3. DISCRETE PROBABILITY DISTRIBUTIONS

Probability distributions may be discrete or continuous. This week we examine two discrete distributions commonly used in biology: the binomial and Poisson distributions. We will use JMPin to generate random samples from these distributions and explore their characteristics.

The Binomial Distribution

The binomial distribution is one of the most commonly encountered discrete probability distributions in biology. It is based on nominal scale data that come from a population with only two categories. One of the two categories is arbitrarily referred to as a "success" and the other a "failure". These categories are mutually exclusive events (i.e. female [success] vs male [failure]; black vs white; left vs right). A proportion p of the individuals in the population are in the success category, and a proportion q are in the failure category. The process of randomly selecting an individual from the population is called a **trial**. The probability of a success p remains constant from trial to trial. Since success and failure are mutually exclusive events, and represent the universe of possibilities, the probability of failure in any one trial is $q = 1 - p$. The outcome of any particular trial is not affected by the outcome of any other trial (i.e. trials are independent). Under these conditions, the probability of obtaining X successes in an independent sequence of n trials has a binomial distribution, where:

$$P(X) = \frac{n!}{X!(n-X)!} p^X q^{n-X}$$

The mean of X is $\mu = np$. The variance of X *is* $\sigma^2 = npq$ (the standard deviation is the square root of this).

Binomial Test

The binomial distribution can be used to estimate, and test hypotheses about, proportions of events in populations. What fraction of individuals in the country are infected with HIV? Are left-handed and right-handed individuals equally common? With your notes and the course textbook, review your knowledge of the following concepts:

- standard error
- null hypothesis (Ho) and alternate hypothesis (Ha).
- P-value
- Type I errors and Type II errors
- significance level

The Poisson Distribution

Another discrete probability distribution commonly encountered in biology is the Poisson distribution. This distribution is important in describing random occurrences of events in space or in time. For example, imagine that you would scatter seeds over a vast field from an airplane. Imagine also that you have divided the field up into blocks of equal size, say 10×10 metres in area.

If the probability that a given square millimetre of soil receives a seed is low (you haven't dropped a trillion seeds, just a few thousand), and if this probability is the same everywhere across the entire field, and if seeds are independent of each other, then the number of seeds per block, X, should follow a Poisson distribution:

$$P(X) = \frac{e^{-u} u^{X}}{X!}$$

In this equation, $P(X)$ is the probability of seeing X successes in a given block and μ is the mean number of occurrences. The constant $e = 2.718$ is the base of the natural logarithm.

A useful property of the Poisson distribution is that the mean and variance of the number of events (X) in a block are equal: $\sigma^2 = \mu$. Thus when sampling from a Poisson distribution, the sample variance to sample mean ratio, often called the coefficient of dispersion ($s^2/\overline{X}$) should be close to 1. Sampling from a distribution other than the Poisson should lead to a value of $s^2/\overline{X}$ less than 1 or greater than 1. The variance to mean ratio is therefore a useful index of "randomness". For example, if events were evenly spaced among blocks, then the variance/mean ratio would be less than 1. More commonly, events are clumped (e.g., seeds are sticky and land in your field in groups) producing a variance to mean ratio greater than 1.

Using the Random Number Generator

To experiment with these distributions you will need to use the calculator functions of JMP IN. Start by opening a **new data table**. To use a column to illustrate a probability distribution, create a column and then add the desired number of rows (e.g., 20). Add rows by clicking the "▼" symbol to the left of the **Rows** label on the left side of the data table window. Then select the column (by double clicking at the top of the column, or clicking once at the top of the column and then choosing **Column Info** from the **Cols** pull-down menu). Click **New Property->Formula** in the window that pops up. Finally, click the **Edit Formula** button. This will open the JMP IN calculator window. This platform allows you to create complex formulas to produce the data for a random variable.

In the calculator window, click the **Random** option in the **Functions (grouped)** panel. This will produce a set of options that will allow us to generate random samples from a wide variety of probability distributions. To generate random numbers from the binomial distribution choose **Random Binomial**. This will generate a formula in the formula box. The two parameters of the binomial distribution will need to be specified here: number of trials (n) and the probability of success in any one trial (p). Click to highlight the first box and type in the number of trials (e.g., 10). Select the second box and type in the probability of success, p (e.g., 0.5). Now click the **Apply** button of the calculator window (don't close the window). Examine the data table to see the results. Each time you click the "Apply" button of the calculator window a new random sample is generated. Adding rows also generates new values.

The process for generating random numbers from the Poisson distribution is similar, except that the Poisson distribution requires only one parameter, the mean, μ (the calculator window will refer to this parameter as "lambda" instead).

Using the Calculator

The calculator window may be used to generate new variables (columns) that are functions of variables (columns) already present in the data set. For example, create a new data table having a single column (label it as *X*) with the following numbers entered in consecutive rows: 0, 1, 2, 3, 4, 5. Now create a new column containing the values of e^X where *e* is the base of the natural logarithm. To accomplish this you will need to create a second column, give it the property **New Property -> Formula** as before. Edit the formula in the calculator window. Choose **Functions -> Trancendental -> Exp** to start the formula. Select the box between the parentheses in the formula and then click on the column corresponding to the variable *X* in the **Table Columns** box of the calculator window. Click **Apply** and return to your data table. The new column will contain the values for e^X. Repeat these steps but calculate *X*! instead ("!" refers to the factorial function) using the **Functions -> Trancendental -> Factorial** function in the calculator window.

