THE UNIVERSITY OF MELBOURNE

Student Number

Semester 1 Assessment, 2018

School of Mathematics and Statistics

MAST90044 Thinking and Reasoning with Data

Writing time: 2 hours

Reading time: 15 minutes

This is NOT an open book exam

This paper consists of 8 pages (including this page)

Authorised Materials

- Mobile phones, smart watches and internet or communication devices are forbidden.

Instructions to Students

- You must NOT remove this question paper at the conclusion of the examination.
- Hand-held electronic calculators may be used.
- A single A4 sheet of hand-written notes (both sides) may be used.
- You should attempt all questions.
- There are 7 questions with marks as shown. The total number of marks available is 88.

Instructions to Invigilators

- Students must NOT remove this question paper at the conclusion of the examination.
- Confiscate all magic wands.

Blank page (ignored in page numbering)

Question 1 (30 marks)

Multiple choice: give answers (as a letter A, B, C, D or E only) to each of the 6 questions in your script book. There is no need to show any working.

(a) In a statistics course, a linear regression equation was computed to predict the final exam score (y) from the score on an assignment (x). The equation of the least-squares regression line was

$$y = -10 + 1.1\,x$$

Suppose Jim scored 80 on both the assignment and the final exam. What would be the residual corresponding to this value?

A. -8

B. -2

C. 0

D. 2

E. 80

(b) A 95% confidence interval for the mean reading achievement score for a population of third grade students is $(44.2, 54.2)$. Suppose you compute a 99% confidence interval. Which of the following statements is correct?

A. the intervals have the same width

B. the 99% interval is wider

C. the 95% interval is wider

D. you cannot determine which interval is wider unless you know n and s.

(c) Which **one** of the following is likely to have a binomial distribution:

A. the number of accidents in a large factory during one 8-hour shift

B. the number of spades in a bridge hand (i.e. in a random selection of 13 cards from a pack of 52 cards)

C. the number of tosses of a fair coin until the 10th head is obtained

D. the number of years between floods at a certain location

E. the number of beetles that are killed when a random sample of 40 beetles is subjected to a specified dose of an insecticide.

(d) A random variable X has mean μ_X and standard deviation σ_X . Suppose n independent observations are taken on X and the mean of these n observations is calculated. We can assert that, if n is very large, the sampling distribution of the sample mean is approximately normal. This assertion follows from

A. the law of large numbers

B. the central limit theorem

C. the definition of sampling distribution

D. the bell curve

E. the law of averages

(e) n 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

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

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

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

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

D. 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.

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

Question 2 (9 marks) 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) P-value;

(b) Correlation coefficient;

(c) Q-Q plots;

(d) Testing the assumptions of linear models;

(e) Using AIC in model selection.

Question 3 (13 marks) Do grumpy old men have a greater risk of having coronary heart disease than men who aren't so grumpy? Harvard medical school researchers examined this question in an observational study reported in the November 1994 issue of *Circulation*. For seven years, the researchers studied men between the ages of 46 and 90 years old. All study participants completed a survey of anger symptoms at the beginning of the study period. Among 199 men with no anger symptoms, there were 8 cases of coronary heart disease. Among 559 men who had the most anger symptoms, there were 59 cases of coronary heart disease.

(a) Construct a contingency table for the relationship between the degree of anger and the incidence of heart disease.

(b) Among those with no anger symptoms, what percentage had coronary heart disease?

(c) Among those with most anger symptoms, what percentage had coronary heart disease?

(c) What is (in words) the sensible null hypothesis for this

(d) Verify that the test statistic for the association between anger symptoms and coronary heart disease is 7.777.

(e) Use the output below to test the hypothesis from (d), and make a conclusion. Note that the `df` argument is blank: what should it have been?

```
> 1 - pchisq(7.777, df = )

[1] 0.00529156
```

(f) Based on these results, could we state that anger *causes* coronary heart disease? Why, or why not?

Question 4 (12 marks) A 1986 study classified a sample of psychiatric patients by their diagnosis and by whether their treatment involved prescription drugs. The frequencies were as follows:

Diagnosis	Drugs	No drugs	Total
Schizophrenia	105	8	113
Affective disorder	12	2	14
Neurosis	18	19	37
Personality disorder	47	52	99
Total	182	81	263

(a) An analysis performed in R gave the following output:

```
> diagnosis.by.drugs <- matrix(c(105, 8, 12, 2, 18, 19, 47, 52),
+     nrow = 4, byrow = TRUE)
> chisq.test(diagnosis.by.drugs)

    Pearson's Chi-squared test

data:  diagnosis.by.drugs
X-squared = 60.8793, df = ..., p-value = 3.814e-13
```

(b) What null hypothesis is tested by this analysis?

(c) Make a conclusion based on the result of the hypothesis test.

(d) The `df` argument is blank in the output: what should it have been?

(e) Have the usual requirements concerning the expected frequencies been met? Explain your answer.

Question 5 (6 marks) A 1997 paper in the *Journal of Medicine* reported on a study on the influence of alpha interferon in the treatment of patients with multiple myeloma (MM). Twenty patients with MM were randomly assigned into two groups; ten patients were treated with interferon (treatment group) and the remaining ten were not (control group). The values of the response variable ("serum beta-2-microglobulin" or SB2M) were as follows:

Treatment group	2.9	2.7	3.9	2.7	2.1	2.6	2.2	4.2	5.0	0.7
Control group	3.5	2.5	3.8	8.1	3.6	2.2	5.0	2.9	2.3	2.9

(a) Formulate a statistical model for analysing these data, defining all terms.

