Examiners' commentaries 2011

04a Statistics 1

Important note

This commentary reflects the examination and assessment arrangements for this course in the academic year 2010–11. The format and structure of the examination may change in future years, and any such changes will be publicised on the virtual learning environment (VLE).

Please note that all page references are to the 2011 subject guide.

Specific comments on questions – Zone B

SECTION A

Answer **all** parts of Question 1 (50 marks in total).

Question 1

(a) **Reading for this question**

This question asks for the mean, which is a measure of location, and the variance, which is a measure of dispersion. Specific sections exist for this material in Chapter 3 of the subject guide about data presentation. These sections also contain activities to test your understanding about these measures.

Approaching this question

i. The mean can be calculated to be $\frac{1+2+3+4+6}{5} = 3.2$, which can then be used to calculate the variance.

$$\frac{(1-3.2)^2+(2-3.2)^2+(3-3.2)^2+(4-3.2)^2+(6-3.2)^2}{5-1} = 3.7.$$

ii. In this question you may think about the definition and realise that if you subtract the same number from all the sample values, the location will be shifted whereas the dispersion will remain the same. Hence the mean will change and the variance will remain the same. Some students verified this with a numerical example by subtracting a specific number – say 1 – from all the values. This can be helpful as well.

Weak candidates confused definitions or did not know how to calculate some or all of the measures asked for. It is important that these basics are thoroughly revised: they underpin the rest of the syllabus.

(4 marks)

(b) **Reading for this question**

This question contains material from various parts of the subject guide. Here, it is more important to have a good intuitive understanding of the relevant concepts than the technical level in computations. Part (i) requires material from Chapter 3 in the section about measures of location. For parts (ii) and (iv) you need to know about hypotheses and types of error. You can look at Chapter 7 on hypothesis testing in the section about Type I and Type II errors and the hypotheses respectively. Finally, part (iii) requires knowledge about the chi-squared test that can be found in Chapter 8.

Approaching this question

Candidates always find this type of question tricky. It requires a brief explanation of the reason for a possible/not possible answer and not just a choice between the two. Some candidates lost marks as well for long, rambling explanations without a decision as to whether a statement was 'possible' or not.

i. The key word in this case is the word 'always' which makes a very strong statement. A good way to approach such questions is to think whether there is a possibility that the mean is not larger than the median. Since this is the case, a good answer would be 'No, it can be also smaller or equal'.

ii. Here, the definition of a Type II error is required. It states clearly that the power of a test is equal to 1 minus the probability of a Type II error. So, a good answer would be 'Yes, it is equal to 1 minus the probability of a Type II error'.

iii. Low values of the chi-squared statistic show that there is not much distance between observed and expected values (if we assumed no association), therefore this is possible. An answer here would be 'Yes, low values provide little evidence for association'. Some careful candidates mentioned here that we may still reject the null hypothesis for low values. This is correct but it does not mean that the statement is not possible. Marks were given to candidates when their explanations were valid.

iv. This question points out a key feature of statistical hypotheses. They can be statements about population parameters only. It would be wrong to phrase a statistical hypothesis in terms of characteristics of the sample. Hence, a good answer would be 'No, hypotheses have to be statements about population parameters'.

(8 marks)

(c) **Reading for this question**

This question asks candidates to show their understanding of the basic ideas of correlation that are covered on pp.164–168 of the subject guide with further references given in Chapter 11. Again a good technical level is not needed, rather a good and intuitive understanding of correlation, although it can be useful.

Approaching this question

This question asks you to identify an error about statistics in some sentences. Some candidates were confused and gave answers like 'There cannot be a positive correlation between research and teaching'. First of all, one cannot be sure that this is the case. But, more importantly, this is an examination on statistical concepts so you should try to identify errors about them. In this case the errors are easy to spot if you recall the basic properties of correlation.

i. Correlation takes values between -1 and 1 with values close to 0 indicating weak (low) correlation. A good answer would be 'Correlation of -0.03 is not high'.

ii. Note that correlation is a relative measure and it can quantify the relationship of quantities with different measurement units. Hence a good answer would be 'Correlation does not reflect measurement units'.

iii. Here you have to realise that correlation refers only to variables where 'high' and 'low' make sense. A good answer would be 'Gender is not a continuous variable'.

