

# **THE ROYAL STATISTICAL SOCIETY**

## **2002 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, 2002. Question 1

(i) This layout confounds litters and diets, so the effect of diets cannot properly be estimated. Litters are likely to be a source of systematic variation, which has to be included in a model as blocks. Otherwise the residual term in the model will include this systematic component, contrary to assumptions. Also the proposed design makes litters the experimental unit, not individual mice; so the diets have only 2 replicates, which is inadequate. A randomised block design allocates, at random, one diet to each animal so that all diets are used once in each litter (block). Now each diet has 6 replicates.

$$y_{ij} = \mu + t_i + b_j + e_{ij} \quad i = 1, 2, 3; \quad j = 1, 2, \dots, 6; \quad e_{ij} \sim \text{ind } N(0, \sigma^2).$$

$t_i$  is the effect of diet  $i$ ,  $b_j$  that of litter  $j$ , and randomisation ensures that independence and constant variance can be assumed.

(ii) Digits            1 6 8 2 0 2 3 7 1 4 4 8 6 9 0 0 3 9 .....

Take 1, 2 or 3 as implying Diet 1; 4, 5 or 6 as implying Diet 2; and 7, 8 or 9 as implying Diet 3.

Litter 1:        use 1 → Diet 1 for first animal; 6 → Diet 2 for second; so third animal gets Diet 3.

Litter 2:        use 8 → Diet 3 for first animal; 2 → Diet 1 for second; so third animal gets Diet 2.

Litter 3:        use 0 not at all; 2 → Diet 1 for first animal; 3 implies Diet 1 which has already been allocated; 7 → Diet 3 for second animal; so third animal gets Diet 2.

And so on.

[0 is not used, otherwise allocation of diets is not "at random" since one of them is more likely to be selected than the other two.]

Final arrangement:

Litter	1	2	3	4	5	6
1st animal	$D_1$	$D_3$	$D_1$	$D_1$	$D_2$	$D_2$
2nd animal	$D_2$	$D_1$	$D_3$	$D_2$	$D_3$	$D_3$
3rd animal	$D_3$	$D_2$	$D_2$	$D_3$	$D_1$	$D_1$

Animals in litters can be numbered arbitrarily.

(iii) Residual = observed value – fitted value.

Fitted values are  $\hat{\mu} + \hat{t}_i + \hat{b}_j$ , where  $\hat{\mu}$  is estimated as the overall mean,  $\hat{\mu} + \hat{t}_i$  by each diet mean and  $\hat{\mu} + \hat{b}_j$  by each block mean.

$$\begin{aligned} \text{Hence } \hat{\mu} &= \frac{1}{18}(1566 + m + n); & \hat{t}_1 &= \frac{602 + m}{6} - \hat{\mu}, & \hat{t}_2 &= \frac{490 + n}{6} - \hat{\mu}, & \hat{t}_3 &= \frac{474}{6} - \hat{\mu}; \\ \hat{b}_1 &= \frac{367}{3} - \hat{\mu}, & \hat{b}_2 &= \frac{148 + n}{3} - \hat{\mu}, & \hat{b}_3 &= \frac{240}{3} - \hat{\mu}; & \hat{b}_4 &= \frac{151 + m}{3} - \hat{\mu}; & \hat{b}_5 &= \frac{382}{3} - \hat{\mu}, \\ \hat{b}_6 &= \frac{278}{3} - \hat{\mu}. \end{aligned}$$

$$\text{The residual for } m \text{ is } m - (\hat{\mu} + \hat{t}_1 + \hat{b}_4) = m - \frac{602 + m}{6} - \frac{151 + m}{3} + \hat{\mu},$$

$$\text{i.e. } m - 100.33 - \frac{1}{6}m - 50.33 - \frac{1}{3}m + 87 + \frac{1}{18}m + \frac{1}{18}n = \frac{5}{9}m + \frac{n}{18} - 63.66.$$

$$\text{Similarly for } n, \text{ the residual is } n - (\hat{\mu} + \hat{t}_2 + \hat{b}_2) = n - \frac{490 + n}{6} - \frac{148 + n}{3} + \hat{\mu},$$

$$\text{i.e. } n - 81.67 - \frac{1}{6}n - 49.33 - \frac{1}{3}n + 87 + \frac{1}{18}m + \frac{1}{18}n = \frac{5}{9}n + \frac{m}{18} - 44.$$

These residuals have to be 0 where the observation is missing.  $m = 107.76$  and  $n = 68.42$  satisfy both these requirements.

