Subject CS1

CMP Upgrade 2020/21

CMP Upgrade

This CMP Upgrade lists the changes to the Syllabus objectives, Core Reading and the ActEd material since last year that might realistically affect your chance of success in the exam. It is produced so that you can manually amend your 2020 CMP to make it suitable for study for the 2021 exams. It includes replacement pages and additional pages where appropriate. Alternatively, you can buy a full set of up-to-date Course Notes / CMP at a significantly reduced price if you have previously bought the full-price Course Notes / CMP in this subject. Please see our 2021 *Student Brochure* for more details.

This CMP Upgrade contains:

- all significant changes to the Syllabus objectives and Core Reading.
- additional changes to the ActEd Course Notes and Assignments that will make them suitable for study for the 2021 exams.

1 Changes to the Syllabus objectives

Only minor formatting changes have been made to the Syllabus objectives.

2 Changes to the Core Reading

This section contains all the ***non-trivial*** changes to the Core Reading.

Chapter 10

Section 8.2, Page 46

The notation used in the formula for calculating expected frequencies has been changed for clarity. The paragraph now reads as follows:

The simple rule for calculating the expected frequency for any cell is then:

$$\frac{\textbf{row total} \times \textbf{column total}}{\textbf{table total}}$$

(*ie* the proportion of data in row *i* is $\sum_j f_{ij} \Big/ \sum_i \sum_j f_{ij}$ so if the criteria are independent, the number expected in cell (i, j) is $\left(\sum_j f_{ij} \Big/ \sum_i \sum_j f_{ij} \right) \times \sum_i f_{ij}$.)

3 Changes to the ActEd material

This section contains all the ***non-trivial*** changes to the ActEd text.

Chapter 2

Page 12

Near the top of the page, in the solution to part (ii) of the question, although the answer is correct, the formula in the calculation is incorrect. The calculation should be:

$$P(X=0) \approx \binom{5}{0} p^0 q^5 = \left(\frac{28}{58}\right)^5 = 0.0262$$

Page 62

In Question 2.13 part (a), the ranges given for relating r to the simulated value do not have the correct inequalities. They should be:

$$n = \begin{cases} 0 & \text{if } 0 \le r \le 0.55 \\ 1 & \text{if } 0.55 < r \le 0.8 \\ 2 & \text{if } 0.8 < r \le 0.95 \\ 3 & \text{if } 0.95 < r \le 1 \end{cases}$$

Page 62

In Question 2.13 part (b), the inequalities in the solution are incorrect. The answer to part (b) should read:

Since $0.55 < 0.6221 \le 0.8$, the first simulated value is 1. Since $0 < 0.1472 \le 0.55$, the second simulated value is 0. Since $0.95 < 0.9862 \le 1$, the third simulated value is 3.

Chapter 4

Page 11

At the bottom of the page, the final part to the question should be part (iii) not part (ii).

Page 12

In the solution to part (iii) on this page, the range given for x and y is incorrect. It should be:

$$0 < y < x < 2$$

Page 50

Question 4.15 has been removed.

Pages 59 and 60

Solution 4.15 has been removed.

Chapter 5

Page 5

The conditional expectation in the solution is incorrect as the limits on the integral are wrong. The solution should be:

$$E(Y \mid X = x) = \int_0^x y \, \frac{2}{3}\left(\frac{1}{x} + \frac{y}{x^2}\right) dy$$

$$= \int_0^x \frac{2y}{3x} + \frac{2y^2}{3x^2} \, dy = \left[\frac{y^2}{3x} + \frac{2y^3}{9x^2}\right]_0^x = \frac{x^2}{3x} + \frac{2x^3}{9x^2} = \frac{5}{9}x \qquad 0 < x < 2$$

Page 14

The following question has been added as Question 5.6:

5.6 (i) Two discrete random variables, X and Y, have the following joint probability function:

		X			
		1	2	3	4
Y	1	0.2	0	0.05	0.15
	2	0	0.3	0.1	0.2

Determine $\text{var}(X \mid Y = 2)$. [3]

(ii) Let U and V have joint density function:

$$f_{U,V}(u,v) = 6\left(2uv - u^2\right) \qquad 0 < u < v < 1$$

Determine $E(U \mid V = v)$. [3]

[Total 6]

Page 18

The following solution has been added as Solution 5.6:

5.6 (i) ***Conditional variance***

$$\text{var}(X\,|\,Y=2)=E(X^2\,|\,Y=2)-E^2(X\,|\,Y=2)$$

$$E(X\,|\,Y=2)=\sum xP(X=x\,|\,Y=2)=\sum x\frac{P(X=x\cap Y=2)}{P(Y=2)}$$

$$=1\times\frac{0}{0.6}+2\times\frac{0.3}{0.6}+3\times\frac{0.1}{0.6}+4\times\frac{0.2}{0.6}$$

$$=2\frac{5}{6}$$ [1]

$$E(X^2\,|\,Y=2)=\sum x^2P(X=x\,|\,Y=2)=\sum x^2\,\frac{P(X=x\cap Y=2)}{P(Y=2)}$$

$$=1^2\times\frac{0}{0.6}+2^2\times\frac{0.3}{0.6}+3^2\times\frac{0.1}{0.6}+4^2\times\frac{0.2}{0.6}$$

$$=8\frac{5}{6}$$ [1]

So $\text{var}(X\,|\,Y=2)=8\frac{5}{6}-\left(2\frac{5}{6}\right)^2=\frac{29}{36}=0.80556$. [1]

(ii) ***Conditional expectation***

We require:

$$E(U\,|\,V=v)=\int_u u\,f(u\,|\,v)\,du$$

Now:

$$f(v)=\int_{u=0}^{v}6(2uv-u^2)\,du=6\left[u^2v-\tfrac{1}{3}u^3\right]_{u=0}^{v}=6\left[v^3-\tfrac{1}{3}v^3\right]=4v^3$$ [1]

$$\Rightarrow\quad f(u\,|\,v)=\frac{f(u,v)}{f(v)}=\frac{6(2uv-u^2)}{4v^3}=\frac{2uv-u^2}{\frac{2}{3}v^3}\quad\text{for } 0<u<v<1$$ [1]

So:

$$E(U\,|\,V=v)=\int_{u=0}^{v}\frac{2u^2v-u^3}{\frac{2}{3}v^3}\,du=\left[\frac{\frac{2}{3}u^3v-\frac{1}{4}u^4}{\frac{2}{3}v^3}\right]_{u=0}^{v}=\frac{\frac{2}{3}v^4-\frac{1}{4}v^4}{\frac{2}{3}v^3}=\frac{5}{8}v\quad\text{for } 0<v<1$$ [1]

Chapter 7

Page 29

In the solution to Question 7.5(ii), the first indented equation line has been corrected to read:

$$\frac{U/m}{V/n} \sim F_{m,n}$$

Chapter 11

Page 20

Some clarification has been added about the inference under Spearman's rank correlation. This is given below.

