THE ROYAL STATISTICAL SOCIETY

2006 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.

Note. In accordance with the convention used in the Society's examination papers, the notation log denotes logarithm to base *e*. Logarithms to any other base are explicitly identified, e.g. $\log_{10}$.

Graduate Diploma, Applied Statistics, Paper II, 2006. Question 1

(i) The grand total is 127.03 the "correction factor" is $127.03^2/60 = 268.9437$.

So the total sum of squares $= 301.4107 - \dfrac{127.03^2}{60} = 32.4670$, with 59 df.

$$\text{SS for blocks} = \frac{36.09^2}{20} + \frac{43.27^2}{20} + \frac{47.67^2}{20} - \frac{127.03^2}{60} = 272.3605 - 268.9437$$

$$= 3.4168, \text{ with 2 df.}$$

$$\text{SS for seed rate} = \frac{10.92^2}{12} + \ldots + \frac{31.64^2}{12} - 268.9437 = 25.6476, \text{ with 4 df.}$$

$$\text{SS for row width} = \frac{31.48^2}{15} + \ldots + \frac{29.01^2}{15} - 268.9437 = 0.9166, \text{ with 3 df.}$$

$$\text{Interaction SS} = \frac{1.87^2}{3} + \frac{5.40^2}{3} + \ldots + \frac{7.05^2}{3} - 268.9437 - 25.6476 - 0.9166$$

$$= 0.9976, \text{ with } 4\times 3 = 12 \text{ df.}$$

The residual SS and df follow by subtraction.

Hence:

SOURCE	DF	SS	MS	*F* value
Blocks	2	3.4168	1.7084	
Seed rate	4	25.6476	6.4119	163.6
Row width	3	0.9166	0.3055	7.8
Interaction	12	0.9976	0.0831	2.1
Residual	38	1.4884	0.0392	$= \hat{\sigma}^2$
TOTAL	59	32.4670		

The *F* value of 163.6 is referred to $F_{4,38}$; this is well beyond the upper 0.1% point (about 5.8), so there is extremely strong evidence of an effect of seed rate.

The *F* value of 7.8 is referred to $F_{3,38}$; this is beyond the upper 0.1% point (about 6.7), so there is very strong evidence of an effect of row width.

The *F* value of 2.1 is referred to $F_{12,38}$; this is (just) significant at the 5% level, so there is some evidence of an interaction.

Overall, though the effects of seed rate and row width appear very highly significant, the results should be explained in terms of the interaction.

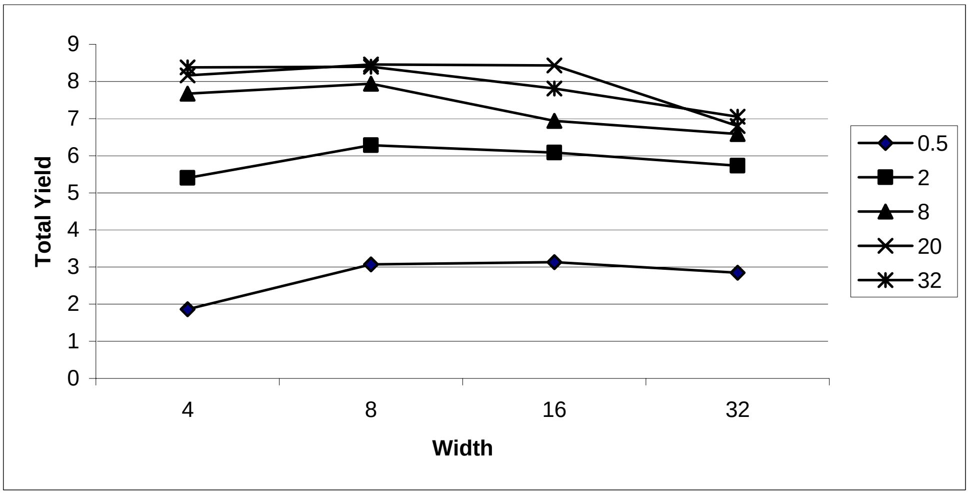
Solution continued on next page

(ii) The partitioning for row width is carried out as follows.

