

## Chapter 2

# Introduction to Statistical Inference

*"Facts are stubborn, but statistics are more pliable" – Mark Twain*

Empiricism requires that we scientists learn about our surroundings by conducting experiments, gathering evidence, and analyzing the results. Although this is by no means a perfect method, it is made more perfect when we conduct ourselves in a critical, unbiased manner and devote ourselves to the pursuit of absolute objectivity. And within the realm of human endeavor, mathematics is about as close as we can come to that lofty goal.

So it should come as no surprise that the sciences have always relied upon the virtue of mathematics to make their own reputation more sterling. Thanks to logic and probability theory, mathematicians have been able to demonstrate that under certain controlled conditions, a small subset of measurements can be used to represent the larger universe from which those measurements were taken. That is exactly what makes science both possible and practical—the fact that we are able to describe order and causation in a natural system without having to gather every conceivable bit of information about that system. Surely, the whole is greater than the sum of its parts, but we can still use some of those parts to discover something new. But to do that, we'll have to learn a bit about statistics first.

### Key Concepts

- Any scientific study that relies on numerical data can be tested for using a combination of descriptive and inferential statistical methods.
- Descriptive statistics include those mathematical analyses that can only describe the data that have been collected.
- Inferential statistics rely on the precision and accuracy of the measured data to make predictions about the system from which they were taken.
- Any research hypothesis can be translated into the language of statistics in an effort to either accept or reject the hypothesis being tested.
- Statistical tests are very robust in their capacity to test hypotheses, but they can never provide absolute proof.

## Different Types of Data

As we have seen in our review of empiricism and the scientific method, data are the currency of science. Without data of sufficient quantity and quality, we swiftly discover that our ability to test the suitability of our chosen experimental method, the appropriateness of our original hypothesis, and the fundamental conclusions of our study have been severely compromised. But that's the problem—the determination of what constitutes “sufficient” data. Fortunately, an elementary review of mathematics, and the associated rules of simple arithmetic, will serve us quite well in this regard.

Essentially all of the data you may gather in the laboratory or in the field will fall into one of three general classifications: meristic, metric, or categorical data. **Meristic** (or **discrete**) data are those measures that are represented by integers. Although these measures are **quantitative** in nature, they are incapable of conveying fractional information and can be somewhat limiting insofar as numerical analyses are concerned. Such data are best utilized for simple counts, frequencies, or assessments of a binary state (such as species presence = 1 or absence = 0 within a particular habitat).

**Metric** (or **continuous**) data are those measures that are represented by real numbers; as such, these data include fractional information. Also quantitative in nature, such data are best utilized for those attributes that naturally fall along some continuum of measurement (for example, the body length of a sardine in centimeters or the density of seawater in grams per milliliter). Metric data are inherently more precise than meristic data and are broadly considered to be the most versatile data for numerical analysis. However, as these data are much more dependent on precise measurement, they are also more susceptible to measurement and rounding errors.

**Categorical** data are those measures that are represented not by numbers, but by some subjective quality as defined by the researcher (such as hurricane intensities categorized as weak, moderate, or intense). As a **qualitative** measure, such data are far more difficult to analyze in an objective manner and are typically disfavored in scientific studies and statistical analyses. However, it is possible for the clever researcher to convert qualitative data into quantitative data, so long as meristic or metric data are used to objectively define the limits of the categories being used. For example, the Saffir–Simpson scale of hurricane intensity ([Figure 2.1](#)) is used not only to convey very easily understandable information about hurricane intensity (as a series of meristic categories ranked from 1 to 5), but statistical analyses may also be used for the metric data (by using the sustained wind speed) that are inherent in all modern hurricane intensity classifications that use this scale.

category	winds	damage
1	74–95 mph	minimal
2	96–110 mph	moderate
3	111–130 mph	major
4	131–155 mph	extensive
5	>155 mph	catastrophic

**Figure 2.1** The Saffir–Simpson scale uses metric data of sustained wind speed to create a more easily understandable ranking system to define categories of hurricane intensity. This is an example of how metric data can be used to create categorical data, thus allowing the researcher to present and analyze the same fundamental data, but in different ways.

Although these three data classifications are applicable in the general sense, when we consider “attribute-specific” data, it is often helpful to define our measurements as being either scale, ordinal, or nominal. To illustrate the differences between these types of measurements, let's assume we were investigating the sedimentary environment at multiple locations, and we decided to study the different sizes and shapes of the sediment grains to accomplish our goals ([Figure 2.2](#)). These are certainly features that we can measure, but it is important to understand what type of measurements we are taking.

**Scale** measurements are those that can be ordered according to a continuous scale, typically with a natural, meaningful metric. In the context of science, scale measurements are the ideal type of measurement, because they

are completely objective in nature. For example, if we were measuring sediment grain sizes, we would expect the diameter of each sediment grain to fall along a continuous length scale (from  $<0.002$  mm for clay particles, all the way to  $>2$  mm for sand grains, and beyond). If we wanted to sort our grain size measurements, it would be a very simple task to order them according to increasing particle diameter. Different researchers might choose to categorize the sediment type using a subjectively defined size class (like “clay” or “sand”), but the fundamental measurement of sediment grain diameter is a natural, meaningful metric of length that is, in itself, immune to subjective definition.



**Ordinal** measurements are those that represent categories within a logically defined (or inherent) ranking. With regard to sediment grains, we might be able to define their overall shape by some logically defined ranking based on how oblong (or round) the grain shape is. That would allow researchers to define the degree of “sphericity,” from low to high. If we assign an artificial number to what we’ve defined as low sphericity (1) versus high sphericity (3), we can still order the sediment grains according to our ranking system. The same could be done for the angularity of the sediment grain, but neither represents a “natural” metric. Even if we analyze sediment type by using the diameter of the grain to define our ranked size classes, ordinal measures are inherently subjective (but at least we’re making a logical attempt to define them as objectively as we possibly can).










**Nominal** measurements are those that represent categories that cannot possibly be ranked in an ordered fashion. As a general rule, these types of measurements are almost useless in a scientific context and should be avoided whenever possible. However, just as we discussed with regard to categorical data, it may be possible to create a system by which nominal measurements can indeed be ranked (thereby “transforming” them into ordinal measurements). For instance, if we had initially graded our sand grains according to their predominant color, we might end up with a jumbled list of brownish-red, yellowish-gold, and olive-black color categories. Although these categories may not at first allow any kind of ranking, if we digitally scanned the images of these sediment grains, our computer would render these colors according to a fixed RGB scale (with each color channel ranging from 0 to 255). Now we have transformed our measurements of color from nominal to ordinal, thereby allowing us to order our sediment grains according to a three-dimensional scale of color (RGB) and along a ranking scale of 0 to 255 in each color channel. This also brings us a bit closer to defining color as an objective, rather than subjective, measure.

### Every Measurement is Limited by Imperfections in Precision and Accuracy

Despite our best efforts as scientists, the natural world has a very frustrating habit of introducing error in just about everything we try to do. Even if we wanted to do something that is conceptually simple (like measure the diameter of a grain of sand), we are fundamentally limited in our capacity to do so. That limitation is borne from two confounding facets of reality that we must confront every time we seek to take a measurement: the notion that neither the **precision** nor the **accuracy** of our measurements can ever be perfect.

No matter what we choose to measure, we must do so by utilizing an instrument specifically designed for that purpose. From a philosophical perspective, instruments of human design and fabrication are inherently flawed. Unfortunately, the imperfections of our measuring devices will ultimately

grain diameter	type	
256 mm and up	boulders	gravel
64–256 mm	cobbles	
2–64 mm	pebbles	
0.0625–2 mm	sand	.....
0.002–0.0625 mm	silt	
0.002 mm and smaller	clay	

		shape		
sphericity	high 3			
	2			
	low 1			
		high 3	2	low 1
		angularity		

**Figure 2.2** Sediment grain diameter, type, and shape (angularity and sphericity) are common measures taken by sedimentary geologists. These measures also serve as an excellent example of how scale and ordinal measurements can be represented within the same dataset.



### MAKE YOUR MEASUREMENTS COUNT

In the context of science, the usefulness of your data will greatly depend upon whether you are taking scale, ordinal, or nominal measurements. As a general rule, the most objective measures are also the best, meaning scale measurements provide the best data while nominal measurements are the least useful.



scale



ordinal



nominal



translate to an imperfect measure. When we consider the precision of an instrument, we are limited to the smallest division on the instrument's scale of measurement. For a typical handheld ruler, the smallest division might be 1 mm. Although it is certainly possible that the sand grain we are measuring might be fractionally smaller than 1 mm, we are limited by the **resolution** of our instrument and can only measure to the nearest millimeter. Thus, the precision of our measurement will be likewise limited to  $\pm 1$  mm (provided that we did not **interpolate** between the divisions on our instrument's scale).

