HONG KONG STATISTICAL SOCIETY

2015 EXAMINATIONS − SOLUTIONS

HIGHER CERTIFICATE – MODULE 1

The Society is providing 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 HC1 paper is, by its nature, open-ended: the questions 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 (i) The report should make points relating to the categorisations of the data, e.g.:

- **EU27 as a whole**. Almost 53% of people use a car as their main mode of transport for daily activities; 22% use public transport, 13% walk and 7% cycle. Other modes of transport (including motorcycle) account for less than 4% of usage.
- **Sex**. Men are heavier users of cars than women (59% versus 47%). Women are more frequent users of public transport than men (25%, 18%) and walk more frequently (16%, 9%). The proportions of men and women cycling are the same (7%). Men outnumber women in motorcycling (3.7%, 0.6%).
- **Age**. Car usage rises steadily across the age groups: 33% for 15-24, 61% for 25-39, 64% for 40-54; it then falls to 47% for ages 55+. Use of public transport shows a corresponding fall then rise: 41%, 19%, 15%, 22%. Walking is pretty steady across the first three age groups at about10%, but rises to 17% for 55+. Cycling is pretty steady at 7 or 8% for all age groups. Motorcycling is most common in the youngest age group.
- **Education**. Car usage is greatest (62%) among those who have had the highest level of education, but lowest (27%) among those still in education – who use public transport in large numbers (47%). Public transport is the main mode (22%) for those with the lowest level of education, followed closely by walking (19%). Cycle usage is constant at 7% across all levels of completed education.
- **Urbanization**. Those in rural areas are the biggest users of cars (64% versus 48% and 43% for urban and metropolitan residents). Metropolitan residents are most likely to use public transport (37% versus 23% for urban and 13% for rural). Walking is more common in urban areas (16%) than in metropolitan (10%) or rural (10%).
- **Occupation**. The self-employed are above average in car use (71%). Those not working are below average (39%) in car use, but above average for public transport (29%) and walking (17%).

There are many possible diagrams. E.g.

- A pie chart for the EU27 data
- Parallel stacked bars to compare data by sex, or possibly two pie charts
- Parallel line graphs to show changes in mode of transport by age
- Parallel stacked bars to compare data by level of education
- Parallel stacked bars to compare data by urbanization
- Parallel stacked bars to compare data by occupation

In the stacked bar charts, a reference bar showing the EU27 proportions would be useful.

In many of these graphs it would be helpful to combine the last three (possibly the last four) transport categories together. (16)

Marking: award up to 8 marks for extracting salient points from the data and expressing them clearly. Award up to 8 marks for relevant, appropriate and well executed diagrams.

(ii) Explanations by category. E.g.

- The EU27 data give the overall picture. Useful in itself and as a reference for other data.
- Sex: differences between men and women of interest as it is likely to relate to different socio-economic status of men and women.
- Age: interesting to see how transport use varies over a lifetime. *(Do not expect to see any consideration of cross-sectional versus longitudinal issues.)*
- Education: higher levels of education are likely to be associated with greater prosperity and so, perhaps, car usage.
- Urbanization: public transport is generally very good in metropolitan areas and poor in rural areas. Interesting to see if this is reflected in the data.
- Occupation: related to prosperity. Also, in the case of self-employment, may be related to the need for own transport. (4)

Award 1 mark each, up to a maximum of 4, for well argued reasons.

