THE ROYAL STATISTICAL SOCIETY

2006 EXAMINATIONS – SOLUTIONS

ORDINARY CERTIFICATE

PAPER II

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.

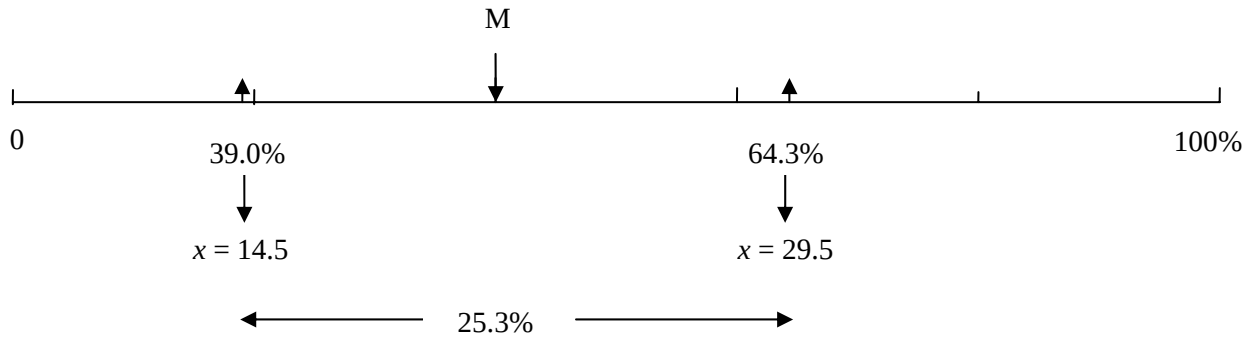
Users of the solutions should always be aware that in many cases there are valid alternative methods. Also, in the many cases where discussion is called for, there may be other valid points that could be made.

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.

Ordinary Certificate, Paper II, 2006. Question 1

(a)



In a large sample, the median can be taken at the 50% frequency point. There is 39.0% of the sample below the value $x = 14.5$ and in the next 15 units of measurement there is 25.3% of the sample. We need to go 11% of this to reach 50%, i.e. $\frac{11.0}{25.3}$ of the way from $x = 14.5$ to 29.5. So $14.5 + \left(\frac{11.0}{25.3} \times 15\right)$ is the required value of the median, i.e. 14.5 + 6.52 = 21.0 (1 d.p.). This assumes that the values between $x = 14.5$ and $x = 29.5$ are uniformly spread.

(b) (i) Again assuming linear interpolation is correct, the critical value for $n = 35$ will be $\frac{1}{2}(55.76 + 43.77) = 49.765$.

(ii)

n	30	35	40
$1/n$	0.03333	0.02857	0.02500
Critical value	43.77		55.76

Using $1/n$ for the interpolation, and the given critical values of the statistic at $n = 30$ and $n = 40$, the value at $n = 35$, where $1/n = 0.02857$, is

$$43.77 + \left(\frac{0.02857 - 0.03333}{0.02500 - 0.03333} \times (55.76 - 43.77)\right)$$

$$= 43.77 + \left(\frac{(-0.00476)}{(-0.00833)} \times 11.99\right)$$

$$= 43.77 + 6.85$$

$$= 50.62 \quad \text{(slightly larger than the value in (b)(i)).}$$

Ordinary Certificate, Paper II, 2006 Question 2

(i) Driver sleeplessness is thought to account for 10% of all UK road accidents. On motorways and dual carriageways, that figure rises to 20%.

(ii) 1 in 10,000 is 0.0001, but as a percentage this is 0.01%.

(iii) The increase as a percentage of the previous year is $\frac{280000-75440}{75440}\times 100$ which is 2.7116×100 i.e. 271%.

This is the only change needed to the quotation.

(iv) The percentages quoted add up to 100. It is not possible for all of them to have increased or remained the same – some must have decreased (as percentages) to balance the increases. Possibly there has been confusion between actual numbers and percentages.

