The University of Melbourne

Semester 1 Assessment — June, 2017

School of Mathematics and Statistics

MAST90044 Thinking and Reasoning with Data

Exam duration: two hours
Reading time: fifteen minutes
This paper has 12 pages.

Authorised materials:
Hand-held electronic calculators may be used.
A single A4 sheet of hand–written notes (both sides) may be used.

Instructions to students:
All answers are to be written in your answer booklet.

This examination has three sections:
Section A: contains nine questions each worth 2 marks.
The total number of marks for section A is 18 marks.
Section B contains two questions, worth 9 marks and 13 marks.
The total number of marks for section B is 22 marks.
Section C contains five questions.
The total number of marks for section C is 60 marks.
All questions may be attempted.
The number of marks for each question is indicated after the question.
The total number of marks available is 100.

There are statistical tables on page 12.

Section A (18 marks)

Section A consists of 9 multiple choice questions, each worth 2 marks.
For each question, one of the alternative answers should be chosen.
No working need be shown.
Write your answers (as a letter A, B, C, D or E only) in your answer book, preferably all on one page.

1. (a) In an opinion poll, 25% of 200 people sampled said that they were strongly opposed to having a state lottery. The standard error of the sample proportion is approximately

   A. 0.0009 **B.** 0.015 **C.** 0.03 **D.** 0.04 **E.** 0.06

   (b) A television station is interested in predicting whether voters are in favour of an increase in GST. It asks its viewers to phone in and indicate whether they support or oppose an increase in the GST in order to generate additional revenue for education. Of the 2633 viewers who phoned in, 1474 (55.98%) were opposed to the increase. The population of interest is

   A. All people who will vote on the GST increase issue on the day of the vote.

   B. All regular viewers of the television station who own a phone and have participated in similar phone surveys in the past.

   C. The 2633 viewers who phoned in.

   D. The 1474 viewers who were opposed to the increase.

   E. None of the above.

   (c) Offspring of a particular genetic cross have an undesirable trait with probability 1/8. Inheritance of this trait by separate offspring is independent. You examine 100 offspring from this cross and count the number X who have the undesirable trait. The mean and standard deviation of X are:

   A. 3.54 10.94 **B.** 12.5 3.31 **C.** 12.5 10.94 **D.** 87.5 3.31 **E.** 87.5 10.94

   (d) In a study to determine optimal cooking temperature, measurements of yield (y) were made at five different cooking temperatures (x). The polynomial regression equation was

$$y = 7.96 - 0.153\,x + 0.001\,x^2.$$

   One measure gave a yield of 3.3. Suppose the cooking temperature was set to 50, what would be the residual corresponding to this value?

   A. -50 **B.** -0.49 **C.** 0 **D.** 0.49 **E.** 3.3

(e) Of the following statements about randomisation in experimental design, which one is *false*?

A. Randomisation is the use of chance to allocate treatments to experimental units.

B. Randomisation is necessary to ensure the validity of an experiment.

C. Randomisation helps to avoid confounding.

D. Randomisation increases the precision of experiments.

E. Randomisation is needed to allocate treatments in Latin square designs.

(f) The following ANOVA was obtained for an experiment that used a randomised block design:

	Df	Sum Sq	Mean Sq	F value	Pr(¿F)
Block	3	2.4486	0.8162	6.22	0.014
Treatment	3	1.8712	0.6237	4.75	0.030
Residuals	9	1.1807	0.1312		

If the same data had been analysed assuming a completely randomized design, then the value of the F statistic would have been

A. 2.06 **B.** 2.45 **C.** 3.57 **D.** 4.75 **E.** 6.22

(g) Of the following statements about P-values, which one is *false*?

A. The P-value is the probability of observing a value of the test statistic at least as extreme as the one observed, assuming that the null hypothesis is true.

