HONG KONG STATISTICAL SOCIETY

2016 EXAMINATIONS – SOLUTIONS

ORDINARY CERTIFICATE – MODULE 2

The Society has published these solutions to assist candidates preparing for the examinations in 2017.

The solutions are intended as learning aids and should not be seen as "model answers".

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.

The OC2 paper contains several open-ended questions which expect candidates to determine their own approach to analysis and interpretation. For this reason, it is not possible to give definitive solutions. The mark scheme below therefore gives, for each question, one possible approach and the corresponding mark allocation. Other approaches will require the marker to exercise judgement.

---

1. Possible weaknesses identified:
   - Respondents will be self-selecting / not random / not representative (1)
   - Few if any respondents will have experience of all 13 banks, so they are not making fully informed judgements (1)
   - Banks will have different numbers of customers, so big banks are likely to attract more positive votes than small banks simply because of size (1)

   (Other answers possible. Award 1 for each distinct weakness identified, up to a maximum of 3.)

   On a newspaper's website it would be difficult to overcome the first weakness. (1)

   Respondents could be asked to indicate all banks that they had accounts with, and then to choose from within that subset. (1)

   Respondents could be asked to rate banks as good or bad. Then banks could be rated on the difference between the proportions of good and bad votes. (1)

(Total 6)

---

2. (i) Actual survival rates for men increased at each stage. (1)
   Actual survival rates for women fluctuated but increased over the 8 years. (1)
   Actual survival rates are higher for women than for men. (1)

   (ii) Estimated survival rates for men fluctuated a little but hardly changed over the 8 years. (1)
   Estimated survival rates for women generally fluctuated a little but rose between 1998 and 2000. (1)
   Estimated survival rates are higher for women than for men. (1)

   (iii) Men's estimates were optimistic, but women's estimates were quite realistic. (1,1)

   (Other sensible answers should be given credit.)

(Total 8)

---

3. (i) The mean is 34124 / 7276 = 4.69 or 4.7 (2)

(1 for working, 1 for answer)

(ii) The mean is not a useful measure (or is of limited usefulness) (1)

As it gives no useful indication of central tendency (1)

Particularly problematic in a U-shaped distribution like this (1)
(Allow 'distribution with greatest frequencies at the ends' or 'heavily skewed')

(iii) The story could make points such as:

The majority of practices (52%) received the highest possible score of 6, indicating low cause for concern

A further 16% were just 1 point off the highest possible score, indicating little cause for concern

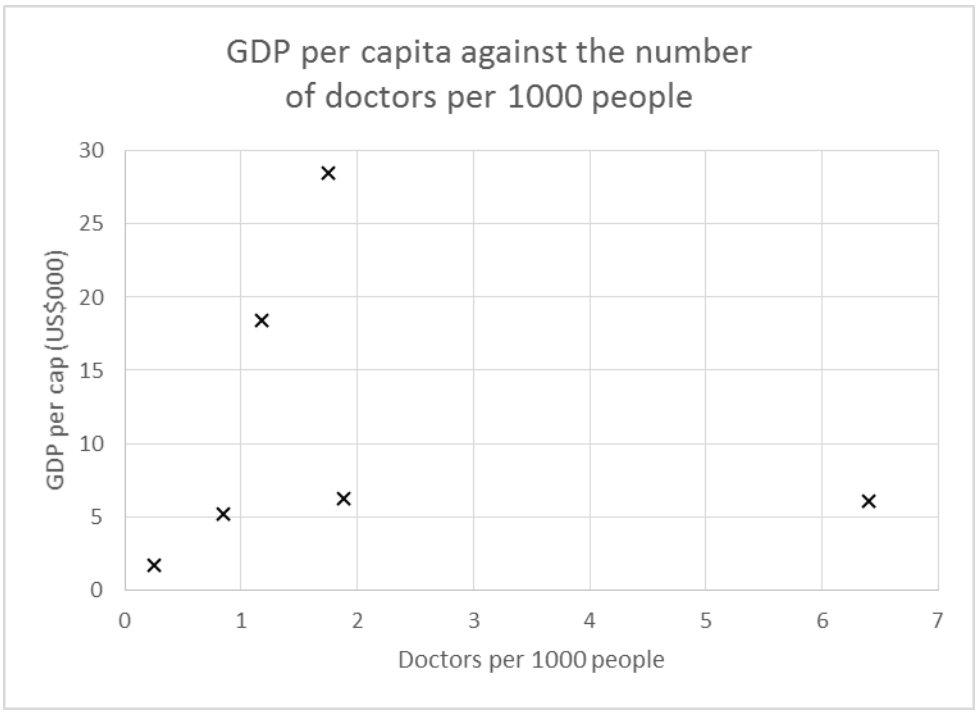
About 20% of practices were given scores of 2, 3 or 4

