THE ROYAL STATISTICAL SOCIETY

2005 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, 2005. Question 1

(i) A complete randomised block is simple to analyse because it is an "orthogonal" design and so the parameters for treatments and blocks can be estimated independently and the sums of squares are also independent. A missing value causes all these properties to be lost. Estimates of treatment parameters are biased according to whether the missing plot was in a good or bad block.

(ii) There are t treatments and b blocks. T_i' is the incomplete treatment total and B_j' the incomplete block total for the treatment and block with the missing observation. Similarly, G' is the grand total of the surviving observations.

(iii) $\hat{x} = \{(4\times176)+(6\times98)-856\}/15 = 29.067$.

We may reasonably take this as 29 when carrying out the analysis as all the other data items are integers. The analysis proceeds as for a complete randomised block but is only approximate. The total and residual degrees of freedom are each reduced by 1. The sums of squares are approximate. The residual on the affected plot is 0, and so F tests are likely to be biased towards significant results. Similarly, individual differences between treatment means can only be tested approximately. The analysis assumes that the lost data item was not related to treatment.

(iv) The analysis gives 14 df for residual (not 15).

For any two means except D, the standard error of a difference is $\sqrt{2\times3.7111/6} = 1.112$.

For D and another mean, $\frac{2}{6}$ is replaced by $\frac{1}{6}+\frac{1}{5}=\frac{11}{30}$ and so the SE is $\sqrt{11\times3.7111/30} = 1.167$.

The treatment means (in ascending order) are

A 32.33 D 34.17 B 37.33 C 43.67.

To examine differences between the underlying population means, we divide each difference by its SE and refer to t_{14} (with some care in interpretation because of "multiple comparisons"). The double-tail 5%, 1% and 0.1% points of t_{14} are 2.145, 2.977 and 4.140 respectively. Thus we find

Solution continued on next page

$$\frac{\text{A}-\text{D}}{SE(\text{A}-\text{D})}: \quad \frac{-1.84}{1.167} = -1.58 \qquad \text{not significant}$$

$$\frac{\text{A}-\text{B}}{SE(\text{A}-\text{B})}: \quad \frac{-5.00}{1.112} = -4.50 \qquad \text{very highly significant}$$

We can reasonably conclude that there is very strong evidence that A is less than B (and C) but no real evidence that A is less than D.

$$\frac{\text{D}-\text{B}}{SE(\text{D}-\text{B})}: \quad \frac{-3.16}{1.167} = -2.71 \qquad \text{significant}$$

$$\frac{\text{D}-\text{C}}{SE(\text{D}-\text{C})}: \quad \frac{-9.50}{1.167} = -8.14 \qquad \text{very highly significant}$$

We can reasonably conclude that there is some evidence that D is less than B and very strong evidence that D is less than C.

$$\frac{\text{B}-\text{C}}{SE(\text{B}-\text{C})}: \quad \frac{-6.34}{1.112} = -5.70 \qquad \text{very highly significant}$$

We can reasonably conclude that there is very strong evidence that B is less than C.

In respect of the additional information about the factor "Bran", the 3 df for treatments could be separated into linear, quadratic and cubic components for equally spaced levels of this factor.

(v) If two missing values x, y are to be estimated, an iterative process can be used. Guess a value for y, say y_1, and use the formula to find x_1. Then put x_1 into the data, take out y_1 and use the formula to find y_2. Repeat alternately for x and y until the process converges to a pair of values x^*, y^* and then complete the analysis using these.

Graduate Diploma, Applied Statistics, Paper II, 2005. Question 2

(i) A complete factorial design for m factors, each at 2 levels, uses all possible combinations of the two levels of the m factors; it has 2^m different treatment combinations.

Confounding into two blocks, each containing half (2^{m-1}) of the combinations, requires one interaction to be lost. Usually the highest order interaction is chosen. Consider an example with 3 factors (A, B and C, each at 2 levels), with the 3-factor interaction ABC used as the basis for confounding. Treatment combinations that are "even" with respect to ABC are placed in one block, and those that are "odd" with respect to ABC in the other block. Thus the blocks are (1), *ab*, *ac*, *bc* and *a*, *b*, *c*, *abc*. However, the interaction ABC would be estimated using the contrast (*abc*+*a*+*b*+*c*) – (*ab*+*ac*+*bc*+(1)) – so we cannot estimate this because it has become part of the block differences.

