**http://www.unc.edu/courses/2010fall/ecol/563/001/docs/lectures/lecture4.htm**

**Lecture 4—Wednesday, September 8, 2010**

**Topics**

* [Probability distributions in regression](http://www.unc.edu/courses/2010fall/ecol/563/001/docs/lectures/lecture4.htm#probability)
  + [The role of the probability model in ordinary linear regression](http://www.unc.edu/courses/2010fall/ecol/563/001/docs/lectures/lecture4.htm#role)
  + [Extending regression to other probability distributions for the response](http://www.unc.edu/courses/2010fall/ecol/563/001/docs/lectures/lecture4.htm#extending)
  + [Things to consider when choosing a probability model](http://www.unc.edu/courses/2010fall/ecol/563/001/docs/lectures/lecture4.htm#some)
* [A review of probability and probability distributions](http://www.unc.edu/courses/2010fall/ecol/563/001/docs/lectures/lecture4.htm#review)
* [Normal distribution](http://www.unc.edu/courses/2010fall/ecol/563/001/docs/lectures/lecture4.htm#normal)
  + [Basic characteristics](http://www.unc.edu/courses/2010fall/ecol/563/001/docs/lectures/lecture4.htm#normbasic)
  + [Importance and use](http://www.unc.edu/courses/2010fall/ecol/563/001/docs/lectures/lecture4.htm#normimport)
  + [R probability functions for the normal distribution](http://www.unc.edu/courses/2010fall/ecol/563/001/docs/lectures/lecture4.htm#probfunc)
* [Bernoulli (binary) distribution](http://www.unc.edu/courses/2010fall/ecol/563/001/docs/lectures/lecture4.htm#bernoulli)
  + [Basic characteristics](http://www.unc.edu/courses/2010fall/ecol/563/001/docs/lectures/lecture4.htm#bernbasic)
  + [Importance and use](http://www.unc.edu/courses/2010fall/ecol/563/001/docs/lectures/lecture4.htm#bernimport)
* [Binomial distribution](http://www.unc.edu/courses/2010fall/ecol/563/001/docs/lectures/lecture4.htm#binomial)
  + [Basic characteristics](http://www.unc.edu/courses/2010fall/ecol/563/001/docs/lectures/lecture4.htm#binbasic)
  + [Importance and use](http://www.unc.edu/courses/2010fall/ecol/563/001/docs/lectures/lecture4.htm#binimport)
  + [Derivation of the formula for the binomial probability mass function](http://www.unc.edu/courses/2010fall/ecol/563/001/docs/lectures/lecture4.htm#binderiv)
  + [Examples of binomial distributions](http://www.unc.edu/courses/2010fall/ecol/563/001/docs/lectures/lecture4.htm#binexamp)
* [Poisson distribution](http://www.unc.edu/courses/2010fall/ecol/563/001/docs/lectures/lecture4.htm#Poisson)
  + [Basic characteristics](http://www.unc.edu/courses/2010fall/ecol/563/001/docs/lectures/lecture4.htm#poisbasic)
  + [Probability mass function for the Poisson](http://www.unc.edu/courses/2010fall/ecol/563/001/docs/lectures/lecture4.htm#poismass)
  + [Importance and use](http://www.unc.edu/courses/2010fall/ecol/563/001/docs/lectures/lecture4.htm#poisimport)
  + [Examples of Poisson distributions](http://www.unc.edu/courses/2010fall/ecol/563/001/docs/lectures/lecture4.htm#poisexamp)
* [Reference](http://www.unc.edu/courses/2010fall/ecol/563/001/docs/lectures/lecture4.htm#reference)
* [R code used in lecture 4](http://www.unc.edu/courses/2010fall/ecol/563/001/docs/lectures/lecture4.htm#Rcode)

**Probability distributions in regression**

**The role of the probability model in ordinary linear regression**

In ordinary regression we assume that the response variable has a normal distribution conditional on the values of the predictors in the model. In a simple linear regression model this would be written as follows.

conditionalmean

The vertical bar is the usual notation for conditional probability. Thus conditional on the value of *x*, *Y* has a normal distribution with mean β0 + β1*x* and variance σ2.

In truth the normal probability model has only a secondary role in ordinary linear regression. Least squares, which is typically used to obtain estimates of the regression coefficients, makes no assumption about the underlying probability model. Least squares finds the values of β0 and β1 that minimize the error we make when we use the model prediction, β0 + β1xi, instead of the actual data value yi. Formally, least squares minimizes the sum of squared errors, SSE.

