Examiners' commentaries 2010

04a Statistics 1

Important note

This commentary reflects the examination and assessment arrangements for this unit in the academic year 2009–10. 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).

Specific comments on questions – Zone B

Section A

Question 1

(a) **Reading for this question**

This question asks candidates to show their understanding of basic ideas of correlation and regression. Basic correlation is covered on pp.108–109 of the subject guide, and regression on p.112. Further references are given in Chapter 11 of the subject guide.

Approaching this question

This question asked candidates to look at a scatter diagram and comment on various aspects of it.

i. No calculation was required for (i.). In fact the calculated correlation coefficient is approximately 0.7. Answers in the range 0.6 to 0.8 were given full marks. Candidates who calculated the correlation coefficient correctly were given credit for doing so, as were those who explained that it must be positive and less than one.

ii. Good candidates explained in answer to (ii.) that, as the correlation coefficient was quite high, this would justify the estimation of a regression line. The line would have a positive slope.

iii. Good candidates answered (iii.) by explaining that the positive line meant that, within the limits of the diagram, the higher the concentration of fertiliser applied to plants, the higher they would grow.

iv. Part (iv.) asked about extrapolation. Good candidates explained that it would not be sensible to go beyond the data. There might be a different, even non-linear, relationship, if higher concentrations of fertiliser were applied. Some explained that higher concentrations might even poison the plants and cause them to die.

[6 marks]

(b) **Reading for this question**

This question requires candidates to think about what they know about sampling both in the context of the estimation they learned in earlier chapters of the subject guide, and in the work on sample surveys. Useful background reading may be found in Chapters 2, 6 and 9 of the subject guide and particularly p.20, pp.61–62, and p.87. See also the references to Moser and Kalton given in Chapter 9.

Approaching this question

This question aimed to test candidates' knowledge of the difference between sampling error and sampling bias by giving simple definitions and then showing that they understood the concepts by stating which was taking place in particular circumstances. Good candidates explained that sampling error arises as part of the process of random sampling and can be measured and used to give the accuracy of estimates. Sampling bias arises from a systematic error and cannot be easily measured. They explained that (i.) was an example of sampling bias — the list of married couples does not include those who have moved over the last year. They could be different from the couples on the list and so their answers to questions may be consistently different from those of the non-movers. (ii.) describes the way a random sample is generally carried out and will have a sampling error which can be estimated and used in inference.

[5 marks]

(c) **Reading for this question**

This, relatively straightforward, 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.27–28 for the arithmetic mean and on p.30 for the standard deviation.

Approaching this question

The total of the data is $(18 \times 4.8) + (23 \times 5.7) = 217.5$. There are $18 + 23 = 41$ data values, so the combined mean is $217.5/41 = 5.305$.

To calculate the standard deviation, find first the 'sum of squares' which is $(17 \times 0.81) + (22 \times 1.21) = 40.39$.

Samples are from the same normal distribution, so their variances are the same, so we can use the pooled variance formula.

$$s_p^2 = \frac{40.39}{18 + 23 - 2} = 1.0356.$$

The standard deviation will be the square root of this $s_p = \sqrt{1.0356} = 1.018$.

Answers to one decimal place were accepted.

[6 marks]

(d) **Reading for this question**

Read pp.63–64 on confidence intervals and limits and note Example 6.2 on p.64.

Approaching this question

Use the formula (see Example 6.2 mentioned above):

$$z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \leq 4.$$

The z value for this level of confidence is 2.326. Solving the equation for n, we find $n = 135.26$.

Remember to **round up**, rather than down, and on no account leave this as a fraction! The completely correct answer has to be $n = 136$.

[4 marks]

(e) **Reading for this question**

These four probability questions are all based on pp.45–49 in Chapter 4 of the subject guide.

