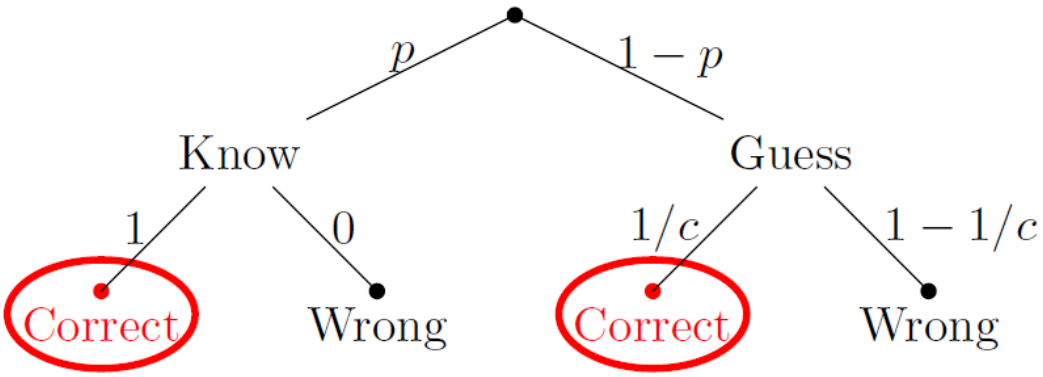
Biometry mid-term exam Spring 2023

(200 pts total)

Section 1 – Short answer

1. (5 pts) If the number of species (X) on each island can be described as $X \sim Pois(17)$, what is the probability of an island containing no species at all?

2. (10 pts) Imagine you are taking an exam. (I know that's tough, but just imagine…) The probability tree describing the probability of getting the correct answer given two scenarios (you know the answer, or you don't and you are guessing) is as follows.



What is the probability that you know the answer conditional on having gotten it correct (i.e. $P(K|C)$)? (Hint: Use Bayes Theorem. Partial credit for writing down Bayes Theorem.)

Biometry mid-term exam Spring 2023

3. (20 pts total) The probability of rejecting the null hypothesis is

$$P(Reject\ H_0) = P(Reject\ H_0|H_0) * P(H_0) + P(Reject\ H_0|H_A) * P(H_A)$$

(2 pts) The left-hand side of this equation is called the

a. joint probability

b. the marginal probability

c. the conditional probability

d. or none of the above?

(3 pts) Rewrite this equation in terms of α and β.

(3 pts) How does α change as the sample size goes to infinity?

(3 pts) How does β change as the sample size goes to infinity?

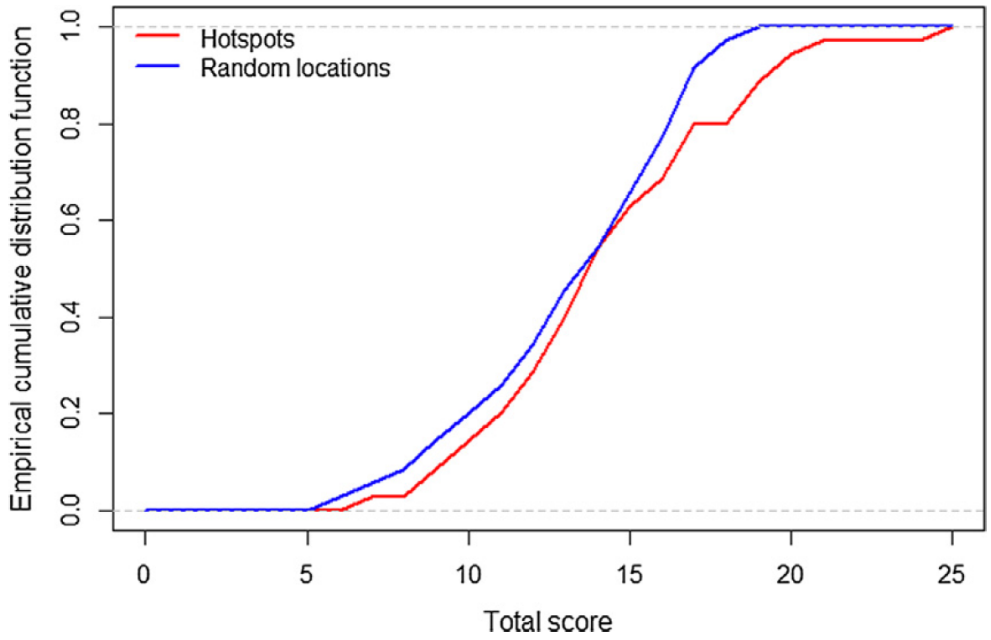
(3 pts) What is the probability that a point null hypothesis about a continuous variable (for example, that $\bar{X} = 1.32$ if $X \sim N(\mu, \sigma^2)$) is true?

