Do These Numbers Mean Anything?

A Beginner's Guide to Interpreting Biological and Ecological Data

data collected June, 2005				
plot id number	number of flowering trees per plot	number of fruiting trees per plot	number of bellbirds per plot	number of quetzals per plot
1	1	1	0	1
2	4	10	6	0
3	2	9	5	0
4	7	1	3	0
5	1	2	0	1
6	2	2	1	2
7	3	2	2	0
8	6	6	4	4
9	2	7	7	2
10	3	4	2	1
11	5	3	0	0
12	1	3	1	0
13	8	3	3	0
14	7	3	4	2
15				2
16				1
17				0
18				0
19				1
20				0
21				1
22				1
23				0
24				5
data collected July, 2005				
1				2
2				0
3				1
4				0
5				1
6				3
7	5	1	3	0
8	5	9	3	1
9	1	5	6	0
10	1	5	5	0
11	1	2	1	2
12	3	2	1	1
13	2	1	1	1
14	5	1	4	0
15	1	6	5	1
16	1	7	4	0
17	1	3	3	0
18	3	3	5	1
19	4	2	5	0

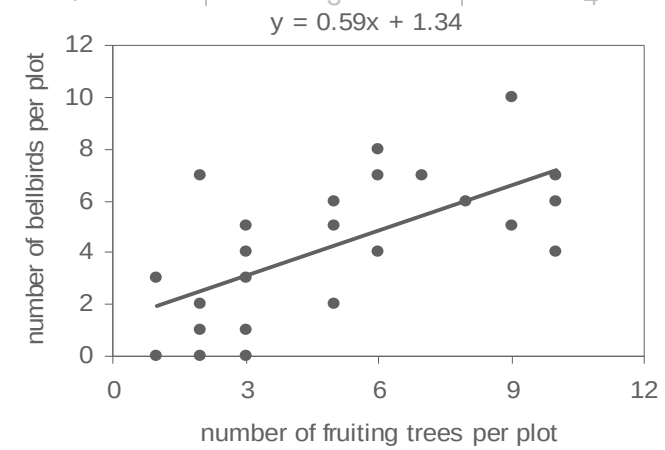

Figure 1. The relationship between the number of bellbirds and the number of fruiting trees in twenty-four 20m x 50m forest plots sampled in June, 2005. Equation and line are from a regression analysis.

by using MS Excel to display and analyze data

Do These Numbers Mean Anything?
A Beginner's Guide to Interpreting Biological and Ecological Data

Rhine Singleton

Departments of Biology
& Environmental Science
Franklin Pierce College

Acknowledgements, July 2005

I would like to thank several colleagues, friends and family members. From the beginning, Catherine Koning, Fred Rogers and Jacques Veilleux encouraged the idea of writing a guide to statistics for undergraduates. The enthusiasm and creative suggestions of Michael Lehner were great sources of motivation while I was writing. April, Jordan and Isaiah were supportive of my efforts even when this manual was a source of distraction. Most importantly, I want to thank Jordan for his thorough reading and insightful comments that helped improve and clarify the text.

Table of Contents

Preface

This manual is for the student who has little or no training in the use of spreadsheets or statistical analysis but has the need or the desire to make sense out of quantitative data. It should be helpful for anyone wishing to analyze relatively straightforward data sets but who finds formal statistics texts "user unfriendly." It is intended to provide the beginner with the tools to start to interpret data without using sophisticated statistical software or complicated equations.

I believe that students can understand many of the key concepts in statistics without learning a lot of statistical theory. I also think that it is possible to successfully perform and interpret statistical tests while not knowing the formulas behind those tests. As my friend Michael Lehner says, a carpenter doesn't need to know how a hammer is made to successfully bang a nail.

In the writing of this manual, I have attempted to make the interpretation of quantitative data as simple as possible without glossing over or over-simplifying the most important concepts. Students may be dismayed to find that even in its simplest form, statistics is a tricky topic with a certain amount of technical language that cannot be eliminated. Statisticians may be dismayed to find that equations are avoided as much as possible, discussion of null hypotheses is left for an appendix, and discussion of assumptions of particular tests is minimized. Nonetheless, I believe that this manual is at an appropriate level for a beginner to start using statistics.

Caveat

It is essential to stress that this manual is not intended as a substitute for a statistics text. Anyone who wants to be sure they have interpreted their data correctly, particularly if they intend to present their analyses to a scientific audience or to publish their work in a professional research journal, should consult a statistics text (see references cited) and/or a statistician.

A Note on the Organization of Chapters

Each chapter includes a realistic example of a research question or hypothesis along with a hypothetical data set. I purposely chose hypothetical data in order to keep the data sets small and to ensure that they illustrate important concepts. Each chapter also includes sections with detailed directions for using MS Excel to make graphs and perform statistical analyses. These directions are distinguished from the rest of the text by use of a different font. After these directions, guidelines for interpreting the graphs and statistical output are included.

It is worth noting that Chapter 1 is the longest and assumes the least knowledge of MS Excel or statistics. Though a reader does not have to read Chapter 1 in order to understand Chapters 2 – 4, a true beginner with spreadsheets will have an easier time with the later chapters after a thorough reading of Chapter 1. On the other hand, those with more experience with spreadsheets and statistics should be able to skim chapters for the information they desire.

Jump in and start analyzing!

It is my hope and belief that a beginner can grasp many of the most important concepts required for basic statistical analysis. One of the best ways to begin this understanding is to jump right in and make some graphs and do some statistical tests. This experience can be a crucial step towards clear thinking about the interpretation of quantitative data.

Introduction

A Few Critical Concepts

Although this manual is designed for the reader with little or no formal training in statistics, it is impossible to discuss basic statistical analysis without first covering some critical concepts. Several are discussed briefly below and reviewed and expanded periodically throughout the manual. In addition, the glossary provides concise definitions for many of the terms presented here and in the following chapters.

Any collection of quantitative data (group of numbers) can be summarized based on **location** and **spread.**

Location of Data – Measures of location summarize where most of the data are found. The **mean** (= average or arithmetic mean), **mode** and **median** are all measures of location. For example, if you measure the height of 100 sugar maple trees, calculating a mean height of 30m would tell you something about the location of the data. This manual will focus entirely on mean as a measure of location. However, mode and median are defined in the glossary and tests incorporating these measures are mentioned at the end of Chapter 1.

Spread of Data – Measures of spread summarize how variable the data are. Are all the measured sugar maples close to 30m in height, or are some considerably shorter and some considerably taller than 30m? Methods for quantifying spread include comparing **maximum values** and **minimum values** and measuring **range**, **standard deviation** and **variance**. As we will see in Chapter 1, summarizing the spread of the data is just as important as summarizing the location.

Pattern, Dependent Variables and Independent Variables – Often the goal of a scientific study is to explain what factors cause a particular pattern in the natural world. For example, if you observe that sunfish grow larger in some lakes than in other lakes, you may want to understand why. You could propose and test hypotheses that focus on different factors that might affect the size of sunfish. For example, one hypothesis could be that sunfish grow larger in lakes with greater food availability.

In this hypothesis, there are two variables, fish size and food availability. Fish size is the dependent variable; it is the variable you wish to explain. Food availability is the independent variable that you hypothesize is affecting fish size. In general, the **dependent variable** (also called the response variable) is the variable that is affected by the independent variable, or responds to or *depends on* the independent variable. The **independent variable** (also called the predictor variable) is the factor that you hypothesize is causing the change in the dependent variable.

If you analyze your data and find that the size of sunfish is indeed associated with food availability, you have identified a pattern in the natural world! This is an exciting and important first step in the scientific process. However, it would require further study to explain the cause of the pattern. For example, maybe food availability does not determine the size of sunfish. Perhaps there is a predator present in some lakes that eats large sunfish and the invertebrates that sunfish feed on. In this case, the presence or absence of the predator is causing the pattern.

Categorical vs. Continuous Variables
In general, both independent and dependent variables can be classified into two types, continuous and categorical. Continuous variables are measured numerically and can have a wide range of values. For example, tree height can be assigned a numerical value in meters and theoretically can vary from 0 all the way up to the maximum height that trees can grow.

Categorical variables are typically assigned to one of a few limited values or categories. For example, fish may come from one of several populations from different lakes such as Dublin Lake, Lake Nubanusit, or Lake Winnepesaki. In this case, fish population is a categorical variable and the three lakes are the possible categories for this variable.

Before analyzing your data, you are going to have to choose what type of statistical test is suitable. Knowing whether your variables are categorical or continuous is an important part of this decision. If the independent variable is categorical and the dependent variable is continuous, a t-Test or Anova may be most appropriate (Chapters 1 and 2). If your independent variable and dependent variable are both continuous, a regression analysis may be most appropriate (Chapter 3). If your independent variable is continuous but your dependent variable is categorical, a logistic regression may be most appropriate (not described in this manual – see Gotelli and Ellison, 2004). If both variables are categorical, some type of Chi-Square analysis may be most appropriate (Chapter 4). Appendices II and III provide more information on choosing the correct test for your data.

Random Chance – Random chance influences any data set that scientists work with. Let's use the height of maple trees to focus on two specific examples of how random chance can affect data. **1.** Imagine a windstorm knocks the branch off of a nearby oak tree that destroys the top of one of our maple trees. This event reduces the height of the maple tree and can be considered an example of random chance (unless we specifically are interested in the effects of damage by neighboring trees on the height of sugar maples). **2.** Measurement error is also considered the result of random chance (unless the focus of our study is to quantify measurement error). For example, if two field workers estimate the height of a sugar maple tree, their estimates may not be exactly the same and can result in random chance affecting the data set.

It is important to remember that random chance affects every data set – we attempt to minimize its effect with good experimental or sampling design, but there is no way to entirely eliminate it. A key question in data analysis is "how much has random chance influenced my data?"

It is possible for random chance to make it look like there are meaningful patterns in a data set. For example, whenever we calculate the means for two groups of numbers (such as the height of maple trees on south-facing vs. north-facing slopes), the means will rarely be *exactly* the same. Imagine that we calculate a mean height of 30.1 m for randomly selected trees on south slopes and 29.9 m for trees on north slopes. On average, are trees on south slopes really taller than on north slopes, or has random chance caused the difference? (As we will see in Chapter 1, measuring variation in the data can resolve this question.) If we go out to the field again and re-sample the two tree populations, will we get *exactly* the same means? Probably not. Maybe we would calculate a greater mean for trees on the north slope than the south slope the second time we collect data. In this case, the differences we observe between the means are most likely due to random chance rather than the slope that trees are growing on.

P-values – Because random chance influences every data set, it is important to quantify its effects by using statistics to find p-values. The purpose of a p-value is to estimate how likely it is that randomness is causing a pattern in our data set. A **p-value measures the probability that the pattern we are interested in** (such as the difference between means) **is the result of random chance**. A high p-value reflects a high probability that random chance is causing the pattern. In the sugar maple example described in the previous section, a high p-value would indicate that any difference in the mean height of trees on south vs. north-facing slopes is simply the result of random chance and therefore is not meaningful.

On the other hand, a low p-value reflects a low probability that the pattern is the result of random chance. Most likely something other than random chance is causing the pattern! If we are comparing mean values, then we would say that the difference between the means is "statistically significant." In the sugar maple example, something associated with south facing slopes may be causing trees to grow taller than on north slopes.

P-values range from 0 to 1. The cutoff value for p most often used in the scientific literature is 0.05. If $p \leq 0.05$, we say that the pattern is statistically significant. In other words, if the probability is less than or equal to 0.05 that random chance is causing the pattern, we consider it likely that something other than random chance is causing the pattern. Yet another way of making this statement – when the probability is 5% or less that the pattern is caused by random chance, then something else is probably the cause. It is important to point out that the 0.05 cutoff is simply a number agreed upon by the scientific community. It is still possible that random chance is causing the pattern in the data when $p < 0.05$; it's just not very likely (see Appendix I on Type I & II Errors). Overall, the lower the p-value → the less likely that random chance accounts for the pattern → the more likely some other factor is causing the pattern.

Parametric Statistics

All of the statistical analyses described in this manual are examples of parametric analyses. Parametric analyses rely on the assumption that the data being tested were sampled from a specified distribution (often the normal distribution – see Appendix III for more on parametric tests and the normal distribution). In addition, like most statistical tests, parametric analyses rely on the assumption of independence. Independence means that the outcome of one observation is not affected by the outcome of another. When you are collecting data, this means that each data point must be independent of the other data points in your study (see Gotelli and Ellison, 2004 for more on independence).

Although this manual does not include a thorough discussion of the assumptions required by particular parametric tests, there is a brief section at the end of each chapter describing when each test is appropriate. In addition, alternative non-parametric analyses are listed. It is the goal of this manual to help you start interpreting your data, a goal that can often be met with parametric analyses even if an assumption is violated. However, be sure to interpret your results with caution! As stated elsewhere, it is critical to consult a statistics text (see references cited) and/or statistician before presenting your results to a professional audience or journal.

Chapters and Appendices

Depending on your previous training in statistics, it may be helpful to refer back to this introduction while reading the following chapters. Also, some of the critical concepts introduced here are repeated or expanded in chapter sections and in the appendices. If the concepts are not entirely clear after your first reading of a certain definition or description, don't worry. You may need to revisit some of these concepts several times before you fully grasp them; but if you want to understand quantitative data, it is worth the effort – it will pay off!

Chapter 1 – Comparing the Means of Two Groups of Numbers: Scatterplots and the t-Test

INTRODUCTION

One of the most common questions encountered in biological and ecological research is whether the averages or mean values for two groups of numbers are different from each another. (Throughout this manual average and mean both refer to arithmetic mean – for definitions, see glossary.) To answer this type of question, it is tempting to simply calculate the mean values, compare them, and if one seems larger than the other, conclude that the means are different. However, there is no way to make any correct conclusions about overall differences between two groups of numbers just by calculating mean values. It is important to consider the spread or variation within each group of numbers. Tools for considering variation include simple descriptive statistics, graphing using scatterplots, and the t-Test.

BACKGROUND EXAMPLE

Research Question: Is there a meaningful difference in the average weight of mountain lions in the northern vs. southern Rocky Mountains?

In order to address this question, let's consider a hypothetical study by the US Department of Fish and Wildlife in which mountain lions have been captured, weighed and released. We access the data and calculate means of 34 kgs. for the northern population and 31 kgs. for the southern population. (For this example, let's assume that these captured mountain lions represent an unbiased random sample of the mountain lion populations.) Based on the averages we've calculated, can we conclude that mountain lions in the northern Rockies are larger than those in the south? ABSOLUTELY NOT! **It is impossible to make any reasonable conclusions about differences between two groups of numbers based only on mean values – we also need to know something about the spread or variation within each group!**

Let's look more closely at the mountain lion data to see how variation can influence our interpretation. We'll consider two hypothetical examples – in both examples the means for the northern and southern populations are 34 kgs. and 31 kgs. However, as Table 1.1 shows, the amount of variation in mountain lion weight is very different in the two examples.

Table 1.1a. Example 1. Data showing high variation in mountain lion weight.

Mountain Lion Weight in Kgs.

	northern population	southern population
	28.5	26.5
	29.5	27.0
	26.0	26.0
	40.0	**37.0**
	34.7	29.0
	37.5	**34.5**
	41.0	**41.0**
	35.0	30.5
	37.0	31.0
	30.8	27.5
mean weight	**34**	**31**

Table 1.1b. Example 2. Data showing relatively low variation in mountain lion weight.

Mountain Lion Weight in Kgs.

	northern population	southern population
	34.8	31.8
	35.2	32.2
	34.2	31.2
	32.5	29.5
	32.8	29.8
	35.5	**32.5**
	33.5	30.5
	33.2	30.2
	34.5	31.5
	33.8	30.8
mean weight	**34**	**31**

Before we start looking closely at the numbers, let's quickly define range and overlap. Range is defined as the difference between the maximum value and the minimum value in a group of numbers. For example, a group of numbers with a maximum value of 9 and a minimum value of 2 has a range of 7. Range is related to the amount of spread or variation within a group of numbers - in general, the greater the range in a group of numbers, the greater the variation in that group.

Overlap between two groups of numbers is determined by comparing the maximum and minimum values of the two groups. If the maximum and minimum values of one group of numbers fall entirely outside of the range of a second group of numbers, there is no overlap. On the other hand, if the maximum and minimum values of two groups of numbers are identical,

they overlap entirely. As you might expect, the greater the overlap between two groups of numbers, the less likely they will have significantly different mean values.

Table 1.2. Maximum values, minimum values, and ranges in weight for the northern and southern mountain lion populations.

	Example 1		Example 2	
	northern	southern	northern	southern
max value	41.0	41.0	35.5	32.5
min value	26.0	26.0	32.5	29.5
range	15	15	3	3

Table 1.2 shows the maximum values, the minimum values, and the ranges in the weights for the two examples. First let's consider the data in Example 1. Here the weights vary from 26.0 to 41.0 for both populations (a range of 15), and the ranges overlap completely. By closely examining the raw data (Table 1.1a), it should be clear that individuals from the northern population are not always larger than those from the south. For example, four of the ten mountain lions (data shown in bold in Table 1.1a) in the northern population are smaller than the mean of 31 kgs. for the southern population. Likewise, three of the ten mountain lions (data shown in bold) in the southern population are larger than the mean of 34 kgs. for the northern population. If we choose two mountain lions at random, one from the north and one from the south, there is a reasonable chance that the southern individual would be the larger of the two.

Now let's look at Example 2 and see how the situation differs. The variation in weight in Example 2 is relatively low as reflected by the range of 3 compared to a range of 15 in Example 1. In addition, there is almost no overlap between the two populations in Example 2 – only at the value 32.5 kgs. (data shown in bold in Table 1.1b) is there any overlap. Furthermore, an examination of the raw data (Table 1.1b) reveals that all of the mountain lions in the northern population are bigger than 31 kgs. (the mean weight of southern mountain lions) and all of the mountain lions in the southern population are smaller than 34 kgs. (the mean weight of the northern mountain lions). Based on these observations, in Example 2 it appears that the difference in the means between the two populations represents a meaningful difference in size. On average, northern mountain lions are in fact larger than southern mountain lions (later we will confirm this with a statistical test). If we select two mountain lions at random, one from the northern population and one from the southern population, chances are very high that the northern mountain lions will be bigger.

So returning to Example 1, why did the means for the two populations differ even though there is no meaningful size difference between the two populations? In Example 1 the difference between 34 and 31 is probably the result of random chance rather than a real difference between the two populations

Random Chance

Does this lack of "meaningful difference" suggest that the number 34 is actually not different than the number 31? No . . . it simply suggests that the calculated means of 34 kgs. and 31 kgs. do not represent an overall size difference between these two mountain lions populations. If a different set of mountain lions from the north and south had been weighed, we certainly would have ended up with different estimates for the means for each population (our calculated averages would not be exactly 34 and 31 with a different set of mountain lions). In fact, we may have even ended up with a greater mean value for the southern population. This scenario is similar to flipping a coin 100 times. The result will rarely be exactly 50 head and 50 tails . . . sometimes there will be a few more heads than tails, other times a few more tails than heads. For Example, recording 54 heads and 46 tails probably doesn't mean that a coin is biased – we could have just as easily recorded 46 heads and 54 tails. In this case, random chance has simply caused minor differences in the number of heads and tails. In our mountain lions in Example 1, random chance was most likely the cause for the different means in the two populations; there is probably no meaningful size difference. On the other hand, in Example 2 the weights of individuals from the two populations are so different that these differences are probably not the result of random chance – the difference is meaningful, and as we will see, "statistically significant."

