3 STATISTICAL DISTRIBUTIONS

Objectives

- Become familiar with properties of the normal distribution.
- Construct a frequency histogram of a trait for a population.
- Become familiar with properties of the binomial distribution.
- Become familiar with properties of the Poisson distribution.

INTRODUCTION

In your studies of ecology and evolution, you will very likely come across a variety of statistical distributions and their uses. If you haven't taken a course on statistics, learning about these distributions may seem like learning a foreign language. However, since they are so widely used in the sciences, it is important that you become familiar with the most common statistical distributions used in ecology and evolution. In this exercise, you will learn about three distributions: the **normal** (or **Gaussian**) distribution, the **binomial distribution**, and the **Poisson distribution**.

Normal Distribution

Let's start with some very basic concepts before introducing the normal distribution. In the biological sense, a **population** is a group of organisms that occupy a certain space and that can potentially interact with one another. In statistics the term population has a slightly different meaning. A statistical population is *the totality of individual observations about which inferences are made, existing anywhere in the world or at least within a specified sampling area limited in space and time* (Sokal and Rohlf 1981). Suppose you want to make a statement about the average height of humans on earth. Your statistical population would then include all of the individuals that currently occupy the planet earth. Usually, statistical populations are smaller than that. For example, if you want to make a statement about the size of a certain fish species in a local stream or pond, your statistical population consists of all of the fish currently occurring within the boundaries of a stream or pond. Other examples of statistical populations include a population of business firms, of record cards kept in a filing system, of trees, or of motor vehicles. By convention, Greek letters are used to describe the nature of a population. For example, the average height of humans on earth would be denoted with the Greek letter μ, and the variance in height would be denoted with the Greek letter $\sigma^{2,}$ and the standard deviation would be denoted as σ. (We will define these terms shortly.)

In practice, it would be very difficult to measure the heights of all the individuals on Earth or even to measure *all* the fish in a local pond. So, we **sample** from the population. A sample is a subset of the population that we can deal with and measure. The goal of sampling is to make scientific statements about the greater population from the information we obtain in the sample. Quantities gathered from samples are called **statistics**. Statistics are denoted by letters from the Latin alphabet (i.e., from the same alphabet we use for writing English). For example, the mean of our sampled population would be denoted by the Latin letter $\bar{x}$, the variance is denoted by S^2, and the standard deviation is denoted by S.

The most important pictorial representation of a set of data that make up a sample is called a **frequency distribution**. If we sampled plants in an area of interest and recorded their biomasses in grams, we could then construct a frequency distribution such as Figure 1 and examine the shape of our data. Biomass would go on the x-axis (on the bottom), and numbers of individuals of a certain biomass would go on the y-axis (the vertical axis).

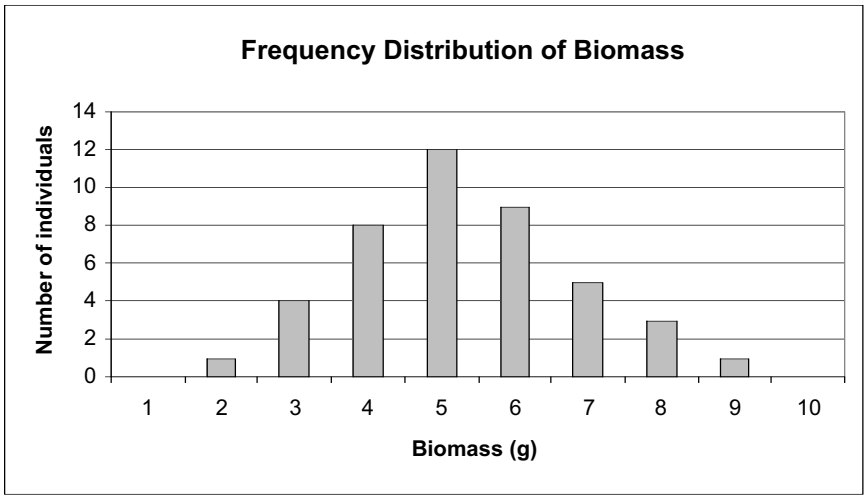

Figure 1

In published papers, you rarely see frequency distributions because they take up too much space in print, and they usually provide more information than a reader needs. Instead, ecologists and evolutionary biologists often report two kinds of summary statistics: (1) measures of **central tendency** (average value, middleness), and (2) measures of **dispersion** (how spread or dispersed the raw data are). Examine Figure 1. How would you characterize the "average plant" in terms of biomass? There are three common measures of central tendency: the mean, the mode, and the median. The **mean**, denoted by $\bar{x}$, is simply the arithmetic average: sum up the total biomass and divide by the number of individuals in the sample.

$$\bar{x} = \frac{\sum_{i=1}^{N} x}{N} \qquad \text{Equation 1}$$

If our sample consisted of the values 4, 6, 10, and 12, those values represent the little x's in equation 1, and $N = 4$ since there are four values in the sample. The average is (4 + 6 + 10 + 12) divided by 4. In Figure 1, the average is 4.3 grams of biomass. The **mode** is the most frequently occurring value. It is the high point of the frequency distribution. In our example, 5 is the mode since this value occurs 12 times. The **median** is the middle number in a data set when the samples are ordered. For example, if our sample consisted of the values 1, 3, 4, 6, and 10, the median would be 4 because it is the middle value. If the data set consisted of an even number of observations, then the median is the average of the two middlemost numbers.

