Analysis of Environmental Data

Probability Distributions

(Written by Kevin McGarigal, but borrowed heavily from Ben Bolker, Ecological Models and Data (2007) and Michael Crawley, The R Book (2007))

The purpose of this lab exercise is to familiarize you with a range of common probability distributions in R. Recall that deterministic functions are used to define the ecological process under consideration, while probability distributions are used to define the stochastic component(s) of the model (i.e., the error distributions). Together, these two components form the basis for most parametric statistical models. It is not practical to provide a comprehensive inventory of existing probability distributions, since there are many. Rather, the intent of this lab is to introduce you to probability distributions by way of examples and give you some of the R skills needed to examine other distributions. See Bolker (2007) and Crawley (2007) for excellent descriptions of the common probability distributions. Here is outline of what is included in this lab exercise:

1. Set up your R work session

Open R and set the current working directory to your local workspace, for example:

```
setwd('c:/work/stats/ecodata/lab/distributions/')
```

Load the biostats library (which is not a formal library) by typing, substituting the appropriate path:

```
source('.../biostats.R')
```

2. What is a probability distribution?

Recall that the deterministic part of the statistical model describes the expected pattern in the absence of any kind of randomness or measurement error. However, to formally estimate the parameters of the model, you need to know not just the expected pattern but also something about the variation about the expected pattern. Typically, you describe the stochastic model by specifying a reasonable probability distribution for the variation about the expected pattern.

The variation about the expected pattern is often termed "noise". Noise affects ecological data in two different ways – as measurement error and as process noise. This distinction can be very important, and separating these two sources of error is the basis for one whole class of multi-level or hierarchical models. Briefly, measurement error is the variability or "noise" in our measurements, which makes it hard to estimate parameters and make inferences about ecological systems. Measurement error leads to large confidence intervals and low statistical power. Even if we can eliminate measurement error, process noise or process error (often so-called even though it isn't technically an error but a real part of the system) still exists. Variability affects any ecological system.

The noise or error about the expected pattern is described by a probability distribution. More formally, a probability distribution is the set of probabilities on a sample space or set of outcomes. Here, we will always be working with sample spaces that are numbers – the number or amount observed in some measurement of an ecological system.

You need the same kinds of skills and intuitions about the characteristics of probability distributions that we developed for mathematical functions in the previous section. In this section we will begin by learning some of the skills needed to work with both discrete and continuous probability distributions and then provide a bestiary of probability distributions commonly used in ecology.

2.1 Discrete distributions

The simplest distributions to understand are *discrete* distributions whose outcomes are a set of integers; most of the discrete distributions we deal with describe counting or sampling processes and have ranges that include some or all of the nonnegative integers.

As an example, one of the most common discrete distributions in ecology and the easiest to understand is the *binomial*. It applies when you have samples with a fixed number of subsamples or "trials" in each one, and each trial can have one of two values (e.g., present/absent, alive/dead), and the probability of "success" (present, alive) is the same in every trial. If you flip a coin 10 times (N=10) and the probability of a head in each coin flip is p=0.7, then the probability of getting 7 heads (k=7) will have a binomial distribution with parameters N=10 and p=0.7. Consider the presence/absence of a focal species across a sample of 10 plots. If we survey all 10 plots (N=10) and the probability of the species being present is p=0.3, then the probability of observing the species present at say 4 plots (k=4) will have a binomial distribution with parameters N=10 and p=0.3. R has a wide range of built-in discrete probability distributions (Table 1).

2.2 Continuous distributions

A probability distribution over a continuous range (such as all real numbers, or the nonnegative real numbers) is called a *continuous* distribution. As an example, the most common continuous distributions in ecology (and all of statistics) and the easiest to understand is the *normal*. It applies when you have samples that can take on any real number and for which there is a central tendency (usually described by the mean). R has a wide range of built-in continuous probability distributions (Table 1).

Table 1. Summary of some common probability distributions (adapted from Bolker 2008).