Ordinary Certificate, Paper II, 2005. Question 3

(i) As the data are already ranked in order, an ordered stem and leaf diagram is easily produced by using 01 to 18 as stems and omitting the final digit when constructing the leaves. For example, 1517 simply becomes 15|1 (see the diagram).

Lifetimes (h) of 100 light bulbs

Stem (hundreds)	Leaf (tens)
1	9
2	7
3	6
4	2
5	15
6	27
7	5599
8	4456778
9	479
10	00000223558889
11	12337788
12	23344455667
13	01244456666677
14	000001122334566678899
15	13
16	25789
17	02
18	9

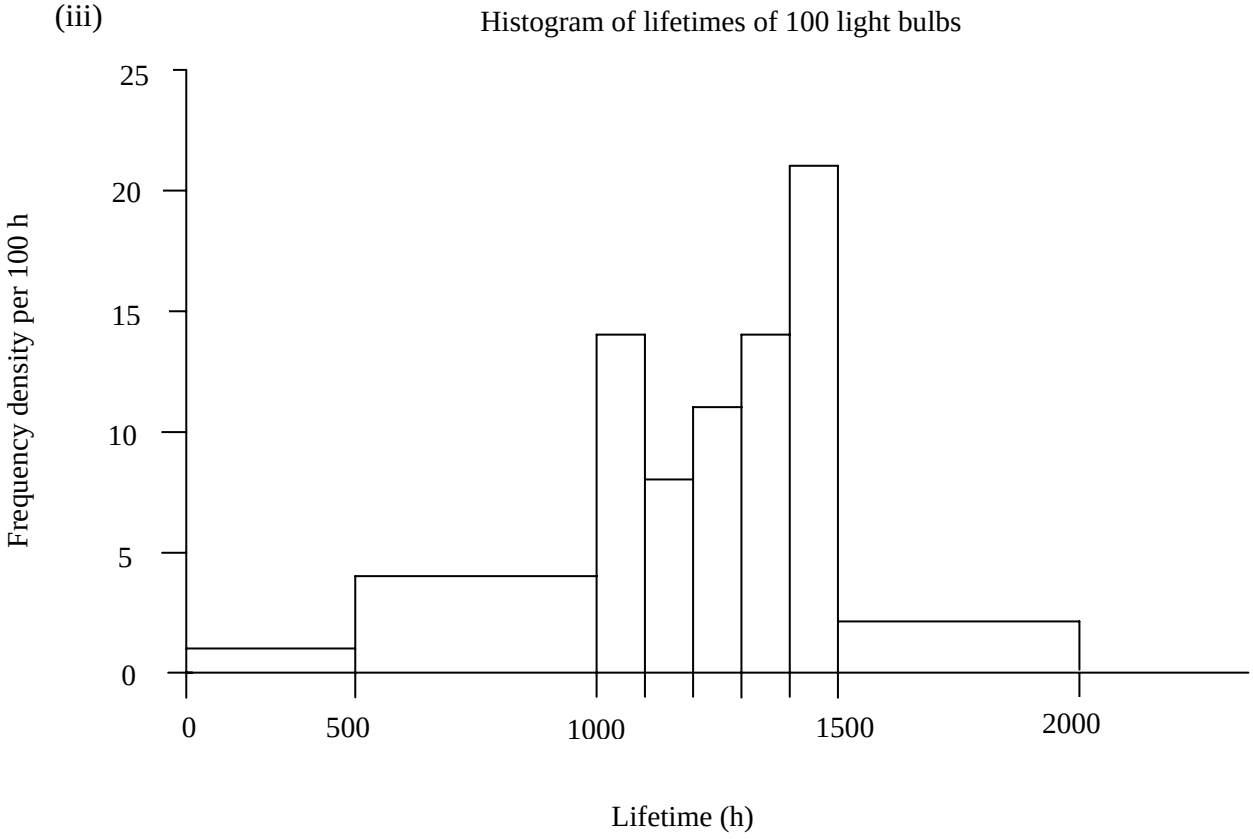
Solution continued on next page

(ii) Grouped frequency distribution of lifetimes of light bulbs:-

Lifetime (h)	*Frequency*	*Density**
0 but less than 500	4	0.8
500 but less than 1000	18	3.6
1000 but less than 1100	14	14
1100 but less than 1200	8	8
1200 but less than 1300	11	11
1300 but less than 1400	14	14
1400 but less than 1500	21	21
1500 but less than 2000	10	2
Total	100	

