NAME: _____

(360 pts total):

Section 1 – Short answer

1. (10 pts) Derive the expression for the probability of obtaining at least one significant ($p<0.05$) result when doing k comparisons. (Define all variables.)

2. (10 pts) Define statistical power (in words or equations).

	Reject H_0	**Do not reject H_0**
H_0 true	**A (Type I error)**	**B (Correct decision)**
H_0 false	**C (Correct decision)**	**D (Type II error)**

3. (10 pts) Define Akaike's Information Criterion (AIC) and explain the logic behind it.

4. (10 pts) (a) What is the null hypothesis for the Kolmogorov-Smirnov test? (b) What is the test statistic for the Kolmogorov-Smirnov test?

5. (10 pts) Killifish have pigmented cells on their dorsal fin called melanopores that disperse and aggregate during certain circumstances. You want to know whether predators affect the melanopore index (darkness) of killifish. You have three tanks with killifish and no predators, three tanks with killifish and a dogfish shark, three tanks with killifish and a heron, and three tanks with killifish and a seal. After one hour, you take 12 killifish from each tank and measure the melanopore index on 5 randomly chosen scales from each fish.

   What analysis should you use?

   What is the associated model equation for the melanopore index?

6. (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.

7. (10 pts) What is the equation for the coefficient of determination r^2? (Define any variables used.)

8. (16 pts) An investigator wanting to study red deer foraging distances attaches telemetry collars to 20 male and 20 female adult deer and collects data for seven days in July and four days in December. Which of the variables in the study are best treated as random effects and which as fixed effects? What is the appropriate null hypothesis associated with each variable?

9. (15 pts)

(A) What kind of response variables are modeled by logistic regression.

(B) Write down the logistic regression equation.

(C) List three reasons why logistic regression is preferred over ordinary linear regression in cases where it is more appropriate.

10. (18 points) Figure 2 illustrates the possible outcomes for a hypothetical experiment testing for the effects of substrate and predation treatment on barnacle recruitment. Each symbol represents a different treatment combination mean. Predation treatments are indicated by the x-axis label and substrate treatments are indicated by the different shades of gray (black=granite; light gray = slate; darker gray = cement). (The lines connect predation levels for each substrate, and each of the three lines represents a different substrate.) The partial error bars represent + or – 2 standard error and should be assumed equal across all three substrates. (Most of the error bars have been left off the plot for clarity. The exact size of the error bars is inconsequential to the question.)

    For each panel (A-F), decide whether the main effect of predation, the main effect of substrate, and the interaction of predation and substrate is significant ($p<0.05$) or not significant.

    Panel A
    Predation: significant / not significant (circle one)
    Substrate: significant / not significant (circle one)
    Predation x Substrate: significant / not significant (circle one)

    Panel B
    Predation: significant / not significant (circle one)
    Substrate: significant / not significant (circle one)
    Predation x Substrate: significant / not significant (circle one)

    Panel C
    Predation: significant / not significant (circle one)
    Substrate: significant / not significant (circle one)
    Predation x Substrate: significant / not significant (circle one)

    Panel D
    Predation: significant / not significant (circle one)
    Substrate: significant / not significant (circle one)
    Predation x Substrate: significant / not significant (circle one)

    Panel E
    Predation: significant / not significant (circle one)
    Substrate: significant / not significant (circle one)
    Predation x Substrate: significant / not significant (circle one)

    Panel F
    Predation: significant / not significant (circle one)
    Substrate: significant / not significant (circle one)
    Predation x Substrate: significant / not significant (circle one)

11. (15 pts) In a famous study analyzed by R.A. Fisher, data were collected on the numbers of ticks found on 60 sheep. As the data are counts, a useful starting model is the Poisson distribution.

Let

$$Y = \{Y_1, Y_2, \ldots Y_n\}$$

denote the observed values in a random sample drawn from a Poisson distribution with parameter λ.

In other words

$$P(Y_i|\lambda) = \frac{e^{-\lambda}\lambda^{Y_i}}{Y_i!}$$

Suppose that you have other "prior" information (perhaps from previous experiments) that λ has a gamma distribution with parameters α_0 and β_0.