Row width	4	8	16	32	r = 15 (for each total)			
Total	31.48	34.15	32.39	29.01	Value	Divisor	SS	F value
Linear	–3	–1	1	3	–9.17	15 × 20	0.2803	7.2
Quadratic	1	–1	–1	1	–6.05	15 × 4	0.6100	15.6
Cubic	–1	3	–3	1	2.81	15 × 20	0.0263	0.7
							0.9166	

Each partitioned term has 1 df and F tests (comparing with the residual mean square as before) have 1 and 38 df, so there is evidence for a linear component of the effect and very strong evidence for a quadratic component.

(iii)



(iv) Looked at overall, the yield is greatest for row width 8; the yield from row widths 4 and 16 are close together at a somewhat lesser value, and the yield from row width 32 is the least. The rise and fall with respect to row width leads to the quadratic component of this main effect.

In terms of overall effect of seed rate, the yield from rate 0.5 is very much less than that from rate 2 which is itself substantially less than that from the others.

However, the interaction has to be taken into account. The diagram shows that there are somewhat different patterns of yields over the row widths at the different seed rates.

A seed rate of 8lb/acre seems adequate and is likely to be economical, although any future work needs to clarify the row width appropriate for this seed rate to give maximum yield.

Graduate Diploma, Applied Statistics, Paper II, 2006. Question 2

A contrast among treatment means is $\sum c_i \bar{y}_i$, where the c_i are a set of constants whose sum is zero. Usually the c_i are integers, for simplicity in calculations.

If the variance of individual observations is σ^2, then that of the mean $\bar{y}_i$ is σ^2/r. The variance of the contrast $\sum c_i \bar{y}_i$, assuming independence of all observations (proper randomisation) is $\sum c_i^2 \operatorname{Var}(\bar{y}_i)$, which is $\sum c_i^2 \sigma^2 / r$. The standard deviation is the square root of this, and thus the standard error is $\sqrt{\sum c_i^2 s^2 / r}$ where s^2 denotes the residual mean square which estimates σ^2.

Two orthogonal contrasts among the same set of means have coefficients c_i and d_i such that $\Sigma c_i = 0 = \Sigma d_i$ and $\Sigma c_i d_i = 0$. The importance of orthogonal contrasts is that the are uncorrelated. Thus they are independent for the case of Normally distributed errors, and represent comparisons among the means that can be independently estimated and tested for.

(i) and (ii)

The required contrasts are

	A	B	C	D	E	F	G	H
Mean	88	198	66	235	265	233	40	41
(a)	–1	1	–1	1	–1	1	–1	1
(b)	–1	–1	1	1	0	0	0	0
(c)	1	1	1	1	1	1	–3	–3
(d)	1	1	1	1	–2	–2	0	0
(e): (b) with (a)	1	–1	–1	1	0	0	0	0
(e): (c) with (a)	–1	1	–1	1	–1	1	3	–3
(e): (d) with (a)	–1	1	–1	1	2	–2	0	0

We have $s^2/r = 3265.8/5 = 653.16$. The SE for each contrast is thus $\sqrt{653.16 \Sigma c_i^2}$. The number of df for the residual is 39 – 7 = 32 (there are 40 observations and 8 treatments). So the statistical significance of each contrast may be tested by referring (value)/SE to the *t* distribution with 32 df, on the assumption of Normality and common variance for the experimental errors. The two-sided critical points of t_{32} are 2.037 for 5%, 2.738 for 1% and 3.622 for 0.1%.

Solution continued on next page

We have, from the above table,

	Value	$\sum c_i^2$	SE	Value/SE
(a)	248	8	72.29	3.431
(b)	15	4	51.11	0.293
(c)	842	24	125.20	6.725
(d)	–409	12	88.53	–4.620
(b) with (a)	59	4	51.11	1.154
(c) with (a)	244	24	125.20	1.949
(d) with (a)	343	12	88.53	3.874

There is strong evidence of an overall difference between the effects of the levels of the fertiliser [contrast (a)]; it appears that high fertiliser level is better than low level.

