THE ROYAL STATISTICAL SOCIETY

2007 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, 2007. Question 1

Part (i)

The slope down towards the river is very likely to be a source of systematic variation in the natural fertility of the reserve, and the distance from the motorway is very likely to be a source of systematic variation due to pollution or other climatic reasons. Both these factors are candidates to be used for blocking, and the Latin square design can deal with both together. The danger of using a completely randomised design is that the layout might coincide with one of the systematic factors and we would not be able to disentangle this in the analysis.

To choose a 4×4 Latin square, start with one of the four standard squares (i.e. with the letters in alphabetical order in the first row and in the first column) chosen at random. Then randomly permute the order of the rows, the order of the columns and the allocation of the treatments to the letters.

Part (ii)

(a) The grand total is 793.98; the "correction factor" is $793.98^2/16 = 39400.265$.

So the total sum of squares $= 41802.6036 - \dfrac{793.98^2}{16} = 2402.3386$, with 15 df.

SS for rows (motorway) $= \dfrac{200.76^2}{4} + \dfrac{199.27^2}{4} + \dfrac{196.05^2}{4} + \dfrac{197.90^2}{4} - \dfrac{793.98^2}{16}$

$= 3.0157$, with 3 df.

SS for columns (river) $= \dfrac{211.46^2}{4} + \ldots + \dfrac{183.59^2}{4} - 39400.265 = 108.9608$,

with 3 df.

SS for treatments $= \dfrac{250.10^2}{4} + \ldots + \dfrac{120.93^2}{4} - 39400.265 = 2275.7726$, with 3 df.

The residual SS and df follow by subtraction.

Hence:

SOURCE	DF	SS	MS	*F* value
Rows (motorway)	3	3.0157	1.0052	0.41
Columns (river)	3	108.9608	36.3203	14.94
Treatments	3	2275.7726	758.5909	311.97
Residual	6	14.5895	2.4316	$= \hat{\sigma}^2$
TOTAL	15	2402.3386		

Solution continued on next page

The *F* values are each referred to $F_{3,6}$; the upper 5% point is 4.76, the upper 1% point is 9.78 and the upper 0.1% point is 23.70. So there is no evidence of an effect due to the motorway, strong evidence of an effect due to distance from the river and extremely strong evidence of differences between treatments.

(b) The required contrasts are as follows (treatment totals are also shown, for use in the next part).

	A	B	C	D
Total	250.10	218.06	204.89	120.93
Hay cut / not cut	1	1	1	–3
Removal from site	2	–1	–1	0
Where hay was left	0	1	–1	0

(c) Using these contrasts and the treatment totals, we get the following table, in which SS = $(\text{value})^2/\text{divisor}$, each with 1 df, and the *F* value is then obtained by dividing by the residual mean square from the analysis of variance above.

	Value	Divisor	SS	*F* value
Hay cut / not cut	310.26	48	2005.44	824.75
Removal from site	77.25	24	248.65	102.26
Where hay was left	13.17	8	21.68	8.92

(It may be checked that these three SSs add to the overall treatments SS in the basic analysis of variance.)

The *F* values are each referred to $F_{1,6}$; the upper 5% point is 5.99, the upper 1% point is 13.74 and the upper 0.1% point is 35.51. All the contrasts appear to be important, the first and second especially so.

Recalling that high values show more effective treatments, we conclude that there is extremely strong evidence that it is better to cut the hay than not, and that it is better to remove the hay from the site. There is also evidence that it is better to flail-cut the hay and leave it in windrows than to scythe it and leave it *in situ*.

(d) It is assumed that the model accounts for all sources of systematic variation and that the random variation is given by independent Normally distributed residuals with zero mean and constant variance over the entire meadow.

The Normal probability plot suggests that the Normality assumption is satisfactory. However, the second plot suggests an increase in variance as the size of the fitted values increases. A transformation, such as logarithmic or square root, should be examined as an alternative to analysing the data in the original units.

Graduate Diploma, Applied Statistics, Paper II, 2007. Question 2

Part (a)

Blocking is a procedure under which experimental units are grouped into "blocks" that are expected to be as alike as possible within themselves but may be consistently different from each other. The blocks should then remove a possible element of systematic variation so that the residual mean square in the usual analysis of variance truly estimates just experimental error and is not inflated by such a source of consistent variation. Comparisons between treatment means are then more precise.

