THE ROYAL STATISTICAL SOCIETY

2009 EXAMINATIONS – SOLUTIONS

HIGHER CERTIFICATE

MODULE 1

DATA COLLECTION AND INTERPRETATION

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. $\log_{10}$.

Higher Certificate, Module 1, 2009. Question 1

(i) In open-ended questions the researcher might obtain unexpected answers or pick up extreme values. Also, the researcher can get opinions more easily than when a similar question is asked in closed form.

(ii) Reasons for preferring closed questions include [only three were required in the examination] that they are easier to code, easier and quicker for respondents to answer, easier to analyse, easier for interviewer to record answers.

(iii) 1. Who pays for your mobile phone calls? ……..

2. How often do you use a mobile phone?

Several times every day	
Once every one to two days	
Once or twice a week	
Less than once a week	

3. How long do your calls usually last? ……

4. Why do you make calls on a mobile phone? Select as many responses as apply.

To chat with my friends	
To contact my parents	
To call a taxi	
To obtain information e.g. the times of buses	

5. Approximately what proportion of calls are ones to you rather than ones that you make? Select the response that is nearest.

None	
About a quarter	
About a half	
About three-quarters	
All	

6. Approximately what percentage of calls that you make are text calls? …….

Questions 1, 3 and 6 are open; 2, 4 and 5 are closed. Qu 4 can be criticised for not including an "Other" category; it can still be regarded as closed if it has an "Other" category, but "Other (please specify)" could be regarded as making it open.

Higher Certificate, Module 1, 2009. Question 2

Part (a)

(i) If the digits had been chosen at random then we would expect each digit to occur approximately 25×228/10 = 570 times. Clearly this is not the case. 0 is of very low frequency of occurrence. Digits 1 to 4 are chosen more frequently than would be expected, especially noticeably so for digit 3. Digits 5, 6, 8 and 9 are chosen noticeably less frequently than expected. Only digit 7 is near (actually it is very near) to the expected frequency of occurrence.

(ii)

Digit (x)	Frequency (f)	fx	x^2	fx^2
0	375	0	0	0
1	637	637	1	637
2	676	1352	4	2704
3	709	2127	9	6381
4	607	2428	16	9712
5	522	2610	25	13050
6	534	3204	36	19224
7	573	4011	49	28077
8	515	4120	64	32960
9	552	4968	81	44712
	5700	25457		157457

Sample mean = 25457/5700 = 4.466.

$$\text{Sample variance} = \frac{157457 - \frac{25457^2}{5700}}{5699} = 7.6790.$$

(iii) The sample mean is near to the supposed population mean of 4.5, but the sample variance is distinctly smaller than the supposed population variance of 8.25. This is largely due to the low observed frequency of occurrence of 0.

(iv) Another property one might look for is whether each digit is followed by each of the ten digits approximately 1/10 of the time.

Part (b)

3-digit random numbers are required. One method which avoids some "wastage" of the generated random numbers (which might be important in a large simulation, for example) would be to number the elements of the population from 001 through to 350, and then again from 351 through to 700, so that each element of the population is allocated two numbers, x and $x + 350$. 3-digit random numbers are then generated; 000 does not lead to selection of any member of the population, nor does any number from 701 through to 999.

Any allocation giving either exactly one choice or exactly two choices to the 350 members of the population is an acceptable procedure.

Higher Certificate, Module 1, 2009. Question 3

Part (a)

Preliminary calculations:

Age group	Class mark	Men			Women		
		Frequency	Freq density	%	Frequency	Freq density	%
19–24	22	61	10.17	8.0	78	13.00	8.1
25–34	30	160	16.00	20.9	211	21.10	22.0
35–54	45	393	19.65	51.3	486	24.30	50.7
55–64	60	152	15.20	19.8	183	18.30	19.1

Note. As the age groups vary in width, the frequency density or % density should be used in frequency polygons or histograms. As the % distributions are almost identical, the two % density polygons would almost coincide if on the same diagram. The diagram below was produced as a joined up scatter diagram in Excel.

