(235 pts total)

Section 1 – Short answer

1. (24 pts) Choose the most appropriate distribution for the types of data shown below and justify your decision.

a) The mass of carbon in the above-ground biomass of a square meter plot.

b) The number of seals on a haul-out beach in the Gulf of Alaska.

c) The presence or absence of an invasive species in forest patches.

d) The probability that a tagged elk will be resighted.

e) The number of tagged elk that are resighted. (Note: Consider carefully the difference between d and e. These are different questions.)

f) The number of individual animals in different wolf packs.

g) The number of diseased individuals in a sample of 100.

h) The age or stage classification (e.g., pup, juvenile, adult, etc.) of individual animals within a wolf pack.

2. (15 pts)
Jared Diamond studied the distribution of fruit pigeons *Ptilinopus rivoli* and *P. solomonensis* on 32 islands in the Bismark archipelago northeast of New Guinea (see Table). Define the event *R* as an island being occupied by *P. rivoli*, and the event *S* as an island being occupied by *P. solomonensis*. The complementary events, that an island is not occupied by *P. rivoli* and not occupied by *P. solomonensis*, respectively, are denoted R^C and S^C.

Table 1: Data on the distribution of species of fruit pigeons on islands

Status	Number of Islands
***P. rivoli* present, *P. solomonensis* absent**	9
***P. rivoli* absent, *P. solomonensis* present**	18
Both present	2
Both absent	3
Total	32

Use the information in Table 2 to answer the following questions:
a) What is the probability $P(R)$?

b) What is the probability $P(S)$?

c) What is the $P(R \cup S)$?

d) What is the $P(R, S)$?

e) What is the $P(R|S)$?

3. (10 pts) What is the test statistic used for the Kolmogorov–Smirnov test?

4. (10 pts) Name one scenario in which you might use a parametric bootstrap rather than a non-parametric bootstrap.

5. (25 pts)

a) (5 pts) Speaking generally, what is a test statistic? (What does it do? What is it for?)

b) (10 pts) For the one sample t-test specifically, what is the test statistic and what is its distribution under the null hypothesis?

c) (10 pts) The one sample t-test in part b is just one possible way to test whether a population mean is a certain value ($\mu = c$). Give an example of an alternative test statistic that would have lower power to test this null hypothesis. Full credit requires that you define "low power" in your answer.

Section 2 – Long answer

6. (54 pts) Consider a scenario in which you are studying the number of seals hauled out on a beach, and wish to model the number of seals using a Poisson distribution.

$$X \sim Pois(\lambda)$$

In this case, the parameter λ can be interpreted as the "intensity" of seals hauling out onto the beach. It is reasonable to suggest that this intensity may not be fixed for all beaches, but may vary according to some statistical distribution. If we assume that λ itself is drawn from a Gamma distribution

$$\lambda \sim Gamma(\alpha, \beta)$$

then we have a Gamma-Poisson mixture distribution. It turns out that this distribution is equivalent to a Negative Binomial distribution.

$$X \sim NegativeBinomial(r, p)$$

where the probability density function is given by

$$f(x|r,p) = \frac{(x+r-1)!}{x!\,(r-1)!} p^x (1-p)^r$$

a) (20 pts) Find the maximum likelihood estimate for p.

b) (18 pts total) Describe two ways in which you could derive 95^{th} percentile confidence intervals for p.

c) (10 pts) Describe in words the interpretation of a 95th percentile confidence interval on p.

d) (6 pts) The classic interpretation of the Negative Binomial (when r is an integer) is the time you need to wait when flipping coins (Bernoulli trials; probability of success=p) to get x successes and r failures. In 2-3 sentences, explain why this makes sense given the probability density function provided.

7. (30 pts) Ecologists are often interested in whether the location of organisms or events (e.g., plants in a field, flu cases in a city) are randomly distributed (H_0) or clustered (H_A). Ripley's K function is one way of assessing whether points are randomly distributed. Ripley's K function is the average number of points within a radius r of a given point, and plots of Ripley's K function include the value of the function at several radii, each of which is compared to the distribution of Ripley's K function under the null hypothesis of complete spatial randomness.

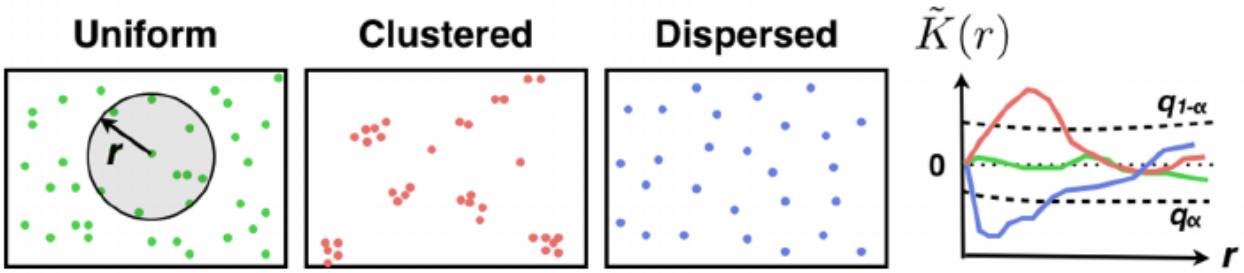

Figure 1: Ripley's K function for three different scenarios. Note that it is not important for this question which scenario is associated with which curve on the right hand plot (details may be lost in printing and photocopying the exam).

a) (10 pts) When Ripley's K function falls outside of the confidence intervals for the null hypothesis at a given radius r, this is often used to reject the null hypothesis of spatial randomness at that spatial scale. Why must we be cautious about interpreting deviations from the null hypothesis in this case? (Hint: How does the number of spatial lags [the values of r at which Ripley's K-function is calculated] play a role?)

b) (20 pts) Describe two different methods that might be used to ensure the correct Type I error rate for using Ripley's $K(r)$ as evidence against complete spatial randomness.

8. (36 pts)

a) (18 pts) Let $X \sim Unif(0,1)$. (This is the Uniform distribution, bounded by 0 and 1.) Find the Var[X].

b) (18 pts) Let $X \sim exp(\lambda)$, the exponential distribution, whose probability density function is given by $\lambda e^{-\lambda x}$. Find the Median[X].

9. (31 points)
a) (9 pts) Complete the sentence: An F-test is commonly used to test....

b) (2 pts) What is the R command for an F test?

c) (20 pts) Write a short R script to generate 1000 draws from the F-distribution using only draws from a normal distribution.