Approaching this question

i. For those who find formulae difficult, (i.) is most easily tackled by working out the number of possible combinations for the scores of two dice (36) and seeing what proportion of them have a sum of at least nine. Listing the pairs we see that (6,3), (3,6), (4,5) and (5,4) have a score of 9, (6,4), (4,6), and (5,5) have a score of 10, (6,5) and (5,6) have a score of 11 and (6,6) a score of 12. So there are ten ways of scoring 9 or more. Of these, four include a four, leaving 6 of the possible 10 to exclude four, thus getting a probability of 0.6.

ii. This requires the use of the formulae given in the subject guide.

$$P(W_1 \cap W_2) + P(W_1^c \cap W_2) = \frac{4}{20} \times \frac{3}{19} + \frac{16}{20} \times \frac{4}{19} = 0.2.$$

iii. This can be tackled as a tree diagram (see p.48 of subject guide) or the appropriate formulae can be used:

$$P(E) = P(E|B)P(B) + P(E|B^c)P(B^c) = 0.1 \times 0.4 + 0.8 \times 0.6 = 0.52.$$

iv. This can be tackled as a tree diagram (see p.48 of subject guide) or the appropriate formulae can be used:

$$P(B^c|E) = \frac{P(E|B^c)P(B^c)}{P(E)} = \frac{0.8 \times 0.6}{0.52} = 0.923.$$

[8 marks]

(f) **Reading for this question**

This question asks candidates to set-up and carry out a one-tailed hypothesis test. Appropriate reading is given in Chapter 7 of the subject guide. Pay particular attention to p.71 on one- and two-tailed tests. Look at Activity A7.4 on p.74.

Approaching this question

Some candidates failed to realise that this question involved a one-tailed hypothesis test. The null and alternative hypotheses should be:

Null: 25% of workers fear the loss of their job,

Alternative: Less than 25% do.

Once candidates wrote this out:

$\mathbf{H_0}$**:** $\pi = 0.25$.

$\mathbf{H_1}$**:** $\pi < 0.25$.

The test statistic is

$$\frac{p - \pi}{\sqrt{\frac{\pi(1-\pi)}{n}}} \sim N(0,1).$$

Note that the formula uses π rather than p in the denominator!

The sample proportion is $p = \frac{132}{600} = 0.22$ and the test statistic evalutes to: -1.697.

The 5% value for a one-tailed (lower tail) test is -1.645. As -1.697 falls in the critical region, we reject our null hypothesis at this level and try the 1% level say, $z = 2.326$. As this is not in the critical region, we fail to reject the null hypothesis at this level, and so our results are ambiguous. There is some evidence that fewer than 25% of workers were worried about losing their jobs, but it is not conclusive.

Good candidates managed to give the logical steps in the argument but too often candidates made silly errors. In particular there was a tendency, having found that results were significant at 5% to neglect to check at a second level as asked, or go to an unhelpful second level. (If a hypothesis is rejected at the 5% level, it will clearly also be rejected at the 10% level!) Candidates are reminded that the Examiners give marks for correct deductions and explanations in this area.

[8 marks]

(g) **Reading for this question**

This question refers to the basic bookwork which can be found on pp.12 and 13 of the subject guide and in particular Activity A1.6 on p.13.

Approaching this question

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

i. $\sum_{i=1}^{i=3}(x_i - 2) = (2-2) + (3-2) + (4-2) = 3.$

ii. $\sum_{i=4}^{i=5} 3x_i = (3 \times 2) + (3 \times 2) = 12.$

iii. $\sum_{i=2}^{i=4} x_i^2 = 3^2 + 4^2 + 2^2 = 29.$

[6 marks]

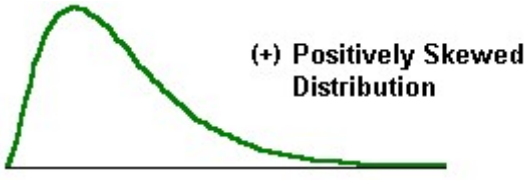
(h) **Reading for this question**

This question also required candidates to think about basic concepts and measures. Pp.27–29 of the subject guide on mean, median and mode cover this question.

Approaching the question

Basic bookwork tells us that we can tell the skew of a distribution if we know whether or not the mean is bigger than the median. In this case, the mean (50) is greater than the median (30) and so the distribution will be right-skewed or positively skewed. Most candidates knew and explained this, though some thought, wrongly, that we could not answer the question without information about the mode.

The distribution should look something like the first of the following two:



[3 marks]

(i) **Reading for this question**

This question requires candidates to revisit measures of location and dispersion. Appropriate reading is contained in pp.27–31 of the subject guide.

