(235 pts total)

Section 1 – Short answer

1. (20 pts) Choose the most appropriate distribution for the types of data shown below and justify your decision.

a)

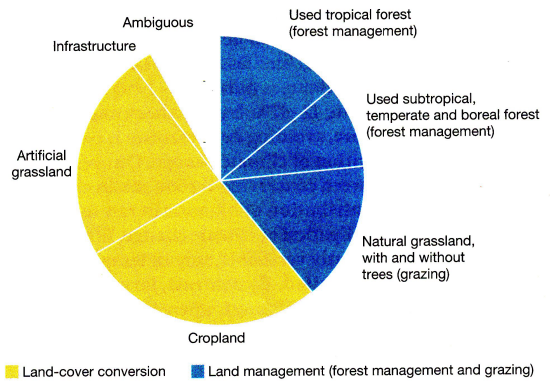


Figure a: Relevant contribution of land-cover conversion and land management to the difference between potential and actual biomass stocks. (Erb et al. 2018)

What would be the best distribution for the data illustrated in the histogram?

b)

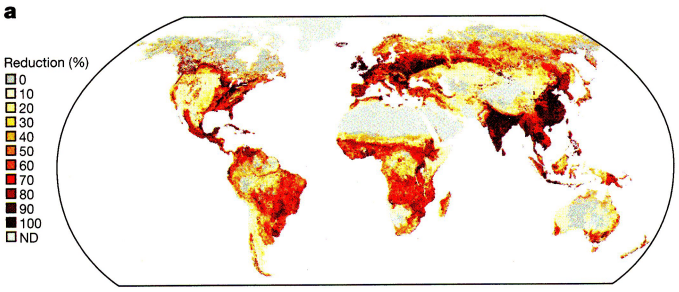


Figure b: Spatial pattern of land-use induced biomass stock differences (expressed as a percentage of potential biomass stocks). (Erb et al. 2018)

What would be the best distribution for the data illustrated in the map?

c

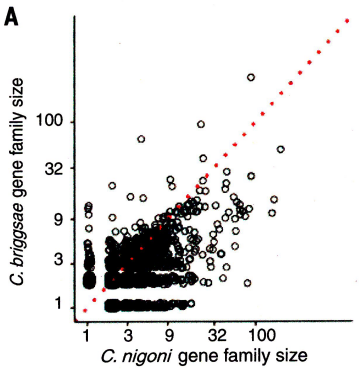


Figure c: Scatterplot of sizes of gene families. Dotted lines indicates equal family sizes. (Yin et al. 2018)

What would be the best distribution for gene family size? (Note that this distribution would apply equally to the data along the x-axis or the y-axis)

d)

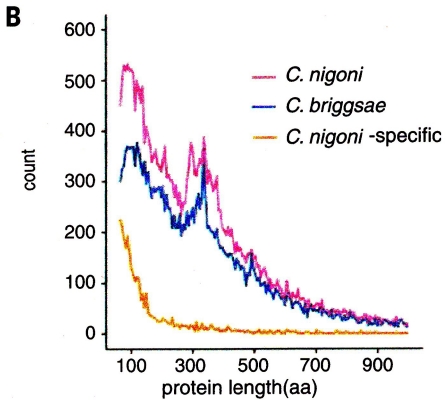


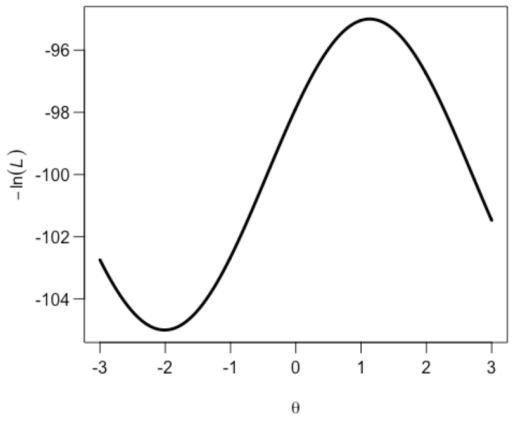
Figure d: Length distributions of *C. nigoni* and *C. briggsae* proteins and of *C. nigoni* proteins that lack *C. briggsae* homologs. (Yin et al. 2018)

What would be the best distribution for the count data (y-axis) representing the frequency of occurrences of proteins of different length?

What would be the best distribution for protein length itself (i.e., the x-axis)?

2. (15 pts)
Two researchers (A and B) are studying the distribution of penguins in an archipelago of 100 islands in the Antarctic. There are two closely related species *Pygoscelis papua* and *Pygoscelis adeliae*, and the researchers want to figure out which species has a competitive advantage on each island (we assume that co-existance is not possible and only one penguin species can persist on each island). The researchers have built a simulation model to model the dynamics and find that on each and every island, *P. papua* has a 51% probability of being competitively dominant and excluding the other species. Researcher A claims that that *P. papua* is predicted (based on the model) to be present (and *P. adeliae* absent) on all 100 islands, while Researcher B says that *P. papua* is expected to be present on only about half of the islands (51% of them). Which one is correct and why?

