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 A

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 0.9. Answers in the range 0.8 to 0.95 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 pupils does not include those who have arrived over the last year: they could be different from the pupils on the list and so their answers to questions may be consistently different from them; (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 $(16 \times 4.4) + (25 \times 5.2) = 200.4$. There are $16 + 25 = 41$ data values, so the combined mean is $200.4/41 = 4.888$.

To calculate the standard deviation, first find the 'sum of squares' which is $(15 \times 0.81) + (24 \times 1.21) = 41.19$.

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

Hence $s^2 = 41.19/(16 + 25 - 2) = 1.0562$.

The standard deviation will be the square root of this, i.e. 1.028.

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):

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

The z-value for this level of confidence is 1.96. Solving the equation for n, we find $n = 216.09$. Remember to **round up**, rather than down, and on no account leave this as a fraction! The correct answer has to be $n = 217$.

[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

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 seven.

(ii.) requires the use of the formulae given in the subject guide.

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

i. Listing the pairs we see that

$$(6,1),(5,2),(4,3),(3,4),(2,5),(1,6)$$

have a sum of 7,

$$(6,2),(5,3),(4,4),(3,5),(2,6)$$

have a sum of 8,

$$(6,3),(3,6),(4,5),(5,4)$$

have a sum of 9,

$$(6,4),(4,6),(5,5)$$

have a sum of 10,

$$(6,5),(5,6)$$

have a score of 11 and (6,6) a score of 12. So there are 21 ways of scoring 7 or more. Of these, 15 have no 3, hence the probability neither face is a 3 is $15/21 = 0.7143$.

[2 marks]

ii. We require the second person asked to watch pay-per-view films, denoted W_2. So we consider two possibilities, i.e. the first person asked either watches or does not watch this way, W_1 and W_1^c respectively.

$$P(W_1 \cap W_2) + P(W_1^c \cap W_2) = \frac{5}{25} \times \frac{4}{24} + \frac{20}{25} \times \frac{5}{24} = 0.2.$$

[2 marks]

iii. Here we apply the 'Total Law of Probability'. Let E denote Rachel enjoys the party and that B be Bill goes.

$$P(E) = P(E|B)P(B) + P(E|B^c)P(B^c) = 0.7 \times 0.6 + 0.2 \times 0.4 = 0.5.$$

[2 marks]

iv. This is a straightforward conditional probability question, where the conditioning event is that Rachel did not enjoy herself, denoted E^c.

$$P(B^c|E^c) = \frac{P(E^c|B^c)P(B^c)}{P(E^c)} = \frac{0.8 \times 0.4}{0.5} = 0.64.$$

[2 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:

$\mathbf{H_0}$: $\pi = 0.20$

$\mathbf{H_1}$: $\pi < 0.20$

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 $p = \frac{83}{500} = 0.166$ and the test statistic evaluates to -1.901.

The 5% critical value for a one-tailed (lower tail) test is -1.645. As -1.901 falls in the critical region, we reject our null hypothesis at this level and try the 1% level say, with critical value -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 20% 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.12–13 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 values of i asked for. This question was generally well done; the answers are:

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

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

iii. $\sum_{i=1}^{i=3} x_i^2 = 2^2 + 1^2 + 4^2 = 21.$

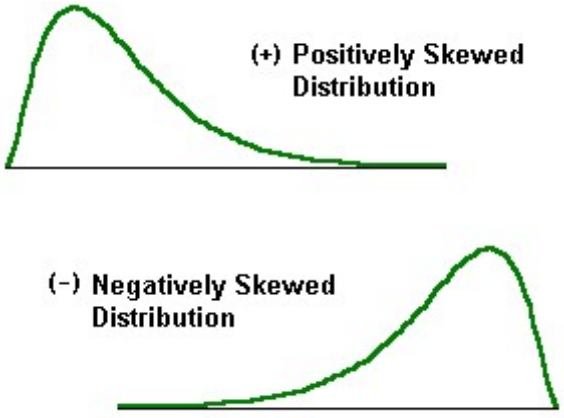
[6 marks]

(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 this 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 (30) is less than the median (45) and so the distribution will be left-skewed or negatively 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 second 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 on pp.27–31 of the subject guide.

Approaching this question

i. All three measures here are ways of calculating dispersion. The range does not use all the data points and is most sensitive to outliers. 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 mean, which measures location or position, were measures of dispersion or deviation/spread and so the mean is the odd one out.

