Biometry mid-term exam Spring 2015

(250 pts total)

Section 1 – Short answer

1. (25 pts) Consider the following function describing the number of breeding attempts X made by a female deer ($0 < F < 1$):

$$P(X = 1) = F$$

$$P(X = 2) = (1 - F) * F$$

$$P(X = 3) = (1 - F)^2$$

A) (5 pts) Is this function a proper probability mass function? Why or why not?

B) (10 pts) What is the expected number of breeding attempts, i.e. $E[X]$?

C) (10 pts) What is the $Var[X]$?

2. (10 pts) The probability density function in Figure 1 could represent which of the following statistical distributions (circle all possibilities)?

Normal	Standard normal	Log-normal
Poisson	Binomial	Gamma
Chi-squared	F	t
Beta	Multinomial	Negative binomial

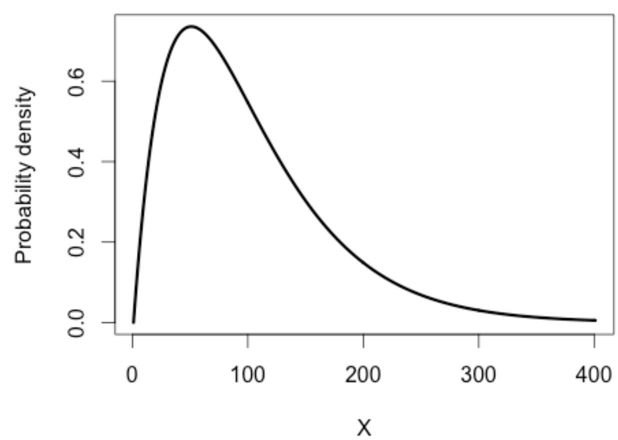

Figure 1

3. (10 pts) If the probability of having green eyes is 10%, the probability of having brown hair is 75%, and the probability of having both green eyes and brown hair is 9%, what is the probability of having brown hair given that you have green eyes?

4. (15 pts) Complete the following equations:

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

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

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

5. (10 pts)

A sample size of $n = 100$ yields a sample mean of $\bar{X} = 16$ and a sample variance of $s^2 = 9$. Compute the 95th percentile confidence interval for the population mean μ.

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

7. A)(8 pts) How would you test whether two datasets were drawn from populations with the same population variance? (Hint: You need to state both the test statistic and its distribution under the null hypothesis.)

B) (2 pts) What is the R function to perform this test?

8. Describe Karl Popper's hypothetico-deductive method (10 pts).

Section 2 – Long answer

9. (75 pts) The Pareto II distribution (also called the Lomax distribution) is a two-parameter continuous probability distribution for positive-valued random variables, which has been used to describe the probability distribution of zooplankton body size.

The Pareto II distribution PDF is given by

$$f(z|\lambda,\alpha) = \frac{\alpha}{\lambda}\left(1+\frac{z}{\lambda}\right)^{-(\alpha+1)}, z \geq 0, \lambda > 0, \alpha > 0$$

A) (10 pts) Find the cumulative density function (a.k.a. the CDF).

B) (20 pts) Assuming you have data on zooplankton body size $z = \{z_1, z_2, \dots, z_n\}$, find the maximum likelihood estimator for the parameter α (assume $\hat{\lambda}$ has already been calculated).

C) (15 pts) Describe in words how would you find the joint confidence intervals for α and λ using maximum likelihood?

D) (15 pts) Describe in words how would you find the confidence intervals for or α and λ using bootstrap?

E) (15 pts) Independent of the method used to construct them, what is the correct interpretation for the 95^{th} percentile confidence intervals for α and λ?

10. (80 pts) (This question is inspired by a recent paper by Stuart Pimm and colleagues published in the *Journal of Biogeography*.) Pimm and colleagues analyzed a dataset on species occupancy of a suite of islands that had been created by the inundation of an artificial reservoir in western Zhejiang Province, China. Pimm and colleagues visited these islands every year for six years; during each visit all the islands in the study were surveyed for birds.

In this "toy" example, let's assume there were three islands and four species surveyed, and the data looked like (P=present, A=absent):

Island 1:

Species	Year 1	Year 2	Year 3	Year 4	Year 5	Year 6
1	P	P	P	A	A	P
2	A	A	A	P	P	P
3	P	P	A	A	A	P
4	A	P	A	P	A	A

Island 2:

Species	Year 1	Year 2	Year 3	Year 4	Year 5	Year 6
1	P	P	P	P	P	P
2	A	P	P	P	P	P
3	P	A	A	A	A	P
4	A	A	A	P	A	A

Island 3:

Species	Year 1	Year 2	Year 3	Year 4	Year 5	Year 6
1	P	A	P	A	A	P
2	A	A	A	A	A	P
3	P	P	A	P	A	P
4	P	P	P	P	P	P

For each island, the Area and Isolation (distance to mainland) were measured: $A_1 = Area\ of\ Island\ 1$

$I_1 = Isolation\ of\ Island\ 1$, $A_2 = Area\ of\ Island\ 2$, etc.

The conditional probabilities representing the transitions are given as follows:

$$P(P_{t+1}|A_t) = probability\ of\ colonization = \frac{1}{1+e^{-(\alpha_0+\alpha_1 A_i+\alpha_2 I_i)}} = \lambda$$

$$P(A_{t+1}|P_t) = probability\ of\ extinction = \frac{1}{1+e^{-(\beta_0+\beta_1 A_i+\beta_2 I_i)}} = \delta$$

$$P(A_{t+1}|A_t) = probability\ of\ remaining\ unoccupied = 1-\lambda$$

$$P(P_{t+1}|P_t) = probability\ of\ remaining\ occupied = \ 1 - \delta$$

(This model has 8 parameters to be fit: α_0, α_1, α_2, β_0, β_1, β_2, λ, δ.

A) (15 pts) Assuming independence among all species, and among all islands, write the likelihood function for the data provided. (Since you have the data provided, you should be able to write the likelihood in terms of λ and δ and numbers [i.e. no additional unknowns].)

B) (10 pts) In the original manuscript's description of the analysis (which was actually based on a study of 37 islands), the authors debate whether they should consider the 'pool' of species to include only those 71 species that were actually found in the survey period (on any island, in any year) or whether they should consider all 93 species that are resident of the mainland community and thus might have been found. Does the likelihood increase or decrease with the addition of these 22 additional species? Will this affect the maximum likelihood estimates for the 8 model parameters? Why or why not?

C) (20 pts) Let's say the authors had been interested in testing the following null hypothesis:

$$H_o\text{: } There\ is\ no\ correlation\ (in\ occupancy\ status) between\ species$$

In other words, the null hypothesis is that (considering a single island for the moment) the presence of species 1 in year 1 has no effect on the probability of species 2 in year 1.

Describe one parametric method that could be used to test this null hypothesis. Full credit would require a verbal description/explanation of the method, the test statistic, and its distribution under the null hypothesis.

D) (20 pts) Describe one non-parametric method that could be used to test this null hypothesis. (Remember to state the test statistic as well as the procedure for calculating its distribution under the null hypothesis.)

E) (10 pts) Explain why the authors might be concerned about an inflated Type I error rate in testing this null hypothesis. Describe one method that might be used to correct for this problem.