Distribution	Range	Skew	Description (examples)
Discrete			
Binomial (binom)	0, N	any	Applies when you have samples with a fixed number of subsamples or "trials" in each one, and each trial can have one of two values. Used with binary response data (e.g., present/absent, alive/dead, success/failure, male/female) or proportional response data (e.g., number surviving out of total number).
Poisson (pois)	0, ∞	right	The distribution of the number of individuals, arrivals, counts, etc., in a given time/space unit of counting effort if each event is independent of all the others and variance≈mean. Used with count data (e.g., seeds per quadrat, number of young produced).
Negative binomial (nbinom)	0, ∞	right	Generally used when the Poisson is appropriate, but when the variance is larger than the mean (i.e., it is overdispersed). Used same as Poisson.
Geometric (geom)	0, ∞	right	The distribution of the number of trials (with a constant probability of failure) until you get a single failure: it's a species case of the negative binomial, with size (overdispersion parameter) = 1. Used, for example, for the number of successful/survived breeding seasons for a seasonally reproducing organism.
Beta-binomial (betabinom [emdbook]	0, 8	any	Compound distribution similar to binomial but allowing for heterogeneity in per-trial probability. Used with binary response data same as binomial.
Continuous			
Uniform (unif)	0,1	none	Constant probability density rarely used in ecology (e.g., cover proportion).
Normal (norm)	-∞, ∞	none	The familiar bell-shaped distribution and the basis for most classical statistics. Used with continuous response data (e.g., mass)
Gamma (gamma)	>0, ∞	right	Distribution of waiting times until a certain number of events take place. Used with continuous response data when negative values are illogical or when the distribution is skewed to

			the right (e.g., survival time, distance to nearest edge).
Beta (beta)	0, 1	any	Distribution of the probability of success in a binomial trial with *a*-1 observed successes and *b*-1 observed failures. Closely related to the binomial but with a finite range, 0 to 1. Used with probabilities or proportions (i.e., continuous data expressed on a proportional scale, not to be confused with discrete proportional data involving multiple trials, each with a binary outcome) or any time a continuous distribution has a finite range (e.g., cover proportion).
Exponential (exp)	0, ∞	right	Distribution of waiting times for a single event to happen, given that there is a constant probability per unit time that it will happen. Continuous counterpart of the goemetric. Used with anything that decreases exponentially with time or distance (e.g., survival time, dispersal distance).
Lognormal (lnorm)	0, ∞	right	Often used like the gamma with continuous, positive distributions with long tails or variance much greater than the mean (e.g., size, mass (exponential growth)).
Chi-square (chisq)	0, ∞	right	Distribution of the sum of squares of *n* (degrees of freedom) normals each with variance one. Famous for its use in contingency table analysis (cross-classified categorical data) and the analysis of count data.
Fisher's *F* (f)	0, ∞	right	Distribution of the ratio of the ratio of the mean squares (variances) of two independent standard normals, and hence of the ratio of two independent chi-squared variates each divided by its degrees of freedom. Famous for its use in analysis of variance (ANOVA) tables, involving the ratio of treatment and error variance.
Student's *t* (t)	-∞, ∞	none	Overdispersed counterpart for the normal distribution which results from mixing the normal sampling distribution with an inverse gamma distribution for the variance. The *t* distribution has fatter tails than the normal. Famous for its use in testing the difference in the means of two normally distributed samples.

3. Plotting distributions

R has four useful functions for each of the built-in probability distributions: the probability mass or density (which as a *d* prefix); the cumulative probability (*p*); the quantiles of the distribution (*q*); and random numbers generated from the distribution (*r*). Each letter can be prefixed to the R function name for the existing probability distributions (e.g., for the binomial distribution: dbinom, pbinom, qbinom, and rbinom).

3.1 Probability mass or density

For a discrete distribution, the probability distribution (often referred to as the *probability mass distribution*) represents the probability that the outcome of an experiment or observation (called a random variables) X is equal to a particular value x (f(x) = Prob(X=x)). In the case of the *binomial distribution*, for example, we can use the dbinom() function to calculate the probability of any specified number of successes given the number of trials (*N*) and the probability of success (*p*) for each trial. For example, in the species presence/absence example above, we can compute the probability of the species being present at 4 of the 10 plots, as follows:

```
dbinom(4,size=10,prob=.3)
```

Or we can compute the probability of the species being present at 0 through 10 plots, as follows:

```
dbinom(0:10,size=10,prob=.3)
```

Or we can plot the results by embedding the dbinom() inside a call to the barplot() function, and provide some additional arguments for labeling the plot, as follows:

```
barplot(dbinom(0:10,size=10,prob=.3),names=c(0:10), xlab='# presences', ylab='probability',
    main='Probability Mass')
```

The plot depicts the probability of any getting any number of presences given the specified number of trials and probability of presence per trial.