For example, in an industrial experiment that takes some time to perform, the blocks might be days, or shifts, or parts of days, chosen so that every treatment can be examined (at least) once in each block using the same machinery.

(i) A randomised (complete) block design is appropriate. The days are the blocks and the concentrations are the treatments; each concentration is run once every day. Any consistent day-to-day differences are removed by the blocking. The order in which the concentrations are run must be random, with a different randomisation for each day; this is in case there is some systematic variation or trend within each day.

(ii) As only three runs are now possible each day, each block cannot contain all four treatments (i.e. the hardwood concentrations). A balanced incomplete block design retains some symmetry in the layout, such that analysis and interpretation of results remains relatively straightforward; any pair of treatments is compared with the same precision. This is achieved by having every pair occurring together the same number (conventionally denoted by λ) of times in the overall design.

In general, a block contains k units and there are b blocks altogether, so $bk = N$ units are needed in all. There are v treatments, each of which is replicated the same number (r) of times; so $rv = N$. Balanced incomplete block designs only exist when these conditions are satisfied and when λ is an integer; it can be shown that $\lambda = r(k-1)/(v-1)$.

In the present case, we have $v = 4$, $b = 4$ and $k = 3$. Hence $r = 3$, and so $\lambda = 2$ and a design can be found. An example of such a design is

Block I	Block II	Block III	Block IV
ABC	ABD	ACD	BCD

The analysis of variance will have the following numbers of degrees of freedom: total 11, treatments (concentrations) 3, blocks (days) 3, residual 5. The number of df for the residual is not very large; experimental error may not be estimated very accurately.

Solution continued on next page

Part (b)

(i) If the variance of colony diameter increases with mean diameter in such a way that the coefficient of variation of diameter is (roughly) constant, a logarithmic transformation will stabilise the variance and thus might be appropriate for analysis.

(ii) The residual has (4 – 1)(6 – 1) = 15 df, so the residual mean square is 0.496/15 = 0.0331. Thus the estimated variance of the difference between any two of the treatments is (2 × 0.0331)/6 = 0.0110. The double-tailed 5% point of t_{15} is 2.131, so any difference greater than 2.131√(0.0110) = 0.224 is significant at the 5% level.

Arranged in ascending order, the treatment means are

C 2.45 *B* 2.72 *D* 2.88 *A* 3.34.

This suggests that treatment (growth supplement) *C* gives a lower result than either of *B* and *D*, and treatment *A* a higher result than either of those two, but *B* and *D* seem not to be different.

(iii) The sample mean difference for *A* and *B* is 0.62. So a 95% confidence interval for the true mean difference for these treatments is (using the result in part (ii)) 0.62 ± 0.224, i.e. (0.396, 0.844).

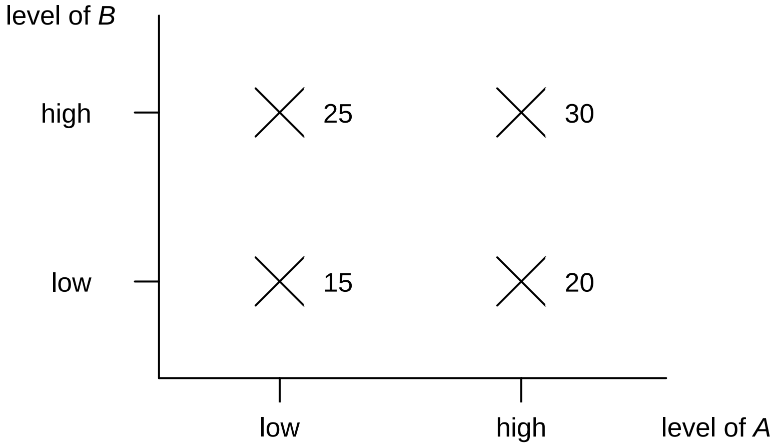
In the original units (i.e. taking anti-logarithms, base *e*), the interval is (1.486, 2.326). These are in terms of the actual diameters, in mm. Similarly, the point estimate of the mean difference in the original units is $e^{0.62}$ = 1.859 mm.