* Density is frequency density per 100 hours. This column is appended to the table to enable the histogram in part (iii) to be constructed, in which the <u>areas</u> standing on each interval represent frequencies.

(iii)



(iv) The histogram shows the overall shape of the distribution, so giving a general picture of the overall quality of the batch. The stem and leaf diagram also does this, but in addition makes it easy to calculate the median and quartiles; it is therefore more useful. The stem and leaf is more easily incorporated into a routine quality control procedure because it can be built up as data are collected.

Ordinary Certificate, Paper II, 2006. Question 4

(i) $P(A|B)$ is $\dfrac{P(A \text{ and } B)}{P(B)}$, i.e. $\dfrac{P(A \cap B)}{P(B)}$.

(ii) If A and B are independent, then

(a) $P(A \text{ and } B) = P(A)\,P(B)$,

(b) $P(A|B) = P(A)$ [i.e. B gives no information on A].

(iii) (a) Device 1 works only when both X and Y work, so the probability is $\dfrac{3}{4} \times \dfrac{7}{8} = \dfrac{21}{32}$.

(b) Device 2 works except when both X and Y fail. The probabilities of failure are $\dfrac{1}{4}$ and $\dfrac{1}{8}$, so $P(\text{both fail}) = \dfrac{1}{4} \times \dfrac{1}{8} = \dfrac{1}{32}$ and $P(\text{Device 2 works}) = 1 - \dfrac{1}{32} = \dfrac{31}{32}$.

(iv) Suppose Device 1 works. Then X and Y both work, so the probabilities are

(a) 1, (b) 0, (c) 1.

Now suppose Device 2 works. We use notation as follows: X for "X works", Y for "Y works", 2 for "Device 2 works" and Y^C for "Y does not work".

(a) $P(X \mid 2) = \dfrac{P(X \cap 2)}{P(2)}$ [note that the event $X \cap 2$ is the same event as X]

$= \dfrac{P(X)}{P(2)} = \dfrac{3/4}{31/32} = \dfrac{3}{4} \times \dfrac{32}{31} = \dfrac{24}{31}$.

(b) $P(X \cap Y^C \mid 2) = \dfrac{P(X \cap Y^C \cap 2)}{P(2)} = \dfrac{P(X \cap Y^C)}{P(2)} = \dfrac{\frac{3}{4} \times \frac{1}{8}}{\frac{31}{32}} = \dfrac{3}{4} \times \dfrac{1}{8} \times \dfrac{32}{31} = \dfrac{3}{31}$.

(c) $P(X \cap Y \mid 2) = \dfrac{P(X \cap Y \cap 2)}{P(2)} = \dfrac{P(X \cap Y)}{P(2)} = \dfrac{\frac{3}{4} \times \frac{7}{8}}{\frac{31}{32}} = \dfrac{3}{4} \times \dfrac{7}{8} \times \dfrac{32}{31} = \dfrac{21}{31}$.

Ordinary Certificate, Paper II, 2005. Question 5

(i) Rank the profit in order also, then calculate *d*, the difference in ranks for each club.

Profit	1	4	2	3	10	6	9	5	8	7
League position	1	5	2	3	4	10	9	8	7	6
d	0	–1	0	0	6	–4	0	–3	1	1

Spearman's rank correlation coefficient

$$r_s = 1 - \frac{6\sum d^2}{n\left(n^2 - 1\right)} = 1 - \frac{6 \times 64}{10 \times 99} = 1 - 0.388 = 0.612 .$$

This shows a fairly strong positive relation between position and profit. The main exceptions are Chelsea (high in table but a very large loss) and Tottenham Hotspur (a smaller loss than some clubs above them in the table; also true of Southampton).