2(i) $a = 436 / 5 = 87.2$ (1)

$b = \sqrt{((5/4)(38092/5 - 87.2^2))} = 4.266\ldots$ (4.27) (formula 1, accuracy 1)

(ii) $c = (3 \times 4247 + \ldots + 40 \times 2492) / 89 = 2645$ (2644.8876) (method 1, accuracy 1)

(iii) SD is zero when the sample size is 1. This is because the SD is not defined in a sample of size 1 *or* because the SD in a sample of size 1 is zero. *(Accept either answer.)* (1)

(iv)(A) The mean examination scores of males and females differ by 3.6, (1)

the sample sizes are (compared with other groupings of the data) a reasonable size, (1)

and the SDs are not too large (SEMs, if anyone chooses to use them, are 1.4 and 0.94). (1)

So there may be some evidence of a difference here. (1)

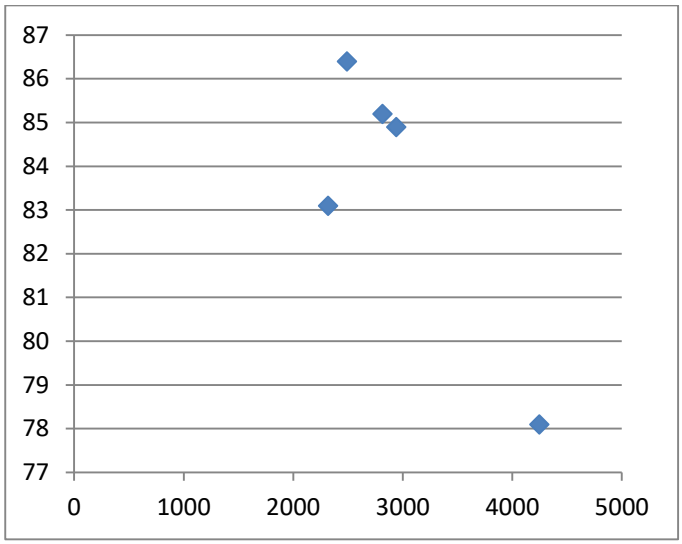
(B) The means for the A and R types are broadly similar, as are the means for K and M. Bigger difference between A+R and K+M. but not enough to be sure of a difference. (1)

Type V has a much larger mean than the rest. But the V sample is very small, (1)

the SD is very large, indicating a very big difference between the people in this group. (1)

It would be unwise to reach conclusions in this case. (1)

(iv) A scatter diagram for pages viewed and mean examination score:



The correlation, albeit based on data points from very different sample sizes) is –0.84. (This calculation is not required: a clear interpretation of the scatter diagram would suffice.)

There is some superficial credibility to the lecturer's claim. (1)(1)

(Award 1 for the conclusion and 1 for supporting reasons.)

Note: calculations are *not* required: a clear interpretation of the scatter diagram would suffice.

However, there is a vital issue of correlation and causation. (1)

It might well be that it is precisely those students who are doing least well who make the greatest use of online resources. Without those resources they might do worse. (1)

Even putting that consideration to one side, there are weaknesses in the lecturer's claim. There are only 5 data points, and this analysis is very heavily influenced by a single point – and that point is based on just 3 individuals. (Without that point, the correlation is positive.) (1)

So the lecturer's claim is, at best, given only weak support by the data. (1)

3(i) Possible points on overall structure include:

- Basic structure of statement with response on a five point scale is good.
- The statements are a mix of positives and negatives; this could be confusing and could easily be misunderstood.
- Though the survey is nominally about health and safety, some questions (6, 13) are about other matters.

Possible points on individual questions:

- Q1 is expressed negatively; possibly better to have a positive form: 'Rules and procedures do need …'
- Q2 could be expressed more directly: 'Management sometimes act slowly …'
- Q3 OK
- Q4 A simpler wording: 'Managers and Supervisors think that people who keep strictly to health and safety rules are over cautious.' Or restructure into a positive form: 'Managers and Supervisors encourage people …'.
- Q5 Could just start at 'people take …'.
- Q6 This question is not about health and safety. It could be amended to '… good communications about health and safety …'.
- Q7 Probably OK.
- Q8 Probably OK.
- Q9 Probably OK, but 'seldom' could be replaced by 'rarely'.
- Q10 A convoluted question. Perhaps: 'Sometimes people do not work to health and safety procedures, but Supervisors ignore it.'.
- Q11 A negative form that could easily be missed. 'Most of the workforce pay attention to health and safety.'.
- Q12 The word 'over' could be missed, giving a different sense to the question. 'My health and safety training gave too much emphasis to …'.
- Q13 Doesn't appear to be relevant to health and safety.