In other words

$$P(\lambda) = \frac{{\beta_0}^{\alpha_0}}{\Gamma(\alpha_0)}\lambda^{\alpha_0 - 1}e^{-\beta_0\lambda}$$

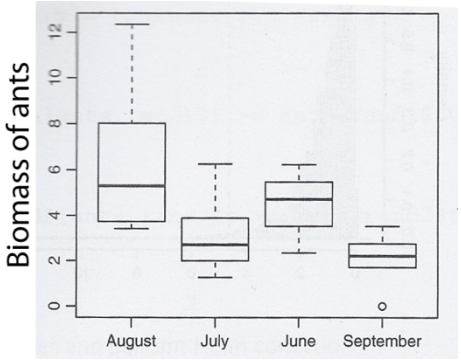
(A) (10 pts) Prove that the "posterior" distribution $P(\lambda|Y)$ is also a gamma distribution. (We would say that the gamma distribution is a "conjugate" distribution for the Poisson likelihood.) (Hint: Focus only on the functional parts of the equation, i.e. the parts involving λ.)

(B) (5 pts) What are the parameters of the gamma distribution for the posterior distribution $P(\lambda|Y)$.

Section 2 – Long answer

12. (55 pts) As part of a study into the diets of the eastern horned lizard (*Phrynosoma douglassi brevirostre*), Powell and Russell (1984,1985) investigated whether the consumption of ants varied over time from June to September. They measured the dry biomass of ants collected from the stomachs of 24 adult male lizards in June, July, August, and September of 1980.

Figure 1



(A) (5 pts) What is the null hypothesis H_0 ?

(B) (5 pts) What is the best parametric statistical test Powell and Russell should use to test the null hypothesis?

(C) (5 pts) Given that test, what is the appropriate alternative hypothesis?

(D) (25 pts) Work out the details of that test, including all equations or ANOVA tables as appropriate. Clearly identify the test statistic and its distribution under the null hypothesis H_0.

Source of variation	SS	dof	MS	F	p
Among groups					
Within groups					
Total					

(E) (15 pts) If Powell and Russell were concerned about inhomogeneous variances across the months, they may be skeptical that the distribution of the test statistic under the null hypothesis H_0 was valid (in other words, they may have the correct test statistic, but be getting the wrong p-value). Describe a non-parametric approach Powell and Russell may use to check the validity of the p-value obtained above.

13. (50 pts)

Consider the following toy model for the relationship between *Growth* (response), *Age* and *Food* (covariates). You have 36 data points.

Model 1: $Growth \sim \beta_0 + \beta_1 Age + \beta_2 Food + error$, $error \sim N(0, \sigma^2)$ / Log-likelihood (Model 1) =

Model 2: $Growth \sim \beta_0 + \beta_1 Age + error$, $error \sim N(0, \sigma^2)$ / Log-likelihood (Model 2) =

Model 3: $Growth \sim \beta_0 + \beta_2 Food + error$, $error \sim N(0, \sigma^2)$ / Log-likelihood (Model 3) =

Model 4: $Growth \sim \beta_0 + error$, $error \sim N(0, \sigma^2)$ / Log-likelihood (Model 4) =

(A) (10 pts) Which combinations of models can be compared by a likelihood ratio test?

(B) (15 pts) What is the test statistic and distribution under the null hypothesis for the likelihood ratio test?

(C) (12 pts) Fill in the AIC model selection table below (for the last column, feel free to leave the answer as a mathematical expression).

Model	# parameters	AIC	AICc
1			
2			
3			
4			

(D) (13 pts) Discuss briefly the pros and cons of using the likelihood ratio test vs. an Information Theoretic approach such as AIC.

14. (40 pts)

Capture-recapture studies were originally developed in the wildlife biology to estimate demographic parameters and trends in population studies. The classical problem of estimating the unknown size of a closed population is the main issue of this case study.

In 1998, biologists sampled, in the surveyed site, the ovipositing female population of *Salamandrina perspicillata* (a salamander) over 11 occasions. Only the oviposition period, which occurs in winter-early spring, was considered so that the population size remains fixed during the study time. Individuals were captured, marked and then released and allowed to mix again with the general population. Subsequent recaptures were performed and the marked individuals were recorded.

The recorded counts of capture-recaptures were: f_1=81, f_2=17, f_3=0, f_4=1, where f_k is the frequency of individuals captured exactly k times in the 11 trapping occasions. (There were no animals caught more than four times, so f_5=0, f_6=0, f_7=0,…, f_{11}=0.) The maximum possible frequency for each individual is the number of trapping occasions (11 in this case). The number of distinct females caught in the experiment was n=99.

