THE ROYAL STATISTICAL SOCIETY

2007 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, 2007. Question 1

(a) Discrete data are measurements that can only take a set of specified values, such as the integers: the number of people in a household much be a whole number, an integer. The values need not have an upper limit: if we count insects or caterpillars in an agricultural plantation there must be a whole number but was cannot say there is a highest possible number. We are not limited to integers; for example, if we consider a currency that includes say a half-penny or half-cent coin, the total count of money in our pocket may be ½, 1, 1½, 2, … pence or cents; this is a discrete set of data points.

Continuous data can take any value whatever within a specified range. For example, the height of a plant may be any value whatever: 25.39827… cm is a possible height even if it might be recorded as 25.4. This is of course subject to the accuracy of available measuring equipment, but this does not detract from the principle that any height is possible. Obviously there is a physical upper limit (though we may be unsure what it is) to the possible height, and likewise a physical lower limit (perhaps taken as 0 for a plant that has died). But between these limits, any value is possible.

(b) As an illustration, if the same person does the same operation (e.g. a calculation) several times, the variation in the times he or she takes is "within subject" or "intra-subject" variation. If several different people each do the same operation once, the variation between these times is "between subject" or "inter-subject" variation. Now suppose that several different people do the same task twice each; the variation between the mean times for these people is "inter-subject" and the variations between the two times taken by the same subject are "intra-subject".

Ordinary Certificate, Paper II, 2007. Question 2

(i) Unordered data, as recorded:

Hundreds	Tens
1	7 5 8
2	1 2 5 0 5
3	5 9 8 6 8 8 9
4	3 6 2 1 3 9 7 6 7
5	1 2 1 5 0 2 0

With the "leaves" ordered:

Hundreds	Tens	(for use in part (ii)) Cumulative Frequency
1	5 7 8	3
2	0 1 2 5 5	8
3	5 6 8 8 8 9 9	15
4	1 2 3 3 6 6 7 7 9	24
5	0 0 1 1 2 2 5	31

(ii) The median is the 16th from the beginning, which is 41. Referring to the original data, it is 419.

The lower quartile is at the 8th position and the upper quartile at the 24th position. Using the cumulative frequencies, these are found as 25 (the second of those) and 49, which give the exact results 253 and 494. [Other conventions for the detailed locations of the lower and upper quartiles also exist.]

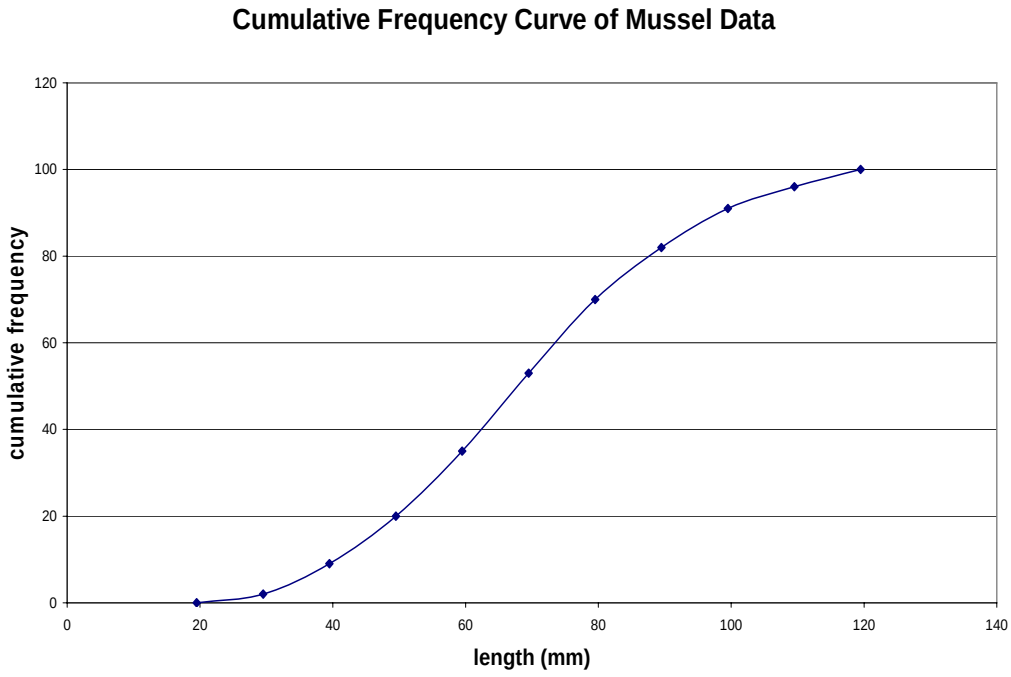
(iii) Taking the days in order, the last two observations in each group of seven (i.e. 6, 7, 13, 14, 20, 21, 27, 28) are much lower than the other five. It seems very likely that these would be the weekend days. A more informative analysis would take these 8 days as one group and the remaining 23 days as another group, and examine the groups separately.

