Homework 4 – Distributions for surviving copepods and abundant tang

The file 'dam_acartia_survival_subset.csv' contains some data on the survival of the copepod *Acartia tonsa*, from a study conducted by Hans Dam and colleagues. In experimental cultures the survival of copepods was tracked from the nauplius 1 stage to the copepodid 6 stage. The data used here is a subset: only the copepods reared at temperature 18ºC, and with survival measured at day 14. The column 'nx' is the number of surviving copepods, out of 25, on day 14. We are going to investigate the distribution of this subset of the data.

(1 – 3 points) First plot a histrogram of the data, using the discrete.histogram() function from the package 'arm'. This function is better than standard histogram functions for our purposes, because the standard binning algorithms makes it hard to see what is happening at the ends of the distribution.

(2 – 4 points) Now calculate the mean and the variance of the number of survivors. This kind of data could potentially be modelled as a binomial distribution. Based on the formula for the binomial distribution (it is in the lecture notes, or on Wikipedia), what *should* the variance be, *if* the data were actually binomially distributed, given the mean number of survivors we observe?

(3 – 4 points) Use the function rbinom() to simulate random draws from a binomial distribution that has the same probability of survival as the observed copepod data. Draw the same number of values as are present in the observed data. Calculate the mean and the variance of the number of survivors. Plot a histogram of your randomly drawn values. Now repeat this four more times – you will have a total five histograms as well as five means and variances. Describe how your simulated binomial data differs from the observed data. Can you imagine reasons why the two distributions might be different?

(4 – 4 points) Now load the package VGAM, which has functions for the *beta-binomial* distribution. This distribution is similar to the binomial, in that it models success/failure of a 'trial', but it allows the probability of success to vary across trials. This extra variability is controlled by the parameter 'rho'. Use trial and error, with the function rbetabinom, to find a value of rho that creates a distribution that looks similar to the observed copepod

data. A value of 0 for rho produces a binomial distribution, while increasing values of rho between 0 and 1 lead to more variability in the data. Display the results of your trial and error simulations, and report what value of rho you think best corresponds to the observed data.

The other common kind of discrete data that we will discuss in this class is count data. The file 'yellowtang.csv' contains counts of the reef fish yellow tang (lau'ipala) from sites around Hawai'i Island collected by NOAA. The column with the counts is called 'count'. We're going to repeat everything we did the copepod data, but instead of using the binomial and beta-binomial distributions, we will use the poisson and negative binomial distributions for the count data.

In brief:

(5 – 3 points) Plot a histogram of the yellow tang counts.

(6 – 4 points) Calculate the mean and variance of the counts. Also calculate what the variance *would* be if the counts were poisson-distributed. How does it differ from the observed data? Why might that be?

(7 – 4 points) Simulate five sets of random draws from the poisson distribution with the appropriate mean and variance (using the function rpois()). Plot histograms of the results, and report the means and variances of the simulated data. Verbally compare these distributions to the observed distribution.

(8 – 4 points) Perform trial and error to find good-fitting parameter values from the negative binomial distribution, which can model counts with more variability than the poisson distribution. Use the function rnegbin() from the package 'MASS'. The parameter 'theta' is the one you will want to manipulate to change the shape of the distribution (keep the value of the mean, mu, the same as the observed data). Hint: smaller values of theta lead to larger variances.