THE ROYAL STATISTICAL SOCIETY

2010 EXAMINATIONS – SOLUTIONS

GRADUATE DIPLOMA

MODULE 5

TOPICS IN APPLIED STATISTICS

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, Module 5, 2010. Question 1

(i) Principal component analysis aims to reduce the dimensionality of the data, so that the overall structure may more easily be described, by summarising the nine correlated variables using a smaller number of orthogonal linear combinations of them.

(ii) The principal components are the eigenvectors of the correlation matrix, obtained in the usual way by solving

$$\boldsymbol{\Sigma}\mathbf{x}_r = \lambda_r \mathbf{x}_r$$

where $\boldsymbol{\Sigma}$ is the matrix, $\mathbf{x}_r$ is the rth principal component and λ_r is the rth eigenvalue.

(iii) Clearly there is collinearity here as these nine variables are not independent since, for each country, the nine percentages add to 100. Thus any one variable can be determined from the other eight. However, principal component analysis remains valid. The 9th eigenvalue will be zero, or very nearly so.

Part (iv)

(a) The largest correlations are negative: those between agriculture (agr) and each of social and personal services (sps), service industries (si) and manufacturing (man). There are some other moderate negative correlations: between agriculture and construction (con), between mining (min) and finance (fin) and service industries, and between agriculture and power supply (ps). The only fairly large positive correlations are those between social and personal services and service industries and between social and personal services and transport and communications (tc), though some others are not very much smaller. Most of the other correlations are positive, but small or very small.

(b) At least two principal components should be chosen as there are two high eigenvalues (3.487 and 2.130); the two corresponding principal components account for 62.4% of the variation in the data.

A rule that is often useful, albeit not always reliable, is to be guided by the number of eigenvalues that are greater than 1 (such components take "more than their share" of the total variability). This would indicate use of the third eigenvalue. However, it is only very slightly greater than 1, and the fourth eigenvalue is only very slightly less than 1. So, if the third eigenvalue is to be used, it would be sensible to use the fourth also, thus giving four principal components. The fifth eigenvalue is much smaller, so there seems no case for using this too.

Thus we arrive at using either two or four principal components. There is no informal way of choosing between these. Choice may well be influenced by the intended use of the analysis.

Solution continued on next page

(c) The first principal component is indeed usually an average of the variables – but a *weighted* average. The coefficients in this principal component can be positive or negative or zero. In this case it seems to be a comparison between agriculture (agr) and a weighted average of the others except mining (min) for which the coefficient is (very nearly) zero and probably also excluding finance (fin). This is entirely possible for the first principal component.

(d) An interpretation of the first principal component is given in part (c) just above. It gives a comparison between agriculture and the rest of the industrial and service economy apart from mining and finance.

The second principal component has negative loadings for "si", "fin" and "sps"; near-zero loadings for "agr" and "con"; and positive loadings for the rest. For the non-agricultural economy, and also excluding construction, it gives a comparison between, broadly speaking, the service and industrial sectors.

(e) All the variables are percentages and thus measured in the same units. Thus use of the (variance-)covariance matrix for principal component analysis might be appropriate.

However, the variables are noticeably very different in their behaviours – note that the standard deviations vary a great deal, from about 0.4 for power supply to about 15.5 for agriculture. This means that the results are likely to be different from those based on the correlation matrix, and be more dominated by agriculture.

Graduate Diploma, Module 5, 2010. Question 2

[Solution continues on next page]

Part (i)

(a) Cluster analysis groups data items into "similar" groups based on a measure of multivariate distance or similarity between items.

Where a population can be split into pre-defined groups, and multivariate measurements are available on a set of data units from the population, linear discriminant analysis produces a linear function of the variables that acts as a classification rule to predict group membership. The formal assumptions for this are that the observations should follow a multivariate Normal distribution, and that the variance-covariance matrices are equal for each group (but the locations will be different).

The formal assumptions for linear discriminant analysis are not easy to check. It is fairly common practice simply to assume that they hold, at least as good approximations, unless there is distinct evidence to the contrary. Choice between cluster analysis and linear discriminant analysis is more likely to be made on the basis of whether the biologist has pre-defined groups for the species. If so, then linear discriminant analysis is likely to be better; if not, cluster analysis should be used.

(b) All of the statements are true. There is no formal hypothesis in cluster analysis, and no testing. The choice of the number of clusters is often highly subjective, depending on the analyst's view of "closeness". The measure of similarity or distance, and the method of agglomeration, all have to be chosen. The apparent cluster structure may differ according to the choices made by the analyst.

