THE ROYAL STATISTICAL SOCIETY

2010 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, 2010. Question 1

The mode is the most frequently occurring value. There may be more than one mode.

The median is the middle value when the observations are arranged in order from lowest to highest (or vice versa). If the sample size is even, the median is halfway between the two middle values.

The (arithmetic) mean is the sum of all the observations divided by the sample size.

Hannah: categorical, unordered

Joshua: categorical, ordered

Sarah: counting, discrete

Hannah: mode only

Joshua: mode and median

Sarah: all three

Ordinary Certificate, Paper II, 2010. Question 2

		Sugar content			
		R	O	G	Total
Fat content	R	2	2	6	10
	O	5	1	5	11
	G	4	1	4	9
	Total	11	4	15	30

(i) 10/30 = 1/3.

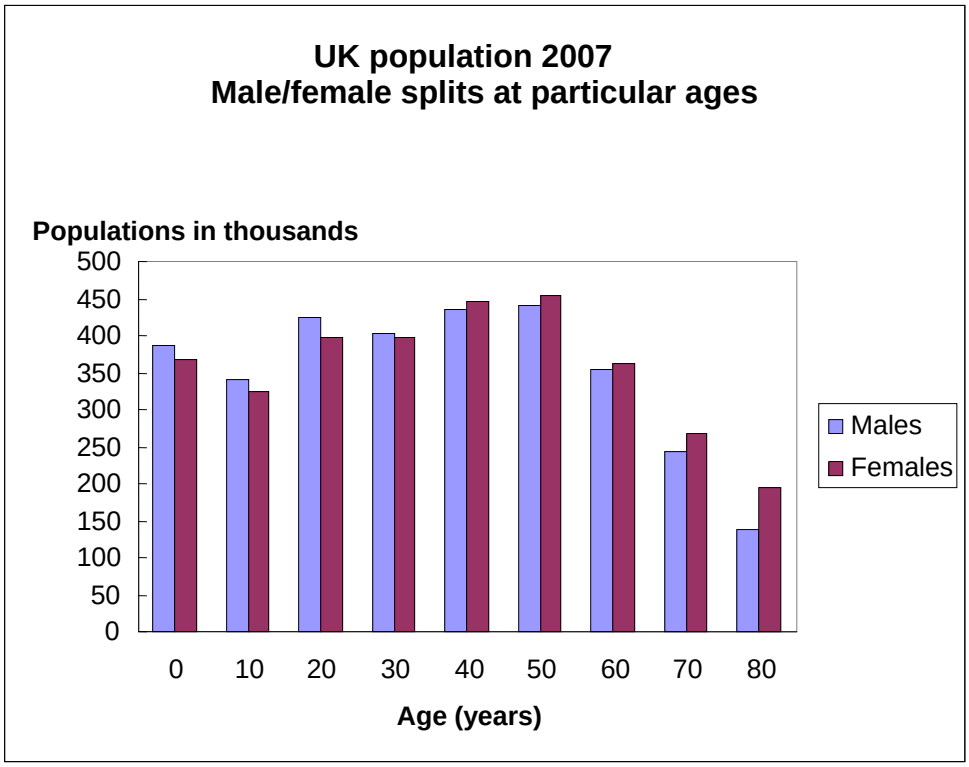
(ii) 4/30 = 2/15.

(iii) G (green).

(iv) Fat R (red), sugar G (green).

(v) 5/15 = 1/3.

Ordinary Certificate, Paper II, 2010. Question 3



[Note. The use of colour is not necessary provided the bars are clearly identifiable and suitably labelled.]

UK population 2007

Age	*Ratio of females to males*
0	0.95
10	0.96
20	0.94
30	0.99
40	1.03
50	1.03
60	1.03
70	1.11
80	1.41

There are more males than females at ages 0, 10, 20, and 30 but the reverse is true at ages 40, 50, 60, 70 and 80. The ratio is the same at ages 40, 50 and 60. The ratio increases quite sharply with age from 60 to 80, and especially from 70 to 80.

Ordinary Certificate, Paper II, 2010. Question 4

The maximum length is 3.2 cm and the minimum length is 2.3 cm, so the range [= max – min] is 0.9 cm.