Ordinary Certificate, Paper II, 2007. Question 3

(i)

Length (mm)	*Upper end-point (mm)*	*Frequency*	*Cumulative frequency*
20–29	29.5	2	2
30–39	39.5	7	9
40–49	49.5	11	20
50-59	59.5	15	35
60–69	69.5	18	53
70–79	79.5	17	70
80–89	89.5	12	82
90–99	99.5	9	91
100–109	109.5	5	96
110–119	119.5	4	100

(ii) Plot cumulative frequencies against upper-end points of intervals, beginning with 0 at 19.5.



(iii) The median length corresponds to a cumulative frequency of 50, and is approximately 68 mm. The quartiles correspond to frequencies 25 and 75, and are approximately 53 and 83 mm.

Hence the inter-quartile range is, approximately, 83 – 53 = 30 mm.

Ordinary Certificate, Paper II, 2007. Question 4

(i) For ammonia, the mean is $(0.533+...+0.278)/6 = 1.9827/6 = 0.330$ (or this can be found directly from a calculator).

For the standard deviation, it is as well to check from another measure that divisor $n - 1$ is intended (it is), and then the standard deviation for ammonia can be found directly from the appropriate key on a calculator (or worked out using the usual formula) as 0.296.

Similarly, for dissolved oxygen the mean is 88.0% and the standard deviation is 8.34%.

The coefficient of variation (%) = $\frac{100 \times \text{standard deviation}}{\text{mean}}$. This can now be calculated for each measure.

Thus the completed table is as follows.

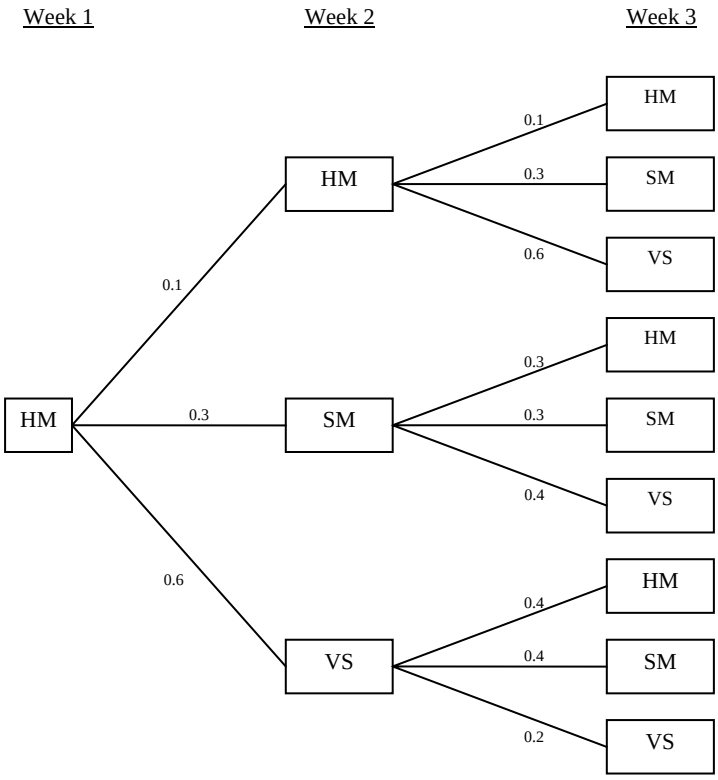
Quality measure	*Mean*	*Standard deviation*	*Coefficient of variation (%)*
Suspended solids	15.0 mg/l	19.7 mg/l	131.0
Ammonia	0.330 mg/l	0.296 mg/l	89.7
Nitrate	2.39 mg/l	0.411 mg/l	17.2
Orthophosphate	0.176 mg/l	0.0858 mg/l	48.7
Dissolved oxygen	88.0%	8.34%	9.5

