NAME: _____

(360 pts total):

Section 1 – Short answer

1. (10 pts) Why is the confidence interval bow-shaped (narrower in the middle) in the linear regression illustrated in Figure 1 below.

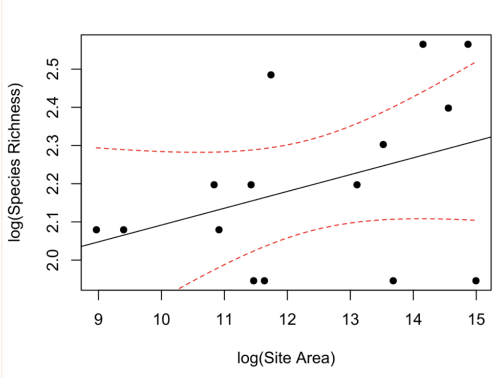


Figure 1

2. (7.5 pts each)

(a) Define statistical power.

(b) Why is statistical power not a meaningful concept for experiments carried out by a computer simulation?

3. (10 pts) What is the equation for the coefficient of determination R^2 (or equivalently, r^2) in the context of linear regression?

4. (15 pts) A researcher is measuring the length of fish (in centimeters) in a series of ponds, and is interested in the relationship between fish length and lake characteristics such as mean depth (measured in meters), peak summer water temperature (in Celsius) and lake area (meters-squared). The researcher measures one fish per pond ($Y_i; i = 1, \dots, n$); her model looks like:

$$Y_i \sim N(\beta_0 + \beta_1 Depth + \beta_2 Temp + \beta_3 Area, \sigma^2)$$

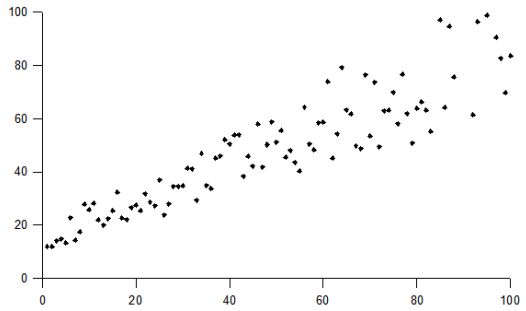
(a) What is the Mean Squared Error (MSE) for this model?

(b) What are the units of β_2?

(c) What are the units of σ^2?

5. (10 pts) The following data (Figure 2) violates an assumption of linear regression. What assumption is violated? How would you test whether the assumption is violated? How would you analyze the data appropriately?

Figure 2



6. (20 pts)

Brosi and Biber wrote a review in 2009 on the use of statistical inference in conservation. I show the abstract below to provide some context for their discussion.

REVIEWS REVIEWS REVIEWS

Statistical inference, Type II error, and decision making under the US Endangered Species Act

Berry J Brosi[1] and Eric G Biber[2]

Critical conservation decisions have been made based on the spurious belief that "no statistically significant difference between two groups means the groups are the same". We demonstrate this using the case of the Preble's meadow jumping mouse (*Zapus hudsonius preblei*), an endangered species in the US. Such faulty statistical logic has been recognized before, but ecologists have typically recommended assessing post hoc statistical power as a remedy. Statisticians, however, have shown that observed power will *necessarily* be low when no differences are found between two populations. Alternatives to assessments of statistical power include equivalence testing (a method rarely used by ecologists) and Bayesian or likelihood methods. Although scientists play a central role in ameliorating this problem, the courts could also assist by requiring litigated federal agency decisions to consider the risks of both Type I and Type II errors.

Front Ecol Environ 2009; 7(9): 487–494, doi:10.1890/080003 (published online 6 Nov 2008)

In their paper, the authors make the following comments in regard to the difficulty of interpreting genetic tests (to determine if two populations are genetically distinct) that fail to find statistically significant differences:

> The fundamental cause of this confusion is that standard statistical tests are set up in a way that gives researchers relative certainty about the result only when the test shows a significant difference between two groups. By definition, you can be at least 95% sure that you are correct when a hypothesis test finds a significant difference (assuming the standard $\alpha = 0.05$; Figure 2, top row). But when the test outcome is not significant (Figure 2, bottom row), there is no way to reliably estimate how likely you are to be wrong (from Type II error) if you conclude that the populations in question are homogeneous (eg Hoenig and Heisey 2001). Thus, if a significant difference is not found in a statistical test, the only appropriate conclusion is that the *null hypothesis cannot be rejected.*

a) (10 pts) The authors make some valid points in this paragraph: Name one of them.

b) (10 pts) The authors make at least one major error in this paragraph: What is it? Why are the authors incorrect?