The lower quartile is the length of the [(15+1)/4]th bean when arranged in order, i.e. the length of the 4th bean. The upper quartile is the length of the [3(15+1)/4]th bean, i.e. the 12th bean.

> [Note. Other conventions exist for defining the lower and upper quartiles. These were acceptable in the examination.]

The data arranged in ascending order are as follows.

Bean	*Length in cm*
1	2.3
2	2.4
3	2.4
4	2.5
5	2.7
6	2.8
7	2.8
8	2.8
9	2.8
10	2.8
11	2.9
12	3.0
13	3.0
14	3.1
15	3.2

So the lower quartile is 2.5 cm
and the upper quartile is 3.0 cm,
and thus the inter-quartile range is
3.0 – 2.5 = 0.5 cm.

The mean is $\bar{x} = \Sigma x/n = 41.5/15 = 2.77$ (cm) to 2 decimal places.

The sample variance is

$$\frac{1}{n-1}\left(\Sigma x^2 - \frac{(\Sigma x)^2}{n}\right) = \frac{1}{14}\left(115.81 - \frac{41.5^2}{15}\right) = \frac{0.9933}{14} = 0.070952$$

and the sample standard deviation is the square root of this, i.e. 0.26637, or 0.27 (cm) to 2 decimal places.

Coefficient of variation = standard deviation/mean (usually expressed as a percentage) = 0.26637/2.77 × 100% = 9.6% to 1 decimal place.

By every measure of variability, the length of the new beans is more variable than the length of the usual beans, even taking into account, with the coefficient of variation, the fact that the new beans have a larger mean than the usual beans. There is evidence to support the merchant's views on the variability in size of the beans.

Ordinary Certificate, Paper II, 2010. Question 5

The base period for the index numbers is January 1987.

The index numbers for July 2007 are RPI = 182.2, PPI = 179.8.

The index numbers for June 2008 are RPI = 193.2, PPI = 193.3.

The percentage rise in the RPI is $\{(193.2 - 182.2)/182.2\} \times 100\% = 6.0\%$ (to 1 d.p.).

The percentage rise in the PPI is $\{(193.3 - 179.8)/179.8\} \times 100\% = 7.5\%$ (to 1 d.p.).

[Thus in June 2008, the annual rate of inflation based on the RPI is 6.0% and based on the PPI is 7.5%.]

The RPI has risen by 11.0 percentage points.

The PPI has risen by 13.5 percentage points.

Rebased to July 2007, the series are as follows (to 1 decimal place).