Although it is impossible to escape the fact that every instrument is inherently imprecise, we can employ our technology to design and construct new instruments of ever-increasing (but never perfect) resolution. A simple way to analyze your instrument's ability to provide sufficiently precise measurements is to determine its **absolute uncertainty (AU)**, which is simply defined as half the instrument's resolution. This value can then be used in Equation 2.1 to determine the **relative uncertainty (RU)** of any measurement taken with that instrument:

$$RU = \left( \frac{AU}{\text{measured value}} \right) \times 100\% \quad (2.1)$$

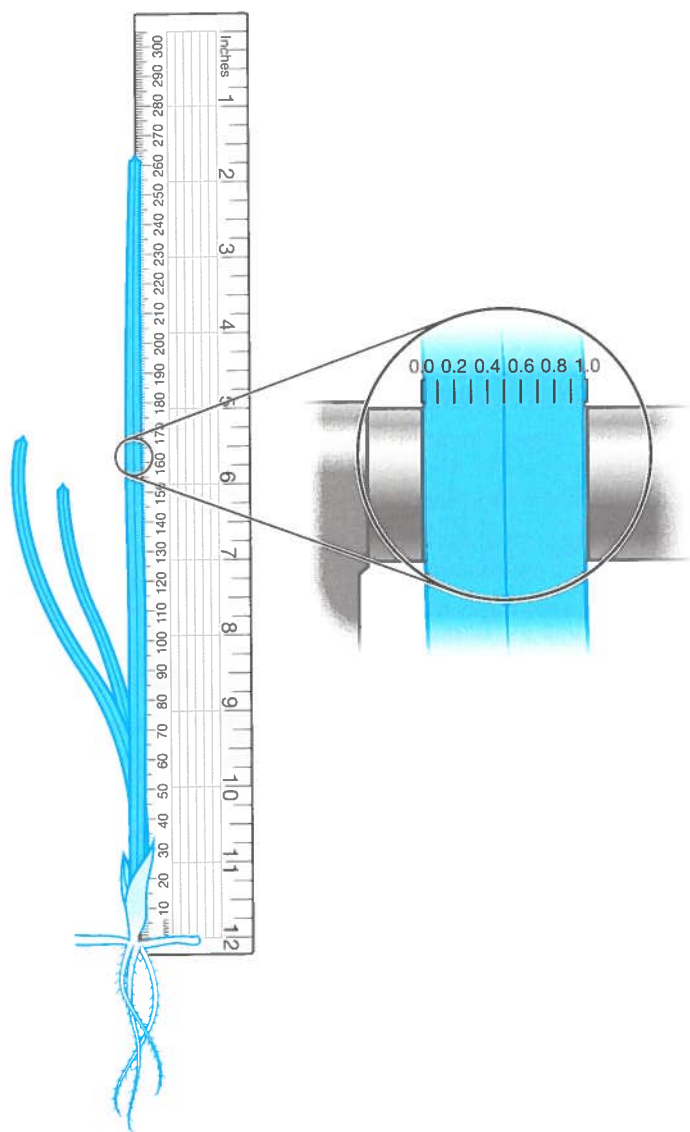
*AU* directly reflects the overall resolution of the instrument, but *RU* offers a more general way to evaluate the information content of a measurement. Generally speaking, the lower the *RU* of a measured value, the more **significant figures** it contains (and that's a good thing, because a lower *RU* is indicative of increased precision, which will be necessary to achieve greater accuracy in our measurements).

Few scientists would be comfortable admitting such a thing, but the quest for accuracy, though noble, is sadly unachievable in science. This notion of accuracy is more fitting in a philosophical context, as accuracy is essentially truth, in the grandest sense of the word. Whenever we seek to measure some aspect of the natural world, we seek a perfectly accurate measure—the truth. Forever just out of reach, we design instruments of greater and greater precision as we court the truth, but with imprecision comes error, and through error accuracy and truth escape.

Overly dramatic? Perhaps. But remember: as a scientist, it is critical that you remain a vigilant skeptic. To understand the limitations of the scientific method, both in theory and in practice, is to be better prepared to describe those phenomena that have escaped the abilities and intellect of scientists whom you now follow.

### The 30–300 Rule is Used to Determine the Proper Instrument of Measure

At some point in your research, you will be forced to choose which instrument of measure you wish to use. Instruments with very high resolution will of course provide very precise measurements, but then the question becomes: what level of precision is “good enough” for the task at hand? As an example, let's say we chose to use a metric ruler ( $\pm 1$  mm resolution) to measure the length and width of seagrass leaf blades (**Figure 2.3**). A length measurement of 97 mm would allow us to use Equation 2.1 to calculate an *RU* of only (0.5 mm / 97 mm), or 0.5%. If our width measurement were only 3 mm, our calculated *RU* for the width measurement would be substantially larger: (0.5 mm / 3 mm), or 16.7%. The dramatic difference in uncertainty between our length and width measurements certainly calls into question



**Figure 2.3** The resolution of your instrument will define both the precision and relative uncertainty of every measurement. Take care not to use the same instrument to quantify very different scales of measurement (for example, the length and width of a single blade of seagrass).

whether we should be using an instrument with greater resolution to measure the widths of the seagrass blades and therefore improve the precision of those measures.

As a common rule of thumb, field researchers often use the “30–300 rule” to determine the most appropriate level of precision necessary when taking measurements. The 30–300 rule simply states that your chosen instrument should provide between 30 and 300 unit steps between the minimum and maximum measured values. For example, if our shortest measured blade length was 48 mm and the longest was 104 mm, we would have  $(104 - 48 =) 56$  unit steps at 1 mm resolution. Since that falls within the 30–300 range of acceptability, we would be well advised to use a ruler with  $\pm 1$  mm resolution to measure leaf-blade lengths. However, if our minimum and maximum measured blade widths were, respectively, 1 mm and 5 mm (that is, only 4 unit steps), it would be clear that our chosen ruler does not possess sufficient resolution. If we instead used a stereoscope with a micrometer plate with  $\pm 0.1$  mm resolution, our measurements would range between 1.0 mm and 5.0 mm: that’s 50 unit steps at 0.1 mm resolution, satisfying the 30–300 rule.

### The Rules of Significant Figures Define the Precision of Any Measurement

One of the most difficult aspects of using metric data stems from the fact that any real number can vary to an infinite number of decimal places, but our measurements will be limited to only a few significant figures within that infinite string of decimal places. In other words, significant figures are those digits that contribute to (and essentially define) the precision of a particular measurement. In practical terms, the rules for identifying the appropriate number of significant digits require that

- All exact numbers (such as counts) have infinite significant digits.
- All nonzero digits are significant.  
( $\pi = 3.14159$  has 6 significant digits, as written)
- Any zero digits located anywhere between nonzero digits are significant.  
(1.002205 has 7 significant digits)
- Any leading zero digits are not significant.  
(0.0067 has 2 significant digits)
- All trailing zero digits in numbers containing a decimal are significant.  
(0.006700 has 4 significant digits)
- Any trailing zero digits in numbers without a decimal are ambiguous.  
(12,500 has either 3 or 5 significant digits)

In that last example, it would be impossible to know the significant digits in the value 12,500 unless we also knew the resolution of the instrument that was used to determine that value. If the precision of measurement is  $\pm 100$ , the value 12,500 would have 3 significant digits. However, if the precision of measurement is  $\pm 1$ , it would have 5 significant digits. In order to avoid confusion, it is always best to cite numbers in **scientific notation**. If the value 12,500 were written instead as  $1.25 \times 10^4$ , it would clearly possess only 3 significant digits. If we wrote 12,500 as  $1.2500 \times 10^4$ , the “trailing zero” rule for decimal numbers would clearly indicate the value possesses 5 significant digits.

Because significant digits are linked so intimately with precision, strict adherence to these rules is one of the most important steps we can take to minimize error, especially when performing arithmetic. This is especially true if we are using measurements obtained from several different instruments, each with its own limit of precision (**Technical Box 2.1**).



#### HOW TO APPLY THE ARITHMETIC RULES OF SIGNIFICANT FIGURES

To prevent rounding errors, it's always a good idea to keep as many digits as possible while performing intermediate calculations. But when the calculations are done, always cite the final answer according to these rules:

##### For multiplication or division:

The final result should have the same number of significant digits as the measurement with the least number of significant digits.

##### For addition or subtraction:

The final result should have the same number of decimal digits as the measurement with the least number of decimal digits.

If both rules are applicable, the multiplication/division rule takes precedence.

### For Every Measure there is Error

As we have discussed, the process of science is not borne from data, but from the original observations made (our perception) and the question we ultimately seek to answer (our hypothesis). Notwithstanding the clarion warnings of Descartes regarding the fidelity of our own senses, the strength of science lies in the fundamental fact that if something is capable of being perceived, it is also capable of being measured. Fallible as the scientific method may be, we are not completely powerless in our ability to suppress error and uncertainty in its practice.

If we were capable of perfect knowledge, our measurements would define the true properties of the universe, with unfailing accuracy. As we discussed earlier, what is truly unfailing is the inescapable introduction of error as a consequence of our own fumbling, the imperfection of our instruments of measure, and the variability that exists in the natural world. Therefore, the cloud that obscures our view of the truth, by corroding the accuracy of our measurements, is a simply conceived but infinitely complex error function.

## TECHNICAL BOX 2.1

### An Example of the Proper Use of Significant Figures in a Calculation

Ol' Dusty is interested in measuring the slope of the beach at his study site, and decides on an antiquated but tried-and-true surveying technique whereby two poles of equal length are used in the vertical (**Figure 1**). Since the poles are exactly the same length, any change in beach slope between positions 1 and 2 can be easily measured (as  $b$ ), so long as Dusty consistently sights the top of pole 2 on the distant horizon. If Dusty has a trusty sidekick to measure the distance between positions 1 and 2 (as  $a$ ), he can use the Pythagorean theorem and some basic trigonometry to compute the slope-angle ( $\theta$ ).

While Dusty marked his survey poles with 1 cm resolution, his trusty sidekick brought a nonmetric tape measure with 1/16-inch resolution. To perform the calculations necessary to determine  $\theta$ , Dusty must first convert inches to centimeters to make sure his units are consistent throughout the calculations:

$$a = 15' 11 \frac{7''}{16}$$

$$b = 9 \text{ cm}$$

Ignoring significant digits:

$$a = \left( \frac{15 \text{ ft}}{1} \right) \left( \frac{12 \text{ in}}{\text{ft}} \right) \left( \frac{2.54 \text{ cm}}{\text{in}} \right) + \left( \frac{11 \text{ in}}{1} \right) \left( \frac{2.54 \text{ cm}}{\text{in}} \right) + \left( \frac{7 \text{ in}}{16} \right) \left( \frac{2.54 \text{ cm}}{\text{in}} \right)$$

$$a = 457.2 \text{ cm} + 27.94 \text{ cm} + 1.11125 \text{ cm}$$

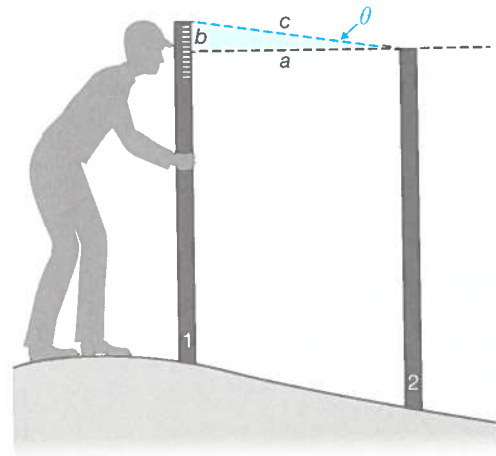
$$a = 486.25125 \text{ cm}$$

$$\theta = \tan^{-1} \left( \frac{9 \text{ cm}}{486.25125 \text{ cm}} \right) = \tan^{-1}(0.018508949) = 1.060363634^\circ$$

So when (and where) should Dusty round off? Or do we really believe Dusty could measure  $\theta$  to 10 significant digits with such a crude surveying device?