SSE

The SSE criterion is easily extended to the case where there are multiple predictors. The parameter values that minimize SSE can be found either using calculus or by using the concepts of projection operators and subspaces from linear algebra.

Having obtained parameter estimates for the data at hand we might wonder how stable those estimates are. What would happen if we went back and collected new data from the same population that gave rise to our first sample? Obviously estimates based on the new sample would be different but how different? To answer this question we need to know how observations vary from sample to sample and for that we need a probability model.

In the least squares formulation of the regression problem, the regression line is treated as a signal contaminated by noise. The signal plus noise view of regression is written as follows.

signal noise

Typically the xi are treated as fixed and the yi and εi are random. Choosing a probability model for εi also gives us a probability model for yi. In the signal plus noise formulation least squares tries to find the values of β0 and β1 that minimize SSE2, the sum of the "squared errors". Because the errors are squared, positive and negative deviations from the line are treated as being equally important. Consequently the only reasonable probability distributions for the errors are symmetric ones. This makes the normal distribution an obvious choice. (In the engineering literature various integrated functions of the normal distribution are often referred to as error functions, or erfs.) We assume that

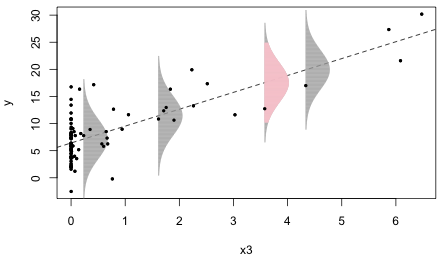
error distribution

From this we obtain that the response, yi, is also normally distributed with variance σ2 but with a mean given by the regression line.

Having assumed that the response has a normal distribution, statistical theory tells us that both the intercept and slope will also have normal distributions. This result is the basis for the statistical tests that appear in the summary tables of regression models. Still, it's important to realize that the least squares estimates themselves are obtained without specifying a probability model for the response.

**Extending regression to other probability distributions for the response**

Fig. 1 illustrates the role that the normal distribution plays in ordinary regression. The data used in generating Fig. 1 are the same simulated data that were used in lecture 1. Letting http://www.unc.edu/courses/2010fall/ecol/563/001/images/lectures/lecture4/normalerrors.gifI generated observations of the form linear modeland used ordinary least squares to estimate the regression lineestimated regression . Fig. 1 displays the raw data as well as the regression line that was fit to those data.



**Fig. 1** The normal distribution as a data-generating mechanism. [R code for Fig. 1](http://www.unc.edu/courses/2010fall/ecol/563/001/notes/lec4%20fig1Rcode.txt)

The estimated errors, ei, are the vertical distances of the individual points to the plotted regression line. The estimated normal error distributions at four different locations are shown. Each normal distribution is centered on the regression line. The data values at these locations are the points that appear to just overlap the bottom edges (left sides) of the individual normal distributions. I've drawn the normal distributions so that they extend ± 3 standard deviations above and below the regression line. (The pink normal curve extends ± 2 standard deviations with the gray tails then extending out the remaining 1 standard deviation.)

Instead of thinking of the normal curves as error distributions, we can think of them as data generating mechanisms. At each point along the regression line, the normal curve centered at that point gives the likely locations of data values for that specific value of x3 according to our model. Recall that 95% of the values of a normal distribution fall within ± 2 standard deviations of the mean while nearly all of the observations (approximately 99.9%) fall within ± 3 standard deviations of the mean. The four selected data values associated with the four displayed normal curves are all well within ± 2 standard deviations of the regression line. Thus they could easily have been generated by the model.

The way to generalize the data-generating approach to regression is obvious. Replace the normal curves by some other probability distribution and think of the new probability models as the data generating mechanisms. How should we choose these probability models?

**Things to consider when choosing a probability model**

1. Is the measured response variable discrete or continuous? An extreme version of discrete is categorical where the recorded numbers just label categories.
2. Are the possible values of the response constrained to some interval or are they theoretically unbounded?
3. If we group the values of the response variable with respect to the categories of putative predictors, do the mean and variance of the response show an obvious relationship?
4. Does theory suggest that a particular probabilistic mechanism may have generated the data we obtained?

**A review of probability and probability distributions**

A mathematical function that maps the outcomes of a random experiment (domain) to a set of real numbers (range) is called a random variable. Formally a random variable is a function on the sample space of possible outcomes. Typically in applications the response variable is the random variable whereas the predictors are fixed. If the range of a random variable can be put into one-to-one correspondence with the set of non-negative integers (or if the range is just a finite set) we call the random variable discrete. If the range is the set of real numbers (or a set isomorphic to the set of real numbers) we call it a continuous random variable.

