FAMILY NAME: . . . .
OTHER NAME(S): . . . .
STUDENT NUMBER: . . . .
SIGNATURE: . . . .

THE UNIVERSITY OF NEW SOUTH WALES

SCHOOL OF MATHEMATICS AND STATISTICS

Semester 2, 2016

MATH2859/MATH2099

All answers must be written in ink. Except where they are expressly required pencils may only be used for drawing, sketching or graphical work.

1. **[20 marks]**

Answer in a separate book marked Question 1

a) **[7 marks]** "Spot or Save" is a game played by Australian school children. When one child spots a yellow car they declare "Spot or Save!" and then punch their friend on their arm until their friend says "Save!" The mean number of yellow cars passing a given school yard is 4.8 per hour.

i) **[2 marks]** What is the probability that at least one child can declare "Spot or Save!" in a 15 minute free time break?

ii) **[2 marks]** There are ten 15 minute free time breaks in one week. What is the probability that "Spot or Save!" can be declared in at least nine 15 minute free time breaks in one week?

iii) **[3 marks]** The school nurse has to spend 5 minutes consoling punched children for each 15 minute free time break period in which a child is punched by playing "Spot or Save". What is the expected number of minutes per week that the school nurse will spend consoling children because of this game? What is the standard deviation?

b) **[6 marks]** One measure of the impact of a twitter feed is the proportion of tweets that are retweeted within 10 minutes of posting. Out of 100 tweets in one week, one engineering researcher had 27 tweets that were retweeted within 10 minutes of posting.

i) **[3 marks]** Construct a two-sided 90% confidence interval for the true proportion of tweets that are retweeted within 10 minutes of posting for this researcher.

ii) **[3 marks]** State two assumptions you need to make in order to determine the above confidence interval. Explain whether each seems reasonable in this situation.

c) **[7 marks]** A battery manufacturer claims that their novel production process allows their batteries to produce more than 100 hours of continual high-power usage on average. A random sample of 50 batteries was tested, and their usable duration recorded. The sample mean duration was $\bar{x} = 105.3$ hours with a sample standard deviation of $s = 15.3$ hours.

Do these data support the hypothesis (1% significance level) that the true mean duration of the manufacturers batteries is greater than 100 hours?

(*Write the detail of the test: null and alternative hypotheses, the distribution of the test statistic under the null hypothesis, an expression for the p-value, and your conclusions in plain language. You may use bounds for the p-value. You may use a test statistic and rejection region.*)

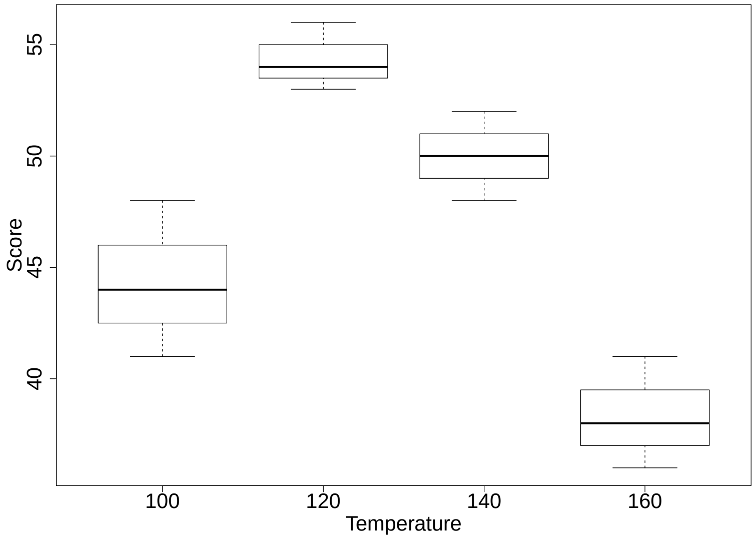
2. [20 marks]

Answer in a separate book marked Question 2

Following a study given in Bassett *et al.* (2000), an industrial plant was maintained at a sequence of increasing temperatures over four successive days. On each day three product samples were taken from the production process and analysed for quality. A score was awarded for each sample, and these are summarised in the table below:

Sample	Temperature (Day)			
	$100°C$ (1)	$120°C$ (2)	$140°C$ (3)	$160°C$ (4)
1	41	54	50	38
2	44	56	52	36
3	48	53	48	41
	$\bar{x}_1 = 44.33$	$\bar{x}_2 = 54.33$	$\bar{x}_3 = 50.00$	$\bar{x}_4 = 38.33$
	$s_1 = 3.51$	$s_2 = 1.53$	$s_3 = 2.00$	$s_4 = 2.52$

Comparative boxplots are given in the figure below.



a) **[3 marks]** What do the boxplots tell you about the scores for different temperatures? Comment on the location, spread and shape.

b) **[3 marks]** State three assumptions that need to be valid for an Analysis of Variance (ANOVA) to test whether there is a difference in mean scores among the four temperatures. Comment on the suitability of these assumptions here, where applicable.