Part (ii)

(a) Observations (1, 9) are closest, with dissimilarity 2.71. So at stage one of the process, we take (1, 9) as a cluster. The next closest are (2, 8), with dissimilarity 2.98. So these also form a cluster, and now we have two clusters.

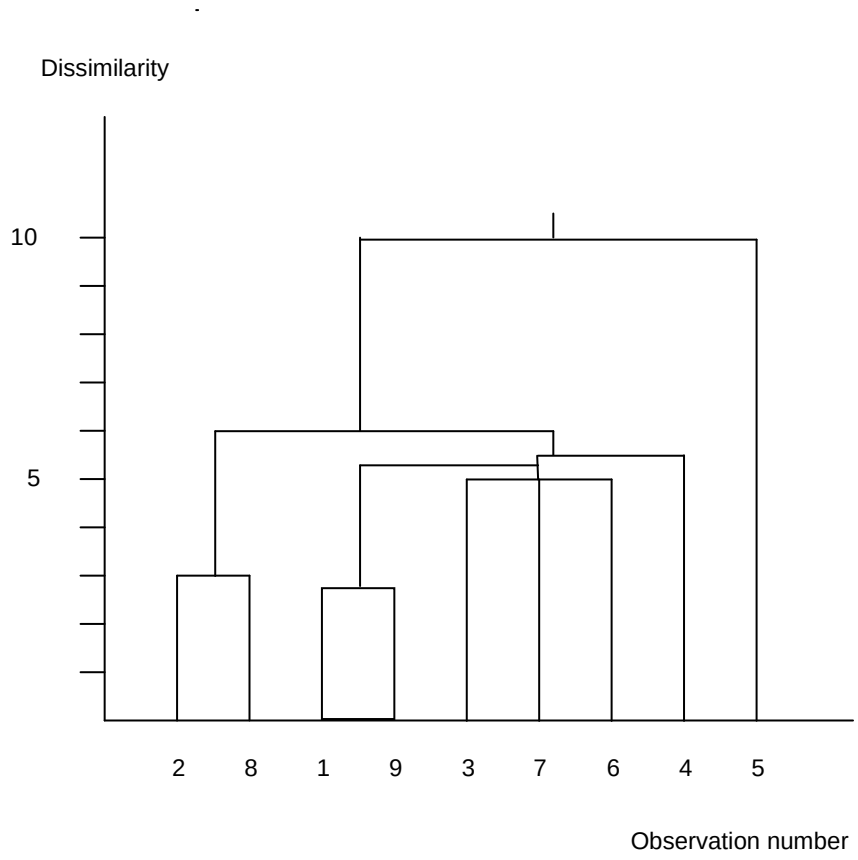
(3, 7) and (3, 6) have almost equal dissimilarities (5.08 and 5.10), so we take (3, 7, 6) as forming the next cluster.

(7, 9) have dissimilarity 5.32, so the next stage is for the clusters (1, 9) and (3, 7, 6) to join. (4, 7) have only slightly greater dissimilarity (5.41), so 4 joins this new cluster almost immediately.

(2, 3) have dissimilarity 5.96, so the clusters (2, 8) and (1, 9, 3, 7, 6, 4) join at this level of dissimilarity.

Observation 5 has been remote from all the clusters so far, but finally joins at dissimilarity 10.05 (the value for (5, 6)).

The dendrogram is shown on the next page.



(b) It is difficult to identify a clear cluster structure here. Evidently observation 5 does not cluster with the others. Perhaps (1, 9) and (2, 8) should be regarded as clusters, but the remaining observations all enter at about the same level of dissimilarity and so do not really join clusters. Overall, it is difficult to tell.

Graduate Diploma, Module 5, 2010. Question 3

Part (i)

(a) The survivor function is $S(t) = P(T > t) = 1 - F(t) = \int_t^\infty f(x)\,dx$, where F represents the cdf and f the pdf. It is the probability that an item survives until at least time t.

(b) $$S(t) = \int_t^\infty \frac{\beta}{\theta^\beta} x^{\beta-1} e^{-(x/\theta)^\beta}\,dx = \left[-e^{-(x/\theta)^\beta}\right]_t^\infty = e^{-(t/\theta)^\beta}.$$

(c) We have $\log S(t) = -(t/\theta)^\beta$ and so $\log\left(-\log S(t)\right) = \beta \log t - \beta \log \theta$. So a plot of $\log(-\log S(t))$ against $\log t$ has slope β and intercept $-\beta \log\theta$.

Thus estimates of θ and β can be found from this plot.

