2

MATHEMATICS AND PROBABILITY THEORY

> How can it be that mathematics, being after all a product of human thought which is independent of experience, is so admirably appropriate to the objects of reality? . . . As far as the laws of mathematics refer to reality, they are not certain, and as far as they are certain, they do not refer to reality.
>
> (Einstein, 1922)

In this chapter, we review some of the essential elements of mathematics and probability theory that the reader may have learned in prior courses, or at minimum, has had some exposure through self-study. We reserve Chapter 3 for a review of the elements of essential statistics that is generally required for an understanding of the rest of the book. Our distinction between *mathematics and probability* versus *statistics* is not a sharp one. In this chapter, we use mathematics as a vehicle to understanding applied statistics rather than see it as a field in its own right, which, of course it is, with a variety of branches and subdisciplines.

Our brief mathematics review draws material sparingly from introductory courses such as precalculus, calculus, linear and matrix algebra, and probability. Such topics constitute the very bedrock of mathematics used in applied statistics. Elements such as *functions*, *continuity*, *limits*, *differential* and *integral calculus* and others are (very) briefly reviewed. We also present some of these fundamentals using R where appropriate. For an excellent review of essential mathematics for the social sciences, refer to Gill (2006). Barnett, Ziegler, and Byleen (2011) also provide a very readable

Applied Univariate, Bivariate, and Multivariate Statistics, First Edition. Daniel J. Denis.
© 2016 John Wiley & Sons, Inc. Published 2016 by John Wiley & Sons, Inc.
Companion Website: www.wiley.com/go/denis/appliedmultivariatestatistics

overview of mathematics covering a wide range of topics. Fox (2008a) is also a useful monograph. Refer to Gemignani (1998) for how calculus is used in statistics.

We do not pretend to cover any of these topics in any respectable depth whatsoever, having only the space to provide brief and relatively informal overviews of these essential concepts. If you lack familiarity with such fundamentals, a bit of time taken to study and appreciate these elements can be of great help in understanding material covered in this book and beyond. This is not to say that without this knowledge you cannot learn and apply principles presented in the book, but the deeper your knowledge of these concepts, the more confident you will likely be in applying your skills to data analysis because you will be better familiar with the "rules of the game."

For instance, informally, the definition of *continuity* is that of not lifting your pencil as you draw a line on a piece of paper. The line is said to be "continuous" since at no point did you impose a *discontinuity* by lifting your pencil. And if it were not for mathematicians' quest to provide a rigorous logical justification for the calculus in the 1800s, thereby formalizing such things as continuity and limits, "naïve" definitions such as that for continuity would still do us just fine. However, as a result of the advance in fields such as real analysis, complexity and rigor have been introduced in order to make these definitions extremely *precise*. Having even some appreciation and understanding of this level of precision and rigor can only benefit the student of applied statistics and social science if for no other reason than to sharpen one's analytical skill and ability to differentiate and decipher among objects one deals with—a skill that *is* required of anyone who purports to do any kind of quantitative analysis or research, either elementary or advanced.

Furthermore, having an understanding of essential mathematics also serves to "demystify" what can otherwise seem like a quite arcane field of study. Perhaps this sentiment was best expressed by Mulaik (1972) when advising the reader on the mathematical training required to study a topic such as factor analysis:

> Ideally, one begins a study of factor analysis with a mathematical background of up to a year of calculus. This is not to say that factor analysis requires an extensive knowledge of calculus, because calculus is used in only a few instances, such as in finding values of an independent variable which will maximize or minimize a dependent variable. But having calculus in one's background provides sufficient exposure to working with mathematical concepts so that one will have overcome reacting to a mathematical subject such as factor analysis as though it were an esoteric subject comprehensible only to select initiates to its mysteries. (p. 16)

With Mulaik's thoughts in mind, I strongly encourage you to embrace as much technical information as possible, even for its own sake, and even if you may be currently unaware of the answer to the longstanding question students *love* to ask—*When will I use this*? If you learn to embrace rather than shy away from difficult, technically rich material replete with symbols, it puts you on course to being able to understand and comprehend virtually *anything* that is put in your path in terms of complexity. On the other hand, if you regularly shy away from complexity, you forever weaken your neuronal ability to disentangle and otherwise *figure out* things. Our general advice is that *if you want to make the difficult simple to understand, study the difficult until it is simple*.

2.1 SET THEORY

A *set* in mathematics is simply a group of objects or elements. What is key to the definition of a set is that it is sufficiently descriptive and identifiable such that we can know which objects belong to the given set and which do not. For example, the set of persons of at least 5 ft, 10 in. in a room is a *precisely* defined group of persons. Consequently, given this precision of the elements in question, we are in a position to know who belongs to the set and who does not. Anyone whose height is 5 ft 10 in. or more is a member of the set. Anyone who is less than that height does not belong to this set. The concept of a set is fundamental to mathematics, for one, because it helps immensely to organize the subject, and second, it provides a language for relating mathematical fields. It has been said that sets are so fundamental to mathematics that the theory of sets forms the foundation of *all* of mathematics.

To communicate that x is in the set A, we write that x is an element (or a member) of set A:

$$x \in A$$

where the notation "$\in$" means "is an element of." To denote that y is *not* an element of the set A, we write (see *Venn diagram* in Figure 2.1):

$$y \notin A$$

To denote that set A is a subset of set B, we write

$$A \subseteq B \text{ or } B \supseteq A$$

What does it mean to say A is a subset of B? It means that the set A is, informally, smaller than or equal to the set B. That is, if A is a subset of B, it implies that elements in A are also contained in B but that elements in B are not *necessarily* contained in A. Note that there are two possibilities here, and "$\subseteq$" can logically be used to qualify

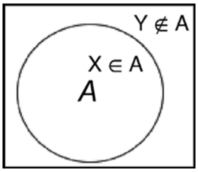


FIGURE 2.1 X is an element of the set A. Y is not an element of set A.

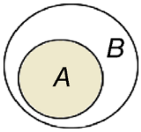


FIGURE 2.2 Set A is a proper subset of set B.

both, “smaller than” or “equal to”. However, if we know definitively that there is at least one element in set B that is not in set A, then we say that A is a *proper subset* of B, and write

$$A \subset B$$

An example of a proper subset, $A \subset B$, is given in Figure 2.2.

When you hear the phrase “proper subset,” just think to yourself that this implies a “true” subset, meaning there is no possibility of the two sets being equal. If we say set A is a proper subset of set B, then A cannot equal B; set A must actually be less than B such that the elements in A are also in set B but that there is at least one element in B that is not contained in A.

The equality of two sets is written as $A = B$. To show that two sets are equal, one must be able to show that A is a subset of B and B is a subset of A.

Thinking up applied examples for sets is not difficult. For instance, consider the set A of numbers 1 through 6 on a die. The set E of even numbers plus the set O of odd numbers between 1 through 6 can be considered a subset of this set, that is, $E + O \subseteq A$, whereas the set of only odd numbers O would be considered a proper subset of the set A. That is, $O \subset A$. To demonstrate equality between these sets, it would be a simple matter to show that $E + O \subseteq A$ and $A \subseteq E + O$. Other simple examples of sets in a research setting are as follows:

- The set of those suffering from schizophrenia is a proper subset of the collection of human beings.
- The set of those gainfully employed is a proper subset of those desiring a job.
- The set of students passing a course is a subset of students enrolled in that course.

Note that for the first two examples, the phrase *proper subset* was used to denote the fact that those suffering from schizophrenia and employed individuals are surely less than the sets of human beings and those seeking work. However, in the third example, since it is a sure possibility that everyone in a given course passes the course, we do not automatically assume this set to be smaller than the set of students enrolled in the course, which is why we used the word “subset” here instead of the identifier “proper subset.” If we knew, on the other hand, that at least one student in the class failed the course, then the set of students passing the course would be a proper subset.

A *countable* set is one in which elements of the set can be put into one-to-one correspondence with the positive integers. A *finite* set is one that has a *noninfinite*

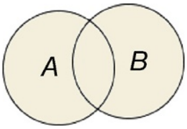


FIGURE 2.3 *A* union *B* is the set of elements in *A* or *B* or both, denoted by the totality of the shaded area.

number of elements. See Gill (2006) for examples along with further characteristics of sets.

2.1.1 Operations on Sets

We can define various operations on sets. For instance, the *union* of two sets *A* and *B* is given by

$$A \cup B := \{x : x \in A \text{ or } x \in B\}$$

where in words, the above statement reads "*x* is an element of *A* or *x* is an element of *B*." For example, if set *A* is the set of unemployed and set *B* is the set of students passing a course, the union of these sets, $A \cup B$, is the set of unemployed *or* those passing a course, *or* those belonging to both sets. See Figure 2.3 for an example of a union of two sets.

We can perform simple set operations using R. For example, consider sets *A* and *B*, and the computation of their union:

```
> setA <- c("3", "4", "5", "6")
> setB <- c("5", "6", "7", "8")
> union(setA, setB)

[1] "3" "4" "5" "6" "7" "8"
```

The intersection of sets *A* and *B* is denoted as

$$A \cap B := \{x : x \in A \text{ and } x \in B\}$$

and is the set of elements in *both* sets *A* and *B*. That is, for an element to belong in the intersection, it cannot simply be in one of the sets. It must be in *both* to qualify for the intersection. An example of two intersecting sets is given in Figure 2.4.

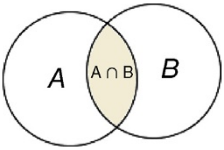


FIGURE 2.4 The shaded area represents the intersection of sets *A* and *B*.

In R, we can easily find the intersection of sets A and B:

```
> intersect(setA, setB)
[1] "5" "6"
```

A practical example of intersecting sets might be the set of college students who are also enrolled in a statistics course. Another example is the set of men who are also married.

2.1.2 Denoting Unions and Intersections of Many Sets

We have thus far expressed the union and intersection of a very small number of sets (only two sets). We can, however, represent the union and intersection for any number of sets, even if infinite in number. For unions, we can represent this by

$$\bigcup_{n=1}^{\infty} A_n$$

The above is the set of elements belonging to *at least* one of the sets $A_1, A_2, \ldots, A_n$. We use the qualifier "at least one" because we are dealing with the *union* of sets, meaning that the element in question can be in one set, *or* another set, *or* another set, etc. The expression reads to start with set $n = 1$ of A_n and to sum to ∞. An example of a union of three sets is given in Figure 2.5, $\bigcup_{n=1}^{3} A_n$, where $A_1 = A$, $A_2 = B$, $A_3 = C$.

Analogously, we can represent intersections by

$$\bigcap_{n=1}^{\infty} A_n$$

Figure 2.6 is the set of elements belonging to the intersection of the sets $A_1, A_2, \ldots A_n$ (where $A_1 = A$, $A_2 = B$, $A_3 = C$). Notice the difference between the qualifier "or" versus "and," when discussing unions versus intersections. When we use the word "and," as we will soon see, we are specifying a *joint probability* in probability theory.

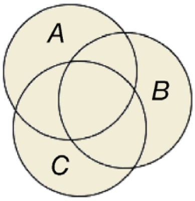


FIGURE 2.5 Union of sets A, B, C, denoted by shaded area.

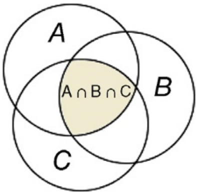


FIGURE 2.6 Shaded area is the intersection of sets *A*, *B*, *C*.

2.1.3 Complement of a Set

The complement of set B relative to set A can be denoted as

$$A \backslash B := \{x : x \in A \text{ and } x \notin B\}$$

In words, this means all the elements in set A that are not in set B. For example, if we defined a set of employed and unemployed individuals, with no other possibilities, then the complement of the set of employed individuals is the set of unemployed persons. If we had defined the full set to include such circumstances as individuals looking for work but unable to find a job, then the complement of the employed individuals would still be "all others," only that now, all others would be defined as the unemployed and those actively seeking work.

Substantively, in research, complements are useful especially computationally when we want to specify an "else" category on such things as questionnaires and other measures. For instance, the complement of all those individuals earning up to $60,000 per year are those individuals earning more than $60,000 per year. The complement to individuals suffering from a psychiatric diagnosis are those individuals not suffering from a psychiatric diagnosis.

2.2 CARTESIAN PRODUCT $A \times B$

The Cartesian product $A \times B$ is defined as

$$A \times B := \{(a, b) : a \in A, b \in B\}$$

and is the set comprising of ordered paired elements in A and B, such that each element in A has a pairing with another element in B. The Cartesian product, or more generally the *Cartesian coordinate system*, originated with René Descartes (1596–1650) and forever linked algebra to geometry, and gave birth to the field of *analytic geometry*. The Cartesian coordinate system was a major stepping stone in the history of mathematics and science in general. An example of a Cartesian coordinate system in two dimensions is given in Figure 2.7.

When one takes a number from the first set and pairs it with a number from the second set, one obtains a new number. For example, if we take the number 3 from

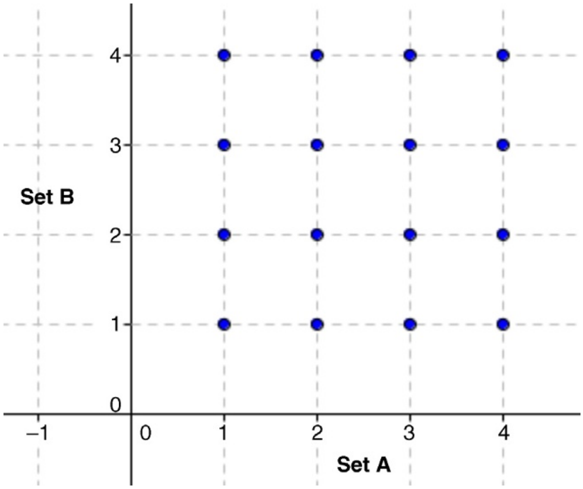


FIGURE 2.7 Two-dimensional Cartesian plane.

set *A* (along the abscissa) and number 4 from set *B* (along the ordinate), we obtain the new number (3,4). We can extend coordinate systems to many more dimensions than simply 2, and mathematically, there is no limit on how many dimensions we may have. Indeed, when considering some of the multivariate techniques in this book, we will regularly work in three and higher dimensions.

A *mathematical relation* is a subset of the Cartesian product. By deleting a few points in the Cartesian product, we can appreciate the nature of a mathematical relation (Figure 2.8).

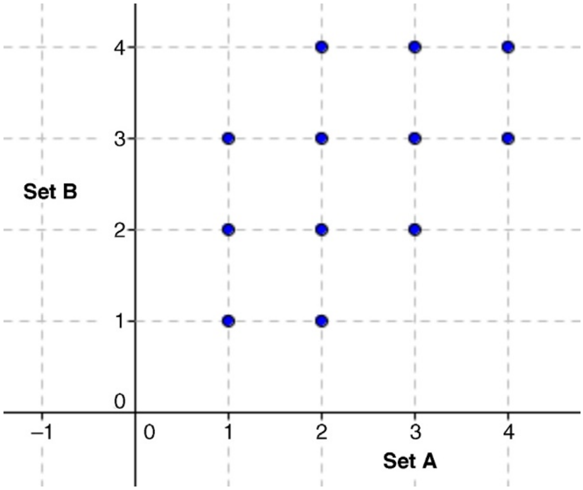


FIGURE 2.8 Mathematical relation as a subset of the Cartesian product.