ENTERING AND DESCRIBING DATA

So how do we determine *objectively* whether **1.** the difference between the means for two groups of numbers is due to random chance OR **2.** the difference is probably not due to random chance and therefore reflects something meaningful or interesting? The answer is to use graphs to visualize the data and a statistical test (such as a t-Test) to get a quantitative estimate of the likelihood that random chance is causing the differences between the means. In order to use MS Excel to graph and test our data, first we need to enter the data into a spreadsheet.

Entering the Data

I recommend entering the data as shown in Table 1.3, with one column for each group of numbers. Although it may be tempting to simply enter the numbers without column headings, it will save time and mistakes later if you enter column headings. It is also helpful to give your worksheet a new name. Initially it will be called "Sheet1" at the bottom of you spreadsheet. If you double-click on "Sheet1", it should become highlighted so you can give it any name you like. In this case, title your worksheet "data table". While entering your data, I recommend saving your workbook periodically (shortcut is holding down the ctrl and s keys at the same time). It is also a good idea to save a master copy of your data in a separate Excel file and title it "raw data" or something like that. You can then make a working copy of the data for doing the tests described below. If anything happens to your working copy, you can always go back to the file with your raw data without having to re-enter the data.

Table 1.3. Spreadsheet of raw data corresponding to Example 1.

Mountain lion weight (kgs.)	
northern	southern
28.5	26.5
29.5	27.0
26.0	26.0
40.0	37.0
34.7	29.0
37.5	34.5
41.0	41.0
35.0	30.5
37.0	31.0
30.8	27.5

Describing the Data

Once you have entered the data, it is helpful to calculate some basic descriptive statistics. Before beginning, save a new copy of your data by using the "File, Save As" command and give it the name "data analyses".

To summarize your data, I recommend calculating the statistics shown in Table 1.4. The mean values will summarize the "location" of your data. The standard deviation, min value, and max value will summarize the "spread" or variation in your data (see the Introduction to this manual for a discussion of location and spread). I also recommend doing a "count" on your data. Count is simply the total sample size – the number of observations or data points in the group of numbers. On small data sets the count or sample size may seem obvious, but for larger sample sizes it is helpful to have count appear in your summary table.

Table 1.4. Raw data from Example 1 and formulas for calculating summary statistics.

Mountain lion weight (kgs.)				
northern	southern	descriptive statistics	northern	southern
28.5	26.5	mean (average)	=AVERAGE(A3:A12)	=AVERAGE(B3:B12)
29.5	27.0	stdev	=STDEV(A3:A12)	=STDEV(B3:B12)
26.0	26.0	max	=MAX(A3:A12)	=MAX(B3:B12)
40.0	37.0	min	=MIN(A3:A12)	=MIN(B3:B12)
34.7	29.0	range	=D5-D6	=E5-E6
37.5	34.5	count	=COUNT(A3:A12)	=COUNT(B3:B12)
41.0	41.0			
35.0	30.5			
37.0	31.0			
30.8	27.5			

In order to calculate these summary statistics, you will need to enter formulas into the Excel spreadsheet. If you have already used Excel formulas, this should be relatively straightforward. If not, you will see that with a little practice, it's pretty easy.

So, type in the information shown in Table 1.4 to the right of your raw data. To use formulas, type text into each cell exactly as it appears in the 4th and 5th columns of the table. Once you type "=", MS Excel goes into formula mode and anything you type or click will get entered into the formula. If you type carefully, the formulas in Table 1.4 will work. When first using formulas in spreadsheets, patience and practice are helpful.

A useful shortcut to know when entering text within parentheses (such as "A3:A12") is to click and drag down the range of values you wish to enter. However, regardless of how you enter your formulas, it is important to check to be sure you that you have entered them correctly. To display your formulas in your spreadsheet, hold down the ctrl and ` buttons (the ` button is usually on the upper left corner of the keyboard). To switch back to viewing the calculated values, hold down ctrl and ` again. Once you have entered your formulas, check to be sure that your calculated values appear as shown in Table 1.5. If not, you have made a mistake in entering your formulas (the formulas will only work if you've set up and entered your data exactly as shown in Tables 1.3 – 1.5).

Table 1.5. Raw data and descriptive statistics showing calculated values for Example 1.

Mountain lion weight (kgs.)				
northern	southern	descriptive statistics	northern	southern
28.5	26.5	mean (average)	34	31
29.5	27.0	stdev	5.082	5.011
26.0	26.0	max	41	41
40.0	37.0	min	26	26
34.7	29.0	range	15	15
37.5	34.5	count	10	10
41.0	41.0			
35.0	30.5			
37.0	31.0			
30.8	27.5			

After summarizing your data in a table, it is extremely helpful to visualize it with a graph. It is much easier to observe range, overlap, and the overall pattern in your raw data with a graph than with a data table. It is also much easier for a reader or an audience to understand your data by looking at a graph than by looking at a table of numbers.

GRAPHING THE DATA

The most straightforward way to visualize the data for two groups of numbers is to use a scatterplot. This type of plot has the advantage of showing all of the raw data rather than grouping it into categories as is done in a frequency distribution or histogram (see Appendix V).

A scatterplot typically includes an independent variable along the x-axis and a dependent variable along the y-axis. In the mountain lion example, the independent variable is the categorical variable source population which has two possible values (northern or southern). The dependent variable is the continuous variable mountain lion weight.

Making the Graph

In order to make a scatterplot in MS Excel, all of the measured values for the dependent variable need to be in the same column. In the mountain lions data set, this means that you need to combine all the weight values for both populations into one column.

I recommend setting up a second worksheet in your MS Excel workbook. Give Sheet2 a new title such as "scatterplot" (remember, you can re-name worksheets by double clicking on the "Sheet" tab at the bottom of your screen). As shown in Table 1.6, title column 1 "category" and column 2 "mountain lion weight". You then need to know how many categories or groups there are (for the mountain lion examples, there are 2 groups – the northern and southern populations). You also need to know how many data points you have within each group (10 mountain lions per population). Enter "1" into the first ten rows in the "category" column and "2" into the next ten rows as shown in Table 1.6. The "1" and "2" are category variables that stand for the northern and southern populations.

Table 1.6. Correct format for creating scatterplot for data from Example 1.

category	mountain lion weight
1	28.5
1	29.5
1	26.0
1	40.0
1	34.7
1	37.5
1	41.0
1	35.0
1	37.0
1	30.8
2	26.5
2	27.0
2	26.0
2	37.0
2	29.0
2	34.5
2	41.0
2	30.5
2	31.0
2	27.5

To fill your "mountain lions weight" column you can copy and paste the data for the northern mountain lions population into the first ten rows of that column and the data for the southern mountain lions population into the next ten rows. There are two very convenient shortcuts that will help at this stage: after highlighting the cells you want to copy by clicking and dragging, hold down the ctrl and c keys (this is the copy shortcut). Next highlight the first cell in the row or column where you want to paste the data and hold down the ctrl and v keys (this is the paste shortcut). Don't forget to periodically save your data by simultaneously holding down the ctrl and s keys! Now your data are formatted correctly for creating a scatterplot.

Making a Scatterplot (i)

- Click the chart wizard button from your toolbar. It's the button that looks like a bar graph with a blue, yellow, and red bar (if it is not present in your toolbar, you can find additional buttons by clicking on the triangle below the two right pointing arrows on the right side of your toolbar).
- On the Standard Types window of the Chart Wizard, click the XY (Scatter) option, then click Next. Usually the default graph that is created is not the graph you want, so follow the steps below for a reliable method to make a scatterplot (you may discover short-cuts later).
- Click the "Series" tab at the top of the "Chart Wizard – Step 2 of 4" window.
- If there are any existing series, use the "Remove" button to remove them from the box on the left.
- Make sure that the "Chart Wizard – Step 2 of 4" window is not covering the data in your spreadsheet. If it is, drag the window to the side so that you can see your data.
- Click the "Add" button to create a new series.
- In the "Name:" box, type "raw data".
- Click inside the "X Values:" box, then to enter your data click and drag down all of the values entered in your category column in your spreadsheet (the column that contains the 1s and 2s).
- Click inside the "Y Values:" box, then drag down all of the values in the "mountain lion weight" column.
- Assuming you've set up and named your worksheet as described above, the text in the boxes should look as follows: in X Values box: =scatterplot!$A$2:$A$21; in Y Values box: =scatterplot!$B$2:$B$21
- Click Next, then during Step 3 you can modify Titles, Axes, Gridlines, Legend and Data Labels. On the Titles window, be sure to name your graph and label your axes. Give your graph a title such as "Mountain Lion Weight in Northern and The southern Rockies" (or you can wait and give your graph a figure number, title and description after you paste it into MS Word). The x axis can be labeled "mountain lion population", and the y axis can be labeled "weight (kgs.)".
- For this figure you probably don't need to change the Axes or Data Labels settings, but I recommend turning off the Gridlines and the Legend as they don't help clarify this graph. Then Click Next, choose where you want your graph to appear (I recommend As object in:), then click Finish.
- Now you need to change the numbers on the x-axis to population labels. To turn off the numbers, double click on the x-axis so that the Format Axis

window appears. Then on the Patterns view, click None under Tick mark labels, then click OK.

- There are two ways to add the categories "northern" and "southern" to your graph. One is to use text boxes. To insert a text box, click on the text box button at the bottom of the screen (the one that looks like an upper case "A" with lines of text. To find the text box button, you may need to install the drawing toolbar by going to Tools, Customize, Toolbars and checking the drawing toolbar.), then click on your graph where you want text and type. The other method for adding category labels to the x-axis of a scatterplot is described in Appendix VIII.
- I recommend turning off the gray background as it doesn't make the graph any clearer and it wastes ink. To do this, right click within the graph, left click Format Plot Area, change the Border to Automatic and the Area to None.
- If you would like to change the scale of an axis, double click on it, select the Scale tab, then change the Minimum, Maximum, etc.
- For graphs appearing in papers or presentations, you should remove the title above the graph, then below the graph assign it a figure number such as "Figure 1", and give your figure a title and description. This is done most easily by copying the graph and pasting it into MS Word, then adding the information below the graph as text in Word.
- Once you have completed these steps for the data from Example 1, you should get a graph that looks like Figure 1.1 shown below. If you repeat this for the data from Example 2, you should get a graph that looks like Figure 1.2.

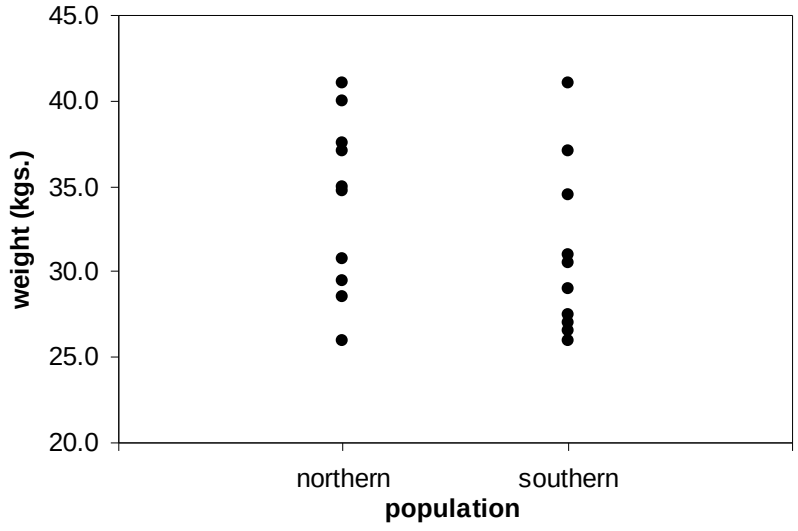

Figure 1.1. Scatterplot showing mountain lion weight for the two populations in Example 1.

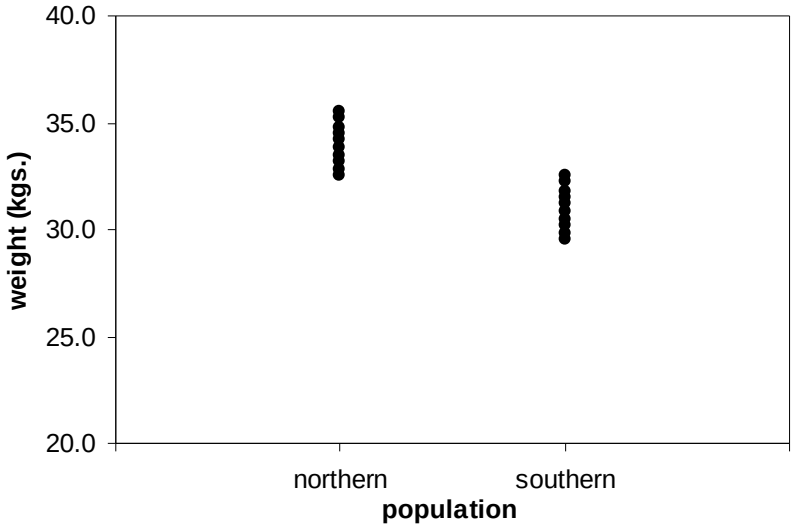

Figure 1.2. Scatterplot showing mountain lion weight for the two populations in Example 2.

Interpreting Your Scatterplot

There are many types of scatterplots, and interpretation is largely a matter of experience and practice. However, there are some general guidelines that will help you to interpret scatterplots that represent data from two groups of numbers (such as the scatterplots in Figures 1.1 and 1.2). The first thing to look for is whether there is overlap between the two groups of numbers. The more overlap, the less likely the two groups are different; the less overlap, the more likely they are different. It is also helpful to look at the amount of variation or spread in the numbers for each group. In other words, within each group, are the values all clustered near each other, or are they spread out?

It should be relatively clear just by looking at Figure 1.1 that mountain lion weight for the northern population is not consistently greater than for the southern population. There is a lot of overlap between the two populations and the data are spread out. In fact it should be easy to see (as we discovered by examining the raw data from the original data tables) that four of the ten mountain lions in the northern population are smaller than the mean of 31 kgs. for the southern population and three of the ten mountain lions in the southern population are larger than the mean of 34 kgs. for the northern population. On the other hand, Figure 1.2 shows that for Example 2, there is little overlap between the two populations - almost all of the mountain lions from the northern population weigh more than those from the southern population. Of course, scatterplots are most useful for interpreting your data when they are combined with a statistical test like the t-Test described in the following section.

Important Note on Hidden Data

Another extremely important issue to be aware of when creating and interpreting Scatterplots is hidden data. When there are two or more identical data points (such as two or more mountain lions that have the same weight), scatterplots created in MS Excel only show one data point. It is critical to find hidden data points and represent them on your graph because multiple identical

data points can radically change the interpretation of a scatterplot. You should always visually scan your raw data for identical data points. Appendix VII describes how to find hidden data and modify scatterplots to include them.

TESTING THE DATA

Once you've visualized your data with a scatterplot, you should have a much better idea of whether there are real differences between your two groups of numbers. However, to make an objective assessment of whether the differences are meaningful (or in technical terms, to estimate the probability that the difference between the means in your two groups of numbers is the result of random chance), you need to perform a statistical test. For now, let's assume that our data meet the appropriate assumptions and that a t-Test is the correct test.

Doing a t-Test

Fortunately, doing a t-Test is quicker and simpler than creating a scatterplot – in a t-Test, interpretation is the tricky part.

To perform a t-Test, follow the steps below:

- Be sure your file "data analyses" is open to the worksheet that you re-named "data table" and that contains the working copy of your raw data.
- Under the Tools menu at the top of your spreadsheet, select Data Analysis. (If this option does not appear, select Add-Ins, check Analysis ToolPak and click OK. Data Analysis should now appear as an option under the Tools menu.)
- Scroll down the options, highlight "t-Test: Two-Sample Assuming Equal Variances", and click OK.
- The cursor should now be blinking in the box next to Variable 1 Range. Click and drag down the data in your northern column. In the box you should now see "$A$3:$A:$12".
- Click the mouse in the box next to Variable 2 Range. Now click and drag down the data in your southern column. In this box you should now see "$B$3:$B:$12".
- Click OK and your data will appear in a new worksheet within your workbook.
- In order to easily find your output in the future, give your new worksheet an appropriate title such as "t-Test results".

Table 1.7. Unformatted output from a t-Test in MS Excel.

t-Test: Two-Sample Assuming Equal Variances		
	Variable 1	*Variable 2*
Mean	34	31
Variance	25.83111	25.11111
Observations	10	10
Pooled Variance	25.47111	
Hypothesized Mean Difference	0	
df	18	
t Stat	1.329175	
P(T<=t) one-tail	0.100196	
t Critical one-tail	1.734063	
P(T<=t) two-tail	0.200393	
t Critical two-tail	2.100924	

Formatting Your Statistical Table

Your output should look like Table 1.7. You will need to modify your output table before presenting it in a paper or talk. Right away you will be able to delete several rows from the table that contain information that is not critical for the overall interpretation of the t-Test (for explanation of all output, see Appendix IX). First, let's format the table and this will simplify the discussion of interpretation.

- First you need to make the left column wider so that you can read what appears in each row. To do this, you can use the mouse to place the cursor (which now looks like a white cross) on top of the line to the right of the A in the first column heading. The cursor should now look like a black vertical line with an arrows pointing to the right and the left. Double-click the mouse and that will make the column as wide as it needs to be to show all the text in that column.
- Go ahead and highlight and delete the rows titled Pooled Variance, Hypothesized Mean Difference, t Stat, P(T<=t) one-tail, t Critical one-tail, and t Critical two-tail. To delete a row, click in the row number on the left so that the entire row is highlighted. Then click Delete from the Edit menu.
- Highlight the cell called "P(T<=t) two-tail" and re-name it simply "p-value".
- You need to highlight and re-name the columns headings above your output titled "Variable 1" and "Variable 2". Type in something accurate and meaningful, such as n. mtn. lion population and s. mtn. lion population. I also recommend changing the font so that it does not appear in italics.
- Clarify the row titled Mean by adding words so that it reads "Mean Weight (kgs.)".
- **Reporting Digits:** You need to reduce the number of decimal places reported for some of the output variables. Report all zeros to the right of the decimal point until a digit with a value greater than zero is reached, then report the next two following digits with a value greater than zero. (For

example, report 5.07832 as 5.078, 0.00012497333 as 0.00012, 0.2003921 as 0.20039. Also, round up if the digit to the right of the last digit you report is 5 or greater. For example, report 32.34511 as 32.35). If you get a number in scientific notation such as 1.87283E-06, it means 1.87283 x 10^{-6} (so the decimal place should be moved to the left 6 places for non-scientific notation). You should report that number either as 1.87 x 10^{-6}, or as 0.0000019.

- For papers and presentations, it is important to give tables a clear title and description. Table titles and descriptions always appear above the table (Figure titles and descriptions always appear below the figure). Here you could change the title to read "Table 1. Results of t-Test comparing weights of mountain lions captured in northern and the southern Rockies during the fall of 2004".
- To make a lot of text fit into one cell at the top of your table, you can use the "Merge and Center" button (it looks like a 3 with arrows pointing to the side). Highlight the cells you wish to merge, then click the button. To make the text appear on several lines, click Cells on the Format window, then on the Alignment window, click Wrap text. To change the height of the row, click and drag on the lower border below the row number.
- Your table should now look like Table 1.8.

Table 1.8. Example of re-formatted output from the t-Test.

Results of t-Test comparing weights of mountain lions captured in northern and the southern Rockies during the fall of 2004.		
	n. mtn. lion population	s. mtn. lion population
Mean Weight (kgs.)	34	31
Variance	25.83	25.11
Observations	10	10
df	18	
p-value	0.20	

Interpreting Your Output

- Mean: These values are the calculated averages or arithmetic means for each of the two groups of numbers.
- Variance: This is a measure of the variation within each group of numbers. The greater the value for variance, the greater the variation. It is also helpful to know that the variance = the standard deviation squared. If you wanted to know the standard deviation, you could take the square root of the variance.
- Observations: This is simply the sample size or the number of data points within each group.
- df: This stands for "degrees of freedom" and is related to sample size. It is necessary when you use statistical table to find your p-value. Even though you have used the

computer to find p, it is still important to report df. In general, the greater your df, the better you are able to detect differences between means.