(iv) Allowing for the two missing observations, the residual degrees of freedom

will be 8. For  $D_1$  vs.  $D_2$ ,  $S.E.(\bar{Y}_1 - \bar{Y}_2) = \sqrt{\frac{2}{5} \times 92.846} = 6.094$ . The 5% point of  $t_8$  is

$$2.306. \quad \text{The means are } \bar{y}_1 = \frac{709.76}{6} = 118.29 \quad \text{and} \quad \bar{y}_2 = \frac{558.42}{6} = 93.07;$$

$\bar{y}_1 - \bar{y}_2 = 25.22$ . Hence an approximate 95% confidence interval for the true difference between means of Diets 1 and 2 is  $25.22 \pm 2.306 \times 6.094$ , i.e.  $25.22 \pm 14.05$  or (11.17, 39.27).

$$\text{For } D_2 \text{ vs. } D_3, S.E.(\bar{Y}_2 - \bar{Y}_3) = \sqrt{\left(\frac{1}{5} + \frac{1}{6}\right)(92.846)} = 5.835.$$

Mean of  $D_3$  is  $\bar{y}_3 = 79.00$ ;  $\bar{y}_2 - \bar{y}_3 = 14.07$ , and the confidence interval is  $14.07 \pm 2.306 \times 5.835$ , i.e.  $14.03 \pm 13.46$  or (0.61, 27.53).

Although these limits are very wide, showing high variability in the data, 0 is not included and so both these pairs are significantly different. Hence  $D_1$  vs  $D_3$  is clearly significant. So all three diets differ.

Graduate Diploma, Applied Statistics, Paper II, 2002. Question 2

- (a) (i) All combinations of all levels of the factors are used as the set of "treatments". Here "factors" are Varieties, Soil types, Moisture level. If interactions among factors may be present, examining one factor at a time will not give valid results for inferring what happens when they are used together. Even when factors do not interact, they have been examined over a wide range of conditions and the results should have more general validity.

$$(ii) \quad VSM = \frac{1}{4r}(v-1)(s-1)(m-1) = \frac{vsm - vs - vm - sm + v + s + m - (1)}{4r}$$

$$= \frac{(182 - 153 - 131 - 113 + 122 + 98 + 96 - 50)}{(4 \times 4)} = \frac{51}{16} = 3.1875.$$

[There are 4 comparisons of (+, -) to be arranged.]

$$(iii) \quad \sum y = 945, \text{ correction term is } \frac{945^2}{32} = 27907.03125.$$

Total SS is therefore 3605.9688.

$$\text{Blocks SS} = \frac{1}{8}(218^2 + 256^2 + 212^2 + 259^2) - \frac{945^2}{32} = 228.5938.$$

Each factorial term has  $SS = 2r \times (\text{effect estimate})^2 = 8(\text{effect}^2)$ .

Hence:

SOURCE	DF	SS	MS	F-value
Greenhouses	3	228.5938	76.198	2.59 not significant
V	1	1667.5313	1667.531	56.70 very highly sig
S	1	675.2813	675.281	22.96 very highly sig
M	1	306.2813	306.281	10.41 highly sig
VS	1	9.0313	9.031	< 1
VM	1	16.5313	16.531	< 1
SM	1	3.7813	3.781	< 1
VSM	1	81.2813	81.281	2.76 not significant
Residual	21	617.6559	29.4122	$= \hat{\sigma}^2$
TOTAL	31	3605.9688		

[NOTE: SS values are given to greater accuracy here than is possible from the information given on the paper.]

Testing each 1 d.f. MS against the residual, we find large main effect differences but no significant interactions. The higher yields were obtained when  $V_2$  was used, grown in  $S_2$ , at high moisture level  $M_2$ .