Part (ii)

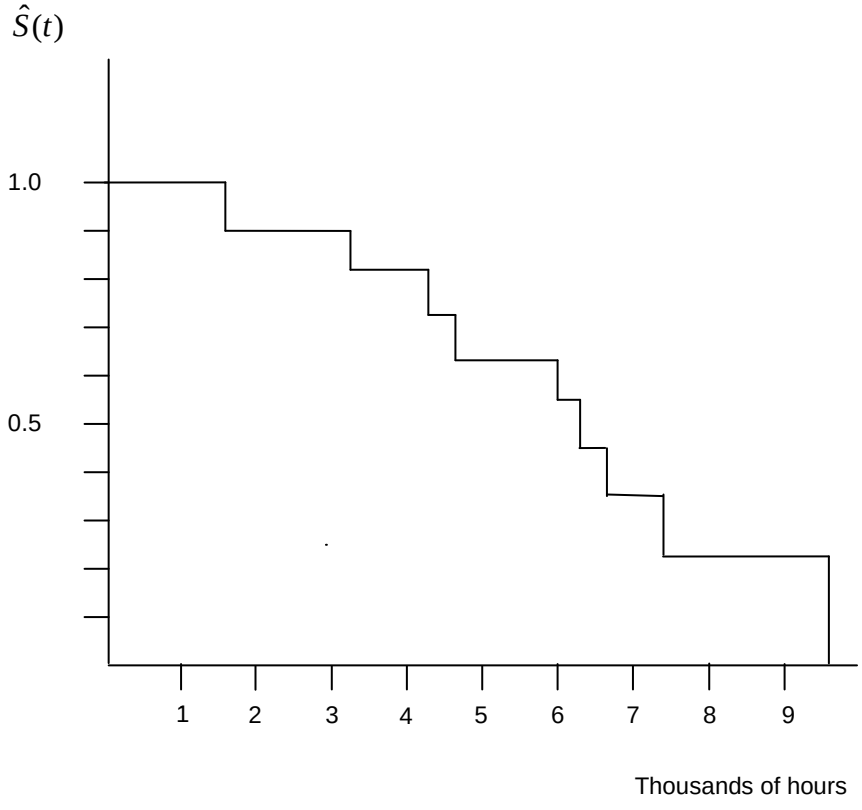
(a) Survivor data are often (very) skewed, so the mean is not a very useful measure of "average". Further, in the present case some of the data items are censored, so the sample mean cannot be directly calculated (though an estimate of the population mean could be found if the underlying distribution were known). In any case, the median would be preferred because of the likely skewness. The median for these data is the sixth ordered observation, i.e. 6.148 (thousand hours).

(b) [Note. The method of calculation of the Kaplan-Meier survivor function is set out in detail in the solutions to question 4 of each of Graduate Diploma Specimen Paper A, Graduate Diploma Specimen Paper B and the 2009 paper, all of which may be downloaded from the "examinations" section of the Society's website.]

The calculation is shown in the table below. Rows for the censored observations, marked with an asterisk, have been included.

Solution continued on next page

Time $t_{(j)}$	$n_{(j)}$ [number remaining in service just before time $t_{(j)}$]	$d_{(j)}$ [number of failures at time $t_{(j)}$]	$\frac{n_{(j)} - d_{(j)}}{n_{(j)}}$	Estimate $\hat{S}(t)$ at each $t_{(j)}$
1.537	11	1	10/11	0.909
3.169	10	1	9/10	0.818
4.219	9	1	8/9	0.727
4.700	8	1	7/8	0.636
5.993	7	1	6/7	0.545
6.148	6	1	5/6	0.455
6.369 *	–	–	–	–
6.683	4	1	3/4	0.341
7.432	3	1	2/3	0.227
9.400 *	–	–	–	–
9.550	1	1	0	0



Solution continued on next page

(c) and (d)

Excluding the last observation and the two censored observations, we have the following (subject to possible slight rounding error in the last column).

t	$\hat{S}(t)$	$\log t$	$\log\left(-\log \hat{S}(t)\right)$
1.537	0.909	0.430	–2.35
3.169	0.818	1.153	–1.60
4.219	0.727	1.440	–1.14
4.700	0.636	1.548	–0.79
5.993	0.545	1.791	–0.50
6.148	0.455	1.816	–0.24
6.683	0.341	1.900	0.07
7.432	0.237	2.006	0.36

A graph of $\log(-\log S(t))$ against $\log t$ [**not shown** in this published solution] indicates that the points lie fairly near a straight line but with distinct evidence of curvature. This implies that the Weibull distribution may not model these data very well – though it must be noted that there are only a few data points.