Graduate Diploma, Applied Statistics, Paper II, 2007. Question 3

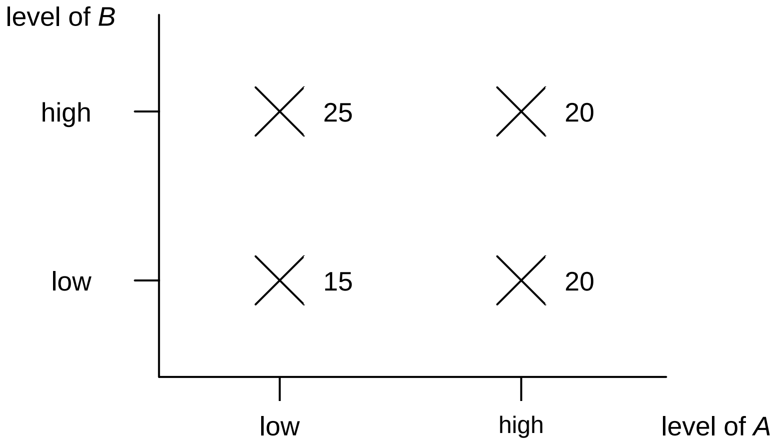
We use the usual nomenclature and notation. The main effect of a factor in a 2^2 experiment is the difference between the results with the factor at its high level and those with it at its low level; thus, for factor A, it is given by $ab + a - (b + (1))$ [an average difference might be used, i.e. with a divisor of 2]. Similarly, for B it is given by $ab + b - (a + (1))$.

The remaining independent comparison that is possible is $ab + (1) - (a + b)$. By rearranging this as $(ab - b) - (a - (1))$, it can be seen to measure the difference between the "responses" to factor A at the high level of B and those at the low level of B. Equivalently, the roles of A and B can be interchanged throughout this. It is called the interaction between A and B.

The first diagram below gives an example of results from a situation where there is no interaction (and no experimental error). Observations are indicated for the four combinations of a level of A and a level of B. At *each* level of B, the "response" to A increases by 5 as we move from the low to the high level. Similarly, at *each* level of A, the "response" to B increases by 10 as we move from low to high.



The next diagram illustrates a situation where there is interaction. The response to either factor varies, depending on the level of the other factor.



Solution continued on next page

Note that alternative forms of diagrams may be used. A common form is similar to what is shown in the two halves of the diagram on the next page. Each half shows a two-factor diagram for factors *A* and *B*, with "yields" plotted, a line joining the points for *A* at high and low levels at the low level of *B*, and another line joining the points for *A* at high and low levels at the high level of *B*. If these lines were parallel, it would indicate absence of interaction. The diagrams on the next page show strongly non-parallel lines (even to the extent that they cross each other), giving a strong indication of the presence of interaction.

Parts (i) to (iv)

The grand total is 175; the "correction factor" is $175^2/16 = 1914.0625$.

So the total sum of squares $= 2605 - \dfrac{175^2}{16} = 690.9375$, with 15 df.

SS for hospitals $= \dfrac{69^2}{8} + \dfrac{106^2}{8} - \dfrac{175^2}{16} = 85.5625$, with 1 df.

To find the SS for *A*, we need the totals for the low and high levels of *A*. These are 85 (= 5 + 25 + 15 + 40) and 90 respectively. So we have

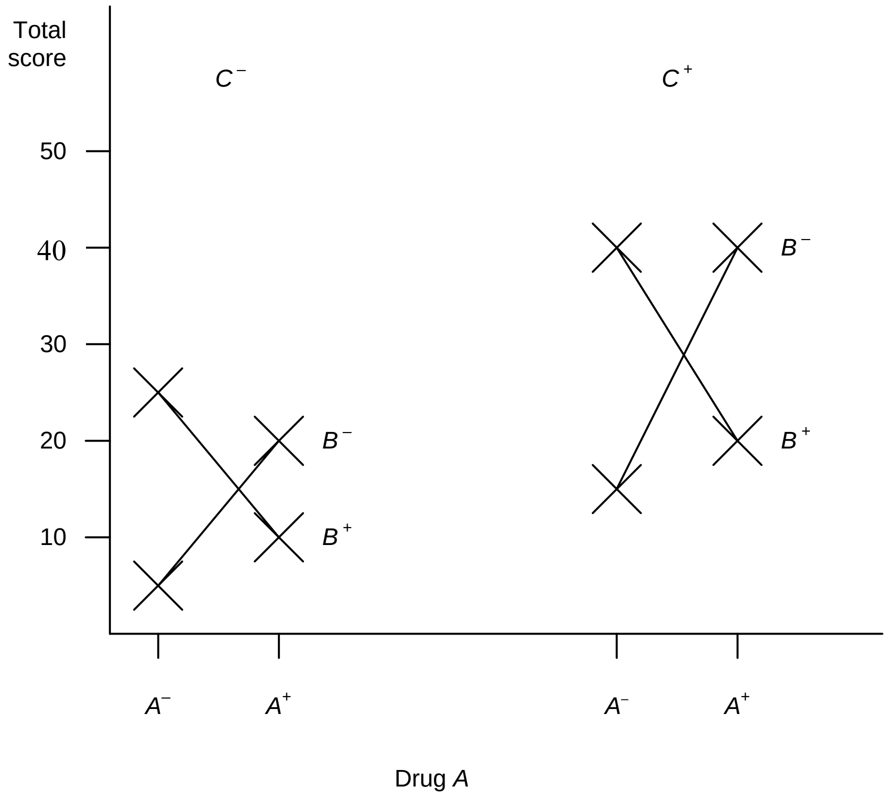
$$\text{SS for } A = \frac{85^2}{8} + \frac{90^2}{8} - \frac{175^2}{16} = 1.5625, \text{ with 1 df.}$$