There is no evidence of difference between the effects of the cultures [contrast (b)].

There is very strong evidence for an effect of inoculation [contrast (c)]; it appears that inoculation gives higher yield.

Likewise there is very strong evidence for an effect of the two strains of Rhizobium [contrast (d)]; it appears that CC511 performs better than R3644.

However, interpretations must take account of any interactions. There is no evidence of interaction between the two cultures of R3644 and fertiliser level [(b) with (a)]. There is also not (quite) sufficient evidence to suggest an interaction between the effect of inoculation and fertiliser level [(c) with (a)]. There is very strong evidence of an interaction between the two strains of Rhizobium and fertiliser level [(d) with (a)]: it appears that R3644 performs better at the high fertiliser level than at the low level, but CC512 somewhat better at the low fertiliser level than the high.

Graduate Diploma, Applied Statistics, Paper II, 2006. Question 3

An "incomplete block" scheme of some sort is needed when block size (the number of "plots" in a "block") is less than the number of treatments to be compared. It is obviously not possible in such circumstances for every treatment to appear in every block. A balanced incomplete block is a design where a degree of symmetry is nevertheless preserved, and is useful when it is desired that comparisons between each pair of treatments are to be made with the same precision. It requires all blocks to be the same size. The design is such that every pair of treatments occurs together in the blocks the same number of times. This is illustrated in the example in this question, where there are 7 treatments in blocks of size 3, and each pair of treatments occurs together in the blocks just once.

If the blocks cannot all be the same size, or if some comparisons are more important than others, less balanced designs may be necessary or preferred.

(i) This is a balanced incomplete block design (see discussion above) with structural parameters as follows.

N is the number of observations: $N = 21$.

b is the number of blocks: $b = 7$.

k is the size of each block: $k = 3$.

v is the number of treatments: $v = 7$.

r is the number of replicates of each treatment: $r = 3$.

λ is the number of times each pair of treatments occurs together in a block: $\lambda = 1$. [Note: $\lambda = r(k-1)/(v-1) = 3\times2/6 = 1$; this has to be an integer for the incomplete block design to be *balanced*.]

(ii) The total SS is $8341 - \dfrac{413^2}{21} = 218.667$.

The SS for blocks is

$$\frac{1}{3}\left(\frac{59^2}{3} + \frac{66^2}{3} + \dots + \frac{63^2}{3}\right) - \frac{413^2}{21} = 90.000.$$

The SS for treatments adjusted for blocks is [formula quoted in question]

$$\frac{7\times1}{3}\sum_i \hat{\tau}_i^{\,2} = \frac{7}{3}\times 41.3908 = 96.579.$$

Solution continued on next page

Hence:

SOURCE	DF	SS	MS	F value
Blocks	6	90.000	–	
Treatments adjusted for blocks	6	96.579	16.097	4.01
Residual	8	32.088	4.011	$= \hat{\sigma}^2$
TOTAL	20	218.667		

The F value of 4.01 is referred to $F_{6,8}$; this is significant at the 5% level (critical point is 3.58), so there is some evidence that there are differences between the treatments, having adjusted for the blocks. The differences are explored in part (iii).

(iii) The variance of the difference between any pair of treatment means is estimated by [formula quoted in question]

$$\frac{2k\hat{\sigma}^2}{v\lambda} = \frac{2\times 3\times 4.011}{7\times 1} = 3.438 .$$

Least significant differences are therefore given by $t\times\sqrt{3.438}$ where t denotes the appropriate critical point from the t_8 distribution: 2.306 for 5%, 3.355 for 1%, 5.041 for 0.1%. So the respective LSDs are 4.28, 6.22, 9.35.

The table below shows the estimated treatment effects (adjusted for blocks) in ascending order of size.