The slope of the line is about 1.75 and the intercept is about –3.35. So the parameter estimates are given by

$$\hat{\beta} = 1.75$$

and

$$-\hat{\beta}\log\hat{\theta} = -3.35 \quad \text{so that} \quad \hat{\theta} = e^{3.35/1.75} = 6.78\,.$$

Graduate Diploma, Module 5, 2010. Question 4

Part (i)

Let T be the random variable giving time to failure, with pdf $f(t)$ and cdf $F(t)$, and with survivor function $S(t)$ defined by $S(t) = P(T > t) = 1 - F(t)$.

The hazard function $h(t)$ is given as follows.

The probability that a unit fails in the short time interval $(t, t + \delta t)$ is

$$P(t < T \le t + \delta t) \approx f(t)\,\delta t\,.$$

Thus, for the corresponding probability conditional on not having failed by time t, we have

$$P(t < T \le t + \delta t \mid T > t) \approx \frac{f(t)\,\delta t}{S(t)}.$$

This may be described as the probability of imminent failure at time t, and the function $h(t) = \dfrac{f(t)}{S(t)}$ is the *hazard function* (the "failure rate function"). Thus $h(t)$ can be formally defined by

$$h(t) = \lim_{\delta t \to 0} \frac{P(t < T \le t + \delta t \mid T > t)}{\delta t}$$

and we have

$$h(t) = \frac{f(t)}{S(t)} = \frac{f(t)}{1 - F(t)} = -\frac{S'(t)}{S(t)} = -\frac{d}{dt}\log S(t).$$

The Cox proportional hazards model is that the hazard function is

$$h(t) = h_0(t)\exp(\beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_n X_n),$$

where the X are the explanatory variables, the β are the coefficients and $h_0(t)$ is the baseline hazard function (all $X = 0$). The key point is that this implies that the hazard function is proportional to the baseline hazard function at all time points. This may be checked for any variable in a model either graphically or by statistical tests.

In the case in the question, which has two covariates *X1* and *X2* and a factor *F*, the model becomes

$$h(t) = h_0(t)\exp(\beta_1 X_1 + \beta_2 X_2 + \beta_3 F).$$

Solution continued on next page

Part (ii)

(a) The estimates and intervals are for the hazard ratios e^{β}, where each β is the coefficient as in part (i). The null value of β is 0, but that for e^{β} is 1. All of the 95% intervals do contain 1.

(b) The hazard functions for the three components are as follows.

$$h_1(t) = h_0(t)\left(1.59310\right)^1\left(0.99581\right)^{1000}\left(1.01292\right)^{100}$$

$$h_2(t) = h_0(t)\left(1.59310\right)^1\left(0.99581\right)^{2000}\left(1.01292\right)^{100}$$

$$h_3(t) = h_0(t)\left(1.59310\right)^0\left(0.99581\right)^{2000}\left(1.01292\right)^{100}$$

These give $\dfrac{h_2(t)}{h_1(t)} = \left(0.99581\right)^{1000} = 0.015$ and $\dfrac{h_2(t)}{h_3(t)} = 1.593(1)$.

For the factor *F*, we consider components 2 and 3 as these have the same values for *X1* and *X2*. Component 2 has *F* = 1 and component 3 has *F*= 0. Thus the hazard for *F* = 1 is about 1.59 times the hazard for *F* = 0.

For covariate *X1*, we consider components 1 and 2 as these have the same values for *F* and *X2*. Component 1 has *X1* = 1000 and component 2 has *X1* = 2000. Thus the hazard for *X1* = 2000 is about 0.015 times the hazard for *X1* = 1000.

(c) If the assumption is not valid for *F*, the model is not appropriate and the results will be incorrect, the seriousness of this depending on how bad the model is and therefore on how poor the assumption is for *F*.

(d) If the assumption is valid, the curves in the given plot should be parallel. There are only 18 data points here, so care is needed to avoid over-interpretation. Apart from the higher times, when the two curves seem to join together, parallelism might be a reasonable hypothesis. Further, the results for *X1* and *X2* from the stratified analysis are similar to those using *F* as a predictor variable. So there does not seem to be any convincing evidence that the proportional hazards assumption is a serious problem here.

(e) It is true that the sample size is small. There are 3 predictor variables and only 18 data points. The lack of statistical significance might be real or simply a consequence of a study with low power.

The correlation of *X1* and *X2* has not been taken into account in the analysis. A model with one but not both of these variables might be as good as the present one. The suggestion from the output would be to remove *X2*, but knowledge of what the variables actually represent should be considered if possible. More data points would of course be useful.