Under the null hypothesis that the variables are uncorrelated:

$$\frac{r_s\sqrt{n-2}}{\sqrt{1-r_s^2}} \div t_{n-2} \quad \text{provided } n > 20$$

For very large values of n, the sampling distribution of r_s can be approximated by the $N\left(0, \frac{1}{n-1}\right)$ distribution.

4 Changes to the X Assignments

Overall

Some of the questions have been modified slightly to bring them into line with the style used by the IFoA. Other minor wording changes were also made. See 2021 versions for new wording.

More significant changes are listed below.

Assignment X1

Solution 1.2

The following alternative solution has been added to part (ii):

Alternatively, let N be the random variable that represents the number of calls arriving between 8am and 9am. Then $N \sim Poi(2.5)$ and $P(N=0)=\dfrac{2.5^0 e^{-2.5}}{0!}=e^{-2.5}=0.08208$. *[2]*

Solution 1.6

The following alternative solution has been added to part (ii)(b):

Alternatively, using the probability function:

$$P(X=1)=\binom{1,000}{1}\times 0.01^1\times 0.99^{999}=0.000436$$

$$P(X=2)=\binom{1,000}{2}\times 0.01^2\times 0.99^{998}=0.00220$$

$$\Rightarrow \quad P(X>2)=1-P(X\le 2)=1-[P(X=0)+P(X=1)+P(X=2)]=0.997$$

The alternative solution for part (ii)(c) using the Poisson distribution has been deleted.

Question 1.9

The paragraph before part (ii) has been updated. The question from this paragraph onwards now reads:

It is subsequently discovered that the random variables, X and Y, are in fact continuous over the intervals $0 < x < 2$ and $0 < y < 2$. Their joint probability density function has the same structure as the joint probability function above, *ie*:

$$f(x,y) = \begin{cases} k(x+2y) & 0<x<2 \text{ and } 0<y<2 \\ 0 & \text{otherwise} \end{cases}$$

where k is an appropriate constant.

(ii) Determine the conditional distribution of X given $Y = y$. [3]

[Total 6]

Solution 1.9

The solution to part (ii) has been adjusted. The constant is now represented by k instead of c. The solution now reads:

Using $\int_x \int_y f(x,y)\,dxdy = 1$ gives:

$$\int_{x=0}^{2} \int_{y=0}^{2} k(x+2y)\,dxdy = \int_{x=0}^{2} k\left[xy+y^2\right]_{y=0}^{2} dx = \int_{x=0}^{2} k(2x+4)\,dx = 1$$

So:

$$k\left[x^2+4x\right]_{x=0}^{2} = 12k = 1 \quad \Rightarrow \quad k = \frac{1}{12} \qquad [1]$$

Now $f_{X|Y=y}(x,y) = \dfrac{f_{X,Y}(x,y)}{f_Y(y)}$:

$$f_Y(y) = \int_{x=0}^{2} \frac{1}{12}(x+2y)\,dx = \frac{1}{12}\left[\frac{1}{2}x^2+2xy\right]_{x=0}^{2} = \frac{2}{12}(1+2y) \qquad [1]$$

$$f_{X|Y=y}(x,y) = \frac{\frac{1}{12}(x+2y)}{\frac{2}{12}(1+2y)} = \frac{x+2y}{2(1+2y)} \qquad 0<x<2 \qquad [1]$$

[Total 3]

Note: Again students do not actually need to calculate the value of the constant, k, to get the answer and full marks.

Assignment X2

Question 2.9

Part (i) has been split into two parts worth 2 marks and 1 mark. Part (ii) (now part (iii)) has also been split into (a) and (b) worth 1 mark each. The question now reads:

A random variable X has probability density function:

$$2e^{-2(x-\theta)} \qquad x \geq \theta$$

where the value of θ is unknown.

(i) Derive a formula for the maximum likelihood estimator of θ based on a sample of n values. [2]

Five observations of X are:

1.90, 2.97, 1.88, 2.94 and 1.56

(ii) Calculate the maximum likelihood estimate for this sample. [1]

(iii) (a) Show that $E(X) = \theta + \frac{1}{2}$.

(b) Calculate the method of moments estimate of θ. [2]

(iv) Comment briefly on your results. [1]

[Total 6]

Solution 2.9

The solution has been restructured in line with the changes to the question. Mark allocations have been adjusted to confirm with the new structure. The solution now reads:

(i) ***Maximum likelihood estimator***

The likelihood function based on a sample of n observations is given by:

$$L = \prod_{i=1}^{n} 2e^{-2(x_i-\theta)} = 2^n e^{-2(\sum x_i - n\theta)} \text{ , provided } x_1, \cdots, x_n \geq \theta \quad (ie \min_i x_i \geq \theta) \qquad [1]$$

So $\hat{\theta}$, the maximum likelihood estimate of θ, is the value of θ that maximises $2^n e^{-2(\sum x_i - n\theta)}$ subject to the condition that $\theta \leq \min x_i$ (otherwise the likelihood is zero).

When $\theta \leq \min x_i$, we have:

$$L = 2^n e^{-2(\sum x_i - n\theta)} \Rightarrow \ln L = n\ln 2 - 2(\sum x_i - n\theta) \Rightarrow \frac{d}{d\theta}\ln L = 2n > 0 \qquad [½]$$

ie L increases as θ increases.

So the maximum likelihood estimate of θ is the highest value of θ subject to the condition that $\theta \le \min x_i$, and hence $\hat{\theta} = \min X_i$ is the maximum likelihood estimator of θ. [½]

[Total 2]

(ii) ***Maximum likelihood estimate***

From this sample, the maximum likelihood estimate of θ is 1.56. [1]

(iii)(a) ***Expectation***

Since $X - \theta \sim Exp(2)$:

$$E[X] = E[X - \theta] + \theta = \frac{1}{2} + \theta$$ [1]

Alternatively, from first principles:

$$E(X) = \int_{\theta}^{\infty} 2xe^{-2(x-\theta)}\, dx$$

$$= \left[-xe^{-2(x-\theta)} \right]_{\theta}^{\infty} + \int_{\theta}^{\infty} e^{-2(x-\theta)}\, dx \qquad \text{by parts}$$

$$= \theta + \left[-\frac{1}{2} e^{-2(x-\theta)} \right]_{\theta}^{\infty} = \theta + \frac{1}{2}$$

(iii)(b) ***Method of moments estimate***

Equating $E(X)$ to the sample mean, $\bar{x}$, gives:

$$\tilde{\theta} + \frac{1}{2} = \bar{x} = \frac{1}{n} \sum x_i = 2.25$$ [½]

So the method of moments estimate of θ is $2.25 - 0.5 = 1.75$. [½]

[Total 2]

(iv) ***Comment***

One of the observed values is less than the method of moments estimate of θ. So the method of moments gives an estimate of θ that is not 'possible' in this case. [½]

This contrasts with the situation for maximum likelihood estimators, which, provided they exist, must, by definition, give feasible estimates. [½]

[Total 1]

Assignment X3

Question changes

The order of the questions has been changed. A new question has been added as 3.9 in 2021.