E	D and F	A	C	B	G
–2.8571	–1.8571	–0.2857	–0.1429	2.5714	4.4290

Interpretation is difficult. Recall that the overall test in the analysis of variance in part (ii) was only significant at the 5% level. In LSD terms, we see that *all* the treatments could be considered the same if judged at the 0.1% level. At the 1% level, G is "detached" from (and better than) E and (D and F), but no better than A, C or B which are themselves no better than E or (D and F). At the 5% level, G is "detached" from (better than) all but B, while B is also "detached" from E and (D and F).

Graduate Diploma, Applied Statistics, Paper II, 2006. Question 4

(a) (i) In response surface analysis, a quadratic surface is fitted when it appears that the experimental region is near to the maximum or minimum of the response variable. This requires the factors to be at 3 (or more) levels, as is the case here. The location of the turning point (in this case a minimum) of the surface can be estimated; in the present context, this gives an estimate of the operating conditions that minimise failure stress. Replication gives a proper base for calculating an estimate of natural (random) variation in order to assess goodness of fit of the model.

(ii), (iii) and (iv)

The total SS is $1605355 - \dfrac{5321^2}{18} = 32408.278$, with 17 df.

The A$_{\text{linear}}$ SS is

$$\left\{-1(682+554+657)+0(530+449+491)+1(654+618+686)\right\}^2/12$$

= 352.083, with 1df.

The S$_{\text{quadratic}}$ SS is

$$\left\{1(682+530+654)-2(554+449+618)+1(657+491+686)\right\}^2/36$$

= 5826.778, with 1 df.

The SS for A$_{\text{linear}}$ × S$_{\text{linear}}$ can now be found by subtraction; this also has 1 df.

> [Note. This could also be calculated directly, using (in an obvious notation) $\mathrm{A_L S_L} = a_0 s_0 - a_2 s_0 - a_0 s_2 + a_2 s_2$ (as can be found by multiplying the coefficients of the treatment combinations in the A$_\text{L}$ row of the contrast table by the corresponding coefficients in S$_\text{L}$). Hence the SS for A$_\text{L}$S$_\text{L}$ is $(682 - 654 - 657 + 686)^2/8 = 406.125$.]

The residual SS can also be found by subtraction; this has 9 df.

Thus the completed analysis of variance table is as follows.

Solution continued on next page

Source of variation	DF	Sum of squares	MS	*F* ratio
A_{linear}	1	352.083	352.083	1.98
$A_{\text{quadratic}}$	1	23053.361	23053.361	129.55
S_{linear}	1	85.333	85.333	0.48
$S_{\text{quadratic}}$	1	5826.778	5826.778	32.75
$A_{\text{linear}} \times S_{\text{linear}}$	1	406.125	406.125	2.28
Other AS components	3	1083.098	361.033	2.03
Treatments	8	30806.778		
Residual	9	1601.500	177.944	
TOTAL	17	32408.278		

The *F* ratios shown above are for comparisons with the residual MS, which is "pure error".

The "Other AS components" entry measures the lack of fit of a second-order linear model (because all components in it are of order at least three). Its *F* ratio of 2.03 is referred to $F_{3,9}$ and is not significant – we have no evidence for lack of fit of a second-order linear model. [It is sometimes argued that, in these circumstances, the SS and df for this should now be combined with the "pure error" residual to give a new residual SS with 12 df which the other effects should be compared with.]

The second-order linear model used is

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_{12} x_1 x_2 + \beta_{11} x_1^2 + \beta_{22} x_2^2 + \varepsilon$$

where x_1 and x_2 represent the respective levels of *A* and *S* and the usual assumptions apply to the error terms ε.

The 5 single df components in the above table are used for β_1, …, β_{22}. On referring to $F_{1,9}$ we find that there is extremely strong evidence for the two quadratic effects but no evidence for linear effects or the linear × linear interaction (the upper 0.1% critical point of $F_{1,9}$ is 22.86; the upper 5% point is 5.12).