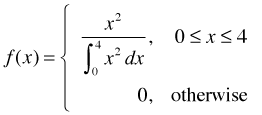
Generally we associate with the random variable *X* another function called a probability function that describes the probability distribution of *X*. If *X* is a discrete random variable we call this function a **probability mass function** and if *X* is continuous we call it a **probability density function**. We will denote the probability function by *f*(*x*) or densitywhen we wish to emphasize the fact that the distribution depends on certain parameters that need to be specified. Here the symbols α and β denote the parameters. As an example, in a normal distribution the parameters are the mean μ and the variance σ2.

In an axiomatic formulation of probability theory a probability function must satisfy three properties.

1. The probability function is non-negative, i.e.,non-negativity. If the function is a probability mass function then we also must have range. (This is not the case for density functions.)
2. If we sum the probability function over all possible value of *x* we get 1.
   1. *X* discrete: sum to 1
   2. *X* continuous: integrate to 1
3. If *A* and *B* are sets of possible values of the random variable *X* such that they are disjoint, i.e., disjoint, then additivity. So the function is additive on disjoint sets.

A large number of probability distributions have been studied and been given names. A flowchart of the relationships between those distributions commonly used in scientific applications appears [here](http://www.johndcook.com/distribution_chart.html). A more complete version of this flowchart appears on p. 47 of [Leemis and McQueston (2008)](http://www.math.wm.edu/~leemis/2008amstat.pdf).

In truth any non-negative portion of a discrete continuous function can be reformulated into a probability function. Consider for instance the function *x*2, but restricted to the interval 0 ≤ *x* ≤ 4. If we define



then *f*(*x*) satisfies the three properties listed above and hence is a probability density function.

I examine in detail five distributions that are commonly used to model biological data: normal, Bernoulli (binary), binomial, Poisson, and negative binomial distributions. Other useful distributions such as the lognormal, gamma, multivariate normal, and Wishart distributions will be discussed later as the need arises.

**Normal Distribution**

**Basic characteristics**

* A continuous distribution.
* It has two parameters, denoted μ and σ2, which also happen to be the mean and variance of the distribution.
* We write normal.
* One hallmark of the normal distribution is that the mean and variance are independent. There is no relationship between them. Knowing one tells us nothing about the value of the other. This characteristic makes the normal distribution unusual.
* The normal distribution is symmetric.
* The normal distribution is unbounded both above and below. Hence the normal distribution is defined for all real numbers.
* R normal functions: **dnorm**, **pnorm**, **qnorm**, and **rnorm**. These are described below.

**Importance and use**

The importance of the normal distribution stems from the central limit theorem. In words, if what we observe in nature is the result of adding up lots of independent and identically distributed things, then no matter what their individual distributions are the distribution of the sum tends to become normal the more things we add up. As a practical application sample means tend to have a normal distribution when the sample size is big enough because in calculating a mean we have to add things up.

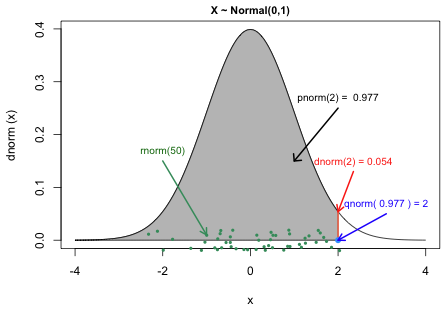
Even if the response is a discrete random variable such as a count, a normal distribution may be an adequate approximation to its distribution if we’re dealing with a large sample and the values we've obtained are far removed from any boundary values. On the other hand count data with lots of zero values cannot possibly have a normal distribution nor be transformed to approximate normality.

**R probability functions for the normal distribution**

For each probability distribution in R there are four basic probability functions. Each of R's probability functions begins with one of four prefixes—**d**, **p**, **q**, or **r**—followed by a root name that identifies the probability distribution. For the normal distribution the root name is "norm". The meaning of these prefixes is as follows.

* **d** is for "density" and the corresponding function returns the value of the probability density function (continuous distributions) or the probability mass function (discrete distributions).
* **p** is for "probability" and the corresponding function returns a value from the cumulative distribution function.
* **q** is for "quantile" and the corresponding function returns a value from the inverse cumulative distribution function, also know as the quantile function.
* **r** is for "random" and the corresponding function returns a randomly drawn value from the given distribution.

To better understand what these functions do I'll focus on the four probability functions for the normal distribution: **dnorm**, **pnorm**, **qnorm**, and **rnorm**. Fig. 2 illustrates the relationships among these four functions.