Please see over . . .

Assume from now on that these assumptions are valid.

c) **[3 marks]** An ANOVA table was partially constructed to summarise the data:

Source	df	SS	MS	F
Treatment	**(1)**	**(2)**	**(3)**	23.16
Error	**(4)**	**(5)**	6.25	
Total	**(6)**	484.25		

Copy the ANOVA table in your answer booklet. Complete the table by determining the missing values (1)–(6) **without using the value of** $F = 23.16$, stating how you computed the missing entries. Confirm the value of $F = 23.16$ and explain how this is obtained from the above table.

d) **[5 marks]** Using a significance level of $\alpha = 0.05$, carry out the ANOVA F-test to determine whether there is a difference in mean scores among the four temperatures.

(*You can use the numerical values found in the above table; however, you are required to write the detail of the test: null and alternative hypotheses, rejection criterion or observed value of the test statistic and p-value, conclusion in plain language - you may use bounds for the p-value.*)

e) **[4 marks]** From the previous results, construct a 95% two-sided confidence interval on the difference between the "true" scores for temperatures $100°C$ and for $120°C$, that is, $\mu_1 - \mu_2$. Would you conclude that there is a significant difference between the "true" scores for temperatures $100°C$ and for $120°C$? Explain.

f) **[2 marks]** Six pairwise two-sample t-tests were carried out for comparing the "true" mean score for each temperature. The p-values were also obtained and given in the table below:

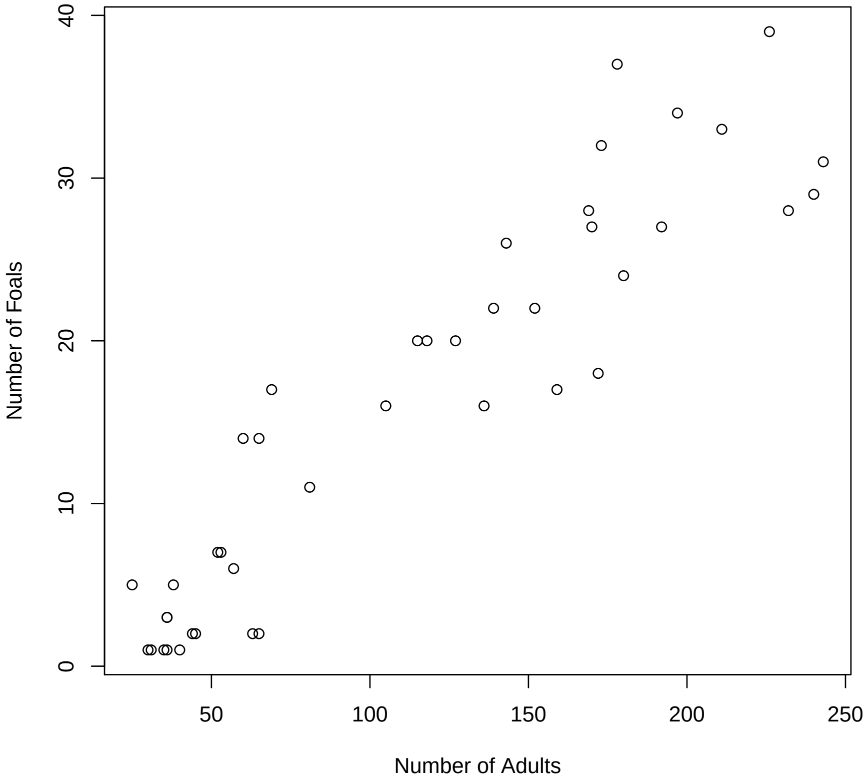
Pairwise comparison	p-value
$100°C$ vs $120°C$	0.00120
$100°C$ vs $140°C$	0.02407
$100°C$ vs $160°C$	0.01873
$120°C$ vs $140°C$	0.06653
$120°C$ vs $160°C$	0.00005
$140°C$ vs $160°C$	0.00045

Does simultaneously analysing the six pairwise comparisons (i.e., those given in the table above) allow you to come to the same conclusion as the ANOVA F-test in (d), at overall significance level $\alpha = 0.05$? Explain.

3. **[20 marks]**

Answer in a separate book marked Question 3

Large herds of wild horses can become a problem on some federal lands in the West. Researchers hoping to improve the management of these herds collected data to see if they could predict the number of foals (young horses) that would be born based on the size of the current herd. They observed 42 herds and recorded how many adult horses and foals were born in each herd. A scatter plot of the `Number of Adults` and the `Number of Foals` is shown below.



The following summary statistics were obtained for the `Number of Adults`

$$\sum_{i=1}^{42} x_i = 4738 \quad \text{and} \quad s_{xx} = \sum_{i=1}^{42}(x_i - \bar{x})^2 = 199970.5$$

In order to study the herd growth, researchers attempted to fit a linear regression model given by

$$\texttt{Number of Foals} = \beta_0 + \beta_1(\texttt{Number of Adults}) + \epsilon.$$

Please see over . . .