But there was a jump in numbers (to almost 12%) at the lowest score of 1, indicating high cause for concern (4)

(Other answers possible. Award 1 mark for each of three distinct statistical points, plus 1 mark for a well constructed news story)

(Total 9)

4. (i)



(3)

(1 for title, 1 for axes, 1 for points plotted reasonably accurately)

(ii) The ranks and squared differences are:

1	6	2	3	5	4	
3	6	4	1	5	2	Sum:
4	0	4	4	0	4	16

Spearman: 0.543

(Award 1 for ranks, 1 for use of $1 - 6\Sigma d^2/(n(n^2 - 1))$*, 1 for answer)* (3)

(iii) I mark for each sensible point made. E.g.:

One is positive and one negative, but the negative is too close to zero for its sign to indicate anything important

The two values are very different, but the sample is too small to expect agreement (2)

(iv) 1 mark for each sensible point made. E.g.

The pmcc indicates that there is no linear relationship

Spearman shows a positive association …

… which is supported to some extent by the graph (bottom left to top right)

But there are big exceptions to the trend. E.g. Cuba and DR very similar on GDP but very different on doctors, or DR and PR very similar on doctors but very different on GDP. (4)

(Total 12)

5. (i) (a) $(0.6)^3 = 0.216$ *(1 for the power, 1 for the answer)*

(b) $(1 – 0.6)^3 = 0.064$ *(1 for complementary probability, 1 for answer)*

(c) $3! \times 0.2 \times 0.1 \times 0.1 = 0.012$ *(1 for factorial, 1 for product, 1 for answer)*

(7)

(ii) (a) $0.6 \times 0.1 + 0.2 \times 0.2 + 0.1 \times 0.3 + 0.1 \times 0.4 = 0.17$

(sum of products method 1, answer 1)

(b) $0.1 \times (0.2 + 0.1 + 0.1) + 0.2 \times (0.1 + 0.1) + 0.3 \times 0.1 = 0.11$

(sum of products 1, contents of brackets 1, answer 1)

(5)

(Total 12)

---

6 (i) Some people may be unwilling to respond at all (1)

Some people may not have a main source of news (1)

Some people may have another source of news (1)

(Award 1 each for any sensible alternative answers)

(ii) For the 18-29 age group, T + I + N + R = 87. (1)

Divide given figures by 0.87 to get (1)

57.5, 31.0, 8.0, 3.4 *(accept answers rounded differently to sum to 100)* (1)

(iii) Points such as:

- In all age groups, television is the main source of news
- Television usage rises with increasing age
- Internet is second most popular for first three age groups, but newspapers is second for the oldest group
- Radio is least popular for the youngest and oldest age groups. For 30-49 and 50-64 radio and newspapers have about the same popularity

(4)

(Award 1 for each distinct sensible point identified)

(Total 10)