  
**Fig. 2** The four probability functions for the normal distribution ([R code for Fig. 2](http://www.unc.edu/courses/2010fall/ecol/563/001/notes/lec4%20fig2Rcode.txt))

**dnorm** is the normal probability density function. Without any further arguments it returns the density of the standard normal distribution (mean 0 and standard deviation 1). If you plot dnorm(x) over a range of *x*-values you obtain the usual bell-shaped curve of the normal distribution. In Fig. 2, the value of dnorm(2) is indicated by the height of the vertical red line segment. It's the *y*-coordinate of the normal curve when *x* = 2. Keep in mind that density values are not probabilities. To obtain probabilities from densities one has to integrate the density function over an interval. Alternatively for a very small interval, say one of width Δ*x*, if *f*(*x*) is a probability density function, then we can approximate the probability of *x* being in that interval as follows.

probability from a density

**pnorm** is the cumulative distribution function for the normal distribution. By definition pnorm(x) = *P*(*X* ≤ *x*) which is the area under the normal density curve to the left of *x*. Fig. 2 shows pnorm(2)as the shaded area under the normal density curve to the left of *x* = 2. As is indicated on the figure, this area is 0.977. So the probability that a standard normal random variable takes on a value less than or equal to 2 is 0.977.

**qnorm** is the quantile function of the standard normal distribution. If qnorm(x) = *k* then *k* is the value such that P(*X* ≤ *k*) = *x* . **qnorm** is the inverse function for **pnorm**. From Fig. 2 we have, qnorm(0.977) = qnorm(pnorm(2)) = 2.

**rnorm** generates random values from a standard normal distribution. The only required argument is a number specifying the number of realizations of a normal random variable to produce. Fig. 2 illustrates rnorm(50), the locations of 50 random realizations from the standard normal distribution, jittered slightly to prevent overlap.

To obtain normal distributions other than the standard normal, all four normal functions support the additional arguments **mean** and **sd** for the mean and standard deviation of the normal distribution. For instance, dnorm(x, mean=4, sd=2) is a normal density with mean 4 and standard deviation 2. Notice that R parameterizes the normal distribution in terms of the standard deviation rather than the variance.

**Bernoulli (binary) distribution**

**Basic characteristics**

* The Bernoulli distribution is almost the simplest probability model imaginable. (An even simpler model is a point mass distribution in which all the probability is assigned to a single point. A point mass distribution is only useful in combination with other distributions as part of a mixture.)
* A Bernoulli random variable is discrete.
* There are only two possible outcomes: 0 and 1, failure and success. Thus we are dealing with purely nominal data. Since there are only two categories, we also refer to these as **binary data**.
* The classical example of the Bernoulli distribution is an experiment in which we record the outcome of the single flip of a coin.
* The Bernoulli distribution has one parameter *p*, the probability of success, with 0 ≤ *p* ≤ 1.
* The notation we will use is Bernoulli, to be read *"X* is distributed Bernoulli with parameter *p*."
* The mean of the Bernoulli distribution is *p.*
* The variance of the Bernoulli distribution is *p*(1 – *p*).
* There is no special Bernoulli function in R. Just use the binomial functions described below with **size** = 1 (*n* = 1).
* WinBUGS has both Bernoulli and binomial mass functions: **dbern** and **dbin**.

**Importance and Use**

An example of its use in ecology is in developing habitat suitability models of the spatial distribution of endangered species. Typically we record the presence-absence of the species in a habitat (using perhaps a set of randomly located quadrats). We then try to relate the observed species distribution to characteristics of the habitat. Each individual species occurrence is treated as the realization of a Bernoulli random variable whose parameter *p* is modeled as a function of habitat characteristics.

**Binomial distribution**

**Basic characteristics**

* A discrete distribution
* A binomial random variable records the number of successes out of *n* trials.
* The classical example of a binomial distribution is an experiment in which a coin is flipped *n* times and we record the number of heads.
* The set of possible values a binomial random variable can take is bounded on both sides—below by 0, above by *n*.
* Formally a binomial random variable arises from a binomial experiment, which is an experiment consisting of a sequence of *n* independent Bernoulli trials. If X1 to Xnare independent and identically distributed Bernoulli random variables each with parameter *p*, then

Bernoulli sum

is said to have a **binomial distribution** with parameters *n* and *p*. We write this as binomial.