(6 marks)

(d) **Reading for this question**

The entire content of Chapter 6 is relevant and in particular Sections 6.6 and 6.7. Try Activity A6.4.

Approaching this question

This asks for a 95% confidence interval. This was straightforward once the correct distribution, t, was identified. Weak candidates did not notice that the variance was unknown and used the normal distribution.

The working is given below:

- Confidence interval formula: $\bar{x} \pm t_{\alpha/2,n-1}\frac{s}{\sqrt{n}}$.
- Degrees of freedom: 14.
- t-value: ≈ 2.14.
- Confidence interval: $(10.92, 12.08)$.

(4 marks)

(e) **Reading for this question**

These are both probability questions. Read Chapter 4 about probability and in particular the sections about the definition of probability and probability trees.

Approaching this question

i. In such questions it is essential to start by defining the events and then list what you know about them. In our case one can define
   - B_1: The first ball is red.
   - B_2: The second ball is red.

   It may be of help to write down some things that are immediately known about B_1 and B_2. For example, as there are 9 balls in total and 5 of them are red, $P(B_1) = \frac{5}{9}$. Now, to get to the specific question of this part, the event that both balls are yellow is $B_1 \cap B_2$. So, we can write

$$P(B_1 \cap B_2) = P(B_1)P(B_2|B_1) = \frac{5}{9} \cdot \frac{4}{8} = \frac{5}{18}.$$

ii. If the balls are of different colours, then only one ball is red. This can happen if the first ball was red and the second ball was green $(B_1 \cap B_2^c)$, or if the first ball was green and the second ball was red $(B_1^c \cap B_2)$. Adding the probabilities for these two cases gives

$$P(B_1 \cap B_2^c) + P(B_1^c \cap B_2) = \frac{5}{9} \cdot \frac{4}{8} + \frac{4}{9} \cdot \frac{5}{8} = \frac{5}{9}.$$

iii. Most candidates had difficulty in this part although it had similarities with (i.) and (ii.). As before, the best way to start with such exercises is to define the relevant events. In this case we have
   - A : Test positive.
   - B : Person has HIV.

   The next step is to write down what is given for the above events, or their complements, or combinations of events. In our case, for example, 5% of people have HIV, so $P(B) = 0.05$. Another way that information can be given is through conditional probabilities. Typical phrases to identify such cases are 'given ..., the probability of ... is' or 'if ..., then the probability of ... is' etc. In this question we are told that if the person has HIV (or else, given B) the diagnostic test is correct (hence positive, or else A) with probability 95%. This means that $P(A|B) = 0.95$. Similarly, we obtain that if the person does not have HIV (given B^c) the test is correct (hence negative, or else A^c) which leads to $P(A^c|B^c) = 0.90$.

   We may now use the formula

$$P(B^c|A) = \frac{P(A|B^c)P(B^c)}{P(A|B^c)P(B^c) + P(A|B)P(B)},$$

   as we know all the quantities. We get

$$P(B^c|A) = \frac{(1-0.90) \times 0.95}{(1-0.90) \times 0.95 + 0.95 \times 0.05} = \frac{2}{3}.$$

(8 marks)

(f) **Reading for this question**

This question refers to the basic bookwork which can be found on pp.15–16 of the subject guide, and in particular Activity A1.6 on p.19.

Approaching this question

Be careful to leave the x_is in the order given and only cover the values of i asked for. This question was generally well done; the answers are:

i. $\sum_{i=3}^{i=5}(x_i - 4) = (2-4) + (4-4) + (3-4) = -3.$

ii. $\sum_{i=1}^{i=4} 2x_i = (2 \times 4) + (2 \times 1) + (2 \times 2) + (2 \times 4) = 22.$

iii. $\sum_{i=2}^{i=3} x_i^3 = 1^3 + 2^3 = 9.$

(6 marks)

(g) **Reading for this question**

This question asks candidates to go back to first principles and calculate a mean and standard deviation using summary statistics. The bookwork is given on pp.36–37 for the arithmetic mean and on pp.40–41 for the variance.