- (b) (i) Block size is now smaller than the number of treatments used, so the blocks (greenhouses) can each only contain half of the full set. Eight blocks are available. If there is a high-order interaction which is unimportant, it can be made to have the same pattern of  $\pm$  signs as a comparison between 2 blocks. Blocking can thus still be used to take out greenhouse differences without losing required information from the experiment.
- (ii) Confounding VSM, the treatments  $v$ ,  $s$ ,  $m$ ,  $vs$ *m* will be placed in random order in one greenhouse, and (1),  $vs$ ,  $vm$ ,  $sm$  in an adjacent house; the comparison between these two houses is part of the blocks SS, so VSM is confounded with blocks. The same procedure can be used in each of 4 pairs of houses.

Graduate Diploma, Applied Statistics, Paper II, 2002. Question 3

(i) In a non-factorial experiment where block size must be less than the number of treatments to be included, and all pairwise comparisons are to be made with the same precision (variance), a balanced incomplete block scheme can be used if one of suitable size exists. In this case, patients will be assessed by fewer than 10 examiners each.

(ii) For  $v$  treatments to be compared in  $b$  blocks, of  $k$  units each, each treatment being replicated  $r$  times,  $rv = bk = N$ , the total number of units (units in this case being examiner/patient sessions). Any pair of examiners see the same patient  $\lambda$  times, and for balance  $\lambda = \frac{r(k-1)}{v-1}$ .

(iii)  $v = 10$ ,  $k = 2$  or  $3$  or  $5$  and correspondingly  $b = 45$  or  $30$  or  $18$ , so that  $N = 90$  in all cases. For all these,  $r = 9$ .

$\lambda = \frac{9 \times 1}{9} = 1$ , or  $\frac{9 \times 2}{9} = 2$ , or  $\frac{9 \times 4}{9} = 4$ . So all these designs may exist (subject to a construction method being found).

When  $v = 10$  and  $k = 4$ ,  $rv = bk$  implies that  $10r = 4b$ .  $\lambda = \frac{r(k-1)}{v-1} = \frac{3r}{9} = \frac{r}{3}$ . So  $r$  must be a multiple of 3.

The case  $r = 3$  requires  $b = 7\frac{1}{2}$ ;  $r = 9$  requires  $b = 22\frac{1}{2}$ . Clearly these are impossible.

For  $r = 6$ , we have  $b = 15$ , and a design may exist in  $N = 60$  units.

(iv) Scheme A:  $\text{Var}(\bar{Y}_p - \bar{Y}_q) = \frac{2k\sigma_A^2}{\lambda v}$  for any pair  $p, q$  of examiners. This is  $\frac{10\sigma_A^2}{40}$ .

Similarly, for Scheme B the variance is  $\frac{6\sigma_B^2}{20}$  and for Scheme C the variance is  $\frac{4\sigma_C^2}{10}$ .

Since  $\sigma_A^2 = 1.4\sigma_C^2$ , and  $\sigma_B^2 = 1.2\sigma_C^2$ , variances are  $\frac{7}{20}, \frac{9}{25}, \frac{2}{5}$  times the "unit variance"  $\sigma_C^2$ . (Ratio is 0.35 : 0.36 : 0.40 or 0.875 : 0.9 : 1.)

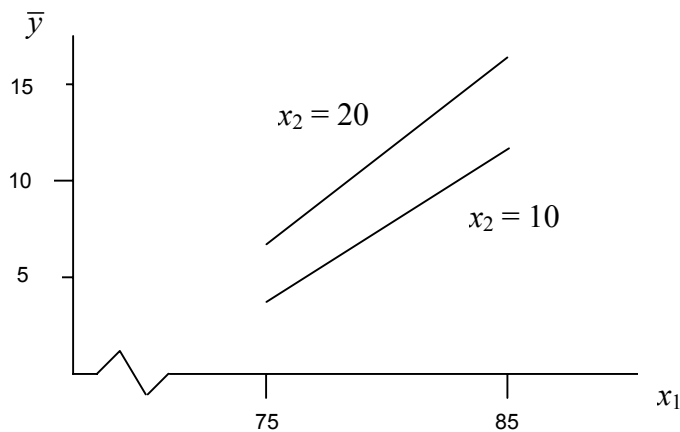
Residual d.f. are  $rv - b - v + 1$ , i.e. 63, 51, 36, all of which are fully adequate. Choose Scheme A as the most precise.