Graduate Diploma, Module 5, 2010. Question 5

(i) Only 16% of the babies in the sample had shoulder dystocia, i.e. suffered from the risk under investigation. The cut-off probability 0.5 (50%) used for the binary predictions is well above this. It has merely happened that none of them has given a value that exceeds this. This does not prevent the actual predicted probabilities on the two models from being different.

(ii) The sensitivity of a diagnostic test for a risk is the proportion of risk-positives that are correctly identified as such:

$$\frac{\text{number who are risk-positive and give positive outcome in the test}}{\text{number who are risk-positive}}$$

(Another way of expressing this is the number of true positives divided by (the number of true positives + the number of false negatives)).

The specificity is the proportion of risk-negatives that are correctly identified as such:

$$\frac{\text{number who are risk-negative and give negative outcome in the test}}{\text{number who are risk-negative}}.$$

(Another way of expressing this is the number of true negatives divided by (the number of true negatives + the number of false positives)).

Alternatively, in probability terms:

Sensitivity is P(predicted value positive | actual value positive).

Specificity is P(predicted value negative | actual value negative).

The overall accuracy can be high if there is either a high or a low incidence of positive cases, as has happened in this example. Sensitivity and specificity refer to the ability to correctly distinguish both positive and negative results individually.

Solution continued on next page

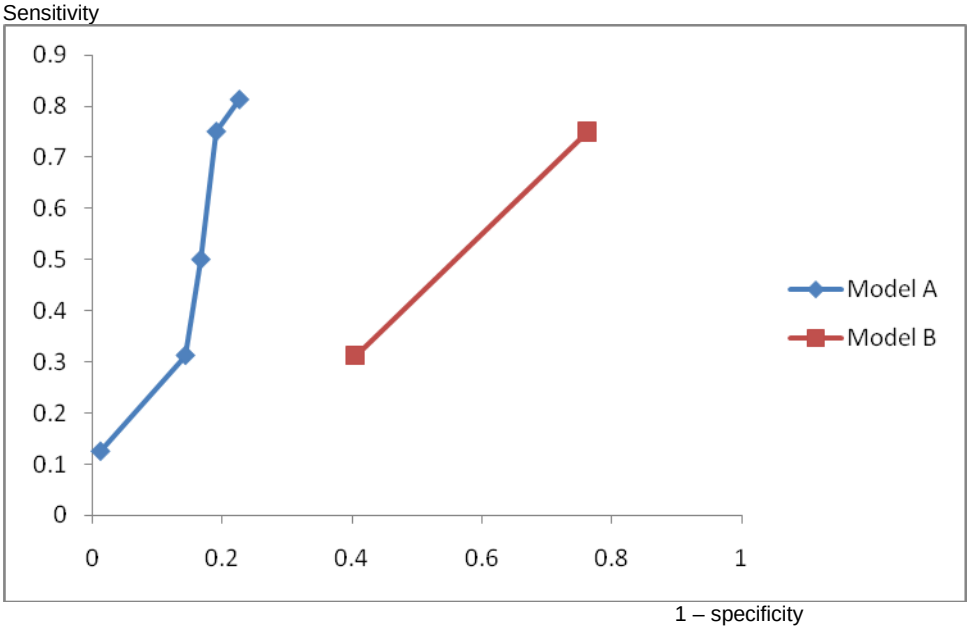
(iii) Model A:

Cut-off	Sensitivity	Specificity
1	2/16 = 0.125	83/84 = 0.9881
2	5/16 = 0.3125	72/84 = 0.8571
3	8/16 = 0.5	70/84 = 0.8333
4	12/16 = 0.75	68/84 = 0.8095
5	13/16 = 0.8125	65/84 = 0.7738

Model B:

Cut-off	Sensitivity	Specificity
1	5/16 = 0.3125	50/84 = 0.5952
2	12/16 = 0.75	20/84 = 0.2381

The ROC curve plots sensitivity against (1 – specificity).



(iv) Although it is based on only two points, it can be seen that the ROC curve for Model B is close to being a diagonal line from (0, 0) to (1, 1), so this is little better than chance. But the curve for Model A is reasonably near to the ideal shape. So we can conclude that birthweight (which appears in Model A but not in Model B) is indeed an important variable here.

Solution continued on next page

(v) Positive predictive value ("PPV") is the proportion with a positive test result for which this result is correct (i.e. the ratio of true positives to all positives).

So for Model B at cut-off 1, PPV = 5/39 = 0.1282; and at cut-off 2, PPV = 12/76 = 0.1579.

Similarly, negative predictive value ("NPV") is the proportion with a negative test result for which this result is correct (i.e. the ratio of true negatives to all negatives).