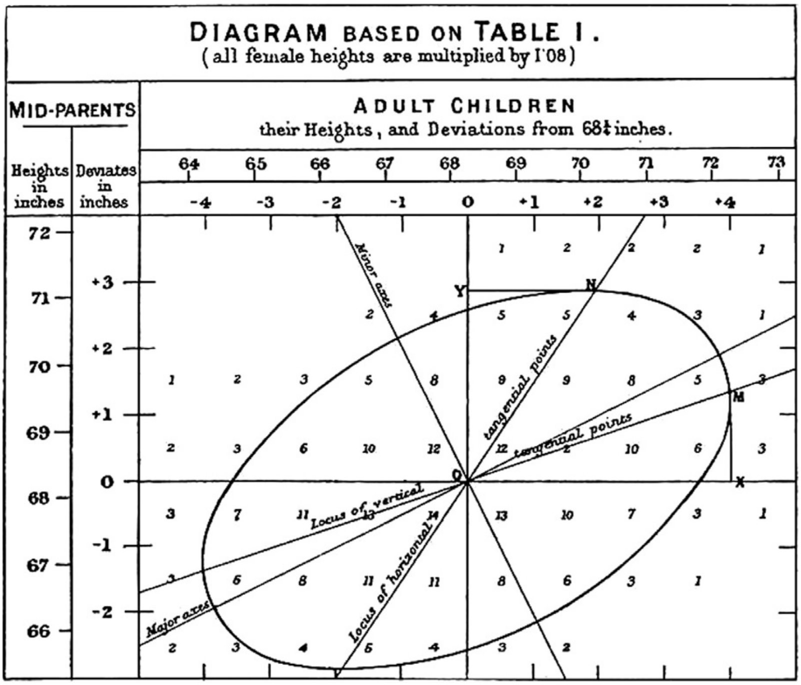


FIGURE 2.9 Galton visually spotted the ellipse that motivated the theory of correlation. That is, Galton detected a subset of the Cartesian product (Galton, 1886).

For the purpose of demonstration, we deleted a few points to reveal a meaningful (in this case, somewhat *linear*) relation between the two sets. Any subset of the Cartesian product is considered a relation between sets. Some relations, of course, are of more interest to us than others, and will be more substantively meaningful to us as scientists. The whole basis of correlational theory rests on the idea of a mathematical relation. Indeed, it was Francis Galton who visually spotted the ellipse in such a subset of the Cartesian product depicted in Figure 2.9 (to be discussed further in Chapter 8 on regression).

In Galton's plot, we can see that not all pairings of mid-parent height and child height are represented. The ellipse represents the general area where *most* of the data lay. Galton's contribution was in visually spotting a subset of the Cartesian product in the form of a linear relation between parent and offspring height. Tall parents tended to have taller children, whereas shorter parents tended to have shorter children.

2.3 SETS OF NUMBERS

As discussed, sets are useful for classifying objects. If those objects are numbers, then we can use the theory of sets to generate groupings of the various kinds of numbers encountered in mathematics.

The set of *natural numbers*, often called *counting numbers*, is given by $\mathbb{N} := \{1, 2, 3, \ldots\}$. Natural numbers are defined by the set as 1, 2, 3, etc. The set of natural numbers are numbers beginning with the number 1 and extending to infinity. It can be shown that the number of counting numbers has no upper limit, meaning that there are an infinite number of natural numbers. Natural numbers are the ones first learned in school to count such things as the number of apples on the table or the number of hours in a day, or the number of beads on an abacus.

The set of *integers* is given by $\mathbb{Z} := \{0, 1, -1, 2, -2, 3, \ldots\}$. Notice that the set of natural numbers $\mathbb{N}$ is a proper subset of the set of integers. Integers are more difficult to grasp than natural numbers since the possibility of *negative* numbers exists, which historically, was actually quite a significant transition in mathematics. But anyone who has purchased a home mortgage, financed a vehicle, or played Vegas (in the *long run* at least) likely has an intuitive (if not painful) grasp of the concept of a negative number.

The set of *rational* numbers is given by $\mathbb{Q} := \{m/n : m, n \in \mathbb{Z}, n \neq 0\}$, where m/n represents a ratio of two integers. The fact that they must be integers is denoted by $m, n \in \mathbb{Z}$. The condition that the denominator cannot equal 0 is imposed (i.e., $n \neq 0$) to avoid dividing by 0, which would make the ratio *undefined*. A rational number is a number that can be expressed by a ratio of two integers such that the quotient has a terminating (finite) or repeating decimal (recurring expansion). If you can write a number as this ratio m/n, then you have a rational number. For example, 4 is a rational number because we can write it as a ratio of 8/2. The number 10 is another example of a rational number because we can write it as a ratio of 20/2 or 100/10, for example.

If m/n does not result in a number that has a terminating or recurring decimal, then such a number is an *irrational* number. The classic example of an irrational number is $\sqrt{2}$, which, try as we may (and the Greeks tried aplenty!), cannot be expressed as a ratio of two integers. For a proof that $\sqrt{2}$ is irrational, see any of the many texts in introductory real analysis (e.g., Bartle and Sherbert, 2011). The union of rational numbers and irrational numbers is known as the set of *real numbers*, denoted $\mathbb{R}$. We will deal exclusively with real numbers in this book.

If you do have an interest in learning more about numbers in general and if questions such as the number of prime numbers that actually exist excite you, you may find the field of *number theory* to your liking.[1]

2.4 SET THEORY INTO PRACTICE: SAMPLES, POPULATIONS, AND PROBABILITY

As discussed, set theory is a field of mathematics unto itself. Set theorists are pure mathematicians who daily study and derive theorems and proofs related to sets, and have little concern in needing to define what the objects in those sets might actually be in a practical sense. That is, the sets need not be *empirical* for them to be

[1] A classic introduction to number theory is Hardy et al. (2008).

mathematically "real." As researchers and scientists however, our study of sets is not motivated by abstraction. Our study of sets is motivated by how we might use sets to group real, empirical objects. From a statistical point of view, there is no better example for making the leap from theoretical sets to real sets than through the concepts of samples and populations.

A *population* is defined as the set of objects or elements (whether they be people, animals, coin flips, etc.) we are interested in studying. This is the group of objects we wish to know something about. In an ideal circumstance, being able to study *all* the elements of a population we are interested in would be best. That way, we could make conclusions about the actual population and we would not need to estimate or infer using inferential statistics.

However, our populations are usually quite large, and collecting observations is usually a timely and expensive endeavor. For these reasons and others, we regularly collect *samples*, which in set language, are simply proper subsets of the wider population. The scientist studies the sample extensively, computing a number of useful numerical characteristics or functions on such samples, called *statistics*, and then uses such functions as *estimators* of population characteristics. The process is one of induction and inference of the sort—*if this is true of my sample, then to what extent can I say it is true of the population from which these data were presumably drawn?* The purpose of inferential statistics is to be able to generalize from the *specific to the whole*, and to be able to quantify how *good*, in some sense, that generalization is.

2.5 PROBABILITY

Probability is the mathematical language of uncertainty. Before reviewing the essentials of probability, it is well worth asking why we even require probability in the first place. We require probability because even if we believe the world is fundamentally deterministic (a viewpoint which in itself can be quite controversial), our knowledge of events that occur in the world is definitely not. Our knowledge of most events is incomplete and uncertain. We can predict events, yes, but our predictions are far from perfect. If there were no uncertainty in the world, we would have little need for probability, and by consequence, much of statistical inference would not be required either.

Probability is intrinsically difficult to define and is a very deep philosophical concern for which there is plenty of disagreement among philosophers and other thinkers. These issues are far beyond the scope of this book. For historical and philosophical accounts, the reader is strongly advised to consult Hacking (1990). Intuitively however, we all *know* what probability means. When we make statements such as "Looks like it will probably rain today," we simply mean that we think it is more *likely* to rain than not. Hence, probability is a statement of *likelihood* of an event occurring. How that likelihood is actually operationalized and quantified is the more difficult part.

2.5.1 The Mathematical Theory of Probability

We have defined probability as the mathematical language of uncertainty. However, we have not yet decided how we will think about probabilities, nor how we will assign probabilities to events. For instance, if I asked you what the probability of rain is today, you might give me a number between 0 and 1. Perhaps you believe the probability of rain today is 0.70. Was your quantification of it correct? How would we know? How did you obtain the number you got? What was your *reasoning* in estimating the probability of rain to be 0.70?

The way to correctly quantify and conceptualize probability is a debate that has existed since the origins of counting and even primitive estimation. That you can give me a number that I can call a probability in no way immediately suggests that the quantification was *correct*, *reasonable*, or in the slightest way *meaningful*. After all, probability is, *mathematically*, just a proportion between 0 and 1 (as we will see when we discuss Kolmogorov's axioms). The numbers do not care where they came from, but we, usually do. Analogously, statistical software cares little about where the numbers came from either, but as scientists, it is virtually all that matters.

2.5.2 Events

Defining an "event" in probability theory is not as easy as it first sounds. To know what an event is, and how it is used, it is first useful to define what it is not. An event is not something that happens. It is not a flip of a coin, it is not getting heads or tails on that coin. *An event is a possible outcome (subset) in a sample space*. For instance, heads and tails are events only if they are possible outcomes (see Figure 2.10) of the experiment or trial.

After you flip the coin, and get tails, we usually call this the *realized event* or simply, *realization*. When we typically speak of events, we associate with each event a *sample point*, which is simply a point that represents the event (see Figure 2.10).

Elementary events are those that cannot be decomposed into smaller events. It is a *singleton* of a sample space. For example, "head" on a coin is an elementary event because it cannot be decomposed into simpler, smaller events. However, *suit of card* is not an elementary event, because it can be decomposed into smaller events (e.g., nine of spades). The *sample space* for an experiment in probability is the set of all possible elementary events defined on that space.

If the event can be decomposed into smaller events, then the event in question is called a *compound* event. Such an event consists of multiple, simpler events.

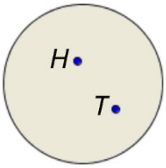


FIGURE 2.10 Sample space for a coin, where H and T are events (possible outcomes in the sample space). Associated with each event is a sample point.

In probability theory, we often speak of *experiments*. They are also called *trials*, and are in no way equivalent to laboratory experiments in the way that we manipulate levels of an independent variable and observe a response on a dependent variable. Examples of experiments in probability theory include the flipping of a coin, a client entering your office, a rat pressing a lever for food, or a bolt of lightning striking a tree. All of these, in a general sense, can be regarded as experiments as they are related to the theory of probability. Fundamentally, they are all repeatable and hence, at least theoretically, we can assign probabilities to their outcomes.

2.5.3 The Axioms of Probability: And Some of Their Offspring

We now survey the mathematical theory of probability by first noting the axioms of probability as developed largely by Andrei Kolmogorov, a Russian mathematician who lived from 1903 to 1987 and who made several contributions to both mathematics and statistics. Kolmogorov suggested the following axioms:

- The probability of an event in a given set is greater than or equal to 0, $p(A) \geq 0$ for all elements in A. That is, a probability is a nonnegative real number.
- The probability of the entire sample space must equal 1. That is, $p(S) = 1$, where S denotes the sample space. Pragmatically, what this axiom essentially guarantees is that on a given experiment, something *must* happen. For example, if I flip a coin, a head or tail must occur (assuming the coin cannot land on its edge).
- If events $A_1, A_2, \ldots, A_n$ are pairwise mutually exclusive (or pairwise mutually *disjoint*, which is another name for mutually exclusive, see Section 2.5.5), then the union of their probabilities is equal to the sum of their probabilities. More formally,

$$p(\cup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} p(A_i)$$

Any function that satisfies these three axioms is known as a *probability function*. From these axioms, we can deduce several rules of probability such as the following:

- $p(\varnothing) = 0$ (the probability of the null or empty set is equal to 0)
- $p(A) \leq 1$ (the probability of any event in the sample space S must be less than or equal to 1)
- $p(\overline{A}) = 1 - p(A)$ (the probability of the complement of event A is equal to 1 minus the probability of A). Here, $\overline{A}$ denotes the complement to whatever larger set we are considering.

Why should we believe in such rules of probability? All of these rules, and others, can be justified by tracing their paths back to the original axioms set out by Kolmogorov. For a proof of these rules, see DeGroot and Schervish (2002). The proofs, however, are not tremendously enlightening from an applied point of view, meaning that they will typically not generate any “ah ha!” moments for you. The key thing to recognize from an applied perspective is that the rules used in probability theory are *justifiable* and not simply drawn out of thin air. They are not just “made up.” They are founded on the well-established and well-accepted axioms of mathematics. Of course, should the axioms one day prove to be faulty[2] or otherwise not work and lose their utility, then there is no guarantee that the derived rules will still hold. However, until that day should come, we operate under the assumption that they are true, and proceed to build additional probability rules on them as a foundation.

2.5.4 Conditional Probability

Conditional probability is a very important topic to both the disciplines of mathematics and statistics proper as well as to applied scientific domains. Conditional probabilities are just as they sound, they are probabilities that are *conditional* or *contingent* upon some other event.

For example, suppose the *unconditional* probability of getting cancer is equal to 0.10. Now, if I selected an individual at random from the population, and learned that individual has been smoking two packages of cigarettes *per day* for the past 30 years, we would probably both agree that the probability of cancer for this individual is not equal to the *unconditional* probability of 0.10. That is, what we have just agreed on is that

$$p(C) \neq p(C/A)$$

where $p(C)$ is the probability of cancer and $p(C/A)$ is the probability of cancer given addiction to cigarettes. If, on the other hand, the person we randomly selected had mini-wheats as his favorite cereal without any mention of smoking cigarettes, we would probably agree that

$$p(C) = p(C/M)$$

where again $p(C)$ is the probability of cancer, but now $p(C/M)$ is the probability of cancer given mini-wheat eating. In this case, we would expect the unconditional probability ($p(C)$) to equal that of the conditional probability ($p(C/M)$).

Conditional probabilities allow us to *narrow* the sample space so that we may “zero in” on a more well-defined set of elements and assess their probability. More formally, we can state the conditional probability of an event B given that event A has occurred as

$$p(B/A) = \frac{p(A \cap B)}{p(A)}, p(A) \neq 0 \tag{2.1}$$

[2] An example of what was thought to be an axiom (though different in flavor to the probability axioms we are discussing), but was later proved not to be an axiom is Euclid’s fifth postulate. See Boyer and Merzbach (1991, p. 106) for details.

In words, (2.1) reads that the probability of event B given that event A has occurred is equal to the probability of A and B occurring together relative to the probability of A occurring, assuming that the probability of A cannot equal 0. If $p(A) = 0$, then $p(B/A)$ is undefined, since any number divided by zero is, by definition, undefined.

From the definition of conditional probability in (2.1), we can also calculate

$$p(A/B) = \frac{p(A \cap B)}{p(B)}, p(B) \neq 0$$

In general, the conditional probabilities $p(B/A)$ and $p(A/B)$ will *not* be equal, because they represent different sets of events. For example, it would be unreasonable to think that $p(C/M) = p(M/C)$ should hold (i.e., the probability of cancer given mini-wheat eating is likely not equal to the probability of mini-wheat eating given cancer).

All scientific investigation can be said to ultimately be about conditional probabilities. For instance, we are rarely interested in the probability of schizophrenia. We are much more interested in probabilities such as that of schizophrenia *given* factors as genetic history and environment. The meteorologist is not interested in the probability of rain. She is interested in the probability of rain given certain atmospheric conditions. Likewise, the student is not interested in the probability of passing a course, he is rather interested in the probability of passing given that he studies a certain amount of time, attends lectures, etc. Hence, the probabilities of true interest to us, regardless of field, are usually *conditional* ones. Theoretically, if we know enough about the conditions, we can "zero in" on more precision regarding the probability of interest.

2.5.5 Mutually Exclusive versus Independent Events

Two events A and B are considered *mutually exclusive* or *disjoint* if the probability of their intersection is equal to zero. That is,

$$p(A \cap B) = 0 \tag{2.2}$$

For numerous events, say, $A_1, A_2, \ldots, A_n$, we say that these events are *pairwise* mutually exclusive if $A_i \cap A_j = 0$ is true for all events $A_1, A_2, \ldots, A_n$ such that $i \neq j$ (i.e., an event cannot be mutually exclusive from itself). "Pairwise mutually exclusive" simply means that any two events in the set are mutually exclusive. This saves us the trouble of having to list *all* pairings that are mutually exclusive. In a Venn diagram, two mutually exclusive events are depicted by their events having an intersection of zero, or the null set (Figure 2.11).

Examples of mutually exclusive events are easy to come up with:

- Events night and day are mutually exclusive. That is, it is either night or day and not both (i.e., In Figure 2.11, let set A = night and set B = day; since it cannot be "night and day" simultaneously, the set of elements containing these two events is empty).

