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, 2016

STAT1201 Analysis of Scientific Data

This paper is for Gatton Campus (External), Gatton Campus and 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 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:

There are **90** marks available on this exam from **5** questions.
Marks are indicated for each question.

Write your answers in the spaces provided in Part A (pages 2–15) of this examination paper. Show your working and state conclusions where appropriate. The backs of pages in Part A may be used for rough working but these will not be marked.

Part B (pages 16–19) gives formulas and statistical tables. Pages in Part B will not be marked.

For Examiner Use Only

Question	Mark
1	
2	
3	
4	
5	

Total ________

Part A – Questions

Question 1 [14 marks]

A recent study in Germany estimated the prevalence of doping in recreational triathletes using a randomised-response method. Anonymous questionnaires were distributed to athletes during the registration procedure on the day before a triathlon event. The questionnaire asked for general information (including gender, age, height, weight) and information about training routines but not for any detailed personal information (such as name, address or date of birth). The final page the questionnaire then included the following:

> **A.** Is your best friend's birthday within the first ten days of a month?
>
> If you answer 'Yes' to **A** then please answer the following:
>
> **B1.** Is your best friend's birthday in the first half of the year?
>
> If you answer 'No' to **A** then please answer the following:
>
> **B2.** Have you taken substances to increase your physical performance in the past 12 months that are only available at a pharmacy, at the doctor's office or on the black market (e.g. anabolic steroids, EPO, growth hormones, stimulants)?

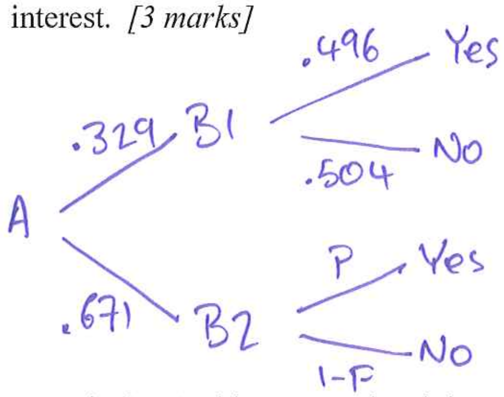
(a) What is the probability that an athlete will be asked to answer the sensitive question **B2**? Identify any assumptions you make in your estimate. *[4 marks]*

Assuming births are equally likely on any day of the year, there are 120 days that are "within the first ten days of a month" so

$$P(B1) = \frac{120}{365} = 0.329,$$

ignoring leap years. Thus $P(B2) = 1 - P(B1) = 0.671$.

(b) Draw a tree diagram to describe the process described above. Indicate the probabilities on each branch of the tree, including the unknown probability, p, of interest. *[3 marks]*



If "first half" is January - June then 181 days so $P(\text{Yes} \mid B1) = \frac{181}{365} = 0.496$.

(c) A total of 534 athletes completed the questionnaire and 102 said 'Yes'. Based on this data, estimate the proportion of recreational triathletes who have used performance-enhancing drugs. *[3 marks]*

Probability of Yes $= .329 \times .496 + .671p$

$= .163 + .671p$

so estimate $.163 + .671p = \frac{102}{534} = .191$,

so $.671p = 0.0278$,

$p = 0.041$,

so estimate 4.1% have used performance-enhancing drugs.

(d) What is the probability that the first 10 athletes will all answer the sensitive question **B2**? What do need to assume to calculate this probability? *[4 marks]*

Assuming independence,

$(.671)^{10} = .0185$.

