

**THE ROYAL STATISTICAL SOCIETY**

**GRADUATE DIPLOMA EXAMINATION**

**NEW MODULAR SCHEME**

**introduced from the examinations in 2009**

**MODULE 5**

**SOLUTIONS FOR SPECIMEN PAPER A**

**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.

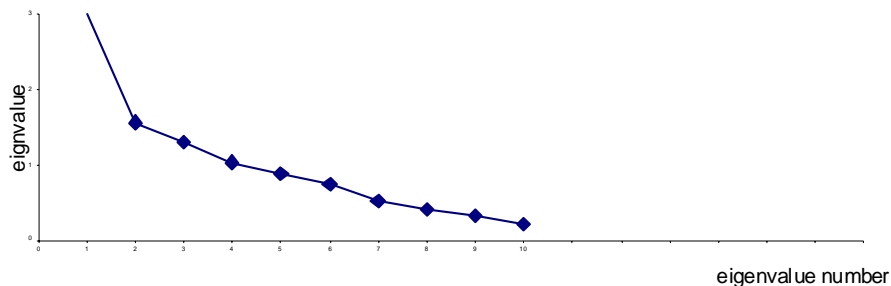
Note. In accordance with the convention used in all the Society's examination papers, the notation  $\log$  denotes logarithm to base  $e$ . Logarithms to any other base are explicitly identified, e.g.  $\log_{10}$ .

Graduate Diploma Module 5, Specimen Paper A. Question 1

- (a) PCA produces uncorrelated components which are weighted linear combinations of  $p$  measurements made on each of  $n$  units, accounting for decreasing proportions of the total variation among the units and corresponding to the eigenvalues (in decreasing order of size) in the correlation (or variance-covariance) matrix of the measurements. There are the same number of components as measurements, but the hope is that a small number of them will account for most of the variation and so reduce the dimensionality of a problem.

Cluster analysis aims, on the basis of  $p$  measurements on each, to group the units into sets that are "similar", again using the correlation matrix.

- (b) (i) m400, m100 and m110h, the "short-run" variables, seem moderately well correlated. Most other correlations are moderate or low. m1500 does not appear to be noticeably correlated with any of the others; nor does high jump. Thus there seems to be only one obvious cluster.
- (ii) There are 4 eigenvalues greater than 1, and they take up 69% of the variation, so this is a reasonable cut-off point to take as a basis for interpretation. Note that short times and long distances show good performance.



- (iii) PC1 is a general average measure of good performance, bearing in mind the remark above. [The first PC of a correlation matrix very often is this.] m100, long jump, m400, m110h, discus and javelin are the major contributors to it.

PC2 is a contrast between high jump and m400.

PC3 is largely m1500, with less contribution from others.

PC4 is pole vault, with some contribution from other jumping measures.

**Solution continued on next page**

- (iv) Raw Euclidean distances give undue weight to variables with a longer time or distance characteristic. It would be better to scale the measures so that each had variance 1. If some were thought more important than others, they could be given greater weighting in the calculation.
  
- (v) The dendrogram gives closest similarity to 15 and 16; this agrees with the plot of PC3 against PC4, but not PC1 and PC2.

Also 18 and 25 are very 'similar' in the cluster analysis, but 25 is rather an outlier on the PC plots. 30 is the last to come in to the clusters; this is similar to PC3/PC4 but not PC1/PC2. On the other hand, 31 joins a cluster quite soon but is an outlier on both PC plots. 10, 22, 30 form a distinctive cluster, they are also separated from most other points on the PC3/PC4 plot but not on PC1/PC2.

In summary, the dendrogram and PC3/PC4 plot give some of the same information; but PC1/PC2 matches neither.

Graduate Diploma Module 5, Specimen Paper A. Question 2

- (i) The test statistic is  $T^2 = n(\bar{\mathbf{x}} - \boldsymbol{\mu}_0)' \mathbf{S}^{-1}(\bar{\mathbf{x}} - \boldsymbol{\mu}_0)$  with the usual notation.