Approaching this question

The total of the data is $(18 \times 5.3) + (15 \times 4.1) = 156.9$. There are $18 + 15 = 33$ data values, so the combined mean is $\frac{156.9}{33} = 4.75$. To calculate the variance, first find the sample variances. They are $1.0^2 = 1.0$ and $1.5^2 = 2.25$. Hence, the 'sum of squares' is $(17 \times 1.0) + (14 \times 2.25) = 48.5$. Samples are from the same normal distribution, so their variances are the same, so we can use the pooled variance formula. Hence $s_p^2 = \frac{48.5}{18+15-2} = 1.564$.
Answers to one decimal place were accepted.

(6 marks)

(h) **Reading for this question**

This section examines the ideas of the normal random variable. Read the relevant section of Chapter 5 and work through the examples and activities of this section.

Approaching this question

The basic property of the normal random variable for this question is that if $X \sim \mathcal{N}(\mu, \sigma^2)$, then $Z = \frac{X-\mu}{\sigma} \sim \mathcal{N}(0,1)$. Note also that

- $P(Z < a) = P(Z \leq a) = \Phi(a)$,
- $P(Z > a) = P(Z \geq a) = 1 - P(Z \leq a) = 1 - P(Z < a) = 1 - \Phi(a)$,
- $P(a < Z < b) = P(a \leq Z < b) = P(a < Z \leq b) = P(a \leq Z \leq b) = \Phi(b) - \Phi(a)$.

The above is all you need to answer the two parts of this question:

i. $P(X \geq 70) = 1 - P(X < 70) = 1 - P\left(\frac{X-55}{\sqrt{81}} < \frac{70-55}{\sqrt{81}}\right) = 1 - \Phi(5/3) = 0.048.$

ii. $P(50 \leq X \leq 59) = P\left(\frac{50-55}{\sqrt{81}} < \frac{X-55}{\sqrt{81}} < \frac{59-55}{\sqrt{81}}\right) = \Phi(\frac{5}{9}) - \Phi(-\frac{5}{9}) = 0.422.$

(4 marks)

(i) **Reading for this question**

This question requires knowledge about sampling and sample surveys. Useful background reading may be found in Chapter 9 of the subject guide. See also the references to Newbold, Carlson and Thorne given on p.135 of Chapter 9.

Approaching this question

This question asked for definitions and an example. The answer is meant to be **short**. Many candidates wrote long answers that in most situations contained irrelevant things. The definitions (also available in the subject guide) are given below:

- Random sampling: Each unit has a known, non-zero probability of being selected.
- Cluster sampling: Roughly speaking, random sampling within a cluster/subgroup of the population (that usually has also been chosen at random). See also p.142 of the subject guide.

Regarding the example, one could mention any kind of sample survey, e.g. data by population density, age, and income **within** London boroughs in order to decide where to locate new convenience stores. An advantage of random sampling is that it allows for more accurate statistical methodology, whereas quota sampling surveys are easier to conduct and have lower cost.

(4 marks)

SECTION B

Answer **two** questions from this section (25 marks each).

