THE UNIVERSITY OF MELBOURNE

Student Number	

Semester 2 Final Exam, 2017

School of Mathematics and Statistics

MAST90058 Elements of Statistics

Writing time: 3 hours

Reading time: 15 minutes

This is NOT an open book exam

Common content with: MAST20005 Statistics

This paper consists of 6 pages (including this page)

Authorised Materials

- Mobile phones, smart watches and internet or communication devices are forbidden.
- Hand-held scientific calculators (not CAS or graphics) may be used.
- Students may bring one double-sided A4 sheet of handwritten notes.

Instructions to Students

- You must NOT remove this question paper at the conclusion of the examination.
- Show full working for each of your answers.
- Some useful R output is given in an appendix on the last page.
- You should attempt all questions. Marks for individual questions are shown.
- This examination contains 8 questions.
- The total number of marks available is 80.

Instructions to Invigilators

- Students must NOT remove this question paper at the conclusion of the examination.
- All graphics or CAS calculators should be confiscated.
- Students may use one double-sided A4 sheet of handwritten notes.

Blank page (ignored in page numbering)

Question 1 (18 marks) A discrete random variable X has the following pmf:

x	0	1	2
$p(x)$	$1-2\theta$	θ	θ

(a) Find $\mathrm{E}(X)$ and $\mathrm{var}(X)$.

(b) Consider a random sample of size n on X.

(i) Write down the log-likelihood function.

(ii) Find the maximum likelihood estimator (MLE) of θ.

(iii) Determine a sufficient statistic for θ.

(iv) Is the MLE biased?

(v) Find the Cramér–Rao lower bound for unbiased estimators of θ.

(vi) Does the MLE achieve this bound?

(c) A random sample of size $n = 20$ produced the following observations:

0 1 0 0 0 2 1 2 0 2 0 2 2 1 0 2 1 2 0 2

(i) Give the MLE for this sample and its standard error.

(ii) Calculate an approximate 95% confidence interval for θ.

(iii) Carry out a goodness-of-fit test for this model, with a significance level of 5%.

(iv) Suppose all of the values of 1 and 2 were relabelled as 3 (meaning that the only possible observations are 0 and 3). Is it still possible to carry out a goodness-of-fit test? Explain your answer. What distribution describes the relabelled data?

Question 2 (10 marks) Consider the same X as in Question 1.

(a) Consider a random sample of size n on X.

(i) Find the method of moments (MM) estimator of θ.

(ii) Is this estimator biased?

(iii) Does this estimator achieve the Cramér–Rao lower bound?

(iv) Which of the MLE or MM estimators is better?

(b) Consider the same random sample of size $n = 20$ as in Question 1.

(i) Give the MM estimate for this sample and its standard error.

(ii) Calculate an approximate 95% confidence interval for θ based on the MM estimator.

Question 3 (10 marks) Researchers are studying the reading speed of people under two levels of light. The 10 people recruited into the study are each tested under both light conditions, giving the following measurements (in words per minute):

Person	Strong light	Weak light
1	85	80
2	84	83
3	80	77
4	89	91
5	86	76
6	82	81
7	89	89
8	88	81
9	81	79
10	97	90

For simplicity, we can assume that these measurements are normally distributed, with a mean of μ_1 under strong light and a mean of μ_2 under weak light.

(a) Calculate a 95% confidence interval for the difference in mean reading speeds, $\mu_1 - \mu_2$.

(b) The researchers want to test if there is a difference between the two light conditions.

(i) What are the null and alternative hypotheses?

(ii) What is an appropriate test statistic?

(iii) Specify the critical region for a test with significance level 5%.

(iv) Carry out the test for the dataset provided above.

Question 4 (14 marks)

(a) A newspaper commissions an opinion poll for an upcoming election. Out of 500 people surveyed, 260 said they intend to vote for the Purple Party. Calculate a 95% confidence interval for the proportion of people in the population that intend to vote in this way.

