BIOS 511: STATISTICAL INFERENCE I
Homework Assignment # 1

Write down all the equations you use. Scale all the plots of evidence so that the highest value is 1.0

1. We desire to learn about the abundance, N, of a certain species of fish living in a lake. A *capture-recapture* experiment is conducted, as follows. A certain number of fish, M, are caught in a net and, after being tagged, are returned to the lake. Some time later, a second catch takes place; denote by n the number of fish captured, of which a certain number, X say, turn out to be tagged. Assume that the probability of being captured is the same for each fish. The experiment is conducted with $M = 20$, $n = 17$, and we observe $X = 13$. Plot the evidence about the abundance of fish. [Hint: assume X has a hypergeometric distribution.]

Table 1: Number of fish caught in each catch

		First catch? yes	First catch? no	
Second catch?	yes	$X = 13$		$n = 17$
	no			$N - 17$
		$M = 20$	$N - 20$	N

2. An experiment was conducted to assess the relative abundances of proteins in the brain tissue of (deceased) Alzheimer's patients. The proteins in the complex mixture were digested into peptides, these peptides were ionized, and the spectral counts of the ionized peptides were recorded using mass spectrometry. A specific protein of interest is known to have 6 peptides, each with a known ionization efficiency. We observe the following spectral counts for these peptides:

Table 2: Spectral counts information.

peptide	ionization efficiency (c)	spectral count (x)
1	0.9	16
2	0.8	10
3	0.6	11
4	0.4	0
5	0.3	0
6	0.2	3

We adopt the spectral count model $X_i \sim \text{Poisson}(c_i\lambda), i = 1, \ldots, 6$, where the X's are independent with means $E(X_i) = c_i\lambda, \lambda > 0$. Plot the evidence about the protein's relative abundance, λ. Compare to the plot of the evidence if the two observations $X_4 = 0, X_5 = 0$, were ignored. Discuss any differences between the two plots.

3. In a study of the prognosis of cancer patients, a random sample of seven patients were observed to live disease-free for the following number of years after chemotherapy:

$$3.5, 8.5, 2.5, 10.5, 1.5, 4.5, 10.5$$

Assume that disease-free survival time follows an exponential distribution with unknown mean $\mu > 0$.

(a) Plot the evidence about μ.

(b) Upon further inquiry, it was discovered that the disease-free survival times were not measured precisely, but rather the midpoints of the annual patient visits were used. For example, the observation recorded as $X_1 = 3.5$ years should more properly be regarded as $X_1 \in (3.0, 4.0)$. Replot the evidence about μ to take into account the interval-censoring of the data.

(c) As a continuation of part (b) above, it was discovered furthermore that the study terminated after 10 years, so the two observations (X_4, X_7) recorded as 10.5 years were really the events $X_4 > 10, X_7 > 10$. Replot the evidence about μ to take into account the right-censoring of X_4 and X_7 and the interval-censoring of the other five observations. Discuss any differences between the three plots in (a), (b), and (c).

4. $X_1, \ldots, X_5$ are i.i.d. random variables and it is known that the means and variances are 4 and 8, respectively. Consider the hypothesis that the random variables are normally distributed versus the alternative hypothesis that they follow a gamma distribution. If we have the following realizations

$$1.2, 3.6, 8.0, 3.4, 7.7$$

which of the two hypotheses is better supported by the data? What is the strength of the evidence? How would our results be affected if the first observation was -1.2 instead of 1.2?

Bonus Question

5. Observation $X = x$ is used to assess the evidence in favor of hypothesis $X \sim f$ versus the hypothesis $X \sim g$, where the probability density functions f and g are not equivalent (i.e., we exclude the possibility that f and g always agree). The strength of the evidence is given by the likelihood ratio, $LR = f(x)/g(x)$. Show that the probability of obtaining strong misleading evidence is small, i.e., when g is the true density function show that we have inequality

$$\Pr(LR \geq k) \leq \frac{1}{k},$$

for any constant $k > 0$.

6. Consider a random sample of size n from a $N(\mu, \sigma^2)$ distribution, where we know that $\sigma^2 = 1$. We desire to assess the evidence in support of the hypothesis $\mu = 1$ versus the hypothesis $\mu = 0$. Assuming that it is true that $\mu = 1$, find the smallest sample size n such that the likelihood ratio exceeds 8 with probability 0.9.