(ii) (a) Consider the principal block (block 1, because it contains (1)). There are 3 confounded interactions and each has to have 0, 2 or 4 letters in common with each of the contents of this block. CD contains 0 or 2; so does ABD; so does ABC. These three are confounded, in each of the two replicates.

(b)

SOURCE OF VARIATION		D.F.	
Blocks		7	(See * below table)
Unconfounded effects			
A, B, C, D	4		
AB, AC, AD, BC, BD	5		
ACD, BCD	2		
ABCD	1		
		12	
Residual		12	
TOTAL		31	

* The 7 df for the blocks can be broken down as 1 for replicates, 3 for the three confounded interactions (ABC, ABD, CD) and 3 for the interactions between replicates and blocks.

Solution continued on next page

(c) The residual mean square is 9705. Three 3- and 4-factor interactions are not significant: ACD, BCD, ABCD.

The overall mean for the low level of A is $A^- = ½(736.500 + 310.250) = 523.375$.

The overall mean for the high level of A is $A^+ = ½(657.625 + 234.125) = 445.875$.

(These come from the AB two-way table; we may similarly use any other table involving A.)

$\therefore A = A^+ - A^- = -77.5$.

For the BC "effect", we consider the BC table. The mean of the forward diagonal terms is ½(927.125 + 309.250) = 618.1875, and of the backward diagonal terms is 351.0625. The difference is BC = +267.125.

[This could also be expressed more formally in the usual notation for a 2^4 experiment.]

(d) A factorial effect is the difference between the + part of the appropriate contrast and the – part. Each contains 8 items in a 2^4 experiment, and here there are two replicates.

So the required SE is $\sqrt{\frac{s^2}{16}+\frac{s^2}{16}} = \sqrt{\frac{s^2}{8}}$ where $s^2 = 9705$, i.e. we have $\text{SE} = \sqrt{1213.125} = 34.83$.

The double-tail 5% point of t_{12} is 2.179, so significant effects are those $> 2.179 \times 34.83$ (in absolute value), i.e. > 75.89 (in absolute value), which are A (just, at 5%), B, C, D, BC, BD.

We note that A does not feature in any significant interactions, and that low values of y occur at the high level of A.

For the other factors, we must look at interactions, i.e. the two-way tables. Because CD is confounded, we gain no information from the CD table. BC is a "positive" effect (see part (c) above), BD is "negative". In each of the BC and BD tables, we see that the lowest mean occurs at the high level of B, and the low levels of C and D are required.

This suggests the combination $A^+B^+C^-D^-$, but this picks up quite a high mean in the AC table. Revisiting the BC table, we note that the means for B^+C^- and B^+C^+ do not differ significantly (SE = $\sqrt{2s^2/8}$ = 49.257).

Hence the best choice appears to be A^+, B^+, C^+, D^-.

Graduate Diploma, Applied Statistics, Paper II, 2005. Question 3

(i) The experimental units might be agricultural plots, people in a medical trial, industrial items, … . If two units (plots, people, items, …) are treated in exactly the same way (or as nearly so as possible), they will hardly ever give exactly the same yield or response. This is natural variation, which has to be measured by including more than one *replicate* of each treatment in an experiment. Replication also guards against assessing a treatment by a single "rogue" observation. Adequate replication is one factor in obtaining sufficient residual degrees of freedom in an analysis, especially for the estimate of σ^2.

Randomisation avoids bias that would be likely to occur (despite "best efforts" to the contrary) if the experimenter made the choice of which treatment to apply to which unit. If allocation is made completely at random, each unit has the same probability of carrying each of the treatments under test.

Suppose that a new drug is being tried in a medical trial for treating a chronic condition. After a reasonably long period of treatment, it may appear that the condition of patients is no different from what it was at the start. However, chronic conditions not receiving effective treatment would usually become noticeably (measurably) worse. So in fact the new drug was, to some extent at least, successful.