B. The smaller the P-value, the more statistically significant the result.

C. A large P-value does not prove that the null hypothesis is true.

D. In general, a small P-value is evidence against the null hypothesis.

E. The P-value is the probability of a Type I error.

(h) A sample of 20 pairs of values (x_1, x_2) is taken from the random variables X_1 and X_2. The following R output arises from fitting a simple linear regression of x_2 on x_1,

```
Call:
lm(formula = x2 ~ x1)

Coefficients:
             Estimate  Std.Error t value Pr(>|t|)
(Intercept) 18.76842    0.83996  22.345 1.41e-14 ***
x1          -0.68271    0.07012  -9.737 1.34e-08 ***

Residual standard error: 1.808 on 18 degrees of freedom
Multiple R-squared: 0.8404,     Adjusted R-squared: 0.8316
F-statistic:  94.8 on 1 and 18 DF,  p-value: 1.345e-08
```

The correlation coefficient of x_1 with x_2 is

A. -0.92 **B.** -0.84 **C.** 0.83 **D.** 0.84 **E.** 0.92

(i) In the above output, a value of `0.07012` is listed under `Std.Error`. This standard error can be interpreted to mean

A. The standard deviation of X_1.

B. The estimated standard deviation of X_1.

C. The estimated standard deviation of the coefficient of x_1 in the regression.

D. The estimated standard deviation of the error distribution.

E. The standard deviation of the correlation coefficient.

[$2 + 2 + 2 + 2 + 2 + 2 + 2 + 2 + 2 = 18$ marks]

Section B (22 marks)

2. Choose *three* of the following five concepts, and explain the meaning of each. For each concept you choose, write a few sentences. Use a diagram or plot if it helps the explanation.

   (a) The binomial distribution;

   (b) Parameters and estimates;

   (c) Q-Q plots;

   (d) Confidence intervals;

   (e) Sampling distributions.

[3 + 3 + 3 = 9 marks]

3. To compare the lifetimes of four brands of size AA battery, an experiment is to be conducted using portable CD players. A decision has been made to take exactly 20 observations using one of the following three options:

   (1) using 20 CD players once each (all players of the same make and model);

   (2) using 5 CD players four times each (all players of the same make and model);

   (3) using just one CD player, 20 times.

   (a) Answer the following questions for **each** of the three options above **in turn**:

      i. What are the experimental units?

      ii. Explain, briefly, how randomisation could/should be used in the design of the experiment.

      iii. The most appropriate method for analysing the data from the experiment is:

         A. a simple linear regression
         B. a multiple linear regression
         C. a one-way Analysis of Variance
         D. a two-way Analysis of Variance
         E. an Analysis of Covariance
         F. a two-sample t-test
         G. a paired-sample t-test
         H. none of the above.

   (b) Which of the three options would you recommend? Justify your answer.

[3 + 3 + 3 + 4 = 13 marks]

Section C (60 marks)

4. A study compared a group of Alzheimer's patients with a control group of people who did not have Alzheimer's but were similar in other ways. The focus of this study was on the use of antacids that contain aluminium.

	Aluminum-containing antacid use			
	None	Low	Medium	High
Alzheimer's patients	213	17	18	7
Control group	211	27	11	2

R produced the following output in relation to these data:

Test 1

```
> aa = matrix(c(213,17,18,7,211,27,11,2),nrow=2,byrow=T)
> (aa.test = chisq.test(aa))

Pearson's Chi-squared test
data:  aa
X-squared = 6.718, df = .., p-value = 0.081
Warning message:
In chisq.test(aa) : Chi-squared approximation may be incorrect
```

Test 2

```
> aa1 = matrix(c(213,42,211,40),nrow=2,byrow=T)
> (aa1.test = chisq.test(aa1))

Pearson's Chi-squared test with Yates' continuity correction
data:  aa1
X-squared = 0.002, df = .., p-value = 0.966
```

Test 3

```
> aa2 = matrix(c(17,18,7,27,11,2),nrow=2,byrow=T)
> (aa2.test = chisq.test(aa2))

Pearson's Chi-squared test
data:  aa2
X-squared = 6.695, df = .., p-value = 0.035
Warning message:
In chisq.test(aa2) : Chi-squared approximation may be incorrect
```

(a) Three tests have been applied.
   i. What statistic does `X-squared` denote?
   ii. The values for `df` have been omitted: what are they?
   iii. What do the warnings indicate?