7. (5 points each) Define the following (with examples and diagrams and/or equations, as needed) (6 pts each):

a. leverage

b. Pearson's product moment correlation coefficient

c. variance inflation factor

d. Type III Sums of Squares

e. maximum likelihood estimate (of a parameter) [Define this one in words only, no drawings or equations!]

f. Akaike's Information Criterion (AIC)

8. (20 pts) Assume that a researcher has data on the probability of plant germination across species, and she is going to model that probability as a Beta distributed random variable.

$$X \sim Beta(\alpha, \beta)$$

and use maximum likelihood to estimate the parameters of this Beta distribution $\hat{\alpha}$ and $\hat{\beta}$.

a) (10 points) Describe in words the interpretation of the standard error of a parameter estimate ($\hat{\alpha}$ or $\hat{\beta}$).

b) (10 pts) The research says to her colleague "Good news, the contours of the log-likelihood function are elliptical, suggesting the data are not ill-conditioned." What was the researcher concerned about with regards to calculating the confidence intervals of the parameters α and β?

Section 2 – Long answer

9. (50 pts)

In Orrock et al. (2011) *PNAS* 177(5): 691-691, the authors try to understand the role of precipitation ('Precip'), island area ('Area'), and the species richness of rodent predators ('Pred') on the prevalence (% infection) of Sin Nombre Virus (SNV) among deer mice living on the eight islands of the Channel Islands in California.

a) Assume the authors sample 75 mice at random from each island. Write the complete model equation for the most appropriate method of analyzing this dataset using the three covariates *Precip*, *Area*, and *Pred*. Make sure to include all necessary indices and explain all variables used to write the model equation. (20 pts)

b) Describe two methods for testing whether precipitation should be retained in the final model as a statistically significant covariate for SNV prevalence. (10 pts each)

c) Should the authors add ‘Island’ to their model? Why or why not? (5 pts)

d) If the authors do add ‘Island’ to their model, should they add “Island’ as a fixed or a random effect and why? (5 pts)

e) Let's say that instead of sampling 75 mice at random, the authors sampled 5 mice from 15 colonies on each island. Re-write the model equation from part (a) in the most appropriate way to accommodate the new sampling design. (10 pts)

10. (65 pts) A researcher is interested in how the size of wolf pack territories (which reflects, among other things, prey density) in each pack at Yellowstone National Park changes as a function of whether its territory was burned in the 1988 Yellowstone fires (Burned vs. Not Burned) and the dominant tree species in the territory (lodgepole pine, whitebark pine, aspen, Douglas-fir, Engelmann spruce). Assume a balanced design, whereby 4 packs are surveyed in each combination of conditions (40 packs total). Assume fire status is to be treated as a Fixed Effect and Forest Type as a random effect. For the purposes of this question, we will assume that territory size is approximately Normally distributed, and that the data meet the assumptions of ANOVA.

a) Write the equation for this model and complete the ANOVA table (20 pts for equation; 20 pts for ANOVA table). Be sure to define all variables and indices.

Equation:

Source	SS	dof	MS	F ratio
Burned status				
Forest type				
Interaction				
Within				
Total				

b) Name all the null hypotheses (H_0) and their implied alternative hypotheses (H_A) being tested in the ANOVA table. (10 pts)

c) Name two reasons why the authors might prefer the two-way ANOVA as described above rather than two separate one-way ANOVAs (one for fire status and one for forest type). (15 pts)

11. (75 pts) The CDC wishes to estimate the probability of contracting Covid-19 in each of n states (New York, Virginia, Texas, etc.), and they sample M people in each state for antibodies to the virus (indicative of prior infection). k_i is the number of people with antibodies to Covid-19 ($i = 1,2, \dots, n$) recorded.

a) What is the appropriate statistical distribution for the number of people infected by Covid-19 in this study (5 pts)?

b) How many parameters are there in this model that need to be estimated from the data if we assume all states have the same probability of infection? (5 pts)

c) Using the probability distribution in (a), write down the likelihood function describing the likelihood of getting the set of observations k_i ($i = 1,2, \dots, n$) conditional on the parameter(s) of the distribution (10 pts).

d) Using the result from (c), calculate the maximum likelihood estimator (or estimators) for the distribution parameter(s) involved in this model. (Full credit requires that you show all your work for the calculation.) (20 pts)

e) Let's say that the CDC wants to fit a model in which the probability of infection is different for each state. Write down the likelihood function describing the likelihood of getting the set of observations k_i ($i = 1,2, \dots, n$) conditional on the parameter(s) of the distribution (10 pts).

f) How many parameters does the model from (e) have? (5 pts)

g) Is there an advantage to fitting a single model to all the data as compared to fitting a separate model for each state? Why or why not? (10 pts)

h) What if the CDC decides not to estimate an independent parameter for each state, but decides to construct a model in which infection probability itself is modelled as coming from its own distribution. What would be the most appropriate distribution for infection probability? (5 pts)

i) Using the suggested model from (h), how many parameters need to be estimated from the available data? (5 pts)

12. (40 pts) Capture-recapture studies were originally developed in 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) (2.5 pts each) 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) (15 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?)

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