For a satisfactory trial, two groups of patients in similar condition should be used, one to receive the new drug and one to receive a standard drug or a placebo. The condition of each group after the same period of trial should be observed and comparison made to give a valid indication of whether there is any difference between treatments.

(ii)(a) The grand total is 380; the "correction factor" is $380^2/60 = 2406.667$.

So the total sum of squares = $5050 - \dfrac{380^2}{60} = 2643.333$, with 59 df.

The duration totals are: Short 246, Long 134.

$$\therefore \text{ SS duration} = \frac{246^2}{30} + \frac{134^2}{30} - \frac{380^2}{60} = 2615.733 - 2406.667$$

$$= 209.066, \text{ with 1 df.}$$

The weight gain totals are: Mild 49, Moderate 110, Severe 221.

$$\therefore \text{SS weight gain} = \frac{49^2}{20} + \frac{110^2}{20} + \frac{221^2}{20} - 2406.667 = 760.433, \text{ with 2 df.}$$

Solution continued on next page

$$\text{Interaction SS} = \frac{27^2}{10} + \frac{73^2}{10} + \ldots + \frac{75^2}{10} - 2406.667 - 209.066 - 760.433$$

$$= 109.034, \text{ with } 1 \times 2 = 2 \text{ df.}$$

The residual SS and df follow by subtraction.

Hence:

SOURCE	DF	SS	MS	F value
Duration	2	209.066	209.066	7.21
Weight gain	1	760.433	380.217	13.12
Interaction	2	109.034	54.517	1.72
Residual	54	1564.800	28.978	$= \hat{\sigma}^2$
TOTAL	59	2643.333		

The F value of 7.21 is referred to $F_{2,54}$; this is well beyond the upper 1% point (about 5) [but not beyond the upper 0.1% point], so there is strong evidence of an effect of duration.

The F value of 13.12 is referred to $F_{1,54}$; this is beyond the upper 0.1% point (about 12), so there is very strong evidence of an effect of weight gain.

The F value of 1.72 is referred to $F_{2,54}$; this is not significant, so there is no evidence of any interaction.

Clearly short duration leads to higher number of days hospitalised. The means are 8.2 for short duration and 4.5 (4.4667) for long.

For weight gain, the means are 2.45, 5.50 and 11.05 respectively. The SE for the difference between any two of these is $\sqrt{\frac{2\hat{\sigma}^2}{20}} = 1.702$. The double-tail 5% point of t_{54} is slightly greater than 2, so a difference of about 3.5 would be significant at the 5% level. Thus Mild and Moderate are not statistically significantly different (though 3.05 days would be important to a hospital). Severe is very highly significantly greater than the other two.

The worst result overall is thus short treatment duration for cases of severe weight gain.

(b) Underlying Normality might be open to question, but a key feature is that responses are much more variable in some treatments than in others. This seems to be a situation where a log transformation should be applied, as variability increases with size of response, giving a skew pattern in the data.

Since there are zeros among the data, $\log(y+1)$ is appropriate. [It is likely that the very clear result would remain substantially the same, but tests would have more validity.]

Graduate Diploma, Applied Statistics, Paper II, 2005. Question 4

(i) A ¼-replicate design requires choice of two interactions to be confounded. Since A, B, C do not interact, nor do C, D, E, we can confound ABC and CDE. (Their generalised interaction ABDE is automatically confounded as well; presumably this is acceptable.)

The principal block contains (1) and the treatment combinations having 0, 2 or 4 letters in common with the confounded interactions. Block size will be 8.

Principal block :

(1), *ab*, *de*, *abde*, *acd*, *bcd*, *ace*, *bce*.

Use this or one of the other blocks, generated by using *a*, *c* or *d* in turn:

a, *b, ade*, *bde*, *cd*, *abcd*, *ce*, *abce*

OR *c*, *abc*, *cde*, *abcde*, *ad*, *bd*, *ae*, *be*

OR *d*, *abd*, *e*, *abe*, *ac*, *bc*, *acde*, *bcde*.

The order of running the items in the chosen block will be randomised.

This allows a full first-order model to be fitted, without any product terms (e.g. $b_{12}\, x_1\, x_2$), but leaves only 2 df for "residual".

