THE OPEN UNIVERSITY

Linear statistical modelling (M346) Diagnostic Quiz

[Press ↓ to begin]

Last Revision Date: March 2015

Version 1.4

1. Introduction

In embarking on Linear statistical modelling (M346), some basic competence in mathematics and statistics is required and a student who has completed Analysing data (M248) should have both. This quiz covers a number of key skills in these areas which are important for successful study of M346.

M346 follows Analysing data (M248), which is an excellent basis from which to start M346. You need a basic knowledge of the ideas of statistical science at the level of M248: a theoretical grounding is not expected, but you should have a conceptual understanding of basic topics and should be able to apply the ideas and interpret the answers they give. The topics include histograms, boxplots and scatterplots; normal, Poisson and binomial distributions; the central limit theorem; confidence intervals; hypothesis testing; simple linear regression; correlation. All these are reviewed in the early units of M346.

Some experience of using a statistical software package is expected, although you are not expected to be familiar with GenStat. If you used MINITAB in M248 you will have the necessary expertise. M346 also requires a mathematical knowledge similar to that expected in M248. You are expected to be familiar with mathematical notation, to be able to follow short algebraic arguments, to handle the logarithm and exponential functions, and to use formulas.

(continued on following page)

You will not be expected to follow complicated algebraic arguments or to produce a great deal of algebra in your written work.

Try each question, using your calculator where appropriate, then click on the green section letter (e.g. '(a)') to see the solution. Click on the □ symbol at the end of the solution to return to the question. Use the ↑ (or <PgUp>) and ↓ (or <PgDn>) keys to move from Section to Section.

There is some advice on evaluating your performance at the end of the quiz.

2. Summation notation

EXERCISE 1.

(a) Write out in full the expression $\sum_{i=0}^{3} i a_i x_i^2$.

(b) Write out in full the expression $\sum_{i=0}^{3} i a_i x_i^{-2}$.

EXERCISE 2. Let $x_1 = 2$, $x_2 = -1$ and $x_3 = 4$.

(a) Calculate $\sum_{i=1}^{3} x_i^2$.

(b) Calculate $\left(\sum_{i=1}^{3} x_i \right)^2$.

3. Logarithms

EXERCISE 3. This exercise is about logarithms to base e, sometimes called 'natural logarithms' and denoted by 'ln', '$\log_e$' or just 'log'.

(a) Use your calculator to write down $\log 3$ and $\log 4$, each to 4 decimal places.

(b) Without using the logarithm key on your calculator again, can you calculate $\log 12$ to 3 decimal places?

(c) Write out in full, and simplify as far as you can, the logarithm of the expression $2e^{4.5}\sqrt{x}$.

4. Numerical data summary

Exercise 4.