Since the measurement with the least number of significant digits (and the least number of decimal digits) is 9 cm,  $\theta$  should be rounded to a single significant digit; therefore,  $\theta = 1^\circ$ .

Keep in mind that Dusty's instrument of measure was a crude one, with only  $\pm 1$  cm resolution (that is, 1 significant figure). Since our final answer must possess no more than a single significant figure, it stands to reason that any calculated result  $0.5^\circ < \theta < 1.5^\circ$  would be rounded to yield  $\theta = 1^\circ$  according to the arithmetic rules of significant figures. That leaves a lot of room for error! Perhaps Dusty should have paid more attention to the 30–300 rule and chosen a more precise instrument of measure.



**Figure 1** An old surveying technique used to calculate beach slope ( $\theta$ ) involves the use of two vertical rods of standard height, set apart from each other by some distance ( $a$ ). The angle ( $\theta$ ) is then determined by the height difference from some level reference ( $b$ ) and a little trigonometry.

Though we may not be capable of eliminating human frailty from the practice of science, at least we have some control over it (and can therefore define the limits of that uncertainty). Of course, our ability to control variability in the natural world is not so straightforward. In our study of the natural world, we seek not only to measure its basic properties, but also to discover the source of any changes to those properties. And since change implies cause and effect, we are compelled to engage our intellect to describe order in our universe.

That is perhaps the defining purpose of the sciences, and the best argument for a mathematical analysis of order in our universe. It is no mistake that the world's greatest philosophers were originally mathematicians by trade, and



it is likewise no mistake that mathematics should play such a central role in the sciences. Mathematics grants us the essential ability to understand order in our time.

More often than not, it is impossible for us to see the whole picture while we are trying to investigate order in the world around us. That means we usually have to take a more pointed look at a specific element—just one piece of the puzzle—and try to formulate an explanation of the larger whole. But that, in and of itself, presents a problem. Aristotle said it first, best, and most famously: “The whole is greater than the sum of its parts.” And although that may be true, it is an untenable position in the practical world of science. Since it is impossible for us to possess perfect awareness of all the natural factors that could affect our measurements, we are hopeless to know the “truth of the whole” (what we might call Aristotelian holism). So we are reluctantly forced to accept the inherent limitations of reductionism; that is, that we are only capable of measuring some small part of reality and must assume that it faithfully reflects the larger reality we seek to define.

In the world of science and mathematics, we call that “subset analysis.” In a few very special cases, mathematicians can use a subset of numbers to prove a mathematical truism. In the natural world, which is plagued by far more uncertainty, we use subset analyses to determine the applicability (or appropriateness) of our conclusions in a much larger context, reaching far beyond our original subset. Here you enter the world of statistics and probability; a world where nothing can be proven, but where our intellect can find order in nature, as supported by evidence of acceptable significance.

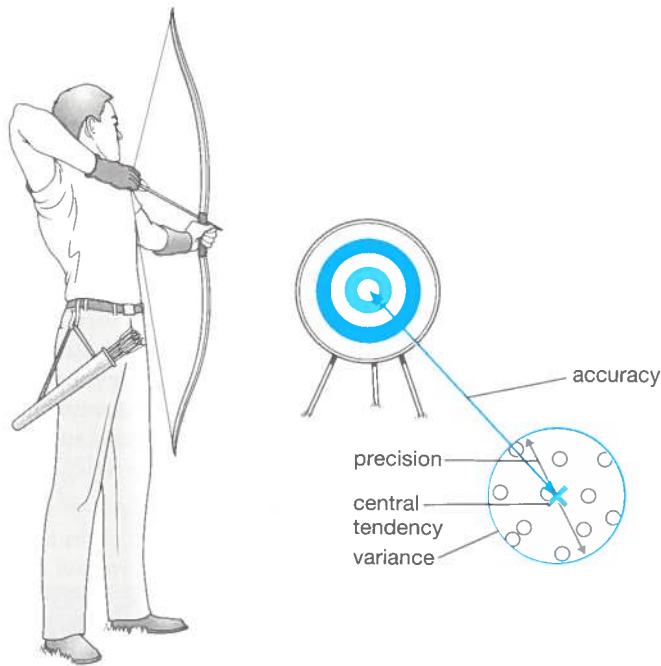
In the world of statistics, we can choose one of two general paths. If we wish to describe only the subset of data that was actually measured (the **sample**), we would use **descriptive statistics**. Descriptive statistics cannot be used to infer any information from other subsets of data for which no measurements were taken, nor can descriptive statistics be used to form conclusions about the larger **population** of data from which the subset was taken. To do that, we must employ **inferential statistics**—a host of mathematical techniques that are used to make predictions about other subsets (or the larger population) of data based upon the subset for which we have data. Let’s begin by focusing on descriptive statistics, and what those numbers can tell us about our data.

### Descriptive Assessments of Central Tendency are Used for Accurate Representations of the Variability in a Sample

With so much talk of uncertainty and error, statistics do offer some good news. As an interesting consequence of number theory, those properties that have been measured are almost completely unimportant; the numbers that represent those properties are what’s important in statistics. What makes statistics such a powerful tool in the sciences is its ability to provide numerical justification for the reliability and completeness of our data (which also helps us to design stronger laboratory and field methods).

Let us imagine that all the data we can gather—the full complement of measurements that can be taken to define some property in nature—is best represented by an archery target (Figure 2.4). The exact center of that target is the one, true value that exists in reality. As we align our bow and fire our arrow at the target, we are besieged by many sources of error, such as an errant breeze, an imperfection in the shaft of our arrow, or the poorness of our aim: all of these will conspire to deflect our arrow from its true path, causing it to strike the target off-center.





**Figure 2.4** The archer attempting to strike the bull's-eye is a fitting analogy to the scientist attempting to take a perfectly accurate measurement. Despite the archer's inability to strike the bull's-eye, he can analyze the central tendency of his arrows and their spread about the central tendency (or variance) to make the appropriate adjustments to improve accuracy. (Courtesy of Bernhard Thierry/CC-BY-SA-3.0.)

Although it is certainly possible, it is highly unlikely that a single shot will strike the bull's-eye. Of course, our chances of hitting the bull's-eye at least once are greatly improved as we take more and more shots—as we take more and more measurements. The better our method (and the better our equipment), the more consistently we will strike the target, and the tighter the grouping of our arrows. By analyzing the **central tendency** of our arrows and how tightly our arrows are grouped (that is, the **variance**), we can say something intelligent about our accuracy and how much error is revealed by the spread of our arrows.

In descriptive statistics, we can determine the central tendency of our measurements in three general ways: by calculating the mean, median, and mode. The most common of these is the **mean** ( $\bar{X}$ ), defined as

$$\bar{X} = \frac{\sum X}{N} \quad (2.2)$$

where  $\sum X$  is the sum of all measurements ( $X$ ) taken to describe a particular property and  $N$  is the total number of measurements taken in the effort to describe that property. In most cases, the mean is the most appropriate measure of central tendency for metric data as demonstrated in **Example Box 2.1**.

The **median** is defined as the midpoint of a range of measures that have been organized in **rank order**. Any odd number of observations will have a natural median, but an even number of observations will require the interpolation of an artificial median. This is typically done by taking the simple average of the two values surrounding the theoretical midpoint in the distribution. In instances where the same value is repeated within the dataset, each value must be included in the rank order when determining the median (as demonstrated in **Example Box 2.2**). Note that the determination of the median depends solely upon whether the number of observations ( $N$ ) is odd or even; beyond that, the median is completely insensitive to the magnitude of  $N$ .



### DOES THE NUMBER OF OBSERVATIONS "COUNT" AS A SIGNIFICANT FIGURE?

In the example given, our mean of 35.4 has three significant digits. But what about  $N$ ? Shouldn't our answer only have one significant digit, like  $N$ ? After all, it was used in Equation 2.1 to calculate the mean—or are we breaking the rules?

Recall that the rules of significant digits apply specifically to our measurements.  $N$  is not really a measurement of the property we are investigating in this example; in fact,  $N$  is an exact number because it represents the exact number of measurements taken. Thus,  $N$  has an infinite number of significant figures (so we can essentially ignore it).

## EXAMPLE BOX 2.1

### Determination of the Mean

12.2   23.8   54.7   17.8   35.2   68.9    $N = 6$

$$\text{Mean} = \left( \frac{212.6}{6} \right)$$

$$\text{Mean} = 35.4$$

Note that when calculating the mean, it is not necessary to rank the data in any logical order. In this case, the mean value of the six measurements above is in fact 35.43333333; however, our rules for significant digits indicate that all of our measures possess three significant digits, so our calculated mean should possess only three significant digits as well ( $\bar{X} = 35.4$ ).

## EXAMPLE BOX 2.2

### Determination of the Median

87.4   88.2   89.6    $N = 3$

$$\text{Median} = 88.2$$

87.4   88.2   89.6   90.2    $N = 4$

$$\text{Median} = \left( \frac{88.2 + 89.6}{2} \right) = 88.9$$

When calculating the median, repetitive measures must be included in the rank order of all values.

87.4   88.2   89.6   89.6   90.2    $N = 5$

$$\text{Median} = 89.6$$

The **mode** is simply defined as the most frequently appearing value within the range of measurements. Although the mode is quite useful in defining frequencies of occurrence, it is also the least stable measure of central tendency, as changes in the mode are highly likely during sample-to-sample comparisons. Modal analyses may also lead to the determination of multiple modes within the same sample, as demonstrated in **Example Box 2.3**.