Graduate Diploma, Applied Statistics, Paper II, 2002. Question 4

(i) This is a first-order design based on a  $2^2$  factorial. It is used as an initial experiment in a response surface study to find the optimum settings of two continuous variables  $x_1, x_2$  to produce best response  $y$ . A linear model  $y = a + b_1x_1 + b_2x_2 + e$  is fitted; assuming the fit is good we are not near an optimum value of  $y$  and we proceed up the surface as steeply (rapidly) as possible using the values of  $b_1, b_2$  which are the gradients in the directions  $x_1, x_2$ . This process can be continued until a linear model stops fitting well. A second-order design, allowing quadratic terms to be fitted, is then needed.

(ii) Means of  $y$  are:

		$x_1$	
		75	85
$x_2$	10	3.75	11.75
	20	6.75	16.25



Estimates of effects and interaction:

	(1)	$a$	$b$	$ab$	Contrast $\div 4$ (because of $\pm 1$ coding)
Mean of $y$	3.75	11.75	6.75	16.25	
$X_1$	-1	+1	-1	+1	4.375
$X_2$	-1	-1	+1	+1	1.875
INTERACTION	+1	-1	-1	+1	0.375

(iii)  $\bar{y} = \frac{1}{8} \sum_{i=1}^8 y_i = 9.625$ . Equation is  $y = 9.625 + 4.375x_1 + 1.875x_2$ . If a convenient step in  $x_1$  is 2.5 units, which is 0.5 coded units, the corresponding step in  $x_2$  by steepest ascent will be  $\frac{0.5b_2}{b_1} = \frac{0.5 \times 1.875}{4.375} = 0.214$  coded units =  $0.214 \times 5 = 1.07$  units in  $x_2$ .

	Base	(1 step)	(2)	(3)	(4)	(5)
$X_1$	80	82.5	85	87.5	90	92.5
$X_2$	15	16.1	17.1	18.2	19.3	20.4

Since the first-order model fits well, we need to go some way up to locate an optimum, and the next experiment might be centred on (87.5, 18.2) or (90, 19.3). A similar design may be used again, and if the model does not fit well some more points are added to allow a quadratic surface to be found.

(iv) We require  $23 + 3x_1 + 4x_2 < 27.5$   
and  $9.625 + 4.375x_1 + 1.875x_2 > 12$ .

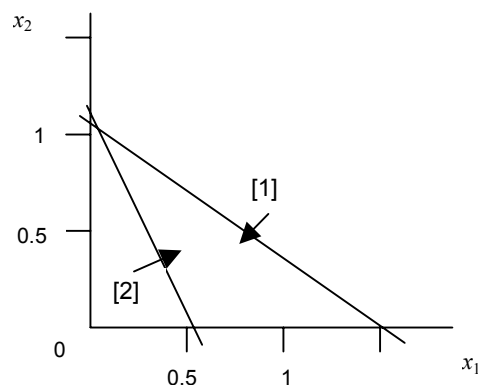
Line  $23 + 3x_1 + 4x_2 = 27.5$  is  $3x_1 + 4x_2 = 4.5$ . [1]

Points (0, 1.125) and (1.5, 0) lie on it. Constraint is to keep below this line.

Line  $9.625 + 4.375x_1 + 1.875x_2 = 12$  is  $4.375x_1 + 1.875x_2 = 2.375$ . [2]

Points (0, 1.27) and (0.54, 0) lie on it. Constraint is to keep above this line.

The lines intersect at  $x_1 = 0.09$ ,  $x_2 = 1.06$  (see graph below). In the "arrowed" region both conditions are satisfied, so a feasible set of conditions can be found.



Final choice of settings depends on balance between cost [1] and life [2]. For example, if cost is the most important the setting (0.54, 0) might be used.



Graduate Diploma, Applied Statistics, Paper II, 2002. Question 5

- (i) (1) A quota sample is a non-random selection, by trained observers, of individuals from a population which has been stratified by characteristics such as age, sex, social class. A specified number from each stratum has to be interviewed but the observer/interviewer is free to select the actual individuals. Their classification is checked by a few questions, and if they belong to an already fully sampled stratum the interview ends. The probability of selection depends on the place and time of interview, e.g. people at work all day cannot be sampled during the day if the interviewer chooses a shopping precinct to operate in.
- (2) In systematic sampling, a list of names/items forming the target population is formed and every  $m$ th name in it is used, systematically from the beginning. This is approximately an equal probability method of selection if the starting point is chosen at random out of the first  $m$  (exactly equal if  $N$  is a multiple of  $m$ ).
- (3) Two-stage cluster sampling first splits a population into groups (clusters), then chooses a random sample from these and a random sample of units from each chosen cluster. Depending on cluster sizes, it can be an approximately equal probability method for individual units; if clusters vary substantially in size, the probabilities of sampling at the two stages can be adjusted so as to make individuals have equal probabilities.

