THE ROYAL STATISTICAL SOCIETY

2009 EXAMINATIONS – SOLUTIONS

GRADUATE DIPLOMA

MODULAR FORMAT

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, 2009. Question 1

(i) The principal components are linear combinations of the variables. It may be that the data are adequately explained by a smaller number of these combinations than the number of original variables. This could usefully reduce the dimensionality of the problem.

The principal components are uncorrelated. This may be a useful objective of the analysis in its own right.

Some of the principal components, likely to be those that correspond with the smallest eigenvalues, may help to indicate variables that do not need to be studied in the future.

(ii) The variance-covariance matrix can be used when all the variables are compatible with each other, typically measured in the same units, and when the magnitude, as opposed to the correlation, of the variables is of interest. However, the components will tend to be dominated by measurements that are large.

(iii) The principal components are the eigenvectors of the correlation (or variance-covariance) 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.

Part (iv)

(a) A rule that is often useful 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). There are two such eigenvalues here, suggesting a dimensionality of 2.

However, the number of variables is 5, so this is also the sum of the eigenvalues. The first three add to 4.95, so the last two are clearly negligible. This would suggest a dimensionality of 3.

There is no informal way of choosing between 2 and 3.

Solution continued on next page

(b) Component 1 is a weighted average of all the variables (this is often the case for the first principal component), with somewhat lower weight for X_2, the income of the second wage-earner in a household, than for the other variables. This component may be a measure of overall wealth.

Component 3 is mainly a contrast between the incomes of the first and second wage-earners.

(c) Component 2 is a contrast between (X_1, X_2, X_3) and (X_4, X_5). This is difficult to interpret in economic terms. It does not follow that this component is wrong. The method is a mathematical one and the components are essentially mathematical constructs; it is frequently difficult to interpret them in the context of the problem. In the present case, the economist's view that the coefficient of X_3 in this component should be +0.48 rather than –0.48 is equivalent to regarding this variable (total debts) as a liability rather than an asset – an arguable point of view.

(d) All the variables are measured in pounds sterling, which is an initial indication that use of the variance-covariance matrix might be appropriate. However, X_3 measures total debts whereas all the others are measured in pounds per month, so the units of measurement are not in fact the same.

Further, there are likely to be major differences of magnitude in the variables. X_1 and (probably) X_2 are likely to be larger than X_4 and (especially) X_5. X_3 could be small or, alternatively, very large and could dominate an analysis based on the variance-covariance matrix.

In short, the results are unlikely to be similar.

(e) It is impossible to be sure about this.

Strictly, principal components analysis requires complete cases of data (i.e. no missing values). Some computer packages simply omit all cases of data for which *any* individual observation is missing. This reduces the sample size (sometimes dramatically) and thus leads to less reliable results. Other packages calculate the correlation for each pair using all the available observations for that pair. This means that the sample sizes for the correlations vary and can lead to the correlation matrix not being positive definite or, more generally, to unstable results.

The reasons for missing values should be investigated. As an example, some people, especially with high debts, might have declined to give a value for X_3; this would certainly affect the results. On the other hand, if a fairly small number of values are missing purely at random, the effect on the analysis may be quite small.

Graduate Diploma, Module 5, 2009. Question 2

(i) If 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.

(ii) The observations should follow a multivariate Normal distribution, and the 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 for equality of variance-covariance matrices exist, they are sensitive to non-Normality. It is helpful to check univariate Normality for each variable; this can be done in the usual ways (e.g. histograms, stem-and-leaf plots, Normal probability plots). However, univariate Normality is a necessary but not sufficient condition for multivariate Normality. It is also sensible to check the sample variance-covariance matrices for each group, to see whether they appear to be roughly similar.

Part (iii)

(a) Fisher's linear discriminant function $y = \mathbf{a}^T\mathbf{x}$ is that linear combination of coefficients $a_1, a_2, ..., a_p$ and the variables $x_1, x_2, ..., x_p$ which maximises the distance between the means $\boldsymbol{\mu}_E$ and $\boldsymbol{\mu}_H$, where $\boldsymbol{\mu}_E$ and $\boldsymbol{\mu}_H$ are the mean vectors for "exaggerators" and "honest" respectively.

(b) Simple multiplication of the quoted inverse matrix and $\boldsymbol{\Sigma}$ immediately gives the identity matrix, confirming that the quoted matrix is the inverse.