The null hypothesis  $H_0 : \boldsymbol{\mu} = \boldsymbol{\mu}_0$  is rejected for large values of the statistic.

It is known as Hotelling's one-sample  $T^2$ .

- (ii) For  $p = 1$ ,

$$\frac{(n-p)}{p(n-1)} T^2 = \frac{(n-1)n(\bar{x} - \mu_0)^2}{1(n-1)s^2} = \left[ \frac{(\bar{x} - \mu_0)}{s/\sqrt{n}} \right]^2 = t^2,$$

where  $t$  is the usual one-sample  $t$  statistic.

In the usual one-sample  $t$  test,  $H_0$  is rejected for  $|t| > t_{n-1; \alpha}$ , where  $t_{n-1; \alpha}$  denotes the customary two-tail critical point for a given significance level. This is equivalent to rejecting  $H_0$  for  $t^2 > t_{n-1; \alpha}^2 = F_{1, n-1; \alpha}$  (in corresponding notation for the  $F$  distribution).

- (iii) The acceptance region for  $H_0$  versus  $H_1$  for a test at significance level  $\alpha$  is

$$n(\bar{\mathbf{x}} - \boldsymbol{\mu}_0)' \mathbf{S}^{-1}(\bar{\mathbf{x}} - \boldsymbol{\mu}_0) \leq \frac{p(n-1)}{(n-p)} F_{p, n-p; \alpha}.$$

A corresponding confidence region with confidence level  $(1 - \alpha)$  consists of all  $\boldsymbol{\mu}_0$  satisfying this inequality. This is an ellipsoidal region centred at  $\boldsymbol{\mu}_0$ .

- (iv) We have  $\bar{\mathbf{x}} - \boldsymbol{\mu}_0 = \begin{pmatrix} -10 \\ -15 \end{pmatrix}$  and  $\mathbf{S}^{-1} = \frac{1}{3600} \begin{pmatrix} 100 & -80 \\ -80 & 100 \end{pmatrix} = \begin{pmatrix} \frac{1}{36} & -\frac{1}{45} \\ -\frac{1}{45} & \frac{1}{36} \end{pmatrix}$

$$\text{So } T^2 = 16 \left[ \frac{100}{36} + \frac{225}{36} - \frac{300}{45} \right] = 37.78 \quad \text{and} \quad \frac{(n-p)}{p(n-1)} T^2 = \frac{14}{30} \times 37.78 = 17.63.$$

Comparing this with  $F_{2,14}$ , the value is well beyond any of the usual critical values, so there is very strong evidence against  $H_0$ , i.e. that the observed data for Down's syndrome babies are not consistent with the expected values for non-Down's syndrome babies.

Graduate Diploma Module 5, Specimen Paper A. Question 3

- (a) Consider a population of homogeneous units which are being studied to observe when they cease to operate correctly, i.e. their *failure time*,  $T$ . The (cumulative) distribution function (cdf) of  $T$  is  $F(t) = P(T \leq t)$ . The *survivor function* is  $S(t) = P(T > t) = 1 - F(t)$ .

Assuming  $T$  is a continuous variable, its pdf is  $f(t) = \frac{dF(t)}{dt} = -\frac{dS(t)}{dt}$ .

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

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

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

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

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

We note that  $h(t) = -\frac{S'(t)}{S(t)} = -\frac{d}{dt} \log S(t)$ .

Using this for the Weibull distribution with  $h(t) = \lambda \gamma t^{\gamma-1}$ , we obtain

$$S(t) = \exp\left(-\int_0^t \lambda \gamma u^{\gamma-1} du\right) = \exp(-\lambda t^\gamma).$$

Hence the pdf is  $h(t)S(t) = \lambda \gamma t^{\gamma-1} \exp(-\lambda t^\gamma)$ .