(b) What is (in words) the sensible null hypothesis for this experiment?

(c) For Test 2, verify that the usual test statistic for the association between Alzheimers and Aluminium is 0.002.

(d) What conclusions do you draw from this output?

[$3 + 1 + 2 + 3 = 9$ marks]

5. Scientists are interested in the effect of a remedial treatment for the amount of blue-green algae in a river. They took samples from a randomly selected side of the river at each of 195 locations and weighed the biomass, then applied the treatment to each location and took samples from the other side of the river after the treatment. The mean weight change for the samples was a loss of 0.37 kg, with standard deviation 1.52 kg.

   (a) State the appropriate hypotheses.

   (b) Carefully define the population parameter(s) that you are testing.

   (c) State the necessary assumptions for this test.

   (d) What is the value of the standard error?

   (e) Provide a value for the test statistic.

   (f) i) Use the output below to test the hypothesis from (a).
   ii) Note that the `df` argument is blank: what should it have been?

```
> qt(0.975, df = )
   [1] 1.972268
```

   (g) State the conclusion based on the results of the test.

$[1 + 1 + 2 + 1 + 1 + 2 + 2 = 10 \text{ marks}]$

6. In order to investigate the relationship between commuting distance and stress a researcher collected data for twelve workers commuting to a job in the city on a daily basis. The twelve subjects had their daily commuting distance measured in kilometres and were given a stress test after arriving at their job, measured on a continuous scale from 1 to 10. Higher scores indicate more stress.

Worker	1	2	3	4	5	6	7	8	9	10	11	12	Mean
Distance	12.8	17.8	9.8	13.6	16.8	4.5	10.4	14.1	18.4	16.3	14.6	15.5	13.72
Stress level	7.5	8.1	4.8	6.3	9.5	4.5	6.3	6.7	8.1	6.9	7.2	7.9	6.98

The following is part of the R output that was obtained for these data:

```
The regression equation is
stress level = 2.78 + 0.306 distance

Predictor         Coef     SE Coef          T
Constant        2.7842      0.8363       3.33
distance       0.30595     0.05875       5.21

S = 0.7687

Analysis of Variance

Source            DF          SS          MS         F
Regression         1      16.027      16.027     27.12
Residual Error    10       5.910       0.591
Total             11      21.937
```

(a) Give the model and state the underlying assumptions. How would you check that the assumptions were satisfied?

(b) Assuming that the model is appropriate, carry out a test of whether there is a linear relationship between stress level and distance travelled, stating the appropriate hypotheses and conclusion.

(c) The estimate of the constant term in the model is 2.7842. Comment on the meaning and usefulness of this estimate.

(d) What proportion of the variation in levels of stress is explained by distance travelled?

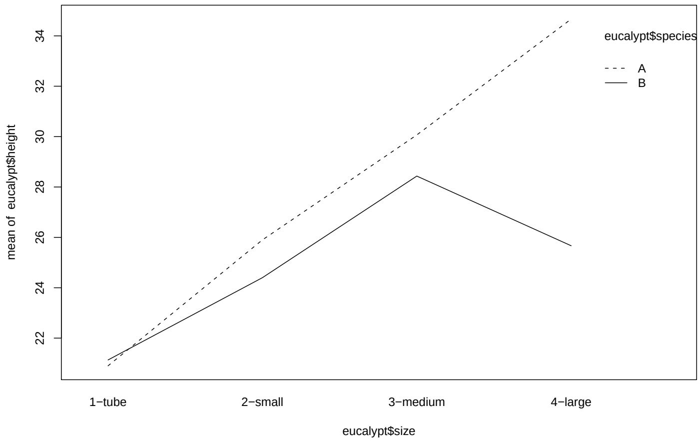
(e) Calculate adjusted R^2 and explain the difference between R^2 and adjusted R^2.