The vector $\mathbf{a}$ is given by $\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_E - \boldsymbol{\mu}_H) = \frac{1}{5767}\begin{bmatrix} 92 & -57 \\ -57 & 98 \end{bmatrix}\begin{bmatrix} 9 \\ 8 \end{bmatrix} = \begin{bmatrix} 0.0645 \\ 0.0470 \end{bmatrix}$.

Therefore Fisher's linear discriminant function is $y = 0.0645x_1 + 0.0470x_2$, where x_1 refers to "exaggerators" and x_2 to "honest".

Solution continued on next page

(c) Inserting $\boldsymbol{\mu}_E$ and $\boldsymbol{\mu}_H$ in the discriminant function gives

for $\boldsymbol{\mu}_E$: $(0.0645 \times 20) + (0.0470 \times 19) = 2.183$

for $\boldsymbol{\mu}_H$: $(0.0645 \times 11) + (0.0470 \times 11) = 1.227$.

As *E* and *H* are being assumed equally likely, the dividing point is simply half-way between these two values, i.e. 1.705, and the decision rule is to assign an observation as "honest" (*H*) if its value of *y* is < 1.705 ("nearer" to $\boldsymbol{\mu}_H$ than to $\boldsymbol{\mu}_E$) and as "exaggerator" (*E*) otherwise.

The variance underlying *y* is given by the following, taking the variances and covariances from the $\boldsymbol{\Sigma}$ matrix quoted in the question:

$$\text{Var}(y) = (0.0645)^2(98) + (0.0470)^2(92) + 2(0.0645)(0.0470)(57)$$
$$= 0.9565.$$

Thus for "honest" people we have $y \sim \text{N}(1.227, 0.9565)$, so the probability that an "honest" person is misclassified is

$$P(\text{N}(1.227, 0.9565) > 1.705) = P(\text{N}(0, 1) > 0.489) = 0.312.$$

By symmetry, the misclassification probability is the same for "exaggerators", and so (as they are equally likely) the overall probability of misclassification is 0.312.

(d) Now we have a cut-off point of 2.

P("honest" misclassified) = $P(\text{N}(1.227, 0.9565) > 2) = P(\text{N}(0, 1) > 0.790) = 0.215$.

P("exag." misclassified) = $P(\text{N}(2.183, 0.9565) < 2) = P(\text{N}(0, 1) < -0.187) = 0.426$.

We also have P("honest") = 0.9, P("exaggerator") = 0.1.

So the overall probability of misclassification is

$$(0.215 \times 0.9) + (0.426 \times 0.1) = 0.236.$$

Graduate Diploma, Module 5, 2009. Question 3

(i) Consider the model with *X* alone in it. The value of the estimated regression coefficient (–0.046) is much greater than its standard error (0.008), which suggests that this is a significant predictor. A better test is based on considering the difference in –2(log likelihood) between this model and the null model. The difference is 639.19 – 591.81 = 47.28, with 1 degree of freedom. This is extremely highly significant as an observation from χ^2_1, so we can conclude that there is extremely strong evidence that *X* is a predictor of survival.

(ii) For *QL1* in model C, the coefficient is estimated by –0.024 and a 95% confidence interval for it is $-0.024 \pm (1.96 \times 0.006)$, i.e. (–0.0358, –0.0122).

The hazard ratio is given by exp(coefficient), so the estimate of the hazard ratio is exp(–0.024) = 0.976 and the 95% confidence interval for the hazard ratio is from exp(–0.0358) to exp(–0.0122), i.e. from 0.965 to 0.988.

Thus two people with the same value of *X* but a difference of 1 point in *QL1* have very nearly the same risk of death at any time. But for a difference of 10 points on the quality of life scale, we calculate 10 × coefficient = –0.24 and then exp(–0.24) = 0.787. The associated confidence interval for this is from exp(–0.358) to exp(–0.122), i.e. from 0.699 to 0.885. This means that, for two people with the same value of *X* but a difference in *QL1* value of 10 points, we estimate that the risk of death for the person with higher *QL1* is 79% [0.787, rounded] that of the other person, and we are 95% confident that the interval 70% to 89% [rounded values again] contains this relative risk of death.

(iii) We examine the differences in –2(log likelihood).

From model B to model C, the difference is 15.07, with 1 df – highly significant as an observation from χ^2_1.

We can therefore reasonably conclude that, after taking account of *X*, *QL1* does add information.