So the dominant terms in the model are $\beta_{11} x_1^2$ and $\beta_{22} x_2^2$.

(b) In a mixture design, the factors (x_i) are components (proportions, or sometimes percentages) of a mixture, for example in manufacture of a detergent or concrete. Thus for concrete there is a mix of cement, sand and water and the strength will depend on the proportions of each. This means that the response surface model for a mixture experiment has a constraint $\Sigma x_i = 1$ (and also all $x_i \geq 0$), whereas in an ordinary factorial the choices of values taken by the x_i in the model are unconstrained (within the experimental region being explored).

(i) (a) The sampling units are households. So, in a standard notation, we have n = 500 and N = 10 000. Also, since x represents numbers of adults and y represents numbers of cars, we have

$$\sum x_i = (1\times 40)+(2\times 280)+(3\times 140)+(4\times 38)+(5\times 2) = 1182$$

and

$$\sum y_i = (0\times 176)+(1\times 243)+(2\times 61)+(3\times 20) = 425 .$$

So $\bar{x} = 2.364$ and $\bar{y} = 0.850$.

The estimated ratio of cars to adults is $\hat{r} = \dfrac{\sum y_i}{\sum x_i} = \dfrac{425}{1182} = 0.3596$.

From the census,

$$\bar{X} = (1\times 0.1)+(2\times 0.5)+(3\times 0.3)+(4\times 0.09)+(5\times 0.01) = 2.41 ;$$

that is, the (population) mean number of cars per household is 2.41.

Thus the ratio estimate of $\bar{Y}$, the mean number of cars per household, is given by

$$\hat{y}_R = \hat{r}\bar{X} = (425/1182)\times 2.41 = 0.8665 .$$

For the regression estimate of $\bar{Y}$ we need $\hat{b} = \dfrac{S_{xy}}{S_{xx}}$.

Now,

$$\sum x_i^2 = (1^2\times 40)+...+(5^2\times 2) = 3078 ,$$

and similarly $\sum_i \sum_j x_i y_j = 1094$. So we have

$$S_{xy} = 1094 - \frac{1182\times 425}{500} = 89.30 \text{ and } S_{xx} = 3078 - \frac{1182^2}{500} = 283.752 .$$

Hence $\hat{b} = 89.30/283.752 = 0.3147$, and so the regression estimate is

$$\hat{y}_{LR} = \bar{y} + \hat{b}\left(\bar{X} - \bar{x}\right) = 0.850 + 0.3147\left(2.41 - 2.364\right) = 0.8645 .$$

Solution continued on next page

(b) We use $\hat{V}$ to denote an estimated variance. For the ratio estimator $\hat{y}_R$, we have

$$\hat{V}(\hat{y}_R) = \frac{1-f}{n}\left(S_{yy} - 2\hat{r}S_{xy} + \hat{r}^2 S_{xx}\right)/(n-1)$$

in standard notation. To calculate S_{yy}, we have

$$\sum y_i^2 \;=\; (0^2 \times 176) + (1^2 \times 243) + (2^2 \times 61) + (3^2 \times 20) \;=\; 667$$

and so $S_{yy} = 667 - \dfrac{425^2}{500} = 305.75$. Hence we have

$$\hat{V}(\hat{y}_R) = \left(1 - \frac{500}{10000}\right)\left(\frac{1}{500}\right)\left(305.75 - (2 \times 0.3596 \times 89.3) + (0.3596^2 \times 283.752)\right)/499$$

$$= \frac{0.95}{500}\left(305.75 - 64.2246 + 36.6926\right)/499 \;=\; \frac{0.95 \times 278.2180}{500 \times 499} \;=\; 0.001059.$$

From the formula quoted in the question for the case of the regression estimator, we have

$$\hat{V}(\hat{y}_{LR}) = 0.95 \times \left(1 - \frac{S_{xy}{}^2}{S_{xx}S_{yy}}\right)\frac{S_{yy}}{n-1}.\frac{1}{n}$$

$$= \frac{0.95}{499 \times 500}\left(305.75 - \frac{89.3^2}{283.752}\right) \;=\; \frac{0.95 \times 277.6463}{499 \times 500} \;=\; 0.001057.$$

The efficiency of the ratio estimator relative to the regression estimator is most easily expressed, as a percentage, as $\dfrac{277.6463}{278.2180} \times 100 = 99.8\%$. The regression estimator is (in this case) only very slightly better than the ratio estimator. Bias in both is negligible in a large sample; the regression estimator does not assume the relation between *x* and *y* is a line through the origin.

