UNIVERSITY OF LONDON | INTERNATIONAL PROGRAMMES

Course information 2015–16
ST104a Statistics 1 (half course)

This half course introduces students to the basic statistical concepts which they may need to understand and use in the other courses they intend to study in their degree or diploma.

Prerequisite

None apply.

Aims and objectives

The emphasis of the course is on the application of statistical methods in management, economics and the social sciences. Attention will focus on the interpretation of tables and results and the appropriate way to approach statistical problems. Treatment is at an elementary mathematical level. Ideas of probability, inference and multivariate analysis are introduced and are further developed in the half course 04b Statistics 2.

Essential reading

For full details, please refer to the reading list

Newbold,P., W. Carlson and B. Thorne *Statistics for Business and Economics.* (Pearson Education)

Lindley, D.V. and W.F. Scott. *New Cambridge Statistical Tables*. (Cambridge: Cambridge University Press)

Assessment

This half course is assessed by a two-hour, unseen, written examination.

Learning outcomes

At the end of the course and having completed the essential reading and activities students should

- ☑ be familiar with the key ideas of statistics that are accessible to a student with a moderate mathematical competence
- ☑ be able to routinely apply a variety of methods for explaining, summarising and presenting data and interpreting results clearly using appropriate diagrams, titles and labels when required
- ☑ be able to summarise the ideas of randomness and variability, and the way in which these link to probability theory to allow the systematic and logical collection of statistical techniques of great practical importance in many applied areas
- ☑ have a grounding in probability theory and some grasp of the most common statistical methods
- ☑ be able to perform inference to test the significance of common measures such as means and proportions and conduct chi-square tests of contingency tables
- ☑ be able to use simple linear regression and correlation analysis and know when it is appropriate to do so.

Syllabus

This is a description of the material to be examined, as published in the *Programme handbook*. On registration, students will receive a detailed subject guide which provides a framework for covering the topics in the syllabus and directions to the essential reading

This course introduces some of the basic ideas of theoretical statistics, emphasising the applications of these methods and the interpretation of tables and results.

Basic background: Elementary summation signs, elementary probability, Venn and tree diagrams.

Data collection: Elements of survey design, the stages of a survey, ideas of randomness, observation and experiment.

Data presentation and analysis: Descriptive statistics, measures of location and dispersion, pictorial and graphical representation.

The Normal Distribution*:* Estimation of mean, proportion, standard deviation, confidence intervals and hypothesis testing. Ideas of testing for differences between means and proportions. The use of Student's *t*.

Goodness of fit: The chi-square distribution and contingency tables.

Regression and correlation: An introduction to the ideas of regression and correlation, least squares, estimation of *a*, *b*, and *r*, scatter diagrams.

Students should consult the *Programme Regulations for degrees and diplomas in Economics, Management, Finance and the Social Sciences* that are reviewed annually. Notice is also given in the *Regulations* of any courses which are being phased out and students are advised to check course availability.

Examiners' commentaries 2015

ST104a Statistics 1

Important note

This commentary reflects the examination and assessment arrangements for this course in the academic year 2014–15. 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). Note that in what follows • corresponds to 1 mark unless stated otherwise.

Information about the subject guide and the Essential reading references

Unless otherwise stated, all cross-references will be to the latest version of the subject guide (2014). You should always attempt to use the most recent edition of any Essential reading textbook, even if the commentary and/or online reading list and/or subject guide refer to an earlier edition. If different editions of Essential reading are listed, please check the VLE for reading supplements – if none are available, please use the contents list and index of the new edition to find the relevant section.

Comments on specific questions – 28 May replacement examination

Candidates should answer **THREE** of the following **FOUR** questions: **QUESTION 1** of Section A (50 marks) and **TWO** questions from Section B (25 marks each). **Candidates are strongly advised to divide their time accordingly.**

Section A

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

Question 1

(a) **Consider the following sample dataset:**

$$\mathbf{8,\ 2,\ 6,\ } x\mathbf{,\ 5.}$$

You are told that the value of the sample mean is 6.

i. **Calculate the value of x.**

ii. **Find the sample variance.**

[4 marks]

Reading for this question

This question contains material mostly from Chapter 3 and in particular Section 3.8 (Measure of location) for part (i) and Section 3.9 (Measure of dispersion) for part (ii) of the subject guide.