The remaining entries can now be found by subtraction, and the complete analysis of variance table is as follows.

Source of variation	df	Sum of squares	MS	*F* value
Hospitals	1	85.5625		18.753
A	1	1.5625		0.342
B	1	14.0625		3.082
C	1	189.0625		41.438
AB	1	351.5625		77.055
AC	1	1.5625		0.342
BC	1	1.5625		0.342
ABC	1	14.0625		3.082
Treatments	7	573.4375	81.9196	17.955
Residual	7	31.9375	4.5625	
Total	15	690.9375		

A suitable diagram to show the relationships between the factors is shown on the next page.

Solution continued on next page



(Superscripts – and + indicate the low and high levels of the factors.)

Solution continued on next page

The diagram on the previous page suggests a possible *C* effect, since the responses at the high level of *C* ("C^+") are higher than the corresponding responses at the low level of *C* ("C^-"). It also appears that there is a two-factor interaction *AB* (so the main effects of *A* and *B* should not be studied in isolation), but probably no *AC* or *BC* interactions. Further, as the sections of the diagram for low and high *C* are quite similar, it seems unlikely that there is a three-factor *ABC* interaction.

For formal significance tests, we can compare each of the single-degree-of-freedom effects in the analysis of variance table on the previous page with the residual mean square in the usual way. The resulting *F* values are shown in the table (mean squares for these effects have not been shown, to avoid cluttering the table; they are of course the same as the sums of squares as they each have 1 df).

The *F* values are each referred to $F_{1,7}$; the upper 5% point is 5.59, the upper 1% point is 12.25 and the upper 0.1% point is 29.25. So there is very strong evidence of a difference between hospitals (and, as high values indicate better quality of life, hospital 2 seems to be the better) and extremely strong evidence for a main effect of factor *C* (psychotherapy: it appears to be better to be given psychotherapy) and for an interaction between factors *A* and *B* (the two drugs; it appears to be better to be given one of them but not both).

Apparently the patients and medical staff responsible for conducting the trial knew which treatment each patient received; this must be true of factor *C* (no explicit information is given about *A* and *B*). This would be very likely to lead to considerable bias in results, especially in a psychiatric study. The observed effect of *C* could be just this – or it might of course be genuine.

Also, the patients should have completed the same questionnaire before the experiment as well as afterwards. The data for analysis would then consist of the differences between the two scores. This would remove personal differences in attitude.

(i) Taking more than one observation at a given point enables an estimate of experimental error to be obtained from the repeat observations, regardless of what model is fitted.

(ii)(a)

$$\mathbf{X'X} = \begin{bmatrix} 9 & 0 & 0 \\ 0 & 4 & 0 \\ 0 & 0 & 4 \end{bmatrix} \text{ and so } \left(\mathbf{X'X}\right)^{-1} = \begin{bmatrix} 1/9 & 0 & 0 \\ 0 & 1/4 & 0 \\ 0 & 0 & 1/4 \end{bmatrix}.$$

(The "dash" indicates the transpose of the matrix. A notation of superscript T is also commonly used.)