Solution continued on next page

(c) The regression estimator is often preferred because it makes fewer assumptions (see above).

The regression estimate of $\bar{Y}$ is $\hat{y}_{LR} = 0.8645$; its variance is estimated by 0.001057, so its standard error is 0.03251. An approximate 95% confidence interval for the mean number of cars per household, based on the regression estimator, therefore has end-points 0.8645 ± (1.96 × 0.03251), so the interval is (0.8008, 0.9282). The corresponding confidence interval for the total number of cars in the 10 000 households in the town is therefore (8008, 9282).

Note that if the ratio estimator $\hat{y}_R = 0.8665$ is used, this has estimated variance 0.001059 and thus standard error 0.03255, so the approximate 95% confidence interval for the mean number of cars per household has end-points 0.8665 ± (1.96 × 0.03255) and therefore is (0.8027, 0.9303). The corresponding confidence interval for the total number of cars in the town is (8027, 9303).

(ii) Use of a telephone directory allows some attention to coverage of the various parts of the town to be given; automatic random digit dialling methods might not give proper (or any) coverage of, for example, inner and outer parts of the town, or larger and smaller houses, where car ownership and household composition could be different. Another advantage of using a telephone directory is that it allows only non-business numbers to be contacted (at least in the UK version of the directory). On the other hand, ex-directory numbers would be picked up by the random method. For either method there will be non-response through people not being at home or refusing to give information.

Graduate Diploma, Applied Statistics, Paper II, 2006. Question 6

(i) N_h is the population size in stratum h; n_h is the sample size in stratum h.

$\bar{y}_h$ is the stratum sample mean and s_h the stratum sample standard deviation in stratum h.

Simple random sampling is carried out in each stratum, so $E[\bar{y}_h] = \bar{Y}_h$, the true stratum mean.

$\therefore E[\bar{y}_{st}] = \frac{1}{N}\sum_h N_h E[\bar{y}_h] = \frac{1}{N}\sum_h N_h \bar{Y}_h$, which is the sum of the stratum totals divided by the total population size N, i.e. the true mean. Thus it is an unbiased estimator.

Assuming that the samples in the different strata are independent, we have

$$\text{Var}(\bar{y}_{st}) = \frac{1}{N^2}\sum_h N_h^{\ 2}\,\text{Var}(\bar{y}_h)$$

where, by standard simple random sampling results,

$$\text{Var}(\bar{y}_h) = \left(\frac{N_h - n_h}{N_h}\right)\frac{S_h^{\ 2}}{n_h}$$

where S_h^2 is the true stratum variance. This is estimated by s_h^2 and thus the variance of $\bar{y}_{st}$ is estimated by (writing $f_h = n_h/N_h$)

$$\hat{V}(\bar{y}_{st}) = \frac{1}{N^2}\sum_h N_h^2 s_h^2 (1 - f_h)/n_h\ .$$

(ii) $$\bar{y}_{st} = \frac{1}{6231}\{(92\times 166.6) + (1612\times 7.7) + (4527\times 0.3)\} = 4.67\,.$$

$$\hat{V}(\bar{y}_{st}) \;=\; \frac{1}{(6231)^2}\left\{\frac{92^2\times 207.7^2}{11}\times\left(1-\frac{11}{92}\right) \;+\; ..\right\} \;=\; 1.015093$$

So the standard error here is √1.015093 = 1.00752. An approximate 95% confidence interval for the mean hazardous waste per company therefore has end-points 4.67 ± (1.96 × 1.00752), so the interval is (2.70, 6.64).

