THE UNIVERSITY OF QUEENSLAND
AUSTRALIA

This exam paper must not be removed from the venue

Venue ____________

Seat Number ______

Student Number |__|__|__|__|__|__|__|__|

Family Name ____________

First Name ____________

Solutions

School of Mathematics & Physics

EXAMINATION

Semester Two Final Examinations, 2019

STAT2203 Probability Models and Data Analysis for Engineering

This paper is for St Lucia Campus students.

Examination Duration: 120 minutes

Reading Time: 10 minutes

Exam Conditions:

This is a Central Examination

This is a Closed Book Examination - specified materials permitted

During reading time - write only on the rough paper provided

This examination paper will be released to the Library

Materials Permitted In The Exam Venue:

(No electronic aids are permitted e.g. laptops, phones)

Calculators - Casio FX82 series or UQ approved (labelled)

One A4 sheet of handwritten notes double sided is permitted

Materials To Be Supplied To Students:

None

Instructions To Students:

Additional exam materials (eg. answer booklets, rough paper) will be provided upon request.

There are 50 marks available on this exam from 5 questions.

Write your answers in the spaces provided on pages 2 – 13 of this examination paper. Show your working and state conclusions where appropriate. Pages 14 – 18 give formulas and statistical tables. Those pages will not be marked.

For Examiner Use Only

Question	Mark

Total ______

1. [10 marks] A pair of random variables (X, Y) has a joint probability distribution in which $X \sim \mathsf{Uniform}(0, 1)$, and $(Y \mid X = x) \sim \mathsf{Uniform}(-x, 2x)$; that is, the conditional distribution of Y given $\{X = x\}$ is uniform on the interval $(-x, 2x)$.

   (a) Write the joint probability density function of (X, Y), clearly specifying the support of the distribution. [1 mark]

$$f_X(x) = \begin{cases} 1, & x \in [0,1] \\ 0, & \text{else} \end{cases} \qquad f_{Y|X}(y|x) = \begin{cases} 1/3x, & y \in [-x, 2x] \\ 0, & \text{else} \end{cases}$$

$$\begin{aligned} f_{X,Y}(x,y) &= f_{Y|X}(y|x)\, f_X(x) \\ &= \begin{cases} \frac{1}{3x}, & x \in [0,1],\ y \in [-x, 2x] \\ 0, & \text{else.} \end{cases} \end{aligned}$$

   (b) Determine the marginal probability density function of Y. [3 marks]

$$\begin{aligned} f_Y(y) &= \int f_{X,Y}(x,y)\, dx \\ &= \int_0^1 \frac{1}{3x}\, \mathbb{I}(-x \le y \le 2x)\, dx \\ &= \begin{cases} \int_{y/2}^1 \frac{1}{3x}\, dx, & y > 0 \\ \int_{-y}^1 \frac{1}{3x}\, dx, & y < 0 \end{cases} \\ &= \begin{cases} \frac{-1}{3} \log(y/2), & y > 0 \\ \frac{-1}{3} \log(-y), & y < 0 \end{cases} \end{aligned}$$

(c) Using the formula $\mathbb{E}Y = \mathbb{E}[\mathbb{E}[Y \mid X]]$, find the expectation of Y. [3 marks]

$$\mathbb{E}[Y \mid X = x] = \int y\, f_{Y|X}(y|x)\, dy \qquad \text{for } x \in (0,1)$$

$$= \int_{-x}^{2x} y \cdot \frac{1}{3x}\, dy$$

$$= \left[\frac{y^2}{6x}\right]_{y=-x}^{y=2x} = \frac{4x^2}{6x} - \frac{x^2}{6x} = \frac{1}{2}x$$

$$\mathbb{E}[Y] = \mathbb{E}[\mathbb{E}[Y|X]] = \mathbb{E}[\tfrac{1}{2}X] = \frac{1}{2} \times \frac{1}{2} = 1/4$$

(d) Compute $\mathrm{Cov}(X, Y)$. [3 marks]

$$\mathrm{Cov}(X,Y) = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y]$$

$$\mathbb{E}[X] = 1/2 \qquad \mathbb{E}[Y] = 1/4$$

$$\mathbb{E}[XY] = \mathbb{E}[\mathbb{E}[XY|X]] = \mathbb{E}[X\, \mathbb{E}[Y|X]]$$

$$= \mathbb{E}[1/2\, X^2] = \frac{1}{2}\int_0^1 x^2\, dx = 1/6$$

$$\mathrm{Cov}(X,Y) = 1/6 - \frac{1}{2} \times \frac{1}{4} = \frac{1}{24}$$

2. [6 marks] Let X be the continuous random variable with probability density function

$$f_X(x) = \begin{cases} -\log(x), & x \in (0,1) \\ 0, & \text{otherwise.} \end{cases}$$

Define the random variable $Y = -\log(X)$.