- p-value: This is the "punch-line" - the output that tells you whether the difference between the means in the two groups of data is statistically significant. If $p \leq 0.05$, then the difference between the means *is* statistically significant. If $p > 0.05$, the difference between the means *is not* statistically significant. It is also helpful to know precisely what p stands for – it is the probability that the difference between the means is due random chance (see Introduction on Critical Concepts). The lower the probability that the difference is due to random chance (the lower the value of p), the more likely that something *non-random* is causing the difference between the means. In the comparison of the weights of mountain in Example 1, the p value is 0.20 (which is greater than 0.05) meaning **there is a 20% chance that the difference in weight between the two populations is due to random chance.** Therefore there is not a meaningful difference between the two groups of numbers; the difference is not "statistically significant."

Repeat the t-Test on the data from Example 2. Your formatted output should look like Table 1.9. Why is the result different from Example 1 even though the mean values for the northern and southern populations in the two examples are the same? If you are unsure how to answer this question, I recommend re-reading the Introduction and Chapter 1 and talking to other students and your professor. Once you fully understand these examples and the interpretation, you are well on your way to understanding some of the most important concepts in statistics!

Table 1.9. The output for a t-Test on the data from Example 2. Note that the means are exactly the same as in Table 7, but the variances are much lower and the p-value of 0.0000016 shows that the difference between the means is highly significant.

Results of t-Test comparing weights of mountain lions captured in northern and the southern Rockies during the fall of 2004 (data from Example 2).

	N.Mountain lions Population	S.Mountain lions Population
Mean Weight (kgs.)	34	31
Variance	1.027	1.027
Observations	10	10
df	18	
p-value	0.0000032	

REVIEW OF KEY CONCEPTS

- It is impossible to determine whether two groups of numbers are meaningfully different in terms of their "location" or "average values" just by comparing the calculated mean values. The "spread" or "variation" within each group must also be considered.

- The lower the range and the less the overlap between the two groups of numbers (and therefore the lower the variation), the more likely there is a significant difference between the means.
- Visualizing data by using a scatterplot is a critical tool for data interpretation.
- The lower the probability that the difference between the means is simply due to random chance (in other words, the lower the p-value), the more likely there is something interesting or meaningful causing the difference between the means. In a t-Test, If $p \leq 0.05$, the difference between the means is considered statistically significant.

MORE ON THE T-TEST AND ALTERNATIVE TESTS

The Concept Behind the t-Test

What follows is a very brief overview of how the t-Test works. For details, consult Snedecor and Cochran (1980).

In a t-Test an equation is used to estimate a statistical parameter called t. The value of t is influenced by both the difference between the means of the two groups being tested and by the variation within the groups: the greater the difference between the means, the greater the value of t; the less variation within groups, the greater the value of t.

Once t is estimated from the data being tested, it is then compared to a known distribution called the t distribution. If the calculated value of t is greater than a critical t value (found using df from a table of t values), then it is unlikely that the difference between the means of the two groups being tested is due to random chance alone. This is the same as saying that the calculated value of t falls beyond a cutoff within one of the tails of the t distribution; the further the calculated t is from the center of the distribution, the less likely random chance alone is causing the difference between the means.

Other Types of t-Tests

The t-Test described above is for comparing two groups that have equal variances. Loosely speaking, this means that the variation within one group is similar to the variation within the other group. MS Excel can perform a similar test called a "**t-Test: Two-Sample Assuming Unequal Variances**" (also found under Data Analysis in the Tools menu). This test should be performed when the variation in the two groups is quite different. Additionally, some texts recommend using the test assuming unequal variances when it is unclear which to use (Gotelli and Ellison, 2004).

Another variation on t-Tests is the one-tailed t-Test (the test used on the mountain lion data was a two-tailed test). **A two-tailed test** is used when there is no prediction about which group should have a higher mean values. It just tests whether the mean values are different from each another. **A one-tailed test** is used when there is a prediction before the data are collected that one group should have a greater mean value than the other. In MS Excel, the table showing the results of a t-Test includes the p-value for both two-tailed and one-tailed tests. The two-tailed value is shown as "P(T<=t) two-tail" while the one-tailed value is shown as "P(T<=t) one-tail".

Yet another type of t-Test is called a **paired t-Test** (another option under Data Analysis in the Tools menu). A paired t-Test is often more powerful than other t-Tests but should only be used when data are paired. Pairing refers to situations when observations from the two groups occur in pairs. For example, if you are comparing the size of male vs. female birds and you measure the size of males and females from mating pairs, the data are naturally paired. To perform a paired t-Test, data must be entered so that both observations for a pair occur in the same row. It is then relatively straightforward to use MS Excel's Data Analysis option to perform a "t-Test: Paired Two Sample for Means".

When the t-Test is Appropriate

(See Snedecor and Cochran (1980) for thorough discussion of assumptions of the t-Test.)

- When the data are distributed normally (see Appendix III).
- When the data are independent (see glossary for more on independence).
- When the data were collected in an unbiased manner. Though this manual does not go into detail on methods for data collection, it is important to stress that statistical tests can not correct for data sets that were collected improperly. See Brower et. al. (1998) or Krebs (1989) for more on unbiased sampling.

Alternatives to the t-Test

There are several non-parametric tests that have less restrictive assumptions and can be used when data are not distributed normally. For details on how to perform these analyses, see Gotelli and Ellison (2004).

- A Median Test or a Wilcoxon Rank Sum Test is analogous to an unpaired t-Test.
- A Wilcoxon Signed Rank Test is analogous to a paired t-Test.

Chapter 2 – Comparing Means Among Three or More Groups of Numbers: Analysis of Variance

INTRODUCTION

The t-Test allows comparison of the means of two groups of numbers. However, in some cases it is useful to compare the means of three or more groups of numbers. Here the question is whether there is a significant difference among the means of the different groups. (The word "among" is used instead of the word "between" when the comparison involves more than two groups.) In these cases an Analysis of Variance (Anova) is required. Throughout this chapter, Anova refers to one-way Anova where only one independent variable is considered (see end of chapter for discussion on other types of Anova).

BACKGROUND EXAMPLE

Imagine you are a researcher in charge of a captive breeding program for an endangered butterfly. It turns out that the caterpillars of this butterfly, the cabbage silverspot, feed on the leaves of plants in the mustard family. You are trying to figure out the most efficient way to raise these caterpillars in captivity and you set up an experiment to compare broccoli, cabbage and wild mustard as food for the caterpillars.

In your experiment, you raise caterpillars in your greenhouse in groups of 100 in ten different patches of broccoli, cabbage and wild mustard. For each patch, you count the number of caterpillars that survive to pupation (when they make their chrysalis) and calculate a percent survival in each patch. Table 2.1 shows the data from this experiment.

Table 2.1. Percent survival of cabbage silverspot butterflies raised on broccoli, cabbage and wild mustard.

% survival to pupation		
broccoli	cabbage	wild mustard
87	88	96
77	76	84
81	81	89
78	77	85
80	80	88
91	90	98
70	71	78
79	78	86
68	69	76
85	84	93

Research Question: Is there a significant difference among the three diets in the survival rate of caterpillars to pupation?

If one of these food plants results in higher survival rates, it may be the best option for the breeding program. In order to answer your question, it makes sense to describe and graph the

data, then perform an Analysis of Variance to determine whether there is a significant difference among the mean survival rates for the three diets.

ENTERING AND DESCRIBING THE DATA

See Chapter 1 for detailed directions on how to enter data and use formulas to calculate basic descriptive statistics (Table 1.4 should be particularly useful). Table 2.2 shows the raw data for the caterpillar diet experiment along with corresponding descriptive statistics.

Table 2.2. Raw data and descriptive statistics for data from the caterpillar diet experiment.

% survival to pupation						
broccoli	cabbage	wild mustard	descriptive statistics	broccoli	cabbage	wild mustard
87	88	96	mean (average)	79.6	79.4	87.3
77	76	84	stdev	7.091	6.74	7.13
81	81	89	max	91	90	98
78	77	85	min	68	69	76
80	80	88	range	23	21	22
91	90	98	count	10	10	10
70	71	78				
79	78	86				
68	69	76				
85	84	93				

As shown in Table 2.2, the mean survival rates for caterpillars raised on broccoli and cabbage are very similar (79.6% and 79.4%), but the mean survival rate on wild mustard is greater (87.3%). Also, the maximum and minimum survival rates for wild mustard are greater than those for broccoli and cabbage. However, the ranges are large (23, 21 and 22), and there is a fair amount of overlap in the three groups of numbers. The next step in understanding any possible effects of diet on survival rate is to graph the data.

GRAPHING THE DATA

Chapter 1 includes detailed directions for making a scatterplot. Table 2.3 shows how the caterpillar survival data must be entered in order to graph these data and to add mean values to the scatterplot.

Table 2.3. Data from the caterpillar diet experiment entered in the correct form for making a scatterplot that shows mean values. Category 1 corresponds to broccoli, category 2 to cabbage and category 3 to wild mustard.

category	% survival	diet	mean value
1	91	broccoli	79.6
1	87	cabbage	79.4
1	85	wild mustard	87.3
1	81		
1	80		
1	79		
1	78		
1	77		
1	70		
1	68		
2	90		
2	88		
2	84		
2	81		
2	80		
2	78		
2	77		
2	76		
2	71		
2	69		
3	98		
3	96		
3	93		
3	89		
3	88		
3	86		
3	85		
3	84		
3	78		
3	76		

Adding Mean Values to a Scatterplot

- Follow the directions for making a scatterplot described on pgs. 11 & 12 in Chapter 1.
- Once you have created the scatterplot, right click within the plot, then click Source Data.
- Make sure you are on the Source Data, Series window, then click Add.
- In the X Values: box, drag down the three rows in the diet column that contain the entries "broccoli", "cabbage", and "wild mustard". Within the box you should now see the name of your sheet and "$C$2:$C$4".

- In the Y Values: box, drag down the three rows in the mean values column that contain those values. Within this box you should now see the name of your sheet and "$D$2:$D$4".
- Click "OK". You should now see the mean values plotted on your scatterplot. It may help to change the symbol used by Excel to represent the mean values (and perhaps the raw data), especially if you are going to print your graph in black and white.
- To change the symbol for a data point, double click on that data point.
- Make sure you are on the Format Data Series, Patterns window. You can then modify the line (in this case there should be None) and the Marker.
- For mean values on black and white graphs, I recommend choosing the Style: that looks like the horizontal line and choosing a black Foreground and a black Background. Try a Size: of 12, though you may have to tweak this to make the mean values stand out but not overwhelm the graph.
- Click OK. Your graph should look similar to that in Figure 2.1.

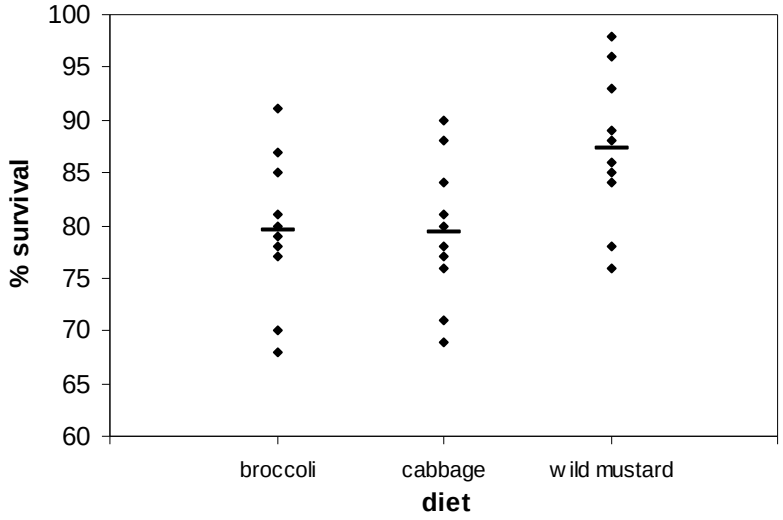

Figure 2.1. Scatterplot showing the raw data from the caterpillar diet experiment along with the mean survival rate to pupation for each diet type.

Interpreting Your Scatterplot

Once you have created a scatterplot that includes mean values, it should be easier to interpret your data. As we discussed earlier when interpreting Table 2.2, the mean value, and the maximum and minimum values for the wild mustard diet are all greater than the corresponding values for the other two diets. Also, there is a fair amount of "spread" in the data as revealed by the relatively large range within each group. However, the scatterplot in Figure 2.1 reveals some additional information. Almost all of the survival rates for caterpillars raised on broccoli and cabbage are less than the mean survival rate for those raised on wild mustard. Likewise, almost all of the survival rates for caterpillars raised on wild mustard are greater than the means for broccoli and cabbage. So perhaps survival rate is significantly greater on wild mustard than on the other two plants. In order to confirm this, we need to do an analysis of variance.

TESTING THE DATA

An Analysis of Variance (Anova) is used to determine whether there is a statistically significant difference among the means of three or more groups of numbers. In the caterpillar example, Anova will allow us to test whether there is a significant difference in the survival rate among the three caterpillar diets.

In order to use MS Excel to do an Anova, the data must be entered as shown in Table 2.2. Then use the following steps to do the test.

Doing an Analysis of Variance

- From the Tools menu, choose Data Analysis.
- Highlight Anova: Single Factor, then click OK.
- While the cursor is within the Input Range: box, drag across your raw data from the upper left entry to the lower right entry. If your data are entered as shown in Table 2, within the box you should now see "$A$3:$C$12".
- Make sure alpha is set to 0.05, under Output options choose where you would like your output to go, then click OK.
- Once you double click on the right side of the top of any column that needs to be widened to show all the text, your output should look like Table 2.4. However, this output needs to be formatted for use in presentations or papers.

Table 2.4. Table showing the MS Excel output from an Anova on the caterpillar survival data.

Anova: Single Factor						
SUMMARY						
Groups	*Count*	*Sum*	*Average*	*Variance*		
Column 1	10	796	79.6	50.26667		
Column 2	10	794	79.4	45.37778		
Column 3	10	873	87.3	50.9		
ANOVA						
Source of Variation	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
Between Groups	405.8	2	202.9	4.153689	0.026744	3.354131
Within Groups	1318.9	27	48.84815			
Total	1724.7	29				

Formatting Your Anova Table

- Delete the top two rows in your table.
- Delete the word "SUMMARY" and give your table a number, name and description such as "Table 1. Output of an Anova comparing the mean

survival rate to pupation of caterpillars raised on broccoli, cabbage and wild mustard."

- Make all of the fonts normal rather than italics.
- Rename the categories under "Groups" so that they read "broccoli", "cabbage", and "wild mustard".
- Replace the word "Groups" with the word "diet".
- Highlight the cells with the entries "Sum", "796", "794", and "873", choose Delete from the Edit menu, make sure "Shift cells left" is chosen, then click OK.
- Highlight the cells with the entries "SS", "405.8", and "1318.9", choose Delete from the Edit menu, make sure "Shift cells left" is chosen, then click OK.
- Highlight the cells with the entries "F crit" and "3.354131", choose Delete from the Edit menu, make sure "Shift cells left" is chosen, then click OK.
- Delete the bottom two rows from the table (the row beginning with "Total" and the blank row above it.
- Retype all numbers so they appear with the correct number of digits to the right of the decimal point (see discussion of reporting digits on page 15 under "Formatting Your Statistical Table".
- Your table should now look like Table 2.5.

Table 2.5. Example of a formatted table showing the Anova output that should be included in a talk or paper.

Table 1. Output of an Anova comparing the mean survival rate to pupation of caterpillars raised on broccoli, cabbage and wild mustard.

Diet	Count	Average	Variance	
broccoli	10	79.6	50.27	
cabbage	10	79.4	45.38	
wild mustard	10	87.3	50.9	
Source of Variation	df	MS	F	P-value
Between Groups	2	202.9	4.15	0.027
Within Groups	27	48.85		

Interpreting Your Output

Table 2.5 includes the minimum information that should be reported for an Anova. It includes the values you will need to determine whether the results show statistical significance as well as additional information that is helpful in understanding how an Anova works.

The Essentials

In Table 2.5, count is simply the sample size – in this case it corresponds to the 10 broccoli, cabbage and wild mustard patches in which you raised the caterpillars. The table also shows the mean (or average) survival rate for each of the three diets, and the variance which is a measure of

the variation in survival rate on each diet . . . the higher the value for variance, the greater the variation. The variance is equal to the standard deviation squared, so if you wanted to know the standard deviation for each diet, you could simply take the square root of the variance (to take a square root in MS Excel, use the "sqrt" function).

The answer to whether there is a significant difference among the means is found by looking at the p-value. If $p \leq 0.05$, then there is a significant difference; if $p > 0.05$, there is not. In this case, $p = 0.027$ which is less than 0.05, so there is a significant difference among the mean survival rate of caterpillars raised on broccoli, cabbage and wild mustard. In precise terms, a p-value of 0.027 means that there is only a 2.7% chance that the differences among the means is the result of random chance. Therefore, since random chance probably is not causing the differences, something interesting is probably going on, such as a real effect of food type on the survival of the caterpillars.

So, you can conclude that diet did have a significant effect on the survival rate of caterpillars in your experiment. By looking at the mean values and your scatterplot, you can conclude that survival rate was higher for caterpillars raised on wild mustard than for those raised on broccoli or cabbage. If you want to formally make pairwise comparisons between just <u>two</u> groups (such as wild mustard vs. broccoli) you need to do some additional testing. MS Excel is not set up to do this type of pairwise comparison. For details on making pairwise comparisons in conjunction with Anova, see Gotelli and Ellison (2004). Alternatively, a simple t-Test can be used to make a pairwise comparison, but should be interpreted with caution because doing multiple t-Tests can artificially increase the chance of finding statistical significance.

Other Parameters

Table 2.5 also includes three other parameters, df, MS and F. The parameter df is an abbreviation for degrees of freedom which is related to sample size. In general, the higher the value for df, the greater the ability to detect significant differences. MS stands for mean square. The two values shown in the table reflect the amount of variation between groups and the amount of variation within groups. In this case the Between Groups MS represents the amount of variation in survival rates among the groups of caterpillars raised on broccoli, cabbage and wild mustard. Likewise, the Within Groups MS represents the amount of variation in survival rates within the caterpillars raised on a given food type. In general, if there is a lot of variation within groups (such as a lot of "spread" in the broccoli data, the cabbage data and the wild mustard data) and not a lot of variation among groups, then there probably is not a significant difference among the means of the groups. On the other hand, if there is very little variation within groups (such as little "spread" in the broccoli data, etc.) but a lot of variation among the groups (such as little overlap in the data among the groups), there probably is a significant difference among the means. As explained in more detail below, F is a ratio that measures the amount of among group variation relative to within group variation.

MORE ON ANOVA AND ALTERNATIVE TESTS

The Concept Behind Anova

What follows is a very brief overview of how Anova works. For details, consult Gotelli and Ellison (2004).

In an Anova, the overall variation in the data is separated into the variation found within groups and the variation found among groups. Variation is quantified using measures called SS (SS is an abbreviation for Sum of Squares - literally the sum of squared differences between individual data points and mean values). Once the within group SS and among group SS values are calculated, they are converted to MS (mean squares) by dividing by df. The MS values can then be used to calculate F (F is a ratio and equals among group MS divided by within group MS). The higher the F ratio, the greater the among group variation relative to the within group variation and the more likely there are differences in the mean values among the groups.