Solution continued on next page

(iii) In proportional allocation, stratum sample sizes are proportional to the population sizes. Thus we have $\frac{n_h}{N_h} = \frac{n}{N}$, where *n* and *N* are the total sample and population sizes respectively (364 and 6231).

$$\therefore n_1 = \frac{92 \times 364}{6231} = 5.37, \quad n_2 = \frac{1612 \times 364}{6231} = 94.17, \quad n_3 = \frac{4527 \times 364}{6231} = 264.46.$$

Rounding these so that the total is still 364, we take (6, 94, 264) or (5, 94, 265).

Because the stratum means are very different, this should give much more precise results than a simple random sample of 364 from the whole population.

The allocation actually used (11, 61, 292) gives more observations to stratum 1 than proportional allocation, considerably fewer to stratum 2 and considerably more to stratum 3. Stratum 1 is clearly very variable, so allocating more observations to it should improve overall precision. On the other hand, stratum 3 has very small variability, so there may be little point in allocating more observations there. A calculation similar to that in part (ii) could be carried out to determine the standard error of the stratified sampling mean for this allocation, but could not of course be done until after the results had been obtained because the new sample standard deviations in each stratum are needed. To obtain an indication of precision before undertaking the survey, it might be a reasonable approximation to assume the sample standard deviations are unchanged.

(a) Convenience sampling would consist of taking crates on or very near the outside of the truckload (in practice perhaps simply from the top of the load), and picking oranges on or very close to the top layer of each crate. There may well be trends in quality from top to bottom, and sides to middle, of the truck (and of course the buyer simply does not know at the outset whether there are any such trends). So there is considerable danger of a biased estimate of whatever is being measured.

Cluster sampling requires identification of "clusters" that are expected to behave reasonably similarly to the entire population. This should avoid the problems outlined above.

One method could be to use the crates as clusters. A one-stage scheme would then consist of taking a simple random sample of the crates and then inspecting all the oranges in each chosen crate. This would be feasible if the crates are all going to be unloaded anyway. It could be extended to a two-stage scheme by defining a further level of clusters within the crates – maybe layers of oranges.

An alternative two-stage scheme that could be envisaged consists of using (say) the layers of crates (or perhaps vertical piles of crates) on the lorry as the first stage clusters, selecting a random sample of these, and then selecting crates from within each chosen cluster.

If the inspection has to be done before the whole load is removed from the truck, any form of random sampling of crates is probably not possible.

(b) (i) If n crates (clusters) are chosen, and random samples of m units are taken from each, the estimator of the mean quantity of juice per orange is

$$\bar{y}_{CL} = \frac{1}{n}\sum_{i=1}^{n}\bar{y}_i = \frac{1}{10}\left(88.4+87.8+\ldots+95.8\right) = 93.54 .$$

The clusters are of equal size, all the samples are the same size, and simple random sampling is used at each stage. Thus every individual unit (orange) has the same probability of selection. We can write $\bar{y}_{CL}$ as $\frac{1}{n}\frac{1}{m}\Sigma\Sigma y_{ij}$ over all the selected oranges. This shows that it is an unbiased estimator of the population mean.

Solution continued on next page

(ii) *n* and *m* are as in (b)(i).

f_1 and f_2 are the cluster and within-cluster sampling fractions (we have in this example f_1 = 10/140 and f_2 = 5/120).

s_b^2 is $\frac{1}{(n-1)}\sum_{i=1}^{n}\left(\bar{y}_i - \bar{y}_{CL}\right)^2$, the between-cluster variance.

$s_w^2 = \frac{1}{n(m-1)}\sum_{i=1}^{n}\sum_{j=1}^{m}\left(y_{ij} - \bar{y}_i\right)^2$, the within-cluster variance.