Approaching the question

First you need to write down the formula for the sample mean. Then, it is important to do the summation carefully and divide with the correct number of observations to obtain the mean. Note that the sum in the numerator will contain the unknown x, hence this will give you a simple equation. The solution of this equation will provide x. The workout of the solution is as follows.

i. • Since the sample mean is equal to 6, we can write:

$$\frac{8+2+6+x+5}{5}=6$$

• or else:

$$21+x=30 \leftrightarrow x=9.$$

ii. • Method:

$$s^2=\frac{(8-6)^2+(2-6)^2+(6-6)^2+(9-6)^2+(5-6)^2}{4}$$

• Correct value: 7.5.

Some candidates divided by 5 in the formula above. In such cases only one mark was awarded for part (ii), provided that the correct value was obtained. The reason is that the formula for the sample variance provided in the subject guide only suggests dividing by $n-1$, where n is the number of observations. In another error that occurred in some cases, candidates subtracted the number $x=9$ rather than the sample mean which is given to be 6.

(b) Suppose that $x_1=7$, $x_2=3$, $x_3=1$, $x_4=0$, $x_5=-6$, and $y_1=-3$, $y_2=5$, $y_3=-8$, $y_4=9$, $y_5=1$. Calculate the following quantities:

$$\textbf{i.}\ \sum_{i=2}^{i=4} 2y_i \qquad \textbf{ii.}\ \sum_{i=1}^{i=3} 4(x_i-1) \qquad \textbf{iii.}\ y_1^2+\sum_{i=3}^{i=5}(x_i^2+2y_i^2).$$

[6 marks]

Reading for this question

This question refers to the basic bookwork which can be found on Section 1.9 of the subject guide and in particular Activity A1.6.

Approaching the question

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

i. $\sum_{i=2}^{i=4} 2y_i=2(5-8+9)=12.$

ii. $\sum_{i=1}^{i=3} 4(x_i-1)=4\sum_{i=1}^{i=3}(x_i-1)=4((7-1)+(3-1)+(1-1))=4(6+2-0)=32.$

iii. $y_1^2+\sum_{i=3}^{i=5}(x_i^2+2y_i^2)=(-3)^2+(1^2+2\times(-8)^2)+(0^2+2\times 9^2)+((-6)^2+2\times 1^2)=$ $9+129+162+38=338.$

(c) In a population 20% of men show early signs of losing their hair and 2% of them carry a gene that is related to hair loss. It is also known that 80% of men who carry the gene experience early hair loss.

i. What is the probability that a man carries the gene and experiences early hair loss?

ii. What is the probability that a man carries the gene, given that he experiences early hair loss?

[4 marks]

Reading for this question

This is a question on probability and targets mostly the material in Chapter 4. It is essential to practise on such exercises through the activities and exercises in this chapter as well as the material on the VLE. In particular you can attempt Activity A4.6 and Sample examination question 4. It is also useful to familiarise yourself with probability trees as they can be quite handy in such exercises.

Approaching the question

The first part was straightforward for those who were familiar with this section as it just requires knowledge of the conditional probability definition. Part (ii) can be done by either using Bayes formula or by a probability tree or even a good understanding of the conditional probability concept. The workout is given below:

i. - $P(G \cap H) = P(G)\, P(H \mid G)$
   - $= 0.02 \times 0.8 = 0.016$.

ii. - $P(G \mid H) = P(G \cap H)/P(H)$
    - $= 0.016/0.2 = 0.08$.

(d) Classify each one of the following variables as either measurable (continuous) or categorical. If a variable is categorical, further classify it as either nominal or ordinal. Justify your answer. (*Note that no marks will be awarded without a justification.*)

i. Classification of a university degree.

ii. Fuel consumption of a car.

iii. Eye colour.

iv. The cost of life insurance.

[8 marks]

Reading for this question

This question requires identifying types of variable so reading the relevant section in the subject guide (Section 3.6) is essential. Candidates should gain familiarity with the notion of a variable and be able to distinguish between discrete and continuous (measurable) data. In addition to identifying whether a variable is categorical or measurable, further distinctions between ordinal and nominal categorical variable should be made by candidates.