* To contrast the binomial distribution with the Bernoulli, we sometimes refer to data arising from a binomial distribution as **grouped binary data**.
* The R binomial functions are denoted **dbinom**, **pbinom**, **qbinom**, and **rbinom**. In R the parameters *n* and *p* correspond to the argument names **size** and **prob** respectively. In WinBUGS the binomial distribution is denoted **dbin**.
* Mean: *binomial mean.* [NB. The notation *E*(*X*) is mathematical notation for the mean of the random variable *X*. The symbol *E* denotes expectation, so *E*(*X*) is also referred to as the expected value of *X*.]
* Variance: binomial variance.

Observe from this last expression that the variance is a function of the mean, i.e., variance function. So one characteristic of a binomial random variable is that the mean and variance are not independent. If you plot the variance of the binomial distribution against the mean you obtain a parabola opening downward in which the maximum occurs at max(and hence when *p* = 0.5).

**Importance and Use**

In order for an experiment to be considered a binomial experiment four basic assumptions must hold.

1. The experiment must consist of a sequence of trials in which each trial is a Bernoulli trial, meaning only one of two outcomes with probabilities *p* and 1 – *p* can occur.
2. The number of trials is fixed ahead of time at *n*, a value that is known to us.
3. The probability *p* is the same on each Bernoulli trial.
4. The Bernoulli trials are independent. Recall that for independent events *A* and *B*, independence.

If (1) or (2) are violated then a binomial model is completely inappropriate and a different probability distribution should be used. If either (3) or (4) are violated, it may be possible to salvage things by coupling a binomial distribution with one of the methods we will discuss in the second half of this course.

A simple illustration of a binomial random variable is the response of a seed germination experiment. Suppose an experiment is carried out in which 100 seeds are planted in a pot and the number of seeds that germinate is recorded. Suppose this is done repeatedly for different pots that are subjected to various light regimes, burial depths, etc. Clearly the first two assumptions of the binomial model hold here.

1. The outcome on individual trials (the fate of an individual seed in the pot) is dichotomous (germinated or not).
2. The number of trials (number of seeds per pot) was fixed ahead of time at 100.

Assumptions (3) and (4), constant *p* and independent trials, would need to be verified. We'll discuss how to do this when we look at regression models for binomial random variables.

**Derivation of the formula for the binomial probability mass function**

Suppose we have five independent Bernoulli trials with the same probability *p* of success on each trial. If we observe the event:

success pattern,

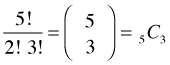
i.e., three successes and two failures in the order shown, then by independence this event has probability

probability of one permutation.

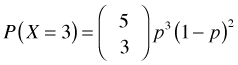
But in a binomial experiment we don’t observe the actual sequence of outcomes, just the number of successes, in this case 3. There are many other ways to get 3 successes, just rearrange the order of *S* and *F* in the sequence *SFSSF*. So, the probability we have calculated here is too small. How many other distinct arrangements (permutations) of three *S*s and two *F*s are there?

* If all permutations are distinguishable, as in *ABCDE,* then elementary counting theory tells us there are 5! = 120 different arrangements.
* Replace *B* and *E* in this permutation by *F* yielding *AFCDF* so that now the second and fifth outcomes are indistinguishable. In the original sequence *ABCDE* and *AECDB* would be recognizable as different arrangements, but now they would be indistinguishable. With five distinct letters every time you write down a different arrangement of the five letters you immediately get another arrangement just by swapping the *B* and *E*. So when *B* and *E* are identical, 5! over counts the number of arrangements by a factor of 2.
* Now suppose we replace *A*, *C*, and *D* by *S*, *SBSSE*. In the original sequence you could write down one arrangement of the letters and then immediately get 3! = 6 more by swapping the letters *A*, *C*, and *D* in all possible ways. Thus when *A*, *C*, and *D* are indistinguishable 5! over counts the number of possible arrangements by a factor of 6.

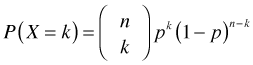
Thus to answer the original question, the number of distinct arrangements of three *S*s and two *F*s is



where the last two symbols are two common notations for this quantity. Carrying out the arithmetic of this calculation we find that there are ten distinct arrangements of three *S*s and two *F*s. The first notation, binomial coefficient, is called a **binomial coefficient** and is read "5 choose 3". The *C* in the second notation denotes "combination" and thus combinationis the number of combinations of five things taken three at a time. Putting this altogether, if binomialthen

.

For a generic binomial random variable, binomial, in which the total number of trials is denoted by *n*, we have

.

Here *k* = 0, 1, 2, … , *n*.