Thus $\text{Var}\left(\hat{\beta}_0\right) = (1/9)\sigma^2$, $\text{Var}\left(\hat{\beta}_1\right) = (1/4)\sigma^2$, $\text{Var}\left(\hat{\beta}_2\right) = (1/4)\sigma^2$, and all covariances are zero.

$$\therefore \text{Var}\left(\hat{Y}\right) = \text{Var}\left(\hat{\beta}_0\right) + x_1{}^2\text{Var}\left(\hat{\beta}_1\right) + x_2{}^2\text{Var}\left(\hat{\beta}_2\right) = \sigma^2\left(\frac{1}{9} + \frac{1}{4}\left(x_1{}^2 + x_2{}^2\right)\right).$$

(ii)(b)

We now have $\left(\mathbf{X'X}\right)^{-1} = \begin{bmatrix} 1/9 & 0 & 0 \\ 0 & 1/3 & 0 \\ 0 & 0 & 1/3 \end{bmatrix}$ and so, similarly, we get

$$\therefore \text{Var}\left(\hat{Y}\right) = \text{Var}\left(\hat{\beta}_0\right) + x_1{}^2\text{Var}\left(\hat{\beta}_1\right) + x_2{}^2\text{Var}\left(\hat{\beta}_2\right) = \sigma^2\left(\frac{1}{9} + \frac{1}{3}\left(x_1{}^2 + x_2{}^2\right)\right).$$

(ii)(c)

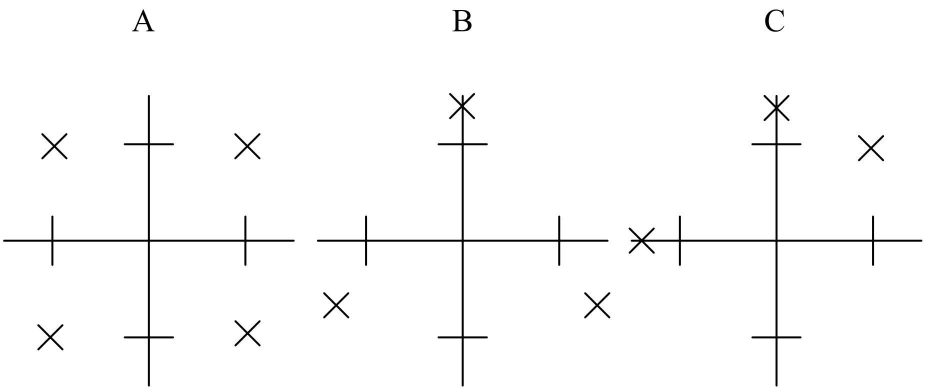
Here there will be non-zero off-diagonal terms in the matrix, indicating that there are non-zero covariances. Thus the estimators are no longer independent of each other.

The additional terms in the formula for Var($\hat{Y}$) are

$$2x_1\text{Cov}\left(\hat{\beta}_0, \hat{\beta}_1\right) + 2x_2\text{Cov}\left(\hat{\beta}_0, \hat{\beta}_2\right) + 2x_1x_2\text{Cov}\left(\hat{\beta}_1, \hat{\beta}_2\right).$$

Solution continued on next page

(iii) Three separate graphs are shown to avoid an excessively cluttered display. The limits of electronic reproduction militate against great accuracy here. Each graph has tick marks at 1 and –1 on each axis and crosses to indicate the design points.



A	B	C
Also five points at (0, 0)	Also six points at (0, 0)	Also six points at (0, 0)

(iv) The experimenter will have used design A. If the first-order model proves inadequate, more experimental points will be needed so as to be able to fit a second-order model

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_{11} X_1{}^2 + \beta_{22} X_2{}^2 + \beta_{12} X_1 X_2 + \varepsilon\ .$$

The experimenter might be near the optimum, i.e. the values of X_1 and X_2 that give the maximum Y. If so, the point (0, 0) would serve as the centre of a design to examine curvature as given by the second-order model. A central composite rotatable design could be used, containing the points in design A and (√2, 0), (–√2, 0), (0, √2) and (0, –√2).

(i) In stratified random sampling, a population is divided into groups (strata). The groups may be fairly homogenous within themselves, but possible systematic differences are expected between them. Independent simple random samples are taken within each group.

For proportional allocation, the sample sizes $\{n_h\}$ in the strata are in the same ratio as the stratum sizes $\{N_h\}$ in the whole population. Optimal allocation chooses the $\{n_h\}$ so as to minimise the variance of an estimator of population mean, total or proportion for a given total cost (budget). It can be shown that the $\{n_h\}$ must be proportional to $N_hS_h/\sqrt{c_h}$, where S_h and c_h are respectively the standard deviation and the cost of sampling a unit in stratum h.