From model B to model D, the difference is 10.19, with 1 df – again highly significant as an observation from χ^2_1.

We can therefore likewise reasonably conclude that, after taking account of *X*, *QL2* does add information.

Solution continued on next page

(iv) We again use the differences in –2(log likelihood).

First, we use forward selection. We have already seen (part (i) above) that *X* should be included and, part (iii) above, that including either *QL1* or *QL2* brings about a further improvement. Clearly *QL1* would be preferred at this stage as the value of –2(log likelihood) is smaller in model C (with *QL1*) than in model D (with *QL2*). Now adding *QL2* to the model with *X* and *QL1* already in it, i.e. moving from model C to model E, we get a further reduction of 1.98 in –2(log likelihood), again with 1 df. This is not significant as an observation from χ^2_1, so we do <u>not</u> move to model E – we choose model C.

We now try backward elimination. We start with the full model, i.e. model E. On removing each of the *QL* variables in turn, we find that the smallest increase in –2(log likelihood) is obtained by eliminating *QL2* and thus moving to model C. The increase is 1.98, which is not significant, so we adopt model C at this stage. We now try eliminating *QL1* by moving to model B. The increase here is 15.07 which is highly significant. So we do <u>not</u> move to model B – we choose model C.

Overall, model C appears best.

[**Note.** In this case, forward selection and backward elimination have arrived at the same answer, and the calculations that have been carried out in the two methods are the same. But this is not necessarily so – the methods are different and need not coincide in this way.]

(v) We use model C. The value of *QL2* is the same for both patients but in any case is irrelevant for model C. The values of *QL1* differ by 10 points for the two patients, patient 1 having a higher (better) value. But survival also depends on the value of *X*, as is shown by the Cox regression model being estimated as $-0.048X - 0.024QL1$.

Suppose that *X* changes by ΔX (measured as the value for patient 1 minus that for patient 2) as *QL1* changes by 10 (measured in the same direction). For there to be no change in the value of the Cox regression function, we would require $-0.048\Delta X = 0.24$, i.e. $\Delta X = -5$. Values of ΔX that are *lower* than –5 (i.e. $\Delta X < -5$) lead to a lower (i.e. "more negative") value of the Cox regression function and hence to a *higher* value for the hazard – i.e. patient 2 is expected to survive *longer*.

Thus, for these two patients with a difference of 10 points in *QL1* in favour of patient 1, it is nevertheless the case that patient 2 is expected to survive longer if the *X* value for patient 1 is more than 5 points lower than that for patient 2.

Solution continued on next page

(vi) The proportional hazards assumption is that the hazard function is

$$h(t) = h_0(t)\exp\left(\beta_1 X + (\text{the other terms in the model})\right)$$

where $h_0(t)$ is the baseline hazard function, the key point being 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.

If the assumption is valid for *X* but not for *QL1* and *QL2*, then part (i) above is unaffected (since it only involves *X*), but the other models are deficient. In particular the interpretation of the hazard ratios is invalid: for example, the effect of *QL1* may change over time. The models, values of –2(log likelihood) and inferences may be unreliable, and the discussion and conclusions may be invalid. The consequences depend to some extent on the type and seriousness of any departures from the assumption.

Graduate Diploma, Module 5, 2009. Question 4

Part (i)

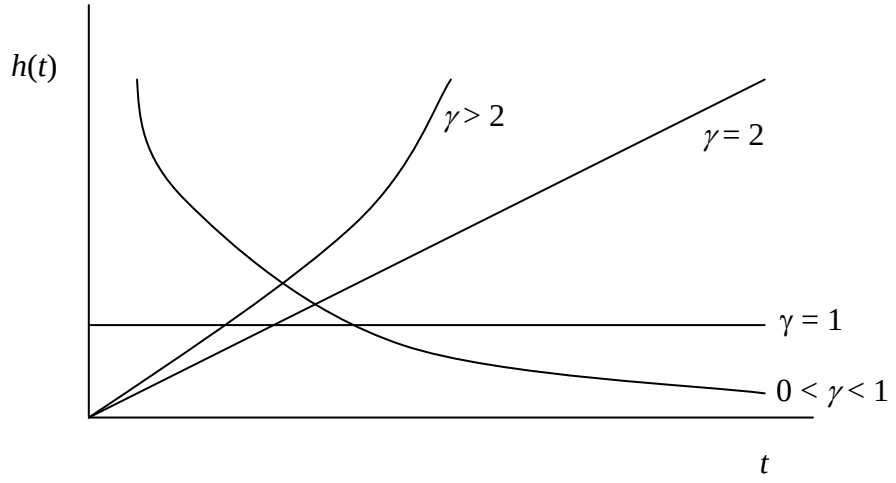
(a) The hazard function $h(t)$ is given by $h(t)=\dfrac{f(t)}{1-F(t)}$ where $f(t)$ is the pdf and $F(t)$ the cdf. Here,