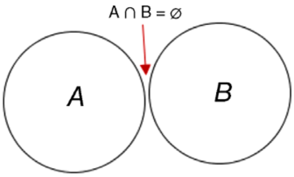


FIGURE 2.11 Sets *A* and *B* have an intersection containing no elements; *A* and *B* are mutually exclusive events.

- Events tall and short are mutually exclusive. One cannot be the event "tall" and also be the event "short."
- Events happy and sad are mutually exclusive, though existentially, I suppose it is possible to be in both states at once, which would violate their disjointness. For instance, you may be both happy and sad about your friend embarking on a new career overseas.

Two events *A* and *B* are considered statistically *independent* if the probability of their intersection is equal to the product of their individual probabilities. That is,

$$p(A \cap B) = p(A) \cdot p(B) \qquad (2.3)$$

We can also express (2.3) as the probability of the joint occurrence of *A* and *B* equal to the product of the respective *marginal* probabilities of *A* and *B*, where the marginal probabilities of *A* and *B* are given by $p(A)$ and $p(B)$. Loosely speaking, when we say two events are independent of one another, it means that the occurrence of one event in no way influences the occurrence of the other event. Examples of independent events that indeed *should* be independent, and if not, could face legal challenges, include the following:

- Employment and gender should be independent; whether one is male, female, or other, should not have any effect on the probability of gaining employment.
- Verdict should be independent of race of defendant in a court trial.

Many court challenges in areas of employment law (e.g., see Zeisel and Kaye, 1997) and racism are actually accusations on the part of the plaintiff that the events under consideration are not empirically independent, but under the law, they should be. Showing a violation of *substantive* independence of two events can be exceedingly difficult in practice. Can you really demonstrate to a jury or judge that a hiring committee purposely did not hire a man or woman of a particular race? While you may have a strong suspicion that such racism has occurred, it is quite another matter to actually demonstrate it. Even in cases where one can show lack of independence quantitatively (e.g., through a statistical test), it is still quite another matter to legally prove that such lack of independence had its origin in the cognition of the accused, that there was *real* discrimination going on. Statistical dependence is only an indicator that an underlying directional *action* or *process* may be occurring. More generally, and in the language of hypothesis testing, rejecting a null hypothesis never *proves* a substantive alternative, a topic we will return to later.

How does independence arise mathematically? If events A and B are independent, then $p(B/A) = p(B)$, so we can rewrite (2.1) as

$$p(B) = \frac{p(A \cap B)}{p(A)}$$

The probability of the joint occurrence of A and B is thus equal to

$$p(A \cap B) = p(A) \cdot p(B)$$

which gives us the rule for independence of (2.3).

Note that if two events A and B are mutually exclusive, it stands that they cannot simultaneously be independent, since if A and B are indeed mutually exclusive, then, by definition $p(A \cap B) = 0$ is true. If A and B were independent however, then as we just saw, (2.3) should be true. Since $p(A) \cdot p(B) = 0$ does not hold unless $p(A) = 0$ or $p(B) = 0$ (or both are equal to zero), then it cannot be the case that two mutually exclusive events are simultaneously independent.

As an example, consider the events “head” and “tail” on a single flip of a coin. They are obviously mutually exclusive events. Are they independent? We usually would not ask the question given that we have only conducted a single trial, but if we did, we would conclude that getting a head on the coin tells us *everything* about the probability of getting a tail. Given that we obtain a head, the probability of obtaining a tail is equal to zero. Again, we ordinarily would not ask such a question, since it seems foolish to ask of the probability of an event *after* it has already occurred (*logically*, the probability does not exist). But when we do push the limits on contrasting the concepts of mutual exclusiveness versus independence, this is what we find.

2.5.6 More on Mutual Exclusiveness

A further distinction on mutual exclusiveness is required. Consider once more the event head and the event tail on a single flip of a coin. Since one cannot obtain a head *and* a tail on a single flip of a coin, the events are deemed mutually exclusive. But this is only so *in this particular context*. Why is the caveat *in this particular context* required? It is necessary because if we consider *two* successive flips of the coin, then the event head and the event tail are *not* mutually exclusive. What this means is that when you define the mutual exclusiveness of two events, you must also state the *context* or *physical model* you are assuming or imposing when applying the property. *On a single flip of a coin, events head and tail are mutually exclusive*” is the correct way to describe the context. Head and tail are not, by themselves, intrinsically mutually exclusive events.

Notice as well that one does not simply apply the "formula" $A \cap B$ to "discover" if two events are mutually exclusive. For instance, on a single flip, are the events H (head) and T (tail) mutually exclusive? The answer is, of course, yes, because one cannot obtain a head and a tail on a single flip of the coin. However, if one naively treated the concept of mutual exclusiveness as a formula and sought to find the answer numerically, one could theoretically compute

$$p(H \cap T) = (.5)(.5) = 0.25$$

and potentially, and erroneously, conclude that since $p(H \cap T) \neq 0$, events H and T are not mutually exclusive! Of course, this would be an error, since we know that $p(H \cap T) = 0$ holds by how we defined our *context* for discussing the mutual exclusiveness. Multiplying probabilities in this case and obtaining $p(H \cap T) \neq 0$ does nothing to counter the fact that heads and tails are mutually exclusive events on a single flip of the coin. The lesson to be learned from this example, one which will in one way or another be repeated throughout this book in a variety of contexts, is the following—*Formulas and equations used blindly will provide blind answers. One must first decide on the correct use of a formula, statistical method, or other computation, including those offered by software, before one can have any confidence in the result. Formulas and equations never speak for themselves. You must oversee their correct interpretation.*

If you blindly "trust" statistical machinery rather than see it as a tool requiring human oversight, then you can easily fall into such traps of trying to solve a problem or conduct an analysis in a context that makes little, if any, sense. Before you apply any statistical model, and before you do any computations, it has to "feel right" that you're proceeding correctly. Statistics and probability are quite slippery, so after all is said and done, if something still feels wrong or out of place with your model, then something likely *is* amiss.

2.6 INTERPRETATIONS OF PROBABILITY: FREQUENTIST VERSUS SUBJECTIVE

Though the *computation* of probabilities is generally agreed upon, the *interpretation* one gives to those probabilities is definitely not. Historically, the traditional interpretation of probability is to take a relative frequency as the "best guess" of the true probability of an event. This is the so-called *frequentist* version of probability. However, as we will soon discuss, relative frequency interpretations of probability carry with them some difficulties.

An even more primitive interpretation of the probability of event A is to simply take the number of elementary events in set A relative to the total number of elementary events in the sample space S. This version of probability is often referred to as *classical* or *analytical*.

For example, consider the probability of drawing any single ball out of a bag containing 10 balls (numbered 1 through 10). The number of elementary events comprising A is equal to 1 (since we are drawing only *one* ball). The denominator of the ratio is equal to the number of *ways* the given event can occur, which in this case is equal to 10. The probability of A is thus 1/10. Now, consider the probability of drawing a 3 *and* a 5 from the bag containing 10 balls, irrespective of order. The denominator will still be equal to the number of ways the event can occur, but now, the question is—*How many ways can you draw 2 objects out of a group of 10 objects, irrespective of order?* For this, we compute the number of *combinations* of choosing 2 out of 10:

$$\binom{n}{r} = \frac{n!}{r!(n-r)!}$$

where n is the number of objects we have, r is the number we are choosing, and $n!$ is defined as $n(n-1)(n-2)(1)$. For the case of choosing 2 out of 10, we obtain

$$\begin{aligned}\binom{10}{2} &= \frac{10!}{2!(8)!}\\ &= \frac{3628800}{80640}\\ &= 45\end{aligned}$$

Hence, the number of ways you can choose 2 objects out of 10, when the order of the sampled balls does not matter, is equal to 45. Therefore, the probability of any one of those ways is 1/45. Notice that the 45 here represents the total number of events in S, only that now, the elementary "unit" is defined as being 2 objects. When we selected only a single ball out of the 10, there were 10 ways in which the event could occur. If you always think about the denominator as representing the number of *ways* rather than simply the number of "things" there are to sample from, you will usually understand the probability problem.

2.6.1 Law of Large Numbers

Why should we believe that the relative frequency of an event is a good estimate of the true probability of that event? This is justified in probability theory by a theorem called the *law of large numbers*. The theorem states that if an experiment or trial is repeated many times under identical conditions, the relative frequency of occurrence of the event is likely to be close to the probability of that event. As the number of repetitions increase, we zero in on the true probability.

We can represent the law of large numbers more formally as follows. If an experiment is repeated n times and r denotes the number of times that the event E occurs in these n repetitions, then the following is true:

$$p[(|(r/n)-p(E)| \geq \varepsilon)] \rightarrow 0 \tag{2.4}$$

What (2.4) means is that as n gets larger and larger, without bound, the absolute distance between r/n and the true probability of an event E, denoted by $p(E)$, being equal to or greater than some small positive number ε (think of this as being some small *magnitude*), goes to zero. That is, in the *long run*, there will be no difference between the relative frequency of an event r/n and its true probability, $p(E)$. We can also express (2.4) using a limit[3] concept more generally for that of sampled random quantity x based on n samplings:

$$\lim_{n\to\infty} p(\,|x_n - x)| < \varepsilon) = 1$$

which says that in the limit as $n \to \infty$, the probability of any absolute distance or difference between our sampled sequence of random values x_n and x being smaller than ε, an arbitrary positive number, is equal to 1 (we will name x a *random variable* in the following chapter). The law guarantees, in a theoretical sense, that as we draw more n, a *convergence* toward x occurs. For further discussion on this particular interpretation, see Casella and Berger (2002, pp. 232–233).

In practical terms, the law of large numbers implies that even if we have only a limited number of trials, we should generally *expect* the probability of any event to be reflected in the relative frequency we actually observe for the given event. In the long run, such an observed relative frequency should approach the true probability.

2.6.2 Problem with the Law of Large Numbers

Since we cannot ever obtain an infinite number of repetitions, no matter how many repetitions we do have, we might expect r/n in (2.4) to be *close* to $p(E)$ but we can never be sure they are *equal*. The idea of letting sample size grow infinitely is quite unrealistic. However, as noted by Casella and Berger (2002), "Although the notion of an infinite sample size is a theoretical artifact, it can often provide us with some useful approximations for the finite sample case, since it usually happens that expressions become simplified in the limit" (p. 232).

Thus, we can tentatively conclude that the law of large numbers, though a useful concept, is entirely unachievable and truly makes sense only theoretically. *Would it not be a good idea then to adopt an interpretation of probability that does not require the law of large numbers?* The subjective interpretation of probability accomplishes this goal, which we now briefly discuss.

2.6.3 The Subjective Interpretation of Probability

According to Winkler (2003, p. 14), "The probability concept acquired an interpretation in terms of relative frequency because it was originally developed to describe certain games of chance where plays [. . .] are indeed repeated for a large number of trials and where it is reasonable to assume that the elementary events of interest

[3] The concept of a limit is briefly discussed later in the chapter.

are equally likely." That is, in gambling games where probability theory began in the mid-1600s, it was usually a safe assumption to make that a game can be repeated over and over, hence having an infinite number of trials. And though even in this context this idea of an infinite number of trials is still unrealistic, it was nonetheless *imaginable*.

There are many situations in research, however, and in daily life where being able to assume an infinite number of trials simply does not make sense, and hence basing a theory of probability on the law of large numbers does not work in such cases. The *subjective* interpretation is especially useful when we cannot use a relative frequency interpretation. The subjective interpretation holds that probability is a *degree of belief*, and not simply a relative frequency. The subjective probalist can still use frequency information in her estimate of the probability of an event, but the point is that she is not *restricted* solely to frequency information. She may draw from a variety of informational sources in deriving her estimate. For instance, consider how you might go about evaluating the probability of the following events:

- Probability of a nuclear world war in the next 5 years
- Probability of an earthquake in California in the next 50 years
- Probability of the earth exploding tomorrow

In these scenarios and others, it is difficult if not impossible to assign a probability based on a relative frequency interpretation. We have never had a nuclear world war, and so we cannot "flip" the event to see how many times a nuclear war shows up as we could with a coin. The "Big One" has yet to occur in California, and thus any estimate of its probability must be based, at least in part, on information external to the event under consideration. The earth has never exploded before, and so estimating the probability that it will explode tomorrow is very difficult to put into relative frequency terms. However, we would still like to provide probability estimates for such events and others like them where relative frequency seems to fail us.

Consider the information we may use in estimating the probability of a nuclear world war in the next 5 years:

- Political climate and stability of nations possessing nuclear arms
- Probability of nonpossessing nuclear countries obtaining nuclear arms in the next 5 years (and the probability that these arms could be or are used)
- Political motivation for any nation or individuals seeking to use a nuclear bomb

The list goes on and on regarding the information we may wish to use in arriving at an estimate of the probability of nuclear world war. If we end up generating an estimate of say, 0.001, it would have been generated based on our *opinion* or *belief* of the probability of war, presumably by incorporating *all* information available, including that of relative frequency information where appropriate. In this sense, relative frequency probability can be considered a *subset* or *special case* of the wider subjective probability framework, as depicted in Figure 2.12.

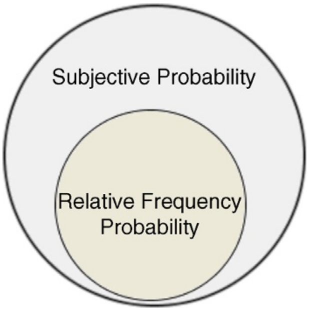


FIGURE 2.12 Relative frequency probability can be interpreted as a special case of a subjective probability.

2.7 BAYES' THEOREM: INVERTING CONDITIONAL PROBABILITIES

Bayes' theorem, sometimes called "Bayes' rule," provides a way to evaluate relationships among various conditional probabilities. More specifically, it allows us to "invert" probabilities such that we begin with $p(B/A)$, and are able to compute $p(A/B)$ (notice that A and B have switched places, they have been *inverted*). The theorem is named after Reverend Thomas Bayes (1702–1761), a British clergyman and great thinker (see Figure 2.13).

FIGURE 2.13 Thomas Bayes (1701–1761).

The theorem can be stated:

$$p(A/B)=\frac{p(B/A)p(A)}{p(B/A)p(A)+p(B/\overline{A})p(\overline{A})} \tag{2.5}$$

where $\overline{A}$ denotes the complement of event A.[4] Bayes' theorem reveals that the conditional probability can be computed in a particular unique, and as we will see, extremely useful way. In addition to being aesthetically pleasing to the probalist or mathematician, Bayes' theorem is also extremely practical, as we shall soon see by constructing some examples of its use. There are entire fields of statistics and philosophy devoted to the study of *Bayesian analysis*, including a variety of procedures developed for the estimation of posterior probabilities and distributions. What follows is only a peak into this vast world of analysis. For more on Bayesian analysis, refer to Gill (2014). Savage (1972) is also a classic resource.

2.7.1 Decomposing Bayes' Theorem

We take a closer look now at the elements of Bayes' theorem and discuss each component. The first probability estimate that enters into Bayes' theorem is $p(A/B)$. This is what we seek to know from our calculation. It is the probability of some event A, given some other event B. It is thus a *conditional probability*. In the paradigm of hypothesis testing, we will denote $p(A/B)$ as $p(H/D)$, where H stands for "hypothesis" and D stands for "data." That is, the question we are asking is: *What is the probability of our hypothesis given obtained data?*

The next estimate that enters into Bayes' formula is $p(B/A)$, or $p(D/H)$, the probability of some data given some hypothesis. This is the probability estimate that dominates classical statistics. When we conduct a statistical significance test (e.g., $p < .05$), we are computing in a general way the probability of the observed data given some null hypothesis. This conditional probability is often referred to as a *likelihood*; it is the likelihood of the obtained data given the hypothesis. Note carefully that it is not the probability of the hypothesis given some data. That is,

$$p(D/H) \neq p(H/D)$$

To get $p(H/D)$ from knowledge of $p(D/H)$, we need to *invert probabilities*. This is what Bayes' theorem allows us to do.