$[8 + 4 + 3 + 1 + 3 = 19 \text{ marks}]$

7. The following data came from an experiment set up to study the effect of pot size and species on the growth of eucalypt seedlings. The height (in cm) of each seedling at 10 weeks is given in the table. Four pot sizes (tube, small, medium, and large) and two species (A and B) were used.

	Species A				Species B			
Pot size	tube	small	medium	large	tube	small	medium	large
Height of seedling	22.1	28.1	28.0	32.5	20.4	25.3	26.6	23.4
	22.5	26.5	29.1	38.0	24.2	22.0	30.1	26.2
	18.1	23.1	33.1	33.5	18.8	25.9	28.6	27.4
Mean	20.9	25.9	30.1	34.7	21.1	24.4	28.4	25.7

The R output below shows a graph of the means, followed by an ANOVA table.



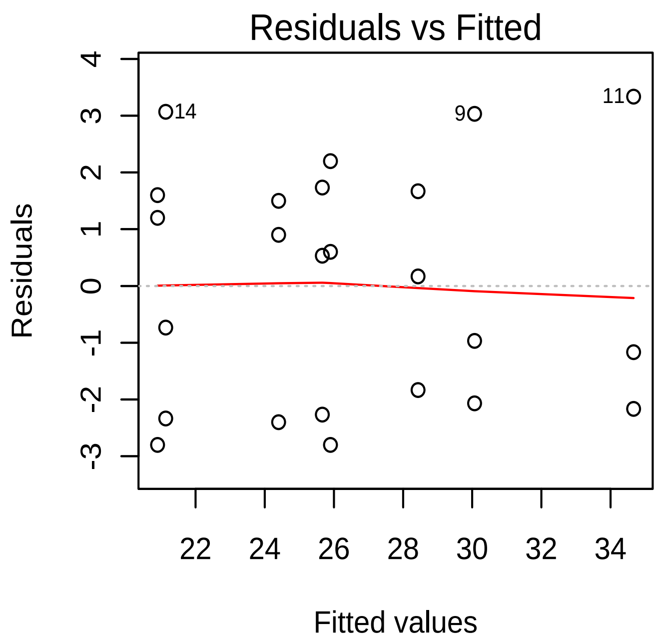

```
> eucalypt.lm <- lm(height ~ species * size, data = eucalypt)
> anova(eucalypt.lm)

Analysis of Variance Table

Response: height
             Df Sum Sq Mean Sq F value    Pr(>F)
species       1  53.10  53.104  8.9200  0.008719 **
size          3 317.12 105.706 17.7558 2.412e-05 ***
species:size  3  75.85  25.285  4.2472  0.021854 *
Residuals    16  95.25   5.953
---
```

(a) Comment on what the graph and the ANOVA table show regarding the factors of interest.

(b) The experimenter would like to make a recommendation to eucalypt growers regarding which pot size (there may be more than one) is best in regard to growing taller seedlings. There may be separate recommendations for species A and B, or an overall recommendation for both. Find a 95% confidence interval for a difference between means which enables the experimenter to make a sound recommendation. State what the recommendation would be. Provide a diagram that summarises your recommendations. [*Hint: the LSD may be of use.*]

(c) Estimate the standard deviation of the distribution of the errors in the statistical model fitted.

(d) Below is a diagnostic plot arising from the fitted model. Which assumption regarding the model can you assess using this plot? Is the assumption satisfied? Briefly explain.



[$4 + 5 + 2 + 3 = 14$ marks]

8. In a series of experiments, n_i beetles were exposed to a fixed quantity of pesticide with concentration x_i, as a result of which y_i beetles were killed.