(ii) One value (April) for suspended solids is very high, while two values (December and February) for ammonia are very low. These lead to very large coefficients of variation as the sets of data for these measures are very variable. Orthophosphate is also quite variable. It is very likely that on some occasions there are pollutants in the stream. The value 104% for dissolved oxygen in March needs checking for possible errors in measurement or technical errors in the analysis or calculation.

Ordinary Certificate, Paper II, 2007. Question 5

(i) (a) From the tree diagram below, P(HM in week 2) = 0.1.

(b) In week 3 there are three ways of finishing up at the hypermarket: (HM, HM), (SM, HM) and (VS, HM). From the tree diagram, these have probabilities $0.1 \times 0.1 = 0.01$, $0.3 \times 0.3 = 0.09$ and $0.6 \times 0.4 = 0.24$ respectively. So the total probability is 0.34.



In an obvious notation, we want

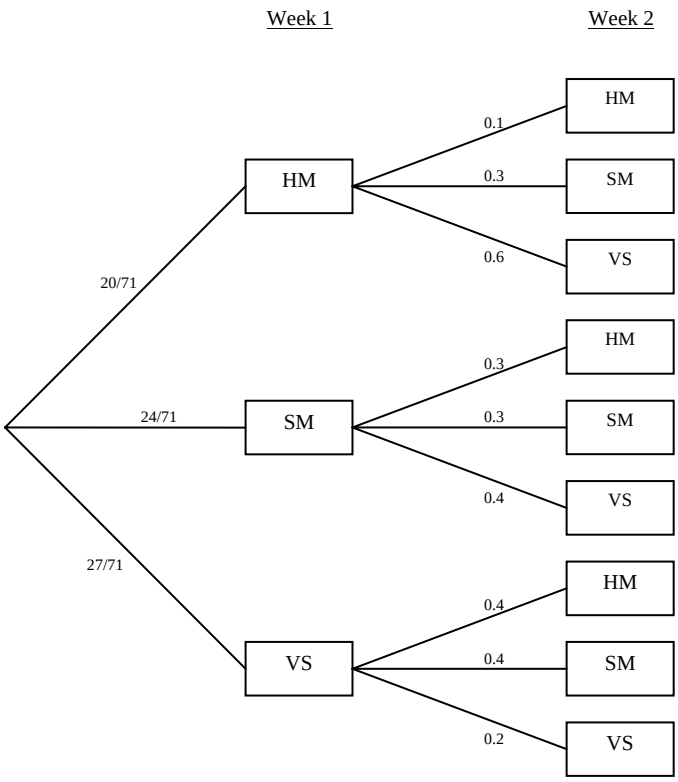
$$P\left(\mathrm{SM}_2 \middle| \mathrm{HM}_3\right) = \frac{P\left(\mathrm{SM}_2 \cap \mathrm{HM}_3\right)}{P\left(\mathrm{HM}_3\right)}.$$

$\mathrm{SM}_2 \cap \mathrm{HM}_3$ means "SM in week 2 followed by HM in week 3" and the tree diagram shows that this has probability $0.3 \times 0.3 = 0.09$.

