THE ROYAL STATISTICAL SOCIETY

2002 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, 2002. Question 1

(i)

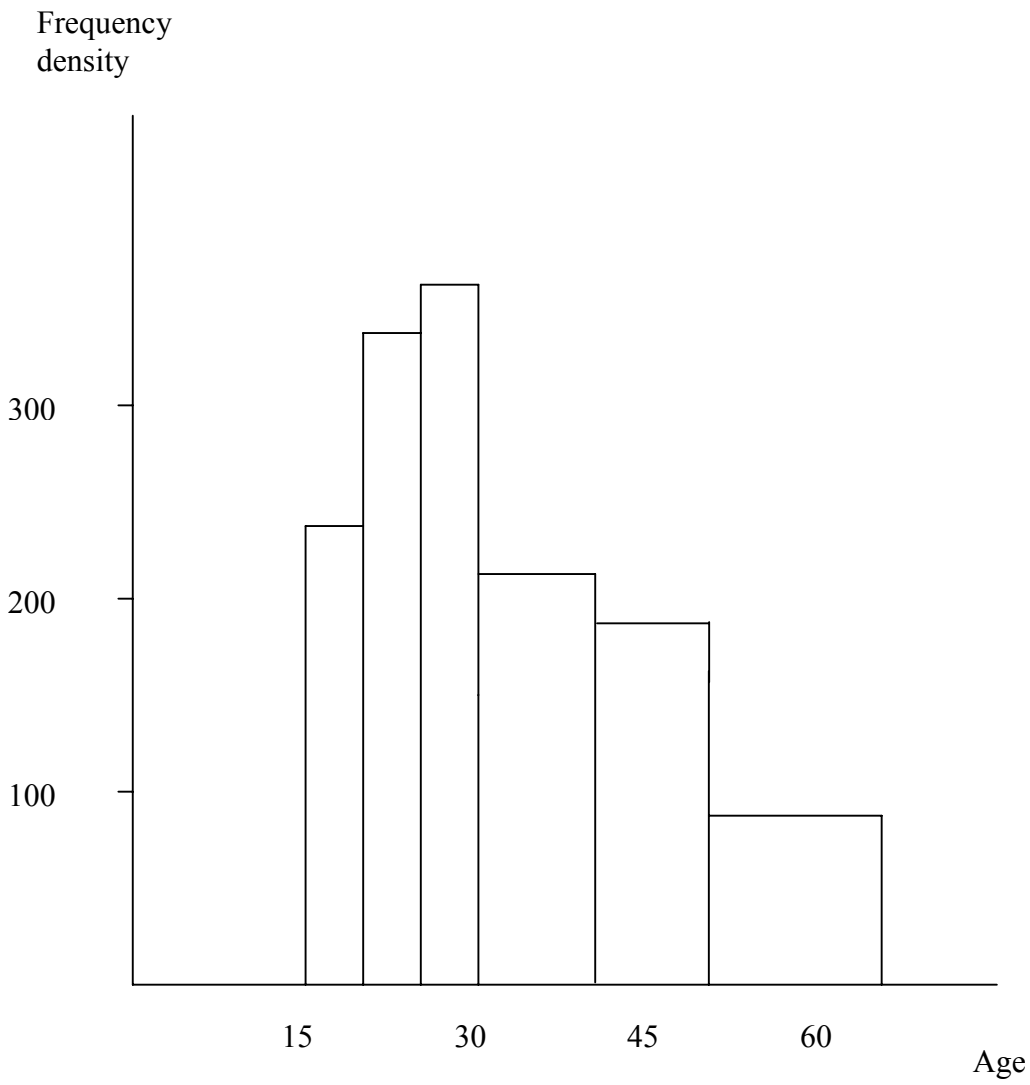
Age last birthday (years)	Employees	Frequency Density
15 – 19	240	240
20 – 24	340	340
25 – 29	360	360
30 – 39	420	210
40 – 49	380	190
50 – 64	240	80

Frequency density is frequency per 5-year interval.

See next page for histogram.

(ii) The histogram is peaked between 20 and 30, and is quite skew to the right. The greatest density of employees is at a younger age, so one possible explanation of this could be "ageist", although there might well be other explanations such as experienced people finding better-paid jobs elsewhere.

(iii) The original table used unequal intervals of age, and it is not until a frequency density based on intervals of the same size (5 years) is calculated that the pattern is easy to see. The difference between absolute frequency and frequency density is unlikely to be known to the casual observer.


Frequency
density
300
200
100
15
30
45
60
Age

Ordinary Certificate, Paper II, 2002. Question 2

The dispersion in a set of data is the variation among the set of data values. It measures whether they are all close together, or more scattered.