This is to be modelled by

$$Y_i \stackrel{d}{=} \text{Bi}(n_i, p_i), \text{ where } \ln(\tfrac{p_i}{1-p_i}) = \alpha + \beta x_i.$$

The following data are available:

number exposed, n_i	59	60	62	56	63	59	62	60
number killed, y_i	6	13	18	28	52	53	61	60
concentration, x_i	91	124	155	184	211	237	261	284

The following R output is obtained:

```
> n=c(59,60,62,56,63,59,62,60)
> y=c(6,13,18,28,52,53,61,60)
> x=c(91,124,155,184,211,237,261,284)

> beetle = glm(y/n ~ x,family = binomial, weights = n)
> summary(beetle)

Call:
glm(formula = y/n ~ x, family = binomial, weights = n)
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.5878  -0.4085   0.8442   1.2455   1.5860
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -5.882647   0.535300  -10.99   <2e-16 ***
x            0.034286   0.002913   11.77   <2e-16 ***
---
Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
(Dispersion parameter for binomial family taken to be 1)
    Null deviance: 284.202  on 7  degrees of freedom
Residual deviance:  11.116  on 6  degrees of freedom
AIC: 41.314
Number of Fisher Scoring iterations: 4}
```

(a) Which statistic in the above output indicates that the concentration significantly affects the mortality?

(b) Which statistic in the above output indicates the model fits the data?

(c) Give a point estimate and an interval estimate for the odds ratio. Give a precise interpretation of your interval estimate within the context of this study.

(d) What is the estimated proportion killed when the concentration is 250?

(e) What is the estimated value of LD50, the dose resulting in 50% mortality?

$[1 + 1 + 2 + 2 + 2 = 8 \text{ marks}]$

Total marks = 100

Some numerical values

Standard normal cumulative probabilities,

$\Pr(Z \le a)$, where $Z \stackrel{d}{=} \mathrm{N}(0,1)$, for $a = 0.1, 0.2, \ldots, 4.0$.

```
> pnorm((1:40)/10)
 [1] 0.5398 0.5793 0.6179 0.6554 0.6915 0.7257 0.7580 0.7881 0.8159 0.8413
[11] 0.8643 0.8849 0.9032 0.9192 0.9332 0.9452 0.9554 0.9641 0.9713 0.9772
[21] 0.9821 0.9861 0.9893 0.9918 0.9938 0.9953 0.9965 0.9974 0.9981 0.9987
[31] 0.9990 0.9993 0.9995 0.9997 0.9998 0.9998 0.9999 0.9999 1.0000 1.0000
```

Standard normal quantiles, $c_q(Z)$, where $Z \stackrel{d}{=} \mathrm{N}(0,1)$.

```
> qnorm(c(0.6,  0.75,   0.8,   0.9,  0.95, 0.975,  0.99, 0.995, 0.999))
[1]    0.2533 0.6745 0.8416 1.2816 1.6449 1.9600 2.3263 2.5758 3.0902
```

0.975 t-quantiles, $c_{0.975}(\mathrm{t}_k)$, for $k = 1, 2, \ldots, 50$.

```
> qt(0.975,1:50)
 [1] 12.706  4.303  3.182  2.776  2.571  2.447  2.365  2.306  2.262  2.228
[11]  2.201  2.179  2.160  2.145  2.131  2.120  2.110  2.101  2.093  2.086
[21]  2.080  2.074  2.069  2.064  2.060  2.056  2.052  2.048  2.045  2.042
[31]  2.040  2.037  2.035  2.032  2.030  2.028  2.026  2.024  2.023  2.021
[41]  2.020  2.018  2.017  2.015  2.014  2.013  2.012  2.011  2.010  2.009
```

Cumulative probabilities $\Pr(T \le 0.025)$, where $T \stackrel{d}{=} \mathrm{t}_k$, for $k = 1, 2, \ldots, 50$.

```
> pt(0.025,df=1:50)
 [1] 0.508 0.509 0.509 0.509 0.509 0.510 0.510 0.510 0.510 0.510 0.510 0.510
[13] 0.510 0.510 0.510 0.510 0.510 0.510 0.510 0.510 0.510 0.510 0.510 0.510
[25] 0.510 0.510 0.510 0.510 0.510 0.510 0.510 0.510 0.510 0.510 0.510 0.510
[37] 0.510 0.510 0.510 0.510 0.510 0.510 0.510 0.510 0.510 0.510 0.510 0.510
[49] 0.510 0.510
```

END OF EXAMINATION