Now let's consider the spread of the data in Figure 1. How can we characterize this spread? One way is to record the range of values the data assume. The lowest observed biomass was 2 grams, and the highest observed biomass was 9 grams. The range of biomass for our sample then is 9 – 2, or 7 grams. The data points at the extremes really affect the range, so it is not a very stable estimate of variability. A second method, called **average error**, describes how far each data point is, on average, from the mean. It is calculated as

$$\frac{\sum\left(x-\bar{x}\right)}{N} \qquad \text{Equation 2}$$

However, because some scores will fall above the mean, and others will fall below it, this sum will always be 0! How can we overcome this problem? By squaring the deviations from the mean, and by subtracting 1 from the total sample size, we end up a definition of **variance**, or S^2:

$$S^2 = \frac{\sum(x-\bar{x})^2}{N-1} \qquad \text{Equation 3}$$

Thus, variance can be defined as (almost) the average squared deviation of scores from the mean. This is a very useful way of describing the spread of data in a given data set. However, all of the units have now been squared (e.g., biomass^2). To get rid of the squaring, we take the square root of both sides and arrive at the equation for computing the **standard deviation** of a sample, or S.

$$S = \sqrt{\frac{\sum(x-\bar{x})^2}{N-1}} \qquad \text{Equation 4}$$

With this background, we can now proceed to talk about the **normal distribution**. This distribution is one of the most familiar in statistics. Let us first return to a statistical population, rather than a sample. For a normally distributed trait, the frequency of distribution takes on a bell-shape that is completely symmetrical and has tails that approach the x-axis. The shape and position of the normal curve is determined by both the mean (μ) and the standard deviation (σ): μ sets the position of the curve along the x axis, while σ determines the spread of the curve. Two normal curves are shown in Figure 2. They have different μ but the same σ; thus they are similar in shape but are positioned in different locations along the x-axis.

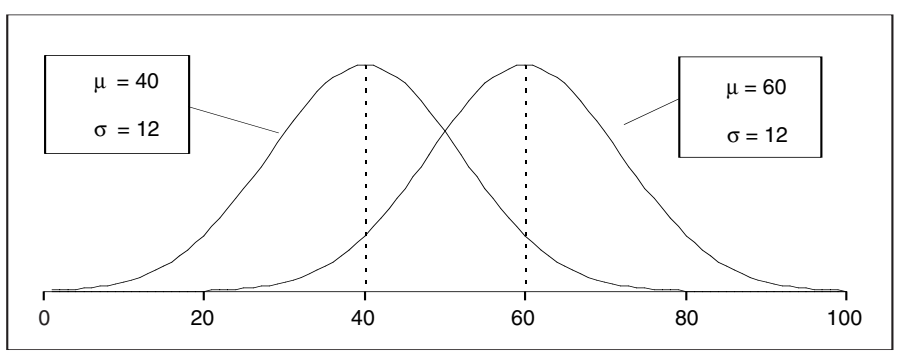

Figure 2

The standard deviation determines the spread of the normal curve. Figure 3 shows two normal curves with the same μ, 40, but different σ. Note that when σ is small, most of the data are distributed close to the mean, and when σ is large, the curve "flattens out" because the data vary more from the mean.

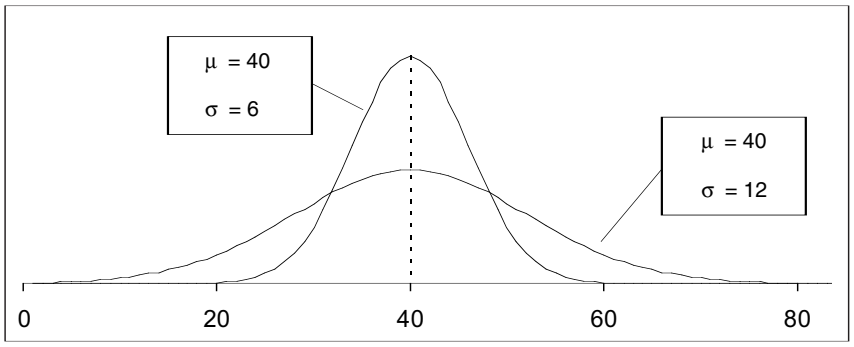

Figure 3

A property of normal curves is that the total area under the curve is equal to 1. (This is true of all probability models or models of frequency distributions.) Another property is that the most of the data fall in the middle of the curve around the mean. Normal distributions are completely symmetrical about the mean, and the mean equals the median and the mode. For normal distributions, approximately 68% of the observations will fall between the mean and plus or minus 1 standard deviation. If we assume, for example, that the mean length for a population of seeds is 10 mm and that *S* is 1.0, and if we assume that seed length is normally distributed, then 68% of the seed length values will fall between 9 and 11 mm (i.e., the mean, 10 mm, plus or minus 1.0, which is 1 standard deviation). And approximately 95% of the observations will fall between the mean and plus or minus 2 standard deviations. These properties make it possible to compute the specific probability that, for example, a seed of 8 mm length will be sampled from the population.

Figure 4 shows that, for a population with a mean of 10 and a standard deviation of 1.0, this probability is 0.054. This probability was computed in Excel with the NORMDIST function. The probability of sampling a seed of 10 mm length is 0.4. The **cumulative probability** gives the probability of sampling a seed of a certain size or less. For example, the probability of sampling a seed of *at least* 10 mm is 0.5. As you can see, with the parameters given, the cumulative probability is 1 when the seed length is 13 mm. This means that there is a 100% chance of sampling a seed of 13 mm or less, given that the population has a mean length of 10 mm and a standard deviation of 1.

