NAME: _____

(360 pts total):

Section 1 – Short answer

1. (10 pts) Describe in words the difference between a “confidence interval” and a “prediction interval” when expressing uncertainty in the fit of a linear model.

2. (10 pts) Describe the process of parametric bootstrap. Under what circumstances might you use parametric bootstrap instead of non-parametric bootstrap?

3. (5 pts) Draw a 4x4 Latin square design for treatments A,B,C,D.

4. (15 pts) Using words and/or equations, explain the difference between treatment contrasts (a.k.a. dummy coding) and sum-to-zero contrasts. Which of these is the default for R's function 'lm'? Which of these is implied by ANOVA?

5. (28 pts) Let $X_1, X_2, \ldots, X_n$ be a random sample of organism lifetimes which we model as being exponentially distributed. In other words,

$$f(X_i|\beta) = \frac{1}{\beta} e^{-X_i/\beta}$$

What is the probability that all of the organisms live more than 2 years? (Make sure you show your work - partial credit for working through the correct steps.)

6. (45 pts)

Question 6A. For each of models A, B, and C below, write down the equation(s) that represent the model described in the text and the R code that would be used to fit that model (you are free to define your own variables as needed). (5 pts for each equation, 5 pts for each R code)

> Model A: We used a generalized linear mixed model to relate variation in abundance (response variable) to geographic origin (a fixed factor, New Zealand or Australia), tidal height (a fixed factor, low tide, mid tide and high tide) and site (a random factor nested within geographic origin, with a total of five sites in New Zealand and five sites in Australia). A Poisson error structure with a log link was appropriate for analyzing the abundance data.

Equation(s):

R code:

> Model B: For body mass as the response variable, two separate generalized linear models were built, one for males and one for females. Geographic origin was treated as a fixed explanatory factor and site (nested within origin) was included in the models as a random explanatory factor. We specified a normal error structure with an identity link in these models.

Equation(s):

R code:

> Model C: To explore sex ratio data, we used a generalized linear mixed model with a binomial error structure and a logit link with geographic origin as a fixed explanatory factor. The number of 'events' in the analysis was the number of males and the number of 'trials' was the total number of male and female individuals.

Equation(s):

R code:

Question 6B. In the above models, why was geographic origin considered a fixed effect and site (within origin) considered a random effect? (5 pts)

Question 6C. How could you rewrite (using an equation, not R code) Model B above to fit the data for males and females together? Would this model yield different parameter estimates? Why or why not? (10 pts)

7. (36 points) Define the following (with examples, as needed) and give the motivation for using or avoiding each (6 pts each):

a. blocking

b. interspersion of treatments

c. Dunn-Sidak Correction

d. Type III Sums of Squares

e. 'variance stabilizing' transformation of the response variable

f. Akaike's Information Criterion (AIC)

Section 2 – Long answer

8. (71 pts) To understand the impact of invasive Argentine ants on the total level of herbivory damage on willow trees, an experiment was carried out as follows: 36 willow trees were randomly chosen, one third from an area that had been invaded by Argentine ants a long time ago, one third from an area that had been invaded only recently, and one-third from an area that had not yet been invaded by Argentine ants. On each tree, 4 branches of similar size and height were randomly selected, and 10 leaves on each branch were collected at random and the percentage of each leaf damaged by herbivores (of all types) was calculated.

A) (5 pts) Should the authors transform the response variable in order to do an ANOVA? Why or why not?

B) (51 pts) Write out the ANOVA table for leaf damage. Which factor(s) should be considered fixed effects and which factor(s) should be considered random effects? (Gray boxes are the ones you have to fill in.)

Source	Degrees of freedom	SS	MS (leave as ratio)	F-ratio (leave as ratio)	p-value
Invasion status					Leave blank
Trees within Status					Leave blank
Branches within Tree					Leave blank
Within groups (residual)				N/A	N/A
Total			N/A	N/A	N/A

C) (5 pts) State the null hypothesis being tested with regards to the main effect of “Invasion status”.

D) (5 pts) What is the implicit alternative hypothesis with regards to the main effect of “Invasion status”?

E) (5 pts) Is there pseudoreplication in this study design or statistical analysis? If not, are there factors in this experiment that may be particularly prone to pseudoreplication?

9. (40 pts) Assume you have a fair coin. How many times would you have to flip the coin to prove that the coin is fair (or, stated more precisely, how many times would you have to flip the coin in order for the width of your 95^{th} percentile confidence intervals to be smaller than 0.01). (Since you don't have calculators, your answer should be an equation that includes DEFINED variables. You must define any variables used in your calculation.)

10. (60 pts) In a study of tadpole predation, *n* identical tanks containing M tadpoles each (and their predators) are monitored and the number of tadpoles eaten k_i in each tank ($i = 1,2, \dots, n$) recorded. (k_i is the number of tadpoles eaten in the i^{th} tank)

Part I: What is the appropriate distribution for the number of tadpoles eaten (10 pts)?

Part II: Using the probability distribution in Part I, write down the likelihood function describing the likelihood of getting the set of observations k_i (i=1,2,…,n) conditional on the parameters of the distribution (10 pts).

Part III: Using the result from Part II, calculate the maximum likelihood estimator (or estimators) for the distribution parameter(s). (Full credit requires that you show all your work for the calculation.) (20 pts)

IV: Defining all necessary parameters, write the R code required to calculate the maximum likelihood estimator(s). Substantial partial credit will be given if you can write down the correct steps, even if you are not sure the correct R functions to use. No points will be deducted for syntax errors that do not reflect errors in statistical thinking. (20 pts)

Section 3 – Essay (40 pts)

11. What is meant by the bias-variance tradeoff in statistical modeling, and how are these two concepts related to expected squared prediction error? (Be sure to define what is meant by “bias”, and what is meant by “variance”). How and why does our concern over bias and variance change between predictive modeling and explanatory modeling?

(PAGE LEFT BLANK AS EXTRA SPACE FOR SHOWING YOUR WORK)