(ii) League position is an ordinal variable, so Pearson's coefficient is not suitable.

Ordinary Certificate, Paper II, 2005. Question 6

(i) For x = number of heart valve operations, $\sum x = 1790$, $\sum x^2 = 437898$.

The sample is of size $n = 8$ (note this is a sample, not a population, so divisor $n - 1$ is used in calculating the variance and standard deviation), so the sample mean is 1790/8 = 223.75 and the sample variance is

$$\frac{1}{7}\sum\left(x-\bar{x}\right)^2 = \frac{1}{7}\left\{\sum x^2 - \frac{\left(\sum x\right)^2}{8}\right\} = 5340.786$$

and the standard deviation is $\sqrt{(5340.786)} = 73.08$.

[Note. The mean and standard deviation can of course be found directly from calculators, but candidates are *strongly* advised to give an indication of the method being used.]

The coefficient of variation (expressed as a percentage) is $\frac{100 \times \text{sd}}{\text{mean}} = 32.7\%$.

(ii) Similarly for y = number of heart bypass operations, we have $\sum y = 14286$ and $\sum y^2 = 28503366$, leading to mean 1785.75, variance 427448.786 and standard deviation 653.80.

Hence the coefficient of variation is $\frac{65380}{1785.75} = 36.6\%$.

(iii) On average, a far greater number of heart bypass operations is carried out, compared with heart valve operations.

The standard deviation is larger for bypasses, so we could say that the between-hospital variation for this operation is greater than for heart valve operations.

But when the coefficients of variations, which compare the standard deviations with the means, are calculated they are quite similar. So, relative to the means, the variability in the number of operations between hospitals is about the same for the two types of operation.

Ordinary Certificate, Paper II, 2006. Question 7

(i) The graph shows percentage changes on the previous quarter. So the chain base index numbers are read directly from the graph.

Year Qtr	North	South
2003 Q1	–	–
2003 Q2	102.0	103.0
2003 Q3	102.5	103.0
2003 Q4	103.0	103.5
2004 Q1	103.5	102.5
2004 Q2	102.5	101.5
2004 Q3	102.0	101.0
2004 Q4	102.0	100.5

(ii) To obtain fixed index numbers based on 2003 Q1 = 100.0, we "scale up" each chain base index number by the corresponding factor for its successor. For example, North 2003 Q3 is calculated as $102.0 \times (102.5/100.0) = 104.55$ (shown as 104.6 in the table); North 2003 Q4 is then this multiplied by (103.0/100.0); and so on.

Year Qtr	North	South
2003 Q1	100.0	100.0
2003 Q2	102.0	103.0
2003 Q3	104.6	106.1
2003 Q4	107.7	109.8
2004 Q1	111.5	112.5
2004 Q2	114.3	114.2
2004 Q3	116.6	115.3
2004 Q4	118.9	115.9

[Each result is given to 1 d.p. and then used as such. Slightly different numbers are obtained in some quarters if higher precision is used throughout.]

(iii) Overall there was an 18.9% increase in house prices in the North over the two years, and a 15.9% increase in the South.

The peak rate of increase in the South was in 2003 Q4, whereas it was a quarter later in the North.