(b) The next week, the newspaper commissions another poll. This time, out of 520 people surveyed, 255 said they will vote for the Purple Party. By comparing it to the previous week, the newspaper reports this as, "Public mood turns against Purple!". Is this an accurate summary of the data? How would you summarise the quantitative evidence here? Your answer should include the calculation of an appropriate 95% confidence interval.

(c) The editor of the newspaper wants to present a strong conclusion based on their next poll. She decides that they need an estimate that will be within (i.e. margin of error) at most 1% with 95% probability. How big should their sample size be?

(d) The finance officer at the newspaper overrules the editor and says they can only afford a poll of 2,000 people. Based on this poll, if there is sufficient evidence that the Purple Party has majority support (greater than 50% of the population), the editor will decide to run a story predicting the outcome of the election in their favour. She will do this based on a hypothesis test with significance level of 5%. What are the null and alternative hypotheses? Determine the power function of the test.

(e) If in fact the true public support for the Purple Party is 53%, what is the probability of a type II error?

Question 5 (8 marks) Damjan, who manages a supermarket, is worried about the quality of milk he is receiving from his suppliers. Managers of other stores, who share the same supplier, have said that about 1% of their stock was sour upon delivery and therefore unsuitable for sale. Damjan decides to test his own stock. He randomly samples 40 bottles and finds that 1 of them is sour. To help him plan his budget for potential customer refunds, as well as collect evidence to complain to his supplier, he would like to estimate the proportion, p, of his stock that is sour.

(a) Write down the maximum likelihood estimate of p.

(b) Damjan would like to incorporate the knowledge he has gained from the other managers. He decides to encode this in the form of a conjugate prior distribution. What type of distribution should he use?

(c) He decides that this information is worth the equivalent of 5 random samples (i.e. as pseudodata) and it should have mean $\mathbb{E}(p) = 0.01$. Determine the prior distribution that satisfies these constraints.

(d) Using this prior, what is the posterior distribution of p?

(e) Calculate the posterior mean and a central 95% credible interval.

Question 6 (6 marks) In a trial of a new drug for treating depression, we observe the following outcomes:

	Symptoms		
	Worse	Same	Better
Placebo	15	10	18
Drug	7	5	45

Is there evidence that the drug has had an effect? Answer by doing an appropriate hypothesis test with a 1% significance level.

Question 7 (8 marks) Robert is conducting a tram reliability study. At a particular tram stop, he observes the following times between successive tram arrivals:

1.7 2.2 0.9 5.7 8.0 4.1 6.8 3.5 2.2

For these data we have $\bar{x} = 3.90$ and $s = 2.46$.

(a) Calculate the statistics in the 'five-number summary': minimum, first quartile, median, third quartile, maximum. Use the 'Type 7' quantiles: $\hat{\pi}_p = x_{(k)}$ where $k = 1 + (n-1)p$.

(b) Sketch a boxplot for these data.

(c) Robert decides to use $\bar{X}$ as an estimator for the mean, μ.

(i) Use the Central Limit Theorem to give an approximate sampling distribution for this estimator.

(ii) Explain clearly the definitions of the following and state what their values are (if known) in this particular problem:

- Population mean
- Sample mean
- The mean of the estimator of μ

(iii) What is Robert's estimate of μ for this dataset?

(iv) Calculate a standard error for Robert's estimate.

(v) Explain clearly the difference between the standard deviation and standard error of an estimator.

Question 8 (6 marks) In a computer games tournament, contestants play a number of games over two rounds and obtain a total score for each round. The games differ in each round, but Jan notices that players tend to perform similarly in both rounds. Using scores from a sample of 20 players, he fits a regression model to predict the scores in the second round from the scores in the first. The following is partial R output from Jan's model fit:

```
lm(formula = score2 ~ score1)

Coefficients:
             Estimate Std. Error
(Intercept) -15.1777    12.5839
score1        0.8542     0.2263
```

(a) Write down the model equation.