Replacement pages for the questions from page 1 and for the solutions from page 7 are included at the end of this document.

Question 3.11 in 2021 (Question 3.9 in 2020)

Part (iv) of the question has been updated to ask for the test to be carried out using the *t* approximation instead of the normal approximation.

Solution 3.11 in 2021 (Solution 3.9 in 2020)

The solution for part (iv) has been updated to reflect the change to the question.

Question 3.10 in 2021 (Question 3.12 in 2020)

Parts (ii) and (iii) from the 2020 question have been deleted.

Assignment X4

Solution 4.11

The solution to part (i)(b) has been adjusted to remove the circumflexes on the parameters until it is stated that we assume that the values that solve the simultaneous equations relate to a maximum.

The solution to part (i)(b) now reads:

(i)(b) ***Maximum likelihood estimates***

Differentiating with respect to α:

$$\frac{\partial}{\partial\alpha}\ln L(\alpha,\beta)=\sum_{i=1}^{10}\left\{\frac{1}{\alpha}-y_i\right\}+\sum_{i=11}^{15}\left\{\frac{1}{\alpha+\beta}-y_i\right\} \;=\; \frac{10}{\alpha}+\frac{5}{\alpha+\beta}-\sum_{i=1}^{15}y_i \qquad [1]$$

Differentiating with respect to β:

$$\frac{\partial}{\partial\beta}\ln L(\alpha,\beta)=\sum_{i=11}^{15}\left\{\frac{1}{\alpha+\beta}-y_i\right\} \;=\; \frac{5}{\alpha+\beta}-\sum_{i=11}^{15}y_i \qquad [1]$$

Setting these two derivatives equal to zero, we obtain the equations:

$$\frac{10}{\alpha}+\frac{5}{\alpha+\beta}-\sum_{i=1}^{15}y_i=0 \qquad \text{eqn (3)}$$

$$\frac{5}{\alpha+\beta}-\sum_{i=11}^{15}y_i=0 \qquad \text{eqn (4)}$$

From (4), we have $\frac{5}{\alpha+\beta}=\sum_{i=11}^{15} y_i$. Substituting this into (3) gives:

$$\frac{10}{\alpha}-\sum_{i=1}^{10} y_i=0 \quad \Rightarrow \quad \alpha=\frac{10}{\sum_{i=1}^{10} y_i} \qquad [1]$$

We can rearrange (4) to get:

$$\alpha+\beta=\frac{5}{\sum_{i=11}^{15} y_i}$$

Now using $\alpha=\frac{10}{\sum_{i=1}^{10} y_i}$ we find that:

$$\beta=\frac{5}{\sum_{i=11}^{15} y_i}-\frac{10}{\sum_{i=1}^{10} y_i} \qquad [1]$$

Assuming these relate to a maximum, the maximum likelihood estimates are therefore:

$$\hat{\alpha}=\frac{10}{\sum_{i=1}^{10} y_i} \quad \text{and} \quad \hat{\beta}=\frac{5}{\sum_{i=11}^{15} y_i}-\frac{10}{\sum_{i=1}^{10} y_i}$$

To actually show these are maximum likelihood estimates, we should really do some second-order differentiation. However, this is complicated in a two-parameter case and would not be expected in the exam.

5 Other tuition services

In addition to the CMP you might find the following services helpful with your study.

5.1 Study material

We also offer the following study material in Subject CS1:

- Flashcards
- Revision Notes
- ASET (ActEd Solutions with Exam Technique) and Mini-ASET
- Mock Exam and AMP (Additional Mock Pack).

For further details on ActEd's study materials, please refer to the 2021 *Student Brochure*, which is available from the ActEd website at **www.ActEd.co.uk**.

5.2 Tutorials

We offer the following (face-to-face and/or online) tutorials in Subject CS1:

- a set of Regular Tutorials (lasting four full days)
- a Block (or Split Block) Tutorial (lasting four full days)
- a Paper B preparation day
- a 5-day Bundle (a Regular or Block Tutorial combined with a Paper B preparation day)
- an Online Classroom.

For further details on ActEd's tutorials, please refer to our latest *Tuition Bulletin*, which is available from the ActEd website at **www.ActEd.co.uk**.

5.3 Marking

You can have your attempts at any of our assignments or mock exams marked by ActEd. When marking your scripts, we aim to provide specific advice to improve your chances of success in the exam and to return your scripts as quickly as possible.

For further details on ActEd's marking services, please refer to the 2021 *Student Brochure*, which is available from the ActEd website at **www.ActEd.co.uk**.

5.4 Feedback on the study material

ActEd is always pleased to get feedback from students about any aspect of our study programmes. Please let us know if you have any specific comments (*eg* about certain sections of the notes or particular questions) or general suggestions about how we can improve the study material. We will incorporate as many of your suggestions as we can when we update the course material each year.

If you have any comments on this course, please send them by email to **CS1@bpp.com**.

Questions 1 to 4 are based on the data set given in the table below:

x	2	4	5	9
y	3.0	6.8	8.2	15.1

$$\sum x = 20 \quad \sum x^2 = 126 \quad \sum xy = 210.1 \quad \sum y = 33.1 \quad \sum y^2 = 350.49$$

It is proposed to fit a simple linear regression model to these data:

$$Y_i = a + bx_i + e_i \text{ where } e_i \sim N(0, \sigma^2)$$

X3.1 Calculate the least squares estimate for the slope parameter b. [1]

X3.2 Calculate an unbiased estimate of the variance parameter. [1]

X3.3 Determine a 95% confidence interval for b based on the sample data. [2]

X3.4 (i) Show that the sample correlation coefficient is 0.9995 to 4 significant figures. [1]

(ii) Calculate:

(a) a 95% confidence interval for ρ, the underlying correlation coefficient

(b) the coefficient of determination, R^2, commenting on your result. [5]

[Total 6]

X3.5 It is desired to test the value of the parameter p for a random variable that has a binomial distribution. In order to test the null hypothesis $H_0: p = 0.4$ against the alternative hypothesis $H_1: p = 0.6$, the following test is devised:

The number of successes, X, in a sample of size 50 is determined. If $X \geq 25$, then H_0 is rejected.