In an Anova, the calculated F ratio is then compared to a known F distribution. Similar to the way *t* works in the t-Test, the greater the value of F, the further it occurs out in a tail of the F distribution, the less likely the difference among the means is due to random chance and the lower the p-value.

Other Types of Anova

There are many different types of Anova. Gotelli and Ellison (2004) give an excellent review of the variety of models that can be used in Anova. Of particular interest is the **two-way Anova** which can test for the effects of two independent variables on a dependent variable. For example, if the caterpillar diet experiment had included patches of young plants and patches of old plants, the resulting data could be analyzed using a two-way Anova. In this case, the two independent variables being tested would be food type (broccoli vs. cabbage vs. wild mustard) and plant age (young vs. old).

When Anova is Appropriate

(see Gotelli and Ellison, 2004 for thorough discussion of assumptions.)

- When the data are distributed normally (see Appendix III).
- When the variances within the groups being compared are equal.
- When the data are independent (see glossary for more on independence).
- When the data were collected in an unbiased manner. Though this manual does not go into detail on methods for data collection, it is important to stress that statistical tests can not correct for data sets that were collected improperly. See Brower et. al. (1998) or Krebs (1989) for more on unbiased sampling.

Alternatives to Anova

The Kruskal-Wallis One Way Anova is a non-parametric alternative to one-way Anova.

Chapter 3 – Looking For Relationships Between Dependent and Independent Variables: Scatterplots and Regression Analysis

INTRODUCTION

As mentioned in the introduction of this manual, biologists and ecologists are often interested in patterns that occur in nature. Once a pattern has been identified, one of the main goals of science is to try to understand the cause for that pattern. This is where it becomes important to understand the distinction between the independent variable and the dependent variable. This distinction is most easily understood by referring to a specific example.

You are interested in the diversity of fish communities in lakes in northern New England. You and a team of field biologists collect data from lakes throughout the region and find that the number of fish species present varies from 1 to 15. What accounts for this variation? In other words, what is causing the differences in the number of fish species in these different lakes?

Well there are probably several factors that influence fish diversity and we could easily come up with several hypotheses that may account for the variation in fish diversity. Each hypothesis would focus on a different factor or possible cause for the differences in fish diversity (or more sophisticated hypotheses might include several independent variables). For example, maybe the size of the lake is a factor influencing diversity such that large lakes have more diverse fish communities than small lakes. In this case, lake size is the independent variable and fish diversity is the dependent variable.

Independent variable – the "predictor" variable; the variable that seems to be causing the observed change in the dependent variable. In the fish example above, one possible independent variable is lake size. Other possible independent variables that might explain fish diversity include lake pH, nutrient levels in lakes, amount of cover along the bottom of lakes, etc.

Dependent variable – the "response" variable that is influenced by the independent variable. This is usually the variable for which you are trying to find an explanation - - - it is often the variable of main interest that you are trying to explain. It can also be thought of as the variable that "depends on" the independent variable.

BACKGROUND EXAMPLE

You are interested in the effects of acid deposition on the diversity of fish communities in the Adirondack Mountains. You visit fifteen lakes that vary in acidity. For each lake you measure water pH and sample the fish community to determine the number of species living in that lake. You decide to measure diversity as the number of fish species. In this example, the hypothesis that you are interested in testing could be stated as follows:

Research Hypothesis: Lake acidity affects the diversity of fish communities in lakes in the Adirondack Mountains; the greater the pH value, the greater the number of fish species.

In this hypothesis, **the independent variable is pH** and **the dependent variable is number of fish species**. (Remember, acidity is measured by pH – the greater the pH value, the lower the acidity level. A pH of 1 is highly acidic while a pH of 7 is neutral.)

You collect the following data:

Table 3.1. Lake pH and number of fish species for 15 lakes in the Adirondack Mountains.

pH	# fish species
4.0	0
5.2	2
4.2	1
6.2	9
4.8	1
6.0	4
5.5	4
3.8	1
6.4	12
5.8	8
5.9	3
6.4	4
6.8	8
4.5	3
5.0	5

ENTERING AND DESCRIBING THE DATA

See Chapter 1 for detailed directions on how to enter data and use formulas to calculate basic descriptive statistics (Table 1.4 should be particularly useful). Table 3.2 shows the raw data for the lake acidity study along with corresponding descriptive statistics.

Table 3.2. Raw data and descriptive statistics for pH and # fish species in 15 lakes in the Adirondack Mountains.

pH	# fish species	descriptive statistics	pH	# fish species
4.0	0	mean (average)	5.4	4.3
5.2	2	stdev	0.95	3.48
4.2	1	max	6.8	12.0
6.2	9	min	3.8	0.0
4.8	1	range	3.0	12.0
6.0	4	count	15	15
5.5	4			
3.8	1			
6.4	12			
5.8	8			
5.9	3			
6.4	4			
6.8	8			
4.5	3			
5.0	5			

Table 3.2 summarizes some basic information from your data set. For example, it shows that lakes ranged in pH from 3.8 to 6.8 with a mean pH of 5.4. For the number of fish species, the range was 0 to 12 with a mean of 4.3. Although this information is useful, we have not yet begun to test our research hypothesis. In order to do that, we need to make a scatterplot and perform a regression analysis.

GRAPHING THE DATA

Although Chapter 1 includes detailed directions for making a scatterplot, detailed directions for graphing data from the lake acidity study are included below.

Making a Scatterplot (ii)

- Enter the data in an excel spreadsheet in two columns as shown in Table 3.1. Be sure to name each column of data with headings such as "pH" and "# fish species".
- Click the "chart wizard" button at the top of your excel window (it is the button that looks like a blue, yellow and red bar graph).
- Click the Chart type "XY (Scatter), then click Next.
- Choose the "Series" window. If there are any columns selected in the Series window, highlight them and remove them.
- Click Add. In the Name window, type in a name for your graph, such as # of fish species and lake pH.
- Click inside the X values box, then drag down the column of data that contains your **independent variable** (in this case, pH). You may have to move the chart wizard window out of the way to see your data. Alternatively, you can type in the

code that identifies where your data are located. In the box you should see the name of your worksheet followed by "$A$2:$A$16".

- Click inside the Y Values box, delete any text in the box, then drag down the column of data that contains your **dependent variable** (in this case, # of fish species). Click Next.
- Now you will see a window titled "Chart Options". Here is where you can adjust your graph to make it as clear as possible. It is always important to include a title, and to label the axes with units when appropriate. However, it is best to avoid including lines, colors, etc. that do not help communicate the main point of the graph to the viewer. In general, the gray background and horizontal gridlines in the default excel scatterplots do not help communicate essential information and should be removed.
- In the Titles window, label your x and y axes, such as"# of fish species" and "lake acidity (pH). Be sure to include units!
- In the Gridlines window, turn off any gridlines that you don't want on your graph.
- Often in simple graphs, a legend is not needed. To remove the legend, de-select "Show legend" in the legend window.
- Once you have adjusted the various options, click Next.
- Choose where you want the graph to appear. I recommend placing it "As object in:" your spreadsheet. Later you can highlight your scatterplot, copy it, and paste it into a word document. If necessary, you can adjust its size by highlighting it and dragging the lower right corner. Click Finish.
- To change the background color, right click within your scatterplot (near, but not on top of a data point), then choose the "Format Plot Area . . . " option. In general, I recommend using no color unless the color helps communicate important information. Also, using no color helps to save ink. To turn of the background color, click "None" in the Area box.
- By right clicking within your graph, you can also change other features of your graph, such as "Chart Options".
- You may need to change the scale of you axes. For this graph, I recommend double clicking on the x-axis, and go to the Format Axis / Scale window. Then type in a minimum of 3, a maximum of 7 and a major unit of 1.
- To paste your graph into a word document, click outside the axes of you graph but within the border (such as near the title). You should then see black boxes around the border of your graph. You can now copy and paste into another document.
- If you need to change the size of your graph, you can click within it, then drag along one of the black boxes on the border. I recommend using the lower right-hand box so that you don't change the proportion of your graph.
- Once your graph is in MS Word, you can add a figure number, a title and a figure description below the graph.
- Your graph should now look like Figure 3.1.

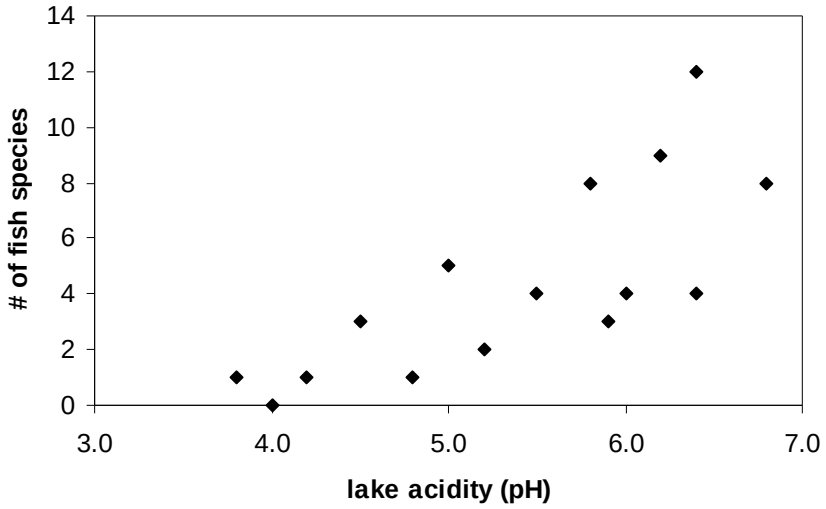

Figure 3.1. Scatterplot of number of fish species vs. lake acidity. In this graph, the data clearly show a positive relationship between the two variables.

Interpreting a Scatterplot

In this example, a scatterplot is used to begin to determine whether there is a relationship between the independent and dependent variables. Here we will only consider linear relationships (variables may also be related in a non-linear way, but that is beyond the scope of this manual). It is possible for the two variables to show a positive relationship (a line drawn through the data points would slope upward from left to right), a negative relationship (line would slope downward from left to right), or no relationship. Sometimes the relationship between the variables is clearly revealed by a scatterplot and regression analysis is then used to quantify specific parameters that describe that relationship. However, some scatterplots can be difficult to interpret and regression analysis is necessary to objectively determine whether there is a relationship between the variables.

Figure 3.1 for the lake acidity data set is an example of a positive relationship. Overall, the number of fish species increases with increasing pH. A line drawn through the data points on this graph would slope upward from left to right. This pattern supports our hypothesis that lake acidity affects the diversity of fish communities in lakes in the Adirondack Mountains. Later we will confirm this positive relationship with a regression analysis and look at some specific statistical parameters that will help us quantify the relationship.

Examples of Other Scatterplots

Examples of other possible outcomes from our lake pH study are shown below. Particularly important is Figure 3.4 that shows a graph that is difficult to interpret. For the data corresponding to Figure 3.4, a regression analysis is critical to help objectively determine whether there is a relationship or not.

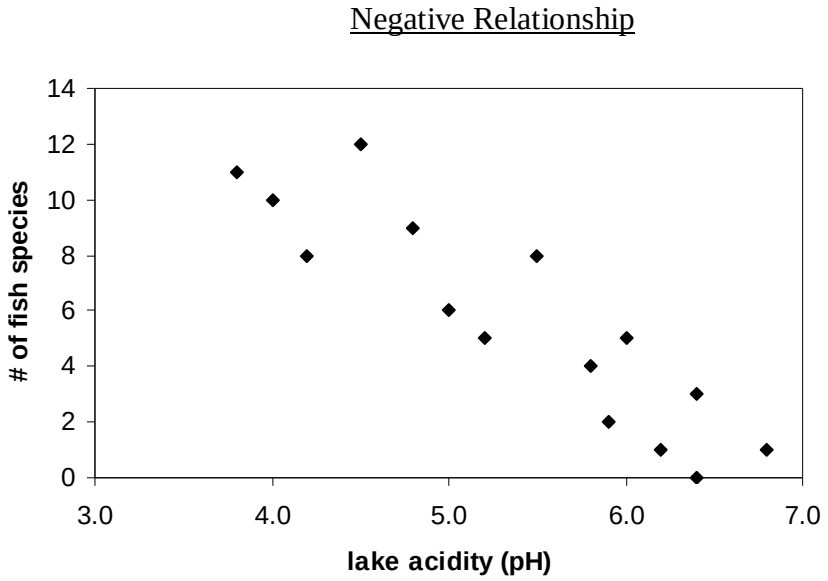

Figure 3.2. Scatterplot of number of fish species vs. lake acidity. In this graph, the data clearly show a negative relationship between the two variables.

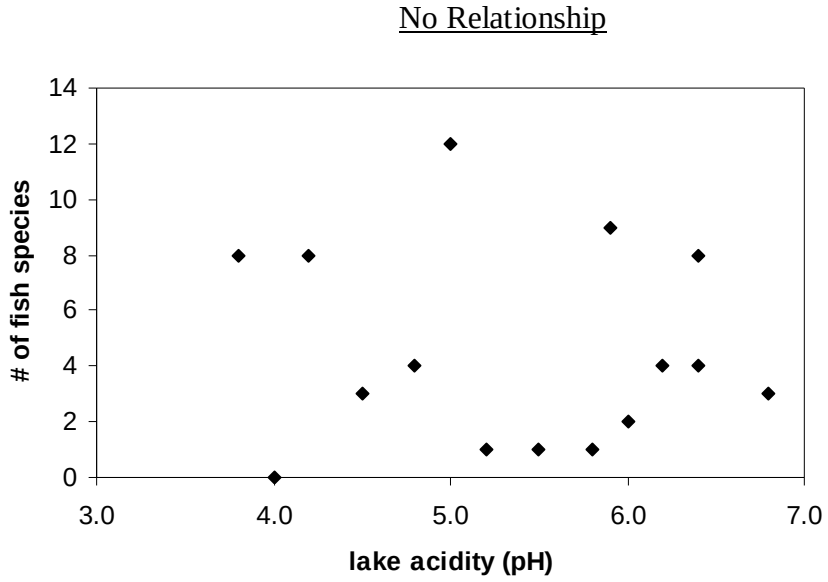

Figure 3.3. Scatterplot of number of fish species vs. lake acidity. In this graph, there is no apparent relationship between the two variables. As pH increases, the number of fish species does not consistently increase or decrease.

Positive Relationship or No Relationship?

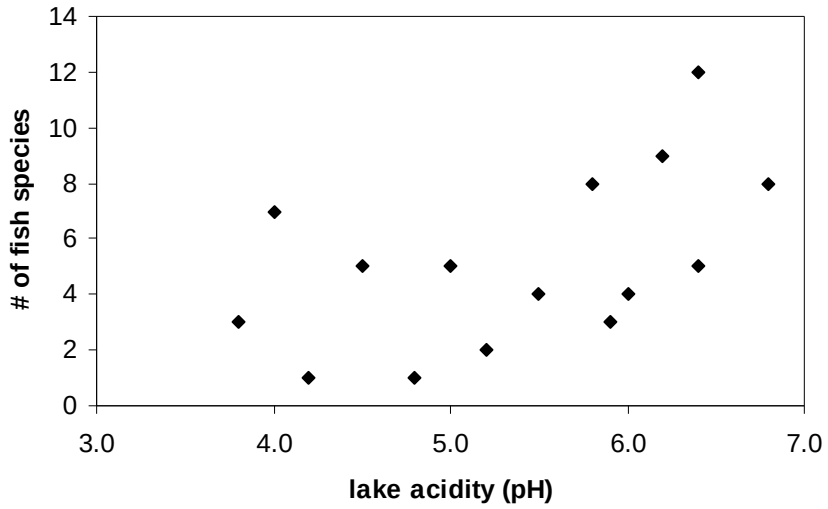

Figure 3.4. Scatterplot of number of fish species vs. lake acidity. This graph is difficult to interpret. Is it an example of a positive relationship, or no relationship?

Without doing a regression analysis (or without a lot of experience interpreting scatterplots), there are two reasonable interpretations of Figure 3.4. One possible interpretation is that the data points are scattered throughout the graph and that lakes with a relatively high pH do not consistently have either more or fewer fish species than lakes with low pH so there is no relationship between the number of fish species and lake pH. However, it is also true that the lakes with fewest fish species have relatively low pH and the lake with the most fish species has a relatively high pH. Also, a line drawn through the center of the scatter of points would slope up from left to right. Perhaps there is a weak positive relationship between the two variables. The best way to objectively quantify this relationship is by using regression analysis (note: A regression analysis on the data for Figure 3.4 reveals a significant positive relationship between the variables; $p = 0.042$, $R^2 = 0.28$; see below for definitions of these parameters).

TESTING THE DATA

A regression analysis estimates a regression line (characterized by slope and intercept) that characterizes the relationship between the two variables. The p-value represents the probability that the relationship between the two variables is due to random chance.

Doing a Regression Analysis

- Under the Tools menu, select "Data Analysis" (if this does not appear as an option, see page 14 in Chapter 1 to add this option).
- Highlight "Regression" and select "OK".
- Click within the "Input Y Range:" box, then drag down the column containing the data for your **dependent** variable (in this case, # of fish species).
- Click within the "Input X Range:" box, then drag down the column containing the data for your **independent** variable (in this case, lake size).

- Make sure the confidence level is set at 95% (this should be the case as it is the default level), then click "OK".
- Your output should look like Table 3.3.

Table 3.3. Output from a MS Excel regression analysis on the lake acidity data set.

SUMMARY OUTPUT								
Regression Statistics								
Multiple R	0.76715672							
R Square	0.588529432							
Adjusted R Square	0.55687785							
Standard Error	2.315095658							
Observations	15							
ANOVA								
	df	*SS*	*MS*	*F*	*Significance F*			
Regression	1	99.65765056	99.65765	18.594	0.000844178			
Residual	13	69.67568277	5.359668					
Total	14	169.3333333						
	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	-10.70404412	3.53813106	-3.02534	0.00975	-18.3477100	-3.06037	-18.347710	-3.06037815
X Variable 1	2.801995798	0.649802081	4.312076	0.00084	1.39818402	4.205808	1.39818402	4.205807577

Presenting and Interpreting the Output

There is a lot of useful information in the "SUMMARY OUTPUT" table from the regression analysis, but at this point there are three main values of importance; the value for "R Square", the value for "Observations" (sample size), and the value for "Significance F" (the p-value for the overall regression analysis). In this case, those values are 0.59, 15, and 0.00084. In addition, the values for intercept and X variable 1 should be reported. These values (-10.7 and 2.8 in Table 3.3) are the estimates for the intercept and slope of the regression line – the line that goes through the center of the scatter of data points. The rest of the information is important for detailed analysis of the regression output, but is not necessary for a basic interpretation of the results. I recommend simplifying the output to report the regression results as shown in Table 3.4.

Table 3.4. Results of a regression analysis on data from the lake pH study.	
Regression Statistics	
sampe size	15
p-value	0.00084
R^2	0.59
intercept	-10.7
slope	2.8

It is important to understand the meaning of these values in order to correctly interpret the analysis. The sample size simply reports how much data you have collected. In general, the more data, the more confident you can be in the results . . . the more data the better. Results from analyses with sample sizes less than 10 should be interpreted with caution.

The p-value is a measure of the probability that the relationship between the two variables is due to random chance. In this case the p-value is 0.00084 which is clearly less than the cutoff of 0.05, so the relationship is statistically significant. It literally means that there is a 0.084% probability that the relationship between number of fish species and lake pH is due to random chance. Because this probability is very low, we can conclude that there is likely something non-random and therefore biologically or ecologically interesting causing the relationship between the two variables. We can conclude that there is a statistically significant positive relationship between the number of fish species and lake pH supporting our hypothesis that lake acidity affects the diversity of fish communities in lakes in the Adirondack Mountains.