(ii) Extra "centre points" are required. A 2-level design only uses points which may be coded ± 1 for each x; e.g. for x_1, it is coded to speed = $\dfrac{75-67.5}{7.5} = +1$ and $\dfrac{60-67.5}{7.5} = -1$. The centre for x_1 is 67.5, which can be coded 0. Similar centres are found for B, C, D, E. Two or more points at (0, 0, 0, 0, 0) can be added to the design, supplying both extra degrees of freedom for a test of fit and some idea of whether the model, linear over (–1, 1), is adequate.

If there are k (≥ 2) centre points, the corrected sum of squares with (k – 1) df calculated between these k responses is a "pure error" term in the analysis and may be regarded as a genuine estimate of the variance underlying an observation y. The remaining part of the residual has 3 df after fitting $y = b_0 + \sum_{i=1}^{5} b_i x_i$, and is "lack of fit". $F_{3,k-1}$ compares lack of fit with pure error.

Solution continued on next page

(iii) If the analysis in (ii) showed evidence of lack of linear fit, so that a model with second order (product and quadratic) terms was required, we may or may not have reached a region where there is an optimum (maximum or minimum) value of the response y. If so, the extra points for fitting the extra terms may be those needed for a rotatable central composite design, namely $(\pm\alpha, 0, 0, 0, 0), \ldots, (0, 0, 0, 0, \pm\alpha)$, where for a 2^5 design $\alpha = 2^{3/4} = 1.68$.

If there is curvature but no evidence of near-optimality, extra values of some or all x variables will be needed. Any knowledge of the operation being studied should be used in deciding on them. A suitable 3^5 design, probably fractional, would be a possibility.

(iv) In a fractional design, undesirable aliasing of main effects and interactions needs to be considered. Here the defining congruence is I = ABC = CDE = ABDE.

Hence A = BC = ACDE = BDE is an alias set, and others are
B = AC = BCDE = ADE
C = AB = DE = ABCDE
D = ABCD = CE = ABE
E = ABCE = CD = ABD

Also we will have AD = BCD = ACE = BE
AE = BCE = ACD = BD

If we had not had the given information that A, B, C do not interact with each other and that C, D, E do not interact with each other, we would not have been able to estimate all the important main effects and two-factor interactions (for example, it would have been unsatisfactory that A and BC were aliased). As it is, the designs in (i) and (ii) are reasonably adequate.

Graduate Diploma, Applied Statistics, Paper II, 2005. Question 5

(i) (a) Simple random sampling requires a target population to be defined, a list of its members to be prepared, and a fully random – not subjective – selection of items from the list. Convenience sampling cannot be either fully random or fully representative, since the people available will not usually be from any particular target population. In simple random sampling, each item has the same probability of selection from the list. Chosen people are only replaced if they no longer belong to the target population, or in some surveys if it proves very difficult to contact them. Simple random sampling requires time and administration to set up and execute; convenience sampling should be quicker and cheaper. But simple random sampling has a proper theoretical base which allows numerical analysis of results, whereas convenience sampling does not.

(b) Volunteer sampling uses people who respond to some public request to take part. In both volunteer and convenience sampling, there is personal choice, either by the survey organiser or by individual respondents, and bias is very likely. Volunteer samples are either very small, as in local authority surveys where households do not return forms that have been distributed, or very unrepresentative because only people very interested in the topic of the enquiry bother to reply (or offer to reply).

A typical sampling frame for volunteer sampling would arise from a radio or television programme where only those actually listening or watching at that particular time can volunteer. Biases may be due for example to work/leisure activities, age group, economic circumstances, level of interest in the topic, ability to make contact (e.g. through a special telephone line which may be overloaded).

(ii) (a) A systematic plan will be much easier to carry out in a field laid out in a regular way, provided the distances apart of selected units do not correspond to any trend in the plants due to the layout – e.g. the sample is not all end-of-row items. If the plan in (c) is carried out in a 20×400 layout, it is difficult to think of any serious trends.

(b) Simple random sampling is used with N = 8000 plants, and we are given that S = 2 kg for a single plant.