Calculate the approximate size of this test. [3]

X3.6 Define the following terms:

(i) a Type I error [1]

(ii) a Type II error [1]

(iii) the size of a test [1]

(iv) the power of a test. [1]

[Total 4]

X3.7 A random sample from a $N(\mu,\sigma^2)$ distribution, where both parameters are unknown, gave the following values:

11.8, 5.4, 8.2, 4.6, 13.6, 10.1, 10.4, 11.2, 12.2, 17.5

Test each of the hypotheses:

(i) $\mu = 9$ [3]

(ii) $\sigma^2 = 8$ [3]

against an appropriate two-sided alternative. [Total 6]

X3.8 Support for the current government is assessed by means of a survey of 5,000 people. Of those questioned 2,185 said that they would vote for the current government in the next election.

(i) Test whether the proportion of people that support the government is significantly greater than 42%. [3]

Following a rather embarrassing scandal a second survey is commissioned. This time, 1,191 out of 3,000 people said that they would vote for the current government in the next election.

(ii) Test to see if there has been any significant change in the proportion supporting the current government. [3]

[Total 6]

X3.9 A researcher is investigating the impact of a new memory technique purported to improve recall. The researcher asks 10 volunteers to complete a memory test before and after teaching them the technique. The table below summaries the scores (out of a maximum of 100) for each participant:

Volunteer	1	2	3	4	5	6	7	8	9	10
Before (B)	45	75	63	80	54	67	39	79	59	83
After (A)	50	70	72	91	53	80	55	94	65	95

(i) Perform a t test to investigate the claim that the new technique improves recall, stating any assumptions made. [5]

(ii) Calculate a 99% confidence interval for the average change in the score after the technique is applied. [2]

[Total 7]

X3.10 A research chemist has discovered a new desiccant that may be more efficient at extracting moisture from chemicals than the existing one. In order to test the claim, equal amounts of a homogeneously mixed compound are put into each of sixteen desiccators. These are divided into two batches of eight, labelled A and B, and in each batch the desiccators are numbered 1 to 8. Into each desiccator is also put a standard amount of the respective desiccant under test. Batch A contains the existing desiccant, whilst the new desiccant is placed in Batch B. The desiccators are sealed for 24 hours and then the increase in weight in grams of each of the sixteen samples of desiccant is measured. The results are:

Sample number		1	2	3	4	5	6	7	8
Existing desiccant	(A)	4.59	5.05	4.49	5.33	4.66	4.98	5.67	5.23
New desiccant	(B)	4.75	5.03	4.66	5.56	4.90	4.88	5.80	5.33

$$\sum A = 40.0 \quad \sum A^2 = 201.1574 \quad \sum B = 40.91 \quad \sum B^2 = 210.3659$$

(i) (a) Construct a dotplot of each data set.

(b) Compare the two dotplots. [2]

(ii) (a) Test whether the variances arising from the use of each desiccant are significantly different.

(b) Carry out a t test to investigate the claim that the new desiccant extracts more moisture than the existing one. [6]

[Total 8]

X3.11 A study was carried out into the effects of smoking on life expectancy. The average number (x) of cigarettes smoked per day from age 50 by 11 individuals was calculated and the number (y) of years from age 50 until their deaths was recorded. The results were as follows:

x	0.0	1.1	17.3	10.6	25.1	5.2	11.8	40.0	15.6	13.8	3.6
y	42.3	30.7	26.3	36.8	8.9	25.1	10.8	10.0	25.2	17.2	29.1

For these data values:

$$\sum_{i=1}^{11} x_i = 144.1, \sum_{i=1}^{11} y_i = 262.4$$

$$\sum_{i=1}^{11} x_i^2 = 3{,}255.91, \sum_{i=1}^{11} y_i^2 = 7{,}481.26, \sum_{i=1}^{11} x_i y_i = 2{,}495.43$$

(i) Calculate Pearson's correlation coefficient, commenting on the value obtained. [2]

(ii) Calculate Spearman's rank correlation coefficient, commenting on the value obtained. [3]

(iii) State a general advantage of using Spearman's rank correlation coefficient. [1]

(iv) (a) Test whether Spearman's rank correlation coefficient is significantly different from zero assuming it is reasonable to use an appropriate t distribution approximation.

(b) Comment on this assumption. [3]

(v) Calculate Kendall's rank correlation coefficient. [3]

[Total 12]

X3.12 1,000 male and 1,000 female subjects were chosen at random by a researcher and cross-classified according to sex and to whether or not they were colour-blind, giving the following table:

	male	female
normal	908	993
colour-blind	92	7

(i) Perform a χ^2 test on this contingency table to show that there is overwhelming evidence against the hypothesis that there is no association between an individual's sex and whether or not the individual is colour-blind. [5]

A genetic model states that the human population is split in the proportions illustrated in the following table, where q $(0 < q < 1)$ is a parameter relating to the distribution of the colour-blindness defect among the relevant genes.

	male	female
normal	$\frac{1-q}{2}$	$\frac{1-q^2}{2}$
colour-blind	$\frac{q}{2}$	$\frac{q^2}{2}$

The maximum likelihood estimate of q calculated from these data is 0.0895.

(ii) Test the goodness of fit of this model to the data. [5]

[Total 10]

X3.13 Following archaeological excavations at a site in Egypt, ten samples of wood were carbon-dated and their ages x (years) estimated as:

4,900	4,750	4,820	4,710	4,760
4,570	4,300	4,680	4,800	4,670

$$\sum x = 46,960 \qquad \sum x^2 = 220,772,800$$

(i) Calculate a 95% confidence interval for the true mean age of the wood found at this site. [3]

(ii) (a) Construct a dotplot of these data points.

(b) Comment on the validity of the confidence interval calculated in part (i). [2]

Ideally the archaeologist would like the 95% confidence interval for the true mean age, calculated in (i) above, to have a width of no more than 200 years.

(iii) Calculate the minimum sample size needed. [3]

At a second site, eight samples of wood gave the following results:

$$\sum y = 36,000 \quad \sum y^2 = 162,280,000$$

(iv) Calculate a 95% confidence interval for the difference between the mean ages of the wood found at the two sites. [3]

(v) (a) Calculate a 90% confidence interval for the ratio of the underlying variances in the ages of the two samples of wood.

(b) Comment on the validity of the confidence interval given in part (iv). [4]

[Total 15]

X3.14 It is thought that a plumber charges £22 per hour plus an administrative charge of £15 per call-out.

A sample of eight invoices was obtained corresponding to jobs with durations of 1 hour, 2 hours, ... , 8 hours. For each invoice the total cost of the job was noted with the following results:

Time x (hours):	1	2	3	4	5	6	7	8
Cost y (£):	40	50	81	89	122	128	151	179

$$\sum(x-\bar{x})^2 = 42 \quad \sum(y-\bar{y})^2 = 16,492 \quad \sum(x-\bar{x})(y-\bar{y}) = 826$$

The following model is used to represent the data:

$$Y_i = a + bx_i + e_i$$

where Y_i $(i = 1,\ 2, \ldots, 8)$ are the costs, x_i $(i = 1, 2, \ldots, 8)$ are the fixed times and e_i $(i = 1, 2, \ldots, 8)$ are independent errors with a $N(0, \sigma^2)$ distribution.