[4] A more general form of Bayes' theorem is the following:

$$P(A_j/B)=\frac{P(B/A_j)P(A_j)}{P(B/A_1)P(A_1)+P(B/A_2)P(A_2)+\ \dots\ +P(B/A_J)P(A_J)}$$

Note the parallel between the first formulation of Bayes' theorem given in (2.5) and the extended form. Both formulations have in common the partitioning of event A, only that in the first case, the partitioning is only between two possibilities, A and $\overline{A}$, whereas in the general formulation, it is between A_1 and A_J possibilities.

Next in (2.5), we come to the term $p(A)$. In hypothesis testing language, this is the probability of our hypothesis, H, that is, $p(H)$. It is the *prior probability* estimate of the hypothesis. How this single estimate is obtained is the source of much controversy and debate (and misunderstanding) concerning Bayes' theorem. The term itself is meant to represent our prior knowledge of the hypothesis under consideration before "revising" this estimate in line with new data and obtaining $p(H/D)$. Often, $p(H)$ can be obtained from base rate information or some similar information, but sometimes (perhaps often) $p(H)$ is not so easily obtained. Bayesian statistics often employ *subjective* probability estimates or *personal* probabilities in their initial computations to get the relevant prior probability. Historically, this has been a contentious issue between the relative frequency and subjectivist camps.

Finally in (2.5), we come to $p(B/\overline{A})$ or $p(D/\overline{H})$, which is again a likelihood. It is the conditional probability of the data given the complement of the hypothesis or theory.

An example will help elucidate how these probabilities can be used.

2.7.2 A Medical Example—Probability of HIV: The Logic of Bayesian Revision

Suppose as a medical doctor, before administering a diagnostic test you attempt to assess the probability that your patient has HIV, the virus that causes AIDS. What would be your best estimate? It would probably be the figure representing the percent of those having the disease in the city or region of interest, otherwise known as the *population base rate*. If it is known that the prevalence of HIV in your population is 1%, then for a randomly chosen individual from this population, I think you would agree that your "best guess" (so far) at the probability of the individual being HIV positive is 0.01.

Now, enter the diagnostic test. The purpose of the diagnostic test will be to *sharpen* your probability estimate based on newly acquired data. Diagnostic tests alone do not tell you whether a person has a disease or not. They are simply an *input* to the final decision. They are a *sign*. Suppose that you give the patient the diagnostic blood test for HIV. The result of this test constitutes *data*, and it is in using such data that you will revise your original probability estimate of 0.01.

What are the probabilities that we need for Bayes' theorem to work? For our example, we have the following:

- $p(H)$ is the probability of the hypothesis that the patient has HIV; it is the probability estimate we would essentially use in the absence of additional information, and as such, is our base rate of 0.01.
- $p(D/H)$ is the probability of the positive reading on the diagnostic test given that the patient has HIV; for our example, suppose this is equal to 0.98.
- $p(\overline{H})$ is the probability that the patient does *not* have HIV; it is the complement of $p(H)$, equal to 0.99 (i.e., 1–0.01).
- $p(D/\overline{H})$ is the probability of a positive reading (data) given no HIV (or "HIV negative"); for our example, suppose this is equal to 0.05 (i.e., relatively error-prone).

We now have all the information required to run Bayes' theorem, and invert probabilities:

$$\begin{aligned} p(H/D) &= \frac{p(D/H)p(H)}{p(D/H)p(H)+p(D/\overline{H})p(\overline{H})} \\ &= \frac{0.98(0.01)}{0.98(0.01)+0.05(0.99)} \\ &= 0.165 \end{aligned}$$

After revising our initial probability estimate of 0.01, $p(H/D)$ is now equal to 0.165. That is, the probability that the patient has HIV given a positive blood test is equal to 0.165, an increase from the initial estimate of 0.01. The data, in the form of the positive blood test, have added to our knowledge of the probability of disease.

2.7.3 Recap of Bayes' Theorem: The Idea of Revising Probability Estimates and Incorporating New Data

In the HIV example just featured, we have a powerful demonstration of how Bayes' theorem can be used to revise a probability estimate. Notice how we proceeded through the example. We first considered the estimate of HIV in the population from which the subject was sampled. This was our *prior probability*, which was equal to 0.01. How did we obtain this information? Although in this example we could simply rely on medical population estimates or real data about the prevalence of HIV (I made the figure up, but 1% is actually a suitable estimate for some populations), the nature of obtaining this probability estimate is not always straightforward, and again is one reason why the Bayesian approach to statistics is sometimes heavily criticized by some. This estimate, $p(H)$ is often referred to as the "prior probability" because it is calculated "prior" to consideration of the obtained data, D. The purpose of Bayes' theorem is to revise $p(H)$ in light of new information, which in the medical example was the diagnostic blood test. That is, as a medical clinician, even before the patient entered your examination room, your best "bet" about him or her being HIV positive would be 0.01, the base rate in the population from which the subject is being sampled.

After administration of the diagnostic blood test however, the probability of the patient having HIV given the result of the diagnostic test $p(H/D)$ is now "updated" to 0.165. This probability estimate is called the *posterior* probability because it is obtained after (i.e., "post") consideration of the obtained data (in this case, the positive blood test).

2.7.4 The Consideration of Base Rates and Other Information: Why Priors Are Important

The consequences of not considering base rates, or any other *prior* information, should be evident from our medical example. For instance, if instead of the prior probability being equal to 0.01, it were equal to 0.90, this would have a drastic influence on the posterior estimate. When we do Bayesian revision, what we start out with in terms of a prior probability is often just as important and in some cases *more important* than the actual data we obtain. And although prior probabilities are often considered to reflect personal opinion, there is nothing in Bayesian philosophy that says they should be "irrational" or otherwise poorly derived. At minimum, however, if you come up with a ridiculous prior, it is there for everyone to see, and hence will not be taken very seriously. For instance, had we started out with a prior probability of 0.90 for HIV, a critic could easily, and rightfully, dismiss our analysis since 0.90 is obviously an irrational prior for even relatively high-risk populations. Prior probabilities, whether in the context of Bayesian revision or other research settings, even if constituting one's "subjective" opinion, should nonetheless be "reasonable."

2.7.5 Conditional Probabilities and Temporal Ordering

With regard to Bayes' theorem and conditional probabilities in general, it should be noted that conditional probabilities do not care about *temporal order*. We, however, usually *do*. For instance, pondering over the probability of something that occurred in the past given something that occurred more recently makes no sense to us *logically*. However, the *calculus* of probabilities themselves care little about time. In other words, they are still computable. They may not be *interpretable*, but they can still be calculated numerically.

For example, the probability of one's suicide 10 years ago could never, in a rational way, be conditional or contingent upon an event that occurs today. The problem simply makes no sense to us.[5] However, one could still obtain a conditional probability of such an event. That is, we could still compute a *number* for it. The computations themselves do not "know" what the events in the problem actually represent. *We give them meaning, otherwise they have none*. This is again a reminder of why you should never blindly trust calculations of any kind when computing probabilities or anything else in a research setting. *The problem must make sense to you first for things to work out as they should, and be correct. Understanding*, *comprehension*, and *judgment are not things you can subject to a computing algorithm.*[6]

[5] Perhaps from a more existential or metaphysical perspective, all events from the beginning of what we call "time" to the end know no temporal order, making all events, regardless of time, inexorably "linked." This would make one's suicide 10 years ago perhaps contingent on an event in the present day. We rule out such possibilities, however.

[6] Of course, those who specialize in artificial intelligence may challenge this statement. My purpose is not to initiate debate or controversy over systems that aid us in decision making. My point is simply that if we do not first have a solid grasp of what we are computing or subjecting *to* the given algorithm, we are on shaky ground when attempting to interpret the output. This holds true in one's use of sophisticated scientific or statistical software just as much as it is true in the use of one's pocket calculator. Statistical software is best conceived as nothing more than a sophisticated calculator. It is the *interpretation* of what the software produces that requires a diligent study of statistical theory and methodological (and philosophical) principles.

2.8 STATISTICAL INFERENCE

Statistical inference is a process by which conclusions and decisions about population parameters are made based on information obtained in samples. It is most easily understood through the use of very simple examples. The classic example is, again, that of a fair coin. Suppose I hold a coin in my hand and for no reason other than my prior experience with coins, I assume the coin is *fair*. That is, I assume that on any given flip of the coin, the probability of getting a head or a tail is equal to the same value, $p(H) = 0.5$ and $p(T) = 0.5$. Note carefully that we are starting off the process with an *assumption*. We are having to first assume something about the coin. The assumption is the hypothesis that is presumed true pending further evidence. Such an assumption often goes by the name of a *null hypothesis*. It is the hypothesis that is assumed to be true in generating the sampling distribution (see Chapter 3) of the test statistic appropriate for the given problem, and is the hypothesis we will attempt to *reject* given evidence that contradicts it. More generally, in the case of the coin, you can think of this assumption or null as the "status quo" or your prior belief (to impose a Bayesian flavor) in the probability characteristics of the coin. Null hypotheses do not need, however, to be statements of *equality*.

We now proceed to sample some data. Suppose our first flip (or our first "trial") of the coin turns up a head. Would you doubt the assumption of a fair coin based on this piece of evidence? Likely not. After all, it's just one flip, and getting a single head is not reason to reject our assumption. Suppose we keep taking trials, and obtain the following sequence of flips for the first 10 trials:

H H H H T H H T H H

We obtained eight heads and two tails. The question we want to ask is: *What is the probability of obtaining 8 heads out of 10 flips under the assumption that the coin is fair?* That is, what is the probability of obtaining 8 heads out of 10 flips under the assumption of the null hypothesis? If the probability of such a sequence is *low*, then it suggests that the obtained data may not have arisen under the model (i.e., assumption) we started out with. Obtaining 8 heads out of 10 flips seems more likely to have occurred under a different model, one for which the coin is *not fair*. This alternative model is typically housed in the *alternative hypothesis*, and may take many forms depending on the given context. It may simply be a statement that $p(H) \neq 0.5$ or even a more specific, well-defined hypothesis such as $p(H) = 0.8$, in which case our obtained sample of 8 heads out of 10 flips would "fit" such an alternative very well.

The key point in this example, as is true of virtually all examples of statistical inference, is that we are evaluating the probability of *data* given some hypothetical situation which we couch in terms of a null hypothesis. If the probability of data is low under our hypothetical situation, then it serves to cast doubt on that hypothetical situation, and probabilistically, we begin to think that perhaps a competing model (i.e., one other than the null hypothesis) is better able to account for the obtained data. Even intuitively, a result of 8 heads out of 10 flips would cause us to *doubt* our

TABLE 2.1 Decision Table for Classical Hypothesis Testing

Quantifying consequences of a decision in terms of losses		State of the world	
		(H_0 true) $\theta \in A$	(H_1 true) $\theta \in B$
Decision	Accept H_0	0	Q_2 [*Loss II*] (type II error)
	Accept H_1	Q_1 [*Loss I*] (type I error)	0

assumption that $p(H) = 0.5$ and we would reject it in favor of the alternative hypothesis (e.g., $p(H) \neq 0.5$.)

2.8.1 Shouldn't the Stakes Matter?

If something does not sit right with you with regard to the previous example, it is a good thing. You may be wondering how we can decide to reject the null hypothesis or infer the alternative hypothesis without any sense of the *consequences* of making the *wrong decision*. Yes, in our example of the coin, the situation is trivial. That is, the costs associated with deciding whether the coin is fair or not fair are likely nonexistent. However, in other paradigms, life or death may be at stake, or at minimum, the consequences of making the wrong decision may be more severe or grave than losing a coin-flip betting game.

In any decision of this sort, where two choices are available to the decision maker, two types of error might occur. The first is that of a *type I error*, which is a *false rejection of the null hypothesis*. The probability of a type I error is set at α, the significance level of the test. The second type of error is that of a *type II error*, which is failing to reject a false null hypothesis. A type II error is equal to β (*Beta*), and will depend on such things as the distance between the null and the hypothesized alternative.

We can display both types of error in the classic decision table featured in Table 2.1. Note that in addition to specifying type I and type II error rates, Table 2.1 also depicts a *loss function* associated with the decision we are making.

Note that the decision to "Accept H_0" is associated with zero loss if in fact H_0 is true (cell in row 1, column 1). Note also that the decision to "Accept H_1" if in fact H_1 ends up being true is likewise associated with zero loss (cell in row 2, column 2).

Losses begin to occur when our decision does not accord with the state of the world. If we "accept" H_1 but H_0 turns out to best reflect reality, then we have made a false rejection of the null hypothesis (or, in the case of decision theory, a false acceptance of H_1). This is the classic type I error of both Fisherian and Neyman–Pearson hypothesis paradigms.[7] Similarly, an incorrect acceptance of H_0 when in fact H_1 is true leads us to make the classic type II error. Which is the more daunting error to

[7] For an historical and very readable account of the Fisherian versus Neyman–Pearson camps, see Denis (2004).

TABLE 2.2 Updated Decision Table for PTSD Example

Quantifying consequences of a decision in terms of losses		State of the world	
		(H_0 true) Population is normal $\theta \in A$ $p(H_0) = 0.70$	(H_1 true) Population suffers from PTSD $\theta \in B$ $p(H_1) = 0.30$
Decision	Accept H_0 [population is normal]	0	Q_2 [*Loss II*] (type II error)
	Accept H_1 [population PTSD]	Q_1 [*Loss I*] (type I error)	0

make? Without associating *costs* with each error, we have no means by which to evaluate which is the more consequential error.

This is where *decision theory* comes in. Decision theorists advise that one should assign a loss function with each type of error. These are represented by Q_1 and Q_2 values in Table 2.1, referred to as *Loss I* and *Loss II*, respectively. If one can quantify these losses somehow (a point we will return to shortly), then one can incorporate such losses into the computation of *expected values* (see Chapter 3) in the decision-making process.

For example, consider the updated decision table in Table 2.2, based on hypothetical data. In this case, the researcher was interested in learning whether the population she is studying suffers from posttraumatic stress disorder (PTSD), which is a mental health condition one may develop after exposure to one or more stressful life events.

From the table:

- The prior probability that the population is normal is $p(H_0) = 0.70$.
- The prior probability that the population suffers from PTSD is $p(H_1) = 0.30$.
- The loss incurred if the decision is made that the population is normal given the population actually *is* normal is 0 (row 1, column 1).
- The loss incurred if the decision is made that the population suffers from PTSD, if the population actually has PTSD, is 0 (row 2, column 2).
- The loss incurred if the decision is made that the population is normal but actually has PTSD is Q_2, constituting a type II error (row 1, column 2).
- The loss incurred if the decision is made that the population has PTSD but is actually normal is Q_1, constituting a type I error (row 2, column 1).

After incorporating losses, our final decision table is that given in Table 2.3.