More on p-value

The p-value can vary from 0 to 1. The higher the p-value, the more likely the relationship is simply due to random chance and the less likely there is a biologically or ecologically meaningful relationship between the two variables. In this case, if the relationship were due to random chance, then we could conclude that there is nothing about lake pH that is influencing the number of fish species.

The lower the p-value, the less likely the relationship is due to random chance and the more likely it is due to something else, such as an influence of lake pH on the number of fish species. It is important to remember that 0.05 is simply a cutoff agreed to among most biologists and ecologists. Using 0.05 does not mean that you will always reach the correct conclusion (see Type I and Type II Errors in Appendix I for more on reaching false conclusions). A p-value of 0.06 is nearly significant and could reflect a meaningful relationship. A p-value of 0.04 is significant, but it is still possible that the relationship between the two variables is just the result of random chance. On the other hand, a p-value of 0.0001 reveals a very low probability that a relationship is caused by random chance, and is considered highly significant. In simple terms, the lower the p-value, the more likely there is a real and meaningful relationship between the two variables.

The R^2 value

The R^2 value (which also varies between 0 and 1) is just as important as the p-value in understanding the results of the regression analysis. R^2 represents the amount of variation in the dependent variable that can be explained by the independent variable. In this case the R^2 is 0.59 which means that 59% of the variation in the number of fish species in the 15 lakes that were sample can be explained by the pH of the lakes. That means that 41% of the variation in the number of fish species is unexplained by pH. Not surprisingly, there are probably other important influences on number of fish species besides pH.

Another way of thinking about R^2 is that is quantifies how tightly the data points are clustered around the regression line. If the data are tightly clustered around the line, the R^2 value will be relatively high, and the relationship between the two variables is relatively strong. If the data are widely scattered around the line, the R^2 value will be low and the relationship between the variables is relatively weak.

R^2 vs. p-value
It is critical to understand both values when interpreting your analysis. With large sample sizes it is quite possible to have a very low p-value (such as 0.001) and a very low R2 (such as 0.08). The correct interpretation would be that there is a highly significant statistical relationship between the two variables (only a 0.1% chance that the relationship is due to random chance), but only 8% of the variation in the dependent variable is explained by the independent variable. Even though the relationship between the variables is probably real and meaningful, there is a lot of variation in the dependent variable that is unexplained by the independent variable.

The intercept and slope are simply estimates for the equation for the regression line that could be plotted through the center of the scatter of data points. They correspond to the parameters b and m in the equation y = mx + b. Often this equation and or the regression line are included on scatterplots that correspond to regression analyses. See Appendix VIII for how to add the equation and regression line to your scatterplot.

An important note on causation
Regression analysis is simply a statistical tool and cannot conclusively determine whether the independent variable that you are testing is *causing* the changes in the dependent variable. It is *always* possible that some unmeasured variable is the true cause of the changes in the dependent variable and is in fact related to both of the variables in your analysis. For example, it is possible that the lakes in the study above were artificially stocked with fish, and lakes with higher pH were stocked with more fish species. In this case, lake pH itself may not be causing the differences in the number of fish species, but it is still related to the number of fish species.

In general, experiments (rather than observational studies) in which the independent variable is manipulated and the effect on the dependent variable is measured are required in order to confirm what is truly causing variation in an independent variable.

MORE ON REGRESSION AND ALTERNATIVE TESTS

The Concept Behind Regression Analysis
What follows is a very brief overview of how regression works. For details, consult Snedecor and Cochran (1980).

In a regression analysis, the slope of the regression line is calculated based on SS (SS is an abbreviation for sum of squares; see section at end of Chapter 2 on the concept behind Anova), then values for SS are converted to MS (mean square) by dividing by df. In order to determine whether this slope is significantly different from zero (in other words whether there is a significant relationship between the dependent and independent variables), the MS for the regression line is divided by the MS residual (an estimate of the variation in the data not related to the regression line) to calculate an F ratio. This F ratio is then used to find a p-value similar to the way p-values are found in Anova. The greater the F ratio, the lower the p-value, the less likely the relationship between the dependent and independent variable is the result of random chance.

Other Types of Regression Analysis
The regression analysis described in this chapter is linear regression. It is also possible that the two variables are related in a non-linear way and a **non-linear regression** model should be used. Another type of regression analysis is called **multiple regression** in which several different independent variables are measured and compared to see which is most strongly related to the dependent variable.

When Linear Regression is Appropriate
(see Gotelli and Ellison, 2004 for thorough discussion of assumptions.)

- When the two variables include a dependent and an independent variable (this is the same as saying the relationship is hypothesized to be a cause and effect). Correlation is used to test for relationships between variables when no cause and effect is hypothesized.
- When the relationship between the dependent and independent variables is linear.
- When the variances are constant along the regression line.
- When the data are independent (see glossary for more on independence).
- When the data were collected in an unbiased manner. Though this manual does not go into detail on methods for data collection, it is important to stress that statistical tests can not correct for data sets that were collected improperly. See Brower et. al. (1998) or Krebs (1989) for more on unbiased sampling.

Alternatives to Regression Analysis
Monte Carlo and Bayesian Analysis are non-parametric alternatives to regression (Gotelli and Ellison, 2004).

Chapter 4 – Comparing Counts With Expected Values: Chi-Square Test

INTRODUCTION

Sometimes data are collected as counts of numbers of individuals or of observations that can be placed into different categories. Often these data are best analyzed using a Chi-Square Test that compares observed counts (also called observed values) to expected values. The exact way the test works depends on how the expected values are determined. This chapter addresses two general approaches for determining expected values. Part I describes how to calculate expected values from contingency tables; Part II describes how to calculate expected values based on another data set or on theory.

Important note on the terms count, value, frequency, and relative abundance: Throughout this chapter, I use the terms count, value and frequency. It is worth defining these terms here in order to avoid confusion. **Count** always refers to raw data that represent counts of observations or individuals.

Value is always preceded by the terms observed or expected. **Observed values** are the same as the counts or the raw data – they are the data collected during the study. **Expected values** represent the numbers we would expect in our data set based on contingency tables, reference data sets, or theory.

In this manual, frequency always refers to a calculated number and never refers to a count. **Observed frequencies** are calculated the number of observations in a particular cell or category divided by the total number of observations. **Expected frequencies** can be determined in several different ways, as described below. Expected frequencies are used to calculate expected values; they are never used in the X^2 equation described later in the chapter. For both observed frequencies and expected frequencies, the sum of the frequencies for all categories is always 1.

For data on the abundance of different species in a community, **relative abundance** is synonymous with observed frequency. For example, the observed frequencies of different tree species in a forest represent the relative abundance of those species.

PART I: CONTINGENCY TABLES

When data are collected as counts of observations that fall into two different types of categories, they typically can be analyzed using a contingency table. The contingency table is then used to calculate expected values so that a Chi-Square Test can be performed. This type of analysis is best demonstrated with a specific example.

BACKGROUND EXAMPLE

Imagine you are interested in whether crayfish spend more time under cover during the day than during the night. You collect data on the number of crayfish found in four different categories according to cover type (under cover vs. in open) and light condition (in daylight vs. in the dark).

Research Hypothesis: More crayfish will be found in the open under dark conditions than under daylight.

In this hypothesis, the independent variable is light condition and the dependent variable is where the crayfish are found (under cover vs. in open). By convention, the independent variable is represented by rows in a contingency table, and the dependent variable is represented by columns. A table for the crayfish example could be setup as follows:

Table 4.1. Example of how to set up a contingency table for the crayfish data set.

	under cover	in open
dark		
light		

The data that you collect can then be entered into this table as the number of crayfish found in each category (Table 4.2).

Table 4.2. Contingency table for the crayfish data set. Numbers represent the number of crayfish found under the specified condition.

	under cover	in open
dark	18	32
light	48	25

Once data are entered into the table, the next step is to calculate the expected values for each cell. The expected values represent the numbers that should be found in each cell of the table if there is no association between the two categories. In the crayfish example, it is the number that should be found in each cell if there is no association between light condition and where crayfish are found.

Calculating Expected Values for Cells in Contingency Tables

- Calculate sums for rows, columns, and the grand total for the all the values in the table.
- The expected value for each cell is calculated by multiplying the row total by the column total, then dividing by the grand total. The calculations are simple enough to do using a calculator, or they can be programmed into MS Excel using equations. See the directions in Chapter 1 for calculating descriptive statistics for examples of how to use equations in MS Excel, or see Table 4.10 at the end of this chapter for directions for setting up spreadsheets to perform a Chi-Square Analysis. In addition, Appendix IV provides some general information for using formulas in MS Excel.

- The procedure for calculating expected values for the crayfish data set is illustrated in Table 4.3. Note: this table does not show equations in the format necessary for use with MS Excel.

Table 4.3a. Contingency table showing the calculated row totals, column totals, and grand total.

	under cover	in open	row totals
dark	18	32	50
light	48	25	73
column totals	66	57	**123** ← ***grand total***

Table 4.3b. Contingency table showing the equations for calculating expected values. These equations represent the mathematical calculations to be performed; they do not correspond to formulas for use with MS Excel.

	under cover	in open	row totals
dark	(R1*C1)/GT	(R1*C2)/GT	R1
light	(R2*C1)/GT	(R2*C2)/GT	R2
column totals	C1	C2	**GT** ← ***grand total***

Table 4.3c. Contingency table showing the values entered into the equations for calculating expected values.

	under cover	in open	row totals
dark	(50*66)/123	(50*57)/123	50
light	(73*66)/123	(73*57)/123	73
column totals	66	57	**123** ← ***grand total***

Table 4.3d. Contingency table showing the calculated expected values.

	under cover	in open
dark	26.8	23.2
light	39.2	33.8

GRAPHING THE DATA

Although a Chi-Square Test is required to determine whether the observed numbers of crayfish in each category are statistically different from the expected numbers, as always it is helpful to graph your data in order to make interpretation easier. In this case we will make a bar graph

(called a Column Chart in MS Excel). First you need to re-enter your data as shown in Table 4.4.

Table 4.4. Crayfish data entered in correct format for creating bar chart.

category	expected #	observed #
dark cover	26.8	18
dark open	23.2	32
light cover	39.2	48
light open	33.8	25

Making a Bar Graph

- Once the data are entered as shown in Table 4.4, click on the chart wizard (button that looks like the blue, yellow and red bar chart), select Chart Type: Column, then click Next >.
- Sometimes the default graph will be correct, but if not (or if you want to be sure to do the graph correctly), go to the Series window and remove all existing Series.
- Click the Add button and type "expected value" in the Name: box.
- Click in the Values: box, delete any existing text, then drag down the column of four numbers under "expected #" on your spreadsheet.
- Click Add again, type "observed value" in the Name: box, then click in the Values: box, remove existing text and drag down the column of data under "observed #" in your spreadsheet.
- Click within the Category (X) axis labels: box, then drag down the four category types in the category column of your spreadsheet and click Next.
- Now you can give your graph a title (or give it a figure number, title and description after you paste it into word) and label the axes. Appropriate titles would be "light condition/cover type category" for the x-axis and "# of crayfish" for the y-axis.
- Turn off the gridlines (don't turn off the legend), then click Next. Choose where you want your graph to appear, then click Finish.
- Now the graph should be done, but you may wish to change the colors, especially if you will be printing in black and white.
- To turn off the gray background (this will save ink), right click within the gray area, choose Format Plot Area, choose Border Automatic, Area None, then click OK. To change the colors or patterns of the bars, you can double click the bars and follow the directions in the Format Data Series window.
- Your graph should now look similar to Figure 4.1. For printing in black and white and saving ink, I recommend using patterns similar to those in this figure.

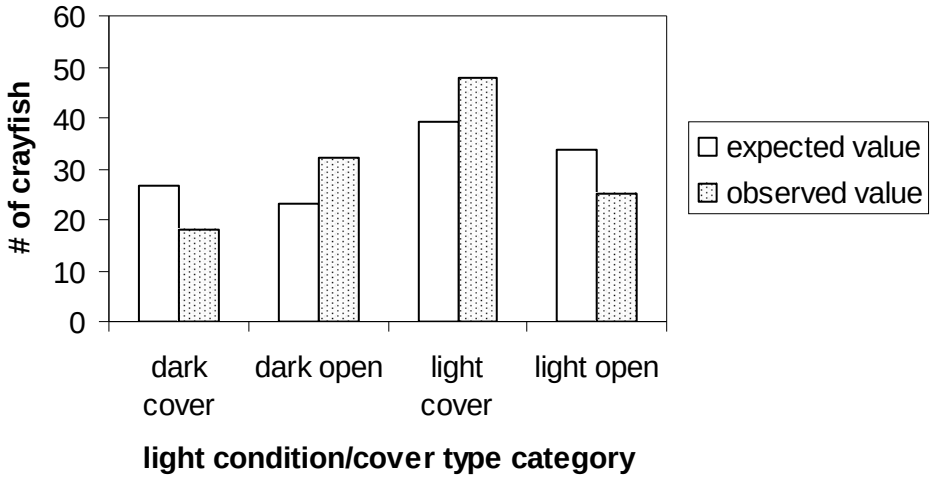

Figure 4.1. Bar graph showing the expected number and observed number of crayfish found in each light condition/cover type catergory.

By examining Figure 4.1, it is clear that under dark conditions, fewer crayfish were found under cover and more were found in the open than expected. In addition, under light conditions more were found under cover and fewer were found in the open than expected. These results are consistent with the hypothesis that more crayfish will be found in the open under dark conditions than under daylight. However, we need to do a Chi-Square Test to determine whether the results are statistically significant.

PERFORMING THE STATISTICAL TEST

Once you have your observed values and your expected values, it is relatively simple to use an equation to calculate a value for Chi-Square (represented by the symbol X^2). By using X^2 and the degrees of freedom, a p-value can be found. In this test the p-value represents the probability that the observed values are different than the expected values as the result of random chance. If there is a low probability that they are different simply due to random chance (if $p \leq 0.05$), we can conclude that the observed values are statistically different from the expected values.

The equation for calculating X^2 is shown below. As described below the equation, it is quite easy to determine X^2 using a table and a calculator. Alterntively, Table 4.10 at the end of the chapter shows how to setup a spreadsheet to have MS Excel calculate X^2 from data in a contingency table.

X^2 Equation: $X^2 = \Sigma (o - e)^2 / e$ (o = observed value, e = expected value)

Using a Table and Calculator to Determine X^2.

- Set up a table with five columns (you can do this by hand on a piece of paper). Give the five columns the following headings: observed #, expected #, o-e, $(o-e)^2$, and $(o-e)^2/e$. (The abbreviations o and e stand for observed # and expected #.)

- Enter your observed values and expected values into the first two columns.
- To fill in the third column, subtract each expected value from each observed value.
- To fill in the fourth column, simply square the values in the third column.
- To fill in the fifth column, divide the values in the fourth column by the corresponding expected values. Then add all of the values in the fifth column to calculate X^2.
- Table 4.5 shows this table for the crayfish data.
- You now need to know the degrees of freedom in order to find the p-value. For contingency table analysis, df = (the number of rows -1) * (the number of columns-1). For the crayfish data, there are two rows and two columns in the contingency table, so the df = 1.
- You can use MS Excel to find the p-value based on X^2 and the df. You can use the chidist option by typing in "=chidist(X^2,df)" In a cell in an MS Excel spreadsheet type in "=chidist(X^2,df)". So, for the crayfish data type in the following: =chidist(10.6,1). After hitting the return button, you should see the value "0.001131".

Table 4.5. Example of how to set up a calculation table for determining a X^2 value. Data are from crayfish example.

obs #	exp #	o-e	$(o-e)^2$	$(o-e)^2/e$
18	26.8	-8.8	78.0	2.9
48	39.2	8.8	78.0	2.0
32	23.2	8.8	78.0	3.4
25	33.8	-8.8	78.0	2.3
			Sum of column 5 = X^2 =	10.6
			df = (R-1) * (C-1) =	1
			to find p-value, type in "=chidist(10.6,1)"	
			p-value =	0.001131

*R and C are the number of rows and columns in the contingency table used to calculate expeceted values.

As shown in Table 4.5, the p-value for this Chi-Square Test is 0.001131 (this should be reported as p=0.0011). So how does this relate to our original hypothesis that more crayfish will be found in the open under dark conditions than under daylight? Well, this p-value literally means that the probability that random chance is causing the observed values to be different from the expected values is 0.11%. There is a very small chance that the differences between observed and expected values are the result of random chance (and p is in fact ≤ 0.05), so we can conclude that the differences are statistically significant. If we refer back to Figure 4.1 and our conclusions based on that figure, it is clear that these data support the hypothesis that more crayfish will be found in the open under dark conditions than under daylight.

Contingency tables can be applied to a wide variety of situations where data can be sorted into two types of categories (i.e. by two variables). For example, the number of isopods found in different experimental conditions of wet vs. dry and dark vs. light, or the number of endangered mammal populations that are declining vs. not-declining in protected vs. non-protected habitats. In addition, tables can be expanded to include more rows or columns depending on the different levels that can be distinguished for each variable. A 3 x 3 table would work for the endangered mammal example if the categories for population status could be divided into declining, stable and growing and the categories for protection could be divided into unprotected, semi-protected, and fully protected.

PART II: EXPECTED FREQUENCIES BASED ON ANOTHER DATA SET OR ON THEORY

Sometimes data consist of counts that are assigned to categories and the expected number for each category is based on expected frequencies from other data sets or from theory. For example, data on the frequency of tree species in the understory of a forest can be compared to expected frequencies based on trees in the canopy. Alternatively, theory can be used to determine expected frequencies such as when Mendelian ratios are used to predict the expected number of individuals with different phenotypes in the F2 generation of a genetic cross.

BACKGROUND EXAMPLE

You are managing a forest as part of a wildlife reserve and are interested in whether the mix of tree species in the forest is likely to change in the future. One way to get at this issue is to collect data on the mix of tree species in the understory and compare it to the mix of tree species in the canopy layer. Species that are under-represented in the understory may become less common in the future.

Research Question: In the wildlife reserve, does the mix of different tree species in the forest understory match the mix of different tree species in the canopy?

To answer this question, you collect data on the number of individuals of different tree species in the wildlife reserve. The raw data are shown in Table 4.6.

Table 4.6. Raw data for study to compare relative abundance of tree species in the understory with those in the canopy

Species	# canopy trees	# understory trees
American Beech	22	79
Sugar Maple	18	68
Red Maple	12	42
White Pine	9	32
Red Oak	6	22

In this example, the mix of tree species in the canopy layer (or more specifically, the relative abundance of tree species in the canopy) can be used to calculate the expected values for the number of individuals of each tree species in the understory. Then the same formula for Chi-Square that we used for the crayfish data can be used to calculate a value for X^2 and determine whether the observed values are different the expected values.

Calculating Expected Values from Expected Frequencies

- The first step is to make sure that your reference data (in this case the canopy tree data) are converted to frequencies. To do this, simply add up the total number of trees to calculate a grand total. Then for each species, divide the number sampled by the grand total. Once you have a frequency for each tree species, you can check your work by confirming that the frequencies for all the species add up to a total of 1. (**Note for analyzing genetic crosses:** This same technique can be used to calculate expected frequencies from expected ratios in genetics crosses. For example, for an expected ratio of 9:3:3:1, each value is divided by the grand total of 16 to give the expected frequencies of 0.5625, 0.1875, 0.1875, and 0.0625.)
- These expected frequencies can then be used calculate expected values for the data you are testing (in this case the understory tree data). First, calculate the grand total for the number of trees found in the understory.
- To calculate an expected value for a given species, multiply the expected frequency for that species by the grand total in the understory. The sum of the expected values should be the same as the grand total for the observed values from the understory.
- It is also worth calculating the observed frequencies for the understory data. These are calculated by dividing the number of individuals of each species in the understory by the grand total of trees in the understory. These observed frequencies are a measure of relative abundance of the different tree species and are very useful for graphing the data in order to make visual comparisons.
- The equations and calculated values for determining expected frequencies and expected numbers are shown in Table 4.7.