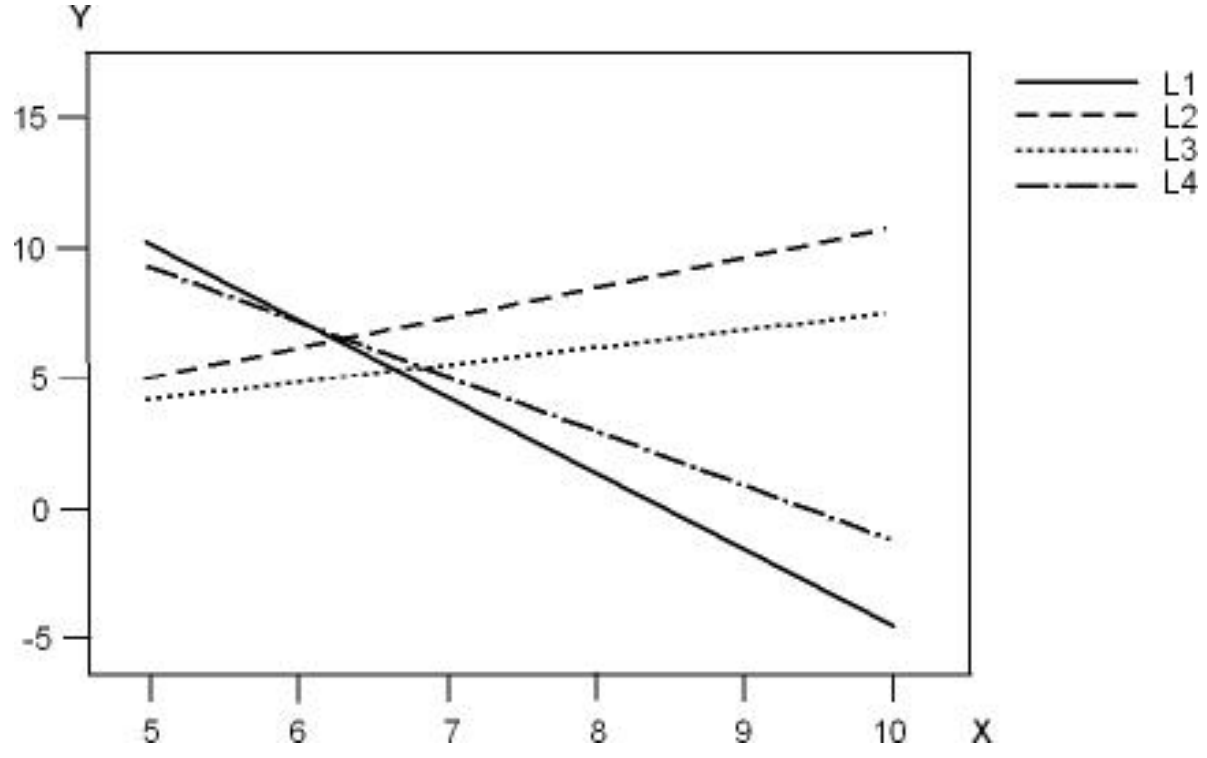
(a) Use your calculator to find the sample mean, sample variance and sample standard deviation of the five numbers 9, 13, 16, 8, 10. Where necessary, round your answers to an appropriate number of decimal places.

(b) Take the natural logarithms of the same five numbers and calculate the sample mean, sample variance and sample standard deviation of these logarithms.

(c) Is the mean of the logarithms the same as the logarithm of the mean?

5. Interpretation of graphs

EXERCISE 5. The graph below shows four straight lines.

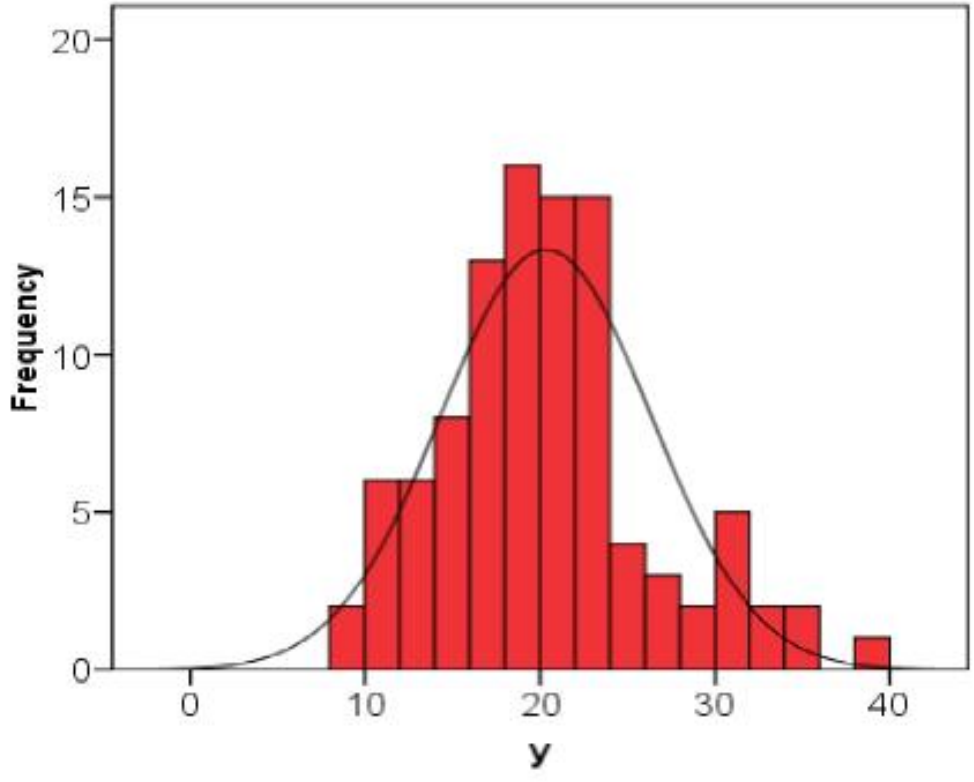


A: $Y = 20 - 2.1X$ B: $Y = 25 - 2.9X$
C: $Y = 1 + 0.6X$ D: $Y = -1 + 1.2X$

(a) Match lines L1 – L4 with equations A – D above.