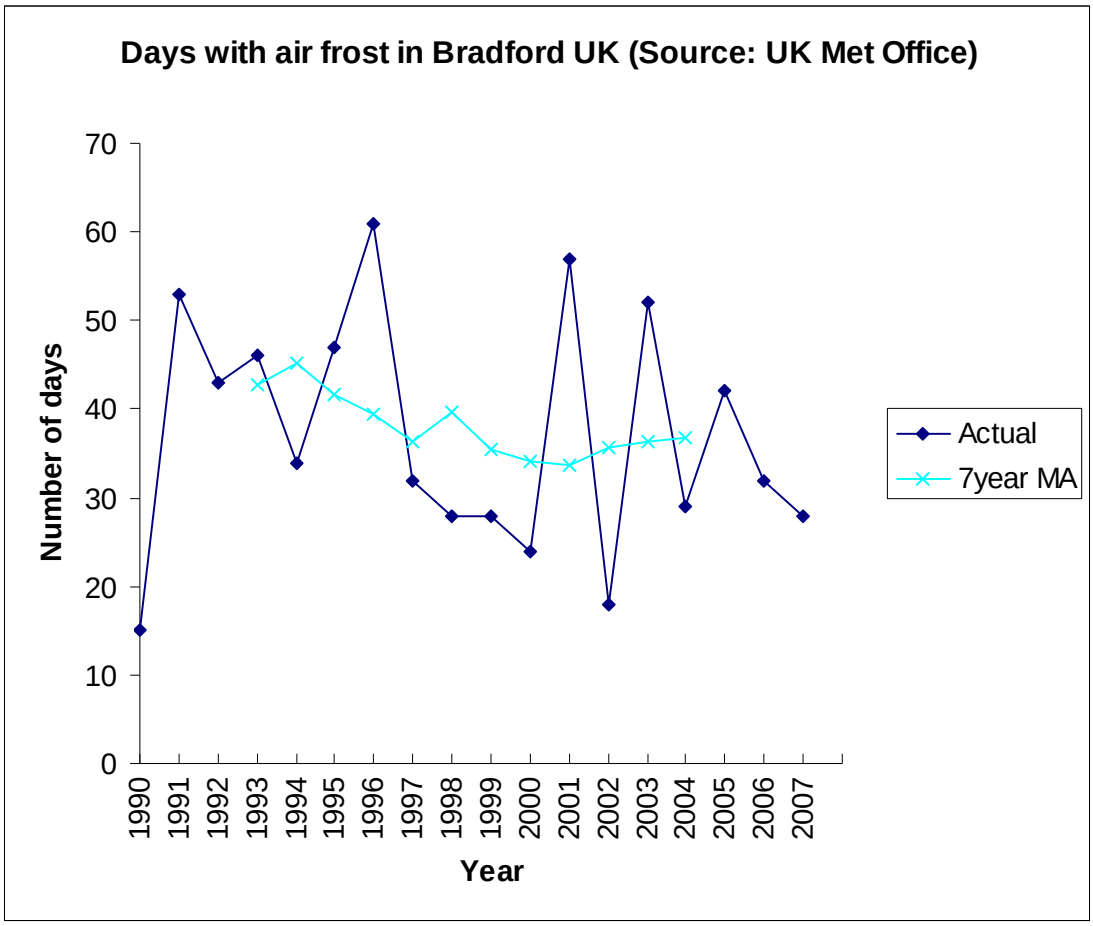
UK Retail Price Index Numbers

Month	RPI (Jul 2007 = 100)	PPI (Jul 2007 = 100)
2007 Jul	100.0	100.0
2007 Aug	100.4	100.3
2007 Sep	100.7	100.9
2007 Oct	101.2	101.3
2007 Nov	101.6	102.0
2007 Dec	102.3	102.6
2008 Jan	101.6	102.2
2008 Feb	102.6	103.6
2008 Mar	103.3	104.2
2008 Apr	104.1	105.1
2008 May	104.9	106.2
2008 Jun	106.0	107.5

The price rise for (two-person) pensioner households (0.3%) was smaller than for general households (0.4%) between July and August 2007. For all other months, the pensioner households have faced a larger percentage increase in prices compared with July 2007 than the general households. Over the whole period from July 2007 to June 2008, the pensioner annual rate of inflation was 7.5% whereas it was 6% for general households.

Ordinary Certificate, Paper II, 2010. Question 6

(The data source is shown for interest.)



[Note. The 7-year moving average is required later in the question. The use of colour is not necessary provided the plots are clearly identifiable and suitably labelled.]

The data fluctuate markedly from year to year with no obvious pattern or trend.

A benefit of using a moving average is that it smoothes out the highs and lows of the data series.

A drawback of using a moving average is that, being an averaging process, there is no estimate for the trend at each end of the data series.

Solution continued on next page

The 7-year moving average is as follows. For illustration, the first MA figure, 42.7, is calculated as (15 + 53 + 43 + 46 + 34 + 47 + 61)/7.