Question 2 [18 marks]

An engineer wished to test two recyclable materials to compare their suitability for making tyres. She tested two materials, A and B, on the left and right rear positions of ten tractors. The wear on the tread was measured in millimetres of wear, with a lower number indicating less wear. The results are shown in the following table:

Tractor	1	2	3	4	5	6	7	8	9	10
Material A	13.2	8.2	10.9	14.3	10.7	6.6	9.5	10.8	8.8	13.3
Material B	14.0	8.8	11.2	14.2	11.8	6.4	9.8	11.3	9.3	13.6
Difference	-0.8	-0.6	-0.3	0.1	-1.1	0.2	-0.3	-0.5	-0.5	-0.3

R was used to obtain the following summary statistics for the differences:

```
summary(tyres$Difference)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 -1.100  -0.575  -0.400  -0.410  -0.300   0.200

sd(tyres$Difference)
[1] 0.3871549
```

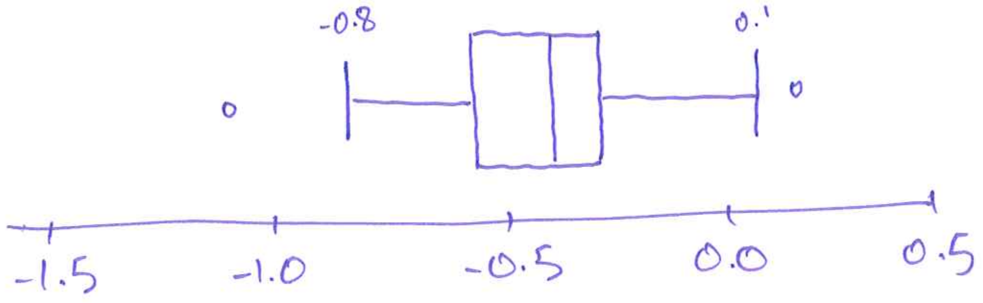
(a) In this experiment, how should the two materials, A and B, be assigned to the two rear positions, left and right, on the tractors? *[2 marks]*

Assigning randomly would help avoid bias in wear.

(b) Using the summary statistics, sketch the boxplot of the differences in tyre wear. *[3 marks]*

$1.5 \times IQR = 1.5 \times (-0.3 - (-0.575)) = 0.4125$

so values below -0.9875 or above 0.1125 flagged.



(c) Describe the distribution of the differences from (b). *[2 marks]*

Roughly symmetric but with two unusual values.

(d) We have two sets of tyre wear results, one for Material A and one for Material B. Briefly explain why we work with the differences rather than carrying out a two-sample *t* test to compare the materials. *[2 marks]*

The materials were on the same tractors so the wear measurements would not be independent.

(e) State the null and alternative hypotheses to address the research question. *[2 marks]*

Let μ = mean difference in wear between A and B.

Test $H_0: \mu = 0$ vs $H_1: \mu \neq 0$.

(f) Carry out the appropriate *t* test to address (e). *[5 marks]*

$$t_9 = \frac{-0.410 - 0}{0.387/\sqrt{10}} = -3.350$$

One-sided P-value is $P(T_9 \leq -3.350)$, between 0.005 and 0.001,

so two-sided P-value is between 0.01 and 0.002,

strong evidence to suggest a difference in mean wear between A and B.

(g) Was this study sufficiently powerful? Briefly justify your answer. *[2 marks]*

Yes. We were able to find evidence of a difference (so cannot have made a Type II error).

Question 3 [22 marks]

Studies suggest that reliving sad memories leads to a reduction in blood oxytocin. In an exploration of 'pet therapy', a student project within the Islands investigated if dogs or cats had a greater feel-good effect by comparing the increase of oxytocin after petting a cat and after petting a dog.

A sample of 40 subjects, ages ranging from 21 to 40 years, were randomly split into two equal groups of 20 and both groups were subjected to reliving sad memories. One group then petted cats while the other group petted dogs, each for 10 minutes. Blood was collected to measure oxytocin levels (pg/mL) for baseline, post-sad and post-pet treatment.

(a) Out of the 40 subjects, 30 experienced a reduction in blood oxytocin after reliving sad memories. Does this give any evidence that reliving sad memories tends to reduce blood oxytocin? *[3 marks]*

If $X = \#$reductions then $X \sim \text{Binomial}(40, 0.5)$ if no effect. P-value is $P(X \geq 30) = 0.001$, strong evidence to suggest blood oxytocin tends to decrease after sad memories.