- (ii) What items are to be included in income? Winning the lottery? Bonus from workplace? Payments in kind?

What constitutes a "family"? Is it the same as a "household"? If not, should a multi-household dwelling have only one family sampled from it?

Should we stratify into urban/rural? If so, by what method? Stratification will be required if the characteristics in the two parts of the population are likely to be different. Clustering is less likely to be useful unless the rural part consists of several similar villages and/or the urban part consists of neighbourhoods (e.g. estates) having similar characteristics.

Stratification by other factors such as age can be useful if the information is available.

- (iii) Post-stratification would be a convenient way of grouping the selected members according to age because that question could be asked without causing refusal to answer. The process sorts the replies according to any question that gives useful information. But we cannot control the numbers that will arise in each 'stratum' because we have not designed the sample to do so. Some strata may therefore not be sampled at all, and the results will have precision that can vary substantially between strata. The method can be useful in fairly large samples, as it should then be similar to proportional allocation in stratified random sampling.

(iv) Houses:  $N = 4000$ ,  $p$  expected to be in range  $(0.45, 0.65)$ , the largest variance occurring when  $p = 0.5$ .

$$\text{Var}(p) = \frac{(N-n)}{(N-1)} \frac{p(1-p)}{n} \leq (0.02)^2.$$

$$\text{Hence } \frac{4000-n}{3999} (0.5)^2 \leq (0.02)^2 n,$$

$$\text{or } \frac{4000}{3999} \times 0.25 \leq n \left( 0.0004 + \frac{0.25}{3999} \right) = 0.0004625n.$$

So  $n > 540.66$ , i.e. at least 541.

Cars:  $p$  expected to be in range  $(0.05, 0.10)$ , largest variance for  $p = 0.1$ .

$$\text{Var}(p) = \frac{4000-n}{3999} \frac{(0.1 \times 0.9)}{n} \leq (0.01)^2.$$

$$\text{This leads to } \frac{4000}{3999} \times 0.09 \leq n \left( 0.0001 + \frac{0.09}{3999} \right) = 0.0001225n.$$

So  $n > 734.84$ . Required  $n$  at least 735.

Graduate Diploma, Applied Statistics, Paper II, 2002. Question 6

(i)  $N = 90. \quad n = 15. \quad \bar{M} = \frac{3510}{90} = 39.$

(a)  $\bar{z} = \frac{\sum \bar{z}_i}{n} = \frac{183.64}{15} = 12.243.$

(b)  $\bar{y}_R = \frac{\sum y_i}{\sum x_i} = \frac{7806}{667} = 11.703.$

(c)  $\bar{y} = \frac{\sum y_i}{n\bar{M}} = \frac{7806}{15 \times 39} = 13.344.$

(ii) (a) 
$$\begin{aligned} \text{Var}(\bar{z}) &= \frac{(N-n)}{Nn} \frac{1}{(n-1)} \sum (\bar{z}_i - \bar{z})^2 \\ &= \frac{75}{90 \times 15} \frac{1}{14} \left\{ 2395.8018 - \frac{183.64^2}{15} \right\} = 0.5855. \end{aligned}$$

(b) 
$$\begin{aligned} \text{Var}(\bar{y}_R) &= \frac{N-n}{Nn} \frac{1}{(n-1)\bar{M}^2} \sum (y_i - \bar{y}_R x_i)^2 \\ &= \frac{75}{90 \times 15 \times 14 \times (39)^2} \sum (y_i^2 - 2\bar{y}_R x_i y_i + \bar{y}_R^2 x_i^2). \end{aligned}$$

The bracketed term =  $4759890 - 2 \times 11.703 \times 393716 + (11.703)^2 \times 34883$   
 $= 322156.27.$