Range is the difference between the largest and smallest values in the data set. It is very easy to calculate, but it depends only on the largest and smallest values which may sometimes be extremes or outliers; it does not use the whole pattern including the central values.

Inter-Quartile Range (IQR) is the difference in value between the upper quartile (Q_3) and the lower quartile (Q_1). Q_3 has 75% of the total number of observations below it, Q_1 has 25%. (The semi-inter-quartile range, SIQR, is $\frac{1}{2}(Q_3 - Q_1)$, and is often used instead).

Once the data are arranged in rank order, Q_1 and Q_3 are easy to locate, and the value $(Q_3 - Q_1)$ is not affected by the most extreme data values. However, it is not easy to develop theory for using these measures in mathematical statistical methods.

Variance is an "average" deviation from the mean, squared. For a sample, the definition is $s^2 = \frac{1}{(n-1)}\sum(x-\bar{x})^2$; if the data are regarded as an entire population, the divisor is often taken as n rather than $n - 1$. It shows whether data cluster round their mean or are more spread out. It does use all the data values in the calculation but only really gives a good measure when data are fairly symmetrical; extreme values or outliers affect s^2 considerably. This measure is the one for which good mathematical theory exists (based on a Normal distribution), and so it is widely used. The standard deviation, s, is in the same units of measurement as x, so is often preferred.

The variance and range will be affected by the extreme values (if any) at each end of the distribution. The IQR (or SIQR) gives a better idea of the spread of wages in the central part of the wage distribution, so it may be a better measure (even though it does not reflect the whole distribution, especially its ends).

Ordinary Certificate, Paper II, 2002. Question 3

(i) $\bar{x} = \frac{47118}{100} = 471.18$.

$s^2 = \frac{1}{99}\left(30710404 - \frac{(47118)^2}{100}\right) = \frac{8509344.76}{99} = 85952.98$, so $s = 293.18$.

(ii)

Class	Tally	Frequency	Mid-point, x	$\frac{x-499.5}{200} = y$
000 – 999	𝍸 𝍸 𝍸 𝍸 |||	23	99.5	–2
200 - 399	𝍸 𝍸 𝍸 𝍸 |||	23	299.5	–1
400 – 599	𝍸 𝍸 𝍸 ||||	19	499.5	0
600 – 799	𝍸 𝍸 𝍸 |	16	699.5	1
800 - 999	𝍸 𝍸 𝍸 ||||	19	899.5	2
Total		100		

(iii) Using the last two columns in the above table, to give a coded measure y, we have $\Sigma fy = -46 -23 + 0 + 16 + 38 = -15$, so $\bar{y} = -15/100 = -0.15$. Now,

$$\bar{y} = \frac{\bar{x} - 499.5}{200}, \quad \text{so } \bar{x} = 200\bar{y} + 499.5 = -30 + 499.5 = 469.5 .$$

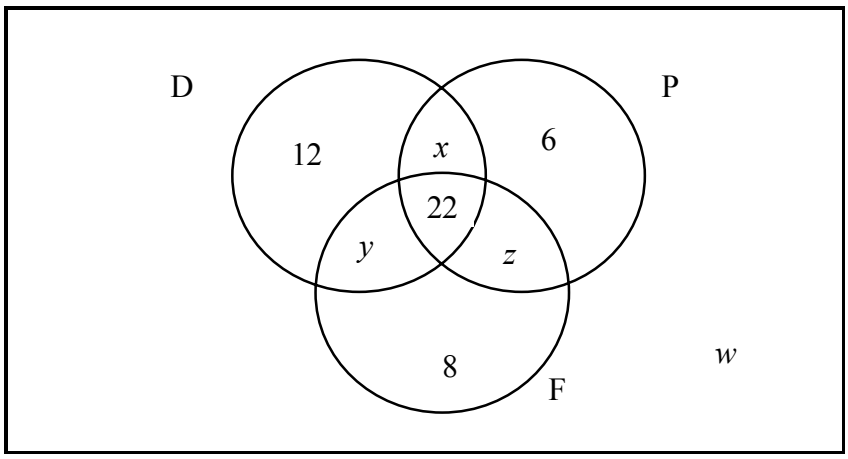
Also, $\Sigma fy^2 = (23\times4) + (23\times1) + 0 + (16\times1) + (19\times4) = 207$.

So variance in coded units = $\frac{1}{99}\left(207 - \frac{(-15)^2}{100}\right) = \frac{204.75}{99} = 2.0682$, and the standard deviation in coded units is 1.438. Hence the standard deviation of x is $200\times1.438 = 287.6$.

(iv) In part (iii) we have had to "group" all the observations at the interval mid-points, assuming uniform spread through each interval. This explains the small differences between results in (i) and (iii); but since the differences are small, the assumption is quite good.