The estimate of the total is $N\,\bar{y}_{\text{SRS}}$, with variance $N^2\left(\frac{N-n}{N}\right)\frac{S^2}{n}$ where n is the sample size. We require 1.96×√(variance) to be ≤ 2000, i.e. we require $1.96\times\left[\sqrt{N(N-n)/n}\right]\times 2 \le 2000$.

Insert n = 240 and check whether $1.96\times\sqrt{\frac{8000\times 7760}{240}}\times 2$ is ≤ 2000. It is in fact 1993, so $n \approx 240$ is satisfactory.

Solution continued on next page

(c) $\frac{8000}{33} = 242$, so n is about right. As noted above, the layout is such that a 1-in-33 scheme has no obvious trend in phase with any feature of the layout. Therefore we can treat it as effectively simple random sampling and use the variance formula as in (b).

Choose at random a number between 01 and 33 inclusive; suppose 22 is chosen. Then the first sample item is row 1, plant 22, and this is followed by plants 55, 88, 121, 154, 187, 220, 253, 286, 319, 352, 385; these are all in row 1. The next number is 385+33 = 418, which is row 2, plant 18. Continue like this to obtain about 12 sample plants from each of the 20 rows.

(iii) If the number of schools in the survey with playing fields is a, then $\hat{P}_a = \frac{a}{n}$ is an unbiased estimator of the proportion P_a in the whole population. The variance of $\hat{P}_a$ is estimated as $\hat{P}_a\left(1-P_a\right)/n$. (A finite population correction factor using $\left(1-\frac{n}{N}\right)$ may be needed is N is not large.)

Suppose there are m_i pupils in the ith school in the survey (i = 1, 2, …, n), and define the variable y_i = 0 if the school has no playing field and y_i = 1 if it has one. The number of pupils attending schools in the survey with a playing field is therefore $\sum_{i=1}^{n} m_i y_i$. The total number of pupils in the whole survey is $\sum_{i=1}^{n} m_i$. $\hat{P}_b = \frac{\sum_{i=1}^{n} m_i y_i}{\sum_{i=1}^{n} m_i}$ is a ratio estimator of P_b. (It is therefore a biased estimator.)

Graduate Diploma, Applied Statistics, Paper II, 2005. Question 6

(i) Stratified sampling splits a complete population into parts or strata, each of which is relatively homogenous within itself, though there are thought to be systematic differences between the strata. Typically, information is required for each stratum, as well as an overall population estimate. It is administratively more convenient to use compact strata.

For an example, there may be different areas in a region, such as urban and rural, coastland and inland, which should form strata. As another example, in agricultural work a useful basis for stratification is often size of farm.

Sampling problems may vary between strata. Sampling costs may vary between strata. The underlying variance may vary between strata, as is shown in the following example; a population estimate can be improved considerably by allowing for this.

(ii) N_h = number of population members in stratum h.
S_h = population standard deviation within stratum h.
n_h = sample size (number of items) taken in stratum h.

(iii) The budget is \$20000. In Design 1, \$4000 is overhead cost and so the sample size is $n = 1600$. In Design 2, overheads are \$10000 and so $n = 1000$.

The overall variance is $S^2 = \frac{1}{N-1}\left\{\sum (N_h - 1) S_h^2 + \sum N_h \left(\overline{Y}_h - \overline{Y}\right)^2\right\}$, and if p_h is the proportion in each stratum (given as % the data table) then we have, approximately,

$$S^2 = \sum_h p_h S_h^2 + \sum_h p_h \left(\overline{Y}_h - \overline{Y}\right)^2.$$

The overall mean is $\overline{Y} = \sum p_h \overline{Y}_h = 135 + 110 + 170 + 180 + 120 = 715$, so the absolute values of $\left(\overline{Y}_h - \overline{Y}\right)$ are 635, 385, 135, 115, 315. Hence

$$S^2 = \left\{(0.1 \times 600^2) + (0.1 \times 400^2) + (0.2 \times 300^2) + (0.3 \times 200^2) + (0.3 \times 160^2)\right\}$$
$$+ \left\{(0.1 \times 635^2) + (0.1 \times 385^2) + (0.2 \times 135^2) + (0.3 \times 115^2) + (0.3 \times 315^2)\right\}$$

$$= 182205.$$

Solution continued on next page

In Design 1, $\mathrm{Var}\left(\bar{y}_{\mathrm{SRS}}\right) = \dfrac{N-n}{N}.\dfrac{S^2}{n} = \left(1-\dfrac{1600}{N}\right)\left(\dfrac{182205}{1600}\right) = 113.878 - \dfrac{182205}{N}$.

