THE ROYAL STATISTICAL SOCIETY

GRADUATE DIPLOMA EXAMINATION

NEW MODULAR SCHEME

introduced from the examinations in 2009

MODULE 5

SOLUTIONS FOR SPECIMEN PAPER B

THE QUESTIONS ARE CONTAINED IN A SEPARATE FILE

The time for the examination is 3 hours. The paper contains eight questions, of which candidates are to attempt **five**. Each question carries 20 marks. An indicative mark scheme is shown within the questions, by giving an outline of the marks available for each part-question. The pass mark for the paper as a whole is 50%.

The solutions should not be seen as "model answers". Rather, they have been written out in considerable detail and are intended as learning aids. For this reason, they do not carry mark schemes. Please note that in many cases there are valid alternative methods and that, in 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 the questions and solutions, the Society will not be responsible for any errors or omissions.

The Society will not enter into any correspondence in respect of the questions or solutions.

Note. In accordance with the convention used in all 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 Module 5, Specimen Paper B. Question 1

(i) The original variables, the answers to the questions, are likely to be highly correlated. Principal component analysis (PCA) gives linear combinations of the variables that are uncorrelated. The first PC accounts for the largest amount of variation in the data, the second for the next largest, and so on. If the questions form themselves into relatively distinct clusters then PCs are useful to define subsets, and possibly to suggest ways of combining scores.

PCs are only strictly valid for numeric data, but the data here are nearer to being categorical – at best ordinal. However, PCA is often used for data such as these.

(ii) A cluster analysis could be useful, using correlations (or absolute values of them); perhaps indications of the grouping of questions would be given.

(iii) PCA only works on complete records. If a respondent's answer to one question is missing, that whole set of responses will be omitted. Because PCA is based on analysis of variability in data, missing values cannot easily be imputed. The choice in this case is between analysing a large number of responses on a small number of questions and a small number of responses on a large number of questions. The strategy proposed seems sensible.

(iv) The first three eigenvalues add to 5.14, i.e. 5.14/6 or 85.7% of the total variation, and should be enough.

The first PC (54% of total variation) is an overall score of concern about cost – note that the "direction" of questions 2, 3, 4 is opposite to that of 1, 5, 6. The second PC (23% of total variation) measures the tendency of respondents to answer all questions in the same way, i.e. with similar scores. The third PC (9% of total variation and so relatively much less important) is dominated by question 4, perhaps contrasting its answers with those for question 2, perhaps also taking question 5 into account. The first two PCs therefore give most of the useful, easily understood, information.

(v) The two unsatisfactory features of the data are the large amount of missing information, leading to 9 of the 15 questions being discarded, and the suggestion from the second PC that the respondents do not complete the form validly. Hence these results are not reliable. A fresh start is needed, with reworded questions and boxes to tick as in a survey.

Graduate Diploma Module 5, Specimen Paper B. Question 2

(i) Linear discriminant analysis can be used to produce a classification rule where the groups are known a priori, and data are described by several variables. Linear combinations of these variables $\boldsymbol{x_i}$ can show up relations not obvious from separate, univariate, analyses. Classifications so found can be applied to the new sites.

(ii) Multivariate Normal variance-covariance matrices are required to be equal for each group (but locations will be different). This is not easy to check; although formal tests exist, they are sensitive to non-Normality. Also, relatively small sample sizes do not help. Univariate Normality for each measurement can be checked in the usual ways (e.g. histograms, stem-and-leaf plots, Normal probability plots); univariate Normality is a necessary but not sufficient condition for multivariate Normality.

(ii) The variance-covariance matrices are apparently different, with changes in sign as well as size of individual entries. Normality cannot be checked on the information given.

The means of x_1 and x_4 (and possibly x_5) appear different for the two groups.

(iv) Method 1

After constructing and applying the discriminant function, 14/17 (1) and 12/15 (2) are found to have been correctly classified. This is good, but is likely to be an overestimate of the future success rate (since the same data have been used to construct the function and to "check" it).