Ordinary Certificate, Paper II, 2002. Question 4

(i)



Given that 62% did tick D, then $12 + x + y + 22 = 62$ or $x + y = 28$. ①

Also 58% did tick P, so $6 + x + z + 22 = 58$ or $x + z = 30$. ②

Finally, 56% did tick F, so $8 + y + z + 22 = 56$ or $y + z = 26$. ③

If w is the % ticking none of D, P, F, then $48 + x + y + z + w = 100$, so $x + y + z + w = 52$.

Adding ①,②,③ gives $2(x + y + z) = 28 + 30 + 26 = 84$, i.e. $x + y + z = 42$, and so $w = 10$.

② – ① gives $z - y = 2$. In ③, put $z = y + 2$, giving $2y + 2 = 26$, i.e. $y = 12$.

Then $x = 16$ and $z = 14$.

(ii)

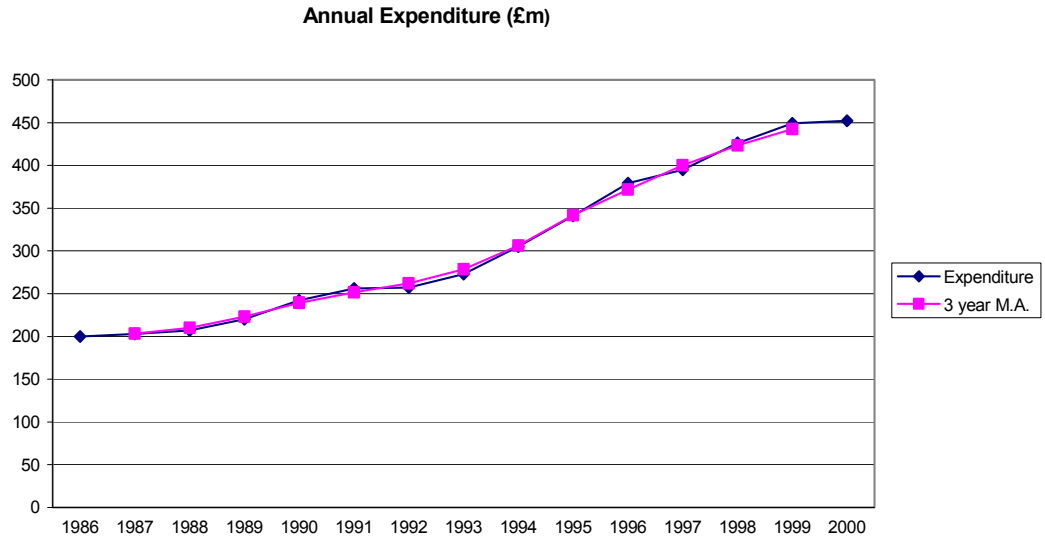
	No factors	1 factor	2 factors	3 factors	Total
	w	$12 + 6 + 8$	$x + y + z$	22	
Probability	0.10	0.26	0.42	0.22	1.00

(iii) Excluding w, the total probability is 0.90, so the conditional probabilities of 1, 2, 3 are

$$\frac{0.26}{0.9} = 0.289, \qquad \frac{0.42}{0.9} = 0.467, \qquad \frac{0.22}{0.9} = 0.244 \ .$$

Ordinary Certificate, Paper II, 2002. Question 5

(i)



(ii)

Year	*Expenditure*	*3-year M.A.*
1986	200	
1987	203	203.33
1988	207	210.00
1989	220	223.00
1990	242	239.33
1991	256	251.67
1992	257	262.00
1993	273	278.33
1994	305	306.33
1995	341	341.67
1996	379	371.67
1997	395	400.00
1998	426	423.33
1999	449	442.33
2000	452	

(iii) A moving average for 2000 requires the 2001 data. Retrospectively, a moving average keeps close to the data curve provided that the data curve is fairly smooth, as it is here. As a forecasting tool, however, a moving average does not pick up changes in trend until some time later – and we have no knowledge whether the trend after 2000 will continue as before. Prediction for a 4-year period 2001–5, using in effect data which end at 1999, is not likely to be reliable.

Ordinary Certificate, Paper II, 2002. Question 6

Diagram 1

(i) There is no indication of what is being measured on the *y*-axis, and no units are given. Therefore we do not know whether the changes are large or small.

(ii) If it is worthy of comment, there seems to be a fairly steady increase in *y* over the 4 years. (If the units are in money, we would like to know whether they have been corrected for inflation.)

Diagram 2

(i) Which is which for 2000/1 – have they crossed over or not?

Note also that the vertical scale does not start at zero.

(ii) The heading implies that they have crossed over, and if so A has gone back to where it started, having initially done much better than B; while B still has a more gentle upward trend.

A key with different symbols for drawing the lines would distinguish between A and B.