If the population can be divided into useful strata, the variance within strata should be much less that the overall population variance. Stratification will then improve the precision of estimates considerably. Examples would be population surveys in urban and rural parts of a region, agricultural surveys in different climatic and soil conditions, social surveys in which age-groups are the strata, industrial surveys of small and large companies, and so on. The improvement by stratification is greatest when the population is stratified by the value of the quantity to be measured in the survey, or some variable highly correlated with it.

(ii) There are N = 100,000 companies; N_1 = 20,000 are large and N_2 = 80,000 are small. The total sample size is n = 1000; let this consist of n_1 large companies and n_2 small ones. Sampling costs are the same for any unit (company).

(a) Ignoring the finite population correction (and for large N),

$$\text{Var}\left(p_{st}\right)=\sum_h W_h^{\,2}\frac{P_h\left(1-P_h\right)}{n_h} ,$$

where $W_h = N_h/N$ (W_h is the "stratum weight" for stratum h).

(b) (1) With proportional allocation, it is immediate that $n_1 = 200$ and $n_2 = 800$.

(2) With optimal allocation with constant cost of sampling, we have (for large N)

$$\frac{n_h}{n} = \frac{W_h\sqrt{P_h(1-P_h)}}{\sum W_h\sqrt{P_h(1-P_h)}}$$

Using the values $P_1 = 0.9$ and $P_2 = 0.5$ (i.e. using the pilot survey estimates), and as $n = 1000$, we have

Solution continued on next page

$$\sum W_h \sqrt{P_h(1-P_h)} = 0.2\sqrt{0.9\times 0.1} + 0.8\sqrt{0.5\times 0.5} = 0.46$$

and thus

$$n_1 = 1000\times\frac{0.2\sqrt{0.9\times 0.1}}{0.46} = 130.435$$

$$n_2 = 1000\times\frac{0.8\sqrt{0.5\times 0.5}}{0.46} = 869.565\,.$$

So we take $n_1 = 130$ and $n_2 = 870$.

(c) The relative efficiencies are measured as (inverse) ratios of variances.

For a simple random sample of size *n*, for large *N* and ignoring the finite population correction, the variance of the estimator of the population proportion is given by $P(1 - P)/n$ where *P* is here the overall proportion which recognise trade unions which (using the information from the pilot survey) we take as 58000/100000 = 0.58. Thus we have

$$\text{Var}(p_{srs}) = (0.58)(0.42)/1000 = 0.0002436.$$

Using the result in (ii)(a), the variances for stratified sampling are

for proportional allocation:

$$\text{Var}(p_{prop}) = 0.2^2\,\frac{0.9\times 0.1}{200} + 0.8^2\,\frac{0.5\times 0.5}{800} = 0.000218$$

for optimal allocation:

$$\text{Var}(p_{opt}) = 0.2^2\,\frac{0.9\times 0.1}{130} + 0.8^2\,\frac{0.5\times 0.5}{870} = 0.0002115\,.$$

Therefore the relative efficiencies are

for proportional allocation: $\dfrac{0.0002436}{0.000218} = 112\%$

for optimal allocation: $\dfrac{0.0002436}{0.0002115} = 115\%\,.$

Solution continued on next page

(d) For this survey, stratified sampling with either form of allocation gives a modest but noticeable improvement in the efficiency of estimation of the proportion for the whole population; optimal allocation is only slightly better than proportional allocation.

However, the improvement is far from spectacular. Stratification will not give great benefit unless the $\{P_h\}$, and hence the stratum variances, vary considerably between strata. When means of a measured variable are being estimated, there is more scope for substantial benefit through large differences in variances than there is for proportions.

Graduate Diploma, Applied Statistics, Paper II, 2007. Question 6

(i) A large region in a developing country is very likely to have communication and transport problems which will slow down the sampling process.

There may be maps which will help to identify major areas of agriculture, although most of the region probably grows some crops where it can. If available, aerial photographs (or satellite images) could be very useful.

Stratification might be by geographically or climatically different parts of the region. Stratification would ensure that all such sub-regions are studied, which might not be the case under simple random sampling.

Clustering would seek to identify parts of the region that exhibit most of the characteristics of the whole region. Survey work could then be restricted to a few clusters (maybe only one), instead of having to try to cover the whole region.

One possibility would be to carry out an initial stratification using all available information and local knowledge, then form clusters within each stratum and choose a sample of these. An administrative base for the survey could be set up in each stratum, where the work for the chosen clusters would be coordinated. It is likely to be important not to have to visit isolated parts of the region more than once, and to maximise the information obtained from each such visit.