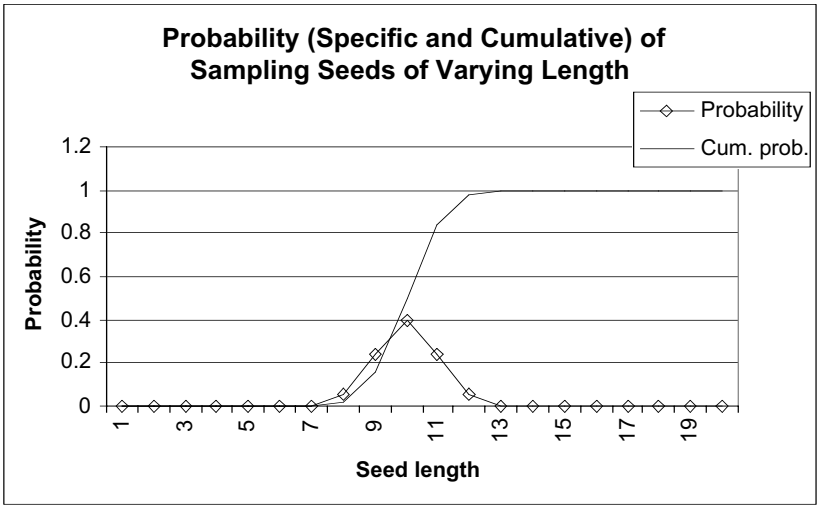

Figure 4

If we change the standard deviation to 3 mm, and keep the mean at 10 mm, the probabilities will be different (Figure 5).

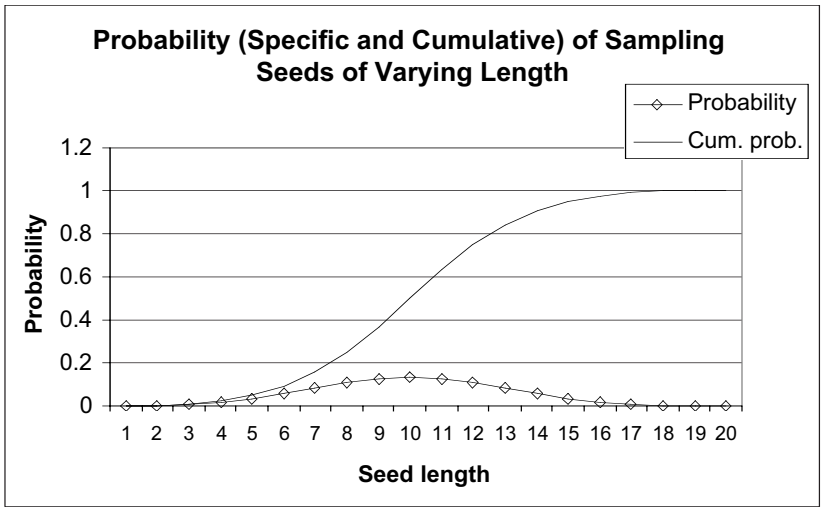

Figure 5

Knowledge about the normal distribution is important to ecologists because many statistical procedures, such as a *t* test, assume that the sampled data are normally distributed. These properties are handy from a modeling perspective; in many of the exercises in this book, we will "draw" samples from a normal distribution whose mean and standard deviation are specified.

Binomial Distribution

Some situations in ecology are binary: There are only two possible outcomes. For example, suppose we are tracking the fates of individuals over time and are interested in the number of deaths. During this period, there are only two outcomes: death or survival. In this situation, a binomial distribution can be used to describe the relative number of times that a particular event will occur (death) among groups of observations. Another example may be the relative numbers of trees in flower among a series of samples of a particular size. The **binomial distribution** is used when a researcher is interested in the *occurrence* of an event, not in its magnitude. The binomial distribution describes, for instance, the relative numbers of individuals that flower, not how well they flower.

The binomial distribution is specified by the number of observations, n, and the probability of occurrence, which is denoted by p. Here are some things to keep in mind when using the binomial distribution:

- Each outcome must be classified as a "success" (the type of outcome that we're interested in) or as a "failure."
- Since we're dealing with a count of successes, this probability distribution is discrete. (The x-axis is the number of successes, and it cannot be a fraction).
- Each trial is independent. The probability of success (p) and the probability of failure ($1 - p$) is the same for each trial. Thus, if one tree in your sample has fruits, you don't know anything about the next sample, other than it has a probability p of having fruit.

The formula for calculating the probability of x successes out of n trials of a binomial experiment, where the probability of success on an individual trial is p, is

$$f(x) = {}_nC_x \times p^x \times (1 - p)^{n-x} \qquad \text{Equation 5}$$

In this equation, p is the probability of success and its exponent, x, is the number of successes. The probability of failure is $1 - p$, and its exponent, $n - x$, is the number of failures. The term ${}_nC_x$ means "out of n samples, let x succeed." This gives the number of ways of choosing x distinct items from a set of n items, and it is calculated as

$$ {}_nC_x = \frac{n!}{x! \times (n-x)!} \qquad \text{Equation 6} $$

Recall that a factorial, such as $n!$, is calculated by multiplying all the integers (whole numbers) from 1 up to and including n.

For example, assume the probability of surviving is 0.1. If we have a population of 5 individuals, what is the probability that exactly 3 individuals will survive? The success in this problem is an individual that survives. The failure is an individual that dies. We know that $p = 0.1$. This also tells us that $1 - p = 0.9$. Since our population consists of 5 individuals, $n = 5$. And we are specifically interested in knowing the probability that 3 individuals will succeed, so $x = 3$. First, let's compute ${}_5C_3$. It is

$$ \frac{5!}{3! \times 2!} = \frac{5 \times 4 \times 3 \times 2 \times 1}{(3 \times 2 \times 1)(2 \times 1)} = \frac{20}{2} = 10 $$

We can compute the binomial probability that exactly 3 of 5 individuals will survive when $p = 0.1$ as

$$ \begin{aligned} f(3) &= {}_5C_3 \times (0.1)^3 \times (0.9)^2 \\ &= (10) \times (0.001) \times (0.81) \\ &= 0.0081 \end{aligned} $$

The probability that exactly 3 of these 5 individuals survive is 0.0081. Similarly, the probability that 0, 1, 2, 4, and 5 individuals survive could be calculated (rather easily with the **BINOMDIST** function). We can graph these binomial probabilities as shown in Figure 6.