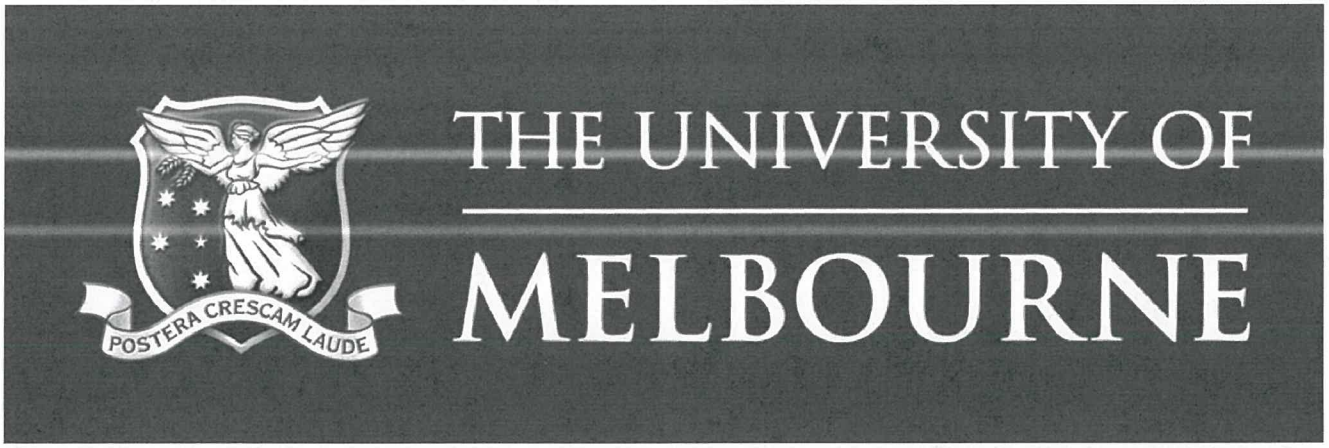
(b) Calculate a 95% confidence interval for the slope.

(c) Is there evidence that the scores from the two rounds are related?

End of exam questions—Total Available Marks = 80
Turn the page for appended material

Appendix (R output)

```
> p1 <- c(0.01, 0.025, 0.05, 0.95, 0.975, 0.99)

> qbeta(p1, 1.00, 39.00)
[1] 0.0002577 0.0006490 0.0013143 0.0739376 0.0902511 0.1113762
> qbeta(p1, 1.00, 40.00)
[1] 0.0002512 0.0006327 0.0012815 0.0721575 0.0880973 0.1087491
> qbeta(p1, 1.00, 45.00)
[1] 0.0002233 0.0005625 0.0011392 0.0644043 0.0787051 0.0972748
> qbeta(p1, 1.05, 43.95)
[1] 0.0002909 0.0007022 0.0013781 0.0678952 0.0826436 0.1017460
> qbeta(p1, 1.05, 44.95)
[1] 0.0002844 0.0006866 0.0013475 0.0664369 0.0808825 0.0996003
> qbeta(p1, 1.05, 45.00)
[1] 0.0002841 0.0006858 0.0013460 0.0663656 0.0807964 0.0994954

> qchisq(p1, df = 1)
[1] 0.0001571 0.0009821 0.0039321 3.8414588 5.0238862 6.6348966
> qchisq(p1, df = 2)
[1] 0.02010 0.05064 0.10259 5.99146 7.37776 9.21034
> qchisq(p1, df = 3)
[1]  0.1148  0.2158  0.3518  7.8147  9.3484 11.3449
> qchisq(p1, df = 4)
[1]  0.2971  0.4844  0.7107  9.4877 11.1433 13.2767
> qchisq(p1, df = 5)
[1]  0.5543  0.8312  1.1455 11.0705 12.8325 15.0863
> qchisq(p1, df = 5)
[1]  0.5543  0.8312  1.1455 11.0705 12.8325 15.0863

> p2 <- c(0.9, 0.95, 0.975)

> qt(p2, df = 8)
[1] 1.397 1.860 2.306
> qt(p2, df = 9)
[1] 1.383 1.833 2.262
> qt(p2, df = 10)
[1] 1.372 1.812 2.228
> qt(p2, df = 18)
[1] 1.330 1.734 2.101
> qt(p2, df = 19)
[1] 1.328 1.729 2.093
> qt(p2, df = 20)
[1] 1.325 1.725 2.086

> qnorm(p2)
[1] 1.282 1.645 1.960

> pnorm(c(0.72, 1.04, 1.4))
[1] 0.7642 0.8508 0.9192
```