For men: $\Sigma fx = 32947$, $\Sigma fx^2 = 1516549$.
Sample mean = 32947/766 = 43.01.
Sample variance = $[1516549 – (32947^2/766)]/765 = 129.99$,
standard deviation = 11.40.

For women: $\Sigma fx = 40896$, $\Sigma fx^2 = 1870602$.
Sample mean = 40896/958 = 42.69.
Sample variance = $[1870602 – (40896^2/958)]/957 = 130.40$,
standard deviation = 11.42.

Median of men's ages = $35 + \{(383 – 221)\times 20/393\} = 43.24$.
Median of women's ages = $35 + \{(479 – 289)\times 20/486\} = 42.82$.

Region	Total	Men	% men in region	Women	% women in region	% for region in sample	Angle for pie chart
Scotland and North	574	248	43.2	326	56.8	33.3	120
Central, South West and Wales	624	274	43.9	350	56.1	36.2	130
London and South East	526	244	46.4	282	53.6	30.5	110
Total	1724	766	44.4	958	55.6		

Solution continued on next page

Report:

There are more women than men in the sample: out of the total of 1724, 44.4% are men and 55.6% are women.

The numbers of men in the three regions are fairly similar to one another, and this is also true of the numbers of women. This is shown in the bar chart. The relative proportions of men and women in the three regions are close to one another and also close to the overall percentages.

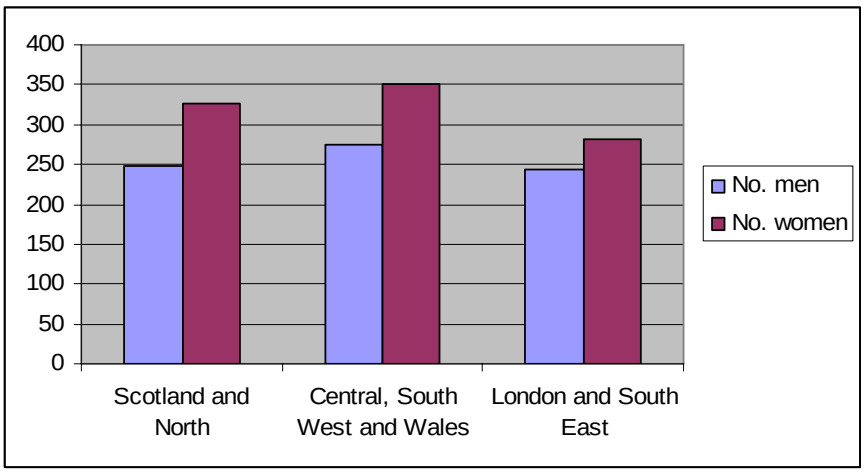
Of the current smokers, 43.0% are men and 57.0% are women.

Of the men, 30.8% are current smokers; of the women, 32.7% are.

The sample sizes in the regions are fairly similar to one another, each region contributing about one third of the overall sample; this is shown in the pie chart.

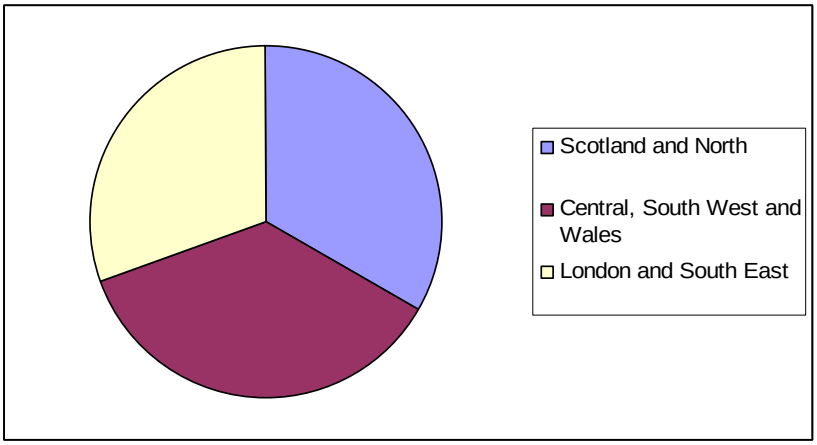
The distributions of ages of the men and women are very similar, as can be shown by frequency density diagrams. On average the men are slightly older than the women, by about one third of a year; the average age of men is 43.0 years and of women is 42.7 years. The pattern for the median ages is similar, that for men being 43.2 years and that for women 42.8 years. The spreads of the ages are approximately the same, with standard deviations for both men and women being 11.4 years (to one d.p.).