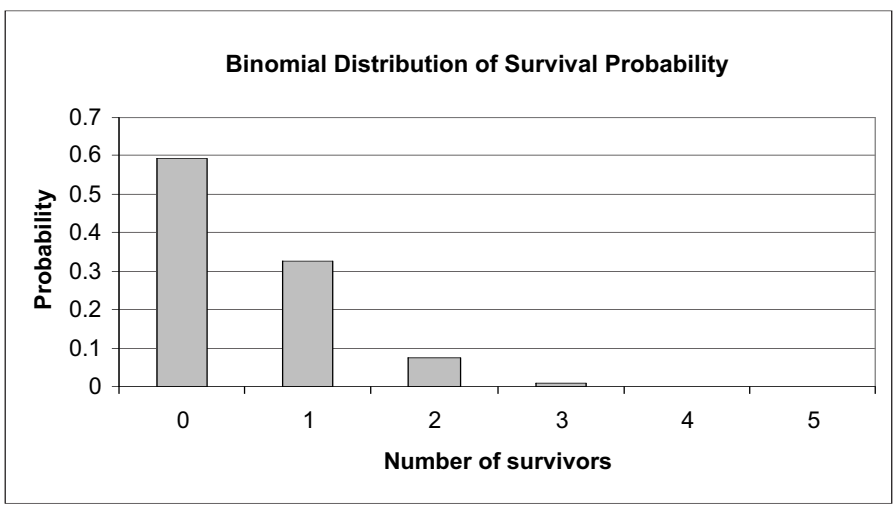

Figure 6

If we change our survival probability to 0.7, our binomial distribution of probabilities will differ, as shown in Figure 7. As with the normal distribution, we could also plot the *cumulative* probability that *at least x* number of individuals survive.

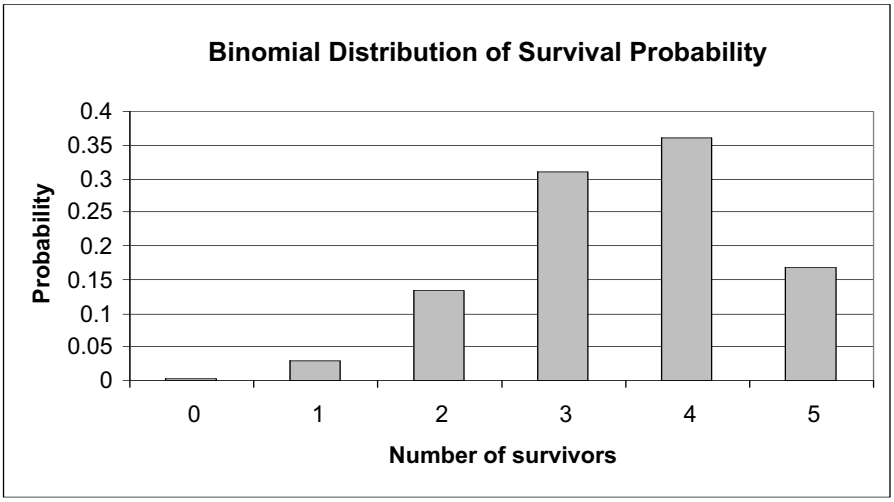

Figure 7

Poisson Distribution

The **Poisson distribution** is similar to the binomial distribution in that the number of events is counted. However, the events are not limited to two outcomes. For example, ecologists may be interested in the number of birth events in a given period of time. The Poisson distribution is a mathematical rule that assigns probabilities to the number occurrences. The French mathematician Poisson derived this distribution in 1837, and evidently its first application was the description of the number of deaths in the Prussian army due to horse kicking (Bortkiewicz 1898). The only thing we need to know to specify the Poisson distribution is the mean number of occurrences, such as the mean number of births. Contrast this to the binomial distribution, in which both the probability that an event will occur and the total number of individuals in the population must be known. For example, in the binomial distribution all individuals are studied to see whether they had survived or not, whereas using the Poisson distribution only the individuals that survived are studied.

The formula for calculating the Poisson probability is

$$f(x) = \frac{\lambda^x \times e^{-\lambda}}{x!} \qquad \text{Equation 7}$$

where λ is the mean number of successes in a given period of time, x is the number of successes we are interested in, and e is the natural logarithm constant (approximately 2.718). As an example, suppose the average number of offspring produced per individual in a population is 2.1; what is the probability that an individual will have exactly 4 offspring? The probability would be calculated as

$$f(4) = \frac{2.1^4 \times e^{-2.1}}{4 \times 3 \times 2 \times 1} = 0.0992 \qquad \text{Equation 6}$$

We could calculate the probability that exactly 0, 1, 2, 3, 5, 6, 7, … offspring were produced, given the average, with the **POISSON** spreadsheet function. Our Poisson distribution is shown in Figure 8.

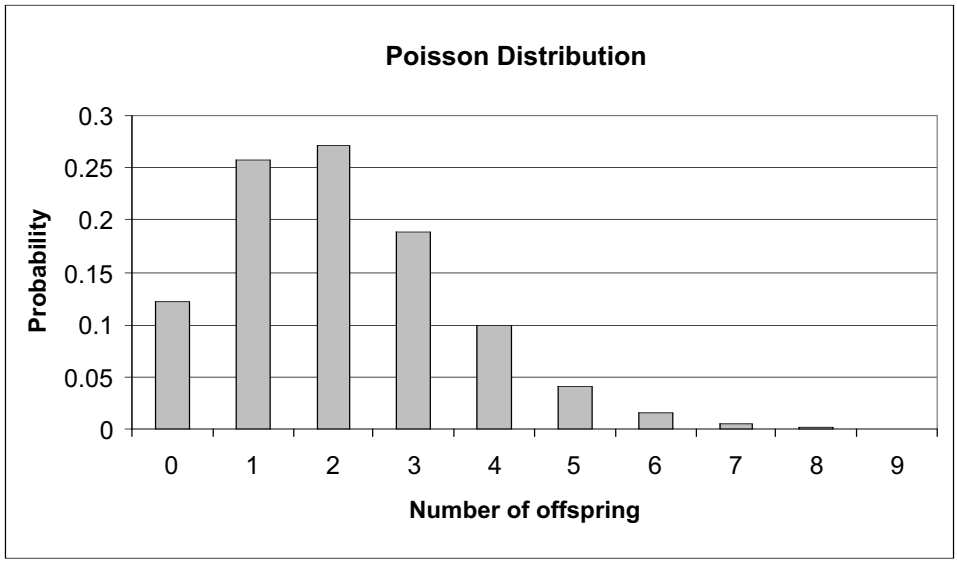

Figure 8

In this exercise, you'll use a spreadsheet to explore properties of the normal, binomial, and Poisson distributions. As always, save your work frequently to disk.