**Examples of binomial distributions**

Because the binomial distribution is discrete (unlike the continuous normal distribution), the proper way to display it is as a bar chart in which the bars are centered on the discrete values and the heights of the bars give the probability of that value. One way to obtain this in R is with the **type='h'** option, 'h' for height, of the **plot** function. The **dbinom** function is the probability mass function for the binomial and returns the probability of the specified value. The syntax of **dbinom** is dbinom(x, size, prob) where **size** is what we have been calling *n* and **prob** is what we've been calling *p*.

I start by placing three binomial distributions side-by-side. All have *n* = 10, but *p* varies from *p* = 0.1, 0.5, to 0.9. I use the **mfrow** argument of the **par** function of R to obtain a graphical display consisting of 1 row and 3 columns.

**#set graphics window to have one row and three columns**

par(mfrow=c(1,3))

**#set bar ends to have square ends**

par(lend=2)

plot(0:10, dbinom(0:10,10,.1), type='h', lwd=6, xlab='#successes', ylab='probability')

mtext(side=3,line=.5,'p = 0.10')

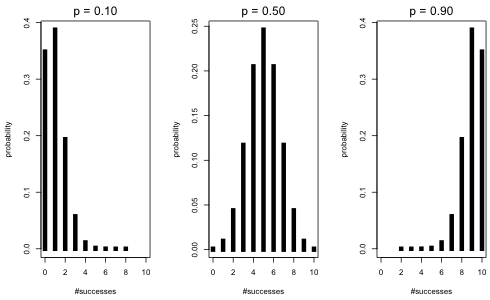
plot(0:10, dbinom(0:10,10,.5), type='h', lwd=6, xlab='#successes', ylab='probability')

mtext(side=3,line=.5,'p = 0.50')

plot(0:10, dbinom(0:10,10,.9), type='h', lwd=6, xlab='#successes', ylab='probability')

mtext(side=3,line=.5,'p = 0.90')

par(mfrow=c(1,1))



**Fig. 3** Three binomial probability mass functions with n = 10 and p = 0.1, 0.5, and 0.9

Observe that as *p* approaches the endpoint values of 0 and 1 the distributions are more skewed. When *p* = 0.5 the distribution looks quite symmetric.

I repeat this for *p* = 0.5, 0.8, and 0.9 but this time with a much larger number of trials, *n* = 100.

par(mfrow=c(1,3))

plot(30:70, dbinom(30:70,100,.5), type='h', lwd=2, xlab='#successes', ylab='probability')

mtext(side=3,line=.5,'p = 0.10')

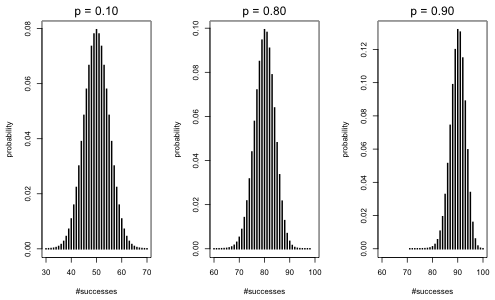
plot(60:100, dbinom(60:100,100,.8), type='h', lwd=2, xlab='#successes', ylab='probability')

mtext(side=3,line=.5,'p = 0.80')

plot(60:100, dbinom(60:100,100,.9), type='h', lwd=2, xlab='#successes', ylab='probability')

mtext(side=3,line=.5,'p = 0.90')

par(mfrow=c(1,1))



**Fig. 4** Three binomial probability mass functions with n = 100 and p = 0.5, 0.8, and 0.9

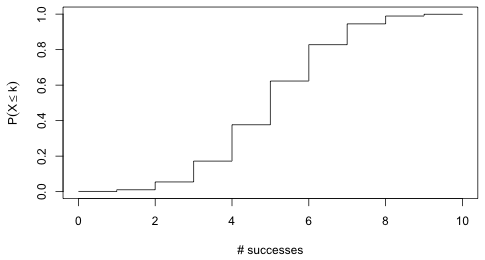
Notice that the distributions for *p* = 0.5 and 0.8 are nearly perfectly symmetrical, while the distribution for *p* = 0.9 is nearly symmetrical showing only a hint of left skewness.

Because a binomial distribution is the sum of *n* independent Bernoulli random variables, the central limit theorem applies. From the central limit theorem when we add up a large number of independent and identically distributed random variables, no matter what their distribution, the sum will tend to look normally distributed. Fig. 4 is an illustration of the central limit theorem in practice and is an empirical rationale for the normal approximation to the binomial. Practically speaking, if you have binomial data and your *n* is big enough, you could analyze your binomial counts as if they came from a normal distribution. In order for the approximation to be good it turns out that both the products *np* and *n*(1–*p*) need to be large. If both products exceed 10 that is typically large enough for the normal approximation to be a reasonable one.