Table 4.7a. Example of how to calculate expected frequencies from reference data and then expected numbers from those expected frequencies.

Data on Canopy Trees				Data on Understory Trees			
Species	# individuals	expected frequ		Species	observed value	expected value	observed frequ
American Beech	22	=22/67		American Beech	79	=exp frequ*243	=79/243
Sugar Maple	18	=18/67		Sugar Maple	68	=exp frequ*243	=68/243
Red Maple	12	=12/67		Red Maple	42	=exp frequ*243	=42/243
White Pine	9	=9/67		White Pine	32	=exp frequ*243	=32/243
Red Oak	6	=6/67		Red Oak	22	=exp frequ*243	=22/243
Grand Total	67	sum of #s above		Grand Total	243	sum of #s above	sum of #s above

Table 4.7b. Table showing the calculated numbers for the expected frequencies and expected values.

Data on Canopy Trees				Data on Understory Trees			
Species	# individuals	expected frequ		Species	observed value	expected value	observed frequ
American Beech	22	0.33		American Beech	79	79.79	0.33
Sugar Maple	18	0.27		Sugar Maple	68	65.28	0.28
Red Maple	12	0.18		Red Maple	42	43.52	0.17
White Pine	9	0.13		White Pine	32	32.64	0.13
Red Oak	6	0.09		Red Oak	22	21.76	0.09
Grand Total	67	1		Grand Total	243	243	1

GRAPHING THE DATA

The frequencies (columns titled expected frequ and observed frequ) in Table 4.7b represent the relative abundance of different tree species and can be used to make a graph showing relative abundance in the canopy and in the understory. This graph will provide a way to visually inspect your data and begin to figure out whether the relative abundance (or "mix") of different tree species in the understory matches the canopy.

The relative abundance graph can be made by following the directions earlier in the chapter for **Creating a Bar Chart.** The main difference is that you need to use the tree data (you can substitute Table 4.8 for 4.4), and the graph will need to be labeled differently. Once you create this graph, it should look like Figure 4.2.

Table 4.8. Data for creating a bar graph comparing the relative abundance of canopy tree species with the understory. Data correspond to the calculated frequencies in Table 4.7.

Species	relative abundance in canopy (from expected frequ.)	relative abundance in understory) (from observed frequ)
American Beech	0.328	0.325
Sugar Maple	0.269	0.280
Red Maple	0.179	0.173
White Pine	0.134	0.132
Red Oak	0.090	0.091

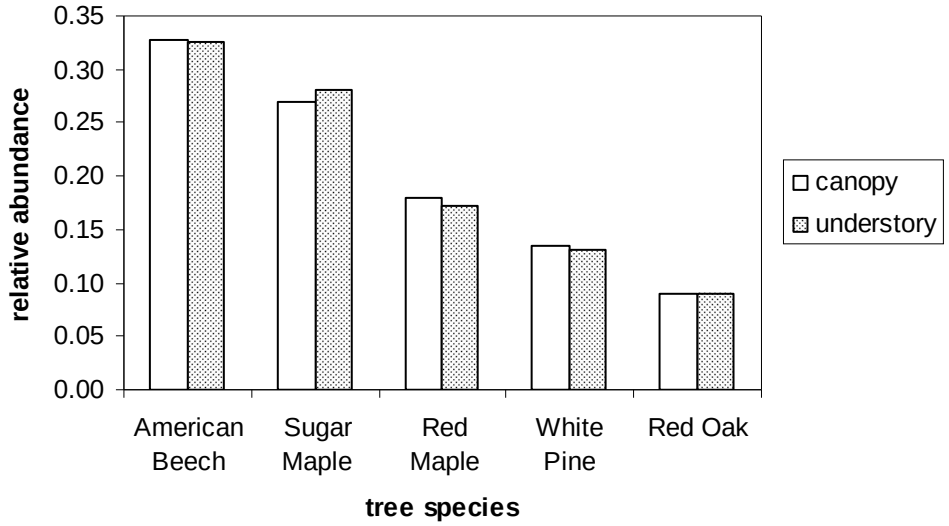

Figure 4.2. Relative abundance of tree species in the canopy and in the understory for the wildlife reserve data set.

It is easy to see from Figure 4.2 that the relative abundance of tree species in the understory is very similar to the canopy. However, it is still important to perform a Chi-Square Test in order to confirm that the differences are not statically significant and to report the values for X^2 and for p.

PERFORMING THE STATISTICAL TEST

Now that you have observed values and expected values for the understory trees (these are found in Table 4.7 and summarized in Table 4.9), you can calculate X^2 exactly the same way as described earlier in this chapter under directions for **Using a Table and Calculator to Determine** X^2. Of course your estimate for X^2 will be different than the estimate for the crayfish data. (Alternatively you can use Table 4.12 at the end of the chapter in order to set up a spreadsheet in MS Excel to calculate X^2 when expected frequencies are based on reference data sets or theory.) Also, **it is important to know that df is determined differently for this type of data set than for contingency table analysis. Here, df = # categories – 1.** So, for the tree data there are a total of five species, so df = 5 – 1 = 4.

Important Note: Even though we used the relative abundance data (or frequency data) to create a bar graph, X^2 should always be calculated using count data. Therefore, be sure to use the observed values and expected values in your analysis, not the observed and expected frequencies.

Table 4.9. Data for use in Chi-Square Test (values are from Table 4.7).		
Data on Understory Trees		
Species	observed #	expected #
American Beech	79	79.79
Sugar Maple	68	65.28
Red Maple	42	43.52
White Pine	32	32.64
Red Oak	22	21.76

Interpreting the Output

When you calculate X^2, you should get a value of 0.19. When you use the chidist option in MS Excel to find the p-value for a X^2 of 0.19 with 4 degrees of freedom, you should get a p-value of 0.99. This p-value means that there is a 99% probability that random chance is causing the differences between the observed and expected values for numbers of different tree species in the understory. Clearly this is not a significant difference, and we can conclude that the relative abundance of different tree species in the understory is no different from the relative abundance of different species in the canopy. At least based on these data, the mix of tree species in the wildlife reserve is not likely to shift in the near future.

MORE ON CHI-SQUARE TEST AND ALTERNATIVE TESTS

The Concept Behind the Chi-Square Test

In a Chi-Square Test, the equation for X^2 is used to quantify the difference between observed and expected values; the greater the difference between observed and expected, the greater the calculated value of X^2. The calculated value of X^2 is then compared to the X^2 distribution to find a p-value. In general, the greater the value of X^2, the lower the p-value, the less likely random chance is causing the differences between the observed and expected values, the more likely something interesting is causing those differences.

Other Types of Chi-Square Test

It is important to mention that for 2 x 2 contingency tables, the estimate for p as described above is subject to minor error. To calculate the exact value of p, Fisher's exact test must be used (Gotelli and Ellison 2004).

Contingency table analysis can be expanded to include more rows and or columns per category (such as a 3 x 3 table) or even to include more categories in a multi-way contingency table (such as 2 x 2 x 2).

When A Chi-Square Test is Appropriate

The biggest concern in using the Chi-Square Test is when many of the expected values are near zero. Snedecor and Cochran (1980) recommend the following guidelines to avoid problems associated with low values.

- None of the expected values should be less than one.
- Two of the expected values can be near 1 if most other values are greater than five.

Alternatives to the Chi-Square Test

Bayesian analysis can be used as an alternative to the Chi-Square Test as described in Gotelli and Ellison (2004).

Setting Up Spreadsheets to Calculate Chi-Square Values

Tables 4.10 through 4.13 show how to use formulas to setup spreadsheets to calculate X^2 values. In order to use those tables successfully, you need to enter the formulas exactly as shown in the exact same cells as shown. Alternatively, you can setup up the tables differently if you have experience using formulas in MS Excel and if you understand how the X^2 formula works. Good luck!

Table 4.10. Spreadsheet showing formulas to calculate X^2 for data entered into a contingency table. Formulas must be typed exactly as shown into the exact same cells as shown in order for calculations to be correct (see Table 1.4 for an introduction on how to use formulas). The spreadsheet below is "split" in two parts in order to fit on the page below, but on your computer screen, column F should appear to the right of column E.

	A	B	C	D	E
1	Spreadsheet to calculate X2 from contingency tables. Works for 2x2, 2x3, 3x2, and 3x3 tables.				
2		observed values			
3		**A**	**B**	**C**	
4	**1**				=SUM(B4:D4)
5	**2**				=SUM(B5:D5)
6	**3**				=SUM(B6:D6)
7		=SUM(B4:B6)	=SUM(C4:C6)	=SUM(D4:D6)	*grand total:*
8					
9		expected values			
10		**A**	**B**	**C**	
11	**1**	=IF(F7=0,0,(E4*B7)/F7)	=IF(F7=0,0,(E4*C7)/F7)	=IF(F7=0,0,(E4*D7)/F7)	
12	**2**	=IF(F7=0,0,(E5*B7)/F7)	=IF(F7=0,0,(E5*C7)/F7)	=IF(F7=0,0,(E5*D7)/F7)	
13	**3**	=IF(F7=0,0,(E6*B7)/F7)	=IF(F7=0,0,(E6*C7)/F7)	=IF(F7=0,0,(E6*D7)/F7)	
14					=SUM(B11:D13)
15					
16	**# rows:**				
17	**# columns:**				

F	G	H	I	J	K
	obs #	**exp #**	**o-e**	**$(o-e)^2$**	**$(o-e)^2/e$**
	=B4	=B11	=G4-H4	=I4*I4	=IF(H4=0,0,J4/H4)
	=B5	=B12	=G5-H5	=I5*I5	=IF(H5=0,0,J5/H5)
	=B6	=B13	=G6-H6	=I6*I6	=IF(H6=0,0,J6/H6)
=SUM(B4:D6)	=C4	=C11	=G7-H7	=I7*I7	=IF(H7=0,0,J7/H7)
	=C5	=C12	=G8-H8	=I8*I8	=IF(H8=0,0,J8/H8)
	=C6	=C13	=G9-H9	=I9*I9	=IF(H9=0,0,J9/H9)
	=D4	=D11	=G10-H10	=I10*I10	=IF(H10=0,0,J10/H10)
	=D5	=D12	=G11-H11	=I11*I11	=IF(H11=0,0,J11/H11)
	=D6	=D13	=G12-H12	=I12*I12	=IF(H12=0,0,J12/H12)
	=SUM(G4:G13)	=SUM(H4:H13)		**X^2=**	=SUM(K4:K13)
				df=	=IF(B16=0,"enter",(B16-1)*(B17-1))
				p-value=	=IF(K15="enter","# rs & # cs",CHIDIST(K14,K15))

Directions for using the spreadsheet shown in Table 4.11

- enter the observed values (counts) into the contingency table shown in gray
- data from a 2 x 2, 2 x 3, 3 x 2, or 3 x 3 table can be entered
- enter the number of rows and columns into the gray cells B16 and B17
- the value for X^2 is shown in cell K14 and the p-value is shown in K16
- Table 4.11 shows the values you should see in your spreadsheet if you enter the data shown within the gray cells in that table.

Table 4.11. Spreadsheet showing the output you should get if you setup the spreadsheet shown in Table 4.10 and enter the data shown in the gray cells below.

Spreadsheet to calculate X2 from contingency tables. Works for 2x2, 2x3, 3x2, and 3x3 tables.										
	observed data									
	A	**B**	**C**			**obs #**	**exp #**	**o-e**	**$(o-e)^2$**	**$(o-e)^2/e$**
1	18	32		50.00		18.00	26.83	-8.83	77.96	2.91
2	48	25		73.00		48.00	39.17	8.83	77.96	1.99
3				0.00		0.00	0.00	0.00	0.00	0.00
	66	57	0	*grand total:*	123.00	32.00	23.17	8.83	77.96	3.36
						25.00	33.83	-8.83	77.96	2.30
	expected data					0.00	0.00	0.00	0.00	0.00
	A	**B**	**C**			0.00	0.00	0.00	0.00	0.00
1	26.8	23.2	0.0			0.00	0.00	0.00	0.00	0.00
2	39.2	33.8	0.0			0.00	0.00	0.00	0.00	0.00
3	0.0	0.0	0.0							
				123.00		123.00	123.00		**X^2=**	10.56
									df=	1
# rows:	2								p-value=	0.0011527
# columns:	2									

Table 4.12. Formulas to setup a spreadsheet to calculate X^2 when expected frequencies are based on reference data or on theory. Formulas must be typed exactly as shown into the exact same cells as shown in order for calculations to be correct (see Table 1.4 for an introduction on how to use formulas). The spreadsheet below is "split" in two parts in order to fit on the page below, but on your computer screen, column F should appear to the right of column E.

	A	B	C	D	E
1	Spreadsheet to calculate X2 when expected frequencies are based on reference data or theory. Will allow calculation for up to ten categories.				
2			**exp count,**		
3	**categories**	**obs data**	**ratio or frequ**	**obs frequ**	**exp frequ**
4				=IF(B14=0,0,B4/B$14)	=IF(C14=0,0,C4/C$14)
5				=IF(B14=0,0,B5/B$14)	=IF(C14=0,0,C5/C$14)
6				=IF(B14=0,0,B6/B$14)	=IF(C14=0,0,C6/C$14)
7				=IF(B14=0,0,B7/B$14)	=IF(C14=0,0,C7/C$14)
8				=IF(B14=0,0,B8/B$14)	=IF(C14=0,0,C8/C$14)
9				=IF(B14=0,0,B9/B$14)	=IF(C14=0,0,C9/C$14)
10				=IF(B14=0,0,B10/B$14)	=IF(C14=0,0,C10/C$14)
11				=IF(B14=0,0,B11/B$14)	=IF(C14=0,0,C11/C$14)
12				=IF(B14=0,0,B12/B$14)	=IF(C14=0,0,C12/C$14)
13				=IF(B14=0,0,B13/B$14)	=IF(C14=0,0,C13/C$14)
14	**sum =**	=SUM(B4:B13)	=SUM(C4:C13)	=SUM(D4:D13)	=SUM(E4:E13)
15					
16					
17		**# categories:**			

F	G	H	I	J
obs #	**exp #**	**o-e**	**$(o-e)^2$**	**$(o-e)^2/e$**
=B4	=E4*$B$14	=F4-G4	=H4*H4	=IF(G4=0,0,I4/G4)
=B5	=E5*$B$14	=F5-G5	=H5*H5	=IF(G5=0,0,I5/G5)
=B6	=E6*$B$14	=F6-G6	=H6*H6	=IF(G6=0,0,I6/G6)
=B7	=E7*$B$14	=F7-G7	=H7*H7	=IF(G7=0,0,I7/G7)
=B8	=E8*$B$14	=F8-G8	=H8*H8	=IF(G8=0,0,I8/G8)
=B9	=E9*$B$14	=F9-G9	=H9*H9	=IF(G9=0,0,I9/G9)
=B10	=E10*$B$14	=F10-G10	=H10*H10	=IF(G10=0,0,I10/G10)
=B11	=E11*$B$14	=F11-G11	=H11*H11	=IF(G11=0,0,I11/G11)
=B12	=E12*$B$14	=F12-G12	=H12*H12	=IF(G12=0,0,I12/G12)
=B13	=E13*$B$14	=F13-G13	=H13*H13	=IF(G13=0,0,I13/G13)
=SUM(F4:F13)	=SUM(G4:G13)		**X^2=**	=SUM(J4:J13)
			df=	=IF(C17=0,"enter",C17-1)
			p-value=	=IF(J15="enter","# categories",CHIDIST(J14,J15))

Directions for using the spreadsheet shown in Table 4.13

- enter the observed data (counts) into the column titled obs data
- enter reference data (such as data on canopy trees in the example above) or frequencies based on theory (such as a Mendelian ratio for a genetic cross) into the column titled exp count, ratio or frequ
- enter the number of categories into the gray cell at the bottom of the sheet
- the value for X^2 is shown in cell K14 and the p-value is shown in K16
- Table 4.13 shows the values you should see in your spreadsheet if you enter the data shown in the gray cells in that table.

Table 4.13. Spreadsheet showing the output you should get if you setup the spreadsheet shown in Table 4.12 and enter the data shown in the gray cells shown below.

Spreadsheet to calculate X2 when expected frequencies are based on reference data or theory. Will allow calculation for up to ten categories.									
		exp count,							
categories	**obs data**	**ratio or frequ**	**obs frequ**	**exp frequ**	**obs #**	**exp #**	**o-e**	**$(o-e)^2$**	**$(o-e)^2/e$**
a. beech	79	22	0.33	0.33	79.00	79.79	-0.79	0.63	0.01
sugar maple	68	18	0.28	0.27	68.00	65.28	2.72	7.38	0.11
red maple	42	12	0.17	0.18	42.00	43.52	-1.52	2.32	0.05
white pine	32	9	0.13	0.13	32.00	32.64	-0.64	0.41	0.01
red oak	22	6	0.09	0.09	22.00	21.76	0.24	0.06	0.00
			0.00	0.00	0.00	0.00	0.00	0.00	0.00
			0.00	0.00	0.00	0.00	0.00	0.00	0.00
			0.00	0.00	0.00	0.00	0.00	0.00	0.00
			0.00	0.00	0.00	0.00	0.00	0.00	0.00
			0.00	0.00	0.00	0.00	0.00	0.00	0.00
sum =	243	67	1.00	1.00	243.00	243.00		**X^2=**	0.19
								df=	4
								p-value=	0.99
	# categories:	5							

APPENDIX I: The Language of Statistics

One of the trickiest parts of learning statistics is getting used to the language. The formal terms make it difficult to understand statistical writing or statisticians when they are talking. Also, it is difficult for the beginner to use terms correctly when referring to results in a written paper or during a talk.

Part of the reason that the terms are so difficult is that it is important to be precise. Also, scientists try to be careful not to over-state or over-interpret their results. It is always possible that new information will become available that can change the interpretation of data. As a result, scientists are reluctant to ever say that something has been proven.

Below is a list of definitions and guidelines that hopefully will help the reader to understand and use the language of statistics. Some of these terms and definitions might not be entirely clear upon first reading. It may be worth returning to this section after reading the results section in a paper from the scientific literature, or before reporting your own results in a paper or talk.

Common Terms Used in Reporting Results from Statistical Tests

- **null hypothesis** – a statement that any observed variability or pattern in the data is caused by random chance. Specific examples are shown in Appendix II. Statistical analyses work by testing specific null hypotheses. If the evidence against the null hypothesis is strong enough (if $p \leq 0.05$), we reject or disprove the null hypothesis. If the evidence is not strong enough (if $p > 0.05$), we accept the null hypothesis that random chance is causing the observed variability or pattern in the data.
- **accept** – to accept a hypothesis is to consider that based on the current data, it is likely to be true.
- **reject** – to reject a hypothesis is to say that it is probably untrue.
- **disprove** – often the term disprove is used instead of the term reject. It is important to note that many scientist take the position that hypotheses can never be proven (only supported) because it is always possible that new and better information may come along that will change the interpretation of a given set of data.
- **statistically significant** or **significant** If $p \leq 0.05$, the results of the statistical test are said to be statistically significant. Often the term significant is used alone when context makes it clear that significant means statistically significant.
- **alternative hypothesis** – in contrast to a null hypothesis, an alternative hypothesis provides an explanation (other than random chance) for the observed variability or pattern in the data. Often there are several alternative hypotheses each providing a different possible explanation for the observed variability or pattern in the data; each focuses on a different independent variable.
- **research hypothesis** – I have used the term research hypothesis in this manual to emphasize that I am not referring to null hypotheses. I use the term to refer to a scientific explanation proposed by a researcher to explain the observed variability or pattern in the data. A research hypothesis includes an independent and dependent variable (defined in the Introduction to the manual and in the glossary). In the research hypothesis, the independent variable (rather than random chance) is the proposed explanation for the observed variability or pattern in the data.