Approaching the question

A general tip for identifying continuous and categorical variables is to think of the possible values they can take. If these are finite and represent specific entities, the variable is categorical. Otherwise, if these consist of number corresponding to measurements, the data are continuous and the variable is measurable. Such variables may also have measurement units or can be measured to various decimal places.

i. The classification of a university degree can be 1st, 2.1, 2.2, 3 or fail in some countries. Clearly these values represent categories and by definition these classifications are ordered. Hence, this variable is a categorical ordinal variable.

ii. Fuel consumption is a variable that can be measured in miles/gallon or kilometres/litre to some decimal places. Hence it is a measurable variable.

iii. Each eye colour is a category, so the possible values are one for each colour. Hence, the variable is categorical. Note also that colours do not have a natural ordering, so this represents a categorical nominal variable.

iv. The cost of life insurance is a variable that can be measured in \$, £ etc. to two decimal places. Hence it is a measurable variable.

(e) **In the past, the mean telephone call time of customers to a computer helpline has been 16.0 minutes. The computer company conducts a training scheme for its telephone consultants with the intention of reducing this mean call time. After training, a random sample of 20 calling times had a sample mean of 14.3 minutes and a sample standard deviation of 5.0 minutes. Carry out a hypothesis test, at two suitable significance levels, to decide if the training scheme has been successful. State your hypotheses, the test statistic and its distribution under the null hypothesis, and your conclusion in the context of the problem.**

[7 marks]

Reading for this question

This question refers to a one-sided hypothesis test examining whether the telephone call time of customers to a computer helpline is less than 16.0 minutes. While the entire chapter on hypothesis testing is relevant, candidates can focus on the relevant sections for a single mean (7.12 and 7.13) and in particular 7.13. The question refers to one-sided hypothesis tests that are located in Section 7.10 of Chapter 7.

Approaching the question

It is essential to identify the type of hypothesis test required for this question. Since there is only one variable involved it will have to be a single mean test, and the test statistic can be found in the formula sheet. Make sure to substitute the relevant quantities carefully and avoid any numerical errors in the calculation.

The next step is to identify the distribution of the test statistic. The fact that a sample standard deviation is given, indicates that the variance is unknown. Hence, since $n < 30$ the t distribution should be used.

The remaining steps involve finding the critical values from the corresponding statistical table for the relevant significance levels, deciding whether to reject H_0, and interpreting the results in the context of the problem. The working of the exercise is given below:

- $\mathrm{H}_0 : \mu = 16$ vs. $\mathrm{H}_1 : \mu < 16$. (No $\bar{X}$s, accept $\mathrm{H}_0 : \mu \geq 16$.)
- Test statistic value:
$$\frac{\bar{x} - 16}{s/\sqrt{20}} = \frac{14.3 - 16}{5/\sqrt{20}} = -1.52.$$
- The variance is unknown and $n < 30$ so the t distribution should be used.
- For $\alpha = 0.05$, the critical value is -1.729.
- Decision: do not reject H_0.
- Choose larger α, say $\alpha = 0.1$, hence -1.328, hence reject H_0.
- Weak evidence that the training has been successful in reducing the mean call time.

(f) **The amount of coffee dispensed into a coffee cup by a coffee machine follows a normal distribution with mean 125 ml and standard deviation 8 ml.**

i. Find the probability that one cup is filled above the level of 137 ml.

ii. What is the proportion of cups with coffee contents between 117 ml and 133 ml?

[4 marks]

Reading for this question

This section examines the ideas of the normal random variable. Read the relevant section of Chapter 5 and work out the examples and activities of this section. The sample examination questions are quite relevant.

Approaching the question

The basic property of the normal random variable for this question is that if $X \sim N(\mu, \sigma^2)$, then $Z = (X - \mu)/\sigma \sim N(0, 1)$. 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 find the requested proportions:

i. • We can write:

$$P(X > 137) = P\left(\frac{X - 125}{8} > \frac{137 - 125}{8}\right) = P(Z > 1.5).$$

• Continuing from above, we get $P(Z > 1.5) = 1 - \Phi(1.5) = 1 - 0.9332 = 0.0668$.

ii. • We can write:

$$P(117 < X < 133) = P\left(\frac{117 - 125}{8} \leq Z \leq \frac{133 - 125}{8}\right) = P(-1 \leq Z \leq 1).$$

• Continuing from above, we get:

$$P(-1 \leq Z \leq 1) = \Phi(1) - \Phi(-1) = 0.8413 - 0.1586 = 0.6827.$$

(g) The variable X takes the values 1, 2, 3 and 5 according to the following distribution

x	**1**	**2**	**3**	**5**
$p_X(x)$	**0.1**	**0.3**	**0.4**	**0.2**

i. What is the probability that X is negative?

ii. Find E(X), the expected value of X.

iii. Find the probability that $X^2 > 8$.