$$F(t)=\int_0^t \lambda\gamma x^{\gamma-1}\exp\left(-\lambda x^{\gamma}\right)dx=\left[-\exp\left(-\lambda x^{\gamma}\right)\right]_0^t=1-\exp\left(-\lambda t^{\gamma}\right),$$

and so the hazard function is

$$h(t)=\frac{\lambda\gamma t^{\gamma-1}\exp\left(-\lambda t^{\gamma}\right)}{\exp\left(-\lambda t^{\gamma}\right)}=\lambda\gamma t^{\gamma-1}.$$

(b)



(c) $\gamma = 1$: hazard is constant, a product in routine use where failure is random.

$0 < \gamma < 1$: a new product with high early incidence of failures but then the hazard declines (i.e. the product "gets better") over time.

$\gamma = 2$ and $\gamma > 2$: products for which the hazard of failure increases with time ("wearing out"), linearly in the case of $\gamma = 2$ and increasing steeply with time for $\gamma > 2$.

Solution continued on next page

Part (ii)

(a) The solid line corresponds to stress 700. The long dashes correspond to stress 750. The short dashes correspond to stress 800.

(b) The Kaplan-Meier survival curve is constructed as follows. We seek the estimated cumulative survival (i.e. "still working") function $\hat{S}(t)$.

The Kaplan-Meier method requires the ordered failure times $t_{(1)}$, $t_{(2)}$, …, $t_{(r)}$ to be considered. For $j = 1, 2, \ldots, r$, let $n_{(j)}$ be the number of springs still working just before time $t_{(j)}$, and let $d_{(j)}$ be the number that fail 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.

> [**Note.** There are no censored failure times for stress 800. If the largest failure time had been 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.]

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. If there were any rows for censored observations, they might be omitted, but care must be taken to ensure that $n_{(j)}$ is always correct.

There are 8 springs. The data (failure times, in ascending order) are as follows: 365 400 462 523 625 1053 1432 2024.

Time $t_{(j)}$	$n_{(j)}$ as defined in text above [i.e. number working just before time $t_{(j)}$]	$d_{(j)}$ as defined in text above [i.e. number of failures at time $t_{(j)}$]	$\dfrac{n_{(j)} - d_{(j)}}{n_{(j)}}$	Cumulative survival estimate $\hat{S}(t)$ at each $t_{(j)}$
365	8	1	7/8	0.875
400	7	1	6/7	0.750
462	6	1	5/6	0.625
523	5	1	4/5	0.500
625	4	1	3/4	0.375
1053	3	1	2/3	0.250
1432	2	1	1/2	0.125
2024	1	1	0	0

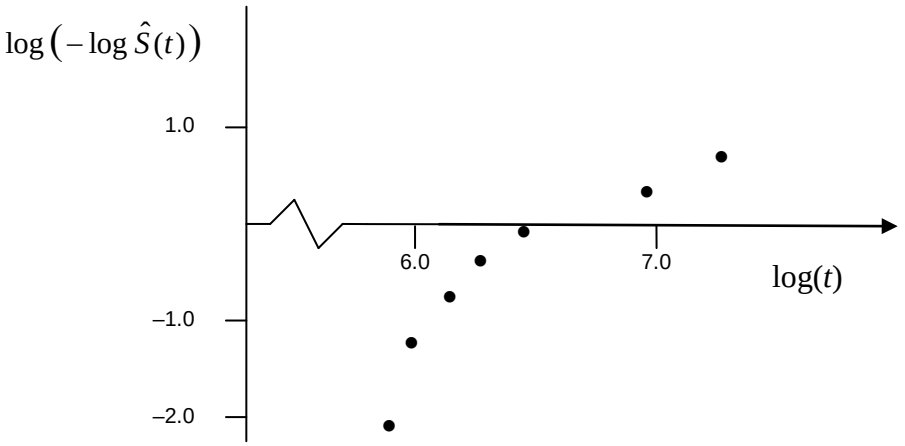
Solution continued on next page

(c) We need values of log(time) and $\log\left(-\log \hat{S}(t)\right)$ from the table in part (b).