For a continuous distribution, the probability distribution is referred to a *probability density function* (or pdf for short), which is a bit more confusing to understand. You may imagine that a measurment of (say) pH is *exactly* 7.9, but in fact what you have observed is that the pH is between 7.82 and 7.98 – if your meter has a precision of ±1%. Thus, continuous probability distributions are expressed as probability densities rather than probabilities – the probability that random variable X is between x an x+Δx, divided by Δx (Prob(7.82 < X < 7.98)/0.16, in this case). Dividing by Δx allows the observed probability density to have a well-defined limit as precision increases and Δx shrinks to zero. Unlike probabilities, probability densities can be larger than 1. In practice, we are concerned mostly with relative probabilities or likelihoods, and so the maximum density values and whether they are greater than or less than 1 don't matter much.

In the case of the *normal distribution*, for example, we can use the dnorm() function to calculate the probability density for any specified value of X given the mean (*mu*) and standard deviation (*sd*). For

example, we can compute the probability density for an individual with mass=12 given an underlying mean mass of say 10 and a standard deviation of say 3, as follows:

```
dnorm(12,mean=10,sd=3)
```

Or we can compute the probability density for a range of masses, as follows:

```
dnorm(0:20,mean=10,sd=3)
```

Or we can plot the results by embedding the dnorm() inside a call to the curve() function, and provide some additional arguments for labeling the plot, as follows:

```
curve(dnorm(x,mean=10,sd=3),0,20,xlab='z',ylab='Probability Density',main='Probability Density')
```

The plot depicts the probability density for any value of z given the specified mean and standard deviation. Note that we used the curve() function instead of the barplot() function because we have a continuous distribution.

3.2 Cumulative probability distribution

The cumulative probability distribution represents the probability that a random variable X is less than or equal to a particular value of x ($F(x) = Prob(X \leq x)$). Cumulative distribution functions are most useful for frequentist calculations of tail probabilities. In the case of the binomial example above, we can use the pbinom() function to calculate the probability of the species being present *n* or fewer times given the specified sample of size and per trial probability, and plot the result, as follows:

```
barplot(pbinom(0:10,size=10,prob=.3),names=c(0:10),
    xlab='# presences',ylab='Cumulative probability',main='Cumulative Probability')
```

We can do the same for the normal distribution example above, as follows:

```
curve(pnorm(x,mean=10,sd=3),0,20,
    xlab='z',ylab='Cumulative  probability', main='Cumulative Probability')
```

3.3 Quantile distribution

The quantile distribution represents the value of x for any given quantile of the cumulative probability distribution; i.e., it is the opposite of the cumulative probability distribution. For example, in the binomial example above, we can use the qbinom() function to calculate the number of times the species is expected to be present *n* or fewer times for any specified probability level, and plot the result, as follows:

```
barplot(qbinom(seq(0,1,.1),size=10,prob=.3),names=seq(0,1,.1), xlab='P', ylab='quantiles(z)',
    main='Quantiles')
```

We can do the same for the normal distribution example above, as follows:

```
curve(qnorm(x,mean=10,sd=3),0,1,xlab='P',ylab='quantiles(z)',main='Quantiles')
```

3.4 Random numbers

Random numbers can be drawn from a specified probability distribution. For example, in the binomial example above, we can use the rbinom() function to draw a random sample from the specified binomial distribution, as follows:

```
rbinom(1,size=10,prob=.3)
```

Note, the result is the number of successes (or species' presences, in this case) in a single sample of size=10 given a per trial probability of 0.3. If we repeat this process, we would expect to observe a different result each time owing to chance. If we can draw a large number of samples from this binomial distribution and plot the result, as follows:

```
y<-rbinom(1000,size=10,prob=.3)
hist(y,xlab='# successes',ylab='frequency',main='Random Numbers',col='gray')
```

Note, this plot displays the distribution of the number of successes (species presences) that we would observe in a sample of size 10 (i.e., 10 trials or plots) if we were to repeatedly sample the underlying population 1000 times. The mode of the distribution is the most likely number of successes (or presences) given the specified parameters of the distribution.