[5 marks]

Reading for this question

This is another question on probability, exploring the concepts of relative frequency, conditional probability and probability distribution. Reading from Chapter 4 of the subject guide is suggested with a focus on the sections on these topics. Try Activity A4.1 and the exercises on probability trees.

Approaching the question

i. • X only takes positive values, so the probability is 0.

ii. •• $\text{E}(X) = \sum_i x_i \, P(X = x_i) = 1 \times 0.1 + 2 \times 0.3 + 3 \times 0.4 + 5 \times 0.2 = 2.9$.

iii. • The probability distribution of $Z = X^2$ will be:

Z	1	4	9	25
$p_Z(z)$	0.1	0.3	0.4	0.2

• Hence, the correct probability is 0.6.

Note that this part may be answered without deriving the probability distribution table of Z. One can note that only the values $X = 3$ and $X = 5$ will give $X^2 > 8$, hence the requested probability is $0.4 + 0.2 = 0.6$.

(h) State whether the following are true or false and give a brief explanation. (*Note that no marks will be awarded for a simple true/false answer.*)

i. The chance that a normal random variable is less than one standard deviation from its mean is 95%.

ii. Quota sampling is free of selection bias.

iii. Increasing the level of confidence will decrease the width of a confidence interval for a population mean (assuming that everything else remains constant).

iv. **Failing to reject a false null hypthesis is known as a Type I error.**

v. **In a chi-squared test of association, the larger the test statistic value, the larger the corresponding p-value.**

vi. **The upper quartile of a sample dataset is never smaller than the lower quartile.**

[12 marks]

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) concerns normal random variables that can be found in Chapter 5 of the subject guide. Part (ii) requires material from Chapter 9 and in particular Section 9.7 on types of sample, whereas Part (iii) is about correlation and regression (see Sections 11.8 and 11.9). Parts (iv) and (v) target specific concepts of hypothesis testing; namely Type I error (see Section 7.7) and a p-value (covered in Section 7.11) respectively. Finally, Part (vi) is on measure of spread and location that are located in Sections 3.8 and 3.9.

Approaching the question

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

i. False; it is approximately 68%.

ii. False. Selection is non-random and therefore introduces selection bias.

iii. False. A higher level of confidence would increase the z/t value making the interval wider.

iv. False. It is a Type II error.

v. False. The larger the test statistic value, the smaller the p-value.

vi. True. $Q_1 \leq Q_3$.

Section B

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

Question 2

(a) **The following data show the periods (in minutes) that a random sample of students needed to complete a statistics assignment:**

76	**59**	**93**	**87**	**38**
50	**56**	**123**	**45**	**67**
102	**34**	**54**	**85**	**85**
50	**44**	**33**	**51**	**40**
82	**92**	**79**	**38**	**86**
34	**29**	**107**	**63**	**46**

i. **Carefully construct a stem-and-leaf diagram of these data.**

ii. **Find the median and the quartiles.**

iii. **Comment on the data given the shape of the stem-and-leaf diagram without any further calculations.**

iv. **Name two other types of graphical displays that would be suitable to represent the data.**

[12 marks]

Reading for this question

Chapter 3 provides all the relevant material for this question. More specifically, reading on stem-and-leaf diagrams can be found in Section 3.7.4, but the entire Sections 3.7, 3.8 and 3.9 are highly relevant.

Approaching the question

i. A stem-and-leaf diagram, which was compatible with what the examiners were expecting to see, is shown below. Marks were awarded for including the title, correct labelling, vertical alignment and reasonable accuracy.

Stem-and-leaf diagram of time needed to complete a statistics assignment

Stem = 10s of minutes | Leaf = minutes

2	9
3	34488
4	0456
5	001469
6	37
7	69
8	25567
9	23
10	27
11	
12	3

ii. • Median = 57.5.

• $Q_1 \approx 44.25$. Note: Any reasonable quartile method was accepted.

• $Q_3 = 85$.

iii. Based on the shape of the boxplot you have drawn, we can see that the distribution of the data is positively skewed.

iv. A boxplot, histogram or dot plot are other types of suitable graphical displays. The reason for this is that the variable income is measurable and these graphs are suitable for displaying the distribution of such variables.