## EXAMPLE BOX 2.3

### Determination of the Mode

15   20   20   20   25   35   35    $N = 8$

$$\text{Mode} = 20$$

15   20   20   20   25   35   35   35    $N = 9$

$$\text{Modes} = 20 \text{ and } 35$$

Although it is certainly possible that all three assessments of central tendency will yield similar results, it is far more likely that they will differ from each other. In these cases, it is worthwhile to examine not just the central tendency of our data, but its **variance** as well. Luckily, there are several varieties of statistical software available that can be used to import our data and render it graphically so we can literally see what the totality of our data look like, and analyze how they are distributed around the central tendency.

As we saw in the analogy of the archer in Figure 2.4, when we take a look at the distribution of our data, there should be a higher probability that a particular measurement will fall close to the central tendency. That also means that any measurements in our dataset that are quite distant from the central tendency should be few and far between. By simply looking at the patterns of how our data are distributed around the central tendency, we can decide how best to proceed in our statistical analyses of that data.

### “Normal” Data Can be Plotted As a Gaussian Curve

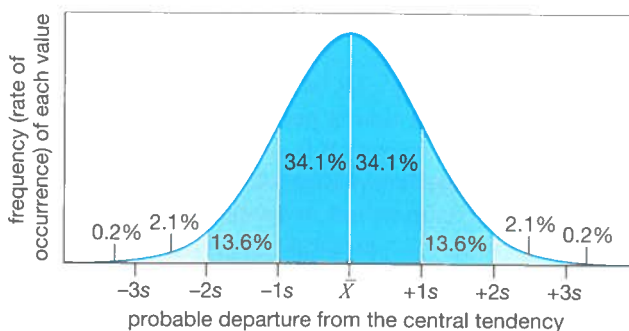
Within the context of statistics, if the mean, median, and mode all share the same value, we say that the data are **normal**. When plotted, normally distributed data will possess the classic bell shape—this is what’s known as a **Gaussian curve** (Figure 2.5). What defines the Gaussian (normal) curve is that it is symmetrical about the mean ( $\bar{X}$ ) and that its first **standard deviation** ( $\pm 1s$ ) is the distance between  $\bar{X}$  and the inflection point on the curve. In a perfectly normal distribution, 68.2% of all the data will fall within one standard deviation ( $\pm 1s$ ) of  $\bar{X}$ , 95.4% will fall within  $\pm 2s$ , and 99.6% will fall within  $\pm 3s$ .

The standard deviation is the most frequently used estimate of variability in the sample data, as it represents the basic tendency of the sample data to depart from the sample mean. In mathematical terms, the standard deviation  $s$  is calculated using Equation 2.3:

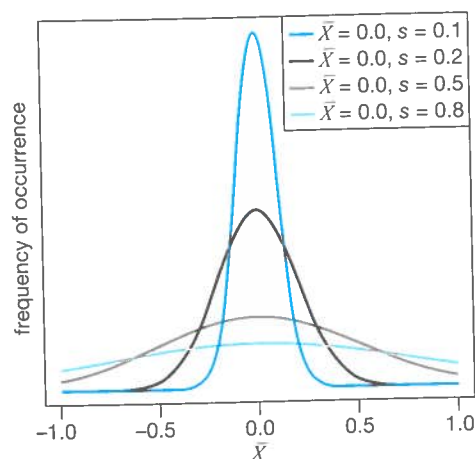
$$s = \sqrt{\frac{\sum d^2}{N - 1}} \quad (2.3)$$

where  $d$  represents the departure of a measured value  $X$  from the mean ( $d = X - \bar{X}$ ) and  $\sum d^2$  is the sum of all calculated values of  $d^2$  in the distribution. As in Equation 2.2,  $N$  is simply the total number of measurements in the dataset.

As a practical matter, the magnitude of the standard deviation  $s$  is an excellent indicator of just how well each measurement within the distribution comports with the mean. From an experimental point of view, a small standard deviation is indicative of low variability, which is exactly what we want. Let’s assume for the sake of argument that our central tendency is in fact



**Figure 2.5** Among normally distributed data, all but 0.4% of the data are described by the classic Gaussian curve,  $\pm 3s$ . This is particularly powerful as a descriptor of data, as this mathematical relationship can be used to make statistical inferences about any particular measure within the data (in relation to where it falls along the Gaussian curve). (Courtesy of Jeremy Kemp / CC-BY-2.5.)



**Figure 2.6** Although all four of these Gaussian curves exhibit the same mean, their distributions are hardly equal. When standard deviations are very small (the preferred case in the sciences), the data are distributed very near to the sample mean (a leptokurtic condition). As the magnitude of the standard deviation increases, so too does the dispersion of the data (a platykurtic condition). (Courtesy of Matthieu / CC-BY-SA-3.0.)



### THE PROPER WAY TO CITE DATA

To better convey the true variability of data, it is always advisable to cite both the sample mean and the standard deviation. For the data in Figure 2.6, we would cite these findings as

$$\bar{X} = 0.0 \pm 0.1$$

$$\bar{X} = 0.0 \pm 0.2$$

$$\bar{X} = 0.0 \pm 0.5$$

$$\bar{X} = 0.0 \pm 0.8$$

the “true” value of what we’re trying to measure in nature. If we took multiple measurements, and every single measurement hit the bull’s-eye, every measurement would be exactly equal to the central tendency, there would be no variability in our measurements, and our standard deviation would be zero. Alternately, if we fired an arrow exactly 10 cm to the right of the bull’s-eye, and the next arrow was exactly 10 cm to the left, the central tendency would still be right smack in the middle of the bull’s-eye, but there would be a lot more variability in our measurements, as evidenced by a large standard deviation.

This is exactly why it is common practice in the sciences to cite both the mean and the standard deviation of each sample, as it is the combination of these two measures of dispersion that will reveal the most information about the reliability of the data. The smaller the standard deviation, the more closely our measurements converge on the central tendency, which in turn provides us with greater confidence that our measures are indeed accurate. Sometimes it is more helpful to visualize the relationship between the central tendency and the standard deviation by analyzing the **kurtosis** of the Gaussian curve (Figure 2.6).

Another useful measure is the **standard error of the mean**, which represents the standard deviation of the error in the sample mean, relative to the true mean of the population. In other words, the standard error can be viewed as an indicator of how well (or how poorly) the mean of your sample represents the larger population from which the sample was taken. Mathematically, the standard error (SE) is calculated using Equation 2.4:

$$SE = \frac{s}{\sqrt{N}} \quad (2.4)$$

where  $s$  is the standard deviation of the sample and  $N$  is the number of observations within the sample. Just as lower values of the standard deviation  $s$  are indicative of greater precision in our measurements, lower values of  $SE$  provide increasing confidence that our sample mean is an accurate estimate of the true population mean.

Instead of looking at the dispersion of the data as a whole, sometimes it is helpful to determine how many standard deviations a particular value deviates from the mean. This can easily be accomplished by using Equation 2.5 to convert any measurement of  $X$  into a **z score**, where

$$z = \left( \frac{X - \bar{X}}{s} \right) \quad (2.5)$$

You can use  $z$  scores to determine the position (or performance) of a particular value in relation to the rest of the data in terms of the standard deviation (Example Box 2.4).

### The Coefficient of Variation Can be Used to Measure Variability

Another estimate of variation that is particularly useful in the sciences is the **coefficient of variation**  $cv$ , which is calculated using Equation 2.6 and represents the relative magnitude of the standard deviation to the mean as the ratio:

$$cv = \left( \frac{s}{\bar{X}} \right) \cdot 100\% \quad (2.6)$$



## EXAMPLE BOX 2.4

### Using z Scores to Determine Single-Value Deviations from the Mean

Amy and Joe are thoroughly enjoying their Marine Field Methods class, despite their somewhat lackluster performance on the exams:

Test	$\bar{X}$	$s$	Amy	Joe	Amy	Joe
			$\bar{X}$	$\bar{X}$	$z$	$z$
1	75	14	97	82	1.6	0.5
2	40	15	31	14	-0.6	-1.7
3	22	4	28	20	1.5	-0.5
4	144	38	148	194	0.1	1.3
$\bar{X} =$	70.25		76.0	77.5		
$\Sigma z =$					+2.6	-0.4
$\bar{z} =$					+0.65	-0.1

If the instructor merely considered each student's mean score on the exams, it would appear as though Joe ( $\bar{X} = 77.5$ ) did slightly better than Amy ( $\bar{X} = 76.0$ ). However, upon further analysis of their z scores, it is obvious that Amy's performance is much more consistently above the mean (as indicated by the positive z score). Amy's overall performance in the class was +0.65 standard deviations above the mean, whereas Joe consistently underperformed compared with the rest of the class (averaging -0.1 standard deviations below the mean).

Since the area beneath a normal curve contains 100% of the data, the number of standard deviations from the mean also defines the percent of data that fall above, or below, that standard deviation. That is why it is more important to consider the percentile ranking of a student's performance on a standardized test, rather than the raw score itself: z scores are what make that analysis possible.

Since a small standard deviation is the preferred condition, a small *cv* is likewise indicative of low variability and is therefore preferred above large *cv* values.

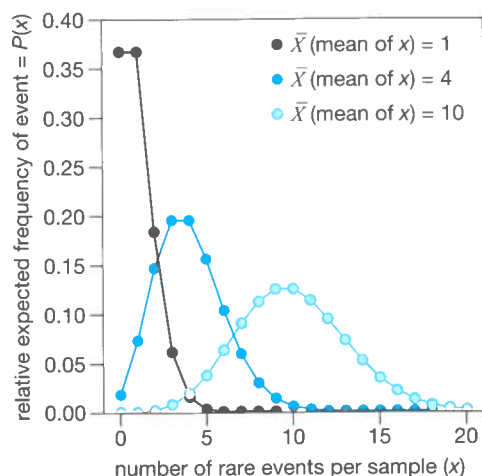
Perhaps the simplest measure of data dispersion is the sample **range** *R*, simply calculated by Equation 2.7 as

$$R = \text{largest value} - \text{smallest value} \quad (2.7)$$

Although it can be a useful measure in certain circumstances, the range is generally considered to be an inferior measure of dispersion because it lacks information about the distribution of data that falls within that prescribed range.