So for Model B at cut-off 1, NPV = (84 – 34)/61 = 0.8197; and at cut-off 2, NPV = (84 – 64)/24 = 0.8333.

PPV and NPV are both affected by the prevalence of the disease. Sensitivity and specificity are characteristics of the test, not of the population to which the test is applied. Their values will generalise to other settings and populations, whereas PPV and NPV do not.

(vi) Assuming a 16% prevalence of the problem, then whatever cut-off point is used in Model B is likely to give a high proportion, say 85% or more, of false positives, which would cause unnecessary distress to many people. The false negative rate is also poor, being at least around 17%.

Graduate Diploma, Module 5, 2010. Question 6

(i) A life table describes the survival pattern of a group of individuals throughout life using the age-specific death rates currently observed in a particular community. It is a convenient summary of current mortality rather than a description of the actual mortality experience of any group.

A current life table summarises current mortality and may be used as an alternative to standardisation for comparing mortality patterns in different communities.

A cohort life table describes the actual survival experience of a group or cohort of individuals born at about the same time.

Age-specific death rates refer to a single specified age group in a life table.

Part (ii)

The table given in the question is expanded as shown below, where

$_{10}q_x$ = probability that a person aged x years dies within the next 10 years (values of this are given in the question)

l_x = number of each year's cohort (of 1000) attaining age x

$_{10}d_x$ = number dying within 10 years of attaining age x ($= l_x \times {}_{10}q_x$)

$_{10}L_x$ = number living between ages x and $x + 10$ ($= 10 \times \frac{1}{2}(l_x + l_{x+10})$)

T_x = number of persons aged x or greater ($= \sum_{y \geq x} {}_{10}L_y$).

[Note. An additional column of e_x = average future lifetime of persons aged x ($= T_x/l_x$) is sometimes added. These values are explicitly calculated for $x = 20$ in part (b) and for $x = 0$ in part (c).]

Age (x)	$_{10}q_x$	l_x	$_{10}d_x$	$_{10}L_x$	T_x
0	0.250	1000	250	8750	54490
10	0.024	750	18	7410	45740
20	0.040	732	29	7175	38330
30	0.051	703	36	6850	31155
40	0.062	667	41	6465	24305
50	0.091	626	57	5975	17840
60	0.172	569	98	5200	11865
70	0.335	471	158	3920	6665
80	0.624	313	196	2155	2745
90	1.000	118	117	590	590
100		0		0	0

[Slightly different values may be obtained in these calculations depending on rounding.]

Solution continued on next page

(a) The age distribution ($= \frac{100({}_{10}L_x)}{54490}$ %) is as follows. [Note that there are small rounding errors in the calculations: the sum of these percentages is 99.99.]

Age	0 –	10 –	20 –	30 –	40 –	50 –	60 –	70 –	80 –	90 –	100 –
%	16.06	13.60	13.17	12.57	11.86	10.97	9.54	7.19	3.95	1.08	0

(b) Expected age at death for a group at present of age x is $x + \frac{T_x}{l_x}$.

For age 20, this is 20 + (38330/732) = 72.36.

(c) The life expectancy is the expected age at death if at present of age 0, i.e. the result of the calculation in (b) for x = 0:

0 + (54390/1000) = 54.49.

(d) There is an assumption of the same death rate in both sexes, which is unlikely. There is also an assumption of uniform death rate within each 10-year age group, which is also unlikely (in later life, for example). Epidemics are not explicitly allowed for, though the probabilities are presumably based on a large amount of data so there is, to some extent, implicit coverage of this factor.

Part (iii)

Population *B* will start with a cohort $(1 + 0.01)^{x+5}$ times the size for population *A*.

A suitable form of age distribution for *B* that would permit comparison with *A* is

$$\frac{100(1+0.01)^{-x+5}\,{}_{10}L_x}{\sum_i (1+0.01)^{-x_i+5}\,{}_{10}L_{x_i}}.$$

The increased birth rate in *B* leads to higher proportions in the lower age groups as compared with *A*.

Graduate Diploma, Module 5, 2010. Question 7

Preliminary calculations:-

From the census, the population total is $X = 680900$ cattle in $N = 1238$ villages, so the population mean number of cattle per village is $\overline{X} = 680900/1238 = 550$.

From the sample of 16 villages, we have the following.