(b) **A random sample of 512 unionised workers found that 38 had been made redundant in the last twelve months. An independent random sample of 654 non-unionised workers found that 67 had been made redundant over the same period.**

i. **Give a 95% confidence interval for the difference in the rates of redundancy between unionised and non-unionised workers.**

ii. **Carry out a hypothesis test, at two suitable significance levels, to determine whether unionised workers are less likely to be made redundant compared to non-unionised workers. State the test hypotheses, and your test statistic and its distribution under the null hypothesis. Comment on your findings.**

iii. **State any assumptions you made in (ii.).**

[13 marks]

Reading for this question

Look up the sections in the subject guide about hypothesis testing and confidence intervals for differences in proportions; more specifically Sections 6.13, 7.14 and 7.15.

Approaching the question

i. Let p_1, n_1 refer to the proportion of redundant unionised workers and to the total number of unionised workers, respectively. Similarly, denote by p_2 and n_2 the corresponding quantities for non-unionised workers. The calculation for the confidence interval is straightforward given the formula sheet; make sure to be able to recognise the relevant formula. First, the standard error needs to be calculated:

$$\text{s.e.}(p_1 - p_2) = \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}} = 0.0166.$$

Then, the lower and upper bounds can be found to be -0.0604 and 0.0044, respectively. Finally, the above should be presented as an interval $(-0.0604,\ 0.0044)$.

ii. As before, let π_1 denote the proportion of unionised workers made redundant and π_2 the corresponding proportion for non-unionised workers. Also denote by p the overall proportion of redundant workers. Regarding hypotheses, note that the wording 'less likely' suggests an one sided test: $\text{H}_0 : \pi_1 = \pi_2$ vs. $\text{H}_1 : \pi_1 < \pi_2$. The next step is to identify the test statistic which is $(p_1 - p_2)/\text{s.e.}(p_1 - p_2)$, and follows a standard normal distribution.

$$\text{s.e.}(p_1 - p_2) = \sqrt{p(1-p)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)} = 0.0169.$$

Based on the above, the value of the test statistic is -1.6576. The critical value at the 5% level is -1.645, hence we reject H_0 at the 5% level. Testing at the 1% level gives a critical value of -2.323. Therefore, we do not reject H_0 concluding that there is moderate evidence that unionised workers are less likely to be made redundant.

iii.
- Sample size is large enough to justify the normality assumption.
- Equal variances.

Some candidates stated assumptions in this part that were not made in part (ii). Marks were not awarded in such cases.

Question 3

(a) **A survey was conducted to investigate the relationship between the frequency of newspaper readership and readers' educational background. The following table shows the results of this survey:**

	Educational background			
	Graduate	**A-levels**	**Less than A-levels**	**Total**
Low readership	**19**	**32**	**49**	**100**
Moderate readership	**25**	**52**	**23**	**100**
Frequent readership	**46**	**40**	**14**	**100**
Total	**90**	**124**	**86**	**300**

i. Based on the data in the table, and *without conducting a significance test*, would you say there is an association between the frequency of newspaper readership and reader's educational background?

ii. Calculate the χ^2 statistic and use it to test for independence, using two appropriate significance levels. What do you conclude?

[14 marks]

Reading for this question

This part targets Chapter 8 on contingency tables and chi-squared tests. Note that part (i.) of the question does not require any calculations, just understanding and interpreting contingency tables. Candidates can attempt Activity A8.4 to practise. Part (ii) is a straightforward chi-squared test and the reading is also given in Chapter 8. Look also at Activity A8.4.

Approaching the question

i. There are some differences in the distributions within readership levels. More specifically, graduates appear more frequent readers than low readership compared to those with lower educational attainment than A-levels (46% vs. 19% and 14% vs. 49%, respectively). For those with A-levels, most are of a moderate readership type (52%). Hence, there seems to be an association between readership levels and reader's educational background, although this needs to be investigated further. (Note: the conclusion of the last sentence must be stated to get full marks).

ii. Set out the null hypothesis that there is no association between readership level and educational background against the alternative – that there is an association. Be careful to get these the correct way round!

H$_0$: No association between readership level and educational background vs.

H$_1$: Association between readership level and educational background.

Work out the expected values to obtain the table below:

30.00	41.33	28.67
30.00	41.33	28.67
30.00	41.33	28.67

The test statistic formula is:

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

that gives a value of 41.350. This is a 3 × 3 contingency table so the degrees of freedom are $(3-1) \times (3-1) = 4$.