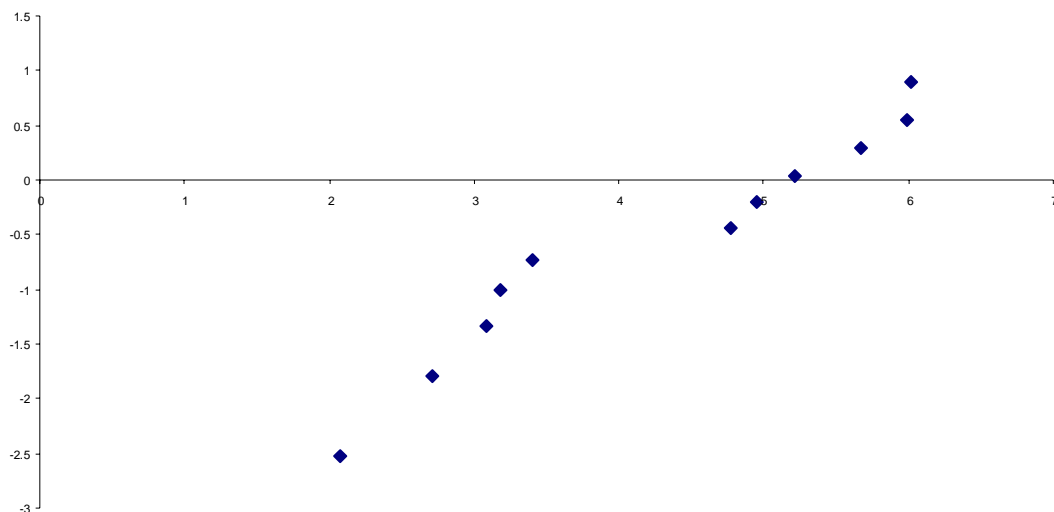
- (b) (i) The cumulative hazard function is  $H(t) = -\log S(t)$  so, for the Weibull distribution [see part (a)], we have  $H(t) = \lambda t^\gamma$ . Hence, for this case, we have  $\log(\text{cumulative hazard function}) = \log(-\log(S(t))) = \log \lambda + \gamma \log t$ . So, if a Weibull model fits the data, a plot of values of  $\log(-\log(S(t)))$  against  $\log t$  should be approximately a straight line. The table in the question shows the (estimated) value of  $S(t)$  for each  $t$  in the column headed "Cumulative Survival".

**Solution continued on next page**

Thus the data give the following, omitting the censored observation.

$\log t = x$	$-\log S(t)$	$\log(-\log S(t)) = y$
2.079	0.0800	-2.5257
2.708	0.1670	-1.7898
3.091	0.2624	-1.3379
3.178	0.3677	-1.0005
3.401	0.4855	-0.7226
4.779	0.6396	-0.4469
4.949	0.8219	-0.1961
5.220	1.0453	0.0443
5.677	1.3329	0.2874
5.996	1.7384	0.5530
6.103	2.4316	0.8885
6.284	—	

The graph ( $y$  against  $x$ ) is as follows.



A straight line can be fitted, although there is some suggestion of two different lines being needed, one for the shorter times and one for the longer. However, a Weibull model *might* be adequate for the whole set.

- (ii) The slope of the fitted line is about 0.7, so the estimate of  $\gamma$  is approximately 0.7.

The intercept on the  $y$ -axis is about  $-3.6$ , so the estimate of  $\lambda$  is approximately  $e^{-3.6} = 0.027$ .

Graduate Diploma Module 5, Specimen Paper A. Question 4

[Solution continues on next two pages]

- (i) The survival time of an individual is *censored* when the end-point of interest (healing 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). In the example, the three observations of 52 (weeks) fall into the first of these categories as the study is terminated after one year, while the observation of 11 falls into the second.

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.

The *hazard function* is closely related to the characteristics of survival. It is a function of time  $t$  giving the risk of reaching the end-point of interest (often "dying") in a very short time interval after  $t$ , assuming survival up to  $t$ . It can be interpreted as the risk of "dying" at time  $t$ .

- (ii) The Kaplan-Meier survival curve is constructed as follows. The word "death" is used in this description generically; here, of course, it refers to time to healing of the leg ulcer.