In Design 2, with the choice in (ii) ("Neyman allocation"),

$$\mathrm{Var}\left(\bar{y}_{\mathrm{ST}}\right) = \frac{1}{n}\left(\sum_h p_h S_h\right)^2 - \frac{1}{N}\sum_h p_h S_h^{\,2} = \frac{1}{n}\left\{\left(60+40+60+60+48\right)^2\right\} - \frac{89680}{N}$$

$$= \frac{268^2}{1000} - \frac{89680}{N} \;=\; 71.824 - \frac{89680}{N}.$$

Therefore we have

$$\mathrm{Var}\left(\bar{y}_{\mathrm{SRS}}\right) > \mathrm{Var}\left(\bar{y}_{\mathrm{ST}}\right) \text{ when } 113.878 - \frac{182205}{N} > 71.824 - \frac{89680}{N},$$

i.e. when

$$42.054 > \frac{1}{N}\left(182205 - 89680\right) = \frac{1}{N} \times 92525$$

i.e. when

$$N > \frac{92525}{42.054} = 2200.$$

This is almost certain to be true, as we are told the district is "large". So in any case we choose stratified sampling, even if it was not essential to have figures for each stratum.

[Note: Neyman allocation gives n_h = 224, 149, 224, 224, 179.]

Graduate Diploma, Applied Statistics, Paper II, 2005. Question 7

(i) Simple random sampling will be a tedious process when only 1% of names have to be selected from a very large list. Some wards may be represented much better than others, some residents will not be on the current list if they have moved recently, some may not be available for interview at a convenient time. Since the list is almost a year old, removals in and out, and deaths, will make the list itself a less satisfactory basis for a sample. Either a number of "reserve" units must be selected to replace those not at their given address, or a new person at the same address may be accepted instead.

Stratification into wards would ensure all were properly represented, would discover any systematic differences between wards and would speed up the sampling process.

In either method, there is proper theoretical basis for estimates obtained, and no bias due to non-random selection. Results give known probabilities of selecting individuals, and can be generalised to the complete population of the area. But non-response can occur in using either method, and the problem of old lists is the same.

(ii) (a) $\hat{p}_A = \dfrac{720}{1600} = 0.45$ and the interval is given by $0.45 \pm 1.96\sqrt{\dfrac{0.45\times 0.55}{1600}}$, i.e. it is 0.45 ± (1.96 × 0.0124) or 0.45 ± 0.024, i.e. (0.426, 0.474).

(b) We may expect the current intention for voting for A to be too high, as the recollection is too high. This may or may not be due to the sample chosen (people's memories could also be inaccurate), but a regression estimator will adjust the current intention down by an amount $b(\mu - \bar{x})$, where $\mu = 0.30$ and $\bar{x} = 0.35$ in this case.

$$\hat{b} = r\frac{s_y}{s_x} = 0.6\frac{s_y}{s_x} \text{ (correlation is given as 0.6), and } s^2 = \frac{np(1-p)}{n-1}.$$

$$\frac{s_y}{s_x} = \sqrt{\frac{0.45\times 0.55}{0.35\times 0.65}} = 1.043, \quad \text{so } \hat{b} = 0.6258.$$

$$\therefore \hat{p}_{\text{LR}} = 0.45 + 0.6258(0.30 - 0.35) = 0.419.$$

Using the formula given in the question,

$$\text{Var}(\hat{p}_{\text{LR}}) = (1-f)(1-r^2)s_y^2/n = \left(1-\frac{1600}{160000}\right)(1-0.36)\left(\frac{1600}{1599}\right)\left(\frac{0.45\times 0.55}{1600}\right)$$
$$= 0.99\times 0.64\times 0.2475/1599$$
$$= 0.000098, \quad \text{and therefore the SE is } 0.0099.$$

Therefore an approximate 95% confidence interval for p_{LR} is given by 0.419 ± (2 × 0.0099), i.e. (0.399, 0.439).