The complete capture history for each female is expressed as a sequence of 0's and 1's, where 0 denotes "not captured" and 1 denotes "captured". So we have a 99x11 matrix = $\overleftrightarrow{X} = x_{ij}$, where x_{ij}= [the *i*th individual is caught (1) or not (0) in the *j*th trapping occasion] *i* = 1, 2,…, 99; *j* = 1, 2,…,11.

Animal	Trapping occasion 1	Trapping occasion 2	Trapping occasion 3	Trapping occasion 4	Trapping occasion 5	Trapping occasion 6	Trapping occasion 7	Trapping occasion 8	Trapping occasion 9	Trapping occasion 10	Trapping occasion 11
1	0	0	1	1	0	0	0	1	0	0	0
2	0	0	0	0	0	0	0	0	1	0	0
3	0	1	0	0	0	0	0	1	0	0	0
4	1	1	0	0	0	0	0	0	0	0	1
5	0	0	1	0	0	0	0	1	0	0	0
:	:	:	:	:	:	:	:	:	:	:	:
99	0	0	0	0	0	1	0	0	0	0	0

The number of individuals never observed (caught zero times) f_0 is unknown. The total population size (N) is also unknown but can be expressed as:

$$N = f_0 + f_1 + f_2 + f_3 + f_4 = f_0 + n = f_0 + 99$$

To estimate *N*, we will use the non-parametric estimator proposed by Chao (1984):

$$\widehat{N_c} = n + \frac{f_1^2}{2f_2}$$

Assuming random recaptures, the capture frequencies contain all the information to estimate the number of missing individuals in the samples. With the data of this experiment the value $\widehat{N_c} = 292$ has been obtained.

(A) (4 pts) What would be a reasonable sampling distribution for

i. A single capture event?

ii. A single capture history (capture success over 11 capture attempts)?

iii. The total population size N?

iv. The non-parametric estimator $\widehat{N_c}$?

(B) (15 pts) Describe how you would use a parametric bootstrap to calculate the bias for $\widehat{N_c}$. (I am particularly interested in knowing how you would do the bootstrap sampling, but please include the formula for bootstrap bias as well.)

(C) (21 pts) Describe two methods of non-parametric bootstrap sampling to calculate the bias for $\widehat{N_c}$. One of these two methods is preferred, which one? (Hint: What happens when an animal dies?)

15. (40 pts)

Hecnar et al. (2002) studied the community composition of amphibians and reptiles among several archipelagoes in the Great Lakes of North America with the goal of documenting and explaining patterns of community similarity across a variety of spatial and taxonomic scales. They sampled four island groups (Apostle Islands, Georgian Bay Islands, Lake Erie Islands, St. Lawrence Islands) and found a total of forty-five species across four Orders (snakes, turtles, frogs, salamanders). For each Order x Archipelago combination, they calculated the similarity between the community of that sampled unit and the species represented in the larger geographic region (the 'source' pool of species). (In other words, each cell contains a single number which is the community similarity. There is no replication in this design.)

Hecnar et al. did a 2x2 factorial ANOVA of species similarity without interaction. They considered Order and Archipelago random effects. Use the following table to answer the questions that follow:

Source	d.f.	SS	MS	*F*	*P*-value
Two-way					
Orders	A	D	255.1	G	I
Archipelagoes	B	E	54.9	H	J
Error	C	F	35.1		

(A) (2 pts) What is A for this analysis?

(B) (2 pts) What is B for this analysis?

(C) (2 pts) What is C for this analysis?

(D) (4 pts) What is the formula for D (i.e. the actual sum-of-squares equation)?

(E) (4 pts) What is the formula for E (i.e. the actual sum-of-squares equation)?

(F) (4 pts) What is the formula for F (i.e. the actual sum-of-squares equation)?

(G) (2 pts) What is G (leave as a fraction)?

(H) (2 pts) What is H (leave as a fraction)?

(I) (3 pts) Comparing I and J, which is the smaller p-value and why? How would we interpret this result?

(J) (5 pts) How would the analysis change if you assumed Order and Archipelago were fixed effects?

(K) (10 pts) What is the difference in interpretation between assuming Order and Archipelago are random effects vs. assuming they are fixed effects (Hint: Write down the null hypothesis for each)?

Section 3 – Essay (41 pts)

16. Describe the steps involved in hypothetico-deductive reasoning. In what way could this be said to convolve predictive modeling with explanatory modeling?