Another simple way to examine the dispersion of data is to calculate the sample **variance**, which is related to the standard deviation by the simple arithmetic function described by Equation 2.8, where

$$\text{variance} = s^2 \quad (2.8)$$



**Figure 2.7** As a probability distribution, Poisson curves can become normal as  $\bar{X} \rightarrow x$  as the number of true counts ( $x$ ) increases. (Courtesy of Skbkekass / CC-BY-SA-3.0.)



### TRANSFORMING YOUR DATA ISN'T CHEATING—IT'S NECESSARY!

At first, it may seem as though data transformations are a way to “cook the books.” Before you get too troubled by this notion, remember that statistics utilize the variance and central tendency of the data, regardless of what those data are.

So long as mathematical transformations are applied uniformly to the dataset, they will not change the information content of your data in the least. Consider this: if you converted a temperature reading from degrees Celsius to Kelvins, you would have to apply the mathematical function

$$K = ^\circ C + 273.15$$

The data have been mathematically transformed, but because the transformation has been applied uniformly to the entire dataset, the information contained within those data remains unchanged.

From a mathematical perspective, it is easy to see why a small standard deviation is preferred, since the variance is an exponential function of  $s$ . A low standard deviation (and therefore, a low variance) indicates that all the values in the sample are tightly centered about the mean.

What makes the variance so special is that it has been proven by mathematical **axiom** that the distribution of any sample can be described if it is normal and both the mean and variance are known. This also means that if any two normally distributed samples possess the same mean and the same variance, they are considered to be statistically identical to each other. That critical relationship is the foundation upon which inferential statistics has been built—it allows researchers to infer relationships between a subset of data and the larger population from which the subset was gathered.

### Data Distributions That Aren't Normal Often Follow a Poisson Distribution

While normally distributed data are the bread and butter of statistics, real data are almost never normal. Very often, data instead follow a **Poisson** distribution (Figure 2.7), a geometric curve that represents the true count of independent or rare events ( $x$ ) that occur in a specified interval of time against the expected mean  $\bar{X}$  over that same interval of time. Poisson distributions are generally expected when your data are meristic (that is, discrete measurements represented by integers).

Since the Poisson curve describes how often an event will occur relative to its expected rate of occurrence, it is as much a probability distribution as it is a frequency distribution (rather than a classic data distribution). However, if the data being gathered are time sensitive or possess some probability of being included (or excluded) from the sample, the results might resemble a Poisson curve rather than a Gaussian curve. If the variance of your sample is equal (or nearly equal) to the mean, it is a Poisson distribution.

Another common type of frequency distribution is the **chi square**, which most often resembles the extreme case of  $\bar{X} = 1$  in the spectrum of Poisson curves depicted in Figure 2.7. The chi-square distribution is indeed a special case and is most often used to test statistical hypotheses, or the “goodness of fit” between a measured distribution and a theoretical distribution.

### Nonnormal Datasets Can be Normalized Using Data Transformation

Although a more accomplished statistician may find Poisson and chi-square distributions to be useful in their own right, most scientific applications of statistics are best served if the data are both continuous (metric) and normal. Frequently, nonnormal datasets can be normalized to better resemble a Gaussian curve through the use of **data transformation**.

Data transformation is a simple process by which all the values within a non-normal dataset are recalculated according to a uniformly applied mathematical function and then analyzed for their new resemblance to the Gaussian curve. Although the number of potential transformation methods is only as limited as the imagination of the researcher, Equations 2.9–2.12 are just a few common transformation methods that seem to work best (and should be attempted first).

$$X' = \log_{10}(X + n) \quad (2.9)$$

The log transformation in Equation 2.9 helps to make **skewed** data more symmetrical and is most appropriate for data that are either related to some measure of biological growth or for counts, especially when the variance of the counts is larger than the mean. If zeroes in the dataset ( $X = 0$ ) should be counted, use the condition  $n = 1$ ; otherwise,  $n = 0$ .

$$X' = X^k + n \quad (2.10)$$

The square root transformation described by Equation 2.10 is also useful for biological counts, particularly when the researcher wants to give less weight to numerically abundant species. To count zeroes ( $X = 0$ ), use  $n = 0.5$  rather than 1; otherwise,  $n = 0$ . The default value  $k = 1/2$  represents the square root of  $X$  and should always apply, unless the researcher wishes to transform counts into presence/absence data (in which case, use  $k = 1/3$ ). To convert from a Poisson distribution to a Gaussian distribution, use  $n = 0$  and  $k = 2/3$ .

$$X' = \sinh^{-1} X \quad (2.11)$$

The arcsinh transformation described by Equation 2.11 is most useful when the dataset is dominated by zeroes.

$$X' = \sin^{-1} \left( \frac{\sqrt{X}}{100} \right) \quad (2.12)$$

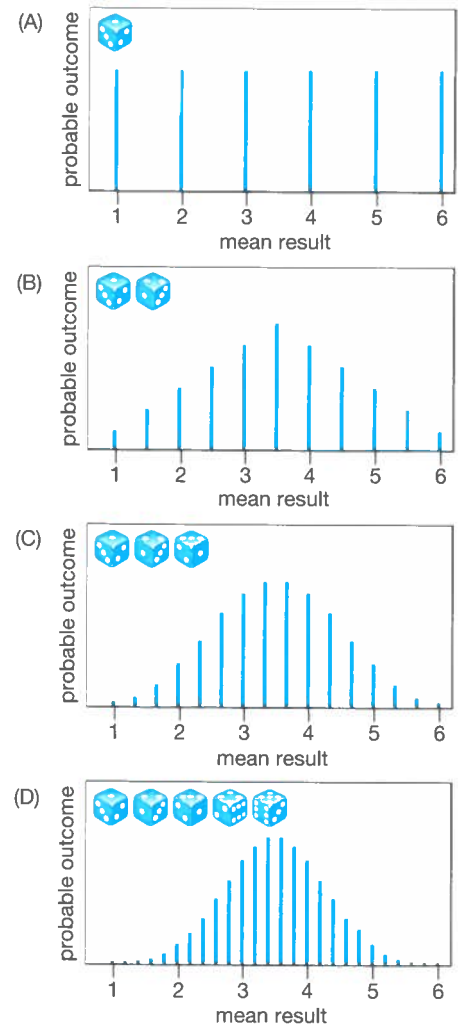
The arcsin transformation defined in Equation 2.12 should only be used when  $X$  is proportional data, given as percentages.

### The Central Limit Theorem Can be Employed to Increase the Predictability of Your Data Distributions

One last little bit of good news before we dive into the realm of inferential statistics and put all this information to good use: just about any data you could conceivably gather from the natural world can be considered normal, so long as you gather enough data. Although that may not sound like a terribly exciting revelation, it is a fundamental axiom of mathematics and probability that as we increase the number of observations ( $N$ ) in each computed mean ( $\bar{X}$ ), the distribution of those averages becomes more and more Gaussian. Though not a rigorous mathematical definition, this is what's known as the **central limit theorem** and its significance in the sciences cannot be overstated (**Example Box 2.5**).

This may seem like quite a fuss, just to get our data to behave more normally. But remember that much of what we will show by using inferential statistics ultimately depends not on our actual measurements but on the distribution of our data. So the normality of our data factors quite heavily on what statistical analyses we can (and can't) perform.

Attendant to the central limit theorem is the tendency of our data to become more normal as the number of observations ( $N$ ) increases (**Figure 2.8**). This is a very important consideration because the task of deciding how much data is "enough" is not an easy one. Since the very practice of science requires us to perform subset analysis, we can't measure everything—we are forced to limit ourselves according to some decision as to just how big our subset should be. Not an easy thing to do, especially when you're trying to investigate new (or underexplored) phenomena.



**Figure 2.8** Assuming we are not using loaded dice, the result of each die roll should be random. In fact, each specific result has a one-in-six chance of occurring. If we were using a single die (A), the central tendency of our results would not be normally distributed at all. If we instead rolled two dice (B) and calculated the mean result, the probability distribution of our results would look more normal. As we increase the number of dice (C–D), the distribution of data becomes more and more Gaussian—an elegantly simple demonstration of the central limit theorem.

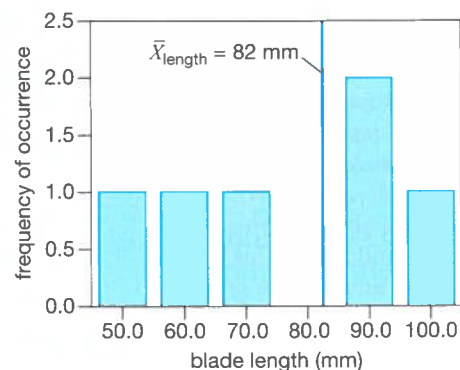
## EXAMPLE BOX 2.5

### Using the Central Limit Theorem to “Create” Normal Data

As an example, let's say we randomly selected 30 seagrass blades from a seagrass meadow and wanted to investigate the central tendency and standard deviation of the measured lengths (see Figure 2.3). As we saw in our earlier example of the 30-300 rule, the range of seagrass blade lengths was 56 mm (48 mm to 104 mm) (Table 1).