How these numbers were computed is not essential for understanding the point we wish to make (for computations of this sort, see Winkler (2003)). For our purposes

TABLE 2.3 Final Decision Table for PTSD Example

Quantifying Consequences of a Decision in Terms of Losses		State of the World	
		(H_0 true) Population is normal $\theta \in A$ $p(H_0) = 0.70$	(H_1 true) Population suffers from PTSD $\theta \in B$ $p(H_1) = 0.30$
Decision	Accept H_0 [Population is normal]	0	Q_2 *Loss II* = 3 (type II error)
	Accept H_1 [Population PTSD]	Q_1 *Loss I* = 1 (type I error)	0
Posterior odds = prior odds × likelihood ratio × loss ratio		$\frac{p(H_0/D)}{p(H_1/D)} \cdot \frac{Q_1}{Q_2} = \frac{0.70}{0.30} \cdot 1.20 \left[\frac{1}{3}\right]$ $= 0.93$	
Final decision		Since $\frac{p(H_0/D)}{p(H_1/D)} \cdot \frac{Q_1}{Q_2} < 1.0$, reject H_0, conclude PTSD in population	

here, it is enough to note that by combining the prior probabilities and likelihood with a ratio of the losses (i.e., Q_1 to Q_2), a decision was made to reject the null hypothesis and conclude there to be PTSD in the population. The relevant equation is –*Posterior odds = prior odds × likelihood ratio × loss ratio*, which in Table 2.3 is given by

$$\frac{p(H_0/D)}{p(H_1/D)} \cdot \frac{Q_1}{Q_2} = \frac{0.70}{0.30} \cdot 1.20 \left[\frac{1}{3}\right] = 0.93$$

The fact that the potential losses incurred in accepting H_0 if H_1 were actually true (3) are greater than the potential losses incurred in accepting H_1 if H_0 were actually true (1) helped *decrease* the prior odds of 0.70/0.30 down to 0.93, within the threshold required to reject the null hypothesis, as indicated in the "Final Decision" of Table 2.3. How the loss ratio of 1/3 was determined in this case was merely through an informal weighting of the consequences of the two types of errors. For this example, the clinician simply deemed concluding the population to be normal when in fact it is not to be more *serious* of an error than concluding the population suffers from PTSD when in fact they do not. One can easily imagine why this may be so. The failure to treat a population suffering with PTSD may have more threatening consequences than treating a population that does not truly require it.[8]

[8] For example, the suffering of a war veteran not diagnosed may have more serious consequences than treating that war veteran even if in actuality he or she does not suffer from the disorder. Of course, the costs associated with treating someone who does not have a disorder may in itself constitute a *financial* cost, but perhaps not as serious as the error of failing to diagnose someone with a real mental health condition.

Such a simple example highlights the importance of considering statistical conclusions within a larger decision analysis framework, one in which costs associated with decisions are incorporated in the hypothesis testing methodology. Statistical analysis, in the end, is often about *decisions*. Excellent books on the topic include Clemen and Reilly (2004) and Winker (2003). For a more technical read, consult Berger (1993).

2.9 ESSENTIAL MATHEMATICS: PRECALCULUS, CALCULUS, AND ALGEBRA

We suspend our discussion of further probabilistic concepts and statistics until Chapter 3 where we survey and review essential statistics in more detail in preparation for the rest of the book. We instead turn now to surveying some of the elements of modern mathematics, focusing primarily on concepts from precalculus up to and including calculus. Much of what we cover only skims the surface, and the reader is strongly advised to refer to sources cited within the text, or simply refer to Labarre (1961) for a classic and very readable overview of fundamental mathematics (the book is somewhat dated, but extraordinary in its clarity). Central to our brief overview and discussion of modern mathematics is the development of the idea of a *function*. Functions are what mathematics, statistics, and research are all about. To set the stage for such a discussion, we first begin with a review of polynomials, of which functions are a special case.

2.9.1 Polynomials

A *nomial* is a single mathematical expression usually with a variable in it. For example, ax is a nomial having the variable x and where a is a constant.

A *polynomial* of the form

$$f(x) = a_n x^n + a_{n-1} x^{n-1} + \cdots + a_1 x + a_0$$

is an equation that contains many (i.e., *poly*) nomials. The degree of the polynomial is a nonnegative integer n, while the coefficients of the polynomial are $a_0, a_1, \ldots, a_n$. The domain is the set of all real numbers. Polynomials are everywhere in mathematics and science. One very special type of polynomial is that of a *function*, an extremely important topic we now develop.

2.9.2 Functions

The concept of a function literally pervades all of mathematics, statistics, and applied scientific research. It is so central to virtually all areas of investigation that one has little if any chance at understanding any kind of mathematics related to science, including statistical modeling techniques, if one does not first understand the nature of a function. Indeed, as Labarre (1961) noted:

> The first man to introduce the word *function* in mathematics seems to have been Gottfried Leibniz, in about 1694. Since Leibniz's time, the notion of a function has undergone many refinements, but it has survived these and, without undue exaggeration, we say that it can lay serious claim to being the most important single concept in mathematics. (p. 73)

For some examples of functional statements, consider the following research questions:

- Is Alzheimer's disease a function of age?
- Is mental illness a function of stress?
- Is intelligence a function of genetics?
- Are school shootings in America a function of current gun laws?

Each of the above research questions imparts a *function statement*. However, as we will see when we review the precise definition of a function, the above statements are far from *perfect* functional forms. Why? Because they are messy English, and not precise mathematics. We all know what it means to ask the question of whether school shootings are a function of current gun laws. Intuitively, it implies that if indeed functional, if we changed the gun laws, it would have an *influence* or *effect* on school shootings.[9] Likewise, if you argue that mental illness is a function of stress, then you are presumably suggesting that stress in some way "contributes" to the prevalence of mental illness. Again, these concepts are very vague and imprecise, yet the word "function" is used in all of them. Functions in mathematics proper have been defined in a much more precise and rigorous fashion.

2.9.3 What is a Mathematical Function?

Mathematically, a function can be defined as *a correspondence between two sets of elements such that to each element in the first set, there corresponds one and only one element in the second set*. The first set is called the *domain* of the function, and the second set (which corresponds to elements in the first set) is called the *range* or the *codomain* of the function.

Perhaps the easiest function is that of $y = x$, given in Figure 2.14, which is a linear function that passes through the origin (0, 0) with slope equal to 1. In this function, y is known as the *dependent variable* and x is known as the *independent variable*.

It is easy to recognize the line in Figure 2.14 as a function since for each value of x that we can choose, there is one *and only one* value of y. In a very big way, the job of statistical modeling is to fit functional forms to data that do not follow functional forms *perfectly*, but do so in a manner close enough that we can nevertheless

[9] As we shall discuss however, functions do not necessarily imply a cause-and-effect relation.

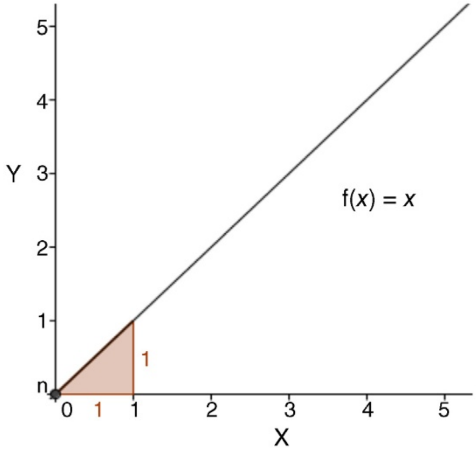


FIGURE 2.14 Linear function between dependent variable Y and independent variable X (slope = 1).

adequately account for the data using the function. Virtually all of statistical modeling constitutes the imposing of a functional form on imperfect, messy data. Regression, analysis of variance, structural equation modeling, as we will see, are ultimately all examples of this process. Quantifying how much the functional form *does not fit* is usually of interest to us just as much as we are interested in how much data it does account for.

The linear function is but one of an infinite number of possible functional forms. Other functional forms include quadratic functions, $f(x) = ax^2 + bx + c\ (a \neq 0)$, cubic functions, $f(x) = x^3$, square root functions, $f(x) = \sqrt{x}$, and absolute value functions, $f(x) = |x|$. The absolute value function is more precisely denoted by

$$|x| = \begin{cases} -x, x < 0 \\ x, x \geq 0 \end{cases}$$

where the above reads that x takes on a negative sign (i.e., $-x$) when $x < 0$ and a positive sign (i.e., x) when $x \geq 0$. However, because of the absolute value sign $|x|$ it means that we will simply take the *magnitude* of the number without regard to sign. For example, −3, in absolute value, is simply equal to 3. A graph of the absolute value function appears in Figure 2.15.

Previewing what is to come, the absolute value function is an example of a function that is *continuous* at all points, yet as we will see, the derivative cannot be defined at every point (i.e., the derivative does not exist at $f(0) = |0|$). We discuss these matters shortly.

2.9.4 Spotting Functions Graphically: The Vertical-Line Test

Having defined what is a function, it would be helpful to be able to identify functions rather easily. That is exactly what the *vertical-line test* is for. The vertical-line test for a

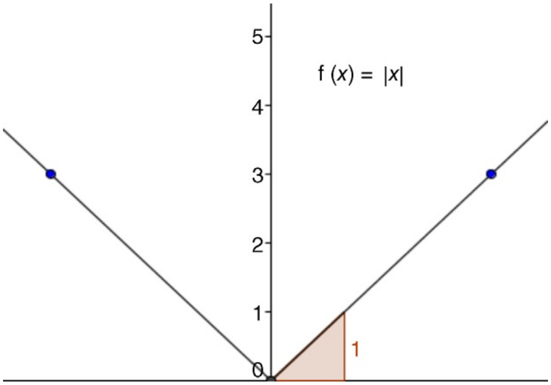


FIGURE 2.15 Absolute value function.

function is a test one can use to verify or confirm that a line or curve constitutes a function. It works as follows: If you are able to draw a vertical line through the line or curve (i.e., the polynomial form) without it passing through more than a *single* point on the line or curve, then it is a function. In other words, the vertical-line test assures us that for a given value of *x*, there is associated one and *only one* value for *y*. If, however, you are able to draw a vertical line through the line or curve and it passes through more than a single point, then it is not a function. It is still considered to be a mathematical relation, but it is does not earn the title of function. As an example, consider the linear function once more in Figure 2.16.

It is easy to see that if we drew a vertical line on the graph, the line will pass through a maximum of a single point, as shown in (b) graph.

An example of a relation that is not functional is that of a circle, $r^2 = x^2 + y^2$ (Figure 2.17).

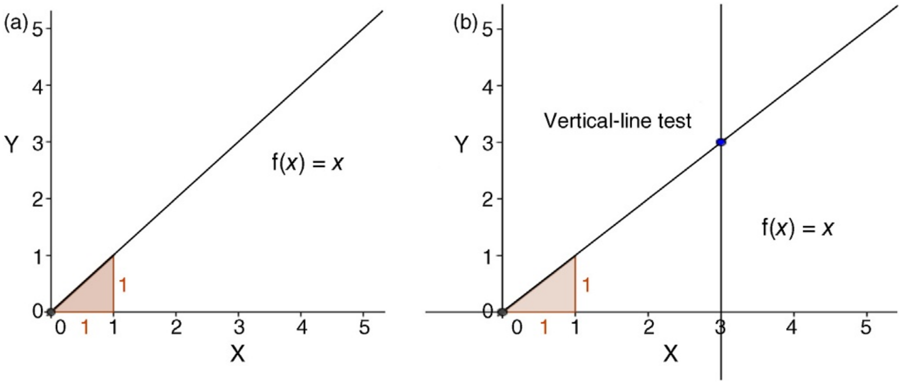


FIGURE 2.16 Vertical-line test confirming that for each value of X, there is only a single value of Y.

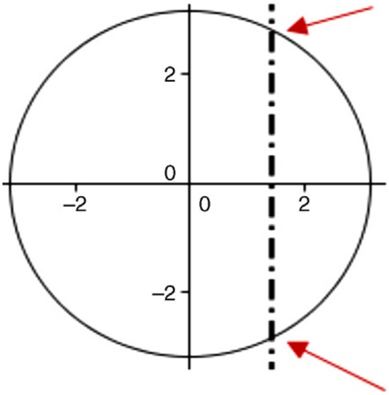


FIGURE 2.17 Vertical-line test for circle. Each arrow represents a point at which the vertical line is touching the curve. Since it touches at more than a single point, it fails the vertical-line test. The circle is hence not a function.

Since we are able to draw a vertical line through the relation and it crosses at more than a single point, the subset of points defining a circle is not a relation that is functional. Again, it is still a mathematical relation since the set of points defining the circle is still a subset of the Cartesian product, but it is not a *functional* relation.

The motivation for probabilistic statistical analysis is the fact that very few phenomena follow pure functional rules. More often, the best we can do is approximate these data using functional forms. Francis Galton's linear regression (Figure 2.9) is a classic example of this. Though the correspondence between heights of parents and their offspring is far from functional, a linear function nonetheless provided him a useful summary or "model" of the messy data. More generally, one can say the goal of statistical modeling is to impose rationally derived structures (e.g., lines, curves, etc.) on imperfect empirical observations. Much of the rest is in the details of the particular model used.

2.9.5 Limits

The idea of a *limit* in calculus and mathematical analysis has a very deep and rich history. The "discovery" of limits is usually associated with the invention of the calculus, however informal conceptions of limits date way back to the Greeks. Very few concepts in history arrive on the scientific scene without any "marinating" and development over time by noteworthy thinkers. The concept of a limit is probably one of the best examples of this. Limits are also a pillar and passage way to higher mathematics. For a discussion of limits from a historical perspective and how they relate to the development of the calculus, see Boyer (1949).

When we compute the limit of a function, we are not so interested in how the function behaves *at a particular point* as much as we are on how the function behaves

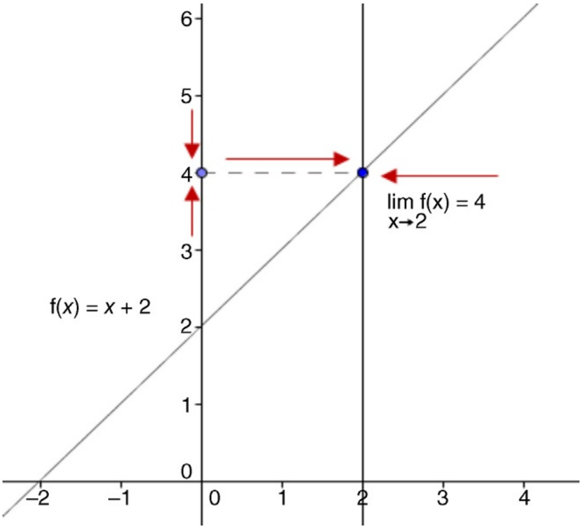


FIGURE 2.18 Depiction of a limit in a simple linear function.

near a particular point on the curve. A simple example is all we need to illustrate the concept of a limit. Consider the linear function:

$$f(x) = x + 2$$

What questions might we ask of this function? Well, we might ask how the function behaves at different values of x. For instance, we might ask what the value of the function is at $x = 1$. The answer is, of course, $f(x) = 1 + 2 = 3$. We now ask a similar question, but one that is yet different enough from the one we just asked as to call for a new historically ground-breaking concept. Consider the following question:

> What is the behavior of the function when x is close to 2?

This question does not ask us for the value of the function *at* $x = 2$, but instead requests the value of the function *near* $x = 2$. To help us answer this question, consider the simple linear function in Figure 2.18.

We can see that as x gets closer and closer to 2, the function f gets closer and closer to 4. Notice that this is true regardless of which side we approach $x = 2$, from the left- or right-hand side (this will be important in a moment). The way we express this idea of the value of the function as x *nears* 2 is to say the *limit* of the function as x approaches 2 is 4. More formally, we write

$$\lim_{x \to 2} f(x) = 4$$

which reads, *as x approaches 2, the limit of the function is equal to 4.*

Notice as well that the value of the function at $x=2$ is also equal to 4, since

$$f(2)=2+2=4$$

The fact that the limit of the function as $x \rightarrow 2$ and the value of the function at $x=2$ are the same implies that there is no break in the graph. As we will see shortly, we have just described the idea of *continuity at a point* for the given function.

It is very important to note that even if the function were not defined at $x=2$, we may have still been interested in the limit of the function as $x \rightarrow 2$. Remember, when speaking of and evaluating limits, we are not concerned with what the function does *at* a particular point, but rather are interested in what the function does *near* a particular point. The value of the function at $x=2$ is really of no interest to us when we are concerning ourselves with limits.

2.9.6 Why Limits? How Are Limits Useful?

As with many mathematical ideas, at first glance, their actual pragmatic use may appear arcane and perplexing. Limits are a prime example of this. For a full understanding of how limits are used and how they are employed in a wide variety of applications, one must study differential and integral calculus. We will very briefly review these techniques shortly, but a simple example for now will suffice to demonstrate the idea of *how* and *why* limits are useful.

The Greeks used to love computing areas. They could calculate areas of squares and rectangles quite easily. However, computing areas within other shapes that were not so "ordinary" caused them great difficulty. For instance, one question they asked was how to compute the area in a shape that did not follow "ordinary" forms, such as that given in Figure 2.19 (we are using the bell-shaped curve for obvious reasons, but could have chosen from a wide variety of curves).