Cross-validation may be carried out by, for example, a jack-knife method: calculate the function omitting one observation, and use the function to predict class membership of that item; repeat this for each item in turn and observe the number of correct predictions. [In a large data-set, the discriminant function would be calculated on some of the data and then used to check the success rate of the remainder. Here we do not have enough data for that.] This gave 12/17 (1) and 9/15 (2) correct.

Method 2

Note that x_4 was identified in (iii) as a useful variate. This method correctly classifies 12/17 (1) and 12/15 (2), and the numbers on cross-validation are the same. This seems the better method.

With these sample sizes, using 5 variables (Method 1) may be over-fitting. The univariate (as it has turned out) Method (2) is more successful.

Graduate Diploma Module 5, Specimen Paper B. Question 3

If $f(t)$ is the probability density function and $F(t)$ the cumulative distribution function for the lifetime, then the hazard function $h(t)$ is defined by $h(t) = f(t)/(1 - F(t))$.

The hazard function in this case is

$$h(t) = h_0(t)\exp(\beta_1 x_1 + \beta_2 x_2 + \beta_{12} x_1 x_2)$$

where $h_0(t)$ is the baseline hazard function.

(i) The log likelihood (log L) is not given for the null model, but the coefficient for x_1 in model A (which contains x_1 only) is large relative to its standard error; further, the hazard ratio is high. These suggest that x_1 is important.

The difference in $-2\log L$ between models B and C is small and certainly not statistically significant. Thus it seems likely that there is not an interaction effect (the interaction term is the only difference between the two models).

The difference in $-2\log L$ between models A and B is $-2(71.518 - 75.640) = 8.244$. This is significant as an observation from χ^2_1. So we conclude that the best model is model B.

(ii) Hazard ratio = exp(coefficient).

For x_2 in model B, a 95% confidence interval for the coefficient is given by $1.172 \pm (1.96 \times 0.429)$, i.e. it is (0.331, 2.013).

The corresponding 95% confidence interval for the hazard ratio is from exp(0.331) to exp(2.013), i.e. from 1.393 to 7.484.

(iii) The fault reduces the expected lifetime. Given two items, both with the same level of the chemical (x_1), the one with the fault is about 3 [exp(1.172)] times as likely to fail at any time.

(iv) $(1.528 \times 1.4) + 1.172 = 3.311$ $(1.528 \times 2.9) + 0 = 4.431$

So the second is more likely to fail first.

(v) If the proportional hazards assumption is not valid then the model is deficient. In particular the interpretation of the hazard ratios is invalid, and the answers to parts (iii) and (iv) may be inaccurate. The consequences depend to some extent on the type and seriousness of any departures from the assumptions.

Graduate Diploma Module 5, Specimen Paper B. Question 4

(i) The survival time of an individual is *censored* when the end-point of interest (death in this example) has not been observed, either because the trial is terminated before the end-point took place or because the individual has been lost to the trial for some reason (e.g. does not respond to follow-up). The phrase *right-censoring* refers to the censoring occurring after (i.e. to the right of, in natural time order) the last known survival time.

(ii) The Kaplan-Meier survival curve is constructed as follows. The word "death" is used in this description generically; here it does in fact refer to death, but in other examples it might be healing, general recovery, etc.

We seek the estimated cumulative survival function $\hat{S}(t)$.

The Kaplan-Meier method requires the ordered death times $t_{(1)}$, $t_{(2)}$, …, $t_{(r)}$ to be considered. For j = 1, 2, …, r, let $n_{(j)}$ be the number of individuals alive just before time $t_{(j)}$, and let $d_{(j)}$ be the number of deaths at $t_{(j)}$.

An estimate of the probability of survival from $t_{(j)}$ to $t_{(j+1)}$ is $\dfrac{n_{(j)} - d_{(j)}}{n_{(j)}}$.

Thus (assuming independence) the probability of surviving through all the intervals up to $t_{(k+1)}$ is estimated by

$$\hat{S}(t) = \prod_{j=1}^{k} \left(\frac{n_{(j)} - d_{(j)}}{n_{(j)}} \right),$$

and this is the Kaplan-Meier estimate.