2017

① (a) $E(X) = 0\times(1-2\theta) + \theta + 2\theta = 3\theta$ $E(X^2) = \theta + 4\theta = 5\theta$.

$Var(X) = E(X^2) - E(X)^2 = 5\theta - (3\theta)^2 = 5\theta - 9\theta^2$ ✓

(b) (i) $L(\theta) =$

$P_{mf} = \begin{cases} 1-2\theta & x=0 \\ \theta & x=1 \\ \theta & x=2 \end{cases}$

$L(\theta) = (1-2\theta)^{n_0} + \theta^{n-n_0}$.

$n_0+n_1+n_2 = n$

$l(\theta) = \log L(\theta) = n_0 \log(1-2\theta) + (n_1+n_2)\log\theta$ ✓

(ii) $\frac{\partial l(\theta)}{\partial\theta} = \frac{-2n_0}{1-2\theta} + \frac{(n_1+n_2)}{\theta} = 0$.

$\frac{2n_0}{1-2\theta} = \frac{n-n_0}{\theta}$

$2n_0\theta = (1-2\theta)(n-n_0)$

$= n - n_0 - 2\theta n + 2\theta n_0$.

$2\theta n = n - n_0$.

$\hat{\theta} = \frac{n-n_0}{2n}$ ✓

(iii) Sufficient statistic for θ is n_0.

(iv) $N_0 \sim Bi(n, 1-2\theta)$

$E(n_0) = n(1-2\theta)$

$n_0 = n - 2\theta n$

$= n\cdot(1-2\theta)$

$E(\hat{\theta}) = \frac{n-n_0}{2n} = \frac{n-n(1-2\theta)}{2n} = \frac{2n\theta}{2n} = \theta$

unbiased.

(v) $\frac{\partial^2 l(\theta)}{\partial\theta^2} = -2n_0\cdot\frac{(-1)\cdot(-2)}{(1-2\theta)^2} - \frac{n-n_0}{\theta^2}$

$(-1)^{(-2)}\ (1-2\theta)^{-1}\ -2$.

$V(\theta) = -\frac{4n_0}{(1-2\theta)^2} - \frac{n-n_0}{\theta^2}$.

$I(\theta) = -V(\theta) = \frac{4n_0}{(1-2\theta)^2} + \frac{n-n_0}{\theta^2} = \frac{4n(1-2\theta)}{(1-2\theta)^2} + \frac{n-n+2\theta n}{\theta^2}$

$= \frac{4n}{1-2\theta} + \frac{2n}{\theta} = \frac{4n\theta + 2n(1-2\theta)}{(1-2\theta)\theta}$.

$\frac{1}{I(\theta)} = \frac{(1-2\theta)\theta}{2n}$ ✓

$= \frac{2n}{(1-2\theta)\theta}$.

(vi) Yes. MLE.

$Var(\hat{\theta}) = Var\left(\frac{n-n_0}{2n}\right) = \frac{1}{(2n)^2}\cdot Var(n_0) = \frac{1}{4n^2}\cdot n(1-2\theta)\cdot 2\theta$.

$= \frac{\theta(1-2\theta)}{2n}$ ✓.

(c) (i)

0	1	2
8	4	8

$E(X) = \frac{0\times 8 + 1\times 4 + 2\times 8}{20} = 1.2\theta$

$\therefore \theta = 0.5$

$E(\hat{\theta}) = \frac{n-n_0}{m} = \frac{20-8}{2\times 20} = \frac{12}{40} = 0.3$.