Exercise 6.

The histogram below, with superimposed normal curve, displays a set of data containing 100 real, continuous values.

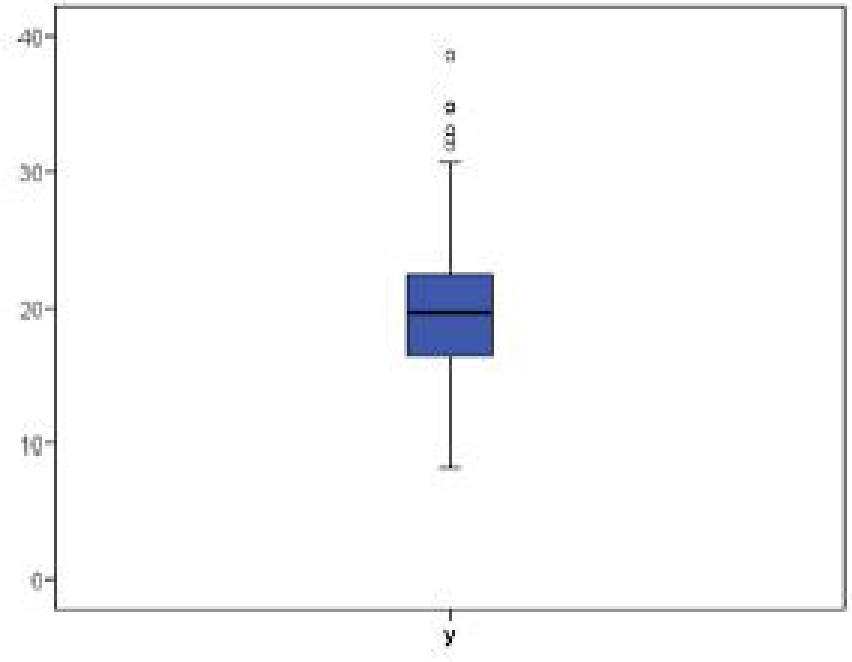


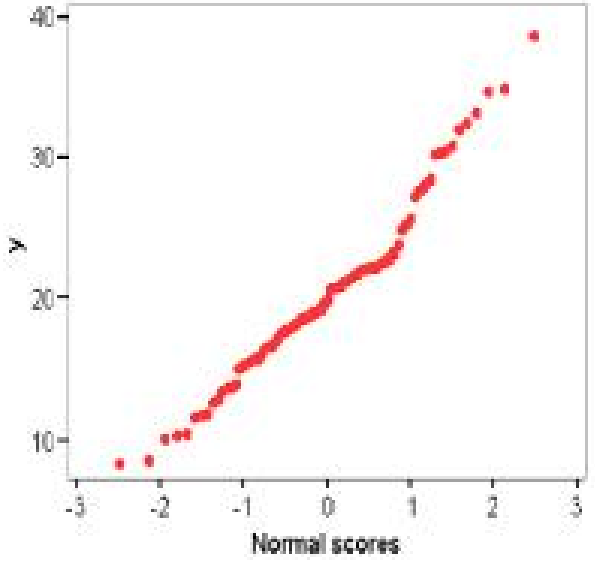
(a) Is the mean, to 2 decimal places, 20.33, 21.69 or 22.78?

(b) Is the standard deviation, to 2 decimal places, 3.79, 5.98 or 9.23?

Exercise 7.

The boxplot and normal plot below summarise the dataset used in Exercise 6 above.





(a) Referring to these two plots and the histogram in Exercise 6, would you say that the data come from a normal distribution? Give reasons for your choice.

6. Modelling with probability distributions

EXERCISE 8. In M346, the notation $N\left(\mu, \sigma^2\right)$ is used for the normal distribution with mean μ and variance σ^2. Similarly, $B\left(n, p\right)$ is used for the binomial distribution with parameters n (a positive integer representing the number of independent Bernoulli trials) and $p \in (0, 1)$ representing the probability of success in a single trial. Poisson(μ) refers to the Poisson distribution with mean and variance μ.