**Table 1** Blade Length of Randomly Selected Seagrasses ( $N = 30$ ) within a Subtidal Shoal Grass (*Halodule wrightii*) Meadow

Sample	length (mm)	Sample	length (mm)	Sample	length (mm)
1	104	11	95	21	78
2	59	12	98	22	100
3	68	13	48	23	101
4	49	14	96	24	104
5	94	15	73	25	103
6	94	16	104	26	66
7	75	17	99	27	63
8	102	18	62	28	62
9	75	19	101	29	94
10	80	20	69	30	48
$\bar{X}_{\text{length}} = 82$ $s = 19$					



**Figure 1** Distribution of seagrass blade-length data pooled into one bin of  $N = 30$  observations.

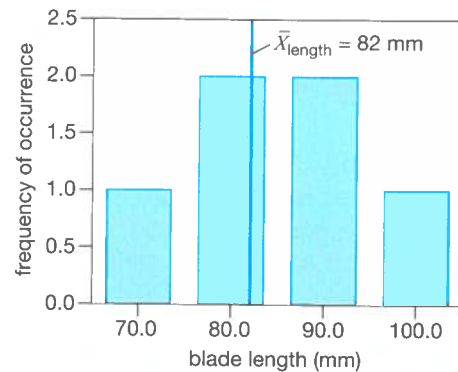
Note that the measures of blade length as depicted in Figure 1 do not appear to be normally distributed about the central tendency (in this example,  $\bar{X}_{\text{length}} = 82$  mm). However, if we binned our measurements into random groups (six groups of  $N = 5$ ) and calculated the means for each of those groups, our results would be

$$\begin{array}{lll} \bar{X}_{1-5} = 75 & \bar{X}_{11-15} = 82 & \bar{X}_{21-25} = 97 \\ \bar{X}_{6-10} = 85 & \bar{X}_{16-20} = 87 & \bar{X}_{26-30} = 67 \end{array}$$

If we then analyzed the descriptive statistics of these means, we would see in Figure 2 that the data appear much more normally distributed, with a greater number of observations grouped more tightly about the calculated mean.

$$\bar{X}_{\text{Means}} = 82 \quad s = 10$$





**Figure 2** Distribution of the same seagrass blade-length data randomly pooled into six bins of  $N = 5$  observations.

This demonstrates how the central limit theorem can be used to transform an otherwise nonnormal dataset into one that exhibits a more Gaussian distribution. This is only possible if

1. The result of one measurement is not dependent upon the result of another;
2. The data are randomly selected when binned into groups; and
3. The sample size in each of the new bins is equal for all bins (for example, six bins of  $N = 5$ ).

It is also important to note that the normality of our transformed data will improve as we increase the number of observations ( $N$ ) in our bins. So 6 bins of  $N = 5$  would yield a more normal distribution than 10 bins of  $N = 3$ .

## Inferential Statistics: A Brief Primer

Now that we have laid the necessary groundwork, it is time to leave the theoretical realm behind and actually try to apply some of what we have learned. As we proceed, keep in mind that the variety of statistical methods available to the scientific community is vast. Since a full treatise on statistics is far beyond the scope of this text, readers are strongly encouraged to familiarize themselves with additional resources on this subject.

That being said, what follow are a number of helpful examples of the simplest, most common statistical methods used in the sciences. As we work through these examples, we shall utilize the same dataset to illustrate how the same data can be analyzed in a variety of different ways. Ultimately, the fundamental goal of science is to distinguish chaos from order; to distinguish pure chance from “cause and effect.” By using statistical tests, we can demonstrate just how confidently we can make those distinctions, in strict mathematical terms.

### The Probability Level Will Define Which Results Are Statistically Significant, and Which Are Not

So what’s your chosen probability level? Although that line would definitely earn some confused looks at your next cocktail party, it is nonetheless a very important question to ponder. The **probability level** ( $\alpha$ ) is a threshold value chosen by the investigator used to define when a statistical result can be considered “significant” (in other words, when a result is unlikely to occur by chance). If you were testing your data for a particular cause-and-effect relationship,  $\alpha = 0.05$  (5%) simply means that you would expect that relationship to occur only 5 times out of 100 (5%) due to pure chance.

For most scientific applications,  $\alpha = 0.05$  is the preferred probability level. However, there are instances where it might be advisable to have a stricter



### WHEN IT COMES TO SAMPLE SIZES, THE MAGIC NUMBER IS 30

More is better. When it comes to collecting scientific data, it’s hard to argue with that statement. But it’s also a statement of no real help to us. What we really want to know is “How much data is enough?”

Using the central limit theorem as inspiration, we can demonstrate that a minimal sample size of 30 ( $N \geq 30$ ) is a good general rule of thumb. What we’re essentially looking for here is a subset of data that are sufficiently representative of the larger population. True in all cases? Of course not. But when in doubt, shoot for  $N \geq 30$ .

probability level. For instance, if you were testing a new design in automobile brakes, a failure rate of 0.05 (5 out of every 100 brake pads) could be absolutely tragic. In such cases, you might wish to use a probability level of 0.01 or 0.001 (or even lower).

### The Use of Confidence Intervals and Outliers Helps in the Analysis of Statistical Significance

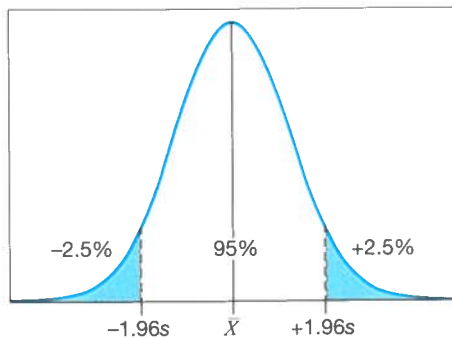
Once we have established the preferred probability level of our investigation, we can use that same value to define the number of standard deviations within which our normally distributed data should fall. If  $\alpha = 0.05$  is chosen, we are using that value to define the threshold of acceptable variance within our data. This also has tremendous implications for the predictability of our data, because we can use Equation 2.13 to establish that

$$CI = \bar{X} \pm \frac{d_{\alpha} \cdot s}{\sqrt{N}} \quad (2.13)$$

where our **confidence interval** (*CI*) describes the accepted probability that our sample mean  $\bar{X}$  accurately represents the true mean of the population from which it was sampled. For Equation 2.13, use  $d_{\alpha} = 1.96$  when  $\alpha = 0.05$ . Any single measurement that exceeds the confidence interval also falls outside the range of values we would expect for 95% of our data (Figure 2.9). In the context of statistics, we may consider those values to be “significantly different” from the dataset from which they were taken.

Sometimes, there may be values that dramatically exceed the accepted variance in the data. As a matter of convenience,  $\pm 3s$  is the accepted threshold to determine the presence of what are called **outliers** in the data. Data that are at least three standard deviations from the mean (or possess  $z \geq 3$ ) exceed the range of values in 99.6% of the data (see Figure 2.5). Because outliers represent extreme values in the dataset, they also represent very significant departures from the “background” distribution of data.

As a practical matter, outliers in the dataset must be handled very carefully. Although outliers can certainly be due to extremely rare or atypical natural events captured in the dataset, they are more commonly attributed to experimental or data-recording errors. Therefore, it is usually a good idea to analyze your data for the presence of any single measurement that exceeds the  $\bar{X} \pm 3s$  threshold.



**Figure 2.9** A graphical representation of normally distributed data, indicating the 95% confidence interval of the data distribution ( $\bar{X} \pm 1.96s$ ), assuming  $\alpha = 0.05$ . The shaded areas of the curve indicate those data that fall outside the 95% confidence interval (*CI*).

If outliers are identified and can be shown undeniably to have arisen from experimental or data-recording errors, those values should be purged from the dataset and new descriptive statistics should be calculated from the cleansed dataset (Example Box 2.6). If it is impossible to attribute human error to the presence of outliers, it may be useful to perform the same statistical analyses twice: once including all outliers, and again with all outliers removed. The effects of these outliers on the final statistical results should then be reported, allowing the reader to draw the conclusions.

### The Strength of Our Statistical Analyses Will Depend Heavily on Our Assumptions of the Data

As we have already discussed in this chapter, the use of statistics requires that we make all sorts of assumptions about our data—the most fundamental assumption is that the subset of data we have collected is representative of the larger population. There are a few more assumptions that we must address, particularly if we wish to place any confidence in the results of our statistical tests. With rare exception, the statistical tests most commonly

## EXAMPLE BOX 2:6

### Using z Scores to Identify (and Purge) Erroneous Outliers in a Dataset

Richard has been measuring the lengths of emergent seagrass blades in a dense patch of shoal grass (*Halodule wrightii*) in a subtidal seagrass meadow. After three hours of monotony in the hot sun, Richard is bound to make a few mistakes. Sure enough—a few of his observations were recorded in centimeters when they should have been entered in millimeters.

128	145	132	121	122	128	131	126
124	12.1*	142	137	129	127	121	127
143	139	126	11.8*	123	134	141	132
12.4*	131	144	126	124	128	129	128

Janet is handling the postprocessing and is clever enough to analyze the data for outliers by calculating z scores (see Equation 2.5). Her initial results indicate three possible outliers:

$$N = 32 \quad (12.1) \text{ z score} = -3.01$$

$$\bar{X} = 120 \text{ mm} \quad (11.8) \text{ z score} = -3.02$$

$$s = 36 \text{ mm} \quad (12.4) \text{ z score} = -3.00$$

Although Janet might be tempted to “correct” Richard’s mistakes, it’s better to simply eliminate the three outliers and reanalyze:

$$N = 29$$

$$\bar{X} = 131 \text{ mm}$$

$$s = 7 \text{ mm}$$

Note that both the central tendency and standard deviation of the data have changed quite a bit, now that the outliers have been purged from the dataset. Perhaps most importantly, we can see how much more improved our standard deviation has become.

employed in the sciences require that we first certify that our data agree with the following:

Assumption of independence

Assumption of single population representation

Assumption of normality

Assumption of homoscedasticity

### Independence

Under most conditions, it is important that the values that comprise our dataset have been randomly selected from the larger population and that those values are independent of each other. An independent sample is one that, when its value is measured, will not influence (or be influenced by) the value of any other sample.

For example, the measured length of one seagrass blade is not likely to influence the length of any other seagrass blade (particularly if we took care to

always measure blades from separate plants). This would be a perfect example of our length measurements being truly independent of each other and therefore satisfying the assumption of independence. Likewise, we would expect that the leaf-width measurements would also be independent of each other.