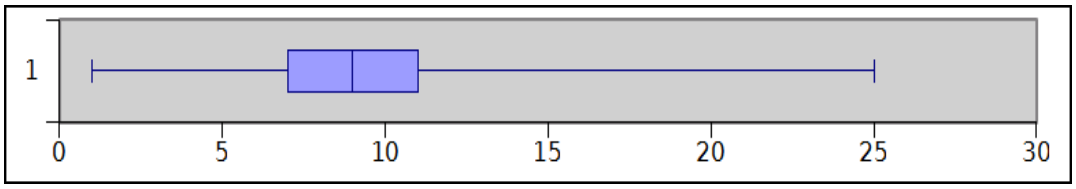
---

7 (i) Mean: 1094334 / 118611 = 9.23 (1)

Standard deviation: $\sqrt{(11091638 / 118611 - 9.226\ldots^2)} = 2.90$ (*method* 1, *answer* 1) (2)

(Other methods of calculation are acceptable.)

(ii)



(median = 9 (1), quartiles = 7 and 11 (1), extremes = 1, 25 (1), plot (1)) (4)

(iii) Mode is 8.

Points such as: (1)

Median and mean are similar (despite the long right hand tail in the data) so they convey much the same information about central tendency. (1)

The mode is lower than the other two measures so could be slightly misleading, particularly as the frequency for 9 is only slightly less than the frequency for 8. (1)

(iv) The words most commonly used will have a different distribution of word lengths. In this case the mean word length is smaller. (explanation 1, conclusion 1) (2)

(Total 12)

---

8. (i) Price relative is 483.1 / 234.0 = 2.06452….

(Accept as index 206.5) (method 1, answer 1) (2)

(ii) (a) Relative increase is (22812 – 21208) / 21208 = 0.07563….

Percentage increase is 7.56% or 7.6% (method 1, answer 1) (2)

(b) Relative increase is 12427822.3 / 6742056.6 – 1 = 0.8433…

Percentage increase is 84.3% (method 1, answer 1) (2)

(Or, equivalently, use the same method as (a))

(iii) Inflation factor over 10 years is 659.6 / 428.3 = 1.5400…. (1)

The tenth root of this factor is 1.04412…. (1)

So the average rate of inflation is 4.4% or 4.41% (1)

(iv) Laspeyres = 11548569.6 / 6742056.6 = 1.7129 … (method 1, answer 1) (2)

(Total 11)

---

9. (i) Clear increasing trend, linear (though only for 8 years) (1,1)

Clear (and stable) seasonal variation (1)

Highs in May to June, lows in September to October (1)

An additive model would be more appropriate (1)

as there appears to be fairly constant amplitude of variation (1)

(Allow: difficult to decide between additive and multiplicative models as it is not possible to judge whether or not the amplitude is constant from relatively small number of data points.)

(ii) With a zero origin on the vertical axis, the graph would look almost horizontal with little seasonal variation. (That is both the trend and the variation would be very difficult to see.) (1,1)

The total change of about 15 units over 8 years is small in relation to a mean figure of about 390 (1)

However, it is possible that small changes can have big effects. (1)

(iii) Intermediate figures: means (t, y): $84/8 = 10.5$, $3126.86/8 = 390.8575$ (1)

Regression line: $y - 390.8575 = (8 \times 32920.98 - 84 \times 3126.86)/(8 \times 924 - 84^2)\ (t - 10.5)$

(1) *for numerator and* (1) *for denominator of slope*

Simplifies to: $y = 2.1179\ t + 368.62$ (1)

PMCC: $(8 \times 32920.98 - 84 \times 3126.86)/\sqrt{((8 \times 924 - 84^2)(8 \times 1222346 - 3126.86^2))} = 0.9975$

(NB: very susceptible to rounding so accept between 0.997 and 0.999) *working* (1) *ans* (1)

(iv) $t = 16$ gives $y = 402.505 \ldots = 402.5$ (1)

$y = 500$ gives $t = 62.03\ldots$, so the year would be 2062 (1)

First estimate is only small extrapolation, and the trend over the given interval is very nearly a perfect straight line, so there is a good chance it will be accurate. (1)

The second extrapolation is over a very long interval, and it would be unwise to suppose that the trend will continue. (1)

(Total 20)