Universiti Tun Hussein Onn Malaysia

UNIVERSITI TUN HUSSEIN ONN MALAYSIA

FINAL EXAMINATION
SEMESTER II
SESSION 2014/2015

COURSE NAME	:	STATISTICAL MODELLING FOR BIOLOGY
COURSE CODE	:	BWB 42703
PROGRAMME	:	3 BWQ
EXAMINATION DATE	:	JUNE 2015/JULY 2015
DURATION	:	3 HOURS
INSTRUCTION	:	ANSWER **ALL FOUR** QUESTIONS

THIS QUESTION PAPER CONSISTS OF **EIGHT(8)** PAGES

Q1 (a) State the precise definition for each of the following terms.

(i) Factor levels

(1 mark)

(ii) Treatments

(1 mark)

(iii) Experimental unit

(1 mark)

(iv) Designed experiment

(1 mark)

(b) List down three principles of designing an experiment.

(3 marks)

(c) The dry shear strength of birch plywood bonded with different resin glues was studied with a completely randomised designed experiment. The data of the different resin glues could be seen in the **Table Q1(c)(i)** while the ANOVA table in the **Table Q1(c)(ii)**.

Table Q1(c)(i) : Different Resin Glues

Glue A	**Glue C**	**Glue F**
102	100	220
58	102	243
45	80	189
79	119	176
68		176
63		
117		

Table Q1(c)(ii) : ANOVA of Different Resin Glues

Source	df	SS	MS	F	p
Treatment	A	D	F	37.99	0.000
Error	B	8168	628		
Total	C	E			

(i) What are the error degrees of freedom for the shear strength of birch plywood data ?

(2 marks)

(ii) What is the standard error of the difference between the mean of Glue C and the mean of Glue F ?

(2 marks)

(iii) What is the sum of squares for Glue ?

(2 marks)

(d) Independent random samples were selected from three populations with results shown in the **Table Q1(d)**. Create an appropriate ANOVA table.

Table Q1(d) : Three populations

Sample 1	Sample 2	Sample 3
2.1	4.4	1.1
3.3	2.6	0.2
0.2	3.0	2.0
	1.9	

(12 marks)

Q2 (a) Basically, in the factorial approach, the researcher compares all (i) that can be formed by combining the (ii) of the different (iii). Factorial experimentation is highly efficient, because every (iv) supplies information about all the factors included in the (v). It is also a systematic method of investigating the (vi) and (vii) between the effects of different factors.

(7 marks)

(b) State the definition of following terms.

(i) Factors

(1 mark)

(ii) Levels

(1 mark)

(iii) Complete factorial experiment

(2 marks)

(c) State the purpose of having interaction in factorial experiment.

(1 mark)

(d) State two disadvantages of using factorial design.

(2 marks)

(e) Students often complain that having final exams at scheduled times that are different from regular class sessions disrupts their circadian rhythms and affects their performance. You suspect that this might be so and that

the impact would be greatest for more difficult exams. To test this, you conduct a 2 x 3 between subjects experiment in which difficulty level and time of testing are varied for a sample of students who are taking final exams at "different" times. Using the data presented in **Table Q2(e)**, test whether performance depends on time of day and difficulty level of the exam at 0.05 level of significance.

Table Q2(e) : Design Factor A and Factor B

	Time of Testing (Factor B)		
Test of Difficulty (Factor A)	9am	12pm	3pm
Difficult	5	8	10
	4	6	10
	1	6	12
	1	5	8
Easy	11	6	6
	5	9	4
	14	8	3
	10	7	1

(11 marks)

Q3 (a) State either the following statements are **True** or **False**.

(i) A simple linear regression model with explanatory variable, x and outcome variable y, we have these summary statistics for sample means and standard deviations; $\bar{x} = 10$, $s_x = 3$, $\bar{y} = 20$ and $s_y = 5$. For a new data point with $x = 13$, it is possible that the predicted value $\hat{y} = 26$.

(1 mark)

(ii) A standard multiple regression model with quantitative predictors, x_1 and x_2, a factor predictor T with four levels, an interaction between x_1 and T, and an intercept has for its model coefficients an 11 x 1 vector β:

(1 mark)

(iii) In a standard multiple regression model, if a plot of residuals versus fitted values shows a fan-shaped pattern with residuals becoming more spread out as fitted values increase, a log transformation of the response variable may result in data more consistent with model assumptions.

(1 mark)

(iv) If the outcome variable is quantitative and all explanatory variables take values 0 or 1, a logistics regression model is most appropriate.

(1 mark)

(v) In a greenhouse experiment with several predictors, the response variable is the number of seeds that germinate out of 60 planted with each treatment combination. A Poisson regression model is most appropriate for this data.

(1 mark)

(vi) A 3x3 between-subjects design is a two-way ANOVA.