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

The Kaplan-Meier method requires the ordered death times  $t_{(1)}, t_{(2)}, \dots, t_{(r)}$  to be considered. For  $j = 1, 2, \dots, r$ , 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  $\frac{n_{(j)} - d_{(j)}}{n_{(j)}}$ .

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.

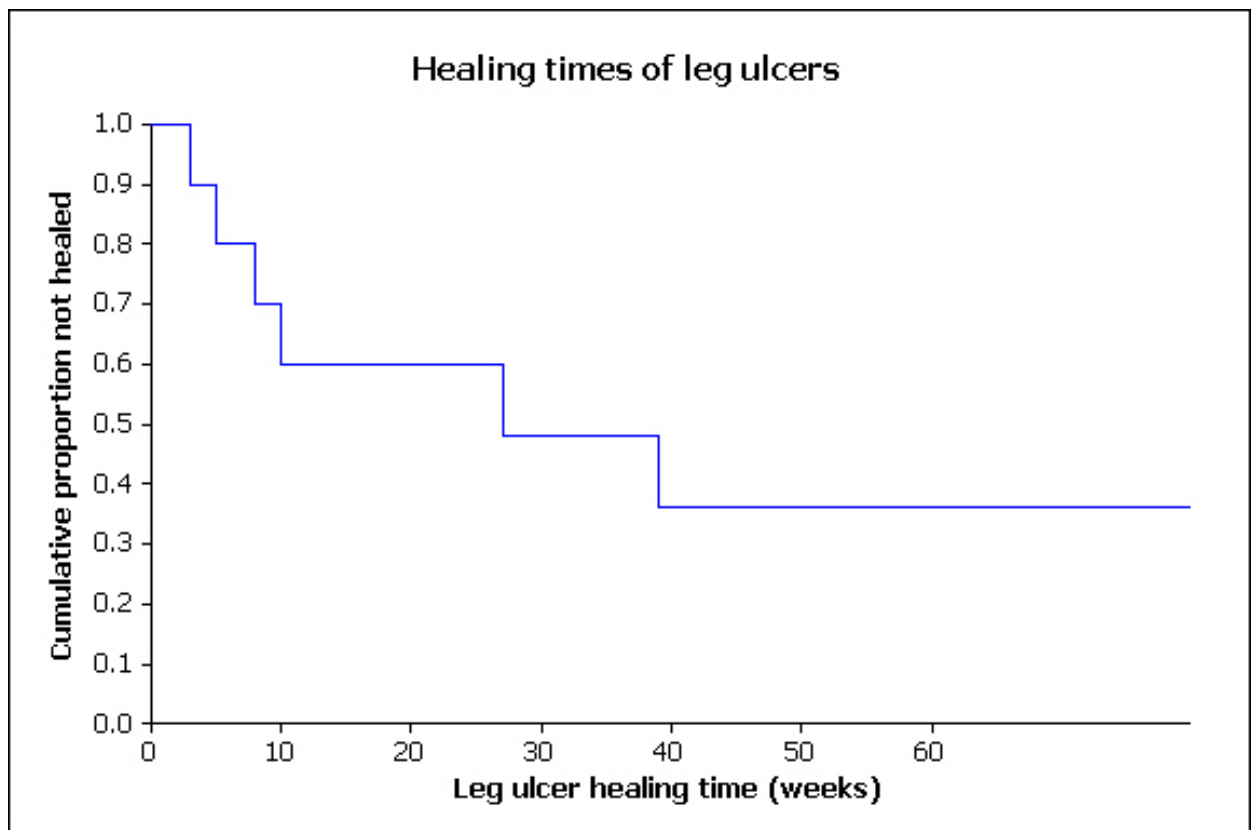
In the present example,  $t_{(r)}$  is censored. The calculation is shown in detail on the next page.

There are 10 patients. The data (healing times) are as follows, where right-censored observations are indicated by \* against the times.