(i) (a) Derive formulae for the least squares estimators of a and b.

(b) Explain how your answer to part (i)(a) would have differed if you had been asked to calculate the maximum likelihood estimators, justifying your answer. [6]

(ii) Calculate the regression coefficients $\hat{a}$ and $\hat{b}$. [2]

(iii) Test the hypothesis that the slope parameter is equal to £22 per hour. [4]

(iv) Calculate a 90% confidence interval for the:

(a) average cost of a job lasting 4 hours

(b) cost of an individual job lasting 6 hours. [6]

(v) Comment on relative widths of the two intervals calculated in part (iv). [1]

[Total 19]

END OF PAPER

Comparing this with the percentage points of the standard normal distribution (given on page 162 of the *Tables*), we see that the *p*-value is $2 \times 0.022\% = 0.04\%$ (since it is a two-sided test). So we have sufficient evidence to reject H_0 at the 0.1% level, therefore it is reasonable to assume that the proportion of those who would vote for the current government is not the same as before the scandal. [1]

[Total 3]

Alternatively, testing at the 5% significance level gives critical values of ± 1.96*, and testing at the 1% significance level gives critical values of* ± 2.5758*. Since the test statistic is more extreme than the critical values, we reject the null hypothesis at both the 5% and 1% significance levels and reach the conclusion stated above. Technically, we would have to carry out a one-sided test to say it was less than before.*

Solution X3.9

(i) ***Paired t test***

We perform a paired t test, looking at the differences. Assuming the differences follow a normal distribution, we use the following test statistic:

$$\frac{\bar{D}-\mu_D}{S_D/\sqrt{n}} \sim t_{n-1} \qquad [1]$$

Markers: award only half a mark if the assumption that the differences follow a normal distribution is not stated.

Let D represent the difference in score, so that $D = A - B$. The observed values of D are:

$$5, \quad -5, \quad 9, \quad 11, \quad -1, \quad 13, \quad 16, \quad 15, \quad 6, \quad 12 \qquad [1]$$

We are testing:

$$H_0 : \mu_D = 0 \quad vs \quad H_1 : \mu_D > 0$$

The test is one-sided as we are specifically investigating whether the technique improves recall.

Calculating the sample mean and variance of the differences:

$$\bar{d} = \frac{1}{10}\sum_i d_i = 8.1$$

$$s_D^2 = \frac{1}{9}\left\{\sum d_i^2 - n\bar{d}^2\right\} = \frac{1}{9}\{1{,}083 - 10 \times 8.1^2\} = 47.4 \qquad [1]$$

Under H_0, $\dfrac{\bar{D}-0}{S_D/\sqrt{10}}$ follows the t_9 distribution. The observed value of the test statistic is:

$$\frac{8.1}{\sqrt{47.4/10}} = 3.72 \qquad [1]$$

The upper 5% point of the t_9 distribution is 1.833. As $3.72 > 1.833$, there is sufficient evidence to reject H_0 at the 5% level (and indeed at the 0.5% level). This suggests that the technique does improve recall. [1]

[Total 5]

(ii) ***Confidence interval***

Assuming the differences follow a normal distribution, the relevant pivotal quantity is:

$$\frac{\bar{D}-\mu_D}{S_D/\sqrt{n}} \sim t_{n-1} \qquad [½]$$

Now, using the t_9 distribution:

$$0.99 = P\left(-3.250 < \frac{\bar{D} - \mu_D}{S_D / \sqrt{n}} < 3.250 \right) \qquad [½]$$

ie $$0.99 = P\left(\bar{D} - 3.250\frac{S_D}{\sqrt{n}} < \mu_D < \bar{D} + 3.250\frac{S_D}{\sqrt{n}} \right)$$

Using the values we calculated in part (i), a 99% confidence interval for μ_D is therefore:

$$\bar{d} \pm 3.250\frac{s_D}{\sqrt{n}} = 8.10 \pm 3.250 \times \frac{\sqrt{47.4}}{\sqrt{10}} = (1.02, 15.2) \qquad [1]$$

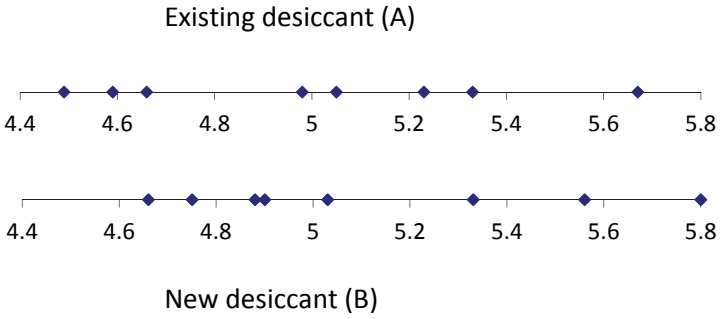
[Total 2]

Solution X3.10

Two sample tests on means, variances and paired data are covered in Chapter 10.

(i)(a) ***Dotplots***

The dotplots are as follows:



[1]

(i)(b) ***Comparison***

The plots suggest that the new desiccant may extract more water. The spread of values is similar for each desiccant. [1]

[Total 2]

(ii)(a) ***Test equality of variances***

We use the F test to test for equality of variance. The hypotheses are:

$$H_0 : \sigma_A^2 = \sigma_B^2 \quad vs \quad H_1 : \sigma_A^2 \neq \sigma_B^2$$

Under H_0, the statistic $\dfrac{S_A^2}{S_B^2} \Big/ \dfrac{\sigma_A^2}{\sigma_B^2}$ follows the $F_{7,7}$ distribution.