Obviously we exclude the last row of the table as this would require log(0). The other values are as follows.

log(time)	$\log\left(-\log \hat{S}(t)\right)$
5.900	–2.013
5.991	–1.246
6.136	–0.755
6.260	–0.367
6.438	–0.019
6.959	0.327
7.267	0.732

These are graphed below. Clearly the plot does not approximate to a straight line, so we may conclude that a Weibull distribution is not a good model here.



Graduate Diploma, Module 5, 2009. Question 5

(i) (a)

		Cases (birthweight < 2500 g)		Controls (birthweight ≥ 2500 g)		
		Number	*% of all cases*	Number	*% of all controls*	Total
Smoked	No	29	*49*	86	*66*	115
	Yes	30	*51*	44	*34*	74
	Total	59		130		189

About one half of "case" mothers smoked during pregnancy, compared with only about one third of "control" mothers. This suggests that smoking is a risk factor for low birthweight.

(b) Using the table in part (a) above, the odds ratio for the risk factor smoking status (smoked compared with did not smoke) can be calculated as

$$\frac{30\times 86}{29\times 44} = 2.02\,.$$

To obtain a 95% confidence interval, we work via the log of the odds ratio and use the formula (in which *a*, *b*, *c*, *d* represent the frequencies in the table)

$$\text{Var}(\log \text{ of odds ratio}) = \frac{1}{a}+\frac{1}{b}+\frac{1}{c}+\frac{1}{d} = \frac{1}{29}+\frac{1}{86}+\frac{1}{30}+\frac{1}{44} = 0.1022\,.$$

Thus a 95% confidence interval for the log of the odds ratio is given by

$$\log(2.02) \pm \left(1.96\times\sqrt{0.1022}\right) \;=\; 0.7031 \pm 0.6265\,,$$

i.e. it is 0.077 to 1.330; so the interval for the odds ratio is 1.08 to 3.78.

(c) The odds ratio will be a good approximation to the relative risk if the incidence of low birthweight is relatively low in this population.

(ii) Ignoring any (likely) confounding, the following interpretations apply. Smoking appears to be associated with increased risk of low birthweight (note that the confidence interval for this odds ratio does not contain the value 1.00). There is also some evidence that "black" and "other" ethnic origins are associated with increased risk of low birthweight compared with "white" – the confidence intervals for these odds ratios do contain 1.00, but the lower limits are only slightly less than 1.00 and the upper limits are substantially greater than 1.00. In the case of zero ante-natal visits compared with 1, the confidence interval is wide and comfortably contains 1.00; there may perhaps be slight evidence of an association with increased risk of low birthweight, but no more than slight evidence, at most. Somewhat less intuitively, the same is true for 2 or more ante-natal visits compared with 1: there is some appearance of increased risk, but the confidence interval is wide and easily contains 1.00. Perhaps increased visits are due to concern about other known risk factors.

Solution continued on next page

(iii) The Mantel-Haenszel method is a simple way of adjusting for other factors. [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 "levels" of the other factors separate, the Mantel-Haenszel estimate of the odds ratio is

$$\frac{\sum a_i d_i / n_i}{\sum b_i c_i / n_i}$$ where the summations are over all "levels".

A typical table for 2 or more ante-natal visits compared with 1 is as follows, which is for non-smokers of white ethnic origin:

	Case (< 2500)	Control (≥ 2500)	
2 or more	2 (= a_i)	9 (= c_i)	
1	0 (= b_i)	20 (= d_i)	Total 31 (= n_i)

Similar tables can easily be constructed for each of the other combinations of "levels", and are summarised in the following display.

Smoked	Ethnic origin	a_i	b_i	c_i	d_i	n_i
No	White	2	0	9	20	31
	Black	2	1	2	3	8
	Other	3	5	7	6	21
Yes	White	4	4	9	5	22
	Black	1	1	1	1	4
	Other	0	0	2	1	3

This gives that the Mantel-Haenszel estimate is

$$\frac{\frac{2\times 20}{31}+\frac{2\times 3}{8}+\frac{3\times 6}{21}+\frac{4\times 5}{22}+\frac{1\times 1}{4}+\frac{0\times 1}{3}}{\frac{0\times 9}{31}+\frac{1\times 2}{8}+\frac{5\times 7}{21}+\frac{4\times 9}{22}+\frac{1\times 1}{4}+\frac{0\times 2}{3}} = \frac{4.057}{3.803} = 1.07 \ .$$

The corresponding unadjusted odds ratio (quoted in the question) is 1.31.