(ii) (a) The simple random sample estimate of the population total is

$$\hat{Y}_{srs} = N\bar{y} = 75308 \times \frac{25751}{2055} = 943677(.04)\,.$$

To find the estimated variance underlying this, we first calculate

$$s^2 = \frac{1}{2054}\left(596737 - \frac{25751^2}{2055}\right) = 133.4244\,.$$

Thus the estimated variance is

$$N^2(1-f)\frac{s^2}{n} \;=\; 75308^2\left(1-\frac{2055}{75308}\right)\frac{133.4244}{2055}$$

and, taking the square root, the standard error is therefore 18925.4.

Solution continued on next page

The ratio estimate is $\hat{r} = \frac{\Sigma y_i}{\Sigma x_i} = \frac{25751}{62989} = 0.4088$. So the ratio estimate of the population total is $\hat{Y}_R = \hat{r}X = 0.4088 \times 2353365 = 962055(.61)$ [note that the value here may be found slightly but noticeably different depending on the level of accuracy with which "0.4088" is worked].

Using the given formula, the estimated variance underlying this is

$$\frac{75308 \times 73253}{2055 \times 2054}\left\{596737 - (2 \times 0.4088 \times 1146391) + (0.4088^2 \times 2937851)\right\}$$

$$= 1306.9358 \times 150413.8566$$

and taking the square root gives the standard error as 14020.7.

(b) We are given that the linear regression estimate is $\hat{Y}_{LR}$ = 969651.6 and the standard error is 13881.9. So both this and the ratio method give very similar estimates and precisions. The estimate from simple random sampling is a little lower and distinctly less precise. The relative efficiencies, using SRS as base, are

for the ratio estimate: $\left(\frac{18925.4}{14020.7}\right)^2 = 182\%$

for the linear regression estimate: $\left(\frac{18925.4}{13881.9}\right)^2 = 186\%$.

Thus using the supplementary information on *x* has improved precision by more than 80%.

(c) If the supplementary variable is strongly positively correlated with *y*, as here, then the ratio estimator will be more efficient than the estimator using the simple random sample mean. The regression estimator is never less efficient than the simple random sample estimator. Finally, the regression estimator cannot be less efficient that the ratio estimator. It makes fewer assumptions about the relationship between *y* and *x* (in effect, the ratio estimator assumes a regression that passes through the origin).

Both the ratio and the regression estimators are biased in small samples, but the bias is negligible in large samples.

Graduate Diploma, Applied Statistics, Paper II, 2007. Question 7

(i) This may be considered as a cluster sample with the hospitals as the clusters (i.e. as the first-stage units) and the individual patients as the second-stage units, limited to those patients suffering from the specified conditions.

(ii) (a) The number of clusters is $N = 33$, of which $n = 10$ have been sampled.

The total number of persons in the sample is $m = 560 + \ldots + 110 = 2210$.

Thus, for the simple random sample of 10 hospitals, the sample mean is $\bar{m} = 2210/10 = 221$, and this is an unbiased estimate of the population mean of the cluster totals. Hence a point estimate of the total number of persons with these conditions in the 33 hospitals is $33 \times 221 = 7293$.

The variance underlying this estimated total is $N^2\text{Var}(\bar{m})$. We have as usual that $\text{Var}(\bar{m})$ is estimated by $(1-f)s^2/n$ where, in the customary notation, $f = 10/33$ and $s^2 = (1/9)\{680700 - (2210^2/10)\} = 21365.56$.

Thus the standard error of the estimated total is

$$\sqrt{33^2\left(1-\frac{10}{33}\right)\frac{21365.56}{10}} = 1273.44\,.$$

The double-tailed 5% point of t_9 is 2.262, so an approximate 95% confidence interval for the total is given by

$$7293 \pm (2.262 \times 1273.44)$$

i.e. it is (4412.5, 10173.5).

(b) A point estimate of the proportion of people discharged dead is

$$\hat{p} = \frac{4+4+\ldots+1}{560+190+\ldots+110} = \frac{30}{2210} = 0.013575\,.$$

An approximate expression for the estimated variance underlying this estimate is

$$\frac{1-f}{n\bar{m}^2}s_p^{\,2}$$

where, using a customary notation,

Solution continued on next page

$$s_p^{\ 2} = \frac{1}{n-1}\sum\left(a_i - \hat{p}m_i\right)^2$$

$$= \frac{1}{9}\left\{\left(4-(0.013575\times 560)\right)^2 + \ldots + \left(1-(0.013575\times 110)\right)^2\right\}$$

= 7.699869 (note: slightly but noticeably different numerical values will be obtained depending on the accuracy of working).