- **support** or **consistent with** – If the results of the statistical test reject the null hypothesis, they may be said "to support" or "to be consistent with" the research hypothesis that is being tested. As stated above, it is important to avoid stating that hypotheses have been proven (there is almost always another possible explanation for the results).
- **pattern** – whenever the observed variability in an dependent variable seems to be related to an independent variable, there is pattern in the data. If the variability is entirely random, there is no pattern.
- **suggest** – often the term "suggest" is used in sentences describing the pattern in data. For example – "The data in the scatterplot **suggest** that the diversity of fish communities is related to lake acidity." Usually the term "suggest" is more appropriate than the terms "show" or "prove".
- **data** – The word data is plural. The singular form is datum, a term that is almost never used.
- **Type I Error** – when a true null hypothesis is rejected. This may lead the researcher to support an alternative hypothesis that is incorrect. When there is a Type I Error, the results of the data analysis show statistical significance even though the null hypothesis is true. The lower the p-value, the less likely that there is a Type I Error.
- **Type II Error** – when a false null hypothesis is accepted. In this case the researcher may fail to support an alternative hypothesis that is correct. When there is a Type II Error, the results of the data analysis do not show statistical significance even though the null hypothesis is false. In general, Type II Errors are most common when the p-value is greater than but close to 0.05 (i.e. a p-value of 0.06).

APPENDIX II: Hypotheses

All statistical tests described in this manual work by testing null hypotheses, often abbreviated H_o. Null hypotheses are statements that any observed variability or pattern in the data is due to random chance. If $p > 0.05$ in the statistical test being performed there is no "statistical significance" and the null hypotheses is accepted. If $p \leq 0.05$, there is "statistical significance" and the null hypothesis is rejected.

The research hypothesis is the possible explanation that is of interest to the scientist doing the data analysis. It is almost always different than the null hypothesis. Often when the null hypothesis is rejected, the results support the research hypothesis.

The term alternative hypothesis is commonly used instead of research hypothesis. Often there are several alternative hypotheses each providing a different possible explanation for the observed variability or pattern in the data; each focuses on a different independent variable.

Table IIa shows the null hypotheses corresponding to the statistical tests described in this manual. Generalized research questions and hypotheses are also included.

Table IIa. Examples of null hypotheses, research questions and research hypotheses associated with the tests described in this manual. Note that the research questions and hypotheses in this table are generalized. They should always be more specific when referring to specific data sets. See the beginning section of each chapter for specific examples.

Statistical Test	Null Hypothesis	Generalized Research Question	Generalized Research Hypothesis
t-Test	There is no significant difference between the means of the two groups being compared; any difference is solely due to random chance.	Is there a significant difference between the means of the two groups being compared?	The two groups have different mean values. For a one-tailed test, a prediction is made about which group should have a larger mean value.
Anova	There is no significant difference among the means of the groups being compared; any differences are solely due to random chance.	Is there a significant difference among the means of the groups being compared?	The groups being compared have different mean values.
Regression	There is no significant relationship between the dependent and independent variables; the slope of the regression line is not significantly different from zero.	Is there a significant relationship between the dependent and the independent variables?	There is a relationship (usually specified as positive or negative) between the dependent and independent variables.
Chi-Square Test	The observed values are not significantly different from the expected values. In R x C contingency tables, this is synonymous with stating that there is no association between the row and column categories.	Are the observed values different than the expected values?	The observed values are different than the expected values.

APPENDIX III: What test is right for these data?

Choosing the correct statistical test to analyze a data set involves understanding what type of variables you are testing (whether the independent and dependent variables are categorical or continuous), and checking whether the data meet the assumptions of particular statistical tests. Thoroughly checking assumptions is beyond the scope of this manual, though some assumptions are listed at the end of each chapter. As pointed out elsewhere, you should consult a statistics text (see references cited) and/or a professional statistician to be sure you are using the correct statistical test.

This manual focuses on several parametric tests. The general concept behind parametric tests is explained below. For now, let's assume that your data meet the assumptions for one of these tests. Then, Table IIIa should help you decide which statistical test is best for analyzing your data.

Table IIIa. General guide to choosing parametric tests based on whether the independent and dependent variables are categorical or continuous. Definitions of independent, dependent, categorical and continuous variables are included in the introduction to the manual and in the glossary. Table IIa in Appendix II may also help you choose an appropriate test.

Independent Variable	Dependent Variable	Statistical Test
categorical	continuous	t-Test or Anova (Chs. 1 & 2)
continuous	categorical	logistic regression (not in this manual; see Gotelli and Ellison 2004)
continuous	continuous	Regression Analysis (Ch. 3)
categorical	categorical	Chi-Square Test (Ch. 4)

Parametric Statistics

All parametric tests rely on the assumption that the data were sampled from a specified probability distribution. A probability distribution is a distribution of outcomes based on a mathematical equation called a probability distribution function (PDF). If data follow a known PDF, they can be characterized with relatively few parameters and the mathematical calculations for performing a statistical test are relatively simple. Many parametric tests assume that data follow the normal distribution, and fortunately biological and ecological data are often distributed normally.

The Normal Distribution

A graph of the PDF for the normal distribution shows the expected frequency of different values. It shows that values occurring near the mean are most common, while those further from the mean in the tails of the distribution are less common (Figure IIIa). Consequently, if you are collecting data from a set of values that follow the normal distribution (such as height in humans), most measurements will be relatively close to the mean, and extremely low or high values will be rare.

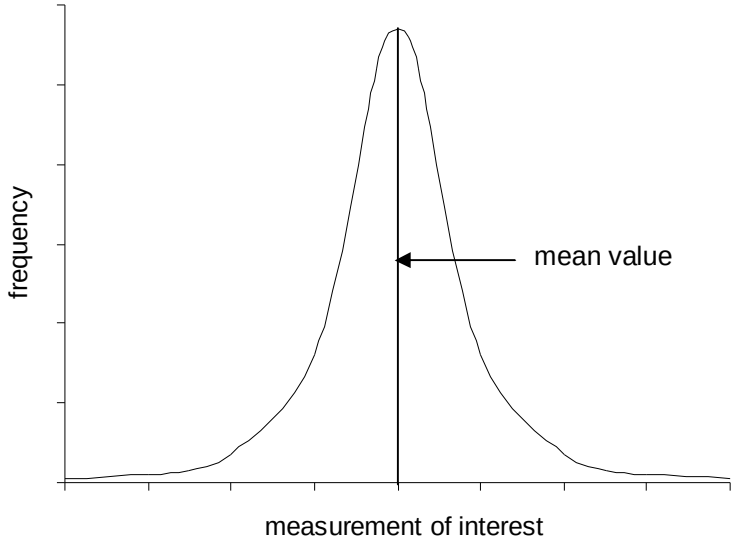

Figure IIIa. The Normal Distribution. The x-axis corresponds to the measurement of interest (such as height in humans), and the y-axis is the frequency of each measurement. Data that follow the normal distribution can be characterized with just two parameters, the mean and the standard deviation. The mean value occurs in the center of the distribution and has the highest frequency. The standard deviation measures the width or "spread" of the distribution.

Data that follow the normal distribution can be characterized with just two parameters, the mean and the standard deviation. As a result, statistical testing is easier for these data than for data that don't follow a known distribution and therefore can't be characterized by two simple parameters. The t-Test and Anova rely on the assumption that data are distributed normally.

APPENDIX IV: Using Formulas in MS Excel

Formulas in MS Excel provide a powerful tool for doing mathematical calculations. Once you get used to using formulas, it is often as quick or quicker to use a spreadsheet than a calculator. In addition, if you save your formulas in a spreadsheet, you can repeat calculations on different data sets very quickly and easily.

Tips for using formulas in MS Excel

- To start a formula in a spreadsheet, type "=" (without the quotation marks) into a cell.
- Once you have started your formula, you have to be very precise about what you type. Any mistake in the formula (such as a stray period or comma or a missing letter) will result in an error. You also need to be careful about clicking with the mouse as that can change a formula.
- To add use "+", to subtract use "-", to multiply use "*", to divide use "/".
- You can type numbers in your formula to do simple calculations. If you type in "=100/2" and then hit the return button, you should see "50" in that cell of the spreadsheet.
- If you want to display the formula rather than the calculated value, hold down the ctrl and the ` button (the ` button is on the upper left side of the keyboard). Displaying formulas can be very helpful in checking that they are entered correctly.
- If you want to perform calculations on data entered in other cells, you can refer to them by code using the letter corresponding to the column heading and the number corresponding to the row heading. For example, to add the values in the first two rows of the first column of a spreadsheet, you could type in "=A1+A2".
- A quick way to enter the value of a cell into a formula is to click on that cell with the mouse.
- If you have a formula entered at the bottom of a column of numbers and you want the same formula entered in the bottom of another similar column of numbers, you can copy and paste the formula from one cell to another (the shortcuts for copy and paste are holding down the "ctrl" and "c" buttons, and holding down the "ctrl" and "v" buttons.) The new formula should be automatically adjusted to correspond to the new column if no dollar signs ($) appear in the formula. So in the example below, copying the formula at the bottom of column 1 and pasting it into the bottom of column 2 should give the following results:

Column 1	Column 2
10	8
5	7
4	2
=(A2+A3)/A4	=(B2+B3)/B4

- If you want to keep the values for a cell constant rather than having them change when you copy and paste them to a new cell, use "$" in the formula. For example, if you type in "=$A$2", then copy and paste that formula into a new cell anywhere in the spreadsheet, it will remain "=$A$2".

- If you want to refer to a range of numbers, you can enter the first cell code, a colon, then the last cell coded. For example "A1:A10" refers to all cells in column A from rows 1 to 10.
- MS Excel has many built in functions that perform calculations. For example, "=AVERAGE(A1:A10)" will calculate the mean value for the data entered in column A rows 1 through 10.
- If you click on the *fx* button on the tool bar at the top of the spreadsheet, you can get a list of functions included in MS Excel.
- Tables 1.4, 4.10 and 4.12 give examples of using formulas to do statistical calculations.

APPENDIX V: Histograms

Histograms or frequency distributions show the frequency of different values in a set of data. They are often used to display data in conjunction with t-Tests or Anova and they can be useful in determining whether data are distributed normally.

Making a histogram in MS Excel requires first using the Histogram option under the Data Analysis in the Tools menu, then choosing Chart type: Column in the Chart Wizard. The trickiest part is choosing what range of values (called "Bin" in Excel) you want each column to display.

If you wanted to create a histogram for the set of numbers shown below, you might choose the following categories or ranges of values: values from zero to 2, values from greater than 2 to 4, etc. These categories are easier to list as 0 – 2, >2 – 4, >4 – 6, >6 – 8, >8 – 10.

2
3
4
4
5
5
5
6
6
10

To set up those categories, type in a second column of numbers as shown below:

2	2
3	4
4	6
4	8
5	10
5	
5	
6	
6	
10	

You are now ready to create a histogram.

Making a Histogram in MS Excel

- Enter the data and Bin categories as shown above.
- Select Histogram from the Data Analysis option under the Tools menu. Click OK.
- In the Input Range: box, drag down the column containing the data; in the Bin Range: box, drag down the column containing the Bin categories. In the two boxes you should see "$A$1:$A$10" and "$B$1:$B$5". Click OK.

- You should now see the following output in your spreadsheet:

Bin	*Frequency*
2	1
4	3
6	5
8	0
10	1
More	0

- I recommend adding another spreadsheet column (see third column below) to more clearly show the range of values that will correspond to the columns in your histogram.

Bin	*Frequency*	Range
2	1	0-2
4	3	>2-4
6	5	>4-6
8	0	>6-8
10	1	>8-10
More	0	

- Click the chart wizard at the top of your spreadsheet (the button that looks like the blue, yellow and red bar graph), select Chart type: Column, then click Next.
- Go to the Series window and remove any existing Series.
- Click Add. In the Values: box, drag down the column of frequency values. You should see "$B$2:$B$6".
- In the Category (X) axis labels: box, drag down the column of range values. You should see "$C$2:$C$6".
- Click Next. In the Titles window, label your x and y axes "range of values" and "frequency". You can also give your graph a title, but for presenting figures in talks or papers, the figure number, title and description go below the graph. This is most easily accomplished by copying and pasting your graph into MS Word when it is finished and then typing the figure information below the graph.
- Turn off the gridlines and legend, then click Finish.
- I recommend turning off the gray background by right-clicking on it. Then click Format Plot Area, choose Automatic for Border and None for Area. I also recommend right clicking outside the graph but within the graph window, choosing Format Plot Area, then in the Patterns window, choose None for Border. Your histogram should now look like Figure 5a.

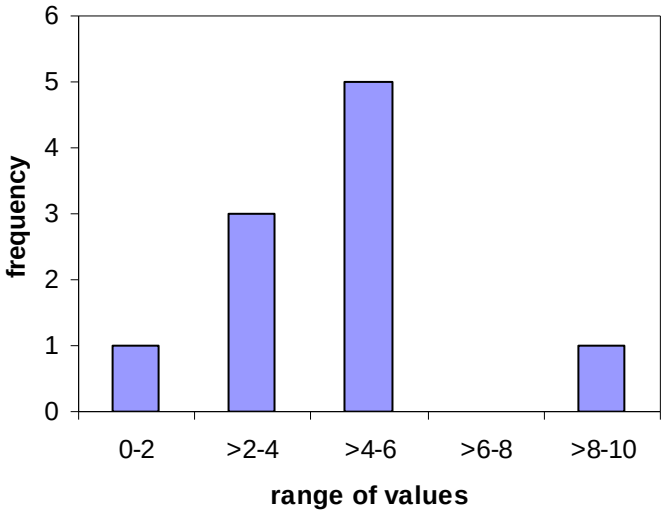

Figure 5a. Histogram showing the frequency of various values for the data shown on the previous page.

APPENDIX VI: Putting Error Bars on Graphs

Sometimes in scatterplots or bar graphs, mean values are shown with error bars to represent the "spread" or variation in the data. In the case of scatterplots, error bars can be used instead of displaying the raw data, especially when there are many data points or many hidden values (see Appendix VIII for information on hidden values). Error bars can represent the standard deviation or the standard error so it is important to state in your figure description what type of error bars appear in your graph.

Table VIa. Data from Example 1 in Chapter I.

mountain lion data, example 1 (data from Table 1.5)		
population	mean values	standard deviation
northern	34	5.082
southern	31	5.011

Adding Error Bars to a Scatterplot

- Enter the data as shown in Table VIa.
- Choose XY (Scatter) for Chart type: from the Chart Wizard (button that looks like a blue, yellow and red bar chart), then click Next.
- Go to the Series window and remove any existing Series.
- Click Add. In the X Values: box, drag down the northern and southern cells. You should see " $A$3:$A$4".
- In the Y Values: box, drag down the mean values. You should see " $B$3:$B$4".
- Click Next. In the Titles window, label your x and y axes "population" and "weight (kgs.)". You can also give your graph a title, but for presenting figures in talks or papers, the figure number, title and description go below the graph. This is most easily accomplished by copying and pasting your graph into MS Word when it is finished and then typing the figure information below the graph.
- Turn off the gridlines and legend, then click Finish.
- Double click on the numbers below the x-axis, then under Format Axis in the Patterns window, click None for Tick mark labels. Click OK.
- Double click a data point, then in the Y Error Bars window under Format Data Series click Custom. In both the + and – boxes, drag down the standard deviation values. In both boxes you should see " $C$3:$C$4". Click OK.
- To add the northern and southern labels below the x-axis, you can either use text boxes as described on pg. 12 in Chapter 1, or you can refer to Appendix VII on tweaking graphs in MS Excel.
- I recommend turning off the gray background by right-clicking on it. Then click Format Plot Area, choose Automatic for Border and None for Area. I also recommend right clicking outside the graph but within the graph window, choosing Format Plot Area, then in the Patterns window, choose None for Border. Your graph should now look like Figure VIa.

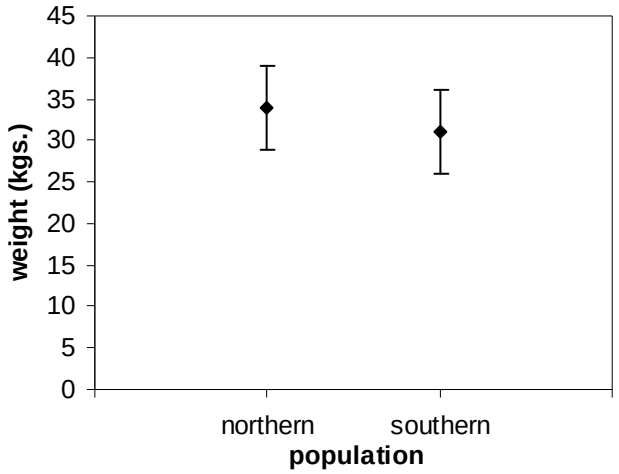

Figure VIa. Scatterplot showing mean weight of the mountain lions from Example 1 in Chaper I. Error bars represent one standard deviation.

By convention, scatterplots are often used for displaying data that are numerical measurements such as size, weight, etc. By contrast, bar graphs are used for displaying data that represent counts, percentages, or frequencies. The data in Table VIb are appropriate for plotting with a bar graph.

Table VIb. The mean and standard deviation of the number of ant nests per sample in field and forest habitats. Data are from Gotelli and Ellison (2004).

Habitat	Mean	Standard Deviation
Forest	7.00	2.19
Field	10.75	1.50

Adding Error Bars to a Bar Graph

- Enter the data as shown in Table VIb.
- Choose Column for Chart type: from the Chart Wizard (button that looks like a blue, yellow and red bar chart), then click Next.
- Go to the Series window and remove any existing Series.
- Click Add. In the Values: box, drag down the mean values. You should see "$B$2:$B$3".
- In the Category (X) axis labels: box, drag down the two habitat types. You should see " $B$2:$B$3".
- Click Next. In the Titles window, label your x and y axes "habitat" and "ant nests/sample". You can also give your graph a title, but for presenting figures in talks or papers, the figure number, title and description go below the graph. This is most easily accomplished by copying and pasting your graph into MS Word when it is finished and then typing the figure information below the graph.
- Turn off the gridlines and legend, then click Finish.

- Double click a bar on the graph, then in the Y Error Bars window under Format Data Series click Custom. In both the + box, drag down the standard deviation values. In that box you should see " $C$2:$C$3". Click OK.
- To add the forest and field labels below the x-axis, you can either use text boxes as described on pg. 12 in Chapter 1, or you can refer to Appendix VIII on tweaking graphs in MS Excel.
- I recommend turning off the gray background by right-clicking on it. Then click Format Plot Area, choose Automatic for Border and None for Area. I also recommend right clicking outside the graph but within the graph window, choosing Format Plot Area, then in the Patterns window, choose None for Border. Your graph should now look like Figure VIb.