(a) Find the probability density function of Y. [3 marks]

$$F_Y(y) = \mathbb{P}(-\log(X) \leqslant y)$$
$$= \mathbb{P}(X \geqslant e^{-y}) = 1 - F_X(e^{-y})$$
$$f_Y(y) = \frac{d}{dy} F_Y(y) = f_X(e^{-y}) \cdot e^{-y}$$
$$= y e^{-y}, \quad y > 0$$

(b) The moment generating function of Y is

$$M_Y(t) = \mathbb{E}(e^{tY}) = (1-t)^{-2}, \quad t < 1.$$

Determine the expected value and variance of Y. [3 marks]

$$\mathbb{E}[Y] = M_Y'(0) \qquad M_Y'(t) = 2(1-t)^{-3}$$
$$= 2 \qquad M_Y''(t) = 6(1-t)^{-4}$$
$$\mathbb{E}[Y^2] = M''_Y(0)$$
$$= 6$$
$$\text{Var}(Y) = \mathbb{E}[Y^2] - (\mathbb{E}[Y])^2 = 6 - 2^2 = 2.$$

3. [10 marks] A study compared the short term effect of playing computer games and watching television on young children's aggression. In the study 28 children aged 4 years to 6 years were randomly assigned to one of two groups with 14 children being assigned to each group. The children were monitored before and after these activities and the level of aggressive behaviour rated on a scale from 1 (no aggressive behaviour) to 100 (very aggressive behaviour).

   (a) The first group of children, those who watched television, experience an average increase of 5.88 in their aggressive behaviour rating with a sample standard deviation of 6.44. Construct a 95% confidence interval for the mean change in aggressive behaviour rating for children after watching television. What does this confidence interval imply in terms of the mean change in aggressive behaviour rating? [4 marks]

Let μ_{TV} be the mean change in aggression rating after watching TV.

observed $\bar{x} = 5.88$, $s = 6.44$, $n = 14$

95% CI $\quad \bar{x} \pm t_{0.975;n-1}\, s.e(\bar{x})$ $\qquad t_{0.975;13} = 2.160$

$$5.88 \pm 2.160 \times \frac{6.44}{\sqrt{14}}$$

$$5.88 \pm 3.7177 \iff (2.162, 9.598)$$

As the 95% CI does not cover 0, there is evidence at the 5% significance level that the mean change in aggression rating is non-zero.

   (b) If the number of children in the study was doubled, but all other sample statistics remained the same, what would happen to the width of the confidence interval? [1 mark]

The width of the CI reduces by a factor of $\frac{1}{\sqrt{2}}$.

(c) The children in the second group, those who played a computer game, experienced an average increase of 3.50 in their aggressive behaviour rating with a sample standard deviation of 8.00. Is there any evidence of a difference in the mean change in aggression rating between the two groups? State the null and alternative hypotheses, and use an appropriate test statistic to determine the P-value. What do you conclude? [5 marks]

Let μ_G be the mean change in aggression rating after playing a computer game.

Test $H_0: \mu_G = \mu_{TV}$ against $H_1: \mu_G \neq \mu_{TV}$

$$\text{test statistic} = \frac{(\bar{x}_G - \bar{x}_{TV}) - 0}{s_p \sqrt{1/n_G + 1/n_{TV}}}$$

pooled variance estimate
$$s_p^2 = \frac{(n_G - 1)s_G^2 + (n_{TV} - 1)s_{TV}^2}{n_G + n_{TV} - 2}$$
$$= \frac{13 \times 8.00^2 + 13 \times 6.44^2}{26}$$
$$= 52.7368$$

$$s_p = 7.262$$

$$\text{test statistic} = \frac{(3.50 - 5.88) - 0}{7.262\sqrt{1/14 + 1/14}} = -0.8671$$

We compare the test statistic with the t_{26}-distribution

p-value $= 2\mathbb{P}(T_{26} \geqslant 0.8671)$ is between 0.2 and 0.5.

There is no evidence of a difference in the mean change in aggression rating between watching television and playing a computer game.

(a) If students do a CI and then comment on whether or not it covers 0, then max 2½ if done correctly.

4. [10 marks] A study was conducted in Germany on the demographics of the online gaming population. Part of this study involved interviewing 688 adult volunteers about their online gaming experience.