(b) The results showed that the mean increase in oxytocin while petting cats was 0.499 pg/mL compared to a mean increase of 0.355 pg/mL while petting dogs, with standard deviations of 0.517 pg/mL and 0.445 pg/mL, respectively. Do these results give evidence that cats are more effective at improving blood oxytocin after reliving sad memories than dogs? *[6 marks]*

Let μ_C = mean increase for petting cats

μ_D = —— " —— dogs.

Test $H_0: \mu_C = \mu_D$ vs $H_1: \mu_C > \mu_D$

using

$$t = \frac{(0.499 - 0.355) - 0}{\sqrt{\frac{.517^2}{20} + \frac{.445^2}{20}}} = 0.944$$

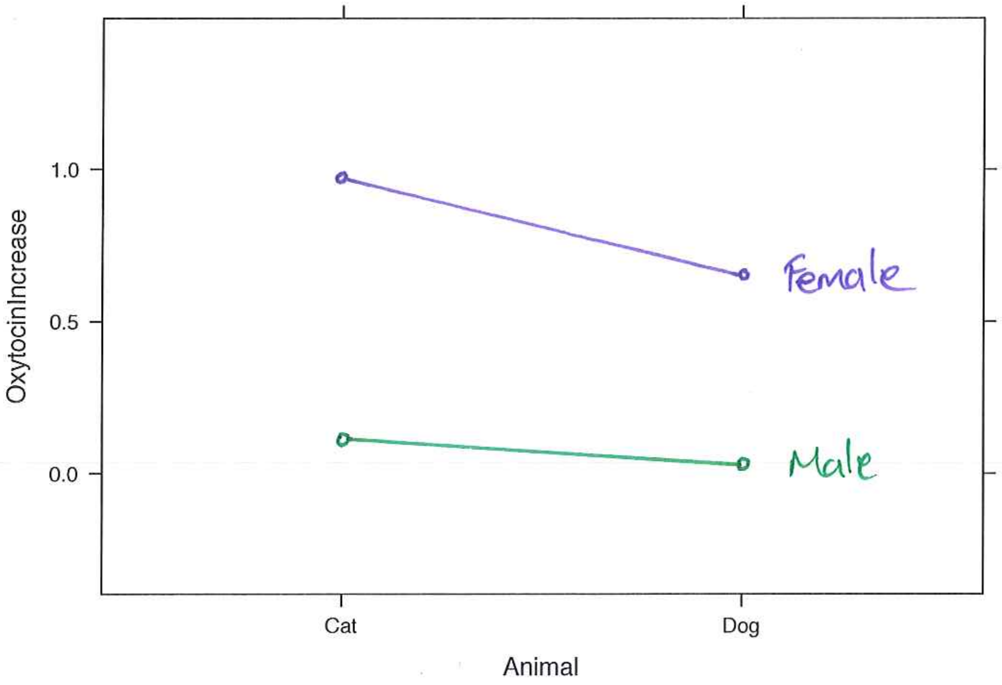
P-value is $P(T \geq 0.944) > 0.10$,

so no evidence that petting cats is more effective at improving blood oxytocin levels.

(c) The table below shows the observed sample means and standard deviations of oxytocin increase for the 10 subjects in each combination of animal and sex.

Animal	Sex	n	Mean	SD
Cat	Male	10	0.013	0.116
	Female	10	0.984	0.166
Dog	Male	10	0.008	0.171
	Female	10	0.702	0.348

Using the axes below, draw an interaction effects plot for this data. *[2 marks]*



(d) Briefly describe the relationships from your plot in ~~(b)~~ (c). *[2 marks]*

There is possibly an interaction since there is little animal effect for males while for females the cat increase seems higher than the dog increase.

(e) Briefly explain why this might undermine the comparison between cats and dogs in ~~(a)~~. *[2 marks]*
(b)

The standard error in (b) may be large because it includes the variability due to sex.