Award 1 mark for each valid point made, up to a maximum of 10. However, award a maximum of 1 mark for identifying that questions3, 7 and 8 are OK.

(ii) The introduction might do some or all of the following:

- Explain the purpose of the questionnaire.
- Emphasise the importance of health and safety.
- Urge workers to complete the survey.
- Guarantee anonymity (particularly as there is scope for criticizing management).
- Give some indication of how the information will be used.
- Explain the procedure for workers with specific concerns about health and safety.

Award 1 mark for each valid point made up to a maximum of 4. Award up to 2 points for the clarity and quality of the heading.

(iii) Obviously, with a 50% completion rate there are people whose views are not known. These people could be the ones with most to be concerned about: they might be worried about criticizing management and current practice.

The completion rate could be improved by having a tracking system for questionnaires (but without breaching anonymity). There could also be incentives for completion.

Award up to 2 marks for identifying the problems with 50% completion, and up to 2 marks for steps to improve the completion rate.

4 (i) Simple random sampling:

- Each possible sample of the required size has equal probability of selection (not: each individual has an equal chance of selection)
- Only practical when the population of interest is small enough / accessible enough for any individual to be selected. E.g. the population of interest might be all the workers on one site of a large company.
- When practical, it is free of bias and not a problematic sampling method.
- Inference is easy because based directly on probability arguments.
- A sampling frame for the whole population is needed.

(ii) Systematic sampling

- If a proportion $1/k$ of the population is to be sampled, systematic sampling proceeds by arranging the population in order, choosing a random starting point in the range 1 to $k - 1$, and choosing every kth item to be in the sample.
- Only practical when the population of interest is small enough / accessible enough for any individual to be selected. E.g. the population of interest might be all the workers on one site of a large company.
- The main potential problem is that the period of the sampling may exaggerate or hide a periodic pattern in the population. But if there are no such patterns then it performs much like simple random sampling.
- A sampling frame for the whole population is needed.

(iii) Quota sampling

- The population is divided into subgroups (e.g. male, female, or different age bands) and the interviewer is give quotas of individuals within each subgroup. The selection of individuals to fulfil the quotas is usually non-random.
- A typical example might be interviewing shoppers in a shopping centre by quota. The interviewer uses his/her own discretion to select individuals until the quotas have been met.
- This is a non-random form of sampling, so is open to bias. For example, an interviewer is unlikely to approach someone who looks aggressive or un-cooperative, so such people are under-represented. Or those who are busy may decline to respond, again being under-represented.

(iv) Cluster sampling

- When the population of interest divides naturally into subsets (clusters) each of which is broadly representative of the whole population, sampling may proceed by taking a random sample of clusters and then randomly sampling from within clusters.
- For example, if the population of interest is university students the natural clusters are universities.
- If the clusters are truly representative then cluster sampling is far cheaper and more practical than simple random sampling and performs much the same. However, bias can creep in if the clusters are not representative – e.g. in the example above, if universities differ from one another in the composition of their student bodies.

(v) Stratified random sampling

- When the population of interest can be divided into mutually exclusive groups (strata), the groups being internally homogeneous but potentially differing from one another, sampling may proceed by choosing simple random samples from within each stratum. (Frequently the sample sizes will be proportional to the stratum sizes.)
- For example, in the debate on greater integration of the EU the populations of different countries would form natural strata: individuals within countries are likely to vary less in their views than individuals chosen at random from the whole population.
- Stratified random sampling can lead to better statistical estimates than other methods as the variability within strata will generally be less than the variability in the population as a whole.
- Get information on each stratum as well as the population as a whole

In each case award 1 mark for a clear description of the method, 1 mark for an example of when it would be used, and up to 2 marks for a clear discussion of advantages and disadvantages.