Solution continued on next page

The adjusted value, 1.07, is close to 1.00. This suggests that any risk associated with 2 or more ante-natal visits compared with 1, after adjusting for all the other factors, is unlikely to be of clinical importance. It is certainly unlikely to be of statistical significance in this (small) sample.

The difference between the unadjusted and adjusted values is appreciable. This suggests that the extra ante-natal visits are likely to be associated with a real causal risk factor such as smoking (or, perhaps, ethnic group). It is interesting to note that the two-way table for "controls" classified by smoking and by 1 or (2 or more) visits distinctly suggests that there is an association here, the smokers appearing to tend to pay more visits:-

	Smoker: no	Smoker: yes
1 visit	29	7
2 or more visits	18	12

(Note that the corresponding table for "cases" shows no such pattern. The interpretations are not straightforward.)

(iv) A good alternative technique is logistic regression for the binary response variable of low birthweight, with the various risk factors being included in the model. This gives adjusted odds ratios (though a constant in the output has no useful meaning for case-control studies).

Graduate Diploma, Module 5, 2009. Question 6

Part (a)

(i) $n_i = n_i' - \frac{1}{2}(l_i + w_i)$ [assuming uniform distributions of losses and withdrawals during the interval].

(ii) $\hat{q}_i = d_i / n_i$ for all intervals except the last; $\hat{q}_i = 1$ for the last interval.

(iii) $\hat{p}_i = 1 - \hat{q}_i$ (as defined in part (ii)).

(iv) $\hat{S}(t_0) = 1$; $\hat{S}(t_i) = \hat{p}_{i-1}\hat{S}(t_{i-1})$ for i = 1, 2, ...

Part (b)

(i) The completed life-table is as shown. (Due to rounding, the fourth decimal place might not be completely reliable.)

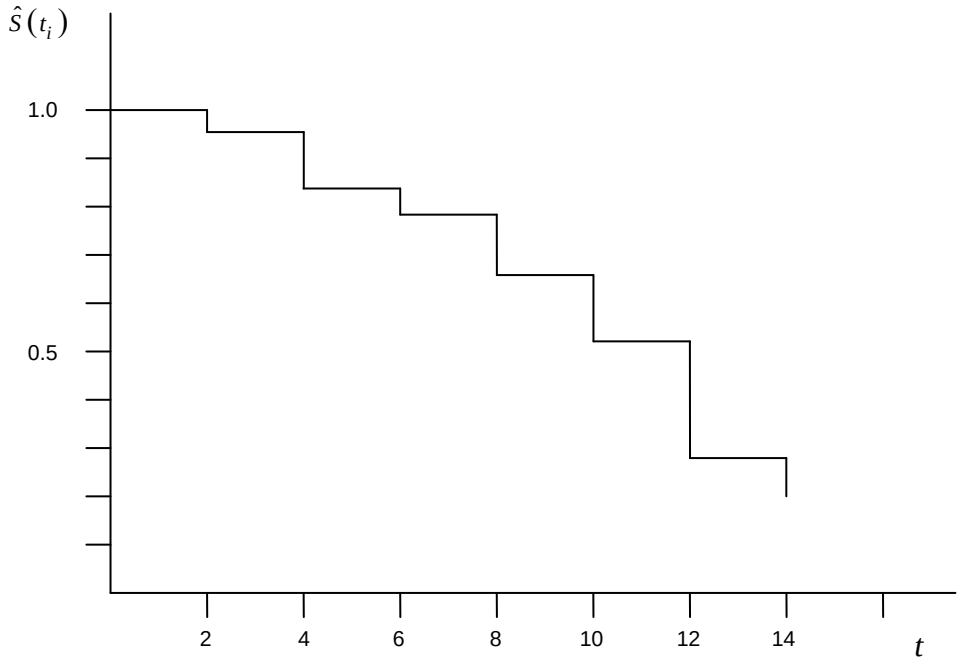
Months after randomisation	$[t_i, t_{i+1})$	n_i'	$w_i + l_i$	d_i	n_i	$\hat{q}_i$	$\hat{p}_i$	$\hat{S}(t_i)$
0	[0, 2)	163	43	7	141.5	0.0495	0.9505	1.0000
2	[2, 4)	113	19	11	103.5	0.1063	0.8937	0.9505
4	[4, 6)	83	11	5	77.5	0.0645	0.9355	0.8495
6	[6, 8)	67	18	10	58.0	0.1724	0.8276	0.7947
8	[8, 10)	39	14	7	32.0	0.2188	0.7812	0.6577
10	[10, 12)	18	4	7	16.0	0.4375	0.5625	0.5139
12	[12, 14)	7	2	2	6.0	0.3333	0.6667	0.2891
14+	14+	3	3	0	1.5	0.0000	1.0000	0.1927