The required probability is therefore $\dfrac{0.09}{0.34} = 0.265$.

Solution continued on next page

(ii)

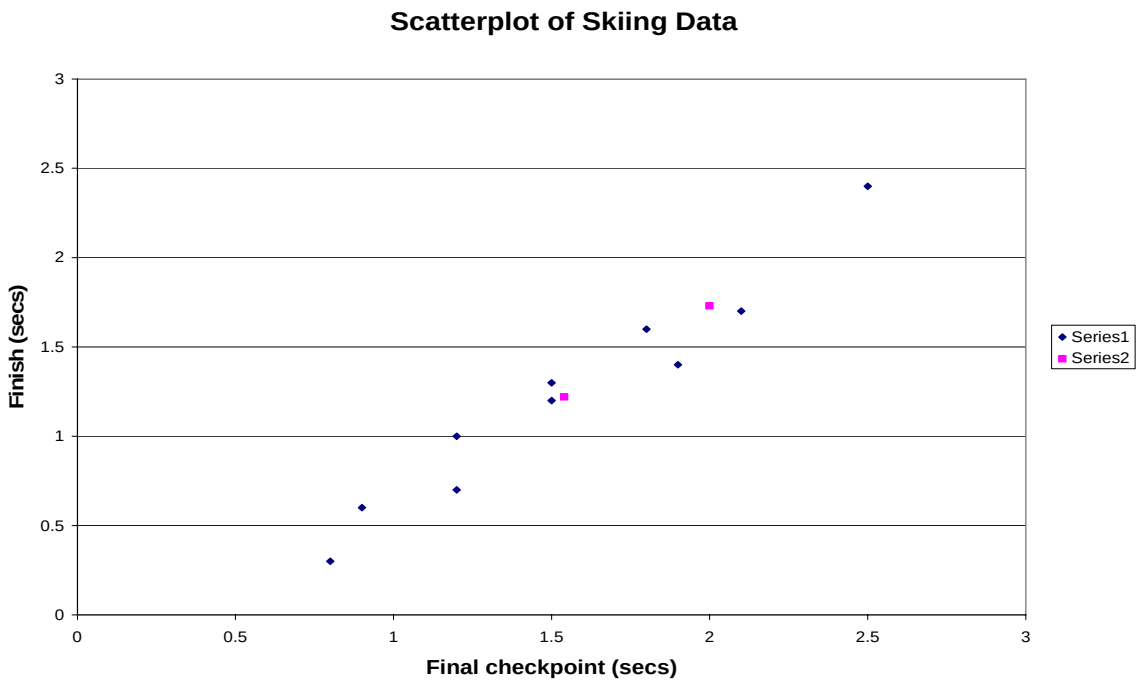


P(HM in week 2) = P(HM, HM) + P(SM, HM) + P(VS, HM)

$$= \left(\frac{20}{71}\times 0.1\right)+\left(\frac{24}{71}\times 0.3\right)+\left(\frac{27}{71}\times 0.4\right) = \frac{1}{71}\left(2.0+7.2+10.8\right) = \frac{20}{71}.$$

Ordinary Certificate, Paper II, 2007. Question 6

(i) ["Series 1" is the given set of data points. For "series 2", see part (v).]



(ii) $\overline{x} = 15.4/10 = 1.54$, $\overline{y} = 12.2/10 = 1.22$.

$$\sum (x-\overline{x})^2 = 26.34 - \frac{15.4^2}{10} = 26.34 - 23.716 = 2.624.$$

$$\sum (y-\overline{y})^2 = 18.24 - \frac{12.2^2}{10} = 18.24 - 14.884 = 3.356.$$

$$\sum (x-\overline{x})(y-\overline{y}) = 21.68 - \frac{15.4\times 12.2}{10} = 21.68 - 18.788 = 2.892.$$

$$r = \frac{2.892}{\sqrt{2.624\times 3.356}} = 0.975.$$

(iii) r is very close to +1, so there is a very close (linear) relation between x and y, both increasing together.