Classify the likely distributions applicable in the scenarios that follow as binomial, Poisson or normal.

(a) Weights of children in Manchester on entering primary school (aged 4–5).

(b) The number of days each week that Mary forgets to take her daily medication.

(c) Counts of vehicles passing my house in 5-minute periods in the mid-afternoon.

Exercise 9.

State the range, mean and standard deviation of

(a) the normal distribution $N(-2, 2)$;
(b) the binomial distribution $B(5, 0.1)$;
(c) the Poisson distribution with parameter 2.6.

Exercise 10.

(a) Calculate the probability that a random observation from the distribution in Exercise 9(a) is positive. (You may need to use statistical tables or a computer.)
(b) Calculate the probability that a random observation from the distribution in Exercise 9(b) is 1.
(c) Calculate the probability that a random observation from the distribution in Exercise 9(c) is less than 3.

EXERCISE 11.

Suppose that the distribution of X is $N(60, 80)$.

(a) What type of distribution does $Y = 100 - \frac{1}{5}X$ have?

(b) Calculate the mean of the distribution of Y.

(c) Calculate the variance of the distribution of Y.

7. Confidence intervals and hypothesis testing

EXERCISE 12. The dataset in Exercise 6 has size 100, mean 20.33 and standard deviation 5.98. The appropriate t-value for calculating a 95% confidence interval for the population mean in this case is 1.984.

(a) Calculate this 95% confidence interval. Does it matter that the data came from a distribution that is not normal?

(b) A 95% confidence interval for the mean of a population is $(7.2, 9.6)$. Which of the following statements is true?

A: The probability that the mean is between 7.2 and 9.6 is 0.95.
B: In repeated sampling, the interval calculated in the same way will include 8.4 nineteen times out of twenty.
C: In repeated sampling, the interval calculated in the same way will include the population mean nineteen times out of twenty.

Exercise 13.

(a) What does it mean to say that the significance probability of a statistical test is 0.045?

(b) The dataset in Exercise 4 has size $n_1 = 5$, mean $\overline{x}_1 = 11.2$ and variance $s_1^2 = 10.7$. Another dataset has size $n_2 = 10$, mean $\overline{x}_2 = 8.2$ and variance $s_2^2 = 14.6$. Assume that both samples come from normal populations with the same variance σ^2. Carry out a test of the hypothesis that the two population means are the same. (The critical value of t on 13 degrees of freedom at the 5% significance level is 2.16.)

8. Least squares and simple linear regression

EXERCISE 14. The table below gives the co-ordinates of six points.

X	0	0	1	1	3	5
Y	5.0	4.0	3.0	5.0	3.0	1.0

(a) Sketch a scatterplot of these six points.

(b) Which of the following three lines is most likely to be the least squares regression line of Y on X?

A: $Y = 4.5 - 1.02x$
B: $Y = 3.5$
C: $Y = 4.6 - 0.66x$

9. Post-mortem: Are you ready for M346?

You should be familiar with the techniques covered by these Exercises before embarking on Linear statistical modelling (M346). In particular, you should have a sound grasp of the concepts and confidence intervals and hypothesis testing. It doesn't matter too much if you haven't encountered boxplots and normal plots before.

You should note that, in this quiz, the datasets are artificial and small for simplicity's sake. In general, M346 uses more substantial and technical datasets from the real world.

The elementary mathematics in Exercises 1 - 3 can be found in the Level 1 module MST124 Essential Mathematics (or its predecessor MST121 Using Mathematics). The other questions are covered in the Level 2 statistics module Analysing Data (M248).

Do contact your Student Support Team via StudentHome if you have any queries about your suitability for M346.