3. (10 pts) Based on the figure below, what is the (approximate) MLE for the parameter θ.



4. (10 pts) In words AND in equations, describe the relationship between the “joint distribution” (for simplicity, let’s assume two variables) and the two marginal distributions.

5. (10 pts) Define “statistical power”.

Section 2 – Long answer

5. (50 pts total) Suppose we model independent observations $Y_1, Y_2, Y_3, \ldots, Y_n$, drawn from the following probability density function:

$$P(Y = y) = \begin{cases} 0 & if\ 0 < y < \theta \\ c(\theta)/y^p & if\ y \geq \theta \end{cases}$$

where p is a constant greater than one, and θ is an unknown positive real parameter.

a) (15 pts) Find the function $c(\theta)$.

b) (15 pts) Find the likelihood function for this model, given observations $Y_1 = y_1, Y_2 = y_2, Y_3 = y_3, \ldots, Y_n = y_n$.

c) (20 pts) Find the maximum likelihood estimate for $\hat{p}$, given observations $Y_1 = y_1, Y_2 = y_2, Y_3 = y_3, \ldots, Y_n = y_n$.

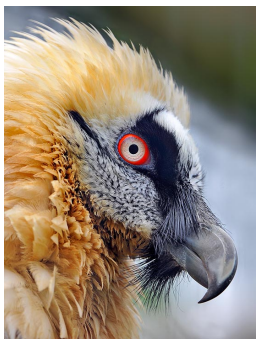
5. (40 pts) A researcher is studying Chytridiomycosis (an infection disease caused by the fungus *Batrachochytrium dendrobatidis*) in Costa Rican variable harlequin toads, and surveys a series of locations for their infection status. The researcher returns from sampling and, following some analysis, reports that the number of sites with Chytridiomycosis infection (Y) can be modelled as

$$Y \sim N(\hat{\mu} = 100, \widehat{\sigma^2} = 25)$$

a) (20 pts) A colleague scratches her head and says: "You should have used a Binomial distribution $Binom(n, p)$ instead of a Normal distribution."

Using the information that you have, estimate n and p.

b) (20 pts) Another colleague scratches his head and says: "Your sample size is too small, because we need the 95th percentile confidence intervals on $\hat{p}$ to be no bigger than 0.01". How many sites should the researcher sample in order to get 95th percentile confidence intervals that are no wider than 0.01.



6. (65 points) Bearded Vulture (*Gypaetus barbatus*) is one of the most charismatic species of birds of prey; as Doug Futuyma says "if you are going to have a favorite vulture, this is it!". Unfortunately, they nest in secluded cliffs; so their nests are hard to find, and even harder to monitor. You have a research team that is responsible with monitoring a small Bearded Vulture population in Beypazari, Turkey. The population consists of only 10 pairs. They are territorial birds and tend to occupy the same sites across years (high site fidelity). However, for various reasons a pair may not breed in any given breeding season and that site will remain unoccupied.

a) (15 points) Let p_i be the probability that the i^{th} site is occupied ($i = (1,2,3 \dots 10)$), and J_i be the number of juveniles produced that season by a pair in site i. Assume that the probability of occupancy and the number of juveniles produced by each occupied nest is independent between nests, and that occupancy and the number of juveniles can be surveyed with no error. Write the joint likelihood for the dataset you have on occupancy status and number of juveniles. (Note: You have to decide on the distribution for J_i.) (Hint: Implicit in the wording is that fact that you have two datasets here. One dataset on occupancy and another on count conditional on occupancy.)

b) (15 points) In order to detect occupancy, your research team goes to previously known nesting sites to monitor the area using telescopes and binoculars. Let π be the probability of detecting vultures for a site that is occupied in a single survey visit.

$$P(detection|occupied) = \pi$$

and that this probability is the same across all sites. If all 10 sites are occupied in a given breeding season, how many visits does your team have to make to each site in order detect (on average) occupancy on 5 sites of those sites? (Hint: Your result will be in terms of π.)

Nearest Neighbor Distance (NND) is one of the indicators used to measure the territoriality of a species. It is the distance of a nest site to the nearest nest site of the same species. Higher distances indicate stronger territoriality. In this regard, it can be considered as measure of intraspecific competition for nest sites. NND can also be calculated between species, as in the distance of a nest site to the nearest nest site of a competitor species. This is a measure of interspecific competition.

The Egyptian Vulture (*Neophron percnopterus*) is a closely related species to the Bearded Vulture. They both nest in cliffs and generally in the same area.

c) (15 points) Using NND as a measure of competition, suggest a parametric method to test whether intraspecific competition has the same strength as interspecific competition for the Bearded Vulture.

d) (15 points) Using NND as a measure of competition, suggest a permutation-based method to test whether intraspecific competition has the same strength as interspecific competition for the Bearded Vulture.

e) (5 points) Why might the permutation-based method be preferred in this case?