(f) To compare the combined effects of animal and sex on oxytocin increase, a two-way analysis of variance was conducted. The results from R are shown below.

```
summary(aov(OxytocinIncrease ~ Sex*Animal, pets))
            Df Sum Sq Mean Sq F value   Pr(>F)
Sex          1  6.931   6.931 144.674 3.59e-14 ***
Animal       1  0.206   0.206   4.299   0.0454 *
Sex:Animal   1  0.192   0.192   4.004   0.0530 .
Residuals   36  1.725   0.048
```

Calculate and interpret the R^2 value of this two-way model. *[3 marks]*

Total SS = 9.054
Model SS = 7.329

so $R^2 = \frac{7.329}{9.054} = 0.81$, so model explains 81% of variability in oxytocin increase.

(g) Does this two-way analysis of variance provide evidence of a difference between cats and dogs on oxytocin recovery? Compare your results and conclusion with what you found in ~~(a)~~. *[4 marks]*
(b)

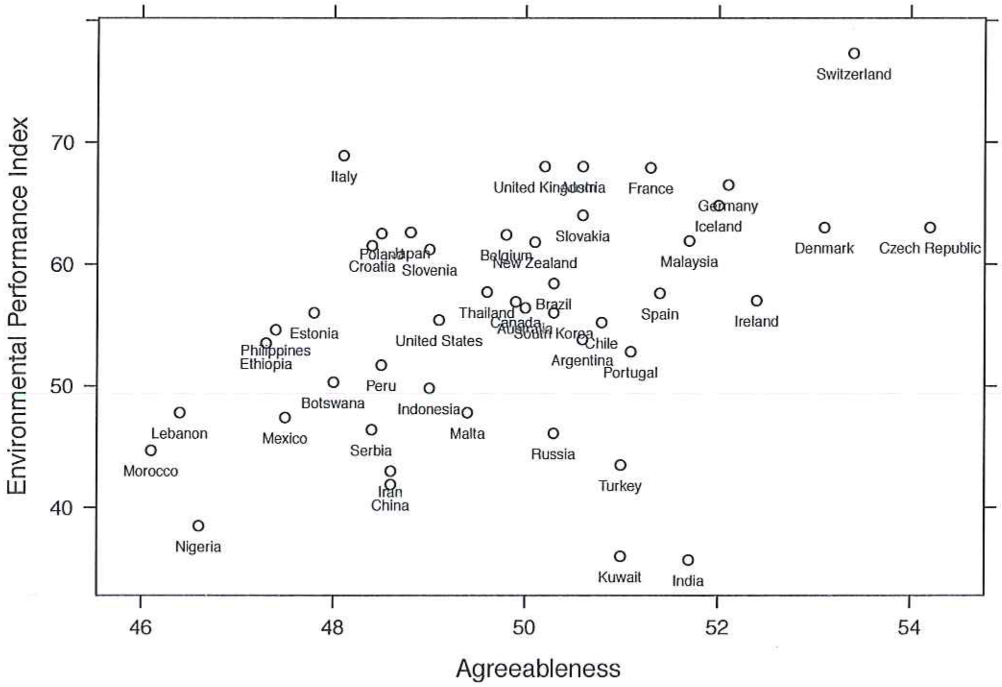
P-value for Animal effect is 0.0454, giving moderate evidence of a difference in oxytocin increase between petting cats and dogs.

This is in contrast to (b) where the two-sample t-test failed to find evidence.

Question 4 [18 marks]

An international study published in 2005 measured personality profiles for a range of countries. For example, Australia received an 'Agreeableness' score of 50.0 compared to 51.7 for Malaysia, suggesting Australians are less 'agreeable' than Malaysians, while the 'Openness' score for Australia was 50.7, higher than the corresponding score of 47.5 for Malaysia.