Approaching this question

i. All three measures here are ways of calculating dispersion. However, the variance is the square of the standard deviation and both these measures are used in statistical inference with the Normal distribution. They also require us to use the whole data set. The range only uses the highest and lowest values of the set and cannot be used for inference. Most candidates spotted this and gave sufficient explanation of the difference between the range and the other two, but a significant number of candidates did not understand that all three measures were of dispersion and that none of them measures location. Candidates are reminded that these basic ideas and measures are part of the syllabus!

ii. This was more straightforward. All the measures, apart from the median, which measures location or position, were measures of dispersion or deviation and so the median was the odd one out. Candidates who said that the standard deviation was the only parametric measure (the others all being descriptive measures using limited numbers of observations) were also given credit.

[4 marks]

Section B

Question 2

(a) **Reading for this question**

Part (i.) is a hypothesis test of the difference between two proportions. This is covered in pp.75–6 of the guide. Note Example 7.6 in the text. Part (ii.) is a confidence interval question. Read pp.66–7 and try Activity A6.4.

Approaching this question

i. The null hypothesis is that the population proportion of own brand purchasers for department stores, π_1, is the same as that for supermarkets, π_2. The alternative hypothesis is that they are different. This is a two-tailed test using sample proportions as estimates. Candidates are expected to test the difference at the 5% level ($z = 1.96$) and then, having found that difference to be significant, at the 1% level. This is also significant, so there is strong evidence that there are different proportions of own brand buyers in the two types of shop.

The working is given below:

- **H$_0$:** $\pi_1 = \pi_2$.
  H$_1$: $\pi_1 \neq \pi_2$.
- Test statistic formula:
$$\frac{p_1 - p_2}{\text{s.e.}(p_1 - p_2)}.$$
- Calculation of standard error (either of the following methods was accepted):
$$\text{s.e.}(p_1 - p_2) = \sqrt{0.4038 \times 0.5962 \times \left(\frac{1}{1200} + \frac{1}{1400}\right)} = 0.019,$$
or
$$\text{s.e.}(p_1 - p_2) = \sqrt{\frac{0.375 \times 0.625}{1200} + \frac{0.429 \times 0.571}{1400}} = 0.019.$$
- Test statistic value = 2.775.

- For $\alpha = 0.05$, critical values are ± 1.96.
- Reject H_0 at the 5% level.
- Choose second (smaller) α, say 1% gives ± 2.576, hence still reject H_0.
- Test is highly significant.
- Strong evidence of a difference between type of shops.

[9 marks]

ii. This asks for a 99% confidence interval between the two proportions. This was straightforward once the correct z-value (2.576) was found.

The working is given below:

- CI formula (can be implicit): $(p_1 - p_2) \pm z_{\alpha/2} \times \text{s.e.}(p_1 - p_2)$.
- Correct z value: 2.576.
- Correct end-points: $0.0536 \pm 2.576 \times 0.019$.
- Report as an interval: $(0.0038, 0.1033)$.

[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 recommended 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

This question looks at the data presented in (a) and asks questions about how it might have been collected.

i. This requires candidates to think how they can establish whether or not the survey which was used to collect the material was random. Some candidates interpreted this as an invitation to think of questions which might have been asked in the actual survey! This was not what was needed and no marks were given for this. Three marks were allocated for sensible questions which might establish whether or not a survey is random. The following are acceptable:

- Is there an up-to-date sample frame?
- Does it cover the target group we wish to question?
- Is the probability of selection from the frame or list known?
- Is there a low non-response rate?

Any three of these were accepted.

[3 marks]

ii. This asked candidates to explain how to carry out a telephone survey using loyalty card holders as a frame. They are also asked whether this could be regarded as a random sample. Taking that question first, the fact that not all shoppers hold a loyalty card does mean that the frame (loyalty card holders) does not include all the target shoppers and so, strictly speaking, this could not be a completely random sample. Good candidates spotted this and also gave good descriptions of the stages of the survey — using the loyalty card list to obtain and telephone loyalty card holders, explaining some of the questions they might ask and methods they might use to raise response. Marks were also given to candidates who explained why this method might not work so well (the lack of pictorial prompts/not-at-homes, etc). Overall there were some excellent replies to this question though some candidates had clearly not revised this area sufficiently.