Use the following regression output to answer the questions below.

```
Estimated Coefficients:

                     Estimate     SE       tStat      pValue
                    ---------   -------   -------   ----------

   (Intercept)      -1.91793    1.32703   -1.445     0.156
   Number of Adults  0.15862    0.01004   15.807     8.44e-19

Root Mean Squared Error: 4.487
R-squared: 0.862
```

a) [**2 marks**]

   i) [**1 mark**] What is the equation of the fitted linear regression line?

   ii) [**1 mark**] What proportion of variability in the response is explained by the predictor?

b) [**1 mark**] Estimate the true average change in the `Number of Foals` for 1 unit change in the `Number of Adults`?

c) [**1 mark**] Determine the observed sample correlation coefficient between the `Number of Foals` and the `Number of Adults`.

d) [**1 mark**] Assume σ is the standard deviation of the error term ϵ. Give an estimate of σ.

e) [**6 marks**] Perform a hypothesis test to determine whether the variable `Number of Adults` is significant in this model, at the 5% level of significance. (*You can use the numerical values found in the above output, however you are required to write the details of the test: null and alternative hypotheses; rejection criterion, or observed value of the test statistics and p-value (specify the degrees of freedom if applicable); conclusion in plain language.*)

f) [**2 marks**] Create a two-sided 90% confidence interval for β_1.

g) [**4 marks**] Suppose that a new herd with 120 adult horses is located.

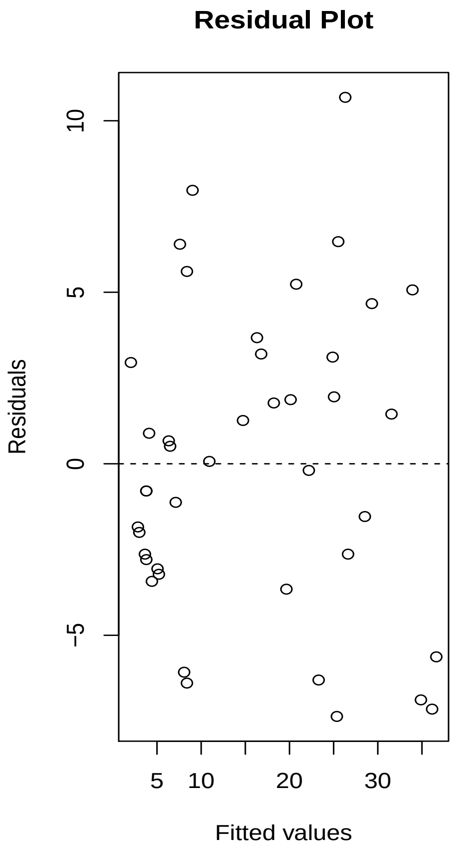
   i) [**1 mark**] What would be a point estimate of the true average `Number of Foals` that may be born?

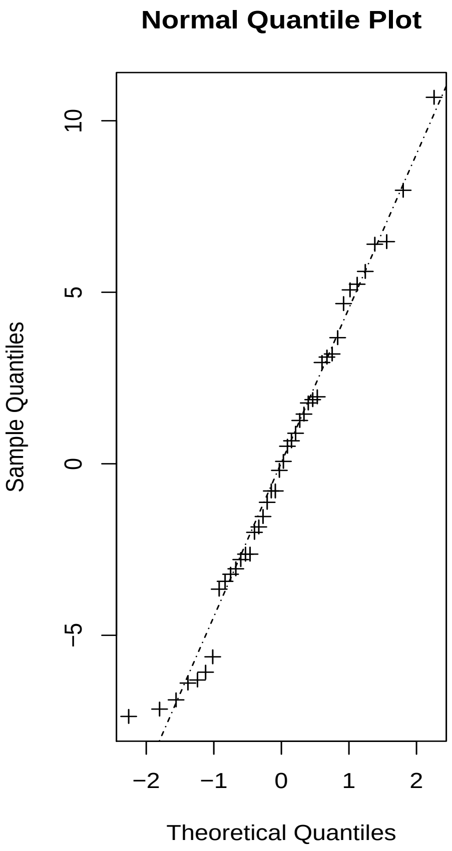
ii) [**3 marks**] Find a two-sided 99% prediction interval for the `Number of Foals`.

h) [**3 marks**]

i) [**2 marks**] For the above regression analysis to be valid, what are the three essential assumptions that the residuals must satisfy?

ii) [**1 mark**] Given the residual versus fitted values plot and the normal quantile plot below, explain why these assumptions are at least approximately valid.


Residual Plot
Residuals
Fitted values


Normal Quantile Plot
Sample Quantiles
Theoretical Quantiles