A researcher in 2014 hypothesised that populations with higher levels of Agreeableness and Openness would be characterized by more sustainable environmental policies. He combined the data from the earlier study with scores on the Environmental Performance Index (EPI), a measure of national environmental sustainability. There were 46 countries for which both sets of scores were available. The following figure shows the relationship between EPI and Agreeableness from this combined data:



In addition to reporting separate correlations between EPI and the two personality scores, the researcher also modelled EPI by Agreeableness and Openness together using multiple linear regression. R gave the following output for this analysis:

```
Response: EPI

Coefficients:
              Estimate Std. Error t value  Pr(>|t|)
(Intercept)  -101.9496    37.5253  -2.717  0.00946 **
Agreeableness   1.3826     0.6810   2.030  0.04854 *
Openness        1.7787     0.6252   2.845  0.00678 **
```

a) The Pearson correlation coefficient for the relationship between EPI and Agreeableness was 0.406. Does this give any evidence of an association between EPI and Agreeableness across the 46 countries? Carry out an appropriate hypothesis test, showing your working and stating your conclusion. *[6 marks]*

Test $H_0: \rho = 0$ vs $H_1: \rho \neq 0$

using $$t_{44} = \frac{0.406 - 0}{\sqrt{\frac{1-0.406^2}{44}}} = 2.947,$$

so one-sided P-value is $P(T_{44} \geq 2.947)$, between 0.005 and 0.001, so two-sided P-value is between 0.01 and 0.002, strong evidence of an association between EPI and Agreeableness.

b) The output from the multiple regression analysis includes three *P*-values. Which of these are relevant to the researcher? Give a brief interpretation of the evidence they provide. *[4 marks]*

The Opennes P-value suggests a fairly significant association between EPI and Openness, after accounting for Agreeableness. Similarly there is some evidence of an association between EPI and Agreeableness, after accounting for Openness.

The intercept P-value is not of interest here.

c) As noted in (a), for the simple relationship between EPI and Agreeableness the correlation was 0.406 so the R^2 value for that model was $0.406^2 = 0.165$. For the multiple regression model, with Openness added, will the R^2 value be higher or lower than 0.165? Briefly justify your answer. *[2 marks]*

Higher, since we have more information to help predict EPI.

d) Australia has an Agreeableness score of 50.0, an Openness score of 50.7 and an Environmental Performance Index of 56.4. Calculate the residual associated with Australia in the multiple regression model. *[2 marks]*

Predicted EPI $= -101.9496 + 1.3826 \times 50.0 + 1.7787 \times 50.7$

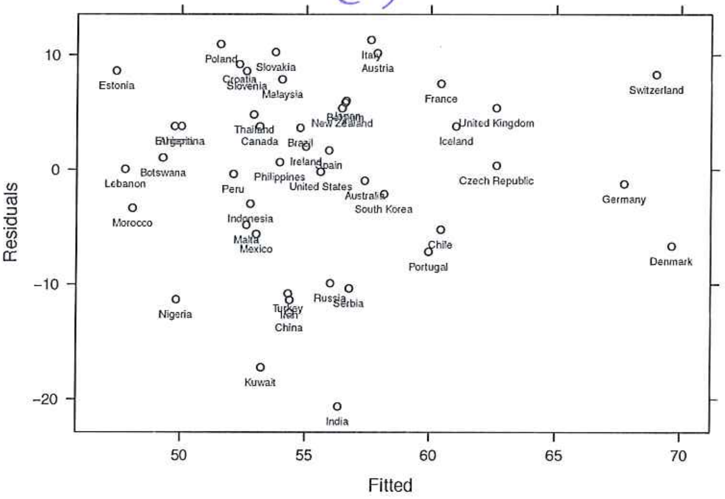
$= 57.4$,

so residual $= 56.4 - 57.4 = -1.0$.