$\Sigma x_i = 3687 + 7808 = 11495$ and so $\overline{x} = 11495/16 = 718.4375$,

$\Sigma y_i = 3815 + 7384 = 11199$ and so $\overline{y} = 11199/16 = 699.9375$.

$$s_{xx} = \Sigma\left(x - \overline{x}\right)^2 = \Sigma x^2 - \frac{\left(\Sigma x\right)^2}{n} = 13462957 - \frac{11495^2}{16} = 5204517.9$$

$$s_{yy} = \Sigma\left(y - \overline{y}\right)^2 = \Sigma y^2 - \frac{\left(\Sigma y\right)^2}{n} = 12773061 - \frac{11199^2}{16} = 4934460.9$$

$$s_{xy} = \Sigma\left(x - \overline{x}\right)\left(y - \overline{y}\right) = \Sigma xy - \frac{\left(\Sigma x\right)\left(\Sigma y\right)}{n} = 12875489 - \frac{11495 \times 11199}{16} = 4829707.4$$

(i) (a) The ratio estimator of the total is

$$\hat{Y}_R = X\,\frac{\overline{y}}{\overline{x}} = 680900 \times \frac{699.9375}{718.4375} = 663366.60\,.$$

(b) The regression estimator of the total is

$$\hat{Y}_{LR} = N\left(\overline{y} + B\left(\overline{X} - \overline{x}\right)\right)$$

where B is estimated by $\frac{s_{xy}}{s_{xx}} = \frac{4829707.4}{5204517.9} = 0.92798$, so that

$$\hat{Y}_{LR} = 1238\left(699.9375 + 0.92798\left(550 - 718.4375\right)\right) = 673015.02\,.$$

Solution continued on next page

(ii) Taken to the nearest integer, these estimates are 663367 (ratio) and 673015 (regression). These are both considerably less than the estimate based on the survey data alone (which is $\hat{Y} = N\overline{y} = 1238 \times 699.9375 = 866523$ to the nearest integer). They both try to take into account the very substantial difference between $\overline{X}$ and $\overline{x}$. It has turned out that the villages in the survey have, on average, considerably more cattle than the average for the whole district, so using the survey data alone would give an inflated estimate for the whole district.

The estimated variance of the estimator based on the survey data alone is

$$\mathrm{Var}\left(\hat{Y}\right) = \frac{N^2\left(1-f\right)s_y{}^2}{n} = \frac{1238^2\left(\frac{1222}{1238}\right)\left(\frac{4934460.9}{15}\right)}{16} = 31104292280.$$

So the (estimated) relative efficiency of the ratio estimator is

$$\frac{\mathrm{Var}\left(\hat{Y}\right)}{\mathrm{Var}\left(\hat{Y}_R\right)} = \frac{31104292280}{2922981150} = 10.64$$

i.e. the ratio estimator is 10.64 times as efficient.

Similarly, the (estimated) relative efficiency of the regression estimator is

$$\frac{\mathrm{Var}\left(\hat{Y}\right)}{\mathrm{Var}\left(\hat{Y}_{LR}\right)} = \frac{31104292280}{2852776962} = 10.90$$

i.e. the regression estimator is 10.9 times as efficient.

(iii) Both adjusted estimators are very much better than that using the survey data alone, but there is little to choose between them. The regression estimator does not require the relation between *x* and *y* to pass through the origin, so it might be preferred. The approximate 95% confidence interval in this case (taking 2 as the approximate double-tailed 5% point) is

$$673015 \pm 2\sqrt{2852776962} = 673015 \pm 106823,$$

i.e. it is (566192, 779838), which might conveniently be presented as 566200 to 779800.

Alternatively, bearing in mind that there is little difference between the two adjusted estimates, it might be argued that the relation between *x* and *y* should pass through the origin, in which case the ratio estimator could be used. In a similar manner, the interval using this is 555200 to 771500.

Graduate Diploma, Module 5, 2010. Question 8

(i) Consider stratum h. Within that stratum, a sample mean $\bar{y}_h$ is an unbiased estimator of the corresponding population mean.

Now define a (0, 1) random variable with value $y = 0$ if the accommodation is not rented and value $y = 1$ if it is rented. Let A_h be the population number of rented homes in the stratum and suppose we find a_h in the sample from it. The population and sample proportions of rented homes, P_h and p_h are given thus:

$$\bar{Y}_h = \frac{A_h}{N_h} = P_h \qquad\qquad \bar{y}_h = \frac{a_h}{n_h} = p_h \ .$$

As $E(\bar{y}_h) = \bar{Y}_h$, we immediately have $E(p_h) = P_h$.

For the whole city, the stratified sample estimate is $p_{st} = \sum_{h=1}^{4} \frac{N_h}{N} p_h$, and

$E(p_{st}) = \sum \frac{N_h}{N} E(p_h) = \frac{1}{N} \sum N_h P_h = P$, the population proportion.