(a) The volunteers were asked what type of games they played online (Role-playing, Action, Strategy, Sports and Racing). Of the 464 men interviewed, 325 said they play role-playing games while of the 224 women interviewed 178 said they play role-playing games. Is there evidence of a difference in the proportion of men and women who play role playing games online? State the null and alternative hypotheses, and use an appropriate test statistic to determine the P-value. What do you conclude? [4 marks]

let p_M and p_W be the proportion of men and women gamers who play role-playing games.
Test $H_0: p_M = p_W$ against $H_1: p_M \neq p_W$] 1 mark

test statistic $= \dfrac{(\hat{p}_M - \hat{p}_W) - 0}{se(\hat{p}_M - \hat{p}_W)}$

$\hat{p}_M = \dfrac{325}{464} = 0.700$ $\quad \hat{p}_W = \dfrac{178}{224} = 0.7946$] ½ mark

$se(\hat{p}_M - \hat{p}_W) = \sqrt{\dfrac{\hat{p}_M(1-\hat{p}_M)}{n_M} + \dfrac{\hat{p}_W(1-\hat{p}_W)}{n_W}}$

$= \sqrt{\dfrac{0.7 \times 0.3}{464} + \dfrac{0.7946 \times 0.2054}{224}} = 0.03436$] ½ mark

test statistic $= \dfrac{0.700 - 0.7946}{0.0344} = -2.742$] ½ mark

It is preferable that $se(\hat{p}_M - \hat{p}_W)$ is computed using pooled proportion.
$\hat{p} = \frac{503}{688} = 0.73$
$s.e(\hat{p}_M - \hat{p}_W) = \sqrt{\hat{p}(1-\hat{p})}\sqrt{1/n_M + 1/n_W}$
$= 0.443 \times 0.0813 = 0.036$
test stat $= -2.622$
p-value $= 0.008$

1 mark bounds on p-value ok.

[We compare the test statistic with $N(0,1)$. p-value $= 2 \times \mathbb{P}(Z \geq 2.742)$ is 0.006.] [There is strong evidence against the null hypothesis suggesting a difference in the proportion of men and women who play role-playing games]

½ mark conclusion

* Cannot get full marks if hypotheses say "significant evidence" (max 3)
* χ^2 test ok, but don't think anyone would do that.

Hypotheses $H_0: p_M = p_W$ against $H_1: p_M \neq p_W$

$$X^2 = \sum_i \frac{(e_i - o_i)^2}{e_i}$$

$$e_1 = \frac{503 \times 464}{688} = 339.2$$

$$= \frac{(339.2 - 325)^2}{339.2} + \frac{(124.8 - 139)^2}{124.8}$$

$$+ \frac{(163.8 - 178)^2}{163.8} + \frac{(46 - 60.2)^2}{60.2}$$

$$= 6.82$$

$e_2 = 464 - 339.2 = 124.8 = \frac{185 \times 464}{688}$

$e_3 = 503 - 339.2 = 163.8$
$= 503 \times 224 / 688$

$e_4 = 224 - 163.8 = 60.2$
$= 224 \times 185 / 688$

p-value $= \mathbb{P}(\chi_1^2 \geqslant 6.82)$ is between 0.005 and 0.01

There is strong evidence against the null, suggesting a difference in the proportion of men and women who play role-playing games.

(b) The volunteers were also asked how many years have they been playing online games. The responses are summarised in the table below.

Sex	Online gaming experience (years)			
	< 3 years	3 – 5 years	> 5 years	
Men	223	120	121	464
Women	132	55	37	224
	355	175	158	688

Based on this table, is there evidence of an association between gender and years of experience in online gaming? [6 marks]

1 mark

Test H_0 : no association between gender and experience in online gaming

against H_1: some association between gender and experience in online gaming.

$e_1 = \frac{355 \times 464}{688} = 239.42$

$e_2 = 355 - 239.42 = 115.58$

$e_3 = \frac{175 \times 464}{688} = 118.02$

$e_4 = 175 - 118.02 = 56.98$

$e_5 = \frac{158 \times 464}{688} = 106.56$

$e_6 = 158 - 106.56 = 51.44$

2 marks.

$$X^2 = \underbrace{\sum \frac{(\text{expected} - \text{observed})^2}{\text{expected}}}$$ $\frac{1}{2}$ mark

$$= \frac{(223 - 239.42)^2}{239.42} + \dots + \frac{(37 - 51.44)^2}{51.44}$$

$$= 9.5712$$

1 mark

$P(\chi^2_2 \geq 9.5712)$ is between 0.005 and 0.01.

