limits so that only a limited part of a whole population can be studied.

When variation between clusters is much greater than variation within them, the results from cluster sampling may be less precise than those using simple random sampling. Therefore some basic information about the clusters is needed. (Conversely, if within-cluster variation is large, the method will probably perform as well as, or better than, simple random sampling.)

Part (ii)

(a) The notation is as follows:

N is the number of clusters in the population;

n is the number of clusters selected (by simple random sampling) to be sampled;

m_i is the number of elements in the ith sampled cluster;

y_i is the sample total for the ith sampled cluster;

$\overline{y}$ is defined as $\frac{1}{n}\sum_{i=1}^{n} y_i$, i.e. it is the average sample *total* over the chosen clusters.

(b) $\hat{Y}_{ub} = N\overline{y}$, so $E\left[\hat{Y}_{ub}\right] = NE\left[\overline{y}\right]$. Now, $\overline{y}$ is the sample mean for a simple random sample of cluster totals, so it is an unbiased estimator of the population mean of the cluster totals; so $NE\left[\overline{y}\right]$ is simply equal to the population total, Y, as required.

(c) $\overline{y}_{cl}$ is a ratio estimator, being the ratio, in the chosen sample of clusters, of the sum of the cluster totals to the sum of the cluster sizes. Thus it is biased. However, if n is large, the bias is small. Also, if all the clusters are the same size, the bias is zero.

Part (iii)

We have a simple random sample of 85 clusters from 828.

In the notation of part (ii), $N = 828$, $n = 85$, $m = 215$ (same for each cluster, so not indexed as m_i). Also, $\Sigma y_i = (57\times0) + (22\times1) + (4\times2) + (1\times3) + (1\times4) = 37$.

(a)

Thus $\bar{y} = (1/85)\times 37 = 0.4353$ and so the required estimate of the total number of errors is $\hat{Y}_{ub} = 828\times 0.4353 = 360.42$.

To find the variance of this, we again regard $\bar{y}$ as the sample mean for a simple random sample (of cluster totals); thus its variance is, in the usual notation, $(1-f)S^2/n$ and so the variance of $\hat{Y}_{ub}$ is $828^2(1-f)S^2/n$. We have $1-f = (N-n)/N = 743/828$.

To estimate S^2, we use the usual $s^2 = \frac{1}{n-1}\sum (y_i - \bar{y})^2 = \frac{1}{n-1}\left\{\Sigma y_i^{\,2} - \frac{(\Sigma y_i)^2}{n}\right\}$. We need only to calculate $\Sigma y_i^2 = (57\times0^2) + (22\times1^2) + (4\times2^2) + (1\times3^2) + (1\times4^2) = 63$. Thus

$$s^2 = \frac{1}{84}\left\{63 - \frac{37^2}{85}\right\} = \frac{46.89412}{84}\ .$$

Thus our estimate of the variance of $\hat{Y}_{ub}$ is $828^2 \times \frac{743}{828} \times \frac{46.89412}{84} \div 85 = 4040.539$, and so the standard error is $\sqrt{4040.539} = 63.57$.

(b) We now consider a population of 178020 (= 828 × 215) fields with a simple random sample of size 18275. Altogether there are 37 fields in error in the sample, so the sample may be considered to consist of 37 data items of 1 and 18238 of zero. So, for this sample, we have $\Sigma y_i = 37$ and $\Sigma y_i^2 = 37$. We here use $\bar{y}$ to denote the usual simple random sample mean, with $\hat{Y}_T = 178020\bar{y}$ the estimator of the population total number of erroneous fields. Thus $\mathrm{Var}\left(\hat{Y}_T\right) = 178020^2\ \mathrm{Var}\left(\bar{y}\right) = 178020^2\left(1-f\right)S^2/n$ in the usual notation, where now $f = 18275/178020$ so that $1-f = 159745/178020$.

As before, we estimate S^2 using $s^2 = \frac{1}{n-1}\sum (y_i - \bar{y})^2 = \frac{1}{n-1}\left\{\Sigma y_i^{\,2} - \frac{(\Sigma y_i)^2}{n}\right\}$, where $n = 18275$.

This gives that our estimate of the variance of $\hat{Y}_T$ is

$$178020^2 \times \frac{159745}{178020} \times \frac{\left(37 - \frac{37^2}{18275}\right)}{18274} \div 18275 \;=\; 3144.3191$$

and so the standard error is $\sqrt{3144.3191} = 56.07$.

A simple random sample of this size would be almost impossible to choose, and very time-consuming. The cluster sample provided 85 large blocks of data, and the result is of the same order of precision.

Graduate Diploma, Applied Statistics, Paper II, 2004. Question 8

[solution continues on next page]

In the table below,

$_{10}q_x$ = probability that a person aged x years dies within the next 10 years (values of this are given in the question)

l_x = number of each year's cohort (of 1000) attaining age x

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

$_{10}L_x$ = number living between ages x and $x + 10$ ($= 10 \times \frac{1}{2}(l_x + l_{x+10})$)

T_x = number of persons aged x or greater ($= \sum_{y \geq x} {}_{10}L_y$).

Hence (part (i) of the question) the age distribution ($= \frac{100\left({}_{10}L_x\right)}{68040}$ %) is as follows (note that there are small rounding errors in the calculations: the sum of these percentages is 99.99):

Age	0 –	10 –	20 –	30 –	40 –	50 –	60 –	70 –	80 –	90 –	100 –
%	14.48	14.20	14.04	13.79	13.33	12.19	9.72	5.85	2.09	0.29	0.01

Age (x)	$_{10}q_x$	l_x	$_{10}d_x$	$_{10}L_x$	T_x
0	0.029	1000	29	9855	68040
10	0.009	971	9	9665	58185
20	0.015	962	14	9550	48520
30	0.020	948	19	9385	38970
40	0.047	929	44	9070	29585
50	0.125	885	111	8295	20515
60	0.291	774	225	6615	12220
70	0.551	549	302	3980	5605
80	0.846	247	209	1425	1625
90	0.979	38	37	195	200
100	1.000	1	1	5	5
110		0		0	0

(ii) Expected age at death for a group at present of age x is $x + \frac{T_x}{l_x}$. Hence:

Age 20 : 20 + (48520/962) = 70.44
Age 40 : 40 + (29585/929) = 71.85
Age 60 : 60 + (12220/774) = 75.79
Age 90 : 90 + (200/38) = 95.26
Age 100 : 100 + (5/1) = 105.00

(iii) The life expectancy is the expected age at death if at present of age 0, i.e.

0 + (68040/1000) = 68.04.

In the abridged table used, it is assumed that deaths occur uniformly throughout each 10-year age group. Clearly this is not true, in the older age groups especially (also for the 0 – 10 group, no doubt), and the results from the unabridged table will be much more accurate in these groups. There is also some effect on the overall life expectancy figure.

If there is an annual growth rate of 1% in addition, the age distribution in 10-year intervals is calculated from

$$\frac{(1+0.01)^{-(x+5)} \times {}_{10}L_x}{\sum_{\text{all groups}} (1+0.01)^{-(x_i+5)} \times {}_{10}L_{x_i}} \times 100\% \ .$$

Because of increasing birth rate, there will be an increase in the proportions in lower age groups as compared with the original population.