If the largest survival time [$t_{(r)}$] is censored, the method above is used to give estimates up to and including the next largest, the value for which is then assumed to apply for all times onward. If the largest survival time is not censored, the estimate drops to zero at that point.

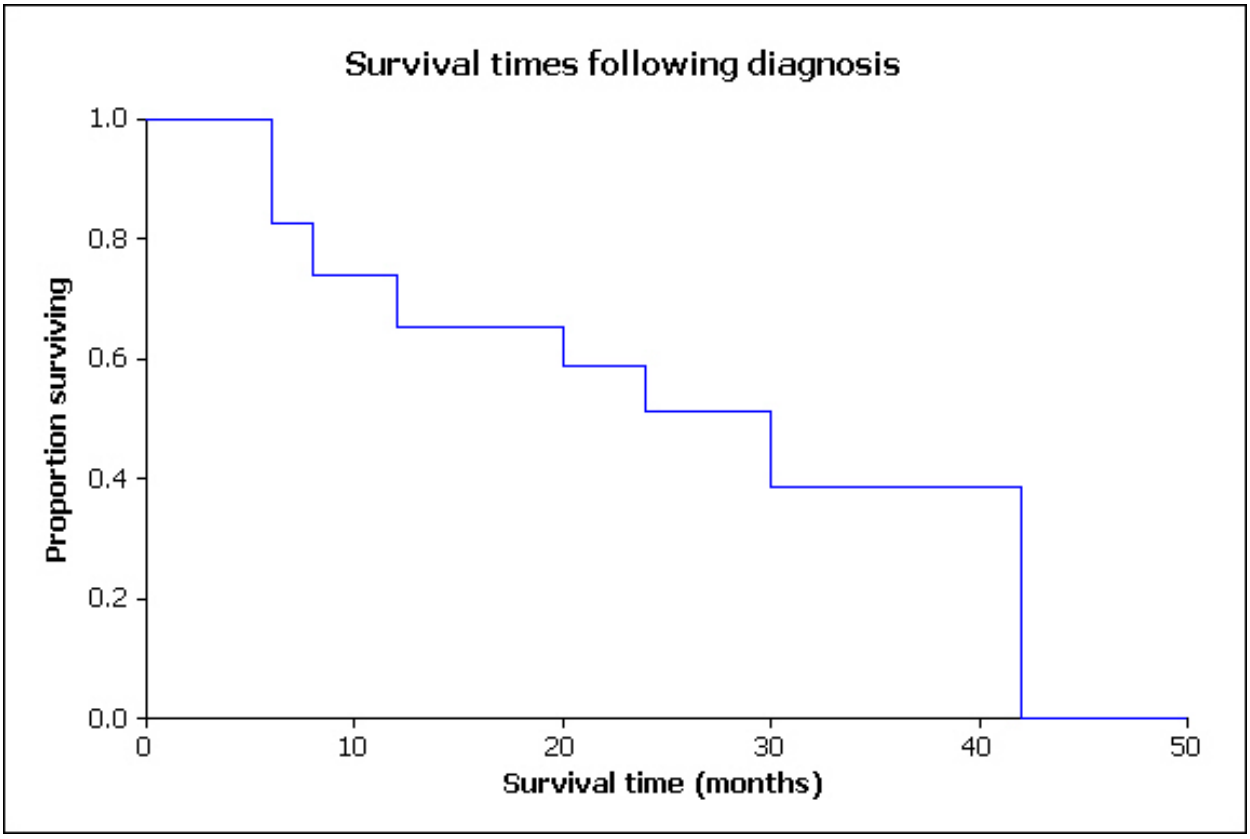
In the present example, $t_{(r)}$ is not censored.

The calculation is shown in detail on the next page.

Solution continued on next page

There are 24 patients. The calculation is shown in detail in the table below. Some of the detail might be omitted in practice, and is often not shown in computer output. Rows for the censored observations have been omitted, but care must be taken to ensure that $n_{(j)}$ is always correct. Users of this solution should carefully verify the values in the table by reference to the data in the question.