Calculating the sample variances:

$$s_A^2 = \frac{1}{7}\{201.1574 - 8 \times 5^2\} = 0.16534 \qquad [½]$$

$$s_B^2 = \frac{1}{7}\{210.3659 - 8 \times 5.11375^2\} = 0.16606 \qquad [½]$$

So the observed value of the test statistic is $\dfrac{0.16534}{0.16606} \Big/ 1 = 0.9957$. [1]

Since this lies between the critical values of 4.995 and 0.2002, we have insufficient evidence to reject H_0 at the 5% level. Therefore it is reasonable to assume that the two population variances are equal. This is not surprising, since the spreads of the two data plots look very similar. [1]

Alternatively, we could have used $\dfrac{S_B^2}{S_A^2}\Big/\dfrac{\sigma_B^2}{\sigma_A^2}$ *to get a ratio of 1.0043. This also has an* $F_{7,7}$ *distribution and we would draw the same conclusion as before.*

(ii)(b) ***Test if new desiccant extracts more moisture***

We now use a two-sample t test. The hypotheses are:

$$H_0: \mu_A = \mu_B \quad vs \quad H_1: \mu_B > \mu_A$$

This test is one-sided. The sample means are $\bar{A} = 5,\ \bar{B} = 5.11375$.

$$s_P^2 = \frac{7s_A^2 + 7s_B^2}{14} = \frac{2.3198}{14} = 0.1657 \qquad [1]$$

The observed value of the test statistic is:

$$T = \frac{(\bar{B} - \bar{A}) - (\mu_B - \mu_A)}{\sqrt{s_P^2\left(\frac{1}{8} + \frac{1}{8}\right)}} = \frac{0.11375}{\sqrt{0.041425}} = 0.559 \qquad [1]$$

We compare this with the 5% point of the t_{14} distribution, which is 1.761. As the observed value of the test statistic is (much) lower than 1.761, we have insufficient evidence to reject H_0 at the 5% significance level. It does not appear that the new desiccant extracts any more moisture than the existing one. [1]

[Total 6]

Solution X3.11

Correlation is covered in Chapter 11.

(i) ***Pearson's correlation coefficient***

For these data values:

$$s_{xx} = \sum x^2 - n\bar{x}^2 = 3{,}255.91 - \frac{144.1^2}{11} = 1{,}368.2$$

$$s_{yy} = \sum y^2 - n\bar{y}^2 = 7{,}481.26 - \frac{262.4^2}{11} = 1{,}221.8$$

$$s_{xy} = \sum xy - n\bar{x}\bar{y} = 2{,}495.43 - \frac{144.1 \times 262.4}{11} = -942.01 \quad [½]$$

So the correlation coefficient is:

$$r = \frac{s_{xy}}{\sqrt{s_{xx}s_{yy}}} = -\frac{942.01}{\sqrt{1{,}368.2 \times 1{,}221.8}} = -0.7286 \quad [½]$$

This value indicates a reasonably strong negative linear relationship. [1]

[Total 2]

(ii) ***Spearman's rank correlation coefficient***

For these data values we have:

x	y	Rank x	Rank y	d	d^2
0	42.3	1	11	−10	100
1.1	30.7	2	9	−7	49
17.3	26.3	9	7	2	4
10.6	36.8	5	10	−5	25
25.1	8.9	10	1	9	81
5.2	25.1	4	5	−1	1
11.8	10.8	6	3	3	9
40	10	11	2	9	81
15.6	25.2	8	6	2	4
13.8	17.2	7	4	3	9
3.6	29.1	3	8	−5	25

Summing, we get:

$$\sum d^2 = 388 \qquad [1]$$

Therefore Spearman's rank correlation coefficient is:

$$r_S = 1 - \frac{6\sum d^2}{n(n^2-1)} = 1 - \frac{6\times 388}{11(11^2-1)} = -0.7636 \qquad [1]$$

This value indicates a reasonably strong monotonically decreasing relationship (as we would have expected given the answer to (i)). [1]

[Total 3]

(iii) ***Advantage of Spearman's rank correlation coefficient***

Since Spearman's rank correlation coefficient considers ranks rather than the actual values, the value of the coefficient is less affected by outliers in the data than Pearson's correlation coefficient. Hence the Spearman's rank correlation coefficient is more robust. [1]

(iv)(a) ***Test***

We are testing:

H_0 : there is no association between age at death and the average number of cigarettes smoked per day

H_1 : there is some association between age at death and the average number of cigarettes smoked per day

Assuming that it is reasonable to use an appropriate t distribution approximation we have:

$$\frac{r_S\sqrt{n-2}}{\sqrt{1-r_S^2}} \div t_{n-2}$$

Here we are using r_S to represent the estimator of Spearman's rank correlation coefficient rather than the estimate that we calculated in part (iii).

The observed value of the test statistic is:

$$\frac{r_S\sqrt{n-2}}{\sqrt{1-r_S^2}} = \frac{-0.7636\sqrt{11-2}}{\sqrt{1-(-0.7636)^2}} = -3.548 \qquad [1]$$

Here we are using r_S to represent the estimate we calculated in part (iii).

This is a two-tailed test. The lower 2.5% point of the t_9 distribution is -2.262. Since $-3.548 < -2.262$, we have sufficient evidence to reject H_0 at the 5% level. It appears that there is some negative correlation. [1]

(iv)(b) ***Comment***

We should only use this approximation when we have a large enough sample size. So we should be cautious about conclusions drawn from this approximate test. [1]

[Total 3]

(v) ***Kendall's rank correlation coefficient***

Writing the data in order of the x values, and showing the ranks for both x and y, we get:

x	y	Rank x	Rank y
0	42.3	1	11
1.1	30.7	2	9
3.6	29.1	3	8
5.2	25.1	4	5
10.6	36.8	5	10
11.8	10.8	6	3
13.8	17.2	7	4
15.6	25.2	8	6
17.3	26.3	9	7
25.1	8.9	10	1
40	10	11	2

[½]

We need to calculate the number of concordant and discordant pairs. From the Course Notes, we have 'The concordant pairs (C) are the number of observations below the current one in the table that have a higher rank for the y and the discordant pairs (D) are the number of observations below which have a lower rank for the y'.

So:

x	y	Rank x	Rank y	C	D
0	42.3	1	11	0	10
1.1	30.7	2	9	1	8
3.6	29.1	3	8	1	7
5.2	25.1	4	5	3	4
10.6	36.8	5	10	0	6
11.8	10.8	6	3	3	2
13.8	17.2	7	4	2	2
15.6	25.2	8	6	1	2
17.3	26.3	9	7	0	2
25.1	8.9	10	1	1	0
40	10	11	2	n/a	n/a
			Sum	12	43

[1½]

For example for $x=0$, the y rank is 11. In the 'Rank y' column there are no values higher than 11, hence C is 0. Similarly, in the 'Rank y' column all ten values are lower than 11, hence D is 10.

From this we can see that $n_c=12$ and $n_d=43$, so the Kendall's rank correlation coefficient is:

$$\tau=\frac{n_c-n_d}{n(n-1)/2}=\frac{12-43}{11\times 10/2}=-0.5636$$ [1]

[Total 3]

Alternatively, we can consider each pair individually. The table showing the concordant and discordant pairs is as follows:

(x, y)	0, 42.3	1.1, 30.7	17.3, 26.3	10.6, 36.8	25.1, 8.9	5.2, 25.1	11.8, 10.8	40, 10	15.6, 25.2	13.8, 17.2	3.6, 29.1
0,42.3		d	d	d	d	d	d	d	d	d	d
1.1,30.7			d	c	d	d	d	d	d	d	d
17.3,26.3				d	d	c	c	d	c	c	d
10.6,36.8					d	c	d	d	d	d	c
25.1,8.9						d	d	c	d	d	d
5.2,25.1							d	d	c	d	d
11.8,10.8								d	c	c	d
40,10									d	d	d
15.6,25.2										c	d
13.8,17.2											d
3.6,29.1											

From this table we can see that $n_c = 12$ *and* $n_d = 43$*, so again the Kendall's rank correlation coefficient is:*

$$\tau = \frac{n_c - n_d}{n(n-1)/2} = \frac{12-43}{11\times 10/2} = -0.5636$$

Solution X3.12

Goodness-of-fit tests and contingency tables are covered in Chapter 10.