(1 mark)

(vii) The same data is fit with two models using exactly the same predictors. The first model uses standard logistic regression (with `glm(...,family=binomial)`) while the second model accounts for overdispersion (with `glm (...,family=quasibinomial)`). The estimated coefficients for the predictors in the two models will be identical.

(1 mark)

(viii) When there is a single categorical response variable, logistic models are more appropriate than loglinear models.

(1 mark)

(ix) When you want to model the association and interaction structure among several categorical response variables, logistic models are more appropriate than loglinear models.

(1 mark)

(x) A difference between logistic and loglinear models is that the logistic model is a GLM assuming a binomial random component whereas the loglinear model is a GLM assuming a Poisson random component. If both are fitted to a contingency table having 50 cells with a binary response, the logistic model treats the cell counts as 25 binomial observations whereas the loglinear model treats the cell counts as 50 Poisson observations.

(1 mark)

(b) A group of dairy scientists is interested in studying the effect that the percentage of crude protein (CP) in the diet and the stage of lactation (Early, Mid and Late) have on the amount of nitrogen excreted in urine. Dairy cows typically produce milk for at least 300 days after calving. The amount milk they produce per day changes over time. In the study, cows in the early lactation stage averaged 123 days in milk (DIM), mid stage

average 175 DIM and the late stage averaged 221 DIM. There are four treatment diets with CP percentage taking values 15%, 17%, 19% and 21% of diet dry matter. CP was treating as quantitative variable in the model. The outcome variable, urea urine nitrogen (UUN) is measured in grams per day. The scientists conduct an experiment in which four cows at each stage of lactation are given each diet for one week with UUN measurements being taken at the end of this period, a total of $n = 4 \times 3 \times 4 = 48$ measurements. The two models, the first with CP and Lactation as inputs, the second with an interaction term included where the results were summarised in the Table **Q3(b)**.

Table Q3(b) : R-Output

```
Model 1
lm(formula = UUN ~ CP + Lactation)
              coef.est coef.se
(Intercept)  -419.3       29.4
CP             33.8        1.6
LactationMid    4.6        8.7
LactationLate -11.6        8.7
---
n=48, k=4
residual sd=24.73, R-Squared=0.91 ---
```

```
Model 2
lm(formula = UUN ~ CP * Lactation)
                  coef.est coef.se
Intercept)          -427.6     46.6
CP                    34.2      2.6
LactationMid   -80.9   65.8
LactationLate         98.9     65.8
CP:LactationMid        4.8      3.6
CP:LactationLate      -6.1      3.6

n=48, k=6
residual sd=22.96, R-Squared=0.93
```

(i) Each model describes the relationship between UUN and CP with a line for each lactation group. Fill in the Table **Q3(b)(i)** indicating the slope and intercepts for each model for each group.

Table Q3(b)(i) : Model 1 & Model 2

	Model 1		Model 2	
Lactation	Slope	Intercept	Slope	Intercept
Early	A	B	C	D
Mid	E	F	G	H
Late	I	J	K	L

(12 marks)

(ii) For each model, predict the UUN for a cow in the mid lactation stage with a diet CP of 20%.

(2 marks)

(iii) For Model 1, provide an interpretation of slope.

(1 mark)

Q4 (a) Define confounder in a single phrase or sentence.

(3 marks)

(b) Identify the following statements as **True** or **False**. For any falsehoods, propose a modification that would make the statement **True**.

(i) To describe the behavior of a continuous random variable X, we specify its probability mass function $M(k) = P(X = k)$.

(2 marks)

(ii) The p-value associated with a hypothesis test is the probability of having seen an experimental result at least as extreme as the one actually observed, given that the null hypothesis is true.

(2 marks)

(iii) If the Pearson product-moment correlation coefficient between two variables is near zero, we may conclude that there is a weak or nonexistent relationship between those two variables.

(2 marks)

(c) What is the key distinction between well designed experiments and observational studies ?

(2 marks)

(d) A volunteer for a mayoral candidate' s campaign periodically conducts polls to estimate the proportion of people in the city who are planning to vote for this candidate in the upcoming election. Two weeks before the election, the volunteer plans to double the sample size in the polls. What is the main purpose of doing this?

(4 marks)

(e) The manager of a Factory AMIDA wants to compare the mean number of units assembled per employee in a week for two assembly techniques Two hundred employees from the factory are randomly selected and each is randomly assigned to one of the two techniques. After teaching 100 employees one technique and 100 employees the other techniques, the manager records the number of units each of the employees assembles in one week. State the appropriate inferential statistical test in this situation.

(6 marks)

(f) Jason wants to determine how age and gender are related to political party preference in his town. Voter registration lists are stratified by gender and age-group. Jason selects a simple random sample of 50 men from the 20 to 29 age-group and records their age, gender and party registration

(Democratic, Republican, neither). He also selects an independent simple random sample of 60 women from the 40 to 49 age-group and records the same information. Based on information given, what is the most important observation about Jason's plan?

(4 marks)

- END OF QUESTION -