Computing areas inside of shapes such as these baffled the Greeks. Their approach to the problem was essentially to "divide and conquer" and they would compute several areas of smaller shapes within the larger shapes as depicted in Figure 2.20.

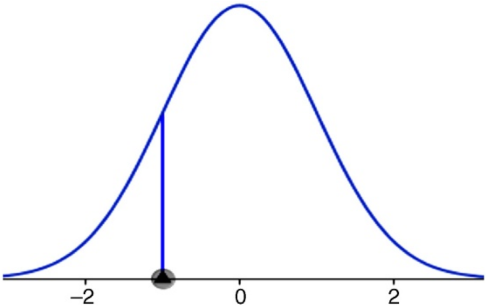


FIGURE 2.19 Area under a curve.

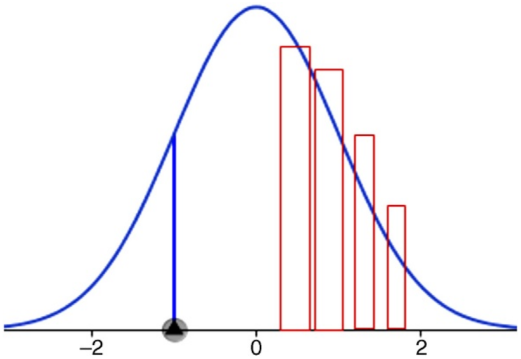


FIGURE 2.20 The Greeks, in trying to capture the area under curves, computed areas of simpler objects such as rectangles to approximate the given area.

By making the rectangles smaller and smaller and computing the area of each, and then *summing* up these rectangles, an approximation to the area under the given curve was obtained. Their method, known as the method of *exhaustion*, was imperfect, because it still seemed impossible to capture the entire area under the curve by approximation. One would have to approximate "infinitely" many rectangles in order to get the true area. This is where the idea of the limit comes in. We say that as the number of rectangles grows successively larger and larger, *in the limit*, the sum of the areas of these rectangles will be equal to the area under the curve. This, in part, was the genius of the calculus "invented" by Newton (1642–1726) and Leibnitz (1646–1716), with much help, of course, from the "giants" on whose shoulders they stood in mid-seventeenth century. They provided a coherent computational system for computing such sums. These sums were later defined as *Riemann sums* (named after the mathematician Bernhard Riemann (1826–1866)), and helped form the theory of *Riemann integration*. We will review integration shortly. The important point for now is to appreciate how the limit concept is employed in defining areas under curves. The application to statistics is obvious—if we are able to compute areas beneath curves, we are able to compute corresponding areas under probability distributions such as the normal curve (to be discussed in Chapter 3).

2.9.7 Asymptotes

Limits are helpful in appreciating a very important property exhibited by some curves, that of asymptotes. To understand what is an asymptote, consider the function (Figure 2.21)

$$f(x) = \frac{1}{x-1}$$

We can see that as x approaches 1 from the right-hand side, the values for the function $f(x)$ become larger and larger and do not seem to "settle" on any particular

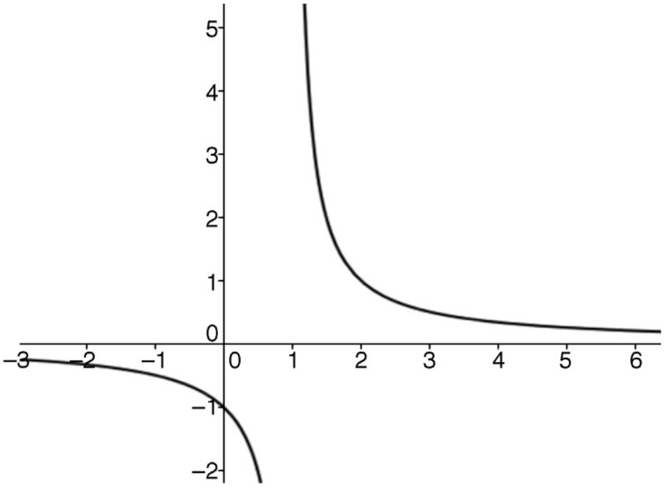


FIGURE 2.21 Graph of function 1/(x − 1).

value. In fact, it seems as though the values of the function go to infinity, so we write the right-hand limit as

$$\lim_{x \to 1^+} = \frac{1}{x-1} = \infty$$

Look at the left-hand limit now. As x approaches 1 from the left-hand side, the values of the function $f(x)$ likewise do not seem to "settle" on any given value of the function, but rather seem to go on to negative infinity, $-\infty$. Hence, we write the limit as

$$\lim_{x \to 1^-} = \frac{1}{x-1} = -\infty$$

However, ∞ is not a real number, and so in each case above, the left-hand and right-hand limits *do not exist*. That we have computed the respective limits as ∞ and $-\infty$ is only to communicate the way in which the limit does not exist. These are called *infinite limits*. The vertical line that extends to positive and negative infinity, in this case (i.e., $x = 1$), is called a *vertical asymptote*. Note that the area between the curve and $x = 1$ tends to 0, about which $x = 1$ is the center of this limiting effect. In statistics, when the asymptotic behavior of an estimator is discussed, it typically refers to the behavior of the estimator as sample size grows without bound, with the properties of the estimator evaluated in the limit as n approaches infinity (i.e., $n \to \infty$).

2.9.8 Continuity

The idea of continuity is intuitive. If I draw a line on a piece of paper without lifting my pencil, the line is a continuous one, it contains no breaks, no *discontinuities*

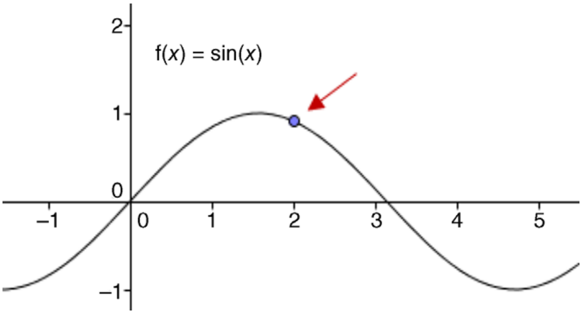


FIGURE 2.22 A point on sin(*x*) function.

(at least not at the macroscopic level). Mathematically, we must be more precise and exact in our definition. When we speak of continuity, we speak of the continuity of a *function*, whether that function be linear, quadratic, cubic, or other. As well, when speaking of continuity, we are usually interested in knowing whether a function is continuous at a particular point on that function. For example, consider the sin(*x*) function (Figure 2.22).

The function surely looks continuous at all points on it, as there does not appear to be any "breaks" in the line. However, in speaking of continuity, visual inspection is not enough. Three conditions need to be satisfied for a function f to be deemed continuous at a point $x=c$, for instance, $x=2$ on the curve:

1. $\lim\limits_{x \to c} f(x)$ must exist (i.e., the limit of the function must exist)
2. $f(c)$ must exist (i.e., the function must be defined at the given point)
3. $\lim\limits_{x \to c} f(x) = f(c)$ (i.e., the limit of the function must equal the function defined at the given point)

If these three conditions are satisfied, we say the function is *continuous* at $x = c$. We can also speak of continuity on an open interval (a, b). What justifies a function being continuous on an open interval is whether the function is deemed continuous at each point on the interval.

Having defined continuity so precisely, we are now in a position to define *discontinuity*. The definition of discontinuity is quite easy, because it is essentially the *complement* of continuity. If any of conditions 1, 2, or 3 do not hold, then we say the function f is *discontinuous* at $x = c$. An example of a function that is discontinuous is

$$g(x) = \frac{x^2 - 4}{x - 2}$$

The graph of the function is depicted in Figure 2.23.

Though the graph at first glance may appear continuous for all values of x, upon closer inspection and zooming in, we notice that at $x = 2$ there is a break. There is a

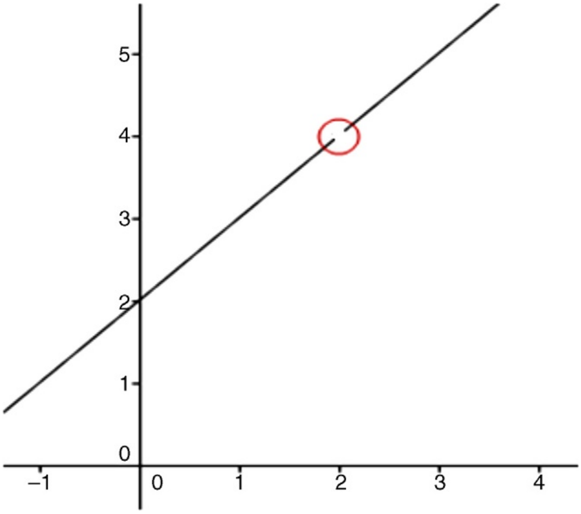


FIGURE 2.23 A discontinuity in an otherwise continuous function.

break at $x=2$ because the function is not defined at that point, and therefore the function does not exist at that point:

$$g(x) = \frac{x^2-4}{x-2}$$

$$g(2) = \frac{2^2-4}{2-2} = \frac{0}{0}$$

When a function results in $0/0$, it is said to be of *indeterminate* form (unless simplified to $x + 2$). Hence, we would say that the function is continuous at all points in the set of real numbers $\mathbb{R}$ except for at $x=2$. Or, more formally, we may write $(-\infty, 2) \cup (2, \infty)$ to indicate that the function is continuous on the real line except for $x=2$, which is why we use round brackets instead of square (which would indicate that 2 *is* included).

2.9.9 Why Does Continuity Matter? Leaping from Rationalism to Empiricism

For the applied social scientist, our discussion of continuity may at first appear quite useless. After all, the phenomena of the social scientist can rarely, if ever, be depicted so "neatly" by such perfect curves and functions. However, even if "real" objects of nature, as opposed to mathematical ones of our minds, are far from continuous, the abstractions we use to model them, in the name of *mathematical functions and distributions*, often are. For instance, when we fit a linear regression line to a plot of data, the line we are fitting is continuous at all points. It is a beautiful, idealized, perfect line with no breaks. Hence, for starters, having an understanding of what continuity means from a purely mathematical point of view is important if for no other reason than to understand some of the properties of the functions we do fit to our empirical data in the process of statistical modeling.

However, an even more important reason for having an understanding of continuity is that in applied statistical analysis, the first question you will need to consider about your variables is whether or not they are measured, or at the minimum can be considered *measurable*, on a continuous scale. For instance, if a medical scientist measured heart rate down to say, two decimal places, where theoretically any heart rate value can exist on the scale, research-wise, we would deem this variable measurable on a continuous scale. This is true even if theoretically we are “chopping” the variable off at two decimal places for each measurement. The important point for the purpose of statistical analysis is whether or not the variable in question can essentially assume a theoretically infinite number of points along the real line. If it can, then it is usually deemed continuous. Now, if we instead measured heart rate more qualitatively using labels “low,” “medium,” or “high,” then the variable would not be considered measurable on a continuous scale. Such a way of measuring heart rate would call for an alternative statistical model, one that does not assume continuity of variables. Such variables would be considered *discontinuous* or *discrete*.

To recap, the point of this discussion of continuity from the perspective of pure mathematics is simply to understand that in the world of rational abstraction (i.e., the mathematical/logical world), continuity really does *exist*. In the world of applied research, achieving the same degree of continuity in our variables is practically impossible, though often times, we will nonetheless use statistical models that assume continuity, such as a normal distribution or linear regression, on data that is less than continuous.

2.9.10 Differential and Integral Calculus

So far, we have reviewed ideas of sets, functions, limits, and continuity. These are the very pillars of modern mathematics. Why is this so? One can argue that they are interesting topics in and of themselves, but the more “practical” purpose of these fundamental building blocks is that they help lay the foundation for *calculus*, which according to most historians of science is *the* crowning intellectual achievement of modern times. Calculus evolved through centuries, as many mathematical and scientific ideas do, but calculus itself is usually associated with Newton and Leibnitz who essentially consolidated prior ideas and made it the reigning champion of science.

There are two branches of calculus, though they are intimately related. The first branch is known as *differential* calculus, while the second branch is known as *integral* calculus. Differential calculus, generally considered, is concerned with such problems as finding tangents to curves at given points along the curve. For instance, consider the graph of the logarithmic function in Figure 2.24.

Differential calculus is concerned with questions of the sort: *What is the slope, or rate of change, of the curve at a given point along the x-axis?* For example, assume this point is $x = 2$, as circled on the graph in Figure 2.25.

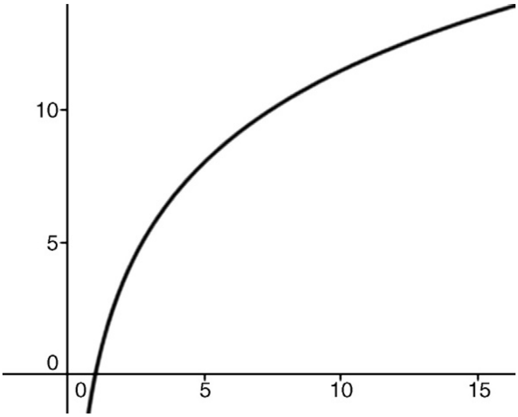


FIGURE 2.24 A logarithmic function.

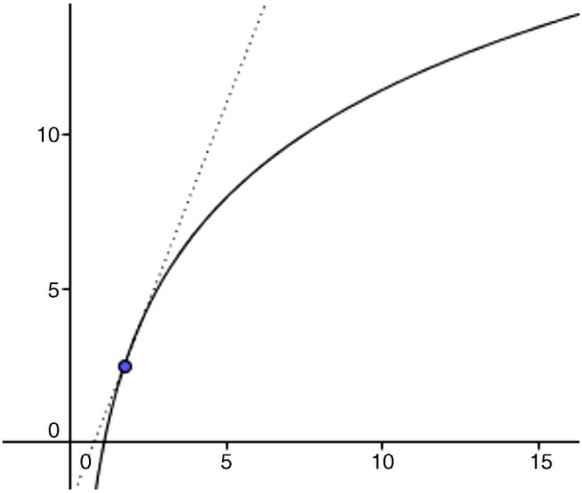


FIGURE 2.25 The line tangent to the point $x = 2$ on a logarithmic curve.

How are we to know the slope *at this particular point*? To know what the slope is at this point, we draw a line *tangent* to the curve at that point. Though somewhat difficult to visualize on the graph, the tangent we drew is touching the curve at *exactly one precise point*, that of $x = 2$. So, our original question of the slope at the given point of $x = 2$ boils down to computing the slope of the tangent to the curve at $x = 2$. This slope of the tangent, once computed, we will call the *derivative* of the function at $x = 2$. Notationally, for a given function $y = f(x)$, the derivative of the function is expressed as $y' = f'(x)$. Differential calculus, fascinating as it is, is very much nothing

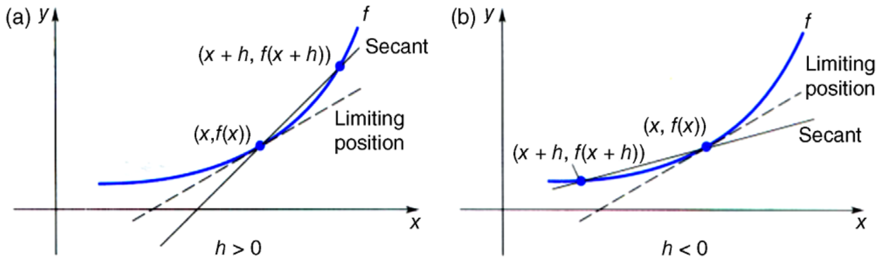


FIGURE 2.26 Defining the derivative as the slope of the tangent (Salas, Hille, and Etgen, 1999). Reproduced with permission from John Wiley & Sons, Inc.

more than a set of computational techniques and methods for finding tangents to curves in a variety of contexts. In other words, it is a set of techniques for finding derivatives.