**Other R binomial functions**

The **pbinom** function is the cumulative distribution function for the binomial. Because the binomial distribution is discrete, its cdf is a step function.

plot(0:10, pbinom(0:10,10,.5), type='s', xlab='# successes', ylab=expression(P(X<=k)))



**Fig. 5** Cumulative distribution function for the binomial distribution (pbinom)

The **qbinom** and **pbinom** functions are inverses of each other To obtain the value of **pbinom** I start on the *x*-axis and read up to the graph in Fig. 5 until I hit a horizontal line and then read off the corresponding value on the *y*-axis. To obtain the value of **qbinom** I choose a probability on the *y*-axis and then move horizontally to the graph until I hit a vertical line at which point I read off the # of successes on the *x*-axis. The **qbinom** function is also a step function that returns only integer values.

pbinom(2,10,.5)

[1] 0.0546875

qbinom(0.0546875,10,.5)

[1] 2

qbinom(0.053,10,.5)

[1] 2

qbinom(0.055,10,.5)

[1] 3

The **pbinom** function can also used to obtain tail probabilities, i.e., probabilities of the form P(*X* > *k*). Because P(*X* > *k*) = 1 – P(*X* ≤ *k*) it follows that P(*X* > *k*) = 1 – pbinom(k, n, p).

**#P(X ≤ k)**

pbinom(7,10,.5)

[1] 0.9453125

**#calculate by hand**

sum(dbinom(0:7,10,.5))

[1] 0.9453125

**#P(X > 7)**

1-pbinom(7,10,.5)

[1] 0.0546875

**#calculate by hand**

sum(dbinom(8:10,10,.5))

[1] 0.0546875

**Poisson Distribution**

**Basic characteristics**

* A Poisson random variable is discrete.
* A typical use for the Poisson distribution is as a model for count data.
* The Poisson distribution is bounded on only one side. It is bounded below by 0, but has no upper limit. This distinguishes it from the binomial distribution which is bounded on both sides.
* Example of a Poisson random variable: the number of cases of Lyme disease in a North Carolina county in a year.
* If the following three assumptions hold then any non-negative discrete random variable recording the number of events in an interval of time or in a region of space will follow a Poisson distribution.
  + **Homogeneity** assumption: Events occur at a constant rate λ such that on average for any length of time *t* we expect to see *λt* events.
  + **Independence** assumption: For any two non-overlapping intervals the number of observed events is independent.
  + If the interval is very small, then the probability of observing two or more events in that interval is essentially zero. (More specifically, this probability is much, much smaller than the probability of observing one event.)
* The Poisson distribution is a one-parameter distribution. The parameter is usually denoted with the symbol λ, the rate.
* The mean of the Poisson distribution is equal to the rate, λ. The variance of the Poisson distribution is also equal to λ. Thus in the Poisson distribution the variance is equal to the mean. So we have if Poisson distributionthen

Mean: Poisson mean

Variance: Poisson variance

Observe that the variance is a function of the mean, i.e., Posson variance. Thus when the mean gets larger, the variance gets larger at exactly the same rate.

* The R Poisson probability functions are **dpois**, **ppois**, **qpois**, and **rpois**.

**Probability mass function for the Poisson**

Let Poissonwhere *Nt* is the number of events occurring in a time interval of length *t*, then the probability of observing *k* events in that interval is

Poisson time

The Poisson distribution can be applied to events occurring in time or space. In a Poisson model of two-dimensional space, events occur again at a constant rate such that the number of events observed in an area A is expected on average to be λA. In this case the probability mass function is written as follows.

Poisson area

Often in application the time interval or the area is the same for all observations. For example, if all the quadrats are the same size then we don't need to know what *A* is. In such cases we suppress *t* and *A* in our formula for the probability mass function and write instead

Poisson probability

**Importance and Use**

For spatial distributions the Poisson distribution plays a role in defining a common null model called **CSR—complete spatial randomness**.

* If we imagine moving a window of fixed size over a landscape then due to the homogeneity assumption of the Poisson distribution, no matter where we move the window the picture should look essentially the same. We will on average see the same number of events in each window. This rules out clumping. If the distribution were aggregated then some snapshots from our moving window would show many events, while others would show none. Note: in a clumped distribution the variance will be greater than the mean.
* Due to the independence assumption the spatial distribution under CSR will not appear very regular. If a regular, equally-spaced distribution is observed then nearby events are interfering with each other to produce the regular spacing. The assumption of the independence of events in non-overlapping regions means that interference of this sort is impossible. In a regular distribution the variance is smaller than the mean.