We may not be so confident in the independence of our data if we measured multiple leaves from the same seagrass plant. Perhaps the length of one leaf would be influenced by the length of another leaf on the same plant, because of an uneven distribution of nutrients, self-shading, or any other scenario where one leaf might impose an influence on the growth (length) of a partner leaf. Why take the chance? If you can design a sampling strategy that maximizes sample independence, the strength of your statistical analyses will be better for the effort.

### Single Population

Since our data are supposed to represent the larger population of values from which they were collected, it is critical that we are sure the sampled population is the same for every observation. That means that we must take great care to avoid sampling across multiple populations. In our seagrass example, that means we would want to collect measurements from random individuals within the same general location. If we were to collect our data from three different sites and try to analyze them all together to yield some grandiose perspective on seagrass growth, we would be violating our “single population” assumption. Of course, we could still use statistics to compare seagrass growth between these three sites; we’d just have to be careful to organize our data so that our measurements were grouped into three single populations.

### Normality

As we have discussed previously in this chapter, normality merely refers to the dispersion of data following a Gaussian curve. Thanks to the central limit theorem, we can often assume our data can be rendered normal if the number of observations in our dataset is robust ( $N > 30$ ). Recall also that several different data transformation methods may be employed to ensure normality. Keep in mind that there are several statistical tests that do not assume (and therefore do not require) that your data are normally distributed, so even nonnormal data can be used for statistical analyses.

### Homoscedasticity

This assumption applies only when comparing two (or more) populations with each other, and refers to the analysis of their variances. If the variances of the populations are equal (or very nearly equal), the assumption of **homoscedasticity** is upheld. Recall that the variance is simply the standard deviation multiplied by itself (see Equation 2.8), so populations whose standard deviations are equal (or very nearly equal) are also covered by this assumption. Keep in mind that the central tendencies of the populations need not be equal; in fact, they can be very different from each other. What is critical to uphold this assumption is that the dispersion of data about the central tendency must be similar.

If none of these assumptions are violated, we can use **parametric** tests to analyze our data. Parametric tests are a special family of statistical methods that generally offer greater power, accuracy, and precision because of the inherent “quality” of the data—quality that has been authenticated by our confirmation that the data are independent, from a single population, normal, and homoscedastic.



If any of these assumptions have been violated, we must instead use **non-parametric** tests to analyze our data. Since nonparametric tests are free from the specific assumptions about data distribution, they are still quite useful but generally considered to be a less powerful alternative to parametric tests. In the natural sciences, it is best to perform parametric tests whenever possible, but nonparametric tests offer a suitable alternative if your data cannot be made to conform with the assumptions outlined above.

### Formulating a Testable Hypothesis is Critical to Both the Scientific Method and the Use of Inferential Statistics

As we learned in Chapter 1, the ability to formulate a testable hypothesis is the very backbone of the scientific method. A research hypothesis is tested in the laboratory or in the field by a variety of well-designed (and well-executed) experiments that will produce data—these data are then used to test a **statistical hypothesis**. It is critical that these two hypotheses are linked in such a way that the acceptance (or rejection) of the statistical hypothesis will likewise allow the investigator to either accept (or reject) the research hypothesis. Because there are only two logical conditions of the hypothesis (either true or false), the statistical hypothesis must be structured in such a way as to yield a binary result.

This is done by translating the research hypothesis into a statistical hypothesis, which can be phrased in two ways: the **null hypothesis** ( $H_0$ ) and the **alternative hypothesis** ( $H_a$ ). The null and alternative hypotheses have to be mutually exclusive of each other, so that if one of them is statistically demonstrated to be true, the other hypothesis is automatically negated. Because of this relationship between the two, we can use statistics to test either the null or the alternative hypothesis; regardless of which one we choose to analyze first, the result of the statistical test will simultaneously answer both hypotheses.

A proper null hypothesis ( $H_0$ ) should be defined in such a way that it: (1) negates  $H_a$ ; (2) denies any relationship between the dependent and independent variable(s) in the study; (3) is assumed to be true unless the results of a statistical test of significance allow it to be rejected; and (4) is phrased so that if  $H_0$  is true, the research hypothesis ( $H_a$ ) cannot possibly be true. That last point is key, and it is why it is called the null hypothesis—if  $H_0$  cannot be rejected by a statistical test of significance, we must assume it is true and that our research hypothesis was incorrect (or incapable of verification). As a practical matter, the null hypothesis generally states that there are no significant differences when making some kind of comparison.

By contrast, the alternative hypothesis ( $H_a$ ) is the research hypothesis we're trying to "prove" with statistics.  $H_a$  should always be defined so that it (1) negates  $H_0$ ; (2) confirms some relationship between the dependent and independent variable(s) in the study; and (3) is assumed to be false unless a statistical test of significance can confirm otherwise. In most cases, the alternative hypothesis is structured so that it implies there are significant differences between two (or more) compared populations, variables, means, variances, or what have you.

The mutually exclusive definitions of  $H_0$  and  $H_a$  will ultimately lead us to a logician's showdown, which is exactly the point. Since  $H_0$  is the condition that is assumed to be true by default, we cannot accept  $H_a$  (and therefore reject  $H_0$ ) unless we have significant mathematical support for doing so—that is the whole purpose of inferential statistics in scientific research.



#### A PROPERLY DEFINED HYPOTHESIS IS CRITICAL TO SUCCESS

Translating the research hypothesis into a valid statistical hypothesis is a critical step, and is perhaps the most confusing aspect of inferential statistics. If there are errors in how you define the null ( $H_0$ ) versus alternative ( $H_a$ ) hypothesis, the results of your statistical tests will be useless.

Take great care that  $H_0$  and  $H_a$  can only be confirmed with a simple "yes/no" or "true/false" answer, and that  $H_0$  and  $H_a$  are mutually exclusive—that is, it is impossible for both of them to be true (or for both of them to be false, for that matter).

**Table 2.1** Confirmation Matrix of Statistical Conclusions for Comparisons Between the  $p$ -Value and  $\alpha$ 

$p\text{-value} > \alpha$	$p\text{-value} \leq \alpha$
Accept $H_0$	Reject $H_0$
Reject $H_0$	Accept $H_0$
Conclusion: No significant differences exist, or variability in the data can be explained as random "noise"	Conclusion: Significant differences do exist, or variability in the data cannot be dismissed as chance occurrences



### WHAT A $p$ -VALUE REALLY MEANS

Remember that  $\alpha$  represents the absolute limit for a result to be significant. Although the line has to be drawn somewhere, you would not want to make a critical decision while teetering on the edge of that decision boundary. For  $\alpha = 0.05$ ,

If  $p \leq 0.05$ , there is sufficient statistical evidence to reject  $H_0$  and accept  $H_a$ .

If  $p \leq 0.01$ , there is strong statistical evidence to reject  $H_0$  and accept  $H_a$ .

If  $p \leq 0.001$ , there is overwhelming statistical evidence to reject  $H_0$  and accept  $H_a$ .

Since statistics is all about probability, we can use our accepted probability level ( $\alpha$ ) to define the probability value (or  **$p$ -value**) in our statistical tests that will indicate when we can define any differences as being statistically significant. In other words, whenever a statistical test returns a  $p\text{-value} \leq \alpha$ , we can claim that the results are significant enough to accept  $H_a$  (Table 2.1).

In order for a  $p$ -value to be deemed significant, it must be no greater than the chosen  $\alpha$ . Since the  $p$ -value returned by all statistical tests will naturally range from 0.00 to 1.00, the realm of significant  $p$ -values (if  $\alpha = 0.05$ ) will be 0.00–0.05; if we had chosen an  $\alpha$  of 0.01, significant  $p$ -values would range from 0.00 to 0.01. Remember that we're living in the land of probabilities here, so there's always a remote chance of erroneously assigning significance when none truly exists. If we chose an  $\alpha$  of 0.01, we would have a 1 in 100 (1%) chance of assigning statistical significance to our result and mistakenly accepting  $H_a$ . However, as  $p \rightarrow$  zero, the stronger our statistical evidence of significance.

Because our choice of  $\alpha$  is subjective (and because no statistical test is 100% infallible), there is always a possibility that we may come to an erroneous conclusion regarding  $H_0$  and  $H_a$ . When a null hypothesis is mistakenly rejected, that is called a **Type I error** and essentially means that the statistician has erroneously found differences where none truly exist. Fortunately, the researcher can reduce the likelihood of committing a Type I error by simply choosing a smaller  $\alpha$ .

By contrast, a **Type II error** occurs when the null hypothesis is upheld when it should have been rejected. In practice, Type II errors are much more difficult to manage, because the likelihood of this error depends on a number of factors that may be beyond the researcher's control, such as the sample size or the variability of the data (evidenced by large standard deviations). Increasing the sample size is a straightforward way to reduce the chance of committing a Type II error, but researchers rarely have the ability to reduce the overall variability of their data. Regardless, it is impossible to completely eliminate the possibility of committing a Type I or Type II error. We must instead try our very best to minimize the likelihood of these errors as much as possible, and keep in mind that our statistical conclusions are a prediction of "correctness," not a guarantee.

### The Most Basic Statistical Tests are Used to Test for Equality (or Inequality) Between Two Populations

Although there are a multitude of different statistical tests available, they are all essentially used with the same goal in mind: to make comparisons and to determine the relative significance of any differences found. Conceptually, it really is that simple.

The simplest comparison that can be made is to test whether one population is indistinguishable from another. Said another way, we are interested in testing whether the difference between two populations can be dismissed as natural variability (or chance). For example, we would expect that the length of seagrass blades would naturally vary, some leaves being shorter and others longer. If those differences can be explained by natural variation, we would expect that same amount of natural variation to exist in other seagrass populations (in other words, the variation between the two populations should be equal, making them indistinguishable from each other). However, if there is some other factor influencing seagrass blade length, those effects should manifest themselves differently in one population when compared to the other.