2.9.11 The Derivative as a Limit

We informally defined the derivative as the slope of the tangent for a given point along the x-axis. Defining what this actually means mathematically is what we must do next. We need a definition of the derivative. To arrive at a definition, we need the idea of limits. The discussion is best motivated by a graphical visualization as depicted in Figure 2.26.

Note that h can be greater (a) or less (b) than 0. What happens as h gets closer and closer to 0? That is, what happens as h gets *nearer and nearer* to 0? Recall that this process of closeness is called *limit*. The limiting position is indicated by the dashed line in Figure 2.26. Hence, we are interested in knowing what happens to the difference or "change" $f(x+h)-f(x)$ relative to the change in h as that change *goes to zero*. We call this rate of change, the *derivative*, and define it formally for function $y=f(x)$ as

$$f'(x) = \lim_{h \to 0} \frac{f(x+h)-f(x)}{h}$$

assuming the given limit actually exists. If by chance the limit does not exist, then the derivative cannot be computed. Note carefully what the derivative actually is. It is simply the rate of change of one difference (i.e., the difference along the ordinate) relative to the rate of change of another difference (i.e., the difference along the abscissa). A more physical interpretation is that it is the *instantaneous* rate of change of the function relative to x.

As already mentioned, sometimes the derivative of a function for a particular range of values does not exist. Such functions are those for which it is impossible to draw tangents at those given points. For instance, one such function is the absolute value function, $f(x) = |x|$, already discussed and depicted again in Figure 2.27.

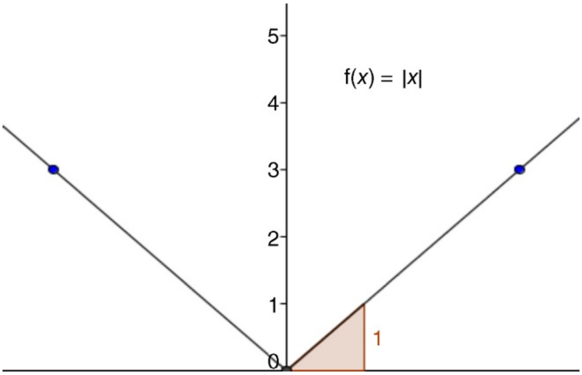


FIGURE 2.27 Absolute value function.

For this function, it becomes impossible to compute a derivative at $x = 0$ because of the sharp turn taken by the curve at that point. Even though the function is continuous at $x = 0$, the function is not differentiable at $x = 0$. Hence, given that not all functions are differentiable at all points, it becomes convenient to be able to specify ranges of intervals for which a function is differentiable. In general, we can say that if the derivative $f'(x)$ exists for each x in the open interval (a, b), then the function f is considered to be *differentiable* over that interval, (a, b).

2.9.12 Derivative of a Linear Function

To better understand just what is a derivative, it is helpful to consider the absolute easiest case. We have said that essentially, interpreted geometrically, the derivative is the slope of a tangent. If this is true, then what might be the derivative of a linear function? If you are understanding the nature of a derivative, then the answer should immediately come to mind. The derivative of a linear function is equal simply to the *slope of the line*. For instance, consider the linear function having slope equal to 5 (i.e., "b") and intercept equal to 3 (i.e., "a"):

$$\begin{aligned} f(x) &= bx + a \\ y &= 5x + 3 \end{aligned}$$

According to the power rule of differentiation,[10] the derivative of this function is equal to $f'(x) = 5$. To demonstrate this algebraically, we start with the original

[10] The power rule is just one of several differentiation rules used for the computation of derivatives. For a function $f(x) = x^n$, the derivative is equal to $f'(x) = n(x^{n-1})$. For a thorough treatment of calculus, see Salas, Hille, and Etgen (1999).

definition of a derivative and simply substitute into that expression our linear function. More generally, we can write:

$$\begin{aligned}\lim_{h\to 0}\frac{f(x+h)-f(x)}{h} &= \frac{[b(x+h)+a]-[bx+a]}{h}\\ &= \frac{bx+bh+a-bx-a}{h}\\ &= \frac{bh}{h}\\ &= b\end{aligned}$$

We see that the derivative of a linear function is simply the slope b.

Oftentimes, we wish to take derivatives of functions while holding certain variables *constant*. For example, suppose we wish to compute the derivative $f(x)$ while holding a second variable z constant. For this we compute what are called *partial derivatives*. As an example, consider the following function:

$$f(x) = 2x^2 + 5z$$

If we wished to compute the derivative of y relative to x while simultaneously holding z constant, we would compute

$$f'(x) = 4x$$

Quite simply, all we did was apply the power rule to $2x^2$. In our discussion of linear regression in Chapter 8, partial differentiation is used in obtaining the least-squares normal equations.

2.9.13 Using Derivatives: Finding Minima and Maxima of Functions

Among the most popular uses of differentiation in applied mathematics is that of learning of the behavior of curves. Just as a psychologist might be interested in learning about the behavior of an individual, the mathematician is interested in learning of the behavior of functions. Consider once more the *sin* function, sin (x) (Figure 2.28).

We may ask many questions about this curve, but one question of high importance in optimization problems is to locate areas on the curve where the function (i.e., the set of y values) attains a *maximum* or a *minimum* value. For instance, between the values of $x = 0$ and $x = 2$, at what value of x does the function achieve a maximum? By "maximum" we mean the greatest value within that open interval. We can see by inspection that it is probably slightly greater than the value $x = 1.5$, but to know for

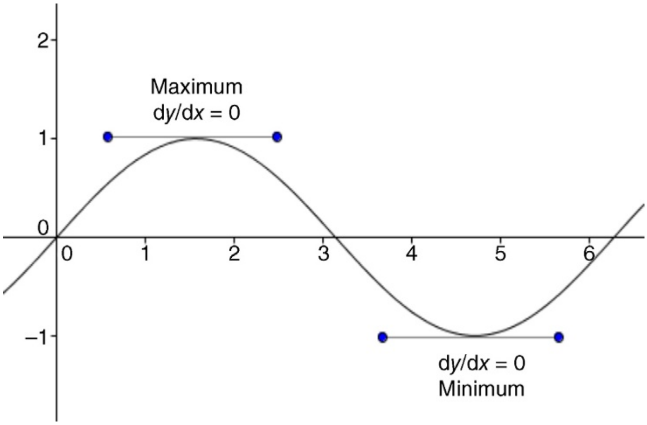


FIGURE 2.28 Graph of sin(x).

certain, and here is where the utility of calculus comes in, we can use the fact that the highest point of the curve must be the location where the derivative (denoted by dy/dx in Figure 2.28) of the function in the interval $(0, 2)$ is equal to 0. This fact is implied by a result known as *Rolle's theorem* (see Bartle and Sherbert, 2011). That is, graphically, the maximum in the interval $(0, 2)$ occurs where the slope of the function is equal to 0, as indicated by the horizontal line.

Hence, when we set the derivative of the function equal to 0, we are able to solve for the *local maximum* (i.e., by "local" we mean in a particular *neighborhood* of points). Likewise, we can also learn where a function attains a *local minimum* through a similar strategy, that of setting the derivative of the function equal to zero. This is one very common use of derivatives in science, that of finding maxima and minima of a variety of functions. Optimization algorithms often feature the search for such extrema. In structural equation modeling, for instance, *Lagrangian multipliers* are regularly used in maximizing functions subject to specific constraints. For a brief discussion, see Mulaik (2009, pp. 60–61). For a general discussion of optimization methods, refer to Meerschaert (2007, Chapter 3).

2.9.14 The Integral

If we let f be a continuous function on a closed interval $[a, b]$, then the definite integral of f from a to b is defined as

$$\int_a^b f(x)\mathrm{d}x$$

The integrand is $f(x)$, the lower limit of integration is a, and the upper limit of integration is b. The definite integral represents the cumulative sum (the limit of Riemann sums) of the signed areas between the graph of f and the x-axis from $x = a$ to

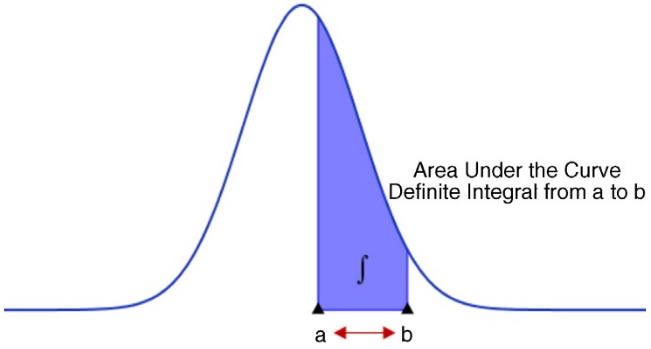


FIGURE 2.29 Area from a to b under the curve in normal distribution.

$x = b$, where the areas above the x-axis are counted positively and the areas below the x-axis are counted negatively (see Figure 2.29, where only positive areas are shown).

There is sometimes a misunderstanding that area can only be computationally positive, but the definition of the definite integral tells us different. In most statistics texts, the areas under the *standard normal distribution* are given in appendices. These areas were computed by integrating under the normal curve. Analogously, areas under the curves of other distributions such as t or F can be computed by integration. These areas under curves are then used to make statements about probability. Hence, even if one does not use integration in applied statistical work and research, it is important nonetheless to recognize where integration plays a role in determining probabilities in normal and other densities.

2.9.15 Calculus in R

Computing derivatives and integrals manually can at best be impractical and time consuming, and at worst, utterly exhausting. In this day of computing power, mental energy spent on computing derivatives is almost akin to doing long division before calculators became widely available. This is not to say that working through differentiation problems by hand is not a great exercise to help master technical skill, but being a "human computer" is a skill less and less required now that we have computing machines that do it far better than us with far better accuracy! Fortunately, we can compute derivatives in R. For example, suppose we wanted to compute the derivative of the polynomial function $f(x) = 5x^4$. Through using the power rule, we know the derivative is equal to $f'(x) = 20x^3$. To perform this computation in R, we compute

```
> D(expression(5*x^4), "x")
5 * (4 * x^3)
```

That is, the derivative of $f(x) = 5x^4$ is $f'(x) = 5(4)x^3 = 20x^3$. Integrals can also be computed in R and other software. See Crawley (2013, p. 339–340) for an example of computing area under the curve and also for an example of computing a *differential equation*, which is an equation that contains one or more derivatives in it. These equations are quite common in areas such as biology and physics. Refer to Kline (1977, pp. 848–865) for a readable introduction.

2.9.16 Vectors and Matrices

Ordinary algebra is done on scalars, which are essentially "ordinary" numbers.[11] For instance, in the linear equation $y = bx + a$, both the "inputs" and "outputs" of the function are scalars. Using scalars works well in one-dimensional data analysis, that is, where $\mathbb{R}$ is understood to be raised to the first power, $\mathbb{R}^1$. We denote higher dimensions and additional axes by raising $\mathbb{R}$ to higher powers. For instance, for pairs of observations (x, y) we are in the realm of $\mathbb{R}^2$. The set of ordered triples is denoted by $\mathbb{R}^3$. Theoretically, there is no limit to the number of dimensions in which we can work. In complex multivariate analysis for instance, we could be working in $\mathbb{R}^7$. And theoretically, a mathematician can work in as many dimensions as he or she chooses. These dimensions or "spaces" are usually referred to as *Euclidean spaces*. The number of dimensions a scientist works in, of course, will typically be dictated by the empirical context in which he or she is working.

Most textbooks on multivariate analysis include either individual chapters or appendices on essential matrix theory. There are also many books that feature the study of matrix theory with applications to statistics. In addition to the brief introduction and overview we provide in Appendix A, some of the better sources for matrix algebra include Searle (1982) and Harville (1997). For a complete and very well-written introductory text on linear algebra and matrix operations the reader should refer to Anton and Rorres (2000). Strang (1993) is also a good reference.

2.9.17 Why Vectors and Matrices?

When we work in higher dimensions such as $\mathbb{R}^2$ and $\mathbb{R}^3$, scalar algebra will not suffice, and we require a new notation to deal with these higher dimensions. Geometrically, a vector is simply a directed line segment on a Cartesian coordinate system. The end or tip of the vector denotes the joint coordinates for the given observation. For instance, consider the following vector $\mathbf{v}$ on variables x, y, respectively:

$$\mathbf{v} = \begin{bmatrix} 1 \\ 3 \end{bmatrix}$$

The vector $\mathbf{v}$ represents the point (1, 3), and can be visualized in Figure 2.30. The vector denotes the "position" on these variables in the $\mathbb{R}^2$ plane. We adopt the convention to use small bold type to denote vectors. Collections of vectors that

[11] A more precise and useful definition of a scalar in mathematics, especially as it relates to multivariate analysis, is that it is a quantity that multiplies vectors in a vector space.

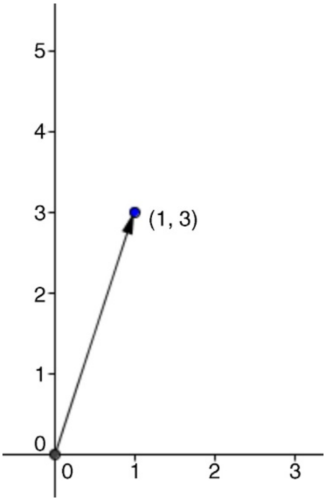


FIGURE 2.30 Vector representation of point (1, 3).

are closed under addition and multiplication are generally known as *vector spaces* in linear algebra. A subset of vectors in that space is referred to as a vector *subspace*.

Let us now add another dimension z and include the point 5:

$$\mathbf{v} = \begin{bmatrix} 1 \\ 3 \\ 5 \end{bmatrix}$$

Figure 2.31 denotes a visualization of this three-dimensional space.

As an applied example, suppose we wanted to represent data on IQ (intelligence) and GPA (i.e., grade point average). We could then extend the vector to be

$$\mathbf{V} = \begin{bmatrix} 100 & 3.8 \\ 90 & 3.6 \\ 85 & 3.4 \end{bmatrix}$$

where 100, 90, 85 are IQ scores, and 3.8, 3.6, 3.4 are GPA scores.

Notice that we have changed the name of the "object" from **v** to **V** to denote what we call a *matrix*. A matrix is simply many vectors in an array. In general, we may denote a matrix **A** as

$$\mathbf{A} = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1p} \\ a_{21} & a_{22} & \cdots & a_{2p} \\ \vdots & & & \\ a_{m1} & & & \end{bmatrix}$$

having m rows and p columns.

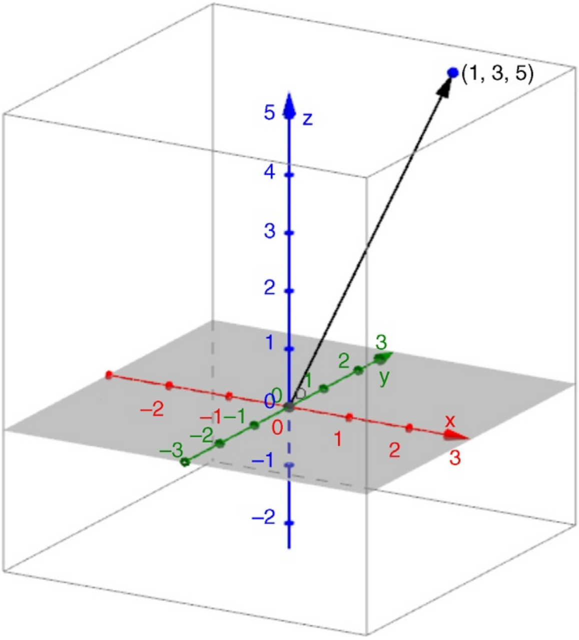


FIGURE 2.31 Point (1, 3, 5) in three-dimensional space.

As in scalar algebra, we are able to add, subtract, and multiply vectors and matrices. Understanding a bit of how these operations work on matrices is important because it helps demystify somewhat what occurs in multivariate techniques such as multiple regression and principal components analysis (Chapter 14). We revisit these multidimensional ideas when we tackle those topics later on in the book. Again, for an immediate overview of these topics, see Appendix A.