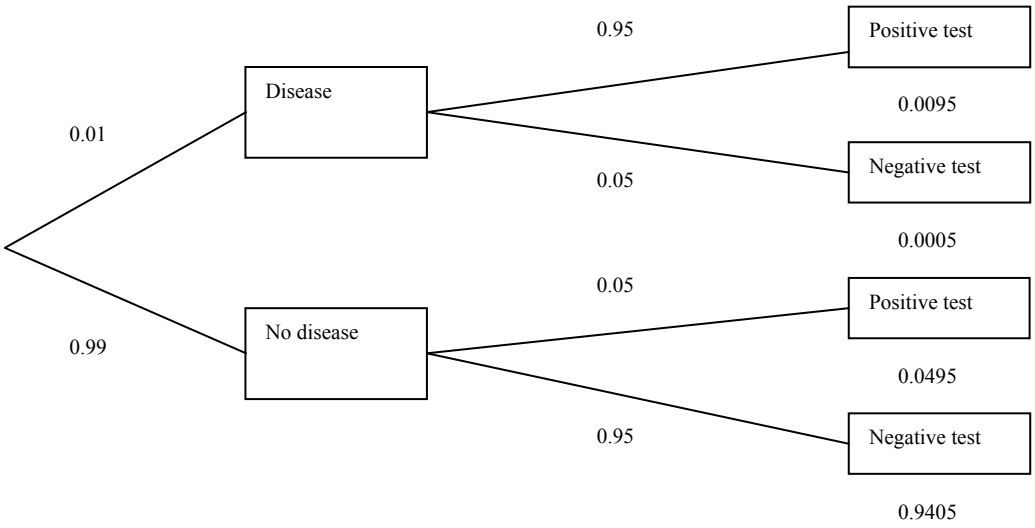
	Days with air frost in Bradford UK	
Year	Actual	7-year MA
1990	15	
1991	53	
1992	43	
1993	46	42.7
1994	34	45.1
1995	47	41.6
1996	61	39.4
1997	32	36.3
1998	28	39.6
1999	28	35.4
2000	24	34.1
2001	57	33.7
2002	18	35.7
2003	52	36.3
2004	29	36.9
2005	42	
2006	32	
2007	28	

A 7-year moving average does not appear to be appropriate as the fluctuations in the data series have not been completely removed.

Ordinary Certificate, Paper II, 2010. Question 7

The probability of an event A given an event B is $P(A \text{ and } B)/P(B)$.

The probability tree with probabilities inserted on the branches is as follows. The probability values for the final outcomes are written at the ends of the branches (eg $0.0095 = 0.01 \times 0.95$).



(i) The probability that a randomly chosen member of the population has a positive test is $0.0095 + 0.0495 = 0.059$.

(ii) The probability that a person has the disease given that this person's test result is positive is $0.0095/0.059 = 0.161$.

(iii) The probability that a person has the disease given that this person's test result is negative is $0.0005/(1 - 0.059) = 0.00053$.

The result of part (iii) indicates that the test is working well with respect to those who have a negative result, as their probability of disease is reduced from 1% to 0.053%. However, for those with a positive result, the test is not so satisfactory: only 16.1% of those with a positive result actually have the disease and the remainder will be subjected to further tests that are in fact unnecessary, thus leading to anxiety and no doubt having cost implications.

Ordinary Certificate, Paper II, 2010. Question 8

(i) $\bar{x} = 581/10 = 58.1$ $\bar{y} = 607/10 = 60.7$

$$\Sigma(x-\bar{x})^2 = \Sigma x^2 - \frac{(\Sigma x)^2}{n} = 37193 - \frac{581^2}{10} = 3436.9$$

$$\Sigma(y-\bar{y})^2 = \Sigma y^2 - \frac{(\Sigma y)^2}{n} = 38795 - \frac{607^2}{10} = 1950.1$$

$$\Sigma(x-\bar{x})(y-\bar{y}) = \Sigma xy - \frac{\Sigma x \Sigma y}{n} = 33426 - \frac{581\times 607}{10} = -1840.7$$

Therefore the product-moment correlation coefficient is

$$r = \frac{-1840.7}{\sqrt{3436.9\times 1950.1}} = -0.711.$$

(ii) r is negative, indicating that increasing age is associated with a decrease in fitness.

r is reasonably close to 1 in absolute value, indicating a relationship that is reasonable close (but not necessarily *very* close) to being linear.

(iii) The appropriate straight line is the usual linear regression line $y = a + bx$.

From the calculations above, we have $b = \frac{\Sigma(x-\bar{x})(y-\bar{y})}{\Sigma(x-\bar{x})^2} = \frac{-1840.7}{3436.9} = -0.5356$

and $a = \bar{y} - b\bar{x} = 60.7 - (-0.5356\times 58.1) = 91.8166$.

Inserting $x = 45$ gives an estimated average fitness score of 67.71 (to 2 d.p.).

(iv) The value is an estimate because the line parameters (a and b) have been estimated from the sample.

It is the average fitness score that is being estimated because, according to the underlying model, any individual fitness score would include an "error" term representing variability about the straight line; this is assumed to have average value 0 and is being ignored here.