(i) ***Contingency table test***

We are testing:

H_0 : there is no association between sex and colour-blindness (*ie* they are independent)

H_1 : there is an association between sex and colour-blindness (*ie* they are not independent)

The expected frequencies for the 2×2 contingency table are:

	male	female	
normal	950.5	950.5	1,901
colour-blind	49.5	49.5	99
	1,000	1,000	2,000

[1]

Since $\dfrac{\text{(row total)}\times\text{(column total)}}{\text{grand total}}=\dfrac{1{,}901\times1{,}000}{2{,}000}=950.5$, *etc.*

The number of degrees of freedom is $(2-1)\times(2-1)=1$. [1]

The observed value of the test statistic is:

$$\sum\frac{(O-E)^2}{E}=\frac{(908-950.5)^2}{950.5}+\frac{(993-950.5)^2}{950.5}+\frac{(92-49.5)^2}{49.5}+\frac{(7-49.5)^2}{49.5} \quad [1]$$

$$=1.9003+1.9003+36.490+36.490$$

$$=76.78 \quad [1]$$

Since this exceeds even the 0.05% critical value of 12.12, we have overwhelming evidence at the 0.05% level to reject H_0. We therefore conclude that there **is** an association between sex and colour-blindness. [1]

[Total 5]

(ii) ***Goodness-of-fit test***

Using the maximum likelihood estimate of $\hat{q} = 0.0895$ in the formulae given in the question, we obtain the following numbers:

	male	female
normal	910.5	992.0
colour-blind	89.5	8.010

[1]

We are testing:

H_0 : the model is a good fit.

H_1 : the model is not a good fit.

Notice here that we are no longer dealing with a contingency table, despite the presentation of the data. We are doing a χ^2 goodness-of-fit test.

The number of degrees of freedom is $(\text{no. of groups}) - 1 - (\text{no. of parameters estimated})$. This is $4 - 1 - 1 = 2$ since q was estimated using the data. [1]

The observed value of the test statistic is:

$$\sum \frac{(O-E)^2}{E} = \frac{(908-910.5)^2}{910.5} + \frac{(993-992)^2}{992} + \frac{(92-89.5)^2}{89.5} + \frac{(7-8.01)^2}{8.01} \quad [1]$$

$$= 0.00686 + 0.00101 + 0.06983 + 0.12735$$

$$= 0.205 \quad [1]$$

Since this is less than the 5% critical value of 5.991, we have insufficient evidence at the 5% level to reject H_0. We therefore conclude that the model *is* a good fit. [1]

[Total 5]

Solution X3.13

The confidence intervals used in this question are covered in Chapter 9.

(i) ***Confidence interval for mean***

Let the true mean age of wood found at this site be μ. Then:

$$\frac{\bar{X}-\mu}{S/\sqrt{n}} \sim t_{n-1} \qquad [½]$$

Now, using the t_9 distribution:

$$0.95 = P\left(-2.262 < \frac{\bar{X}-\mu}{S/\sqrt{n}} < 2.262\right) \qquad [½]$$

ie $$0.95 = P\left(\bar{X} - 2.262\frac{S}{\sqrt{n}} < \mu < \bar{X} + 2.262\frac{S}{\sqrt{n}}\right)$$

From our sample:

$$\bar{x} = \frac{46,960}{10} = 4,696 \qquad [½]$$

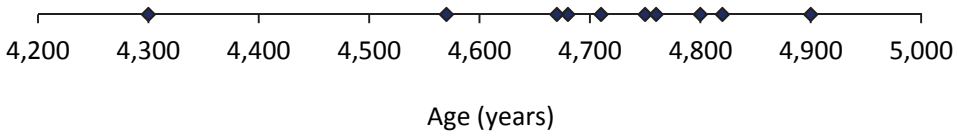
$$s^2 = \frac{1}{9}\{220,772,800 - 10 \times 4,696^2\} = 27,627 \qquad [½]$$

Therefore, a 95% confidence interval for μ is $4,696 \pm 119 = (4,577,\ 4,815)$. [1]

[Total 3]

(ii)(a) ***Dotplot***

The dotplot for the ages of the sample:



[1]

(ii)(b) ***Comment***

Given that the data set is small, the confidence interval in part (i) requires that the ages are normally distributed. [½]

The plot seems to show that 4,300 is very different to the other values and so it may be an outlier. In which case the underlying distribution is not normal, and our confidence interval is not valid. However, more data is needed for us to be sure. [½]

[Total 2]

(iii) ***Minimum sample size needed***

We require:

$$2\times t_{n-1;0.025}\frac{s}{\sqrt{n}}\le 200 \quad\Rightarrow\quad \frac{t_{n-1;0.025}}{\sqrt{n}}\le 0.60164 \qquad [1]$$

Trial and error leads to $\frac{t_{13;0.025}}{\sqrt{14}}=0.5773$ and $\frac{t_{12;0.025}}{\sqrt{13}}=0.6043$. Therefore a sample size of at least 14 is required. [2]

[Total 3]

(iv) ***Confidence interval for difference between means***

We will use μ_X and μ_Y to denote the true mean age of wood at the first and second sites respectively. Then:

$$\frac{(\bar{X}-\bar{Y})-(\mu_X-\mu_Y)}{\sqrt{S_P^2\left(\frac{1}{n_X}+\frac{1}{n_Y}\right)}}\sim t_{n_X+n_Y-2} \qquad [½]$$

Now, using the t_{16} distribution:

$$0.95=P\left(-2.120<\frac{(\bar{X}-\bar{Y})-(\mu_X-\mu_Y)}{\sqrt{S_P^2\left(\frac{1}{10}+\frac{1}{8}\right)}}<2.120\right) \qquad [½]$$

ie $$0.95=P\left(196-2.120\sqrt{S_P^2\left(\frac{1}{10}+\frac{1}{8}\right)}<(\mu_X-\mu_Y)<196+2.120\sqrt{S_P^2\left(\frac{1}{10}+\frac{1}{8}\right)}\right)$$

Now:

$$s_Y^2=\frac{1}{7}\{162,280,000-8\times 4,500^2\}=40,000$$

$$s_P^2=\frac{9\times 27,627+7\times 40,000}{10+8-2}=\frac{528,640}{16}=33,040 \qquad [1]$$