[9 marks]

Question 3

(a) **Reading for this question**

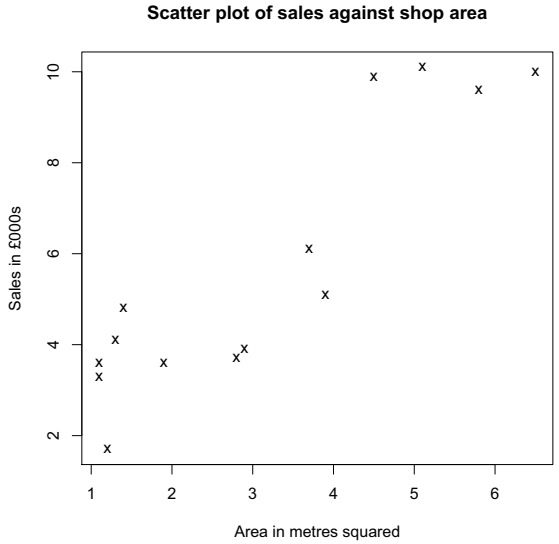
This is a standard regression question and the reading is to be found on pp.110–114 in the subject guide.

Approaching this question

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 give their units in addition. 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 any marks for this part of the question.

Additionally, if asked to calculate a regression line, no marks were given for calculating a correlation coefficient. Far too many candidates wasted time on this!

i. The diagram is given below:



ii. It shows a *positive*, *linear* relationship between sales area and actual sales. The terms in italics are needed!

iii. The line is calculated as follows:

$$b = \frac{\sum xy - n\bar{x}\bar{y}}{\sum x^2 - n\bar{x}^2} = 1.420 \qquad a = \bar{y} - b\bar{x} = 1.296$$

and gives the least squares regression line

$$\hat{y} = 1.296 + 1.420x.$$

iv. Finally we estimate an x of 6 metres squared in the equation:

$$\hat{y} = 1.296 + 1.420(6) = 9.816,$$

and find that we should expect sales to be £9,816. Be careful not to forget to give the units (£)!

(b) **Reading for this question**

The first part is a straightforward chi-squared test and the reading is given in Chapter 8 of the subject guide, in particular pp.80–83. 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 method of computation and gender against the alternative, that there is. Be careful to get these the correct way round!

H_0: There is no association.

H_1: There is an association.

Work out the expected values. For example, you should work out the expected value for the number of males who use no aids from the following: $(95/195) \times 22 = 10.7$.

The formula for calculating chi-squared is:

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

which gives us a calculated value from the data of 9.96.

This is a 4-by-2 table, so the degrees of freedom are $(4-1) \times (2-1) = 3$. This gives us a 5% critical value of 7.815 (looking up in the tables with 3 degrees of freedom). This means we reject the null hypothesis at the 5% level. Taking the 1% value next (remember there is no point whatsoever in moving to the 10% value — you know that you will continue to reject the null hypothesis) we have a critical value of 11.34 and this time fail to reject the null hypothesis.

We have rejected the null hypothesis of no association at the 5% level, but not at the more stringent 1%, so all we can say is that we think there is some association, but not a strong one, between gender and method of computation.

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 a connection 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 given. Many candidates lost marks on missing out on follow-up like this.

[9 marks]

ii. The final part of this question asked for comments on potential gender differences. The point of this is that chi-squared tests only establish association (or the lack of it). Here you are being asked if there are any differences which seem to contribute to the association. Looking at individual 'observed' and 'expected' values, we can see that there is no difference between men and women in their using no aids. Slightly fewer women than men than might have been expected use a computer. But the big difference is that men are much less likely than women to use a statistical function on a calculator than expected, while women are less likely to use a basic calculator compared with men. There were some excellent answers to this, but many candidates ignored this part of the question.

[4 marks]