For $\alpha = 0.05$, the critical value from the chi-squared distribution with 4 degrees of freedom is 9.488, hence reject H$_0$.

Next, for $\alpha = 0.01$, the critical value is 13.277, hence reject H$_0$ again.

We conclude that there is strong evidence of an association between readership level and educational background.

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 do reject at the 5% level, but at 10%' 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 to be earned. Many candidates lost marks by missing out on follow-up like this.

(b) **i. Explain the difference between item non-response and unit non-response.**

ii. State any three factors which could cause non-response.

iii. A travel agency offers customers a range of ways to make holiday bookings – in store, online and through their call centres. To determine the level of customer satisfaction, the company's management has decided to use a survey of all types of customers and has asked you to devise an appropriate random sampling scheme. Explain in detail your recommendation, including how you might address non-response.

[11 marks]

Reading for this question

This question was on basic material on survey designs. 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. It is also a good idea to try the activities in Chapter 9.

Approaching the question

One of the main things to avoid in this part is to write essays without any structure. This exercise asks for specific things and each one of them requires 1 or 2 lines. If you are unsure of 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, keep in mind that **the answer should not be long**. Note also that in some cases there is no unique answer to the question.

The marking scheme and some model answers are given below:

i. • Item non-response occurs when a sampled member fails to respond to a question in the questionnaire.

• Unit non-response occurs when no information is collected from a sample member.

ii. **3 marks:** Any three of:

— Not-at-home.

— Refusals.

— Incapacity to respond.

— Not found

— Lost schedules.

iii. **6 marks:** Possible 'ingredients' of an answer:

— Sampling frame to be the travel agency's customer database.

— Propose stratified sampling since all types of customers are to be surveyed.

— Stratification factors could include booking method, gender, holiday type.

— Take a simple random sample from each stratum.

— Contact method: mail, phone or email (likely to have all details on database).

— Minimise non-response through suitable incentive, such as discount off next booking.

Question 4

(a) **An area manager in a department store wants to study the relationship between the number of workers on duty, x, and the value of merchandise lost to shoplifters, in \$. To do so, the manager assigned a different number of workers for each of 10 weeks. The results were as follows:**

Week	#1	#2	#3	#4	#5	#6	#7	#8	#9	#10
x	9	11	12	13	15	18	16	14	12	10
y	420	350	360	300	225	200	230	280	315	410

The summary statistics for these data are:

Sum of x data: 130	Sum of the squares of x data: 1760
Sum of y data: 3090	Sum of the squares of y data: 1007750
Sum of the products of x and y data: 38305	

i. **Draw a scatter diagram of these data on the graph paper provided. Label the diagram carefully.**

ii. **Calculate the sample correlation coefficient. Interpret your findings.**

iii. **Calculate the least squares line of y on x and draw the line on the scatter diagram.**

iv. **Based on the regression equation in part (iii.), what will be the predicted loss from shoplifting when there are 17 workers on duty? Will you trust this value? Justify your answer.**

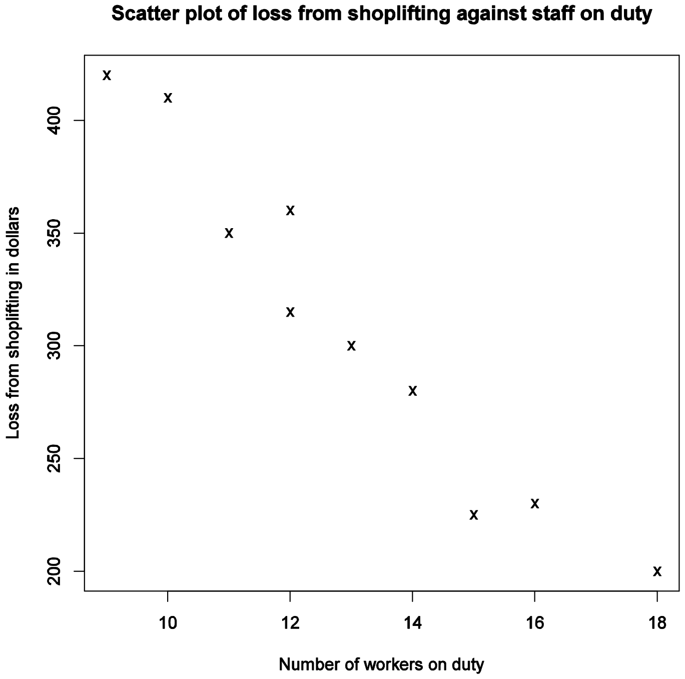
[13 marks]

Reading for this question

This is a standard regression question and the reading is to be found on Chapter 11. Section 11.6 provides details for scatter diagrams and is suitable for part (i) whereas the remaining parts are on correlation and regression that are covered in Sections 11.8–11.10 of the subject guide. Section 11.7 is also relevant. Sample examination question 2 in this chapter is recommended for practice on questions of this type.