Hence  $\text{Var}(\bar{y}_R) = (383292)^{-1} \times 322156.27 = 0.8405.$

(c) 
$$\begin{aligned} \text{Var}(\bar{y}) &= \frac{N-n}{Nn} \frac{1}{(n-1)\bar{M}^2} \sum_{i=1}^{15} (y_i - \bar{M}\bar{y})^2 \\ &= \frac{1}{383292} \left( \sum y_i^2 - 2\bar{M}\bar{y} \sum y_i + 15\bar{M}^2\bar{y}^2 \right). \end{aligned}$$

The bracketed term =  $4759890 - 8124734.6 + 4062492.2 = 697647.6.$

Hence  $\text{Var}(\bar{y}) = 1.820.$

(iii) (a) and (b) both give biased estimators of the population mean; (c) gives an unbiased estimator, but a larger variance due to likely positive correlation of cluster totals.

Graduate Diploma, Applied Statistics, Paper II, 2002. Question 7

(a)  $N_m = 10\ 000, N_f = 5000, N = 15000, n = 75.$

$A$  is a simple random sample;  $B$  is stratified with equal proportions from the two strata "male" and "female".

(i)  $n_m = 50, n_f = 25.$  Let  $\hat{p}_m$  be the estimator of  $p_m$  and  $\hat{p}_f$  that of  $p_f$ ; also  $\hat{p}$  is the estimator of  $p$ , the proportion in the whole population who smoke,  $\hat{p}_{st}$  denoting the estimate found using a stratified sample.

Then 
$$\hat{p}_{st} = \frac{N_m}{N} \hat{p}_m + \frac{N_f}{N} \hat{p}_f = \frac{2}{3} \hat{p}_m + \frac{1}{3} \hat{p}_f.$$

(ii) Design A: 
$$\text{Var}(\hat{p}) = \frac{p(1-p)}{n} \frac{N-n}{N-1}.$$

(a) Clearly  $p = 0.5$ , so 
$$\text{Var}(\hat{p}) = \frac{(0.5)^2}{75} \frac{15000-75}{14999} = 0.003317.$$

(b) 
$$p = \frac{2}{3}(0.5) + \frac{1}{3}(0.1) = 0.3667.$$

So 
$$\text{Var}(\hat{p}) = \frac{0.3667 \times 0.6333}{75} \frac{15000-75}{14999} = 0.003081.$$

Design B:

$$\text{Var}(\hat{p}_{st}) = \frac{1}{N^2} \left\{ \frac{N_m^2 (N_m - n_m)}{N_m - 1} \frac{p_m q_m}{n_m} + \frac{N_f^2 (N_f - n_f)}{N_f - 1} \frac{p_f q_f}{n_f} \right\}$$

$(q_i \equiv 1 - p_i).$

Hence

(a) when  $p_m = p_f = 0.5$ , we find

$$\begin{aligned} & \text{Var}(\hat{p}_{st}) \\ &= \frac{1}{(15000)^2} \left[ \frac{10000^2 (10000 - 50)}{9999} \frac{(0.5)^2}{50} + \frac{5000^2 (5000 - 25)}{4999} \frac{(0.5)^2}{25} \right] \\ &= 0.003317. \end{aligned}$$

This is the same as simple random sampling when the strata proportions are the same.

(b) If  $p_m = 0.5$  and  $p_f = 0.1$ , we have

$$\begin{aligned} & \text{Var}(\hat{p}_{st}) \\ &= \frac{1}{(15000)^2} \left[ \frac{10000^2 (10000 - 50) (0.5)^2}{9999 \cdot 50} + \frac{5000^2 (5000 - 25) 0.1 \times 0.9}{4999 \cdot 25} \right] \\ &= 0.002609. \end{aligned}$$

There is a gain in precision due to stratification.

(iii) Design  $B$  will be better, since if the two population proportions are different this design allows for that. Stratification allows a check of whether  $p_m$  and  $p_f$  are different, giving extra information on the two strata as well as combined information for the whole sample. Also since the strata sizes are different, design  $B$  gives a representative sample of the whole population while  $A$  may accidentally be quite heavily weighted to one stratum.

(b) The questionnaire will only achieve reliable answers if it is administered anonymously – nevertheless it needs to be numbered since we have to chase up initial non-response. This has to be explained, and the number code destroyed after use.