$Var(\hat{\theta}) = \frac{\theta(1-2\theta)}{2n} = \frac{0.3(1-0.6)}{2\times 20} = \frac{0.3\times 0.4}{40} = 0.003$

Std. $se(\hat{\theta}) = \sqrt{Var(\hat{\theta})} = \sqrt{0.003} = 0.055$ ✓

(ii) $0.3 \pm 1.96 \times 0.055$ ✓

$\hat{\theta} \sim N(\hat{\theta}, 0.003)$

(iii)

	0.4	0.3	0.3
O	8	4	8
E	8	6	6

$X^2 = \frac{(8-8)^2}{8} + \frac{(4-6)^2}{6} + \frac{(8-6)^2}{6} = \frac{4}{3} = 1.333$

χ^2_1 $k-p-1$

fail to reject H_0. ✓

$\chi^{-1}(0.95) = 3.84$

(iv)

X	0	3
	8	12
	8	12

0.6×20

$X^2 = 0$ fail to reject ✓

$df = k-p-1 = 2-1-1$

Cannot $= 0$.

E 2 5

(2) (i) $E(\bar{X}) = 0\times(1-2\theta) + 1\cdot\theta + 2\cdot\theta = 3\theta = \bar{x}$

$\bar{x} = \frac{N_1 + 2N_2}{n}$

$\hat{\theta} = \frac{\bar{x}}{3} = \frac{N_1+2N_2}{3n}$

(ii) $E(\hat{\theta}) = E(\frac{1}{3}\bar{x}) = \frac{1}{3}E(\bar{x}) = \frac{1}{3}\times 3\theta = \theta$. unbiased.

(iii) $Var(\hat{\theta}) = Var(\frac{1}{3}\bar{x}) = \frac{1}{9}\cdot Var(\bar{x}) = \frac{1}{9}\times\frac{\theta(5-9\theta)}{n} = \frac{5}{9}\theta - \theta$

$E(\bar{x}^2) = \theta + 4\theta = 5\theta$.

$Var(\bar{x}) = E(\bar{x}^2) - E(\bar{x})^2 = 5\theta - 9\theta^2$ $\quad Var(\bar{x}) = \frac{1\theta - 9\theta^2}{n}$

$Var(\hat{\theta}) = \frac{5\theta - 9\theta^2}{9n}$

$\frac{1}{2(\theta)} = \frac{\theta(1-2\theta)}{2n} = \frac{\frac{1}{2}\theta - \theta^2}{n}$

$= \frac{\frac{5}{9}\theta - \theta^2}{n} >$ not achieve.

(ii) MLE, CR bound Lower Variance

(b) $E(\hat{\theta}) = \frac{N_1 + 2N_2}{2n}$ $\frac{4+2\times 8}{3\times 20} = \frac{20}{3\times 20} = \frac{1}{3}$

$Var(\hat{\theta}) \frac{5\theta - 9\theta^2}{9n} = \frac{5\times\frac{1}{3} - 9\times\frac{1}{9}}{9\times 20} = \frac{2}{3}\times\frac{3}{9\times 20}$

$\frac{21}{3}\times\frac{1}{9}\times\frac{1}{10}$ $se(\hat{\theta}) = \sqrt{\frac{1}{270}} = 0.061$

(c)(i) $\theta \pm 1.96\, se(\hat{\theta}) : \frac{1}{3} \pm 1.96\times 0.061$

3 Paired difference

$\bar{d} \pm 2.262 \times \frac{s}{\sqrt{n}}$ $\bar{x} = 3.4$ $s = \frac{1}{n-1}\sum(x_i - \bar{x})^2$

$(6-3.4)^2 + (1-3.4)^2 + (3-3.4)^2 + (-2-3.4)^2$ $\frac{64}{5}$

$Var = 13.7$ $sd = \sqrt{\quad} = 3.7$

$3.4 \pm 2.262 \times \frac{3.7}{\sqrt{10}}$ $\frac{64}{45}$

(b) $H_0: d = 0$ $H_1: d \neq 0$ (iii) $t = \left|\frac{\bar{D}}{s_D/\sqrt{10}}\right| > t^{-1}(0.975)$