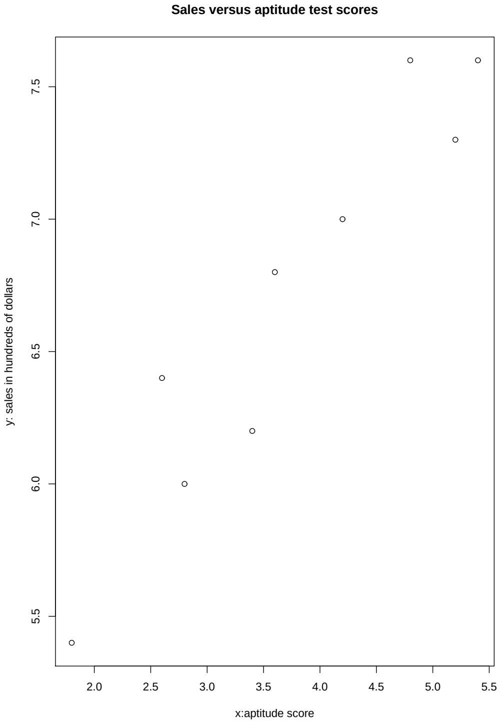
Question 2

(a) **Reading for this question**

This is a standard regression question and the reading is to be found on pp.170–174 in the subject guide. Further references are given in Chapter 11 of the subject guide.

Approaching this question

i. Candidates are reminded that they are asked to draw and label the scatter diagram which should include a title (‘Scatter diagram’ alone will not suffice) and labelled axes which also give their units. Far too many candidates threw away marks by neglecting these points and consequently were only given one mark out of the possible four allocated for this part of the question. Another common way of losing marks was failing to use the graph paper which was provided, and required, in the question. Candidates who drew on the ordinary paper in their booklet were not awarded marks for this part of the question.



(4 marks)

ii. The regression line can be written by the equation $\hat{y} = a + bx$ or $y = a + bx + \epsilon$. The formula for b is

$$b = \frac{\sum x_i y_i - n\bar{x}\bar{y}}{\sum x_i^2 - n\bar{x}^2},$$

and by substituting the summary statistics we get $b = 0.58$.

The formula for a is $a = \bar{y} - b\bar{x}$, so we get $a = 4.52$.

Hence the regression line can be written as $\hat{y} = 4.52 + 0.58x$ or $y = 4.52 + 0.58x + \epsilon$.

(5 marks)

iii. The prediction will be $\hat{y} = 4.52 + 0.58 \times 4.0 = 6.84$ hundreds of dollars. One mark was deducted in cases where the units of measurement were not given.

(2 marks)

iv. This could be a good idea due to a strong, positive, linear relationship but requires extrapolation, so it has to be applied with caution. Answers such as 'No, because it requires extrapolation' were given half credit whereas answers saying yes, but without mentioning extrapolation, were not given any credit.

(2 marks)

(b) **Reading for this question**

The question asks for a two-tailed hypothesis test comparing means. See pp.114–115 of the subject guide.

Approaching this question

i. The null hypothesis is that the mean lives of the two brands (μ_A and μ_B) do not differ, the alternative is that they do differ.

$$\mathrm{H}_0 : \mu_A = \mu_B \quad \text{vs} \quad \mathrm{H}_1 : \mu_A \neq \mu_B.$$

Use the test statistic formula:

$$\frac{\bar{x} - \bar{y}}{\sqrt{\frac{s_A^2}{n_A} + \frac{s_B^2}{n_B}}} \qquad \text{or} \qquad \frac{\bar{x} - \bar{y}}{\sqrt{\frac{s_p^2}{n_p} + \frac{s_p^2}{n_2}}}$$

to find the test statistic value: 2.934 (or 3.071 if pooled variance used). The critical values, assuming a normal approximation as the number of observations is large, are ± 1.96. If a t-distribution with 70 degrees of freedom is assumed, we have $t = 2.00$ (using 60 degrees of freedom, the nearest value in the table). Taking 5%, we reject the null hypothesis and there is therefore evidence for a difference between the two. If we take an α of 1%, the critical values are ± 2.576, so we do not reject H_0. We conclude that there is some evidence of a difference between the brands.

(7 marks)

ii. The assumptions for (ii.) were that:

- Assumption about whether $\sigma_A^2 = \sigma_B^2$.
- Assumption about whether $n_A + n_B - 2$ is 'large', hence t v. z.
- Assumption about independent samples.