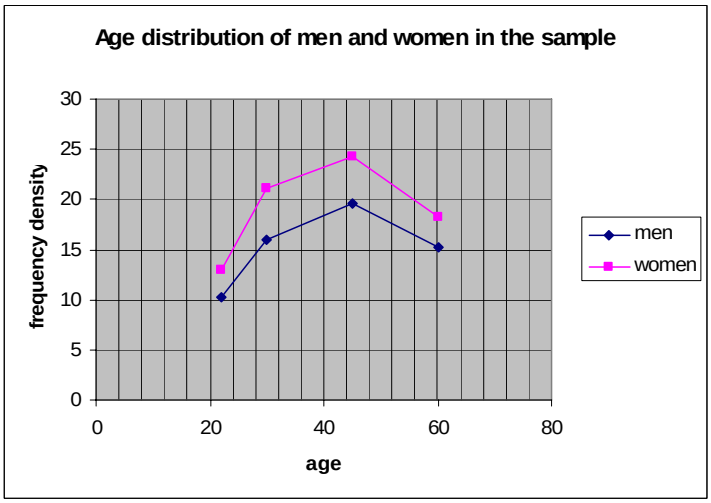
Numbers of men and women in the regions



Solution continued on next page

Breakdown of the overall sample by region





Part (b)

Advantages of collecting dietary information by a diary as described include the following [only four were required in the examination].

Layout should make it easy for respondents to complete.
Should be accurate if completed at time food/drink is consumed.
Need not rely on memory.
Could be fairly easy/cheap to design.
Does not usually involve interviewing costs.

Disadvantages include the following [only four were required in the examination].

Respondents might change eating habits as a result of keeping the diary.
No guarantee that entries are made at the time.
Tedious for respondents to complete.
Diary could get lost.
Low agreement to participate.
High dropout rate.
Could be difficult to compare different respondents.
Coding and analysis could be difficult/time consuming.

Higher Certificate, Module 1, 2009. Question 4

(i) Either a random sampling scheme or a quota sampling scheme would be suitable, or a combination of both. For example, if a list of residents on the estate is available, then taking a simple random sample from this would be suitable. Alternatively, a random sample of addresses could be taken and people living at these asked to participate. If the estate is large, some kind of stratification by type of dwelling (such as flat, terrace, end terrace, semi-detached house, detached house) might be worthwhile before sampling residents. Quota sampling might be done by walking round the streets and knocking on doors, once quotas had been established.

(ii) For example, questions for use in an interview might be as follows.

Do you buy food at the local shops on the estate? Occasionally/often/never

Do you buy food in the town? Occasionally/often/never

Do you use the internet for food shopping? Yes/no

How often do you shop for perishable food items?

Every day	
Once or twice a week	
Once every 7 to 10 days	

Where do you usually shop for these? ……

How often do you shop for non-perishable food items?

Every day	
Once or twice a week	
Once every 7 to 10 days	

Where do you usually shop for these? …….

Do you use a private car when shopping for food? Mostly/Sometimes/Never

If you don't use a private car for food shopping, how do you usually travel back from the shops? On foot/by bicycle/by taxi/by public transport

(iii) Would want all tabulations broken down by sex, age group and household size.

Would want tables showing type of shop (corner shop on estate, in town, out of town, internet, etc) by type of commodity.

For different types of commodity (e.g. food, household cleaning materials, luxury goods) would want a breakdown of geographical location where they usually bought such items by how often they bought them.

Would want tables showing distance travelled and mode of transport used.