(ii) In general, the variance of a stratified sampling mean is

$$\mathrm{Var}(\bar{y}_{st}) = \frac{1}{N^2} \sum N_h (N_h - n_h) \frac{S_h^{\,2}}{n_h}$$

where $S_h^{\,2}$ is the population variance in stratum h,

$$S_h^{\,2} = \frac{1}{N_h - 1} \sum_{i=1}^{N_h} (y_{ih} - \bar{Y}_h)^2 = \frac{1}{N_h - 1} \left(\sum_i y_{ih}^{\,2} - N_h \bar{Y}_h^{\,2} \right).$$

For the (0, 1) random variable y introduced in part (i), we have $y_{ih}^{\,2} = y_{ih}$ and so

$$S_h^{\,2} = \frac{1}{N_h - 1} \left(\sum_i y_{ih} - N_h \bar{Y}_h^{\,2} \right).$$

Now introducing the population proportion in stratum h, we have

$$S_h^{\,2} = \frac{1}{N_h - 1} \left(N_h P_h - N_h P_h^{\,2} \right) = \frac{N_h P_h (1 - P_h)}{N_h - 1}.$$

$$\therefore \mathrm{Var}(p_{st}) = \frac{1}{N^2} \sum_{h=1}^{4} N_h (N_h - n_h) \frac{N_h P_h (1 - P_h)}{n_h (N_h - 1)}.$$

Solution continued on next page

Now consider simple random sampling within stratum h, with s_h^2 denoting the sample variance. This is an unbiased estimator of the population stratum variance, so we have

$$E\left(s_h^{\,2}\right) = E\left(\frac{1}{n_h - 1}\sum_i \left(y_{ih} - \bar{y}_h^{\,2}\right)\right) = S_h^{\,2}.$$

Also, by the same argument as used above for S_h^2, $s_h^{\,2} = \dfrac{n_h p_h (1 - p_h)}{n_h - 1}$.

Thus, for the expression in the question, we have

$$E\left(\frac{1}{N^2}\sum_{h=1}^{4} N_h (N_h - n_h)\frac{p_h(1-p_h)}{n_h - 1}\right) = \frac{1}{N^2}\sum_{h=1}^{4} N_h (N_h - n_h)\frac{S_h^{\,2}}{n_h},$$

which is Var(p_{st}), as required.

(iii) From the data in the question, we have

Stratum (h)	p_h	N_h/N	$p_h(1-p_h)/(n_h-1)$	$1-(n_h/N_h)$
< 50	0.7500	0.5874	0.004808	0.9664
50 – 100	0.5143	0.2581	0.007347	0.9331
100 – 200	0.2000	0.1061	0.004706	0.8372
> 200	0.1250	0.0484	0.002804	0.5918

The estimate (see part (i)) is

$$p_{st} = \sum_{h=1}^{4} \frac{N_h}{N} p_h$$

$$= (0.5874 \times 0.7500) + (0.2581 \times 0.5143) + (0.1061 \times 0.2000) + (0.0484 \times 0.1250)$$

$$= 0.6006, \quad \text{i.e. } 60\%.$$

Using the expression in the question for the variance, we have

$$\text{Var}(p_{st}) = \frac{1}{N^2}\sum_{h=1}^{4} N_h (N_h - n_h)\frac{p_h(1-p_h)}{n_h - 1}$$

$$= \sum_{h=1}^{4} \frac{N_h^{\,2}\left(1 - (n_h / N_h)\right)}{N^2}\frac{p_h(1-p_h)}{n_h - 1}$$

$$= \left\{\left(0.5874^2\right)(0.9664)(0.004808)\right\} + \ldots = 0.0016032 + \ldots = 0.002108,$$

and SE(p_{st}) = 0.0459.

Solution continued on next page

(iv) For optimal allocation we have $n_h \propto N_h S_h$, so we take $n_h \propto N_h\sqrt{\{P_h(1 - P_h)\}}$ and use the sample estimates of P_h. This gives that the ratio of the n_h for the four strata is

515.285 : 261.393 : 86.000 : 32.410.

These add to 895.088. Thus, as the total sample size is 150, we scale by 150/895.088 to obtain

86.35, 43.80, 14.41, 5.43

which we take as

86, 44, 14, 6.

The stratum sample sizes in the survey in the question are 40, 35, 35, 40. These are a long way away from the optimal sizes – roughly half what there should be in the first stratum, somewhat too few in the second, and far too many in the third and fourth – so these are not reasonable.