(2 marks)

iii. In this case the question was whether the mean life of the tyres of brand B is longer than that of the brand A tyres. Hence the hypotheses are

$$\mathrm{H}_0 : \mu_A = \mu_B \quad \text{vs} \quad \mathrm{H}_1 : \mu_A < \mu_B.$$

The statistic to use is the same as before in absolute value but has a different sign, it is 2.934 (or 3.071 if pooled variance used). The critical values take a positive value for any significance level (≈ 1.645 for 5%), so we do not reject the hypothesis that the life of brand B is longer.

This bit was a little confusing as the sample mean of brand B was in fact smaller. Some candidates tested the hypothesis

$$\mathrm{H}_0 : \mu_A = \mu_B \quad \text{vs} \quad \mathrm{H}_1 : \mu_B < \mu_A$$

as they thought it might have been a more interesting question. Usually this is not allowed and candidates should answer the question as set. But given the peculiarity of this case these candidates were awarded full marks if they carried out their test correctly.

(3 marks)

Question 3

(a) **Reading for this question**

Part (i.) is a straightforward chi-squared test and the reading is given in Chapter 8 of the subject guide, in particular pp.122–127. For part (ii.) of the question, look at Activity A8.4.

Approaching this question

i. Set out the null hypothesis that there is no association between performance and pre-school attendance against the alternative, that there is an association. Be careful to get these the correct way round!

H_0 : No association between performance and pre-school attendance.

H_1 : Association between performance and pre-school attendance.

Work out the expected values. For example, you should work out the expected value, if there is no association, for the students below grade that attended pre-school as: $(30/100) \times 51 = 15.3$. Repeat for each cell to get the table below.

	Below Grade Level	At Grade Level	Advanced
Pre-school	15.3	20.4	15.3
No pre-school	14.7	19.6	14.7

The test statistic formula is

$$\sum \frac{(O_{i,j} - E_{i,j})^2}{E_{i,j}},$$

which gives a value of 6.896. This is a 3×2 contingency table so the degrees of freedom are $(3-1) \times (2-1) = 2$. For $\alpha = 0.05$ this gives a critical value of 5.99, hence we reject H_0. For a second (smaller) α, say 1% we get a critical value of 9.21, where we do not reject H_0.

We conclude that there is some evidence of an association between pre-school attendance and algebra marks.

Many candidates looked up the tables incorrectly and so failed to follow through their earlier accurate work. A larger number did not expand on their results sufficiently. Saying 'we reject at the 5% level, but not at 1%' is insufficient. What does this mean? Is there an association or not? If there is one, how strong is it? This needed to be answered if the full nine marks allocated for this question were to be given. Many candidates lost marks by missing out on follow-up parts like this.

(9 marks)

ii. There are a number of statements that can be drawn from the previous results. By checking differences between expected and observed numbers we can extract various arguments that aid in the interpretation of the results. For example, we may say things like:

- Main sources of association: pre-school v. below grade and at grade.
- Students who attended pre-school are more likely to obtain grade algebra marks than students who did not.

- Pre-school attendance reduces the chances of a below grade level algebra mark.

There were some excellent answers to this, but many candidates ignored this part of the question. **(4 marks)**

(b) **Reading for this question**

This was a fairly standard survey design question. Background reading is given in Chapters 9 and 10 of the subject guide which, along with the essential reading, should be looked at carefully. Candidates were expected to have studied and understood the main important constituents of design in random sampling.

Approaching this question

The main thing to note here is that many candidates wrote essays without any structure. This exercise asks for specific things and each one of them requires one or two lines. If you do not know what these things are, **do not write lengthy essays**. This is not giving you anything and is a waste of your invaluable examination time. If you can identify what is being asked for, keep in mind that **the answer should not be long**.