We can do the same for the normal distribution example above, as follows:

```
y<-rnorm(1000,mean=10,sd=3)
hist(y,xlab='z',ylab='frequency',main='Random Numbers',col='gray')
```

3.5 Other plotting functions

The biostats library includes two sets of useful plotting functions for a number of the most common probability distributions:

1. The first set of functions allows you to plot together on the same plot the probability distribution (e.g., dbinom) for several different sets of parameter values. This can facilitate understanding how the parameters affect the shape of the distribution. For example, we can plot the binomial probability distribution for a sample size of N=10 and three different per trial probabilities of success, as follows:

```
dbinom.plot(size=10,prob=c(.1,.5,.9))
```

We can do the same for the normal distribution example above, as follows:

```
dnorm.plot(mean=c(10,12),sd=c(1,2,3),xlim=c(0,20),ylim=c(0,1))
```

2. The second set of functions allows you to plot all four of the distribution plots as subplots on a single plot, including the probability mass, cumulative probability, quantile, and random sample plots. In addition, you can specify multiple values for each of the parameters and a separate four-part plot will be created for each combination of parameter values. For example, we can create several plots for the binomial distribution under different combinations of values for the size and probability parameters, as follows:

```
binom.plot(size=c(10,100),prob=c(.1,.5,.9))
```

We can do the same for the normal distribution example above, as follows:

```
norm.plot(mean=c(10,12),sd=c(1,2,3),xlim=c(0,20))
```

4. Bestiary of probability distributions

In the previous section, we developed some skills needed to examine probability distributions, but our examples were limited to a single discrete distribution (binomial) and a single continuous (normal). In this section, let's use the biostats functions introduced above to examine some of the other available distributions. In the functions below, you can use the range of parameter values given in the provided script or explore different values.

Discrete distributions:

```
dpois.plot(events=25,lambda=c(.5,1,3,12))
pois.plot(events=25,lambda=c(.5,1,3,12))

dnbinom.plot(events=25,mu=c(1,2),size=c(.1,1,10))
nbinom.plot(events=25,mu=c(1,2),size=c(.1,1,10))

dgeom.plot(events=25,prob=c(0.1,0.3,0.5,0.7,0.9))
geom.plot(events=25,prob=c(0.1,0.3,0.5,0.7,0.9))
```

Continuous distributions:

```
dgamma.plot(shape=c(1,2,5),scale=c(1,2,3),xlim=c(0,25),ylim=c(0,1))
gamma.plot(shape=c(1,2,5),scale=c(1,2,3),xlim=c(0,25))

dexp.plot(rate=c(1,.5,.1),xlim=c(0,15),ylim=c(0,1))
exp.plot(rate=c(1,.5,.1),xlim=c(0,15))

dbeta.plot(shape1=c(.5,1,2,5),shape2=c(.5,1,2,5),ylim=c(0,5))
beta.plot(shape1=c(.5,1,2,5),shape2=c(.5,1,2,5))

dlnorm.plot(mean=c(0,2),sd=c(.2,.5,1),xlim=c(0,15),ylim=c(0,2))
lnorm.plot(mean=c(0,2),sd=c(.2,.5,1),xlim=c(0,15))
```

```
dchisq.plot(df=c(1,2,10),xlim=c(0,20))
chisq.plot(df=c(1,2,10),xlim=c(0,20))

df.plot(df1=c(1,2,20),df2=c(10,20))
f.plot(df1=c(1,2,20),df2=c(10,20))

dt.plot(df=c(1,10,100))
t.plot(df=c(1,10))
```

5. Exercise – marbled salamander dispersal

The purpose of the following exercise is to give you additional experience working with probability distributions. You can work individually or in teams to complete the exercise.