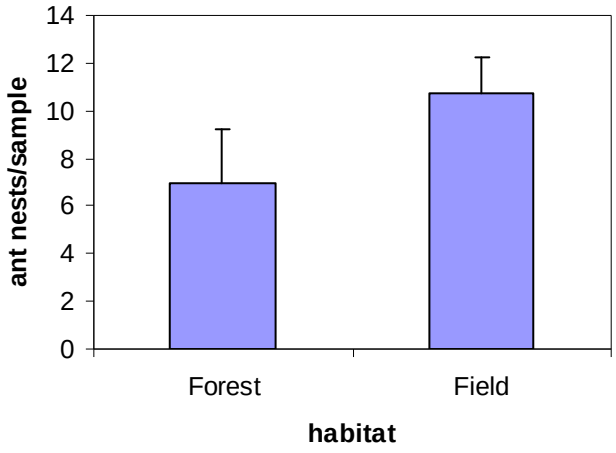

Figure VIb. Bar graph showing the mean number of ant nests/sample in forest and field habitats (Data from Gotelli and Ellison, 2004). Error bars represent one standard deviation.

APPENDIX VII: Finding and Displaying Hidden Data on Scatterplots

Hidden data occur in scatterplots when two or more data points have the same values for the x and y coordinates. Although there are several data points with the same values, the scatterplot only shows one data point on the graph. As we will see below, the presence of hidden data can greatly influence the interpretation of a scatterplot. Therefore, hidden data points need to be identified and displayed.

Table VIIa shows hypothetical data set for fish sampled in lakes of different acidity (data are different from the example in Chapter 3). Figure VIIa is a scatterplot of the data that does not display hidden data points.

Table VIIa. A hypothetical data set showing lake pH and number of fish species from sixteen different lakes.

pH	# fish species
5.0	7
5.0	4
7.0	8
6.0	4
7.0	8
5.0	4
4.0	2
4.0	8
4.0	2
7.0	3
6.0	8
5.0	4
7.0	8
6.0	8
6.0	8
4.0	2

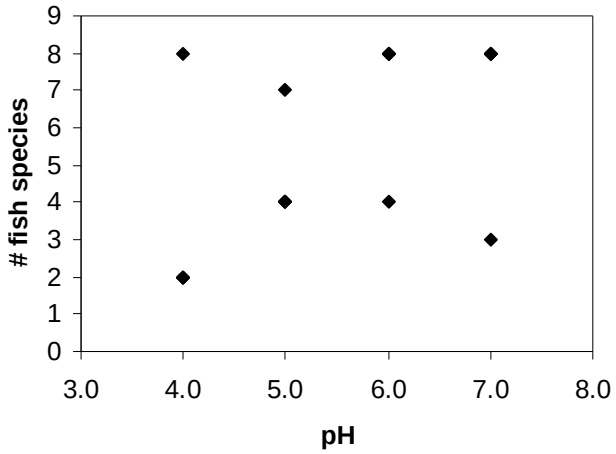

Figure VIIa. The number of fish species plotted against lake pH. Hidden data are not displayed. Notice that there are only eight data points on the graph even though sixteen lakes were sampled.

In Figure VIIa it appears that there is no relationship between the number of fish species and lake pH. Lakes with higher pH do not have consistenly more or fewer fish species than lakes with lower pH. However, this graph is misleading because it does not display hidden data. It is important to always check your scatterplot against your raw data for evidence of hidden data. The first (and often easiest) way to identify the presence of hidden data is to count the number of points on the scatterplot. In Figure VIIa there are only eight data points even though sixteen lakes were sampled. Clearly there must be data that are hidden.

Finding Hidden Data

- To find all hidden data on a scatterplot, you can use the sort command on the corresponding raw data in your spreadsheet.
- First highlight the column headings and all data for your independent and dependent variables. It is important to be sure that both columns are highlighted. Then, from the Data menu, click Sort.
- In the Sort by window, choose the name of your independent variable (in this case, pH). Be sure Ascending is selected. In the Then by window, choose the name of the dependent variable (in this case, # fish species). Again, be sure Ascending is chosen, then click OK.
- Your data should look like those in Table VIIb.
- It should now be easy to see data that share identical x and y coordinates. It turns out that many of the data points share the same values for pH and # fish species, such as the three lakes that had a pH of 4.0 and 2 fish species (data points with shared values are shown in bold in Table VIIb).
- All of these data points need to be represented on the scatterplot. The easiest way to do this is by labeling data points on the graph with a number in parentheses as shown in Figure VIIb. Labeling can be done using the text box option (see directions for Making a Scatterplot in Chapter I). Alternatively, you can use data labels to display numbers in parentheses. Though this procedure is somewhat awkward, it is explained below Figure VIIb.

Table VIIb. Data from Table VIIa that has been sorted. Note that when data are sorted, it is easy to identify and count data with identical x and y coordinates (data shown in bold).

pH	# fish species
4.0	**2**
4.0	**2**
4.0	**2**
4.0	8
5.0	**4**
5.0	**4**
5.0	**4**
5.0	7
6.0	4
6.0	**8**
6.0	**8**
6.0	**8**
7.0	3
7.0	**8**
7.0	**8**
7.0	**8**

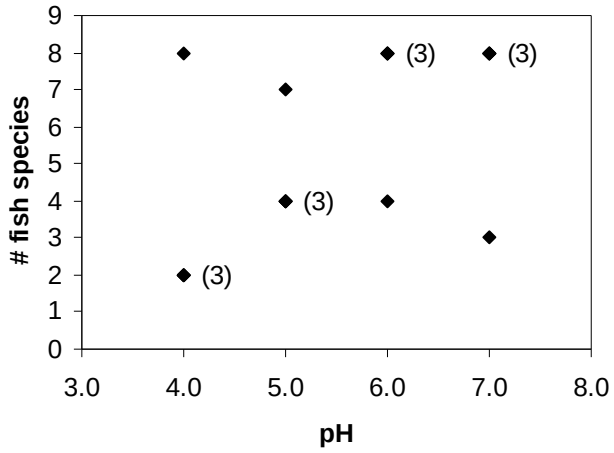

Figure VIIb. Scatterplot showing the same data that are graphed in Figure VIIa except that hidden data are shown by numbers in parentheses. Now the scatterplot shows what may be a positive relationship between number of fish species and pH.

Using Data Labels to Display Hidden Data

- Right click within your scatterplot, then select Source Data. We are now going to Add a new series and use the series name to label hidden data. So, in order to label any points on the graph that correspond to two data points, we need a series named "(2)", to label points on the graph that correspond to three data points, we need a series named "(3)", etc.

- Fortunately, all hidden data in the scatterplot in Figure VIIb need the same label (3), so we only need to create one new series.
- In the Source Data Series window, click Add. In the Name: box, type ="(3)".
- In the X Values: box, you need to select all of the pH values for the points you wish to label. To do this, hold down the ctrl key and click on each individual value that you need. If your data are sorted like those in Table VIIb, you should see "=(Sheet1!$A$2,Sheet1!$A$6,Sheet1!$A$11,Sheet1!$A$15)" in that box.
- Repeat this procedure for the Y Values: box to enter the values for # fish species for the data points that will receive labels. In this box you should see " =(Sheet1!$B$2,Sheet1!$B$6,Sheet1!$B$11,Sheet1!$B$15)". Click OK.
- Your scatterplot should now show different data points on top of the points that need labels. You now need to turn off the new points and add the "(3)" to the right of those points.
- Double click on one of the new data points. On the Format Data Series Patterns window, select None for Marker. On the Data Labels window, check Series Name. Click OK.
- Your graph should now have labels showing the hidden data as in Figure VIIb.

APPENDIX VIII: Tweaking Graphs in MS Excel

Below is a list of tips for modifying how your graph is displayed in MS Excel. Although the list is not comprehensive, it should help with many of the changes you are likely to need to make. Also, I recommend experimenting and exploring the options by clicking and right clicking various parts of your graphs.

Using dummy variables and data labels

- It is often possible to add text to various parts of your graph by adding "dummy" data points and giving labels to those points. For example, to label categories along the x-axis in scatterplots (like those in Chapter 1), you can use dummy variables and give them data labels. To see how this works, try entering the following data.

population	mean	dummy variable
northern	34	0
southern	31	0

- Create a scatterplot with two Series. For the first, use northern and southern for the X Values, and the means for the Y Values. For the second series, use northern and southern for the X Values, but use the dummy variables for the Y Values.
- Once you have finished your graph, you can turn off the numbers associated with the x-axis by double clicking on them, then on the Format Axis Patterns window, select None for Tick Mark Labels.
- Now if you double click on a dummy data point, you can turn it off by choosing None for Marker on the Format Series Patterns window. Then on the Format Series Data Labels window, check X Value and click OK.
- Now double click the word northern (or southern) on the graph, then on the alignment window, choose below for Label Position.
- Important Note: If you change the scale of your y-axis so that the minimum value is not zero, your dummy variables need to be changed. If the minimum value for the y-axis is 20, the dummy variables should have the value of 20.

Other tips for modifying graphs

- To change the scale of an axis, double click on the axis, then go to the Format Axis Scale window. Change the scale by entering the minimum and maximum values you wish to see displayed. Change how many numbers or units are displayed along the axis by changing the Major unit.
- If you wish to go through any of the steps involved in creating the graph, you can click within the chart area (but outside the graph itself) so that the outer boundaries of the chart are highlighted. Then click on the chart wizard button.
- You can also highlight your graph, then choose Chart Options from the Chart menu. This allows you to change the chart type, the source data, the chart options, or even add a trendline to a scatterplot.
- In general, I recommend keeping graphs as simple looking as possible. The goal of the graph is to communicate the essential information. The goal is not to make a flashy looking graph. Often simple graphs with no gridlines and background colors are easiest to interpret. Right clicking on backgrounds, data points, bars, etc. is often the way to find the right options to turn features off or to change colors.

Glossary

alternative hypothesis An hypothesis other than the null hypotheses. Also used to refer to hypotheses when there are several possible explanations for an observed pattern.

analysis of variance A statistical analysis that tests whether there is a significant difference among the means of different groups of numbers.

Anova An abbreviation for an analysis of variance.

bar graph A type of graph where values are represented with bars or columns; also called a bar chart or column chart.

categorical variable A variable that can can be placed into one of a few limited set of values or categories. For example, fur color in labrador retrievers can be categorized as black, yellow, or brown.

Chi-Square Test A statistical analysis that tests whether a set of observed values are statistically different from a set of expected values.

contingency table A table for organizing data that are characterized by two or more categorical variables. Contingency tables are often used for calculating expected values for comparison with observed values in a Chi-Square Test.

continuous variable A variable that is measured numerically and can have a wide range of values. For example, the pH of a solution can vary from 0 to 14 with many possible values along the pH scale.

correlation A method of analysis to explore the relationship between two variables. No cause and effect is hypothesized.

count A statistical parameter referring to the number of observations or data points in a group of numbers. Often synonymous with sample size. Count also refers to a categorical variable in which the numbers of individuals or observations that can be placed into different categories.

data transformation When a mathematical function is applied to data, such as taking the log or arcsine of all data in a data set. Usually this procedure is used so that the data conform more closely to a probability distribution such as the normal distribution.

degrees of freedom A statistical parameter that is related to sample size. In general, the greater the sample size, the greater the degrees of freedom and the greater the ability to detect statistical significance.

df An abbreviation for degrees of freedom.

dependent variable Usually the variable that the researcher is trying to explain. It is the variable that is affected by the independent variable. The dependent variable is sometimes called the response variable. Loosely speaking, it "depends on" the independent variable.

descriptive statistics Statistical parameters such as mean, maximum value, minimum value, etc. that are used to describe the location and spread of a group of numbers.

dummy variable In this manual, dummy variable refers to numbers that are entered into a spreadsheet in order to alter the appearance of a corresponding graph; they are often used to add labels or text to a graph. These numbers do not represent real variables that are part of a data set or statistical analysis.

equal variances Refers to the assumption that the variation within two or more groups of numbers is similar; an assumption for both the t-Test assuming equal variances and Anova.

error bars Vertical lines on a scatterplot or bar graph that represent the amount of variation in a group of numbers. They usually represent either the standard deviation or the standard error.

expected frequency The expected frequency of observations that should fall into a particular category. The sum of expected frequencies for a set of numbers should always equal 1.

expected value The expected number of observations that should fall into a particular category. The sum of the expected values for a set of numbers should always be the same as the sum of the observed values.

F critical In an analysis of variance or a regression analysis, the F critical corresponds to a 0.05 probability that random chance is causing the observed variation or pattern in a set of data. If the calculated value of F is $\geq$ F critical, then $p \leq 0.05$.

frequency distribution See histogram.

hidden data Data that are not represented in a scatterplot because they share identical x and y coordinates with other data points. In an MS Excel scatterplot, only one data point will be displayed in these situations, so hidden data should be represented by adding information to the scatterplot, such as numbers in parentheses near the data point indicating how many observations the point represents.

histogram Histograms (also called frequency distributions) show the frequency of different values in a set of data plotted as a bar graph. The height of the bars represents the frequency of the values on the x-axis. See Appendix V for examples and explanation.

independent Two events are independent if the probability of one occurring is not related to the probability of the other occurring.

independent variable A variable that affects the dependent variable. Also referred to as the predictor variable.

location of data Measures of location summarize where most of the data are found; examples include mean, median and mode.

mean In this manual, mean refers to arithmetic mean. The arithmetic mean is a measure of location that is calculated by taking the sum of all observations (in a group of numbers) and dividing by the number of observations.

mean square A measure of the spread in a group of numbers. In an Anova, the mean square for a particular category is calculated by dividing the sum of squares (SS) by the degrees of freedom.

median The value in a group of numbers that falls in the middle; half of the numbers fall below that value, the other half fall above.

mode The most common value in a group of numbers.

multiple regression A type of regression analysis that includes several independent variables.

negative relationship When there is a negative relationship between two continuous variables and the data are plotted using a scatterplot, a line drawn through the center of the scatter of points will slope downward from left to right (the slope of the line is negative).

non-linear regression A type of regression analysis that does not assume that the relationship between the dependent and independent variables is linear and therefore cannot be represented by a straight line.

non-parametric test A statistical test that does not assume that the data follow a particular probability distribution.

null hypothesis Null hypotheses are statements that any observed variability or pattern in the data is due to random chance. If $p > 0.05$ in the statistical test being performed there is "statistical significance" and the null hypotheses is accepted. If $p \leq 0.05$, there is "statistical significance" and the null hypothesis is rejected.

observation A term used to refer to an individual data point. A group of ten numbers has ten observations.

observed frequency The observed proportion or frequency at which the observations (or raw data) fall into particular categories. It is calculated by dividing the number of observations in a particular category by the total number of observations. The observed frequencies for a given set of data should always sum to a value of one.

observed value The number of observations (or counts) that fall into a particular category.

one-tailed test A type of t-Test where the direction of the difference between the means is predicted prior to analyzing the data; one mean is predicted to have a greater value than the other.

overlap The extent to which the minimum and maximum values of two groups of numbers are similar. If the two groups have identical maximum and minimum values, the two groups overlap entirely. If the maximum value of one group is less than the minimum value of another, there is no overlap.

paired test A type of t-Test that is performed when the observations in one group of numbers are paired with the observation in the second group. For example, if you are comparing the size of male and female birds of a particular species and you take all measurements on mating pairs, the data are naturally paired.

parametric test A type of statistical test that assumes that the data conform to a particular probability distribution.

pattern Any time the variation in a group of numbers is non-random (is related to another variable), there is pattern in the data. Often the goal of scientific research is to identify and explain pattern.

positive relationship When there is a positive relationship between two continuous variables and the data are plotted using a scatterplot, a line drawn through the center of the scatter of points will slope upward from left to right (the slope of the line is positive).

predictor variable An alternate term for the independent variable.

p-value The probability that the variation or observed pattern in the data is the result of random chance. The greater the p-value, the more likely the variation or observed pattern is the result of random chance and the less likely the independent variable is affecting the dependent variable. The lower the p-value, the more likely the independent variable is affecting the dependent variable. If $p \leq 0.05$, the result of a test is said to be statistically significant.

qualitative data Descriptive data that generally cannot be represented numerically and used in statistical analysis.

quantitative data Data that can be represented numerically and used in statistical analysis.

R^2 A statistical parameter in regression analysis that measures the amount of variation in the dependent variable that is explained by variation in the independent variable. In a scatterplot, the higher the R^2 , the more tightly the data will be clustered around the regression line. If all values fall on the line, $R^2 = 1$. The value of R^2 can vary from 0 to 1.

random chance Random chance refers to the effect of random events on a data set. Every data set is subject to the effects of random chance. If the variability in a data set is all the result of random chance, then there is no meaningful influence of the independent variable on the dependent variable.

range The difference between the maximum and minimum values in a group of numbers.

regression A type of statistical analysis that tests for relationships between continuous independent and dependent variables. As used in this manual, the term regression is synonymous with simple linear regression.

regression line In a linear regression analysis, the estimated line that represents the relationship between the dependent and independent variables. The line is characterized by the two variables slope and intercept.

relationship In this manual, I use the term relationship to describe the possible pattern in the data in a regression analysis. If the independent variable accounts for at least some of the variation in the dependent variable, then there is a relationship between the two variables. A regression line with a positive slope indicates a positive relationship; a negative slope indicates a negative relationship.

research hypothesis In this manual, I use the term research hypothesis to distinguish from a null hypothesis. A research hypothesis is any statement that the researcher or scientist proposes to explain the effect of an independent variable on a dependent variable.

response variable A alternate term for the dependent variable.

scatterplot A graph in which data are represented as points in an x y coordinate system. The value of the independent variable is used for the x coordinate and the value for the dependent variable is used for the y coordinate.

spread of data Measures of spread summarize the variability in a group of numbers. If all numbers are near the mean (or median or mode), then there is little spread. If many numbers are far from the mean, there is a lot of spread. Measures of spread include the standard deviation and variance.

standard deviation A measure of spread in a group of numbers. It is equal to the square root of the variance.

standard error A measure of spread in a group of numbers. The standard error is smaller than the standard deviation so it is important to be clear about which measure is being reported.

statistical significance In a statistical analysis, if $p \leq 0.05$, the result is considered statistically significant. This means that the probability that random chance accounts for the variability or the observed pattern in the data is less than or equal to 0.05.

sum of squares (SS) An estimate of the spread or variation in a group of numbers. It is calculated as the sum of the square of the difference between each observation and the mean. This calculation is used in Anova.

t critical In a t-Test, t critical corresponds to a 0.05 probability that random chance is causing the difference in the means between the two groups. If the calculated value of t is $\geq$ t critical, then $p \leq 0.05$ and the difference between the means is considered statistically significant.

transformation See data transformation.

t-Test A statistical analysis designed to test whether the difference in the means of two groups of numbers is statistically significant.

two-tailed test A type of t-Test in which there is no prediction about which group of number has a greater mean value.

unequal variances Refers to the assumption that the variation within the two groups of numbers in a t-Test is not the same; an assumption for the t-Test assuming unequal variances.

unpaired test A type of t-Test in which there is no pairing between the observations in the two groups being compared.

variance A specific statistical parameter that is a measure of spread in a group of numbers. It is calculated by summing the square of the difference between each observation and the mean, then dividing by n-1 where n = the sample size.

variation A general term used to describe the spread in a group of numbers. In general, the greater the spread, the greater the variation.

References Cited

Brower, J.E., J.H. Zar and C.N. von Ende. 1998. *Field and Laboratory Methods for General Ecology*, 4th ed. WCB/McGraw-Hill, Boston, MA.

Gotelli, N.J. and A.M. Ellison. 2004. *A Primer of Ecological Statistics*. Sinauer Associates, Inc., Sunderland, MA.

Krebs, C.J. 1989. *Ecological Methodology*. Harper Collins Publishers, New York.

Snedecor, G.W. and W.G. Cochran. 1980. *Statistical Methods*. 7th ed. The Iowa State University Press, Ames, Iowa.

Sokal, R.R. and F.J. Rohlf. 1995. *Biometry*, 3rd ed. W.H. Freeman & Company, New York.

Southwood, T.R.E. 1988. *Ecological Methods*. Chapman and Hall, New York.

Index