(b) The following R code gives two possible ways of analysing the data:

```
> treatment <- c(2.9, 2.7, 3.9, 2.7, 2.1, 2.6, 2.2, 4.2, 5, 0.7)
> control <- c(3.5, 2.5, 3.8, 8.1, 3.6, 2.2, 5, 2.9, 2.3, 2.9)
> diff <- treatment - control
> t.test(treatment, control)


    Welch Two Sample t-test

data:  treatment and control
t = -1.1518, df = 15.95, p-value = 0.2664
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -2.2159977  0.6559977
sample estimates:
mean of x mean of y
     2.90      3.68


> t.test(diff)


    One Sample t-test

data:  diff
t = -1.073, df = 9, p-value = 0.3112
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 -2.4243875  0.8643875
sample estimates:
mean of x
    -0.78
```

Which is the more appropriate analysis? Briefly explain.

Question 6 (9 marks) The taste of cheddar cheese is related to the concentration of several chemicals in the final product. In a study of cheddar cheese from the La Trobe Valley of Victoria, 30 samples of cheese were analysed for their chemical composition, and subjected to taste tests. The results for the first few samples are shown in the table below. The response variable (`taste`) was obtained by combining the scores from several tasters. The three chemicals whose concentrations were measured were acetic acid, hydrogen sulphide (H_2S) and lactic acid.

taste	acetic	H_2S	lactic
12.3	4.543	3.135	0.86
20.9	5.159	5.043	1.53
39.0	5.366	5.438	1.57
47.9	5.759	7.496	1.81
⋮	⋮	⋮	⋮

The following R output shows some of the results of a multiple regression of `taste` against the other three variables:

```
> cheese.lm3 <- lm(taste ~ acetic + H2S + lactic, data = cheese)
> summary(cheese.lm3)


Call:
lm(formula = taste ~ acetic + H2S + lactic, data = cheese)

Residuals:
    Min      1Q  Median      3Q     Max
-17.391  -6.612  -1.009   4.908  25.449

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -28.8768    19.7354  -1.463  0.15540
acetic        0.3277     4.4598   0.073  0.94198
H2S           3.9118     1.2484   3.133  0.00425 **
lactic       19.6705     8.6291   2.280  0.03108 *
---

Residual standard error: 10.13 on 26 degrees of freedom

F-statistic: 16.22 on 3 and 26 DF,  p-value: 3.81e-06
```

(a) What can you conclude?

(b) Give an interpretation of the coefficient of `lactic` given in the output.

(c) The following R output shows a stepwise method applied to help select explanatory variables.

```
> cheese.lm0 <- lm(taste ~ 1, data = cheese)
> step(cheese.lm0, ~. + acetic + H2S + lactic)


Start:  AIC=168.29
taste ~ 1

          Df Sum of Sq    RSS    AIC
+ H2S      1    4376.7 3286.1 144.89
+ lactic   1    3800.4 3862.5 149.74
+ acetic   1    2314.1 5348.7 159.50
<none>                 7662.9 168.29

Step:  AIC=144.89
taste ~ H2S

          Df Sum of Sq    RSS    AIC
+ lactic   1     617.2 2669.0 140.65
<none>                 3286.1 144.89
+ acetic   1      84.4 3201.7 146.11
- H2S      1    4376.7 7662.9 168.29
```

```
Step:  AIC=140.65
taste ~ H2S + lactic

         Df Sum of Sq    RSS    AIC
<none>                2669.0 140.65
+ acetic  1      0.55 2668.4 142.64
- lactic  1    617.18 3286.1 144.89
- H2S     1   1193.52 3862.5 149.74

Call:
lm(formula = taste ~ H2S + lactic, data = cheese)

Coefficients:
(Intercept)          H2S       lactic
    -27.592        3.946       19.887
```

State which explanatory variable(s) should be included in a model for predicting taste of cheddar cheese. Give reasons for your choice.

(d) Suppose that acetic acid concentration was much cheaper and easier to measure than lactic acid concentration. Would it be worthwhile (on statistical grounds) including acetic acid concentration in a model for predicting taste? Give a reason.

Question 7 (9 marks) Ophthalmologists from Victoria and Western Australia have surveyed children in the Western Desert in Western Australia to assess the prevalence and severity of trachoma. The data below come from two years of a longitudinal survey. There are six stages of trachoma, of increasing severity. In this study, children were observed to have trachoma up to the fourth stage. The data below show the stages of trachoma including an additional level — those with no signs of trachoma.

Stage	1993	2003
None	124	264
Follicular	88	46
Intense inflammatory	7	3
Trachomatous scarring	0	2
Trichiasis	2	0

Another ophthalmologist wants to investigate whether the proportion of children with any sort of trachoma has significantly decreased between 1993 and 2003, and decides to use logistic regression. The output of the analysis is as follows:

```
> trachoma <- data.frame(year = factor(c(1993, 2003)), disease = c(97,
+     51), total = c(221, 315))
> trachoma.1 <- glm(disease/total ~ year, family = binomial, weight = total,
+     data = trachoma)
> summary(trachoma.1)


Call:
glm(formula = disease/total ~ year, family = binomial, data = trachoma,
```

```
    weights = total)

Deviance Residuals:
[1]  0  0

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -0.2456     0.1355  -1.812     0.07 .
year2003     -1.3986     0.2044  -6.843 7.75e-12 ***

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 49.639  on 1  degrees of freedom
Residual deviance:  0.000  on 0  degrees of freedom
AIC: 15.433
```

- Estimate the odds ratio, together with an appropriate measure of uncertainty.
- Test the hypothesis (at the 0.05 significance level) that the odds have halved between 1993 and 2003.
- Make a concluding statement of a few sentences, along the lines of what the ophthalmologist might include in a report.

End of Exam—Total Available Marks = 88