3 5 8 10 11\* 27 39 52\* 52\* 52\*

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. The rows for the censored observations (at times 11 and 52) might be omitted, but care must be taken to ensure that  $n_{(j)}$  is always correct.

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 $t_{(j)}$
3	10	1	9/10	0.900
5	9	1	8/9	0.800
8	8	1	7/8	0.700
10	7	1	6/7	0.600
11	6	–		
27	5	1	4/5	0.480
39	4	1	3/4	0.360
52	3	–		



- (iii) [Note that from this point onwards the solution refers to the results for the full trial shown in the question, not to the results for the sample shown above.]

50% survival is reached at 20 weeks for the Intervention group and 45 weeks for the Control group (NB this is the end of the horizontal line in the Kaplan-Meier step function).

The  $p$ -value of the log rank test statistic is  $<0.05$ , so we may reject the null hypothesis of no difference between groups as regards survival patterns; it appears that there is a difference. The Control group appears to show that half of the "Control" patients were very slow to heal.

- (iv) (a) In analysis 2, we see that only base area and group have  $p$ -values that indicate association of these factors with healing time.

The groups are defined by coding Intervention as 1 and Control as 0. So we see that the "hazard" of ulcer healing is 1.65 times higher in the Intervention group than in the Control group when other factors have been allowed for (the 95% confidence interval for this hazard ratio is 1.16 to 2.36).

Base area is important (lower base area is of course good), but the other factors appear not to be.

- (b) The hazard ratio in the simple model (i.e. in analysis 1) gives a similar result to that in part (iv)(a): 1.48 instead of 1.65, and with fairly similar  $p$ -values. But the fit achieved by including the other variables, as in analysis 2, is much better, with an improvement in the chi-squared value of  $48.25 - 4.85 = 43.4$  on 4 degrees of freedom ( $p$ -value extremely small).

Graduate Diploma Module 5, Specimen Paper A. Question 5  
[solution continues on next page]

In the table below,

${}_{10}q_x$  = 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 ( $= \sum_{y \geq x} {}_{10}L_y$ ).

Hence (part (i) of the question) the age distribution ( $= \frac{100({}_{10}L_x)}{68040}$  %) 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 –
%	14.48	14.20	14.04	13.79	13.33	12.19	9.72	5.85	2.09	0.29	0.01

Age ( $x$ )	${}_{10}q_x$	$l_x$	${}_{10}d_x$	${}_{10}L_x$	$T_x$
0	0.029	1000	29	9855	68040
10	0.009	971	9	9665	58185
20	0.015	962	14	9550	48520
30	0.020	948	19	9385	38970
40	0.047	929	44	9070	29585
50	0.125	885	111	8295	20515
60	0.291	774	225	6615	12220
70	0.551	549	302	3980	5605
80	0.846	247	209	1425	1625
90	0.979	38	37	195	200
100	1.000	1	1	5	5
110		0		0	0

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

Age 20 : $20 + (48520/962) = 70.44$	Age 90 : $90 + (200/38) = 95.26$
Age 40 : $40 + (29585/929) = 71.85$	Age 100 : $100 + (5/1) = 105.00$
Age 60 : $60 + (12220/774) = 75.79$	

(iii) The life expectancy is the expected age at death if at present of age 0, i.e.

$$0 + (68040/1000) = 68.04.$$

In the abridged table used, it is assumed that deaths occur uniformly throughout each 10-year age group. Clearly this is not true, in the older age groups especially (also for the 0 – 10 group, no doubt), and the results from the unabridged table will be much more accurate in these groups. There is also some effect on the overall life expectancy figure.

If there is an annual growth rate of 1% in addition, the age distribution in 10-year intervals is calculated from

$$\frac{(1+0.01)^{-(x+5)} \times {}_{10}L_x}{\sum_{\text{all groups}} (1+0.01)^{-(x_i+5)} \times {}_{10}L_{x_i}} \times 100\% .$$

Because of increasing birth rate, there will be an increase in the proportions in lower age groups as compared with the original population.