To investigate those possible differences, we would have to represent those populations by some consistent measure using descriptive statistics, such as the central tendency, standard deviation, or variance for each population within our comparison. If we chose to compare the variances, the simple mathematical expression of this comparison could be written as

$$H_o: s_n^2 = s_o^2$$

$$H_a: s_n^2 \neq s_o^2$$

In this example, our null hypothesis ( $H_o$ ) states that the variance within one population ( $s_n^2$ ) is the same as the variance in the population to which it is being compared ( $s_o^2$ ); in other words, our two populations possess equal variances and are therefore indistinguishable from each other (when using the variance as our measure of comparison). Since we have to define  $H_a$  in a way that is contrary to  $H_o$ , we simply have to state the alternate hypothesis: that the variances are not equal. This basic test of equality is known as a **two-tail test**, where the default condition ( $H_o$ ) always assumes that the compared measures are equal.

Keep in mind that we can use any type of descriptive statistics in our comparisons. Instead of comparing the variances, we could just as easily compare the standard deviations or the central tendencies. If we chose to compare the population means, it would require that we slightly modify our stated hypotheses, so that

$$H_o: \overline{X}_n = \overline{X}_o$$

$$H_a: \overline{X}_n \neq \overline{X}_o$$

Of course, comparisons of any measure of central tendency are valid, so the medians or modes could be compared in exactly the same fashion. But in this example, we are now testing whether the mean of some population ( $\overline{X}_n$ ) is equal to (or indistinguishable from) the mean of another population to which it is being compared ( $\overline{X}_o$ ).

If our statistical analysis yields a result that indicates  $p \leq \alpha$ , we could use Table 2.1 to reject  $H_o$  and accept  $H_a$ : a result that indicates that the two populations are not equal. Because of the logical consequences of these results, the two-tail test should always be the first statistical test employed. If the results of your statistical analysis indicate that you cannot accept  $H_a$ , then the measures are mathematically indistinguishable from each other,  $H_o$  is upheld, and there is nothing left for you to do—your statistical analyses are complete.

However, if the results of the two-tail test indicate that the measures are not equal, there exists just one of two other possibilities: either  $\overline{X}_n > \overline{X}_o$  or  $\overline{X}_n < \overline{X}_o$ . Each of these represent a special case of the **one-tail tests**, which are strictly defined as follows:

Right-tail test:

$$H_o: \overline{X}_n \leq \overline{X}_o$$

$$H_a: \overline{X}_n > \overline{X}_o$$

Left-tail test:

$$H_o: \overline{X}_n \geq \overline{X}_o$$

$$H_a: \overline{X}_n < \overline{X}_o$$

Mathematically, it is impossible for both the right- and left-tail tests to yield  $p \leq \alpha$  simultaneously. So, because of the logical structure of these one-tail tests, only one of them shall be necessary (chosen at the investigator's discretion) if a two-tail test has already been performed. Essentially what that means is that the result of one will automatically determine the result of the other.

## Choosing the Right Statistical Method

Now that we've covered the basics, it's time to put it all together and formulate an ordered checklist of steps to be followed as a standard recipe for any statistical analysis, whether those analyses shall require parametric (Figure 2.10) or nonparametric (Figure 2.11) tests. How you construct your research hypothesis and how you choose to collect measurements in the lab or in the field will have a tremendous influence on your ability to use inferential statistics, so it is always wise to refer to the following checklist when first designing your lab/field method:

- ☑ Formulate a testable research hypothesis that requires a binary answer (yes/no; true/false).
- ☑ Determine whether you require quantitative (metric or meristic) or qualitative (categorical) data.
- ☑ Research the most appropriate statistical tests you wish to use and how to use them.
- ☑ Acquire and learn how to use a statistical software package appropriate for your goals.
- ☑ Calculate the descriptive statistics of each population you wish to analyze:
  - Central tendency (see Equation 2.2)
  - Standard deviation (see Equation 2.3)
  - Standard error of the mean (see Equation 2.4)
  - Coefficient of variation (see Equation 2.6, if needed)
  - Range (see Equation 2.7)
  - Variance (see Equation 2.8)
  - Confidence interval (see Equation 2.13, if needed)
- ☑ Test for outliers (z score; see Equation 2.5). Recalculate descriptive statistics if outliers were removed.



number of samples to be tested			
tests for:			
	1	2	3
central tendency (mean, median, mode)	1-sample z test (compare $\bar{X}_1$ against a chosen value)	2-sample z test (compare $\bar{X}_1$ against $\bar{X}_2$ )	n-sample F tests (compare multiple $\bar{X}_n$ )
	1-sample t test (compare $\bar{X}_1$ against a chosen value)	2-sample t test (compare $\bar{X}_1$ against $\bar{X}_2$ )	
		paired-sample t test (compare $\bar{X}_1$ against $\bar{X}_2$ when populations are dependent)	
variance or variability	1-sample $\chi^2$ test (compare $s^2$ against a chosen value)	Levene's test (compare $s^2_1$ against $s^2_2$ )	Levene's test (compare multiple $s^2_n$ )
association or correlation	F test for linear regression (compare associations between 2 variables)	z test for 2 correlation coefficients (compare correlation coefficients between 2 samples)	F test for multiple regression (compare associations between $n$ variables)
	z test for 1 correlation coefficient (compare a correlation coefficient against a chosen value)	z test for correlation proportions (compare correlated proportions in 2 same-group surveys)	
goodness of fit	1-sample log-likelihood or $\chi^2$ test (compares the differences between observed vs. expected values)	2-sample log-likelihood or $\chi^2$ test (compares the differences between 2 sample distributions)	n-sample log-likelihood or $\chi^2$ test (compares the differences between $k$ sample distributions)
proportions, ratio, or counts	z test for 1 proportion (compare proportion <sub>1</sub> against a chosen value)	z test for 2 proportions (compare proportion <sub>1</sub> against proportion <sub>2</sub> )	
sufficiency of sampling effort	sequential test for $\bar{X}$ (compares the cumulative trend of $\bar{X}$ against a chosen value)		
	sequential test for $s^2$ (compares the cumulative trend of $s^2$ against a chosen value)		
likelihood of occurrence	z test for uncertainty (compares the likelihood of events based on past occurrences)		

**Figure 2.10** A simplified classification of the parametric statistical tests most commonly used for data analysis in the natural sciences. Note that the use of these parametric tests assumes that the data are independent, from a single population, normally distributed, and of equal variances. If any of these assumptions have been violated, the use of nonparametric tests (see Figure 2.11) is indicated. (From Kanji GK [1999] 100 Statistical Tests. With permission from SAGE Publications Ltd.)

- ☑ Verify data in each population are
  - Independent
  - From a single population
  - Normal
  - Homoscedastic
- ☑ Determine whether parametric or nonparametric tests are warranted.
- ☑ Choose the desired probability level ( $\alpha$ ).
- ☑ Explicitly state the null ( $H_0$ ) and alternative ( $H_a$ ) hypotheses for the two-tail test.
- ☑ Explicitly state the null ( $H_0$ ) and alternative ( $H_a$ ) hypotheses for both the right-tail and left-tail tests.

**Figure 2.11** A simplified classification of the nonparametric (distribution-free) statistical tests most commonly used for data analysis in the natural sciences. (From Kanji GK [1999] 100 Statistical Tests. With permission from SAGE Publications Ltd.)

number of samples to be tested			
tests for:	1	2	3
central tendency (mean, median, mode)	sign test for median (compare median <sub>1</sub> against a chosen value)	sign test for 2 medians (compare median <sub>1</sub> against median <sub>2</sub> )	rank-sum test for means (compare whether $n$ samples came from populations with the same mean)
	rank test for mean (compare $\bar{X}_1$ against a chosen value)	rank test for 2 means (compare $\bar{X}_1$ against $\bar{X}_2$ )	steel test (compare $n$ experimental treatments against the control)
variance or variability		Siegel-Tukey test (compare whether 2 samples came from 2 populations with equal $s^2$ )	
association or correlation	Spearman or Kendall rank correlations (compare association between 2 series of observations obtained in pairs)		Wilcoxon-Wilcoxon or Friedman test (compare differences in treatments given to $n$ subjects)
randomness	run test for randomness in 1 sample (compares whether observations in a sample are independent of order)	run test for randomness in related samples (compares whether related samples have been selected randomly)	
	adjacency test for random fluctuations (compares whether fluctuations in a sequence are random)		
	Wilcoxon-Mann-Whitney randomness test (compares whether + and - signs in a sequence are random)		
	rank correlation time-series test (compares whether trends are present within a time series)		

## References

- Bakus GJ (2007) Quantitative Analysis of Marine Biological Communities: Field Biology and Environment. John Wiley and Sons.
- Jones ER (1996) Statistical Methods in Research. Edward R. Jones.
- Kanji GK (1999) 100 Statistical Tests. SAGE Publications.

Keeping ES (1995) Introduction to Statistical Inference. Dover Publications.

Linton M, Gallo Jr PS and Logan CA (1975) The Practical Statistician: Simplified Handbook of Statistics. Wadsworth Publishing Company.

## Further Reading

- Cassel C-M, Särndal C-E, & Wretman JH (1997) Foundations of Inference in Survey Sampling. John Wiley and Sons.
- Cochran WG (1977) Sampling Techniques, 3rd ed. John Wiley and Sons.
- Dorofeev S & Grant P (2006) Statistics for Real Life Sample Surveys: Non-Simple-Random Samples and Weighted Data. Cambridge University Press.
- Dytham C (2011) Choosing and Using Statistics: A Biologist's Guide, 3rd ed. Wiley-Blackwell.
- Jaisingh LR (2006) Statistics for the Utterly Confused, 2nd ed. McGraw-Hill.
- Keller DK (2006) The Tao of Statistics. SAGE Publications.

- Kenny DA (1979) Correlation and Causality. John Wiley and Sons.
- Mandel J (1964) The Statistical Analysis of Experimental Data. Dover Publications.
- Newman I & Newman C (1977) Conceptual Statistics for Beginners. University Press of America.
- Salkind NJ (2007) Statistics for People Who (Think They) Hate Statistics: The Excel Edition. SAGE Publications.
- Steiner F (ed) (1997) Optimum Methods in Statistics. Akadémiai Kiadó.
- Thompson SK (1992) Sampling. John Wiley and Sons.