It is assumed that the withdrawals are subject to the same probabilities of the event of interest as the non-withdrawals. This is a reasonable assumption for withdrawals who are still in the study and would be available for future follow-up, but perhaps dangerous for those who are lost to follow-up, since failure to examine a participant for any reason may be related to the participant's health.

It is also assumed that the values of $\hat{p}_i$, being obtained from participants who enter the study at different times, remain reasonably constant over time.

Solution continued on next page

(ii)



The median survival time is the time beyond which 50% of the individuals in the population under study are expected to survive. From the above graph, this can be estimated at about 10 months.

The above median survival time is for women in the treatment group. The median survival time for those in the control group is approximately 6 months (given in the question). So those in the treatment group have a slightly longer (by about 4 months) median survival time and so, on the whole, we may say that they tend to survive slightly longer.

(iii) The suggested analysis would not use information on exactly how long the participants survived without the event of interest. It seems likely that all participants would be followed up for the same period of time, so the estimated proportion would be biased.

Graduate Diploma, Module 5, 2009. Question 7

Part (a)

$\bar{y}_{st}$ is the stratified sampling estimator of the overall population mean, and $\mathrm{Var}(\bar{y}_{st})$ is its variance.

There are k strata. N_i is the population size for the ith stratum. N (= ΣN_i) is the total population size. n_i is the size of the random sample taken in the ith stratum. S_i^2 is the population variance in the ith stratum.

For convenience, write $W_i = N_i/N$ (often called the "stratum weight" for the ith stratum), with which

$$\mathrm{Var}(\bar{y}_{st}) = \sum_{i=1}^{k} \frac{W_i^2 S_i^2}{n_i} - \sum_{i=1}^{k} \frac{W_i^2 S_i^2}{N_i} = V, \text{ say.}$$

The total cost of sampling is $C = c_0 + \Sigma c_i n_i$ which is to be minimised subject to fixed V.

This constrained minimisation problem can be solved by the standard technique of introducing a Lagrange multiplier λ. We minimise

$$c_0 + \sum_{i=1}^{k} c_i n_i + \lambda \left(\sum_{i=1}^{k} \frac{W_i^2 S_i^2}{n_i} - \sum_{i=1}^{k} \frac{W_i^2 S_i^2}{N_i} \right).$$

Differentiating with respect to each n_i and setting equal to zero gives that the minimum is attained for $c_i - \frac{\lambda W_i^2 S_i^2}{n_i^2} = 0$, i.e. $n_i = \frac{W_i S_i \sqrt{\lambda}}{\sqrt{c_i}}$.

But λ is a constant, and $W_i = N_i/N$. So n_i is proportional to $N_i S_i/\sqrt{c_i}$. Simply dividing by the sum of all these quantities gives the required answer.

[Note. The result can also be obtained by an application of the Cauchy-Schwarz inequality.]

Solution continued on next page

Part (b)

Using the given estimated standard deviations as the population figures, we find the following (with n denoting the total sample size, $n = \Sigma n_i$).

N= 227

	N_i	N_iS_i	$N_iS_i/\sqrt{c_i}$	n_i/n
District 1	68	2312	731.119	0.4256
District 2	143	2860	764.367	0.4449
District 3	16	944	222.503	0.1295
Total	227	6116	1717.989	1

Translating the second requirement so as to refer to the population mean $\overline{Y}$, it becomes that we require

$$P\left(\left|\overline{y}_{st} - \overline{Y}\right| > \frac{300}{227}\right) \leq 0.05.$$

Assuming a Normal distribution for $\overline{y}_{st}$, we have $\overline{y}_{st} \sim \mathrm{N}\left(\overline{Y}, V\right)$ where V is as in (a). Hence $\frac{300}{227} = 1.96\sqrt{V}$, so $\sqrt{V} = 0.6743$ and $V = 0.45465$.