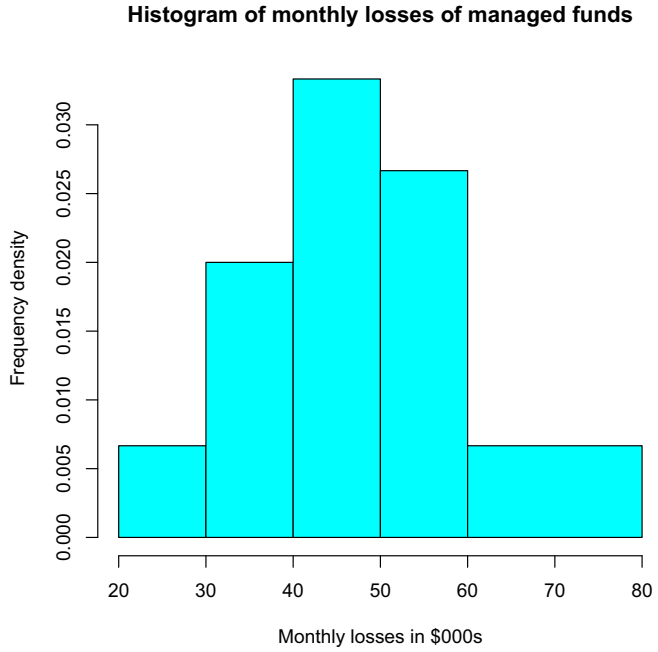
Question 4

(a) **Reading for this question**

Reading is given on p.34 of the subject guide. You should also look at the diagram (Figure 3.3) and the accompanying commentary.

Approaching this question

i. Many candidates did not read the question carefully enough. They are required to draw a histogram with five classes on the graph paper provided. Far too many candidates, seeing the difficulty of managing the unequal widths which would be needed if a conventional five-class histogram were to be drawn with five classes, chose to draw one with six instead. This was penalised. In addition some candidates, as in 3(a), did not use the graph paper as asked. No marks were awarded if this was done. The histogram the Examiners were hoping to see is shown below:



Note that students were awarded marks for:

- labelling the y axis 'frequency density' and not 'frequency'
- labelling the x axis
- giving the histogram an appropriate title
- adhering to five classes and
- accuracy.

Some candidates ingeniously managed to construct and draw a histogram using intervals running in tens from 23 to 73. The examiners accepted this, as it is technically correct!

[6 marks]

ii. Looking at the measures asked for, the mean is \$47,333, median \$48,500, range \$49,000. To get full marks for this, candidates needed to remember to give the units (dollars as shown here).

[3 marks]

iii. The modal group is 40–50.

[1 mark]

iv. Given this information, it is clear that the distribution is negatively (left) skewed (see Question 1(h) earlier!). Some candidates also commented, correctly, that the numbers in the 20–30 categories seemed to be lower than might have been expected.

[2 marks]

(b) **Reading for this question**

Reading for the hypothesis test is given in Chapter 7 of the subject guide, in particular p.74. For the confidence interval, look up p.66 and also p.63 on Student's t distribution.

Approaching this question

i. This asks for a two-sided hypothesis test comparing means. The null hypothesis is that the mean delivery times for Store A and Store B do not differ, the alternative is that they do.

$\mathbf{H}_0$**:** $\mu_A = \mu_B$.

$\mathbf{H}_1$**:** $\mu_A \neq \mu_B$.

The test statistic is either:

$$z = \frac{\bar{x} - \bar{y}}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} = 6.41$$

or, if using the pooled estimate (which was accepted):

$$z = \frac{\bar{x} - \bar{y}}{\sqrt{\frac{s_p^2}{n_1} + \frac{s_p^2}{n_2}}} = 6.64.$$

The tabular value, assuming a normal approximation as the number of observations is large, is 1.96. If a t distribution with 70 (using 60 — the nearest value on the table) degrees of freedom is assumed, we have $t = 2.00$. In either case, the calculated value is much larger and the null hypothesis is therefore rejected. Taking 1% and even 0.01% values, we still reject the null hypothesis and there is therefore a strong difference between the two. For full marks, candidates were expected to explain that rejecting the null hypothesis showed there was a difference between the two stores, and that the levels at which the null hypothesis could be rejected meant that this was a marked difference. Good candidates got full marks for this.

ii. The assumptions were:

- if the pooled estimate for the variance was used, the two variances were equal
- the Normal distribution can be used as sample size was large (or a statement that Student's t was to be used)
- that samples are independent.

iii. This required another confidence interval, but only for Store A. Some candidates insisted on carrying this out for the difference between the two stores and so lost marks.

It was expected that, as the standard deviation figures are based on a sample estimate, and the number in the sample for Store A was 41, t multipliers would be used. Marks were given for this or for an explanation that it was acceptable to use the normal approximation with these numbers.

Using t, we get the confidence interval formula

$$\bar{x} \pm t_{\alpha/2,n-1}\frac{s}{\sqrt{n}},$$

with 40 degrees of freedom. This gives a t value (for a 98% confidence interval) of 2.423 and a confidence interval of (33.38, 35.22), or the statement, 'The confidence interval lies between 33.38 and 35.22'.

[13 marks]