Approaching the 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 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.



ii. The summary statistics can be substituted to the formula for the correlation (make sure you know which one it is!) to obtain the value −0.9688. An interpretation of this value is the following: The data suggest that the higher the number of workers, the lower the loss from shoplifters. The fact that the value is very close to −1, suggests that this is a strong linear negative association.

   Many candidates did not mention all three words (strong, linear, negative). Note that all of these words provide useful information on interpreting the association and are therefore required to obtain full marks.

iii. The regression line can be written by the equation $\widehat{y} = a + bx$ or $y = a + bx + \varepsilon$. 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 $b = -26.64$.

   The formula for a is $a = \bar{y} - b\bar{x}$, so we get $a = 655.36$.

   Hence the regression line can be written as $\widehat{y} = 655.36 - 26.64x$ or $y = 655.36 - 26.64x + \varepsilon$. It should also be plotted in the scatter diagram.

Many candidates reported **incorrectly** the regression line as $y = 655.36 - 26.64x$. This expression is false; one of the two above is required.

iv. The prediction will be $\widehat{y} = 655.36 - 26.64 \times 17 = \$202.48$. Yes we would trust this value, since this point is inside the observed range of x, and therefore the prediction is based on interpolation.

Many candidates did not give the measurement units here. These are essential in answering such questions and a mark is deducted if they are not specified. It is also important to provide the answer in two decimal places.

(b) **A study was conducted to determine the amount of hours spent on Facebook by university and high school students. For this reason, a questionnaire was administered to a random sample of 16 university and 14 high school students and the hours per day spent on Facebook were recorded. Summaries of the data are shown in the table below:**

	Sample size	**Sample mean**	**Sample variance**
University students	**16**	**2.9**	**0.9**
High school students	**14**	**2.1**	**1.1**

i. Use an appropriate hypothesis test to determine whether the mean hours per day spent on Facebook were different between university and high school students. Test at two suitable significance levels, stating clearly the hypotheses, the test statistic and its distribution under the null hypothesis. Comment on your findings.

ii. State clearly any assumptions you made in (i.).

iii. Adjust the procedure above to determine whether the mean hours spent per day on Facebook for university students is higher than that of high school students.

[12 marks]

Reading for this question

The first two parts of the question refer to a two-sided hypothesis test comparing proportions. While the entire chapter on hypothesis testing is relevant, one can of focus on the sections involving proportions (7.14 and 7.15), in particular 7.15. The last part of the question refers to one-sided hypothesis tests that are also located in these sections.

Approaching the question

i. Let μ_A denote the mean hours per day spent on Facebook for university students and μ_B the mean hours per day spent on Facebook for high school students.

The null hypothesis is that the proportions of the two population means (μ_A and μ_B) do not differ, the alternative is that they do.

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

If equal variances are assumed, the test statistic value is 2.1939 (the pooled variance is 0.9929). If equal variances are not assumed the test statistic value is 2.1788.

Since the variances are unknown and the sample size is not large enough, the t_{28} distribution is being used. The critical values at the 5% level are ± 2.048, hence we reject the null hypothesis. If we take a (smaller) α of 1%, the critical values are ± 2.763, so we do not reject H_0. We conclude that there is some but not strong (moderate) evidence of a difference in the mean hours spent on Facebook between university and high school students.

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 vs. z.
- Assumption about independent samples.

Some candidates stated assumptions in this part that were not made in part (i). Marks were not awarded in such cases.

iii. This case corresponds to an one-sided test, therefore the hypotheses would be $\text{H}_0 : \mu_A = \mu_B$ vs. $\text{H}_1 : \mu_A > \mu_B$. The test statistic is the same for this case but the critical values are now 1.701 for 5% and ≈ 2.467 for 1%. As before we reject H_0 at the 5% but not at the 1% level, and we conclude that there is moderate evidence (result is moderately significant) – university students spend more time on Facebook than high school students.