INSTRUCTIONS

ANNOTATIONS

A. Set up the spreadsheet for normal distribution.

We will start our exercise by investigating properties of the normal distribution, and we will compare a trait (height, for example) between two different populations, each consisting of 30 individuals.

1. Open a new spreadsheet and set up column headings as shown in Figure 9.

	A	B	C	D	E	F	G
1	***Normal Distribution***						
2							
3		**Mean**	**Std**				
4	Population 1	50	5				
5	Population 2	30	5				
6							
7	Individual	Pop 1	Pop 2		**Frequency distribution**		
8	1				Bins	Pop 1	Pop 2

Figure 9

2. Set up a linear series from 1 to 30 in cells A8–A37.

Enter 1 in cell A8.
Enter **=1+A8** in cell A9. Copy this formula down to cell A37.

Next we will assign a height to each of the 30 individuals in population 1, drawn from a normal distribution with a mean given in cell B4 and a standard deviation given in cell C4. We don't really have individuals to measure, of course, but the **NORMINV** function allows us to simulate this. The **NORMINV** function consists of three parts, each separated by a comma. It has the form **NORMINV(probability,mean,standard_dev)** where **probability** is a probability (from 0 to 1) corresponding to the cumu-

lative normal distribution, **mean** is the arithmetic mean of the normal distribution, and **standard_dev** is the standard deviation of the distribution.

3. In cell B8, enter a **NORMINV** formula to generate a random height for an individual in population 1. Copy your formula down to cell B37.

In cell B8, enter the formula **=ROUND(NORMINV(RAND(),B4,C4),1)**. Copy this formula down to cell B37.

The formula =**NORMINV(RAND(),B4,C4)** tells the spreadsheet to draw a random cumulative probability between 0 and 1 (the **RAND()** portion of the formula) from a normal distribution that has a mean given in cell B4 and a standard deviation given in cell C4. The formula returns the inverse of this probability; it changes the cumulative probability into an actual number from the distribution. Excel will return a value, which is the height of the individual. You'll note that this formula is embedded within a **ROUND** formula, which consists of two parts that are separated by a comma. The first part is the number that should be rounded **(NORMINV(RAND(),B4,C4**), and the second part is the number of decimal places to which the number should be rounded. Note that when you press F9, the calculate key, the spreadsheet will generate a new random number, and hence will generate a new cumulative probability and height for individual 1 in Population 1.

4. In cells C8–C37, enter a formula to generate a random height for an individual in Population 2.

Enter the formula **=ROUND(NORMINV(RAND(),B5,C5),1)** in cell 8. Copy your formula down to cell C37.
Note that the references to the mean and standard deviation are absolute cell references (indicated by the dollar signs), so that when you copy the formula down to cell C37 the heights will be drawn from a distribution whose mean and standard deviation are fixed in cells B5 and C5.

5. Save your work.

B. Construct the frequency distribution.

The most common way to depict a population's values is as a frequency distribution. A frequency distribution is a plot of the raw data, in this case height, against the frequency that each value appears in the population.

We will use the **FREQUENCY** function to generate a frequency distribution of heights for Population 1 and Population 2. This formula is a bit tricky, so pay attention to these instructions. The **FREQUENCY** function calculates how often values occur within a range of values.

1. In cell E9, enter the number 5. In cell E10, enter **=5+E9**. Copy this formula down to cell E28.

Use the **FREQUENCY** function to count the number of heights that fall 5 mm or lower, within 6 and 10 mm, within 11 and 15 mm, and so on. These groupings are called "bins." The bins may be very small (hold only a few numbers) or very large (hold a large set of numbers). Our bins will cluster heights in groups of 5 mm. The bin labeled 5 (cell E9) will "hold" heights up to and including 5 mm (0, 1, 2, 3, 4, and 5 mm). The bin labeled 10 (cell (E10) will "hold" heights from 6 to 10 mm, and so on.

2. Use the **FREQUENCY** function in cells F9–F28 to compute frequencies of heights for Population 1.

The **FREQUENCY** returns an array of values (in our case the values will be in cells F9–F28), it must be entered as an array formula, which is a bit different than the normal formula entries. It has the syntax **FREQUENCY(data_array,bins_array)**, where **data_array** is the set of values for which you want to count frequencies (heights), and **bins_array** is the array of intervals into which you want to group the values in data_array. You can think of a bin as a bucket in which specific numbers go.

The **FREQUENCY** formula works best when you use the f_x button and follow the cues for entering a formula. Since you will be entering this formula for an array of cells, the mechanics of entering this formula are a bit different than the typical formula entry. *Instead of selecting a single cell to enter a formula, you need to select a series of cells, enter a*

formula, and then press <Control>+<Shift>+<Enter> (Windows) to enter the formula for all of the cells you have selected.