Therefore a 95% confidence interval for $\mu_X-\mu_Y$ is $196\pm 182.8=(13.2,\ 379)$. [1]

[Total 3]

(v)(a) ***Confidence interval for ratio of variances***

Using $\frac{S_X^2 / S_Y^2}{\sigma_X^2 / \sigma_Y^2} \sim F_{n_X-1, n_Y-1}$, we have an $F_{9,7}$ distribution with upper and lower bounds of 3.677 and $1/3.293$. [1]

This gives:

$$\frac{1}{3.293} < \frac{27{,}627/40{,}000}{\sigma_X^2/\sigma_Y^2} < 3.677$$

or $$\frac{27{,}627/40{,}000}{3.677} < \frac{\sigma_X^2}{\sigma_Y^2} < 27{,}627/40{,}000 \times 3.293$$ [1]

This gives a confidence interval for $\frac{\sigma_X^2}{\sigma_Y^2}$ of $(0.188, 2.27)$. [1]

(v)(b) ***Comment***

Since this confidence interval contains 1, the assumption of equal variances used in the confidence interval in part (iv) looks reasonable. [1]

[Total 4]

Solution X3.14

The material tested in this question is covered in Chapter 12.

(i)(a) ***Derive least squares estimators***

Since $Y_i = a + bx_i + e_i \Rightarrow e_i = Y_i - a - bx_i$, the sum of squares is:

$$Q = \sum e_i^2 = \sum (Y_i - a - bx_i)^2$$

Differentiating with respect to a and b, we have:

$$\frac{\partial Q}{\partial a} = \sum 2(Y_i - a - bx_i) \times (-1)$$

$$\frac{\partial Q}{\partial b} = \sum 2(Y_i - a - bx_i) \times (-x_i)$$

[1]

Setting these two expressions to zero, we have:

$$\sum Y = na + b\sum x \quad \ldots (1)$$

$$\sum xY = a\sum x + b\sum x^2 \quad \ldots (2)$$

Multiplying equation (1) by $\sum x$ and equation (2) by n, we obtain:

$$\sum x \sum Y = na\sum x + b\left(\sum x\right)^2 \quad \ldots (3)$$

$$n\sum xY = na\sum x + nb\sum x^2 \quad \ldots (4)$$

Subtracting equation (4) from equation (3), we obtain:

$$\sum x \sum Y - n\sum xY = b\left[\left(\sum x\right)^2 - n\sum x^2\right] \quad [1]$$

Rearranging this gives:

$$\hat{b} = \frac{\sum xY - \frac{\sum x \sum Y}{n}}{\sum x^2 - \frac{(\sum x)^2}{n}} \text{ or } \hat{b} = \frac{S_{xy}}{S_{xx}} \quad [1]$$

and this is the least squares estimator for b. The estimator for a is given by $\hat{a} = \overline{Y} - \hat{b}\overline{x}$. This can be obtained by rearranging equation (1). [1]

Markers: please award the marks for the alternative method using substitution.

(i)(b) ***Maximum likelihood estimators***

The answer would not change at all. For a normal distribution, maximum likelihood and least squares give the same estimators. [1]

Since $Y_i = a + bx_i + e_i$, where $e_i \sim N(0, \sigma^2)$, the distribution of Y_i is $N(a + bx_i, \sigma^2)$. So the likelihood function using maximum likelihood estimation is:

$$L(a,b) = \prod \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{-\frac{1}{2}\left(\frac{Y_i - a - bx_i}{\sigma}\right)^2\right\} \Rightarrow \quad \ln(L) = \text{constant} - \frac{1}{2\sigma^2}\sum (Y_i - a - bx_i)^2$$

Maximising L is equivalent to minimising $\sum (Y_i - a - bx_i)^2$, which is identical to the criterion used above to find the least squares estimators. [1]

[Total 6]

(ii) ***Regression coefficients***

$$\hat{b} = \frac{s_{xy}}{s_{xx}} = \frac{826}{42} = 19.667 \qquad [1]$$

$$\hat{a} = \bar{y} - \hat{b}\bar{x} = 105 - 19.667 \times 4.5 = 16.5 \qquad [1]$$

[Total 2]

(iii) ***Testing slope parameter***

We wish to test: $H_0 : b = 22 \quad vs \quad H_1 : b \neq 22$

Under H_0: $\dfrac{\hat{b} - b}{\sqrt{\hat{\sigma}^2 / S_{xx}}} \sim t_{n-2}$

where $\hat{\sigma}^2$ is an estimator of σ^2 in the above formula. Using the data, the estimated value of σ^2 is given by:

$$\hat{\sigma}^2 = \frac{1}{n-2}\left(S_{yy} - \frac{S_{xy}^2}{S_{xx}}\right) = \frac{1}{6}\left(16,492 - \frac{826^2}{42}\right) = 41.2222 \qquad [1]$$

So the observed value of the test statistic here is:

$$\frac{19.667 - 22}{0.9907} = -2.355 \qquad [1]$$

This is a two-sided test. Using percentage points of the t_6 distribution, the probability value for this test statistic is about $0.03 \times 2 = 6\%$. [½]

So we have insufficient evidence to reject H_0 at the 5% level. Therefore it is reasonable to assume that the slope of the model is £22 per hour. [1]

[Total 4]

(iv)(a) ***Confidence interval for the average cost of a job lasting 4 hours***

The estimate of the mean predicted cost is:

$$\hat{\mu} = \hat{a} + \hat{b} \times 4 = 16.5 + 19.667 \times 4 = 95.17$$ [1]

The standard error is given by:

$$\sqrt{\left(\frac{1}{8} + \frac{(4-4.5)^2}{42}\right) 41.22} = \sqrt{5.3981} = 2.323$$ [1]

This gives a confidence interval, using the t_6 tables, of:

$$95.17 \pm 1.943 \times 2.323 = 95.17 \pm 4.51 = (90.7, 99.7)$$ [1]

(iv)(b) ***Confidence interval for an individual job lasting 6 hours***

The estimate of the individual predicted cost is:

$$\hat{\mu} = \hat{a} + \hat{b} \times 6 = 16.5 + 19.667 \times 6 = 134.5$$ [1]

The standard error is given by:

$$\sqrt{\left(1 + \frac{1}{8} + \frac{(6-4.5)^2}{42}\right) 41.22} = \sqrt{48.583} = 6.970$$ [1]

This gives a confidence interval of:

$$134.5 \pm 1.943 \times 6.970 = 134.5 \pm 13.54 = (121, 148)$$ [1]

[Total 6]

(v) ***Comment***

The confidence interval for the individual job is wider (£27) than the confidence interval for the average cost (£9). So there is greater uncertainty over an individual result than an average one.

[1]