Biometry mid-term exam Spring 2023

(3 pts) What does this mean for the probability of rejecting a point null hypothesis as sample size goes to infinity?

(3 pts) What does this mean for the use of null hypothesis significance testing in simulation-based experiments?

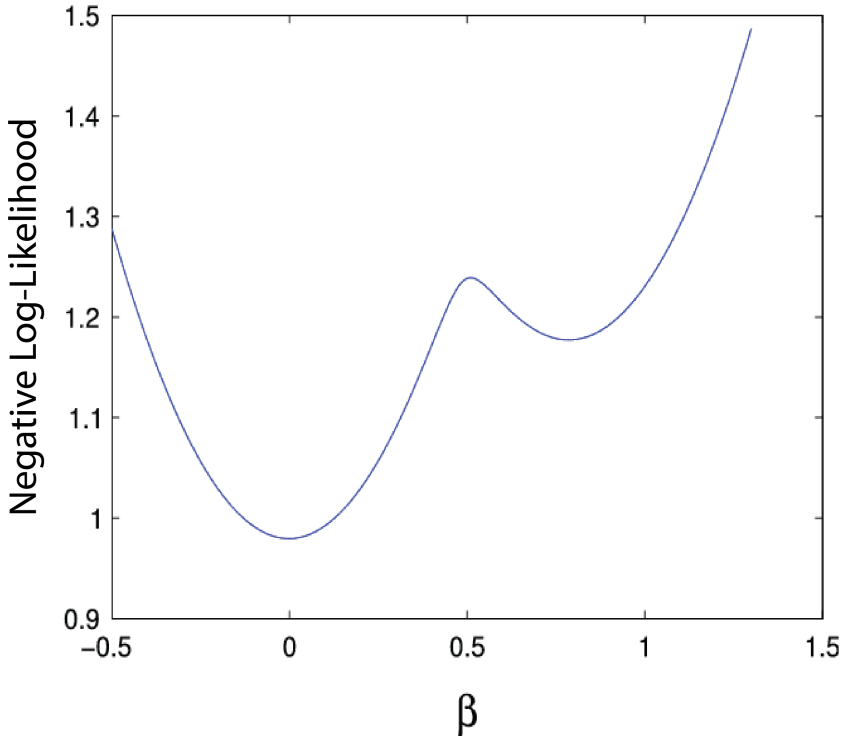
4. (10 pts) Keken et al. (2019) completed a study of what roadside features might promote ungulate-vehicle collisions. For regions considered collision hotspots and control regions randomly sampled from the landscape, they calculated a score that summarizes all the factors that might make an area attractive for ungulates (vegetation cover, etc.). They then used a Kolmogorov-Smirnov test to compare the collision hotspots with the random locations as illustrated below.



a) Based on the information in the figure, what is the approximate test statistic for the K-S test? (2 pts)

b) Another possible approach the authors could have taken would be a t-test. What is the hypothesis that the authors are testing with the K-S test and what is the hypothesis that they would be testing if they had used a t-test instead? (8 pts)

5. (10 pts) Imagine the following curve representing the negative log-likelihood as a function of some parameter β.



(5 pts) What is the MLE for the parameter β?

(5 pts) What complication arises in trying to find the MLE for β ?

Section 2 – Long answer

6. (20 pts) Suppose $X_1, X_2, \ldots, X_n$ are i.i.d. random variables from the standard triangular distribution with probability density function

$$f(x) = \begin{cases} x+1 & when \quad -1 < x < 0 \\ 1-x & when \quad 0 \leq x < 1 \end{cases}$$

A) (5 pts) Find the E[X]. (Hint: It might be helpful to sketch the PDF first.)