Solutions to Exercises

Exercise 1(a)

$$\begin{aligned}\sum_{i=0}^{3} i a_i x_i^2 &= 0 \times a_0 x_0^2 + 1 \times a_1 x_1^2 + 2 \times a_2 x_2^2 + 3 \times a_3 x_3^2 \\ &= a_1 x_1^2 + 2 a_2 x_2^2 + 3 a_3 x_3^2.\end{aligned}$$

□

Exercise 1(b)

$$\begin{aligned}\sum_{i=0}^{3} i a_i x_i^{-2} &= 0 \times a_0 x_0^{-2} + 1 \times a_1 x_1^{-2} + 2 \times a_2 x_2^{-2} + 3 \times a_3 x_3^{-2} \\ &= \frac{a_1}{x_1^2} + \frac{2a_2}{x_2^2} + \frac{3a_3}{x_3^2}.\end{aligned}$$

□

Exercise 2(a)

$$\begin{aligned}\sum_{i=1}^{3} x_i^2 &= 2^2 + (-1)^2 + 4^2 \\ &= 4 + 1 + 16 \\ &= 21.\end{aligned}$$

□

Exercise 2(b)

$$\begin{aligned}\left(\sum_{i=1}^{3} x_i\right)^2 &= (2+(-1)+4)^2 \\ &= 5^2 \\ &= 25.\end{aligned}$$

□

Exercise 3(a) To the limits of calculator accuracy,

$$\log 3 = 1.098612289$$

and is therefore 1.0986 to 4 decimal places.

(If you get something different, check that you have used the right key on your calculator. In particular, if you have used the key which provides logarithms to base 10, you would get 0.4771, to 4 decimal places.)

Similarly, $\log 4 = 1.386294361 = 1.3863$, to 4 decimal places.

(Again, using the 'base 10' key would give 0.6021, to 4 decimal places.)

Exercise 3(b) Using the rule that, for any strictly positive numbers a and b, $\log(a \times b) = \log a + \log b$ yields

$$\begin{aligned} \log 12 &= \log(3 \times 4) \\ &= \log 3 + \log 4 \\ &= 1.0986 + 1.3863 \\ &= 2.485, \end{aligned}$$

to 3 decimal places.

Exercise 3(c) Here, we also need the rules that, for any real number a and any $b > 0$, $\log(e^a) = a$ and $\log(b^a) = a \log b$. This yields

$$\begin{aligned} \log\left(2e^{4.5}\sqrt{x}\right) &= \log 2 + 4.5 + 0.5 \log x \\ &= 5.193 + 0.5 \log x. \end{aligned}$$

(The first constant has been expressed to 4 significant figures.)

□

Exercise 4(a) The sample mean is

$$\begin{aligned} \overline{x} &= \frac{9 + 13 + \ldots + 10}{5} \\ &= \frac{56}{5} \\ &= 11.2, \end{aligned}$$

(exactly).

[Press ↓ or <PgDn> to continue.]

The sample variance is

$$\begin{aligned} s^2 &= \frac{1}{n-1}\sum_{i=1}^{n}(x_i-\overline{x})^2 \\ &= \frac{(9-11.2)^2+(13-11.2)^2+\ldots+(10-11.2)^2}{4} \\ &= \frac{42.8}{4} \\ &= 10.7, \end{aligned}$$

(exactly).

(There is an alternative formula which is easier for hand calculation but ex-M248 students may not have encountered it; the formula above will be more useful in M346 in developing understanding.)

The standard deviation is then $s=\sqrt{10.7}=3.3$, rounded to 1 decimal place. (The range of this small dataset is greater than 1 but less than 10, so there is no justification for more than 1 decimal place.)

□

Exercise 4(b) The natural logarithms are 2.197, 2.565, 2.773, 2.079 and 2.303. Their sum is 11.917 and the sum of squares is 28.720. The same formulas as before give sample mean 2.38, sample variance 0.080 and sample standard deviation 0.28.

In this case, the range of values is more than 0.1 but less than 1 and the data set is very small, so 2 decimal places is appropriate for the standard deviation and 2 significant figures for the variance.

(The calculations were done to the limits of calculator accuracy and the final answers rounded in each case. This is good practice.)

Exercise 4(c) $\log 11.2 = 2.42$, to 2 decimal places. This is not the same as 2.38.

The logarithmic function is not a linear transformation so a difference would, in general, be expected.

□

Exercise 5(a) Equation A is one of two with negative slopes. Its slope is the smaller of the two in modulus and so it corresponds to L4.

Equation B is one of two with negative slopes. Its slope is the larger of the two in modulus and so it corresponds to L1.

Equation C is one of two with positive slopes. Its slope is (marginally) the smaller of the two in modulus and so it corresponds to L3.

Equation D is one of two with positive slopes. Its slope is (marginally) the larger of the two in modulus and so it corresponds to L2.