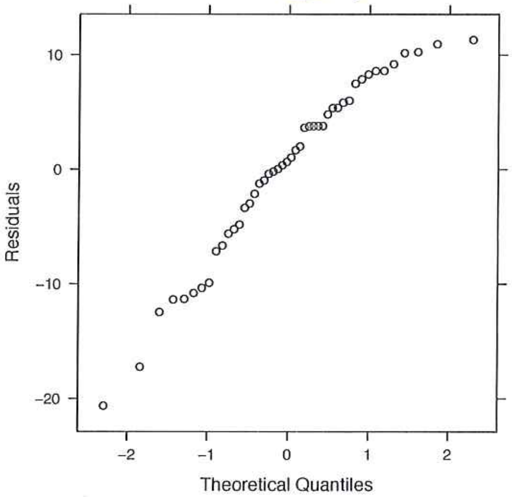
[ie. Australia's EPI is 1 unit lower than predicted by the model with Agreeableness and Openness scores]

e) The following figures were generated by R to help check the assumptions underlying the linear regression:

(1)



(2)



Comment on the validity of the assumptions underlying linear regression for this data with reference to these figures. *[4 marks]*

(1) No strong pattern in residuals and variability seems generally constant. However, residuals seem skewed to the left with more positive values (countries with higher EPI than expected) but then a tail of large negative values (such as India with an EPI 20 points lower than expected). This skewness is reflected in (2).

Question 5 [18 marks]

Children with malignant disease are at increased risk of bone disorders and cardiovascular disease. A study aimed to explore if vitamin D status may influence this risk, by measuring vitamin D levels in children with malignant disease and comparing this to a control group of children (with no malignant disease). The results are shown in the following table:

Vitamin D	Deficient	Not Deficient	
Control children	6	54	60
Children with malignant disease	13	48	61

(a) Calculate the sample proportion of children with malignant disease who have a vitamin D deficiency. *[2 marks]*

$$\frac{13}{61} = 0.2131$$

(b) The researchers believed that the incidence of vitamin D deficiency would be higher for children with malignant disease. Briefly explain why a chi-squared test would not be appropriate for testing this belief. *[2 marks]*

χ^2 test is always two-sided but this is a one-sided question.

(c) Is there evidence that the incidence of vitamin D deficiency is higher for children with malignant disease compared to the control children? *[6 marks]*

Let p_1 = proportion of control who are deficient

p_2 = — " — disease group — " —

observe $\hat{p}_1 = \frac{6}{60} = 0.1$, $\hat{p}_2 = 0.2131$

Test $H_0: p_1 = p_2$ vs $H_1: p_1 < p_2$, or $p_1 - p_2 < 0$.

P-value is $P(\hat{P}_1 - \hat{P}_2 \leq 0.1 - 0.2131)$

$= P(\hat{P}_1 - \hat{P}_2 \leq -0.1131)$

$\simeq P\left(Z \leq \frac{-0.1131 - 0}{\sqrt{\frac{.1\times(1-.1)}{60} + \frac{.2131(1-.2131)}{61}}}\right) = P(Z \leq -1.735)$

$\simeq 0.041$

some evidence to suggest rate of vitamin D deficiency is higher with malignant disease.

Inspired by these initial results, the researchers carried out a further study with new groups of children in which they classified vitamin D levels into *three* groups (deficient, insufficient and sufficient) to gain a better understanding of its role. The results from this second study are given below.

Vitamin D	Deficient	Insufficient	Sufficient	
Control children	11 15.2	28 29.8	50 44.0	89
Children with malignant disease	20 15.8	33 31.2	40 46.0	93
	31	61	90	182

(d) Using this data, is there evidence of an association between vitamin D levels and malignant disease in children? *[8 marks]*

Test using $\chi^2_2 = \frac{(11-15.2)^2}{15.2} + \frac{(28-29.8)^2}{29.8} + \dots + \frac{(40-46.0)^2}{46.0}$

$= 4.09$

P-value is $P(\chi^2_2 \geq 4.09)$, between 0.25 and 0.10, no evidence of an association between vitamin D levels and malignant disease.

END OF EXAMINATION