Note also that there is usually no unique answer to such questions. Below are some good answers for this case.

i. Two possible ways to answer this part:
   - A sampling frame can be an email list from different companies. However, this would probably rule out face-to-face interviews and response bias may become an issue.
   - Alternatively, a list of telephone numbers can be obtained and telephone interviews may be conducted. However, there may not be a telephone for every employee and this could vary depending on the job type.

ii. Make sure you provide a justification for your answers. Possible relevant stratification factors are the following:
   - Income level, as it is obviously related to job satisfaction and varies across job types.
   - Gender, as we see different proportions of women in different job types and the link with job satisfaction is interesting.
   - Education level, as it varies across job types and the link with job satisfaction is interesting.

iii. Make sure you provide a justification for your answers. Possible ways to reduce bias are the following:
   - Incentives, as people are more likely to answer and also to answer with accuracy.
   - Face-to-face interview, to eliminate confusion and reduce the chance of missing values.
   - Length of questionnaire, so that people are more likely to answer and devote a reasonable amount of attention.

iv. Summaries of variables indicating amount of job satisfaction for different job types. If these variables are continuous, boxplots can be used to graphically compare pairs of job types, whereas *t*-tests may be used to test if the observed differences are significant. If the variables are categorical, contingency tables and chi-squared tests can be used instead.

(12 marks)

Question 4

(a) **Reading for this question**

Chapter 3 provides all the relevant material for this question. More specifically read p.35 of the subject guide and look at the stem-and-leaf example and the accompanying commentary.

Approaching this question

i. The stem-and-leaf diagram the Examiners were hoping to see, is shown below. Marks were awarded for including the title, a sensible choice of stems, stem-and-leaf labels, correct vertical alignment, and accuracy.

Stem-and-leaf plot of IQ scores

Stem = \$10s | Leaf = 1s

9	566899
10	112234477
11	1233457
12	13479
13	15
14	3

ii. The mean can be found to be 111.4, whereas the modal group is the one between IQ scores of 100 and 110.

iii. The median is 109, whereas the lower quartile 101.25. The median had to be exactly 109, but for the lower quartile similar values based on different interpolations were also accepted.

iv. There is positive (right) skewness in the distribution of the data. Most of the IQ scores are around 100 and 110.

(12 marks)

(b) **Reading for this question**

Look up the sections about hypothesis testing for proportions (part i.) and confidence intervals for proportions (part ii.) in Chapters 7 and 6 of the subject guide, respectively.

Approaching this question

i. If π_1 is the proportion of males in favour of the new grading system and π_2 the corresponding proportion of females, the hypotheses are:

$$\mathrm{H}_0 : \pi_1 = \pi_2 \quad \text{vs} \quad \mathrm{H}_1 : \pi_1 \neq \pi_2.$$

The test statistic formula is

$$\frac{p_1 - p_2}{\text{s.e.}(p_1 - p_2)},$$

where the standard error can be calculated with either of the following methods:

$$\begin{aligned} \text{s.e.}(p_1 - p_2) &= \sqrt{0.55 \times 0.45 \left(\frac{1}{225} + \frac{1}{275} \right)} = 0.045 \\ \text{or} &= \sqrt{\frac{0.4889 \cdot 0.5111}{225} + \frac{0.6 \cdot 0.4}{275}} = 0.045 \end{aligned}$$

The test statistic value is 2.495. For $\alpha = 0.05$, the critical values are ± 1.96, so we reject H_0 at the 5% level.

We therefore choose a smaller second α to be 1%, which gives critical values of ± 2.576. We therefore do not reject H_0 at this level and conclude that there is some evidence of a difference in the proportions in favour of the new grading system between males and females.

Candidates got full marks for this question if they either:

- provided an interpretation of the findings saying that 'Females are more in favour of the new grading system than males', or

- justified the use of the normal distribution by the large sample.

(9 marks)

ii. This asks for a 98% confidence interval. The normal distribution may be used as before. The working is given below:

- Confidence interval formula: $(p_1 - p_2) \pm z_{\alpha/2} \times \text{s.e.}(p_1 - p_2)$.
- z-value: 2.326
- End-points: $0.111 \pm 2.326 \times 0.045$.
- Report as an interval: $(0.006, 0.216)$.

(4 marks)