(There are other ways of tackling these questions. You could, for example, work out the Y-values for, say, $X = 10$, where the lines are well separated, then read off the approximate corresponding values on the graph.)

Exercise 6(a) The mean is 20.33, as this corresponds with the peak of the normal curve.

□

Exercise 6(b) The standard deviation is 5.98. Most of the values are within 5.98 units of the mean, very roughly in the range $(14, 26)$. (Remember that, very approximately, 70% of values should lie within one standard deviation of the mean.)

The value 3.79 would give an interval of about $(17, 24)$; too many values are outside this range.

The value 9.23 would give an interval of about $(11, 30)$; too few values are outside this range.

□

Exercise 7(a) The distribution looks somewhat skew compared with the normal curve. There are ten values over 30, but only two below 10.

The boxplot shows several outliers at the high end only.

The normal scores plot is not straight and indicates outliers at the top end.

□

Exercise 8(a) This may well be normal. The data are real and continuous, with no fixed limits on their range (although there is a range within which one would sensibly expect weights to lie).

□

Exercise 8(b) This could be binomial. We would have $n = 7$ and p, the probability of Mary's forgetting to take her medication on any given day, would be (hopefully) small.

Exercise 8(c) This could be Poisson. The data are discrete (they are counts) and there is no fixed upper limit.

□

Exercise 9(a) The range is $(-\infty, \infty)$, the mean is -2 and the standard deviation is $\sqrt{2} = 1.414$, to 3 decimal places.

Exercise 9(b) The range is $\{0, 1, 2, 3, 4, 5\}$, the mean is $np = 0.5$ and the standard deviation is $\sqrt{np(1-p)} = \sqrt{0.45} = 0.671$, to 3 decimal places.

□

Exercise 9(c) The range is $\{0, 1, 2, \ldots\}$, the mean is 2.6 and the standard deviation is $\sqrt{2.6} = 1.612$, to 3 decimal places.

□

Exercise 10(a)

$$\begin{aligned} 1 - \Phi\left(\frac{0-(-2)}{\sqrt{2}}\right) &= 1 - \Phi\left(1.414\right) \\ &= 1 - 0.9213 \\ &= 0.0787, \end{aligned}$$

using normal probability tables.

□

Exercise 10(b)

$$\begin{pmatrix} 5 \\ 1 \end{pmatrix} p\,(1-p)^4 = 5 \times 0.1 \times 0.9^4$$
$$= 0.32805,$$

(exactly).

□

Exercise 10(c)

$$\begin{aligned} P(0) + P(1) + P(2) &= e^{-2.6}\left(1 + \frac{2.6}{1!} + \frac{2.6^2}{2!}\right) \\ &= e^{-2.6}(1 + 2.6 + 3.38) \\ &= 0.5184, \end{aligned}$$

to 4 decimal places.

□

Exercise 11(a) The distribution of $Y = 100 - \frac{1}{5}X$ is normal.

□

Exercise 11(b) The mean of the distribution of Y is

$$100 - \frac{60}{5} = 88,$$

(exactly).

□

Exercise 11(c) The variance of the distribution of Y is

$$\frac{80}{5^2} = 3.2,$$

(exactly).

□

Exercise 12(a) The 95% confidence limits are

$$\begin{aligned} \overline{x} \pm \left(1.984 \times \frac{5.98}{\sqrt{100}} \right) &= 20.33 \pm 1.19 \\ &= (19.14, 21.52) \end{aligned}$$

(Note that the mean was only given to 2 decimal places, so we cannot claim any further accuracy in the answer.)

The sample size here is 100, which is sufficient for the Central Limit Theorem to apply to the sampling distribution of the sample mean $\overline{x}$. It makes little difference in such a case whether the t-distribution or the normal distribution is used. (The corresponding critical value for the normal distribution is 1.960 instead of 1.984.)

Exercise 12(b) Option C is correct. You might think of this as entitling you to make a claim that the population mean is in the interval $(7.2, 9.6)$, with the proviso that 5% of statements made on a similar basis will turn out to be false.

(Can you see that Option A is saying something rather different from this?)

Option B is not correct. The valid interpretation of the confidence interval is a statement about the true (but unknown) value of the population mean; there is no reason to expect it to be the mid-point of this or any other interval calculated by a similar procedure.