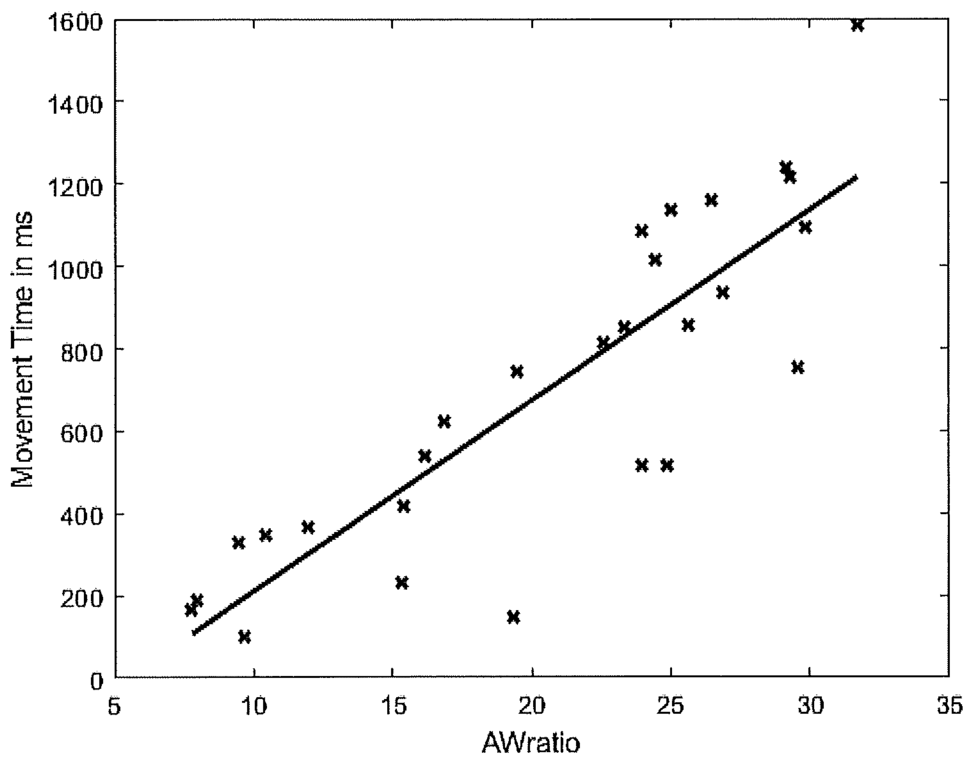
$+\frac{1}{2}$ for correct dof

1 mark

There is strong evidence of an association between gender and experience in online gaming.

* be aware of differences in test statistic due to rounding. Don't penalise this.

5. [14 marks] Twenty seven right-handed university students were recruited into a study to investigate the time taken for a person to move a cursor on a computer screen along a specified path of length A and width W using a mouse. Each student was given a single path along which they were to move the cursor and the time (in ms) taken to complete the task was recorded. The data are displayed in the figure below together with the fitted least squares line.



The output on the next page shows the results of a linear regression fit in MATLAB for the relationship between the time take to complete the task (`MovementTime`) and the ratio A/W (`AWratio`).

```
Linear regression model:
    MovementTime ~ 1 + AWratio

Estimated Coefficients:
                   Estimate      SE       tStat       pValue
                   ________    ______    _______    _________

    (Intercept)    -252.36     120.79    -2.0892     0.047029
    AWratio         46.299     5.5179     8.3906    9.747e-09


Number of observations: 27, Error degrees of freedom: 25
Root Mean Squared Error: 211
R-squared: 0.738,  Adjusted R-Squared 0.727
F-statistic vs. constant model: 70.4, p-value = 9.75e-09
```

(a) Briefly interpret the value 46.299 in the regression output. [1 mark]

An increase in A/W by 1 increases the mean movement time by 46.3 ms.

(b) Briefly interpret the value 211 in the regression output. [1 mark]

The standard deviation of fluctuations about the regression line is estimated as 2.11ms.

(c) Give a 95% confidence interval for the underlying slope of the linear relationship between the time take to complete the task and the ratio A/W. [2 marks]

95% C.I estimate $\pm\ t_{0.975;\,df} \times$ s.e(estimate)

$46.299 \pm t_{0.975;\,25} \times 5.5179$

$46.299 \pm 2.060 \times 5.5179$

$\Leftrightarrow (34.93,\ 57.67)$ ms

(d) Drury's law suggests that the time taken to complete the task will depend on the ratio A/W in the following way:

$$\text{Movement Time} = \text{Constant} \times (A/W)\,.$$

Does the data provide evidence that when the ratio A/W is zero, the time taken to complete the task will be zero? State the null and alternative hypotheses, and report the appropriate test statistic and P-value from the output. What do you conclude? [3 marks]

Let β_0 be the mean movement time when $A/W = 0$.

