Week 3 Problem Set: Probability Distributions

**(30 pts+4 bonus pts possible)**

PART I – Warm up

Question 1 (0.5 pt each)

Let X be a discrete random variable whose pdf is described in the table given here:

|  |  |  |  |
| --- | --- | --- | --- |
| x | -1 | 0 | 1 |
| f(x) | 1/8 | 6/8 | 1/8 |

Find the following:

1. E[X]
2. P(X=0)
3. P(X≤1)
4. F(1)
5. F-1(7/8) [NB: I am using Aho’s notation of F-1 for the quantile function.]

Question 2 (0.5 pt each)

X and Y are independent random variables. Write the following expressions in terms of E[X] and E[Y]

1. E[(X+35)/10]
2. E[X-14Y+E[Y]+7]-E[X+Y+5]

Question 3 (2.5 pt each)

Use R to convince yourself of the Central Limit Theorem using draws from any distribution that is not the Normal distribution. Briefly (2-3 sentences) explain your process and provide your code as well.

PART II – PDFs

Question 4 (1.5 pts each)

Assume that leaf biomass (in grams) from the plant *Salix arctica* can be described as having the following probability density function:

1. Another way of saying that this pdf is restricted to is to say that there is no *support* for . Why would this be the case here?
2. Prove (using the two requirements for a valid pdf) that this is a valid pdf.
3. Manually (i.e., using calculus) calculate the probability that a leaf has biomass between 0 g and 3 g. In other words, find .
4. Use the *integrate* function in R to confirm this result numerically.

Question 5 (2 pts each)

Bliss and R.A. Fisher (1953) examined female European red mite counts (*Panonychus ulmi*) on McIntosh apple trees [*Malus domestica* [McIntosh]). Counts of the mites on 150 leaves are shown here:

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Mites per leaf | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
| Leaves observed | 70 | 38 | 17 | 10 | 9 | 3 | 2 | 1 | 0 |

1. What is the expected value E[X] of mites per leaf in this sample (show your work, either an R script or a calculation)?
2. Assume for the moment that Bliss and Fisher used a Poisson distribution to describe the number of mites per leaf X (). What would be the most reasonable value for the Poisson parameter and why?
3. Make a barplot (the R function histogram should work, but as these are discrete variables, we would normally refer to this as a bar plot) of Bliss and Fisher’s data. Is the expected value also the most common value? Why or why not?

1. What is the standard deviation of the number of mites per leaf? What is the standard error of the mean (SEM) number of mites per leaf? In one sentence, describe the interpretation of the standard error of the mean?
2. To gain better intuition for the meaning of the standard error of the mean, create 1000 new datasets by sampling with replacement from the original dataset. Use those 1000 simulated datasests to calculate the standard error of the mean. Because of sampling error, this will not be exactly the same as what you calculated in part d, but it should be close.

PART III – Applications of distributions

Question 6

We are going to get some practice using these statistical distributions in a biological context, and in the process gain some extra practice writing R code. Read Viswanathan et al. (2008) for the biological context of this problem, but don’t worry too much about the mathematical details (you can skim over Section 4).

This question is designed to get you writing some more sophisticated R scripts, with more emphasis on loops and logic, but there are some statistical and biological goals as well. Before getting to the actual question, I want to emphasize the following “take-home” messages:

1. There are many, many statistical distributions. We have learned only a small subset of all the statistical distributions. Long-tailed distributions (such as the Lévy, Cauchy, and Pareto) have many uses in ecology and evolution.
2. Simulations provide a straightforward way of studying complex stochastic phenomena.
3. Understanding ‘pattern’ goes a long way to understanding the underlying ‘process’. Statistical distributions provide a way to quantify the pattern of your data. In many cases, only certain biological or ecological mechanisms can generate those patterns.

Write a short script to simulate the movement of animals operating according to the movement rules described below. For simplicity, we will assume only one-dimensional motion along a line. Animals can move forward, or backwards, along a line. For each animal, simulate 500 individuals each moving for 100 time steps each. **Note that only Animal #1 and #2 are required for this problem set (8 pts total).** Animals #3 and #4 are BONUS questions (2 pts each).

Animal #1: This animal moves by “Brownian” motion. In each time step, the individual chooses a random direction (50% probability of moving forward or backwards [this is a Binomial process!]) and moves a distance given by (the absolute value of) . (You can do this by combining a Bernoulli for direction with a distance given by the magnitude of the Normal, or just do it in one step by making the movement Normally distributed. The former approach will be more portable to Animals #2-#4.) Cut and paste your script and a histogram of the final location for each of the 500 individuals. Fit a normal distribution to that distribution (hint: use use ‘fitdistr’)– what is and ?

Animal #2: This animal is similar to Animal #1 but it preferentially moves to the right. In each time step, there is a 60% chance the animal will move to the right, the distance moved is still given by . Cut and paste your script and a histogram of the final location for each of the 500 individuals. Fit a normal distribution to that distribution (hint: use ‘fitdistr’)– what is and now? Note that this is still a random walk, albeit one with drift.

Animal #3: This animal walks with a certain amount of inertia. Distances are still given by , but the probability of moving to the left or right depends on the previous move. The first move is random (50% chance of both left and right) but all subsequent moves have a 60% chance of being in the same direction as the previous move. Cut and paste your script and a histogram of the final location for each of the 500 individuals. Is the resulting distribution still normally distributed? If not, how does the final distribution of individuals differ from a normal? Why?

Animal #4: This animal has no preference for direction, so has a 50/50 chance of moving left or right at each time step. However, distances travelled are now NOT normally distributed, but are governed by the Cauchy distribution with location=0 and scale=2 [Hint: Use ‘rcauchy’]. The Cauchy distribution is ‘pathological’ in that is has no mean and no variance; in fact, none of the moments of the Cauchy are defined. The Cauchy is one of the so-called ‘heavy-tailed’ distributions, in that the probability of very large values declines more slowly than the exponential distribution. (Translation: Very large numbers are more likely with heavy-tailed distributions.) In this context, the Cauchy distribution will yield a lot of small moves and some very large moves. Cut and paste your script and a histogram of the final location for each of the 500 individuals. Is the resulting distribution still normally distributed? If not, how does the final distribution of individuals differ from a normal? Why?