B) (15 pts) Find Var[X].

7. (65 pts) Cramer et al. (2008) studied the recovery of 108 agricultural fields that had been under different types of cultivation in the past. Each field was recorded as either fully recovered, partially recovered, or unrecovered.

		Agriculture type			
		Tradit.	*Pasture*	*Intensive*	*Row total*
Recovery type	*Full*	11	2	1	**14**
	Partial	1	20	13	**34**
	None	0	12	48	**60**
	Column total	**12**	**34**	**62**	**108**

a) (3 pts) What is the probability of being fully recovered if the original type of agricultural activity was Intensive?

b) (10 pts) Write the expression for the 95th percentile confidence interval for the probability in part A. While you do not need to work out the math (b/c you don't have a calculator) you should provide an expression in terms of the actual data in the table above. Are there any concerns with this confidence interval?

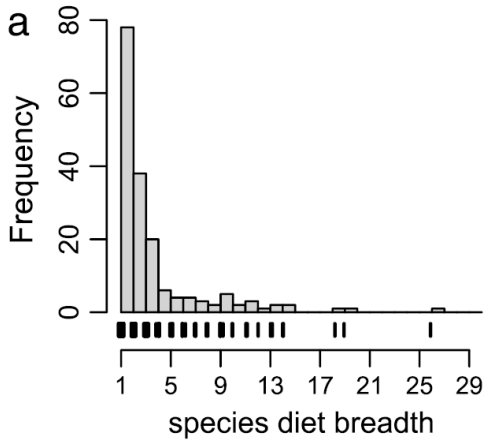
c) (15 pts) Describe in detail how you would use bootstrap sampling to calculate the 95^{th} percentile confidence interval from part b. (You can use words or R code.) Does this approach suffer from the same concern identified in b?

d) (15 pts) Describe a non-parametric permutation-based statistical test that you could use to test the null hypothesis that Intensive agriculture has the same probability of full recovery as Traditional or Pasture. (Here we lump together Traditional and Pasture.). You must state the test statistic and describe in detail how you would generate its distribution under the null hypothesis.

e) (7 pts) Assuming that we did not want to lump together Traditional and Pasture and we wanted to consider the full suite of agriculture types to ask whether there were any differences in the probability of full recovery among the three agricultural use types. In class we introduced two different approaches to controlling family-wise error rates. What per-comparison error rate should be used for each of those two methods to keep the family-wise error rate at 0.05? Which of these approaches is associated with a larger Type II error rate?

f) (15 pts) Finally, let's say we want to consider all three agricultural use types separately and we want to test the null hypothesis that recovery (full vs. partial vs. none) is independent of use type. Describe in detail a permutation-based test for testing this hypothesis. (You need to identify the test statistic and the means by which you would generate the distribution under the null hypothesis.)

8. (60 pts total) The Pareto distribution is a long-tailed distribution that has many uses in ecology, such as for modelling diet breadth among herbivores (i.e. the number of host plants eaten by a single species of herbivore)



From Kozubowski et al. (2015)

Let's assume we have data on diet breadth and assume the following Pareto distribution probability density function

$$P(X = x_i|\alpha) = \frac{\alpha x_i^{\alpha}}{S_o^{\alpha+1}}\ , x_i > S_o$$

where x_i is the number of plant species consumed in sample i, S_o is a constant, and α is a parameter of the distribution.

a) (5 pts) What is the support of this distribution?

b) (15 pts) Find the likelihood function for this model, given observations $X_1 = x_1, X_2 = x_2, X_3 = x_3, \ldots, X_n = x_n$.

c) (20 pts) Find the maximum likelihood estimate for $\hat{\alpha}$, given observations $X_1 = x_1, X_2 = x_2, X_3 = x_3, \ldots, X_n = x_n$.

c) (15 pts) Describe one method of finding 95th percentile confidence intervals for $\hat{\alpha}$.

d) (5 pts) What is the correct interpretation of the interval estimated in part c? (I'm looking for just one sentence.)

e) (5 pts) Given the context of the question and the data provided, why would the Poisson distribution be an inappropriate distribution?