Diagram 3

(i) What is the distinction between ── and ---- ? Do we already have the complete year-end data for 2001/2? And what are the ---- lines based on?

(ii) Starting from a very low base, sales have doubled in the first year, and apparently doubled again in the second year. On this basis, the projection seems to be that they will double again in each of the next two years. This seems highly dubious on the information available.

Ordinary Certificate, Paper II, 2002. Question 7

(i) If 1997 is taken as 100, 1998 is $\frac{100\times 15203}{14590}=104.2$. In the same way, 1999 $\left(=\frac{100\times 15735}{14590}\right)$ is 107.8, 2000 is 111 and 2001 is 113.7 (all relative to 1997).

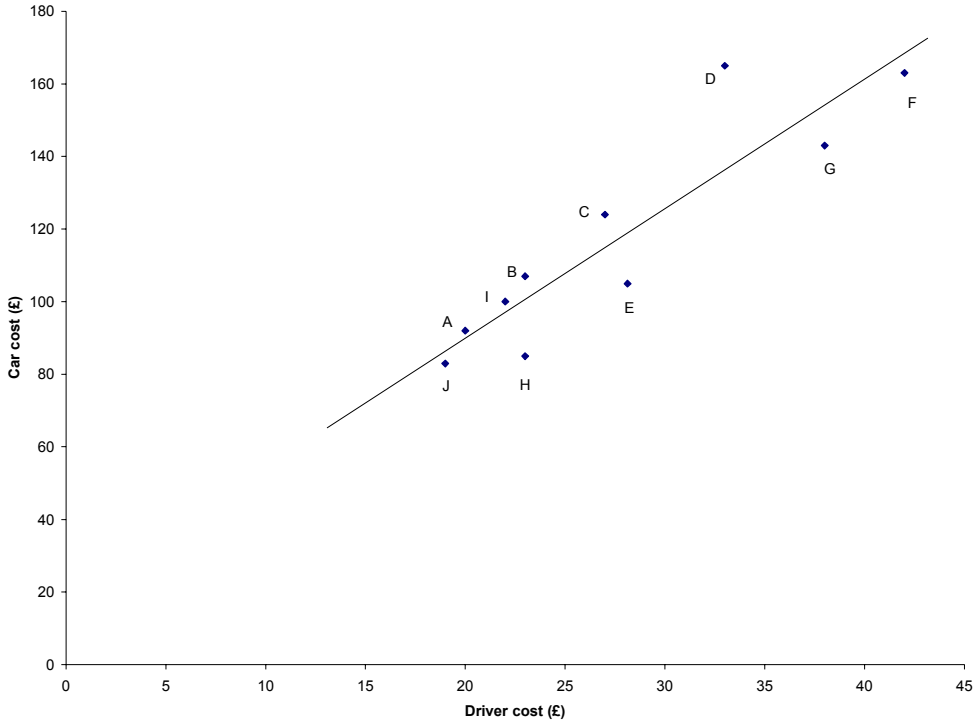
(ii) 1998 is 104.2, as in part (i).

By this method, 1999 is $\frac{100\times 15735}{15203}=103.5$, 2000 is $\frac{100\times 16191}{15735}=102.9$ and 2001 is $\frac{100\times 16596}{16191}=102.5$.

(iii) The index (i) seems to be showing that roughly 3.5% of the 1997 earnings is being added each year. However, the % increase based on present earnings, shown in (ii), is falling by a little more than 0.5% each year. So actual earnings have increased each year, but at a decreasing rate.

Ordinary Certificate, Paper II, 2002. Question 8

(i) Driver and car costs for 10 ferry routes



(ii) $\Sigma x = 275$ $\Sigma y = 1167$ $n = 10$ $\overline{x} = 27.5$ $\overline{y} = 116.7$

$$S_{xy} = \Sigma xy - \tfrac{1}{10}(\Sigma x)(\Sigma y) = 34046 - 32092.5 = 1953.5$$

$$S_{xx} = 8113 - (275)^2/10 = 550.5 \qquad S_{yy} = 144671 - (1167)^2/10 = 8482.1$$

$$r = \frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}} = \frac{1593.5}{\sqrt{550.5 \times 8482.1}} = 0.904$$

This is near +1, showing a good linear relationship between x and y, with both increasing together.

(iii) $y = a + bx$, where $\hat{b} = \dfrac{S_{xy}}{S_{xx}}$ and $\hat{a} = \overline{y} - \hat{b}\overline{x}$.

$$\hat{b} = \frac{1953.5}{550.5} = 3.549 \qquad \hat{a} = 116.7 - (3.549 \times 27.5) = 19.10 \ .$$

The fitted line is $y = 19.10 + 3.549x$.

(iv) H is "cheapest". (E is only slightly "dearer" than H). D is "most expensive".