□

Exercise 13(a) There will be a null hypothesis. If the null hypothesis is true, then the result that has been obtained is rather unusual, meaning that the probability of a result as extreme as, or more extreme than, the one obtained has the (fairly small) probability of 0.045. This provides some evidence against the null hypothesis, but the evidence is not very strong.

□

Exercise 13(b) The pooled estimate of variance is

$$s^2 = \frac{(4 \times 10.7) + (9 \times 14.6)}{4 + 9} = 13.4.$$

The value of the test statistic is $t = \dfrac{\overline{x}_1 - \overline{x}_2}{\sigma\sqrt{\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$, which is estimated as

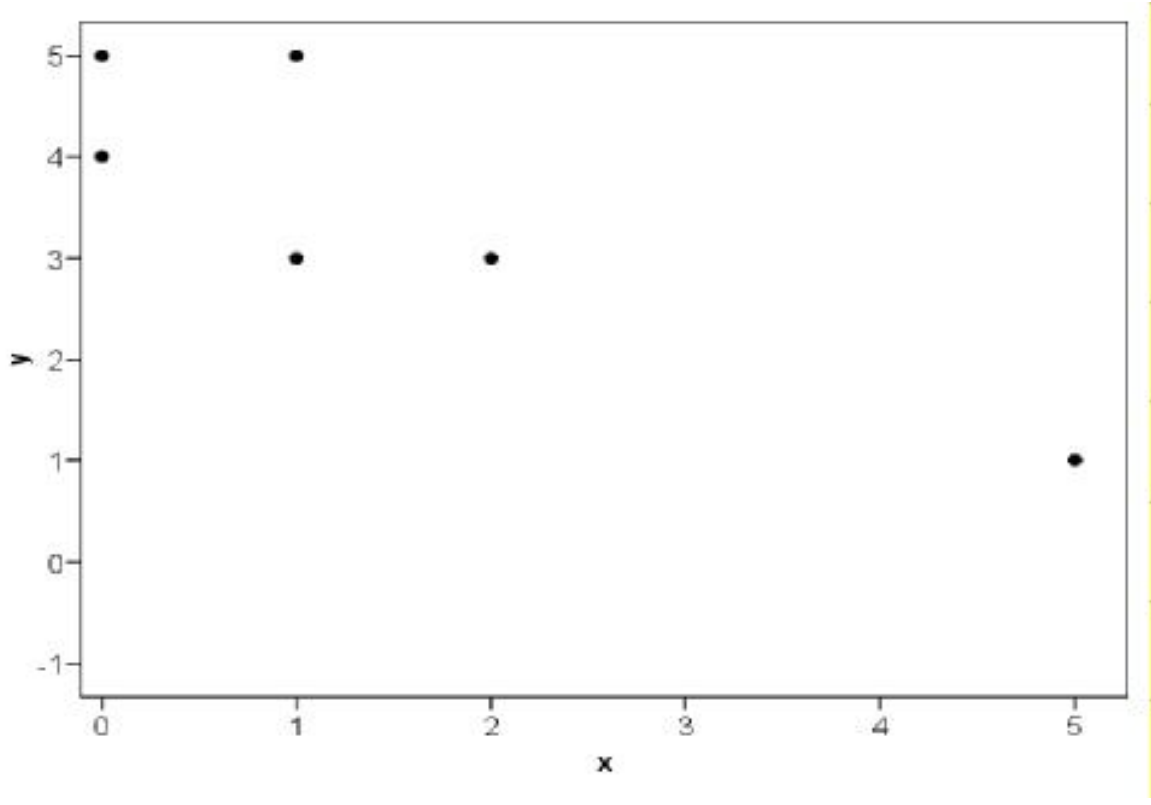
$$\frac{11.2 - 8.2}{\sqrt{13.4}\sqrt{\left(\frac{1}{5} + \frac{1}{10}\right)}} = 1.496,$$

to 3 decimal places.

This is smaller than the given critical value, so there is no evidence against the null hypothesis of equal population means.