Graduate Diploma Module 5, Specimen Paper A. Question 6

[Solution continues on next page]

- (i) Sensitivity is the proportion of positives that are correctly identified by a test:

$$\frac{\text{number who are disease-positive and test-positive}}{\text{number who are disease-positive}}$$

(in the context of testing for a disease).

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

Specificity is the proportion of negatives that are correctly identified by a test:

$$\frac{\text{number who are disease-negative and test-negative}}{\text{number who are disease-negative}}.$$

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

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).

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).

Note.

Alternative expressions for PPV and NPV in terms of the prevalence of the condition and the sensitivity and specificity of the test are

$$\text{PPV} = \frac{\text{prevalence} \times \text{sensitivity}}{(\text{prevalence} \times \text{sensitivity}) + \{(1 - \text{prevalence})(1 - \text{specificity})\}}$$

and

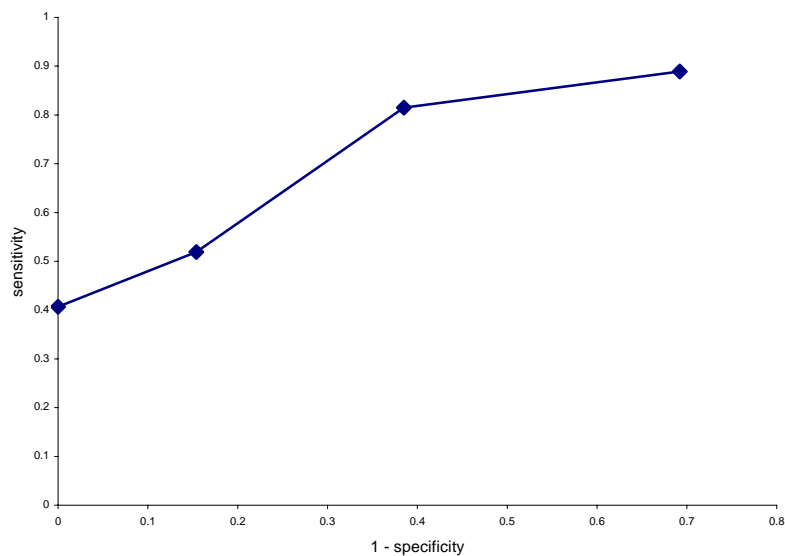
$$\text{NPV} = \frac{(1 - \text{prevalence}) \times \text{specificity}}{\{(1 - \text{prevalence}) \times \text{specificity}\} + \{\text{prevalence} \times (1 - \text{sensitivity})\}}.$$

Sensitivity and specificity are useful statistics because (unlike PPV and NPV, and again using the context of testing for a disease) they give consistent results for the diagnostic test in several patient groups with different disease prevalences. Sensitivity and specificity are characteristics of the test, not of the population to which the test is applied. For a very rare disease they are virtually independent of prevalence, though the accuracy of estimation of sensitivity will be poor.

(ii)

<i>FEV1 value</i>	<i>Sensitivity</i>	<i>Specificity</i>
< 60%	11/27 = 0.407	13/13 = 1.000
< 70%	14/27 = 0.519	11/13 = 0.846
< 80%	22/27 = 0.815	8/13 = 0.615
< 90%	24/27 = 0.889	4/13 = 0.308

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



The "80%" test, i.e. declaring people with FEV1 value  $\geq 80\%$  to be free of pneumoconiosis, might be an acceptable balance. As is illustrated here, high sensitivity has to be balanced against high specificity. The nature of the compromise will be determined with reference to the consequences of false positive or false negative declarations by the diagnostic test.

If the sensitivity were 1, this would mean that the test recognises all people who have the disease as such; this, though desirable, is of course unattainable, but the "80%" test does have quite high sensitivity.

On the other hand, specificity of 1 would mean that all people without the disease are recognised as such, which is also desirable. Again it is of course unattainable, but it is moderately high for the "80%" test. It would almost certainly be regarded as too low for the "90%" test. But moving to the "70%" test to increase specificity leads to unacceptably low sensitivity.