Let's try it to determine the frequencies of heights for Population 1. Select cells F9–F28 with your mouse, then use your f_x button and select the **FREQUENCY** function. (If it doesn't show up in the list of most recently used functions, you will have to view the list of all functions.) To define the data array, use your mouse to highlight the heights of all 30 individuals in Population 1 (cells B8–B37). To define the bins array, select cells E9–E28. Next, instead of clicking "OK," press <Control>+<Shift>+<Enter> to return your height frequencies. After you've obtained your results, examine the formula in cells F9–F28. Your formula should look like this:

{ =FREQUENCY(B8:B37,E9:E28)}

The { } symbols indicate that the formula is part of an array. If for some reason you get "stuck" in an array formula, press the Escape key and start over.

3. Use the **FREQUENCY** function in cells G9–G28 to compute frequencies of heights for Population 2.

Follow steps 1 and 2. Your formula should be **{=FREQUENCY(C8:C37,E9:E28)}** in cells G9–G28.

4. Graph the frequencies of Population 1 and Population 2.

Use the column graph option and label your axes fully. Your graph should resemble Figure 10.

5. Save your work.

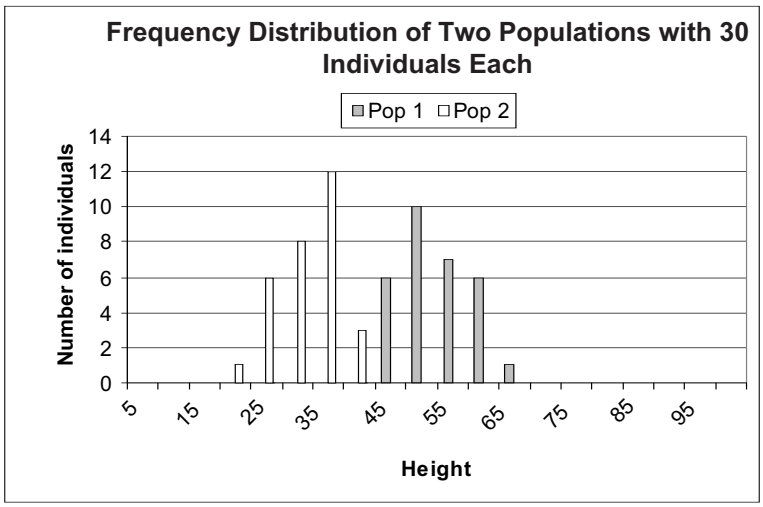

Figure 10

C. Compute statistics.

1. Set up new spreadsheet headings as shown in Figure 11.

	I	J	K
7		***Pop 1***	***Pop 2***
8	Mean		
9	Median		
10	Mode		
11	Standard deviation		
12	Minimum		
13	Maximum		
14	Range		
15	Count		

Figure 11

2. Enter formulae to compute measures of central tendency: the mean, median, and mode height for Populations 1 and 2 in cells J8–K10.

Use the f_x button to guide you through the formulae. Your results should be

- J8 **=AVERAGE(B8:B37)**
- J9 **=MEDIAN(B8:B37)**
- J10 **=MODE(B8:B37)**
- K8 **=AVERAGE(C8:C37)**
- K9 **=MEDIAN(C8:C37)**
- K10 **=MODE(C8:C37)**

If Excel cannot find a most commonly occurring number (i.e., if there is no mode), it will return **#N/A**.

3. Enter formulae in cells J11–K14 to compute measures of dispersion: standard deviation, minimum, maximum, and range.

Use the f_x button to guide you through the formulae. Your results should be:

- J11 **=STDEV(B8:B37)**
- J12 **=MIN(B8:B37)**
- J13 **=MAX(B8:B37)**
- J14 **=J13-J12**
- K11 **=STDEV(C8:C37)**
- K12 **=MIN(C8:C37)**
- K13 **=MAX(C8:C37)**
- K14 **=K13-K12**

4. Enter a formula in cells J15–K15 to count the sample size of each population.

Enter the formulae:

- J15 **=COUNT(B8:B37)**
- K15 **=COUNT(C8:C37)**

5. Save your work, and answer Questions 1–4 at the end of the exercise. Your spreadsheet should now resemble Figure 12, although your numbers will be different. Each time you press the F9 key to generate new heights, the statistics for each population will be automatically updated.

	I	J	K
7		***Pop 1***	***Pop 2***
8	Mean	50.4	29.4
9	Median	50.0	30.2
10	Mode	52.6	31.6
11	Standard deviation	5.3	4.7
12	Minimum	40.1	18.8
13	Maximum	61.2	38.1
14	Range	21.1	19.3
15	Count	30.0	30.0

Figure 12

D. Set up the binomial distribution spreadsheet.

1. Click on Sheet 2 and set up new headings as shown in Figure 13.

	A	B	C	D	E	F	G
1	***Binomial and Poisson Distributions***						
2							
3	Probability of survival =		0.5				
4	Mean number of offspring =		20				
5	Number of individuals =		30				
6							
7	**Binomial**				**Poisson**		
8	# Survivors	Probability	Cum. prob.		# Offspring	Probabity	Cum. prob.

Figure 13

2. Set up a linear series from 0 to 30 in cells A9–A39.

First, we will consider the number of survivors over a period of time in a population that again consists of 30 individuals. There are only two outcomes for each individual (survive or die), which makes survival probabilities an appropriate use of the binomial distribution. We will consider the probability that 0, 1, 2, …, 30 individuals will survive the period with a binomial distribution, given that the survival probability is 0.5 (cell C3) and that there are 30 individuals (cell C5).

Enter **0** in cell A9.
Enter **=1+A9** in cell A10. Copy this formula down to cell A39.

3. In cells B9–B39, enter a formula to calculate the probability that the exact number of individuals given in cell A9 will survive.

In cell B9, enter the formula
=BINOMDIST(A9,C5,C3,FALSE). Copy this formula down to cell B39.