Time $t_{(j)}$	$n_{(j)}$ as defined in text above [i.e. number remaining just before time $t_{(j)}$]	$d_{(j)}$ as defined in text above [i.e. number of events at time $t_{(j)}$]	$\frac{n_{(j)} - d_{(j)}}{n_{(j)}}$	Cumulative survival estimate $\hat{S}(t)$ at each $t_{(j)}$
6	23	4	19/23	0.8261
8	19	2	17/19	0.7391
12	17	2	15/17	0.6522
20	10	1	9/10	0.5870
24	8	1	7/8	0.5136
30	4	1	3/4	0.3852
42	1	1	0	0



Solution continued on next page

Greenwood's formula for the standard error for the Kaplan-Meier estimate at 12 months follow-up (corresponding to the third row in the calculation above) is

$$SE = \hat{S}(12)\left(\sum_{j=1}^{3}\frac{d_j}{n_j\left(n_j - d_j\right)}\right)^{\frac{1}{2}}$$

$$= 0.6522\left(\frac{4}{23\times 19} + \frac{2}{19\times 17} + \frac{2}{17\times 15}\right)^{\frac{1}{2}}$$

$$= 0.6522\left(0.009153 + 0.006192 + 0.007843\right)^{1/2}$$

$$= 0.6522\sqrt{0.023188} = 0.0993.$$

(iii) The interval is $0.6522 \pm \left(1.96 \times 0.0993\right) = 0.6522 \pm 0.1946$, i.e. (0.46, 0.85).

(iv) From the graph, the median survival time is 30 months.

(v) A log rank test could be used for this purpose.

Graduate Diploma Module 5, Specimen Paper B. Question 5

(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].

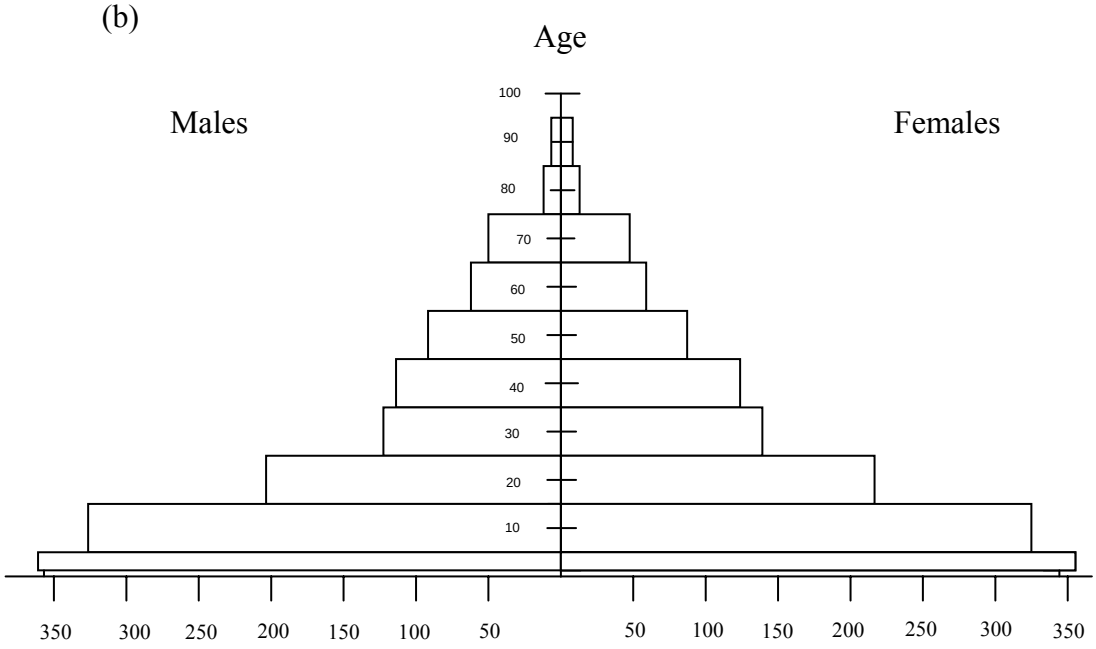
The maternal mortality rate is

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

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 of life than those of *U*.