The given formula (in part (a)) for V can be written as

$$N^2V = \sum \frac{N_i^2 S_i^2}{n\left(n_i / n\right)} - \sum N_i S_i^2 ,$$

so we have

$$227^2 \times 0.45465$$
$$= \frac{1}{n}\left(\frac{68^2 \times 34^2}{0.4256} + \frac{143^2 \times 20^2}{0.4449} + \frac{16^2 \times 59^2}{0.1295}\right) - \left((68 \times 34^2) + (143 \times 20^2) + (16 \times 59^2)\right)$$

from which, after some arithmetic, n = 176. Hence n_1 = 75, n_2 = 78, n_3 = 23. However, this requires sampling of more than 100% in the first and third districts, which is obviously impossible. So we take n_1 = 68 and n_3 = 16 (100% sampling in these two districts).

Thus, adding in the overhead cost, the total cost of sampling is

C = 200 + (10 × 68) + (14 × 78) + (18 × 16) = 2260 units.

Graduate Diploma, Module 5, 2009. Question 8

[Solution continues on the next page]

The notes below are intended as guides to points that might be made in good answers. In the examination, credit was given for any relevant comments that were made and explained.

(i) A good sampling frame of readers of monthly magazines is unlikely to exist (though agencies that build up databases by buying into other organisations' lists may have some useful material). The most likely way in which some sort of frame would be built up is to include a short questionnaire in the magazine and ask readers to return it. A "freepost" facility would be essential, and an incentive such as entry into a "prize draw" would help encourage responses. It would be sensible to include the questionnaire in at least two months' issues, partly to help overcome non-response the first time, partly because there are likely to be many readers who do not buy every issue (for example, some readers may only buy motor magazines when they are considering buying a car); but on the other hand, there might be duplicate replies which would need to be removed, and there is potentially a problem of irritating regular readers. Overall, though, response rates are likely to be poor, though they may be better for magazines that cater for special age- or interest-groups to which they appeal strongly. A further problem is that of multiple readership: a magazine may be read by several members of a household, or shared among friends. A very useful question would be to ask how many other regular readers there are besides the respondent. Even this is very unlikely to cover the case of magazines read in communal places such as libraries or doctors' and dentists' waiting rooms, as there would almost certainly be no response from such places.

All in all, any sampling frame is likely to be incomplete and generally poor in its coverage, and it would be difficult even to have some notion of the nature of the deficiencies. The use of a questionnaire as suggested above may be the only practicable way of selecting a sample.

If, however, the magazine has subscribers, who pay in advance for say a year's issues that are sent by post, there will be good information in the subscription department – names, addresses, other contact points. It should be easy to select a sample from the lists of subscribers. Further, people who subscribe to a magazine are likely to be committed to it and, as such, may well take the trouble to respond to the questionnaire. Subscribers are not likely to have the same characteristics in respect of the magazine as people who buy it more casually in shops, but they may be of special interest to advertisers.

(ii) The monthly circulation figure will be known (it may perhaps vary due to seasonal and other effects, but nevertheless it will be known – and potential advertisers would certainly be interested in any such variations). If the "multiple readership" question discussed in part (i) is successful, an estimate

of total readership can be obtained. It would be very useful to know whether a respondent is a regular reader (and whether or not a subscriber), an occasional reader or merely someone who happened to make a casual purchase. Some indication of why the magazine was bought would be very helpful, but this should be in the form of a question with pre-specified answers for respondents to select, otherwise analysis might be very difficult; and it may turn out that a response such as "general interest" would swamp all others. Especially for specialist readerships, information on their buying habits for various goods and services would be of much interest to potential advertisers, though care should be taken to avoid the danger of the questionnaire being seen as intrusive and thus generating ill-will.

(iii) The questionnaire survey could of course be repeated at intervals. It might not be easy to tell whether or not respondents had replied to earlier circulations (though this should be easy to tell in the case of subscribers), but somehow the marketing manager needs to be in a position to point to any changes in the readership that may be helpful to advertisers. The sampling frame, despite its likely deficiencies, may still be able usefully to be stratified, and this could be very helpful in identifying market segments that might be especially attractive to advertisers. Also, any further survey undertaken by the magazine itself could be conducted as stratified sampling if it appeared appropriate to do so.