(iii) $$s_b{}^2 = \frac{1}{9}\left\{(88.4-93.54)^2 + (87.8-93.54)^2 + \ldots + (95.8-93.54)^2\right\}$$

$$= \frac{1}{9}\left\{88.4^2 + 87.8^2 + \ldots + 95.8^2 - \frac{935.4^2}{10}\right\} = 24.0893\,.$$

The above expression for $s_w{}^2$ can be simplified to $s_w^2 = \frac{1}{n}\sum_{i=1}^{n} s_i{}^2$ (note that complications of unequal values of *m* for the clusters do not occur in this example; note also that this expression clearly shows the form of $s_w{}^2$ as the average within-cluster variance), so we have

$$s_w^2 = \frac{1}{10}(97.3 + 372.2 + \ldots + 97.2) = 150.71\,.$$

$$\therefore \hat{V}\left(\bar{y}_{CL}\right) = \left(\frac{130}{140}\times\frac{1}{10}\times 24.0893\right) + \left(\frac{10}{140}\times\frac{115}{120}\times\frac{1}{50}\times 150.71\right)$$

$$= 2.23686 + 0.20633 = 2.4432\,.$$

So the standard error is √2.4432 = 1.563. Thus an approximate 95% confidence interval for the true mean quantity of juice per orange has end-points 93.54 ± (1.96 × 1.563) and therefore is (90.48, 96.60).

(iv) It appears that the within-cluster variance is considerably larger than the between-cluster variance. So a sensible strategy might be to take fewer clusters (crates) but more observations (oranges) within each chosen cluster (i.e. smaller *n* and larger *m* – say 5 and 10 instead of 10 and 5 respectively). Five crates rather than 10 would also be somewhat more convenient to sample.

Graduate Diploma, Applied Statistics, Paper II, 2006. Question 8

The crude leaving rate = $1000 \times \dfrac{\text{number of leavers}}{\text{total number of employees}}$

$$= \begin{cases} \dfrac{161000}{1285} = 125.29 & \text{for company } A. \\ \dfrac{29000}{252} = 115.08 & \text{for company } B. \end{cases}$$

To calculate the age-adjusted rates, we need the age-specific rates in each category for each company. An age-specific rate is given by

$$1000 \times \frac{\text{number leaving in age category}}{\text{number in age category}}$$

and so the age-specific leaving rates are

Age	*A*	*B*
16 – 24	200.00	111.11
25 – 34	200.00	171.43
35 – 44	68.38	146.67
45 – 54	16.04	100.00
55 +	12.20	55.56

An age-adjusted rate is given by

$$\frac{\sum\left(\text{number of employees in "Standard"} \times \text{ age-specific rate}\right)}{\text{total number of employees in "Standard"}}.$$

Take *A* as "Standard". Then its age-adjusted rate is, immediately, 125.29 as above.

The rate for *B*, using *A* as Standard is

{(115 × 111.11) + (585 × 171.43) + .. + (164 × 55.56)}/1285 = 136.34.

This rate for *B* is greater than that for *A*, whereas the crude rate for *A* was greater than that for *B*. This is because although *A* loses more of its younger employees than *B*, the loss of older ones from *B* is relatively considerably more than for *A*. This loss of older people makes the age-adjusted rate for *B* greater than that for *A*.

Solution continued on next page

Using duration of service we similarly obtain

Years	*A*	*B*
0 – 4	210.31	169.49
5 – 9	160.71	127.66
10 – 14	58.82	68.97
15+	13.62	17.24

The rate for *A* is 125.29 as before. The rate for *B* using *A* as Standard is

$$\{(485 \times 169.49) + (280 \times 127.66) + (153 \times 68.97) + (367 \times 17.24)\}/1285 = 104.92.$$

Most of *A*'s employees leave before they complete 10 years of service, and so the service-adjusted rate for *B* is lower than that for *A*.

A two-way table of age and length of service would allow rates to be adjusted for both together, which would be informative. The type of work employees were doing, and the type to which they went, would be useful where this could be discovered. Salary and status, if known, could also be compared. Any other more personal, individual reasons for leaving would also give information, for example location, housing, etc.