Biometry mid-term exam Spring 2014

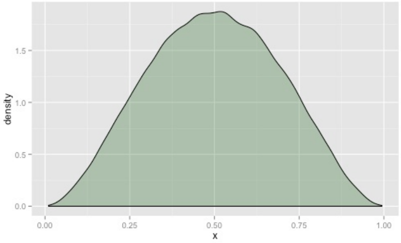
NAME: ______________________________

(250 pts total):

Section 1 – Short answer

1. (10 pts) What is the expected value of a draw from the sequence {1, 2, 3, 4, 5} with probabilities (0.2, 0.1, 0.1, 0.2, 0.4)?

2. (10 pts) Sketch the probability density function for the $Beta(\alpha = 3, \beta = 3)$ distribution. (Your sketch is not going to be exact, but should contain some key features.)



3. (5 pts) Derive the expression for the probability of obtaining at least one significant (p<0.05) result when doing k comparisons. (Define all variables.)

4. (15 pts) Complete the following equations:

$$Beta(\alpha = 1, \beta = 1) =$$

$$If\ X_1, X_2, \dots, X_n \sim N(0,1), then\ \sum_{i=1}^{n} X_i^2 =$$

$$\lim_{n \to \infty} Binom(n, p) \to$$

5. (10 pts) Assume that you have data

$$X = \{X_1, X_2, X_3, \dots, X_{100}\}$$

representing the petal length of a population of flowering plants, which you assume are drawn from a Normal distribution, i.e.

$$X \sim N(\mu, \sigma^2)$$

Using your data, you calculate that

$$\frac{1}{n}\sum_{i=1}^{100} X_i = 3.2\ cm$$

and

$$\sqrt{\frac{1}{n-1}\sum_{i=1}^{100}(X_i - \bar{X})^2} = 0.8\ cm$$

a) Approximately what proportion of the population would be expected to have a petal length greater than 4.5 cm?

b) What is the standard error on the statistic $\bar{X}$?

6. (10 pts) Define statistical power (in words or equations).

7. (6 pts) What function would you use in R to get the likelihood of a single draw $X = 3$ from a Poisson distribution? (Note: I'm not looking for the maximum likelihood, just the likelihood.)

8. (10 pts)

Biometry mid-term exam Spring 2014

a) Using an equation, state the Central Limit Theorem.

b) For what kinds of distributions does the Central Limit Theorem not apply?

9. (6 pts each) A logging company wishes to harvest timber from a forest containing a population of Spotted Owls. An environmental group opposes the timber harvest on the basis that the logging will adversely impact the owl population (defined as a 20% or greater decrease in fecundity). In order to mitigate the dispute, a study is convened to compare the fecundity of Owls in undisturbed habitats with the fecundity of Owls in habitats disturbed by logging.

a. What is the null hypothesis of this study? What is an appropriate alternative hypothesis? Which hypothesis corresponds to the position of the environmental group?

b. What outcome of the study would represent a Type I error? Which of the two parties is most concerned with minimizing Type I error?

c. What outcome would represent a Type II error? Which of the two parties is most concerned with minimizing Type II error?

d. What is the number of owls that we would want to measure (N) in order to detect a difference in fecundity of 0.2 (20%) if the variance in fecundity is 0.5? (You can leave it as a fraction.)

10. (10 pts) How would you test for the equality of means between two sets of data that were paired? (Hint: You need to state both the test statistic and its distribution under the null hypothesis.) What would be the consequence of treating them as unpaired samples?

11. Fill in the three empty boxes.

Part I (15 pts)

Test	Hypothesis (assuming two-tailed tests)	Test statistic T	$f(T\|H_0)$ (Distribution of T under H_0)	Assumptions
Binomial test for proportions				

Part II (15 pts)

Name one problem with the confidence intervals constructed using the Wald test for binomial proportions.

Section 2 – Long answer

12. (40 pts) The Rayleigh distribution is a continuous probability distribution for positive-valued random variables, often used to describe the length of a vector Z when its X and Y components are normally distributed. (For example, if the east-west component of wind is normally distributed, and the north-south component of wind is normally distributed, then the magnitude is Rayleigh distributed.)

The Rayleigh distribution PDF is given by

$$f(z|\sigma) = \frac{z}{\sigma^2} e^{-z^2/2\sigma^2}, z \geq 0$$

a) Find the maximum likelihood estimator for the parameter σ. [Hint: take the square root of the MLE for σ^2.] (20 pts total – 10 for knowing to take the derivative of the NLL and set to 0, another 10 for working through the algebra correctly.)

b) Describe in words how would you find the confidence intervals for σ using maximum likelihood? (7 pts)

c) Describe in words how would you find the confidence intervals for σ using bootstrap (7 pts)?

d) Independent of the method used to construct them, what is the correct interpretation for the 95th percentile confidence intervals? (6 pts – NO PARTIAL CREDIT)

13. (45 pts) Yu and Davidson (1997) looked for associations between ant species and seven members of the *Cercopia* family of tropical trees in Madre de Dios, Peru. The table below lists the numbers (n) of each species encountered in the survey and the number (x) of those trees on which the any *C. balzani* was found. These data may be used to ask if *C. balzani* use one tree species more than another. Species 1 has the highest count of trees with *C. balzani*, but it is also the most abundant *Cercopia* in the study area, so we need to account for host availability. We also need to allow for chance variations in the occurrence of ants.

Species	1	2	3	4	5	6	7
$\boldsymbol{n_i}$	53	6	7	31	17	17	24
$\boldsymbol{x_i}$	16	0	5	3	2	1	12

a) (10 pts) What is the best statistical distribution to model n_i? What is the best statistical distribution to model x_i?

b) (15 pts) Write a short script in R to calculate the 80th percentile confidence interval of the probability of using Species 1 using a non-parametric bootstrap. (Partial credit for a purely verbal description of the right steps.)

c) (20 pts) Describe how you would use a permutation test to see whether *C. balzani* were favoring one type of tree over another. (You must state the test statistic you would use, and how you would generate the distribution of that test statistic under the null hypothesis.)

14.(25 pts)

Suppose that the prevalence (the overall rate of occurrence) of a disease in the general population is 1 in 1,000 people (0.1%). Suppose that a test for the disease has been developed which has the following characteristics. The sensitivity of the test (the probability that the test is positive when it is applied to a person with the disease) is 99%. The specificity of the test (the probability that the test is negative if the person does not have the disease) is 80%.

a) If a person from the population with no other known risk factors tests positive for the disease, what is the chance that the person actually has the disease? This is called the "positive predictive value" of the test. (You can leave your expression unevaluated if you wish.)

b) What is the Type I error rate of this test?

c) What is the Type II error rate of this test?

d) In this scenario, should we be more concerned with Type I or Type II error and why?