The **BINOMDIST** function returns the probability of success (survival) from the binomial distribution, given the number of trials (the number of individuals in the population) and the probability of success (the probability of survival). This function consists of four parts, each separated by a comma. The first part is the number of individuals in the population, the second part is the number of survivors in the population, the third part is the probability of survival for the whole population, and the fourth part tells the spreadsheet whether you want the binomial probability to be a cumulative probability (e.g, the probability that there will be *up to but not more than* 15 survivors), or simply the probability that a given number of individuals will survive (e.g., the probability that 4 out of 30 individuals in the population will survive). The word "True" returns the cumulative probability, while the word "False" returns the specific probability.

For example, the formula in cell B9 returns the binomial probability that there will be 0 survivors (cell A9) when the population consists of 30 individuals (given in cell C5) and the average survival probability is 0.5 (given in cell C3). The **FALSE** part of the formula indicates that the program should return the probability for the exact number of survivors, not the cumulative probability.

4. Enter a formula in cell C9 to calculate the cumulative probability that no more than the number of individuals given in cell A9 will survive.

In cell C9, enter the formula **=BINOMDIST(A9,C5,C3,TRUE)**. Copy this formula down to cell C39.

This formula is identical to the one just entered in cells B9–B39, except that the last part of the formula is **TRUE**, indicating that the program should return the cumulative probability, or the probability that there will be *up to* a certain number of survivors.

Use the column graph option and label your axes fully. You could also use the Scatterplot graph option if you prefer.

5. Graph the probability of survival against the number of survivors (cells B9–B39).

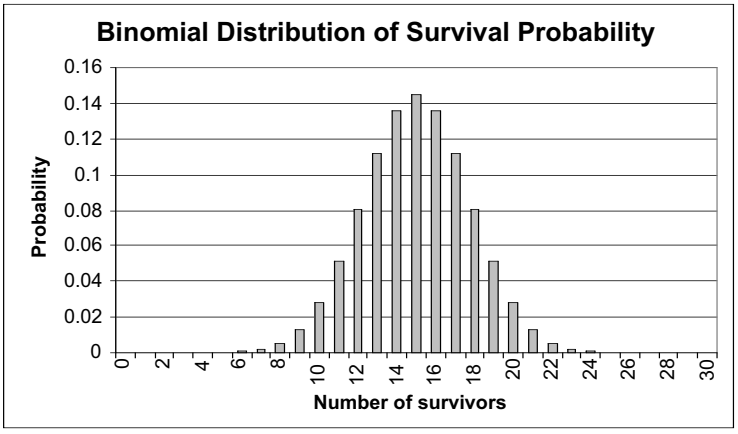

Figure 14

6. Graph the probability of survival, and the cumulative probability of survival, against the number of survivors (cells B9–C39).

7. Save your work.

Use the column graph option and label your axes fully. Your graph should resemble Figure 15.

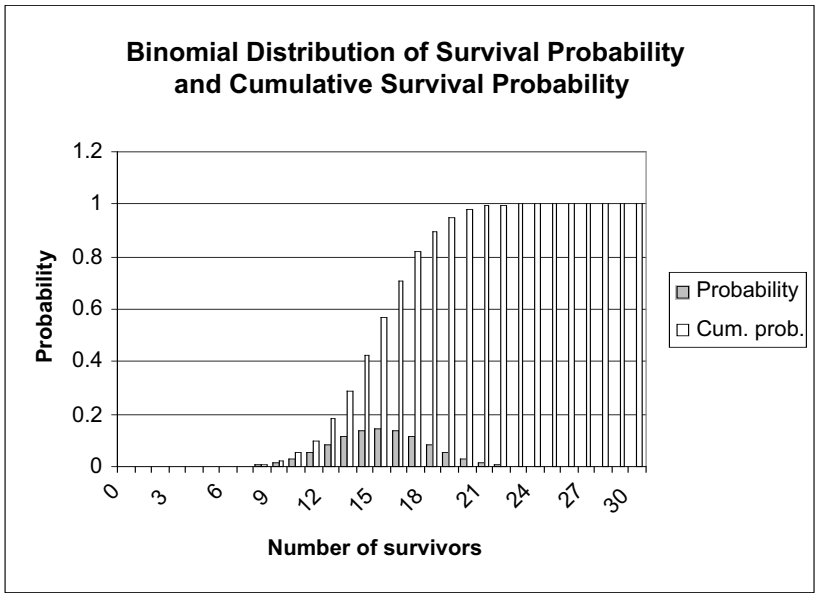

Figure 15

E. Set up the Poisson distribution spreadsheet.

Now we will consider the number of births over a period of time in a population that once again consists of 30 individuals. For this exercise, we will assume that there are between 0 and 30 births possible. Because there are several discrete numbers of births possible, this analysis is an appropriate use of the Poisson distribution. We will consider the probability that 0, 1, 2, …, 30 individuals will be born during a time period of interest, given that the average number of offspring for the population is 20 (cell C4).

1. Set up a linear series from 0 to 30 in cells E9–E39.

Enter 0 in cell E9.
Enter **=1+E9** in cell E10. Copy this formula down to cell E39.

2. In cell F9, enter a formula to calulate the probability that the exact number of young given in cell E9 will be born. Copy this formula down to cells F10–F39.

In cell F9, enter the formula **=POISSON(E9,C4,FALSE)**. Copy this formula down to cell F39.

Cell F9 uses the **POISSON** function to give the probability that a certain number of young will be born, given the average number of young born per period of time. This function has three parts, each separated by a comma. This first part gives the number of young born (e.g., 0 young, cell E9). The second part gives the expected number of young born (cell C4). The third part, like the **BINOMIAL** function, indicates whether you want the cumulative probability (e.g., the probability that up to 8 young will be born) or the probability that a specific number of young are born (e.g., the probability that exactly 10 young will be born). **FALSE** returns the exact probability, whereas **TRUE** returns the cumulative probability.

3. In cell G9, enter a formula to calulate the probability that no more than the number of young given in cell E9 will be born. Copy this formula down to cells G10–G39.