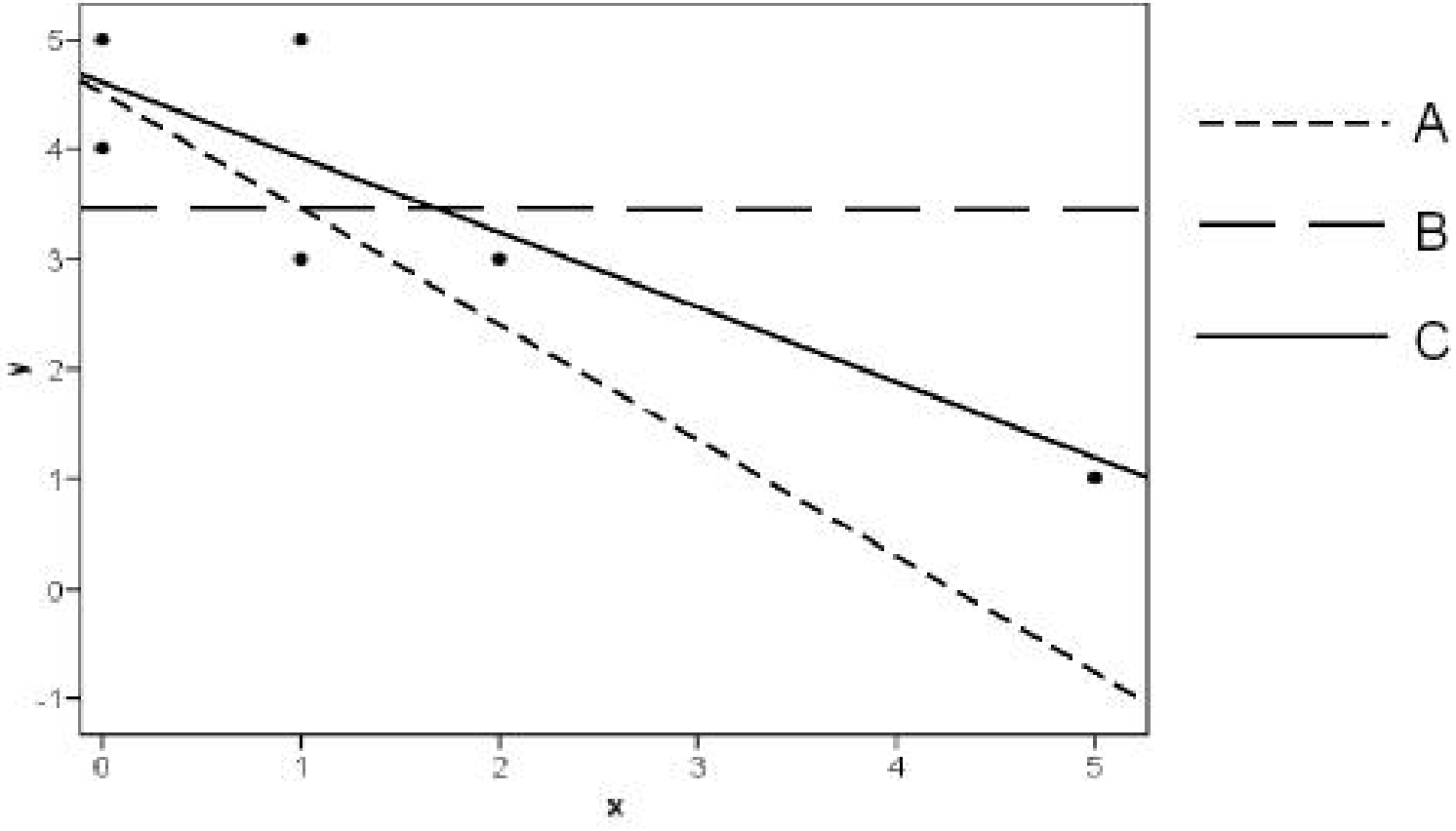
Hence, at the 5% significance level, the null hypothesis is not rejected. □

Exercise 14(a) The scatterplot is shown below.



□

Exercise 14(b) The scatterplot with the three lines superimposed is shown below.



[Press ↓ or <PgDn> to continue.]

The least squares regression line always passes through the point $(\overline{x}, \overline{y}) = \left(1.66\dot{6}, 3.5\right)$. Lines B and C both go through this point but line A does not. Hence line A cannot be the least squares regression line.

Line B is too flat and does not appear to minimise the sum of squared (vertical) deviations. Thus it seems that the line most likely to be the least squares regression line of Y on X is line C: $Y = 4.6 - 0.66x$.

□