Graduate Diploma Module 5, Specimen Paper A. Question 7

(i) In stratified sampling, a population is divided into groups (strata) and the strata each have a simple random sample taken from them. If proportional allocation is used, the fraction of the stratum population that is sampled is the same in every stratum, i.e.  $\frac{n_i}{N_i}$  is the same for all  $i$ . So it is here, equal to  $\frac{1}{6}$ .

(ii) Simple random samples within strata yield unbiased estimates of means,  $\bar{y}_i$ . Weighting these by stratum sizes gives the overall estimate  $\bar{y}_{st} = \frac{1}{N} \sum_{i=1}^M N_i \bar{y}_i$  ( $M = 4$ , the number of strata).

$$\text{Var}(\bar{y}_{st}) = \frac{1}{N^2} \sum_{i=1}^M N_i^2 \text{Var}(\bar{y}_i) = \frac{1}{N^2} \sum_{i=1}^M N_i^2 \frac{S_i^2(1-f_i)}{n_i} ; \text{ SE is square root.}$$

(This simplifies when proportional allocation is used.)

(iii)  $N = 120$ .

$$\bar{y}_{st} = \frac{1}{120} \{(24 \times 99.25) + (36 \times 100.0) + (30 \times 98.0) + (30 \times 100.0)\} = \frac{11922.0}{120} = 99.35.$$

$$\begin{aligned} & \text{Var}(\bar{y}_{st}) \\ &= \frac{1}{120^2} \left[ \left\{ \frac{24^2 \times 9^2}{4} \times \frac{5}{6} \right\} + \left\{ \frac{36^2 \times 7.46^2}{6} \times \frac{5}{6} \right\} + \left\{ \frac{30^2 \times 6.28^2}{5} \times \frac{5}{6} \right\} + \left\{ \frac{30^2 \times 10.61^2}{5} \times \frac{5}{6} \right\} \right] \end{aligned}$$

$$= 2.9541, \text{ and hence SE} = 1.719.$$

An approximate 95% confidence interval for the true mean is given by  $99.35 \pm 1.96 \times 1.719$ , i.e.  $99.35 \pm 3.37$ , i.e. (95.98, 102.72).

**Solution continued on next page**

(iv)  $\text{Var}(\bar{y})$  by simple random sample  $= \frac{5}{6} \times \frac{7.75^2}{20} = 2.5026$ .

The confidence interval is now  $99.35 \pm 1.96\sqrt{2.5026}$ , i.e.  $99.35 \pm 3.10$ , i.e. (96.25, 102.45).

(v) Ratio of variances = efficiency  $= \frac{\text{Var}(\bar{y})}{\text{Var}(\bar{y}_{st})} = \frac{2.5026}{2.9541} = 0.847$  (or 84.7%).

Stratification may not have been well chosen, since within the same chain the sales can vary greatly. Size of store, reflecting size of turnover, may have been a better choice.

(vi) Strata should be internally homogeneous. Construction can be on the basis of past records of the variable being studied, or of something closely correlated to it. Any major variation should be between strata, not within.

Graduate Diploma Module 5, Specimen Paper A. Question 8

[solution continues on next page]

(i) In random sampling from a population, units are selected by a probability mechanism. Simple random sampling from a finite population gives every item the same probability of selection, but in less simple methods the probabilities need not be the same. For example, in stratified sampling a random sample is taken from each stratum, but the strata are usually of different sizes. Other methods include cluster sampling and multi-stage sampling, in which primary units (for example geographical units such as villages) are selected at random from all those available and these units are either studied completely or subsampled.

Exact estimation methods for means, totals or proportions can be developed for a method of sampling that is based on probability rules. However, these sampling methods require setting up carefully and this can be very time-consuming and expensive.