In cell G9, enter the formula **=POISSON(E9,C4,TRUE)**. Copy this formula down to cell G39.

4. Graph the number of offspring and the Poisson probability (exact, cells E9–F39). Use the column graph and label your axes fully (Figure 16). You may also use the Scattergraph option if you prefer.

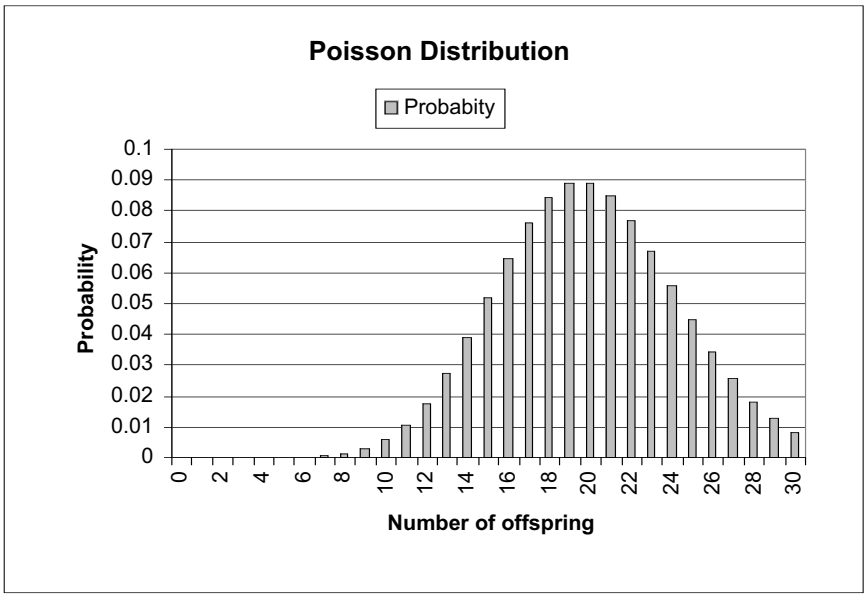

Figure 16

5. Graph the number of offspring and the cumulative Poisson probability (cells G9–G39). Use the column graph and label your axes fully (Figure 17).

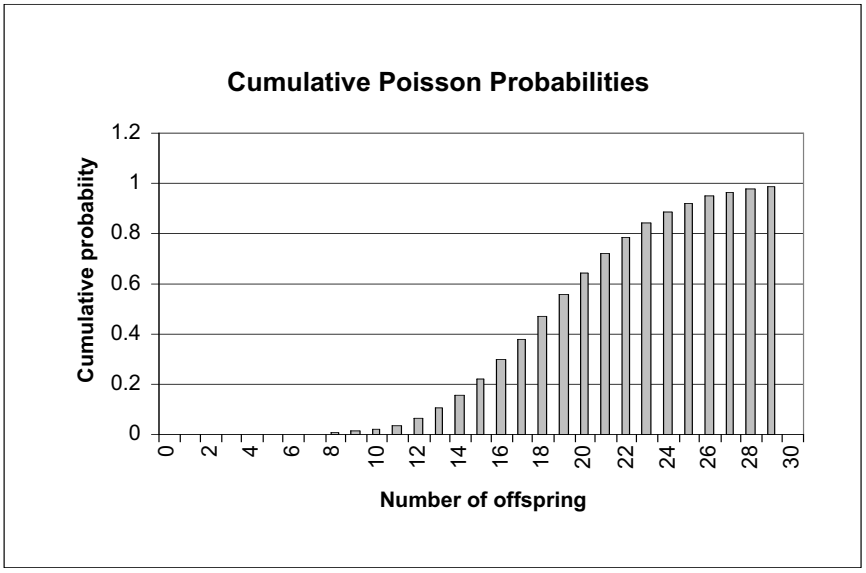

Figure 17

6. Save your work, and answer the remaining questions at the end of the exercise.

QUESTIONS

1. What parameter controls the location along the *x*-axis of the data in your frequency distribution? Change the value in cell B4 (Population 1, try several values) and examine your distribution.
2. What parameter controls the spread of the data in your frequency distribution? Change the value in cell C4 (Population 1, try several values) and examine your distribution. What happens when this value is almost 0 (0.0001)?
3. One of the properties of the normal distribution is that the mean, mode, and median are equal. Why might this not be the case in your spreadsheet? How could you increase the chances that the mean, mode, and median would be equal?
4. Assume that instead of heights, we are comparing the annual salaries (in thousands of dollars) of 30, randomly selected individuals. Set up cell values as shown:

	A	B	C
3		**Mean**	**Std**
4	Population 1	50	5
5	Population 2	50	5

Furthermore, assume that Bill Gates is part of our sample in Population 1, and his salary is entered in cell B8. Enter 1000 in cell B8 (overwrite the formula in that cell). Assess which measure of "middleness" is the most appropriate descriptor of average salaries.

5. Assume you are a biologist working on a mark-recapture study of a population of salmon, and you have tagged 20 salmon. You estimate that 50% of the salmon will survive to the time set for recapture. What is the probability that *exactly* 10 of the marked salmon are still alive when it is time to recapture? What is the probability that *up to* 10 of the marked salmon are still alive?
6. How do your answers from Question 5 change if the survival estimate is 30%?
7. Set cell C3 to 0.5. Change the value in cell C5, starting with 0, and increase by twos up to 20. How does changing cell C5 (n) affect the location and shape of the binomial distribution?
8. How does changing cell C4 (λ) affect the location and shape of the Poisson distribution? Change the value in cell C4, from 0 to 10, in increments of 1. As λ increases, what kind of shape does the Poisson distribution take?

LITERATURE CITED

Sokal, R. R. and F. J. Rohlf. 1981. *Biometry*, 2nd Ed. W. H. Freeman and Company, New York.