(ii) $t = \frac{\bar{D} - 0}{s_D/\sqrt{10}}$ $\frac{9}{4}$ 2.262

(iv) $\frac{3.4}{3.7/\sqrt{10}} = 2.86 > 2.62 \Rightarrow$ reject H_0

4 (a) $p = \frac{262}{500}$ $1-p = \frac{24}{50}$

共 $p \pm c\sqrt{\frac{p(1-p)}{n}}$ $c = 1.96$

$(-0.091, 0.032)$

don't have strong evidence

(b) $p = \frac{245}{520} = 0.49$

$p = 0.48$

$p_1 - p_2 \pm c\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}$ 1.96

(c) $n = \left(\frac{c\sigma}{\epsilon}\right)^2$ $1.96\times\frac{\epsilon^2}{\epsilon^2} = \frac{1.96^2}{4\epsilon^2} = 96.04$

(d) H_1 ... power $= 1-\beta$

(e)

5 : ~~(a)~~ $\sim Bi(n, p)$. $E(\hat{p}) = \frac{1}{40} = 0.025$

(b) Gamma ~~Beta~~ Beta

(c) $\begin{cases} \alpha+\beta = 5 \\ \frac{\alpha}{\alpha+\beta} = 0.01 \end{cases}$ $\begin{cases} \alpha = 0.05 \\ \beta = 4.95 \end{cases}$

prior $Beta(0.05, 4.95)$

(d) $\Rightarrow$ posterior $Beta(x+\alpha, n-x+\beta)$ $(1.05, 43.95)$

(e) mean $= \frac{\alpha}{\alpha+\beta} = \frac{1.05}{1.05+43.95} = 0.023$ $(0.05+1, 39+4.95)$

~~0.023 보다 작다.~~

~~sd = $\sqrt{Var} = \sqrt{\frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}}$~~ $= 0.022278$

$45 \quad 46$

$0.95 = Pr(a < 0.023 < b | x)$ 0.023 ± 0.0436.

$a = \Phi^{-1}(0.025)$ $b = \Phi^{-1}(0.975)$

$= 0.007$ 0.08264

6

$X^2 = \sum \frac{(O-E)^2}{E}$

0.9 1.7 2.2 2.2 3.5 4.1 5.7 6.8 8.0

7 (a) $X_{(1)} = 0.9$

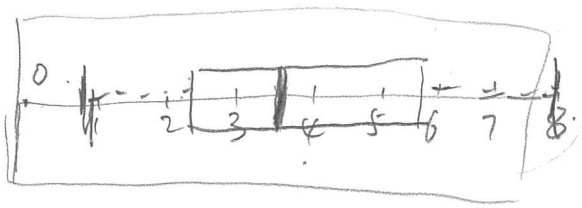
$\ast \pi 0.25 = X_{(3)} = 2.2$ $i = 1 + (8 \times 0.25) = 3$

median $= 3.5$ (5.7) 8.0

$3.5 \times \frac{1}{2} = (5.25)$

중위수 3.5

(b)



7. (i) ~~$\hat{\mu} \sim N(\mu, \frac{\sigma^2}{n})$~~ $\hat{\mu} = \bar{X} \sim N(\mu, \frac{\sigma^2}{n})$

(i) expected value of popular dn. μ.

(ii) every value of observation it sample $\bar{x} = 3.9$

~~(iii)~~ expect. Vale of random Variable $\bar{X}$, some value =

(iii) 3.9

iv. $s(\hat{\theta}) = \frac{s}{\sqrt{n}} = \frac{2.46}{\sqrt{9}} = 0.82$

(v) sd ~~set~~ ~~estimator~~ of sd of sample dn.

estimate of its sd.

8 (a) score 2 = $-15.1777 + 0.8541$ score 1

2. 3.

(b) $0.8542 \pm$ ~~t~~ ~~df~~ 18

Regress t_{n-2}

2.101×0.2263

(c)