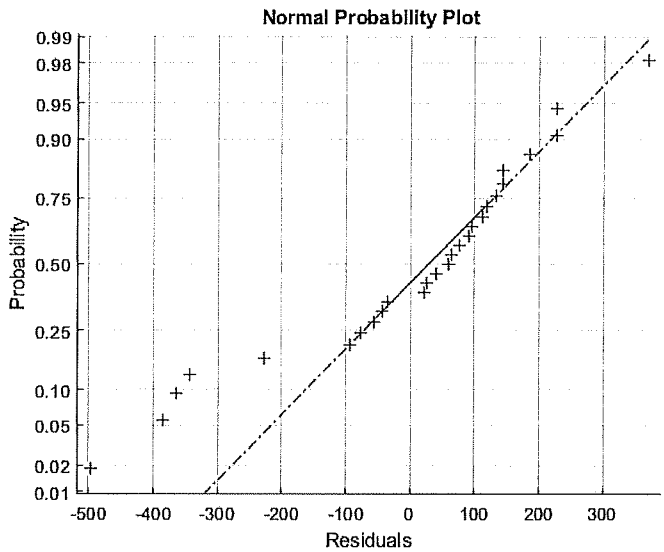
Test $H_0: \beta_0 = 0$ against $H_1: \beta_0 \neq 0$

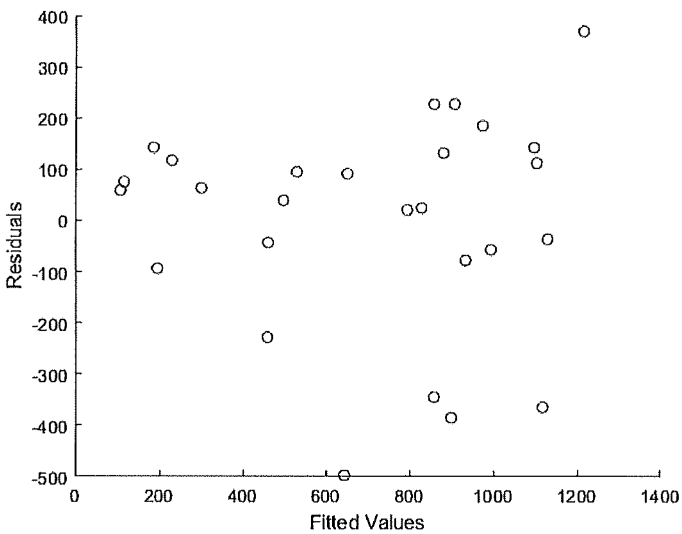
test statistic $= \dfrac{\hat{\beta}_0 - 0}{se(\hat{\beta}_0)} = -2.0892$

p-value is 0.047029.

There is moderate evidence against $\beta_0 = 0$, suggesting the intercept is non zero.

(e) The following figures were generated in MATLAB to help check the assumptions underlying the linear regression. Comment on the validity of the assumptions underlying linear regression for this data with reference to these figures and the figure on page 9. [3 marks]





- From plot on pg 9 and residual vs fitted values, a linear function for the mean movement time appears appropriate.
- Normal probability plot shows a possible departure from normality with longer left tail.

(additional space for answer to part (e))

- Residuals vs fitted values shows possible variance increasing with mean. The regression model assumes constant variance.

(f) The covariance matrix for the estimator of (intercept, slope) is

$$\begin{bmatrix} 14591 & -628 \\ -628 & 30 \end{bmatrix}.$$

Construct a 95% confidence interval for the expected time taken to complete the task when the ratio A/W is 20. [4 marks]

Let μ_{20} be the mean movement time when $A/W = 20$.

$\hat{\mu}_{20} = -252.36 + 46.299 \times 20 = 673.62$

$$s.e.(\hat{\mu}_{20}) = \sqrt{\begin{bmatrix} 1 & 20 \end{bmatrix} \begin{bmatrix} 14591 & -628 \\ -628 & 30 \end{bmatrix} \begin{bmatrix} 1 \\ 20 \end{bmatrix}}$$

$$= \sqrt{1471} = 38.35$$

95% CI $\quad 673.62 \pm t_{0.975;25} \times 38.35$

$(594.61, 752.63)$ ms.

END OF EXAMINATION