Background.–The data for this exercise represent the dispersal of juvenile marbled salamanders from their natal ponds to neighboring ponds – the same data set as used in the previous lab. Recall that the data were derived from a long-term study of marbled salamanders in western Massachusetts in which a cluster of 14 vernal pools were monitored continuously between 1999-2004. All juveniles were marked upon leaving their natal ponds. Subsequent recaptures at non-natal ponds were used to determine dispersal rates between ponds for first-time breeders (ftb). In the data set provided, the dispersal rates have been standardized to account for several factors, including the propensity for dispersal from each pond and the available distances between ponds owing to the unique configuration of ponds. For our purposes, the most important standardization involves rescaling the dispersal rate to range from 0 to 1 (like a proportion). The data set includes three variables: (1) dist.class = distance class, based on 100 m intervals; (2) disp.rate.ftb = standardized dispersal rate for first-time breeders, which can be interpreted as a relative dispersal probability; and (3) disp.rate.eb = standardized dispersal rate for experienced breeders, which can be interpreted as a relative dispersal probability.

Your assignment is to examine the data set provided (...\dispersal.csv) and determine whether any of the existing probability distributions are appropriate for describing the error component of a parametric statistical model designed to describe the apparent relationship between dispersal of *first time breeders* and distance. Your specific tasks are as follows:

1. Examine the data set. Take care of the missing value (i.e., either drop record or impute an appropriate value), which will create problems below if not addressed. Plot the relationship between juvenile dispersal (disp.rate.ftb) and distance (dist.class).

2. Plot a histogram of the dependent variable (disp.rate.ftb) and overlay it with a continuous, smoothed curve using the kernel density function, as follows:

   ```
   hist(disp$disp.rate.ftb, breaks=10, freq=FALSE)
   lines(density(disp$disp.rate.ftb, na.rm=TRUE))
   ```

3. Based on the histogram and the type (and scale) of the dependent variable, consider whether any of the standard parametric probability distributions can be used to describe the observed error

distribution. If not, consider what the alternatives might be. Note, you need to carefully consider the various probability distributions that are now in your toolbox (from the bestiary), and review the use and limitations of each from the lecture notes and Bolker if need be.

4. Select a deterministic function to describe the apparent dispersal-distance relationship (from the previous lab) and based on your previous fit of the function to the data (i.e., either your eyeball estimates of parameters or your ordinary least squares fit from the extended lab) plot the fitted curve on the scatter plot. Note, this was completed during the previous lab exercise, but you should repeat it here to include in your lab report (see below).

5. Compute the residuals (or errors); i.e., the deviation between the observed values and the fitted values for each observation. For example, let's say you have calculated the fitted values for your deterministic function and they are stored in the object y.pred. You can calculate the residuals (or errors) as follows:

```
y.resid<-disp$disp.rate.ftb-y.pred
```

6. Plot a histogram of the residuals (i.e., errors) and overlay it with a continuous, smoothed curve using the kernel density function, as follows:

```
hist(y.resid,breaks=10, freq=FALSE)
lines(density(y.resid, na.rm=TRUE))
```

7. Based on the histogram of the residuals, determine whether the shape of the distribution is plausible give the probability distribution chosen in step 3. Note, be sure not to let the fact that the residuals will take on both positive and negative values influence you; it is the shape of the distribution that is of concern here.

8. Plot a scatterplot of the fitted values (x-axis) versus the residuals (y-axis) and inspect the plot for any discrepancies. In other words, is the pattern of the residuals consistent with the expectations from the tentative error model selected in step 3. Are there any patterns in the residuals that might suggest one error model as being more appropriate than another.

9. Hand in a two-page report (including figures) containing your names, three figures: (1) histogram of the dependent variable with the kernel density curve overlaid (from step 2), (2) histogram of the residuals with the kernel density curve overlaid (from step 6), and (3) ; and scatterplot of the fitted values versus residuals (from step 8); and a brief discussion of the applicability of existing parametric error distributions to use in a plausible statistical model for this data (i.e., choose one or model error models and justify your choice).