Thus the approximation for the estimated variance of $\hat{p}$ is

$$\frac{1-(10/33)}{10\times 221^2}\times 7.699869 \;=\; 0.000010988$$

and the standard error is the square root of this, i.e. 0.0033147.

Now again using the double-tailed 5% point of t_9, i.e. 2.262, an approximate 95% confidence interval for the proportion is given by

$$0.013575 \pm (2.262\times 0.0033147)$$

i.e. it is (0.0061, 0.0211).

This is based on the Normal approximation to the binomial distribution, which is scarcely likely to be valid for a p so small and the samples not especially large. The large-sample formula used for the variance may also not be very valid here. The result should be regarded as very approximate. (There is also the point that $\hat{p}$ is a ratio estimator and therefore biased, though the bias may be small.)

(iii) The main reason why cluster sampling might be preferred is that stratified sampling would need a full population list whereas cluster sampling only needs the information for the chosen clusters. Extracting data items from detailed hospital records is likely to be time-consuming and expensive, so reducing the number required is important. It is also quite likely to be the case that some hospitals would not give permission for their patient records to be explored.

Stratified sampling, however, is likely to lead to more precise estimates. Hospital sizes vary, and larger hospitals may account for a higher proportion of patients in these conditions.

Simple random sampling of the clusters leads to units in small clusters having greater probability of selection that units in large clusters; this can lead to estimates having bias and low precision. Sampling with probability proportional to size would be an improvement, for example based on the numbers of Accident and Emergency admissions at the hospitals.

Graduate Diploma, Applied Statistics, Paper II, 2007. Question 8

Comparison of crude death rates may be confounded by differences in the population structure of the subgroups being compared. Standardisation allows comparisons free of the effect of differences in the numbers of individuals in the subgroups of the populations. These sub-groups are typically defined by age or age and sex.

Direct standardisation involves defining a standard population and applying to it the specific death rates for the subgroups being compared. This gives the number of deaths expected in the standard population if these specific rates were to apply. Indirect standardisation applies a known set of specific death rates for a standard population to the subgroup populations.

Each method assumes that the relative increase in mortality with age is the same in each population, standard and other.

(i) The crude rate for CHD for males in the UK in 2004 is 58555/29270000 or 200.1 per 100,000. For females, it is 47287/30563000 or 154.7 per 100,000.

(ii) The direct adjusted rate is $\Sigma N_i p_i / \Sigma N_i$ where N_i is the number of individuals in age group i of the European Standard Population and p_i is the age-specific death rate in the study population (UK males in 2004) for group i. This is calculated using the following table. The final column of the table is used in part (iii).

Age group i	p_i	$N_i p_i$	$n_i p_i$
< 35	1.0106	0.5053	10.9956
35 – 44	18.6731	2.6142	71.7047
45 – 54	80.4497	11.2630	272.7246
55 – 64	217.3746	23.9112	636.9077
65 – 74	595.9983	41.7199	1233.7165
75 +	1922.5393	76.9016	2576.2027
TOTAL		156.9152	4802.2518

ΣN_i is here 100,000 so the required direct adjusted rate is 156.9 per 100,000.

(iii) Here the standard population is that of UK males. Let n_i be the number (in thousands) of males in age group i in Scotland. Then $\Sigma n_i p_i = 4802.2518$, as shown in the last column of the above table. The actual number of deaths in Scotland is 5814 (stated in the introduction to the question). So the standardised mortality ratio (deaths per thousand individuals per year) is $5814/4802.2518 = 1.211$.

Solution continued on next page

(iv) CHD mortality figures are as follows.

	UK *crude rate*	*UK* *direct adjusted rate*	*Scotland* *standardised* *mortality ratio*
Males	200.1	156.9	1.21
Females	154.7	80.13	1.23

The mortality rates are considerably higher for males than for females in the UK. The crude rates indicate an increase of about 30% for males over females; the direct adjusted rate indicates that it is almost doubled. It is very noticeable in the detailed table given in the question that there are many more deaths for males than females in every age group except "75+".

The standardised mortality ratios indicate that CHD mortality is just over 20% higher in Scotland than in the UK as a whole, for either sex.