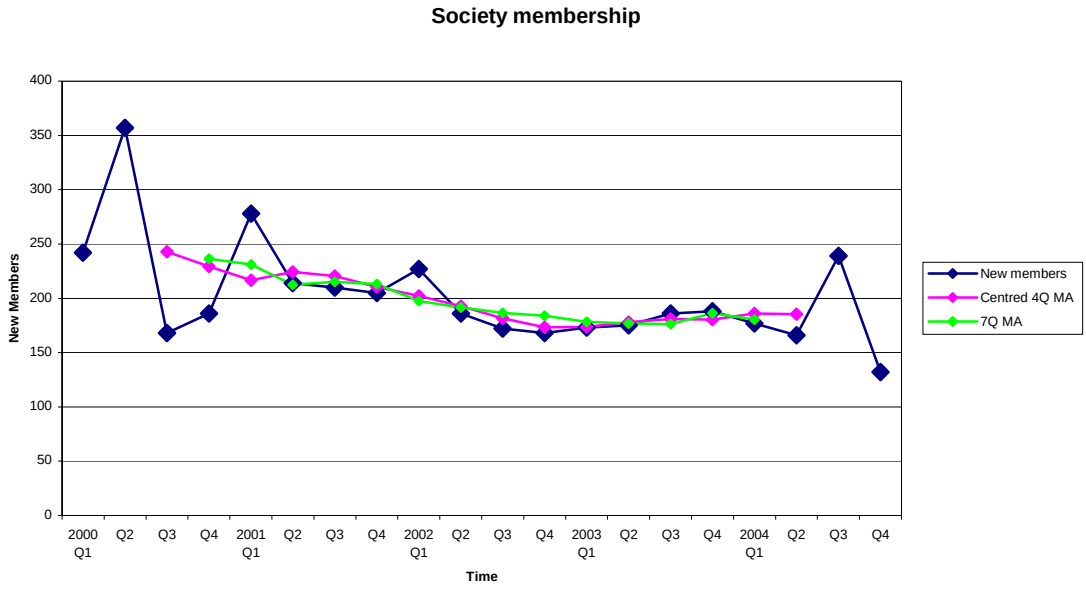
In the North, the rate of increase continued to rise until 2004 Q1, and then began to drop; in the South, there was a more rapid increase until 2003 Q4 and then the rate of increase dropped considerably for the rest of the period.

The rate of increase was greater in the South than in the North in 2003, but the reverse was true in 2004.

Ordinary Certificate, Paper II, 2006. Question 8

In order to draw a useful chart on this scale, a whole sheet of graph paper should be used. The fluctuations in quarterly numbers of new members cause the vertical scale to be inconveniently wide. If it goes down to 0, this wastes space at the bottom end; a "false origin" might be used, but there *must* be a clear indication of it (e.g. by a "break" in the axis).

(i) The line ——◆—— shows the time trend, which is rather irregular but does not seem to have a seasonal component. There may have been a membership campaign occasionally; information would be useful. This could explain the local peaks.



(ii) and (iii) The calculations are shown on the next page. For example, the centred four-quarterly moving average for 2000 Q3 is

$$\frac{1}{2}\left\{\frac{242+357+168+186}{4}+\frac{357+168+186+278}{4}\right\} = \frac{242+2(357+168+186)+278}{8}.$$

The seven-quarterly moving averages do not need centring in this way.

On the graph above, ——◆—— shows the centred four-quarterly moving averages and ——◆—— shows the seven-quarterly moving averages.

Solution continued on next page

(iv) There is some remaining fluctuation in both moving averages, although both suggest that the quarterly number of new members is reducing slowly from the early two quarters. However, the peak in 2004 Q3 causes a slight rise at the end. But the average of this and the next quarter is similar to that for the corresponding quarters in 2003. Since moving averages should not be extrapolated, neither plot for these data is helpful in prediction.

Calculation

Membership numbers			
Time	*New members*	*Centred 4Q MA*	*7Q MA*
2000 Q1	242		
Q2	357		
Q3	168	242.8	
Q4	186	229.4	236.4
2001 Q1	278	216.8	231.1
Q2	214	224.4	212.6
Q3	210	220.4	215.1
Q4	205	210.5	213.1
2002 Q1	227	202.3	197.4
Q2	186	192.9	191.6
Q3	172	181.5	186.6
Q4	168	173.4	183.9
2003 Q1	173	173.8	178.3
Q2	175	178.0	177.0
Q3	186	181.0	176.1
Q4	188	180.4	186.3
2004 Q1	177	185.9	180.4
Q2	166	185.5	
Q3	239		
Q4	132		