Many probability distributions have no theoretical motivation but get used in biological applications because they resemble distributions seen in nature. Unlike these distributions the Poisson distribution can be motivated by theory. There are two primary ways that the Poisson distribution arises in practice from theory.

1. Using the three assumptions listed above one can derive the formula for the Poisson probability mass function directly from first principles. The process involves setting up differential equations that describe the probability of seeing a new event in the next time interval. If you would like to see how this is done see this [document](http://www.isds.duke.edu/courses/Fall06/sta113/poisson.pdf) from a probability for engineers course taught at Duke. The theoretical motivation for the Poisson is important because if a Poisson model does not fit our data we can use it to help us understand why not. We can examine the assumptions and try to determine which one(s) is(are) being violated.
2. Another way the Poisson probability model arises in practice is as an approximation to the binomial distribution. Suppose we have binomial data in which *p*, the probability of success is very small, and *n*, the number of trials, is very large. In this case the Poisson distribution is a good approximation to the binomial where λ = *np*. Formally it can be shown that if you start with the probability mass function for a binomial distribution and let *n* → ∞ and *p* → 0 in such a way that *np* remains constant, you obtain the probability mass function of the Poisson distribution. [Here's](http://www.uoxray.uoregon.edu/ph353/Poisson_distribution.pdf) a document from a physics course taught at the University of Oregon that illustrates the proof of this fact.

**Examples of Poisson distributions**

I graph the Poisson distribution for different choices of the parameter λ.

par(mfrow=c(1,3))

par(lend=2)

plot(0:10, dpois(0:10,1), type='h', lwd=6, xlab='count category', ylab='probability')

mtext(side=3,line=.5, expression(lambda==1))

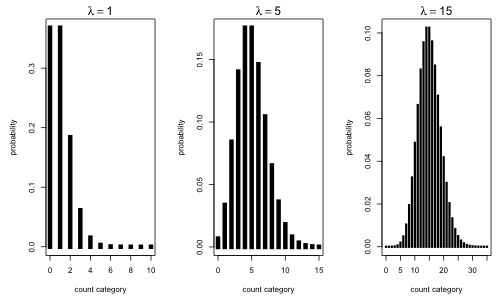
plot(0:15, dpois(0:15,5), type='h', lwd=6, xlab='count category', ylab='probability')

mtext(side=3,line=.5, expression(lambda==5))

plot(0:35, dpois(0:35,15), type='h', lwd=3, xlab='count category', ylab='probability')

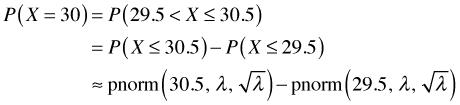
mtext(side=3,line=.5, expression(lambda==15))

par(mfrow=c(1,1))



**Fig. 6** The Poisson distribution for different choices of the rate constant λ

Notice that when the rate constant λ = 1 the zero category is extremely common comprising nearly 40% of the observations. For small values of λ the Poisson distribution is skewed, but as λ gets larger the Poisson distribution becomes more symmetric. Not surprisingly given that counts are really sums, the central limit theorem kicks in when λ gets large enough (usually λ > 30 is sufficient). In such cases it is reasonable to use a normal distribution to approximate the discrete Poisson distribution if desired. Keep in mind that in using the normal distribution it will be necessary to calculate probabilities over intervals to find the probabilities of specific count categories. For example, using a normal distribution we would have to calculate the Poisson probability P(*X* = 30) as follows.



**Reference**

Leemis, Lawrence M. and Jacquelyn T. McQueston. 2008. Univariate distribution relationships. *The American Statistician* **62**(1): 45–53.

**R Code**

A compact collection of all the R code used in this document appears [here](http://www.unc.edu/courses/2010fall/ecol/563/001/notes/lecture4%20Rcode.txt).

[Course Home Page](http://www.unc.edu/courses/2010fall/ecol/563/001/index.html)

|  |
| --- |
| Jack Weiss *Phone:* (919) 962-5930 *E-Mail:* jack\_weiss@unc.edu *Address:* Curriculum in Ecology, Box 3275, University of North Carolina, Chapel Hill, 27516 Copyright © 2010 Last Revised--Sept 11, 2010 URL: [http://www.unc.edu/courses/2010fall/ecol/563/001/docs/lectures/lecture4.htm](http://www.unc.edu/courses/2010fall/ecol/563/001/docs/lectures/lecture4.htm#lecture3) |