The calculator will also compute binomial probabilities. For example, create a new data table having a single column (label it "*X*") with the following numbers entered in consecutive rows: 0, 1, 2, 3, 4, 5. We will use this column to represent the number of successes in n=5 trials, and use the calculator to compute the binomial probability of each *X*. Create a second column beside the first column, and label it "P(*X*)". Double click the mouse at the top of this second column and give it the property **New Property -> Formula** as explained earlier. Finally, click the **Edit Formula** button to open the JMP IN calculator window. Choose **Functions -> Probability -> Binomial Probability** to start the formula. Specify the desired probability of success in any one trial, p (e.g., 0.5), and the number of trials, n (enter the value 5 for this example). In the last box of the formula, labelled "*k*", type the name of the first column you created (e.g., type *X* if this is what you called it). Click "OK" on the **Formula** window and **Column Info** windows to obtain your result.

Poisson probabilities can be computed similarly. Don't hesitate to experiment with other parameters and formulas.

Problems

1. Let's examine the distribution of the number of male offspring in families of a rats with litter sizes of 10 (i.e., the number of trials, n=10) in a hypothetical population. Assume that males and females occur with the same probability, so set p=0.5. Generate a random sample of 50 such families using the random number generator.

   a) Plot a histogram and boxplots of the number of males and describe the distribution (skewed, bimodal, uniform, normal etc.). Does this sample appear to be symmetric? Does it have any outliers?

   b) Note the mean and standard deviation of the random sample. How do they compare with the mean and variance of the population from which you obtained your sample?

   c) Click the "Apply" button in the calculator window to generate a new random sample of families. Plot a new histogram and compare it with the previous one. Are the two distributions identical? Why or why not?

d) In mammals, meiotic drive occurs when the Y chromosome is more successful than the X chromosome during sperm formation, with the result that more sons than daughters are produced after mating. Create a second column in the data table, and using the binomial random generator sample 50 families of n=10 offspring in which the probability of a male offspring in any one trial is p=0.90. Plot the histogram and boxplot for the new column of data. Describe the changes to the distribution of the sample from this new binomial population. How are the mean and variance of the distribution changed? Are these changes expected (i.e., calculate the new mean and standard deviation of the number of males in litters of size 10 in the population of families).

e) Add **950** more families to the two columns. What effect does this have on the shapes of the distributions? On the sample mean and variance? How much do your new sample estimates of mean and standard deviation (in the number of males in families of 10 offspring) differ from population values? Which sample sizes provides a more reliable estimate of population parameters?

f) Change the number of trials in the second column to 100 (keep p=0.90). How does this change the shape of the distribution?

2. Assume that the number of Asian Gypsy moths captured in traps set for several nights across the lower mainland follows a Poisson distribution with mean 0.5. Set up a new column in the data table (label it "moths") and use the random number generator to randomly sample trap counts.

   a) Produce a histogram and boxplots for the sample and describe its shape. Note the mean and standard deviation for this set of values.

   b) Edit the formula for your Poisson distribution under column info, and change the mean to 0.1. Produce a new histogram and boxplots, and describe how the shape of the distribution has changed.

   c) Increase the mean number of insects found per trap to 15 and describe the shape of the curve. Describe what happens to the shape of the distribution. What distribution is this starting to resemble?

   d) Record the means and standard deviations from parts (a) through (c), and compute variances by squaring the standard deviations. What is the approximate relationship between these values? What is the relationship between mean and standard deviation in the populations from which the samples were obtained?

3. 

   It is expected that global warming will lead to changes in the distributions of animals and plants whose current ranges are limited by temperature, and of other animals and plants that depend on these temperature limited species. A group of scientists recently published results of surveys of changes in the northern and southern range limits of butterfly species in different parts of the world (Parmesan et al., 1999. Poleward shifts in geographical ranges of butterfly species associated with regional warming. *Nature* 399:579–583). For butterflies in Britain, they found that over the previous 30

years, the northern limits had extended northward in 18 of 21 butterfly species sampled, whereas northern limits had moved southward in the other 3 species.

a) What should the null hypothesis be for a test of whether a consistent trend existed in the range changes of British butterflies over the past 30 years? What is the alternate hypothesis?

b) Use the binomial test to decide whether the data warrant rejection of the null hypothesis.

4. During the process of sperm and egg formation in most metazoans, deleterious mutations may occur that will be passed on to the next generation. Thus, offspring individuals of each new generation may carry zero, one, two, or more new mutations. In an experimental study of mutation accumulation in *Arabidopsis*, 60 offspring of a cross within an inbred line were screened for new mutations. The following results were obtained:

Number of new mutations	**Number of individuals**
0	25
1	22
2	9
3	3
4	1
>4	0
Total	**60**

a) Calculate the sample mean number of mutations per individual and the sample variance in the number of mutations.

b) If separate mutations are independent and the probability of a particular gene acquiring a new mutation is small and equal between all genes and all individuals, what probability distribution should the number of new mutations per individual be expected to have?

c) Using the formula for this distribution and the JMP IN calculator, calculate the probability that an individual possesses 0, 1, 2, 3, 4, and >4 new mutations (a simple way to calculate the probability of more than 4 new mutations is to sum the probabilities of 0–4 mutations and subtract this sum from 1. Assume that the mean number of new mutations per individual (μ) is exactly 0.8833. Multiply each probability by the sample size, 60, to get the expected number of individuals in the sample having 0, 1, 2, 3, 4, and >4 new mutations. Do these expected frequencies resemble the observed numbers [just examine them; no test is necessary here].