Non-random sampling methods are usually much quicker, particularly quota sampling in which interviewers are typically sent to central points, such as shopping areas, and given a quota of people to be interviewed. These are specified by characteristics, such as age-group or sex or voting intentions, which can be discovered by a few simple questions so that the specified number (quota) in each sub-group of the population can be obtained. There is no restriction on which actual individuals in each sub-group shall be interviewed, and the easiest to obtain (the most co-operative) will usually be included in the sample. Bias often results from this, and usually also the population to be found in the shopping area (if that is indeed the situation) at the time of the survey is not representative of the whole population of the town. Analysis has to use the methods based on probability because no others are available.

Systematic sampling is done from a population whose members are listed in some standard order (such as alphabetical). It consists of choosing a random starting point at the beginning of the list followed by a regular selection of every  $k$ th item, where  $k = N/n = (\text{population size})/(\text{sample size})$ . Systematic sampling (with random starting point) is much quicker and simpler than pure random sampling. There may be refusals, as in any method of choosing individuals, but this is so in random sampling also. Provided enough is known about possible regular trends in the list used, this method does have a reasonable theoretical base. If there are no trends, a systematic sample might behave as if it were a simple random sample, though strictly speaking it is not. Sometimes the methods for cluster samples can be used for analysis, if there are no trends.

(ii) If  $n$  members are selected at random from  $N$ , without replacement, the population variance (defined as  $\frac{1}{N-1} \sum_{i=1}^N (X_i - \text{population mean})^2$ ) is estimated by

$$s^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2 .$$

For the case of a binary variable, each  $x_i$  will be either 0 or 1 according as the characteristic being studied is absent or present. Suppose we take a sample of size  $n$  and find  $r$  individuals with the characteristic, so that  $r/n$  is the sample proportion with the characteristic. Then we will have

$$\sum x_i = r.1 + (n-r).0 = r \quad \text{and} \quad \sum x_i^2 = r.1^2 + (n-r).0^2 = r.$$

Therefore

$$s^2 = \frac{1}{n-1} \left\{ \sum x_i^2 - \frac{(\sum x_i)^2}{n} \right\} = \frac{1}{n-1} \left( r - \frac{r^2}{n} \right) = \frac{r}{n-1} \left( 1 - \frac{r}{n} \right),$$

and now writing  $p = r/n$  we have  $s^2 = np(1-p)/(n-1)$ , as required.

A 95% confidence interval for the population mean is  $\bar{x} \pm 1.96s/\sqrt{n}$ , assuming  $n$  is fairly large. Hence  $1.96s/\sqrt{n} \leq 1.5$ , giving  $\frac{1.96}{1.5} \leq \sqrt{\frac{n}{168.33}}$ , from which we obtain  $n \geq 168.33 \times 1.7074 = 287.4$ .

Similarly, a 95% confidence interval for the population proportion is

$$\hat{p} \pm 1.96 \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}, \text{ so we require } 1.96 \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \leq 0.04, \text{ and from this we obtain}$$

$$1.96 \sqrt{\frac{0.36 \times 0.64}{n}} \leq 0.04 \text{ so that } n \geq \frac{(1.96)^2 \times 0.36 \times 0.64}{(0.04)^2} = 553.19.$$

Thus we need  $n$  at least 554.

(iii) If a population (e.g. a geographical region) can be split into clusters (e.g. towns, villages), sampling can be based on these clusters. Either a random sample of clusters is chosen and these are studied completely, which is "one-stage", or a sub-sample of units may be taken at random for study from each chosen cluster, which is "two-stage". The sample of clusters could be simple random, stratified random or systematic with random starting point.

Stratified sampling splits a population into various groups, according to some specified characteristic such as urban or rural areas, which are expected to be relatively homogeneous within themselves – which clusters might not be. Stratified sampling requires a complete listing of the whole population, whereas cluster sampling only requires that for the chosen clusters (and of course an initial list of clusters). Cluster sampling is often used for administrative convenience, in limiting the area that is to be covered, and in reducing costs, while stratified sampling aims to give a precise estimate of the population parameters through careful choice of homogeneous strata; cluster sampling might not give any better precision than simple random sampling.

In the UK, the Family Expenditure Survey stratifies into quite large geographical areas (by postcode) and uses cluster sampling to locate different communities within the areas.