(c) The recalled behaviour for A may be greater than what actually happened, so the regression estimate for current intention may have been adjusted too much, i.e. biased on the low side.

Graduate Diploma, Applied Statistics, Paper II, 2005. Question 8

(a) The infant mortality rate is

$$\frac{\text{number of deaths between birth and one year (excluding fetal deaths, stillbirths)}}{\text{total number of live births in the same year}}.$$

The neonatal mortality rate is as above but only including deaths up to 28 days.

The perinatal mortality rate is

$$\frac{\text{number of fetal deaths and neonatal deaths}}{\text{total number of live births}}$$

[sometimes divided by the total number of live births and fetal deaths, there being no generally accepted convention for computing this rate].

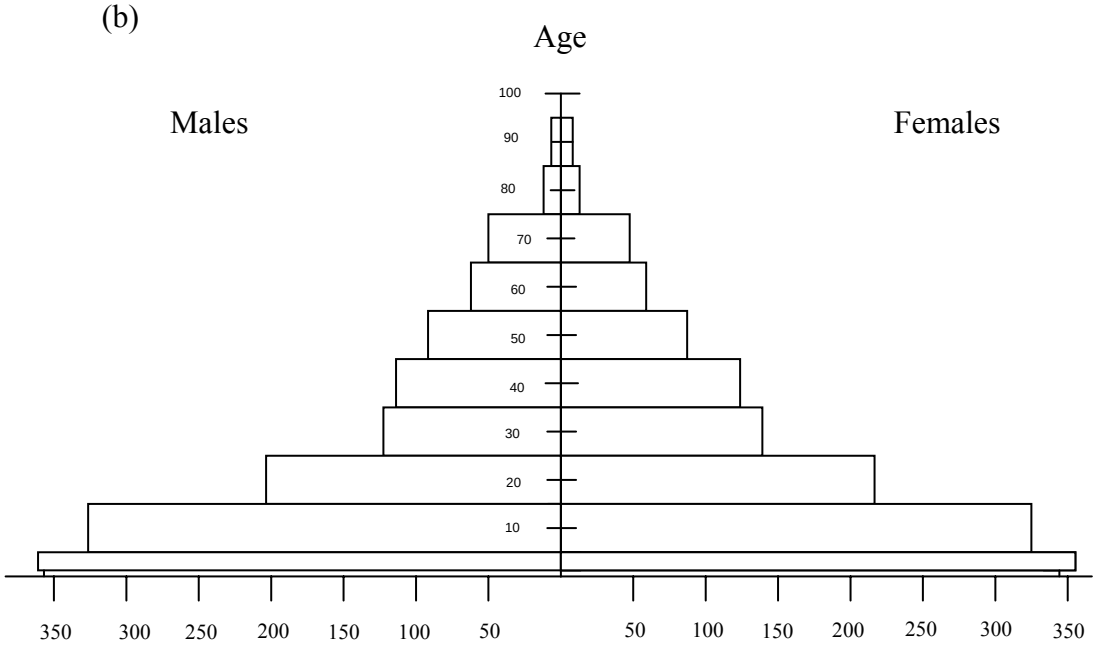
The maternal mortality rate is

$$\frac{\text{number of deaths from puerperal causes}}{\text{total number of live births}}$$

All these rates are usually multiplied by 1000.

Solution continued on next page

(b)



Totals (thousands) in ten-yearly age groups

The 0–1 frequencies are multiplied by 10; those for 1–4 by 10/4; the choice of upper limit for the 85+ group is made so as not to distort the pyramid. [NOTE. The accuracy of representation on the diagram is constrained by the limits of electronic reproduction.]

(c) (i) The sex-age-specific death rate for a country is

$$\frac{\text{number for that sex in age-range of interest during 1 year}}{\text{average number of persons of that sex and age living during the year}} \times 1000$$

This is calculated separately for males and females, using suitable age ranges. "Average" (mid-year) is usually the mean of beginning and end figures for that year.

(ii) The rates for U are generally higher than for D up to 44, and then become lower.

The rates for males are generally higher than for females, in both countries.

Females thus have longer expectation of life than males, and inhabitants of D have longer expectation than U.