Graduate Diploma Module 5, Specimen Paper B. Question 6

(i) This is because a case-control study is retrospective, whereas relative risk is measured in a prospective study.

A retrospective study takes affected persons and explores in detail the history of events that may have led to the condition. For example, it may thereby enquire into whether most of those affected by a cholera epidemic have consumed water from the same source. A retrospective study relies on having fairly complete and reliable data on a variety of topics.

A prospective study begins with unaffected persons (e.g. those without lung cancer), notes various characteristics (e.g. smoking habits, occupation, place of residence) and studies future development of the condition in relation to those characteristics. Thus it may enquire into whether the condition develops more frequently in some groups than in others.

(ii) $$\text{Odds} = \frac{P(\text{event happens})}{1 - P(\text{event happens})}.$$

The odds ratio is the ratio of odds of disease in the exposed group of patients (i.e. here the smokers) to that in the unexposed group, i.e. here

$$\text{odds ratio} = \frac{\dfrac{P(\text{disease, smoker})}{1 - P(\text{disease, smoker})}}{\dfrac{P(\text{disease, non-smoker})}{1 - P(\text{disease, non-smoker})}}.$$

(iii) The combined data are as follows.

	Smokers	*Non-smokers*	*Total*
Cases	89	394	483
Controls	13	434	447
Total	102	828	930

Using the relative frequencies from this table, the odds ratio may be calculated as

$$\frac{89 \times 434}{13 \times 394} = 7.54$$

which is substantially greater than 1 and indicates greater prevalence of cancer among smokers.

Solution continued on next page

(iv) The Mantel-Haenszel method is a simple way of adjusting for another factor, in this case sex. [Note. Other methods for doing this are used in some computer programs.]

Representing each table by $\begin{matrix} a & c \\ b & d \end{matrix}$ with $a + b + c + d = n$, and keeping the two sexes separate as in the question, the Mantel-Haenszel estimate of the odds ratio is

$$\frac{\sum a_i d_i / n_i}{\sum b_i c_i / n_i} \qquad \text{where } i = 1, 2 \text{ for males, females.}$$

This gives

$$\frac{\frac{58\times 271}{580} + \frac{31\times 163}{350}}{\frac{6\times 245}{580} + \frac{7\times 149}{350}} = \frac{41.537}{5.514} = 7.53 \ .$$

This is virtually the same as for the pooled data (7.54). This often turns out to happen when both subsets of the data are large and of the same order of size; also, in this case, we might suppose that sex is not in fact an important factor.

To obtain a 95% confidence interval for the odds ratio, we work via logarithms and first use the pooled data to obtain the standard error of the log odds ratio using the formula

$$\text{Var(log of odds ratio)} = \frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d}$$

(= 0.093001 here) so that the standard error is $\sqrt{0.093001} = 0.305$. The log of the Mantel-Haenszel estimate of the odds ratio is log(41.537/5.514) = 2.0193, so the 95% confidence interval for the logarithm is given by

$$2.0193 \pm (1.96 \times 0.305) = 2.0193 \pm 0.5978 \,,$$

i.e. it is (1.421(5), 2.617), and thus the interval for the odds ratio itself is (4.14, 13.69).

(v) The confidence interval does not contain 1.00, so we may reject the null hypothesis that smoking status and occurrence of lung cancer are unrelated. There is definite evidence of an association. As mentioned above, the odds ratio strongly indicates greater prevalence of cancer among smokers than among non-smokers. It does not appear that sex is an important factor in this.