[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.75–76 of the subject guide. Note Example 7.6 in the text. Part (ii.) is a confidence interval question. Read pp.66–67 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:

$\mathbf{H_0}$**:** $\pi_1 = \pi_2$

$\mathbf{H_1}$**:** $\pi_1 \neq \pi_2$

This is a two-tailed test using sample proportions as estimates. Candidates are expected to test the difference at the 5% level ($z = \pm 1.96$) and then, having found that difference to be significant, at the 1% level (or some other appropriate combination of levels). The 1% test is also significant, so there is strong evidence that there are different proportions of own brand buyers in the two types of shop. Nine marks were allocated to this part of the question. The working is given below:

- Test statistic formula: $\dfrac{p_1 - p_2}{\text{s.e.}(p_1 - p_2)}$.
- Calculation of standard error (either of the following methods was accepted):

$$\text{s.e.}(p_1 - p_2) = \sqrt{0.4333 \times 0.5667 \times \left(\frac{1}{1400} + \frac{1}{1600}\right)} = 0.018,$$

or

$$\text{s.e.}(p_1 - p_2) = \sqrt{\frac{0.3571 \times 0.6429}{1400} + \frac{0.5 \times 0.5}{1600}} = 0.018.$$

- Test statistic value = 7.983.
- 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 98% confidence interval between the two proportions. This was straightforward once the correct z-value (2.326) 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: 2.326.
- Correct end-points: $0.1429 \pm 2.326 \times 0.018$.
- Report as an interval: $(0.1010, 0.1847)$.

[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 part (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 data 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, for the remaining marks, to explain how to carry out a survey using shoppers as a frame. They are also asked how to ensure random selection of shoppers. All of the following needed to be addressed:

- the sampling frame
- how to contact shoppers
- suggested survey questions
- stratification by age, sex, etc.
- incentives to minimise non-response
- potential response bias and interviewer effects.

Marks were given for raising other relevant points. Overall there were some excellent replies to this question though some candidates had clearly not revised this area sufficiently.

[9 marks]

Question 3

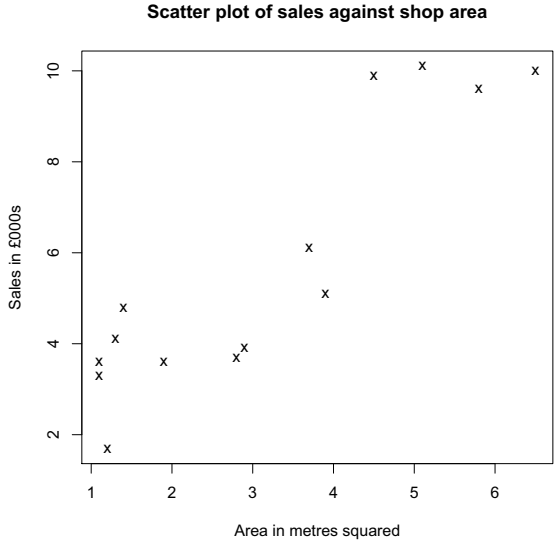
(a) **Reading for this question**

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 marks for this part of the question.

i. The diagram is given below:



It shows a positive, linear relationship between sales area and actual sales.

ii. The correlation is calculated using the formula

$$r = \frac{\sum x_i y_i - n\bar{x}\bar{y}}{\sqrt{\left(\sum x_i^2 - n\bar{x}^2\right)\left(\sum y_i^2 - n\bar{y}^2\right)}}$$

and gives a value of $r = 0.8998$.

iii. A linear regression line is justified, because we have a *strong*, *positive*, *linear* relationship. All three terms in italics are needed!

iv. For the intercept a, any value between 1 and 1.5 was given full marks, and for the slope b any value between 1.25 and 1.75 was given full marks.

[12 marks]

(b) **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.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 9.963.

This is a 4-by-2 contingency 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% critical value next (remember there is no point whatsoever in moving to 10% – 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 evidence of association between gender and method of computation, but the evidence is not terribly strong.

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. 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 managed funds' monthly losses

Stem = \$10000s	Leaf = \$1000s
2	3
3	013569
4	0223447899
5	012334567
6	345
7	2

ii. The mean is $\$4,333$ and the standard deviation is $\$11,287$. To get full marks for this, candidates needed to remember to give the units (dollars as shown here).

iii. The data look approximately normal.

iv. Due to (iii), we can expect around 95% of losses to be within two standard deviations of the mean. We get the interval $\$47,333 \pm 2 \times \$11,287 = (\$24759, \$69907)$. This compares well with the actual figures: 93.3% lie in this interval.

[12 marks]

(b) **Reading for this question**

See p.74 of the subject guide. For the confidence interval, look up p.66 and also p.63 on Student's t.

Approaching this question

i. The question 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:

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

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

Use the test statistic formula

$$\frac{\bar{x} - \bar{y}}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \quad \text{or} \quad \frac{\bar{x} - \bar{y}}{\sqrt{\frac{s_p^2}{n_1} + \frac{s_p^2}{n_2}}}$$

and you find that $z = 5.93$ (the pooled estimate, if used, was accepted; it came to 7.15). The critical value, assuming a normal approximation as the number of observations is large, is 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). 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 strong evidence for a difference between the two. For full marks, candidates were expected to explain that rejecting the null hypothesis showed there was evidence for a difference between the two stores, and that the levels at which H_0 could be rejected meant that this evidence was strong.

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

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

iii. This question 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 about 2.3 and a confidence interval of (37.20, 40.00). A statement like 'the confidence interval consists of numbers lying between 37.20 and 40.00' was also acceptable.

[13 marks]