Age, sex, any medical conditions (which may stop people from smoking) are required. Family smoking history (parents, other relatives) and any known deaths related to smoking are also needed. For those who do not smoke now, have they ever smoked, when did they begin/end, reasons for giving up; for those who do smoke, when did they begin, have they tried to give up. For all who have ever smoked, how much. For all who have never smoked, reasons why not. For everyone, have they been influenced by advertising campaigns (for or against); do they "smoke" drugs, or use drugs in any way. Do smokers do it all the time, or only when in presence of other teenagers. Boxes to tick, with carefully worded labels, will be the best way of completing a questionnaire. A sensitive question would be how they get the money to pay for smoking – part-time jobs etc? (This might not be worth asking).

Graduate Diploma, Applied Statistics, Paper II, 2002. Question 8

(i) Crude rates may be used when the age distributions are similar in the different subgroups being compared. Otherwise the rates must be standardised to take into account differences between age structures of subgroups. Age standardisation adjusts the figures to show what the death rates would be if each subgroup had the same pattern of age distribution. Age standardisation adjusts only for age, not for any other factor that may affect death rate.

*Direct* standardisation involves defining a standard population, and applying to it different 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.

(ii) (a) The crude rates are (actual no. of deaths)÷(total population), i.e.

$$\frac{412}{731177} = 56.3 \text{ per } 100000 \text{ for first-born,}$$

$$\frac{740}{442811} = 167.1 \text{ per } 100000 \text{ for 5th or later.}$$

(b) Indirect adjusted rate =  $c_s \frac{\sum r_i}{\sum n_i P_i}$ , where  $c_s$  is the crude rate for

Down's Syndrome in the standard population [ $c_s = \frac{2529}{2825173} = 89.5166$  per 100000],  $P_i$  is the age-specific prevalence rate for age group  $i$  in the standard population; and, in the study population,  $n_i$  and  $r_i$  are the number of live infant births observed and the number of those with Down's Syndrome, in age group  $i$ .

The necessary calculations are shown in the table following (see part (c) for the last two columns of the table).

We have

$$P_1 = \frac{136}{319933} = 42.5089 \text{ per } 100000; \quad n_1 P_1 = 230061 P_1; \quad \text{etc.}$$

Age group $i$	$P_i$	$n_i P_i$	$p_i$	$N_i p_i$
1	42.5089	97.7964	46.5094	148.7989
2	42.5204	140.0830	42.7987	398.5923
3	52.2561	60.0527	52.2102	410.6392
4	87.6627	34.6154	101.2992	494.5780
5	264.0175	37.5116	274.4932	652.9179
6	864.4170	26.3820	819.1349	502.2362

$$\sum r_i = 412, \quad \sum n_i P_i = 396.4411;$$

$$\text{so indirect rate per } 100000 = \frac{(89.5166)(412)}{(396.4411)} = 93.03 \text{ for first born.}$$

(c) The direct adjusted rate =  $\frac{\sum N_i p_i}{\sum N_i}$ , where  $N_i$  is the observed number of live infant births in age group  $i$  of the standard population, and  $p_i$  is the age-specific prevalence rate of Down's Syndrome in the study population in age group  $i$ .

$$\left[ \begin{array}{l} p_1 = 46.5094 \text{ per } 100\,000 \left( = \frac{107}{230061} \right) \text{ and } N_1 p_1 = 319933 p_1; \text{ etc.} \\ \text{See table above, in part (b).} \end{array} \right]$$

$$\sum N_i p_i = 2607.7625, \quad \sum N_i = 2825173; \quad \text{so the direct adjusted rate is } \frac{2607.7625}{2825173} = 92.30 \text{ per } 100000 \text{ for first born.}$$

(d)

	Indirect	Direct	Crude
1st	93.0	92.3	56.3
5th or later	84.8	75.5	167.1

The crude rates indicate a 3-fold increase for (5th or later) compared with first born, but the adjusted rates indicate a slight decrease.

The size of a family is probably less likely to reach 5 if the first-born is found to have Down's Syndrome, which may account for this slight decrease.

The major difference between crude rates and others is probably due to maternal age distribution. Proportionately more mothers of later-born infants are obviously in the older age categories, where the specific rates are higher. After adjusting for differences in the maternal age distributions, the rate of Down's Syndrome in infants born later seems somewhat less, rather than very much more.