Solution continued on next page

(iv) The regression line is given by $y - \bar{y} = b(x - \bar{x})$, where

$$b = \frac{\sum (x - \bar{x})(y - \bar{y})}{\sum (x - \bar{x})^2} = \frac{2.892}{2.624} = 1.102,$$

so the line is

$y - 1.22 = 1.102(x - 1.54)$

or $y = 1.102x - 0.477$.

The checkpoint at 1 min 30 seconds has an x-value of $x = 2$. So the estimate of y is 2.204 – 0.477 = 1.727, giving that the estimated finishing time is 1 min 54.7 seconds.

(v) The mean ($x = 1.54$, $y = 1.22$) lies on the line. In part (iv) we have found that ($x = 2$, $y = 1.73$) lies on the line. So these can be used. See graph above. These points have been indicated as "series 2"; the line has not been drawn, to avoid confusing the display.

Ordinary Certificate, Paper II, 2007. Question 7

(i) (a) Trend is the basic long-term underlying movement of the series.

(b) Seasonal variation is short-term, usually regular (and in some sense seasonal), variation about the trend.

(c) An additive model assumes that the components Trend, Seasonal and Irregular are added together (rather than multiplied together) to give the time series value, so that the model to explain the time series data actually observed is of the form

Time series value = Trend + Seasonal + Irregular.

(ii) Differences between Sales and Trend are

Year	*Qtr 1*	*Qtr 2*	*Qtr 3*	*Qtr 4*	
2004		92.000	–108.000	–75.750	
2005	87.625	81.125	–86.125	–78.250	
2006	84.000	88.125	–89.875		
Mean	85.8125	87.083	–94.667	–77.000	Total 1.229

Since these means do not add to 0, an adjustment must be made by subtracting $\frac{1.229}{4} = 0.307$ from them to give

	Qtr 1	*Qtr 2*	*Qtr 3*	*Qtr 4*
Adjusted mean	85.506	86.776	–94.974	–77.307

(the total of 0.001 is still non-zero, but this is a minor rounding error).

(iii) Sales are considerably higher in quarters 1 and 2 than they are in quarters 3 and 4.

(iv) When the trend changes rapidly, a multiplicative model may be more appropriate because seasonal variation is then a constant percentage of trend (rather than just an absolute value).

Ordinary Certificate, Paper II, 2007. Question 8

(i) Price relatives, using 1975 = 100, are $\frac{\text{2005 price}}{\text{1975 price}}$. These are as follows.

Sugar	450.00	
Eggs	380.95	
Raisins	256.14	
Ground Almonds	172.76	
Brandy	431.82	
Total	1691.67	Mean = 1691.67 / 5 = 338.3

(ii) Although the price of these ingredients is higher in 2005 by 238.3% compared with 1975, this index tells us nothing about other items in the cost of living, such as other food and drink, housing, heating, clothing, travel and transport.

(iii) Expenditure weights are appropriate. Base weighting requires the 1975 expenditure weights. Expenditure = price × quantity. We work in units of pence.

Ingredient	*1975 price (pence)*	*Quantity*	*1975 Expenditure = Price × Quantity*	*Price Relative (1975 = 100)*	*Price relative × Expenditure*
Sugar (1 kg)	16	0.2 kg	3.20	450.00	1440
Eggs (12)	42	4 eggs	14.00	380.95	5333
Raisins (1 kg)	57	0.45 kg	25.65	256.14	6570
Gr almonds (1 kg)	246	0.1 kg	24.60	172.76	4250
Brandy (70 cl)	220	0.1 bottle	22.00	431.82	9500
			89.45		27093

The base-weighted price relative index number for 2005 using 1975 as base is $\frac{27093}{89.45} = 302.9$.

Overall the prices of the required ingredients have slightly more than trebled during the period. The weighted index is less than the unweighted index because relatively small quantities of the ingredients with the greatest price increases are used.