Graduate Diploma Module 5, Specimen Paper B. Question 7

Part (i)

(a) $\mathrm{Cov}\left(\hat{R},\bar{x}\right) = E\left(\hat{R}\bar{x}\right) - E\left(\hat{R}\right)E\left(\bar{x}\right) = E\left(\bar{y}\right) - E\left(\hat{R}\right)E\left(\bar{x}\right) = \bar{Y} - \bar{X}E\left(\hat{R}\right).$

This gives $E\left(\hat{R}\right) = -\frac{1}{\bar{X}}\mathrm{Cov}\left(\hat{R},\bar{x}\right) + \frac{\bar{Y}}{\bar{X}}$, i.e. $E\left(\hat{R}\right) - R = -\frac{\mathrm{Cov}\left(\hat{R},\bar{x}\right)}{\bar{X}}$.

(b) $f = \frac{n}{N}$ and $\hat{R} = \frac{\bar{y}}{\bar{x}}$, so that $\hat{R}\bar{x} = \bar{y}$ or $\bar{y} - \hat{R}\bar{x} = 0$.

Hence the estimator of $\mathrm{Var}\left(\hat{R}\right)$ given in the question is

$$\frac{1-f}{n\bar{x}^2}\cdot\frac{1}{n-1}\sum\left\{\left(y_i-\bar{y}\right)-\hat{R}\left(x_i-\bar{x}\right)\right\}^2$$

$$=\frac{1-f}{n\bar{x}^2}\cdot\frac{1}{n-1}\left\{\sum\left(y_i-\bar{y}\right)^2-2\hat{R}\sum\left(y_i-\bar{y}\right)\left(x_i-\bar{x}\right)+\hat{R}^2\sum\left(x_i-\bar{x}\right)^2\right\}$$

$$=\frac{1-f}{n\bar{x}^2}\left(s_Y{}^2-2\hat{R}\hat{\rho}s_X s_Y+\hat{R}^2 s_X{}^2\right)$$

in which $s_Y{}^2$, $s_X{}^2$ are the estimated variances of Y and X, and $\hat{\rho}$ is the estimated correlation coefficient for X and Y.

Part (ii)

The ratio method works well when Y is proportional to X, with the relation passing through the origin. It will not be better than a simple random sample when ρ is less than 0 or when the relation does not pass through the origin (in which case a regression estimator is required instead).

See next page for solution to (iii)

Part (iii)

(a) The sugar content of an individual fruit should be roughly proportional to its weight, in fruit from the same source and batch.

(b) Since we are not told *N*, the total number of oranges, a ratio estimator is used rather than regression. Counting the whole batch would take a very long time for what might be a very small improvement in precision.

$$\sum x = 1975, \quad \sum y = 110.9. \qquad X_T = 820.$$

$$\hat{R} = \frac{\overline{y}}{\overline{x}} = \frac{\Sigma y}{\Sigma x} = 0.05615. \quad \therefore \hat{Y}_T = \hat{R} X_T = 46.045 \text{ (kg)}.$$

We have $\text{Var}\left(\hat{Y}_T\right) = X_T{}^2 \text{Var}\left(\hat{R}\right)$.

Also, on neglecting *f* which will be very small (as *n* is only 10), we have that the value of the estimator of $\text{Var}\left(\hat{R}\right)$ is

$$\frac{1}{10\overline{x}^2} \cdot \frac{1}{9} \sum \left(y_i{}^2 - 2\hat{R} x_i y_i + \hat{R}^2 x_i{}^2 \right)$$

$$= \frac{1}{90} \cdot \frac{1}{(197.5)^2} \left(1268.69 \; - \; 2 \times 0.05615 \times 22194.8 \; + \; (0.05615)^2 \times 392389 \right)$$

$$= \frac{1}{351056.25} (1268.69 - 2492.476 + 1237.133) \; = \; \frac{13.34687}{351056.25}$$

which on multiplying by $(820)^2$ gives that the value of the estimator of $\text{Var}\left(\hat{T}\right)$ is 25.5641, i.e. the standard error is 5.056.

(c) The half-width of the interval, $ts/\sqrt{n}$, is to be less than 2. Thus $s/\sqrt{n} < 1$ and 25 oranges will achieve this approximately.

Graduate Diploma Module 5, Specimen Paper B. Question 8

[solution continues on next page]

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

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

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

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

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

so the estimated standard error is 9171.9.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Finally,

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

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

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

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