A vector may be elongated or shortened by multiplying that vector by a constant. For instance, consider once more the vector **v**:

$$\mathbf{v} = \begin{bmatrix} 1 \\ 3 \end{bmatrix}$$

Suppose we wanted to maintain the direction of the vector, but make the vector longer by a factor of 2. To do this, we multiply the elements of the vector by the scalar 2:

$$2\mathbf{v} = \begin{bmatrix} (2)1 \\ (2)3 \end{bmatrix} = \begin{bmatrix} 2 \\ 6 \end{bmatrix}$$

which graphically, is depicted in Figure 2.32.

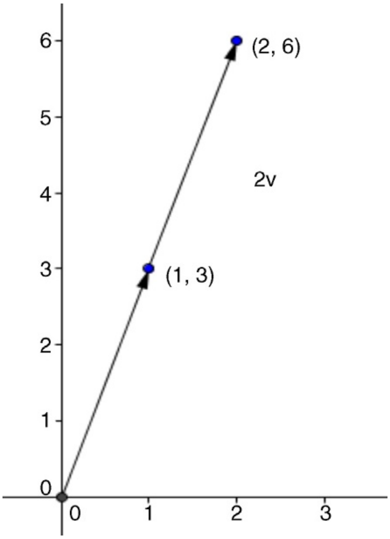


FIGURE 2.32 Multiplication of a vector by a scalar of 2.

In general then, to influence the length of a vector **v** by a scalar λ, we multiply that vector *element by element* by the scalar:

$$\lambda\mathbf{v} = \begin{bmatrix} \lambda(v_1) \\ \lambda(v_2) \\ \lambda(v_3) \end{bmatrix}$$

For two vectors or matrices to be added or subtracted, they must be of the same *dimension*. That is, they must have the same number of rows and columns. For example, consider the vectors **u** and **v**:

$$\mathbf{u} = \begin{bmatrix} u_1 \\ u_2 \\ u_3 \end{bmatrix}, \ \mathbf{v} = \begin{bmatrix} v_1 \\ v_2 \\ v_3 \end{bmatrix}$$

Notice that both **u** and **v** are of the same dimension (3 rows, 1 column). To add these two vectors, we add element by element:

$$\mathbf{u} + \mathbf{v} = \begin{bmatrix} u_1 + v_1 \\ u_2 + v_2 \\ u_3 + v_3 \end{bmatrix}$$

Likewise, to add matrices **A** and **B**, we add element by element:

$$\mathbf{A} = \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix} \quad \mathbf{B} = \begin{bmatrix} b_{11} & b_{12} & b_{13} \\ b_{21} & b_{22} & b_{23} \\ b_{31} & b_{32} & b_{33} \end{bmatrix}$$

$$\mathbf{A} + \mathbf{B} = \begin{bmatrix} a_{11}+b_{11} & a_{12}+b_{12} & a_{13}+b_{13} \\ a_{21}+b_{21} & a_{22}+b_{22} & a_{23}+b_{23} \\ a_{31}+b_{31} & a_{32}+b_{32} & a_{33}+b_{33} \end{bmatrix}$$

When multiplying matrices, say, **A** and **C**, the product **AC** is defined only for matrices for which the *number of columns in* **A** *is equal to the number of rows in* **C**. When the number of columns in the first matrix is equal to the number of rows in the second matrix, we say the matrices are *conformable for multiplication*.

For example, let matrices **A** and **C** be defined as

$$\mathbf{A} = \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix} \quad \mathbf{C} = \begin{bmatrix} c_{11} & c_{12} \\ c_{21} & c_{22} \\ c_{31} & c_{32} \end{bmatrix}$$

Notice that the number of columns in **A** is equal to the number of rows in **C**. That is, there are three columns in **A** and three rows in **C**. They are conformable for multiplication so long as we wish to generate the product **AC**. Notice that in this case, the product **CA** is not defined, since the number of columns in **C** (equal to 2) is not equal to the number of rows in **A** (equal to 3).

To get the product **AC**, we carry on with multiplying each element in respective rows of **A** against each element in respective columns of **C**:

$$\mathbf{AC} = \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix} \begin{bmatrix} c_{11} & c_{12} \\ c_{21} & c_{22} \\ c_{31} & c_{32} \end{bmatrix}$$

$$= \begin{bmatrix} a_{11}c_{11} + a_{12}c_{21} + a_{13}c_{31} & a_{11}c_{12} + a_{12}c_{22} + a_{13}c_{32} \\ a_{21}c_{11} + a_{22}c_{21} + a_{23}c_{31} & a_{21}c_{12} + a_{22}c_{22} + a_{23}c_{32} \\ a_{31}c_{11} + a_{32}c_{21} + a_{33}c_{31} & a_{31}c_{12} + a_{32}c_{22} + a_{33}c_{32} \end{bmatrix}$$

2.9.18 Solving Systems of Linear Equations

At a technical level, the estimation of parameters in many statistical models can be conceptualized as the solving of sets or *systems* of linear equations. Recall that a linear equation is of the form

$$bx + ay = c \tag{2.6}$$

where a and b are constants (or “coefficients” for x and y) and c is another constant. This is generally known as the *standard form* of a linear equation. The difference between equation (2.6) and the equation ordinarily given for a line, that of $bx + a = y$ is that $bx + ay = c$ has two unknowns it in (x and y), whereas $bx + a = y$ has an unknown in only a single variable x.

A set of linear equations can be given by the following:

$$\begin{aligned} a_{11}x_1 + a_{12}x_2 + \cdots + a_{1n}x_n &= b_1 \\ a_{21}x_1 + a_{22}x_2 + \cdots + a_{2n}x_n &= b_2 \\ a_{m1}x_1 + a_{m2}x_2 + \cdots + a_{mn}x_n &= b_m \end{aligned} \tag{2.7}$$

Written more compactly, the matrix equivalent of (2.7) is $\mathbf{Ax} = \mathbf{b}$. Many statistical problems, such as linear regression, boil down to solving for $\mathbf{x}$ given that $\mathbf{A}$ and $\mathbf{b}$ are known. A system of equations that has no solutions is referred to as *inconsistent*. If there is at least one solution, the system is called *consistent*. It is a fact of linear algebra that every system of linear equations has either no solution, only a single solution, or an infinite number of solutions.

If the system is not consistent, one may nevertheless seek values for $\mathbf{x}$ that are a good *approximation* to the actual solution. A good approximation is that of *least-squares*. For a least-squares problem, the vector $\mathbf{x}$ will be called the *least-squares solution*. It is typically solved by computing an inverse for $\mathbf{A}'\mathbf{A}$, such that we have

$$\mathbf{x} = (\mathbf{A}'\mathbf{A})^{-1}(\mathbf{A}'\mathbf{b})$$

Ordinary least-squares is often the estimation method of choice for analysis of variance and regression models (see Chapter 8 for details).

Solving simple systems of linear equations in R is easy. Consider the following system of two equations with unknowns in x and y:

$$\begin{aligned} 5x + 2y &= 20 \\ 8x + 6y &= 31 \end{aligned}$$

We can represent the above system in terms of matrix $\mathbf{A}$ and vector $\mathbf{v}$:

```
> A <- matrix(c(5, 8, 2, 6), nrow = 2)
> A
     [,1] [,2]
[1,]    5    2
[2,]    8    6
> v <- matrix(c(20, 31), nrow = 2)
> v
     [,1]
[1,]   20
[2,]   31
```

To find solutions for x and y that will simultaneously satisfy both equations, we use the `solve` function:

```
> solve(A, v)
           [,1]
[1,]  4.1428571
[2,] -0.3571429
```

We can easily verify that $x=4.14$ and $y=-0.36$ are solutions to the system. For the first equation, that of $5x+2y=20$, we have

```
> eq.1 <- 5*4.14 - 2*0.36
> eq.1
[1] 19.98
```

For the second equation, that of $8x+6y=31$, we have

```
> eq.2 <- 8*4.14 -6*0.36
> eq.2
[1] 30.96
```

We can see that within rounding error, our obtained solutions for x and y satisfy both equations.

2.10 CHAPTER SUMMARY AND HIGHLIGHTS

- Gaining an understanding of essential mathematics and probability is important for any work using or applying statistics to empirical problems.
- The theory of sets, though studied for its own sake, is useful to the applied researcher in understanding the distinction between a sample and a population. Basic set operations such as unions and intersections are also important to master as they are the same tools used in elementary probability theory.
- A mathematical relation is a subset of the Cartesian product. A function is a mathematical relation in which each element of the domain is paired with only one element of the range.
- Sets of numbers include natural numbers, integers, rational numbers, and irrational numbers, among others. In typical data analysis, numbers are assumed to be real numbers (i.e., rational or irrational).
- Probability is the language of uncertainty and is a deep philosophical topic. Though the mathematical theory of probability is well-defined and has been axiomatized, how to conceptualize probability is a more difficult problem. Traditional camps include the frequentist and subjectivist perspectives.
- Virtually all of scientific endeavors, in one way or another, is about estimating conditional probabilities of interest. Scientists are usually not very interested in unconditional probabilities, but are much more interested in the probability of

events given certain specific conditions. Ideas of conditional probabilities pervade virtually all statistical methods.

- Two events in probability are considered mutually exclusive or disjoint if the probability of their intersection is equal to zero.
- Two events are statistically independent if the probability of one event does not alter or change the probability of the second event. A lack of independence, by itself, is not enough to substantiate a physical directional process.
- The law of large numbers says that if an experiment is repeated many times under identical conditions, the relative frequency of the event will approach the true probability of the event as the number of trials increases. In the limit, as sample size grows without bound, the true probability of the event is theoretically attainable.
- The subjective interpretation of probability overcomes some of the logical difficulties with the law of large numbers and instead designates the probability of an event as one's belief or opinion.
- Bayes' theorem is an equation used to invert probabilities and is useful in situations such as when diagnostic information is obtained as a sign that can inform us of the probability of disease. The theorem requires the specifying of a prior probability, which is sometimes considered controversial, but in many cases can be regarded as simply the base rate of the event in question for the given population under study.
- Statistical inference is a process of drawing conclusions about parameters based on information provided by samples on which statistics are computed.
- Deciding to reject a null hypothesis should ideally incorporate, even if informally, an estimate of the costs associated with making a wrong decision.
- The vertical-line test can be used as a quick visual test to ascertain the existence of a function by its graph.
- The idea of a limit in calculus is one of great historical and scientific importance because it laid the groundwork for the fields of differential and integral calculus. Limits concern themselves with the behavior of functions near particular points on a curve rather than the behavior of functions at those points.
- One use of limits is in conceptualizing the process of integration in calculus. When we take the limit of areas under a curve, we are determining the area under that curve. The process is called Riemann integration.
- Differential calculus concerns itself with the determination of the slopes of tangents at a given point on a curve. Like integration, differential calculus is heavily based on the limit concept.
- The idea of continuity is that of drawing a line without lifting one's pencil off the page. A more formal definition of continuity has been developed by mathematicians that make it more precise and exact.
- For the research scientist, an understanding of continuity is important if for no other reason than to be able to recognize when his or her variables can be considered continuous or discrete for the purpose of statistical analysis.
- Vectors and matrices are used to display data in higher than one dimension. Geometrically, vectors are directed line segments with both magnitude and

direction. Matrices are arrays of numbers. Vectors and matrices are needed in the study of multivariate methods in which one regularly works in greater than a single dimension.

REVIEW EXERCISES

2.1. Discuss the quote of Einstein given at the opening of this chapter:

> How can it be that mathematics, being after all a product of human thought which is independent of experience, is so admirably appropriate to the objects of reality? . . . As far as the laws of mathematics refer to reality, they are not certain, and as far as they are certain, they do not refer to reality.

Specifically, why are the laws of mathematics, as they refer to reality, not certain?

2.2. Define what is meant by a set in mathematics, and give a couple practical examples of sets.

2.3. What might be the difficulty in defining the set of all beautiful clouds in the sky? Under what condition(s) could such a set exist, and under what conditions could such a set not exist?

2.4. Distinguish between a subset and a proper subset. Under what condition(s) are they equal?

2.5. Distinguish between the union of sets and the intersection of sets.

2.6. Discuss the following notation with regard to the union of sets. Describe exactly what is specified.

$$\bigcup_{n=1}^{3} A_n$$

2.7. Define what is meant by a Cartesian product, and then define what is meant by a mathematical relation on that Cartesian product.

2.8. Distinguish between the natural numbers and the integers.

2.9. Distinguish between a rational number and an irrational number.

2.10. What makes the number $\sqrt{2}$ irrational? Can you think of another irrational number?

2.11. Define what is meant by a population versus a sample in terms of sets.

2.12. Define what is meant by an event in probability theory.

2.13. Define what is meant by an experiment in probability theory.

2.14. First, discuss what is meant by an axiom in mathematics, then give the three axioms of the theory of mathematical probability.

2.15. Why is it that even if two people disagree on how to define probabilities, they will still likely (though not always) agree with the axioms of probability and the mathematical theory on which probability is based?

2.16. Compare and contrast an unconditional probability with a conditional one.

2.17. Give a research example where the unconditional probability of an event would most likely be equal to the conditional probability of that event, and specify the condition you are imposing.

2.18. Discuss how the idea of statistical independence arises from a consideration of the following conditional probability:

$$p(A/B) = \frac{p(A \cap B)}{p(B)}$$

2.19. Define what is meant by mutual exclusiveness or to say two events are disjoint.

2.20. Compare and contrast mutual exclusiveness with statistical independence. Specifically, why is it usually true that two events that are mutually exclusive cannot simultaneously be independent?

2.21. Why is it that in practice, demonstrating a violation of independence is very difficult? For example, why is it that demonstrating that hiring practices are independent or dependent of race is a very difficult position to substantiate?

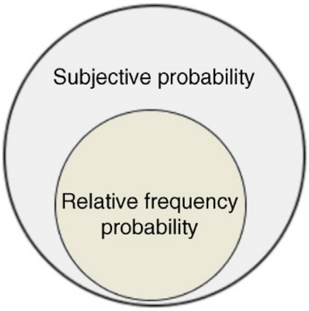
2.22. Give a physical scenario in which the events heads and tails are mutually exclusive, and then one where these same two events are not mutually exclusive. How does the context matter in which we speak of mutual exclusiveness?

2.23. Distinguish between the frequentist and subjectivist interpretations of probability, and why the mathematical theory of probability cannot resolve the debate.

2.24. What is often cited as a philosophical problem with the law of large numbers?

2.25. Give an example where it would be virtually impossible to quantify probability using a frequentist approach and how a subjective version of probability would allow for such a quantification.

2.26. Discuss what is meant by the following picture. Do you agree? Disagree? Why or why not?



2.27. Describe the overall purpose of Bayes' theorem. How is it relevant in the overall scheme of things, especially as relatable to scientific practice?

2.28. Discuss why the following statement is true:

$$p(D/H) \neq p(H/D)$$

2.29. Discuss why prior probabilities needed to effectively use Bayes' theorem can be somewhat controversial. Why might prior probabilities be difficult to estimate?

2.30. How are base rates important in the estimation of probabilities?

2.31. Outline a simple example (other than one featuring a coin) of statistical inference, clearly detailing the steps involved in the process.

2.32. Distinguish between a type I and a type II error. How might one error be potentially no more problematic to make than the other?

2.33. Discuss how ignoring the stakes involved in decision making (i.e., type I and type II error rates) inhibits one to make a truly rational and coherent decision regarding a hypothesis.

2.34. Define what is meant by a polynomial in mathematics.

2.35. Discuss the precise definition of a function in mathematics and why it can be said that functional statements are what virtually every science is about.

2.36. Give three substantive examples of function statements from an area of investigation of your interest.

2.37. Consider the equation $f(x) = x^2 + 3$. Is it that of a function? Why or why not?

2.38. Consider the equation $x = y^2$. Is it that of a function? Why or why not?

2.39. Explain how the vertical-line test is used to determine whether a given equation is or is not a function.

2.40. Define what is meant by a rational function, and give an example of one.

2.41. Define and discuss what is meant by a limit in calculus, and give an example using a linear function.

2.42. Define what is meant by continuity in mathematics. Give three conditions that must exist for a function to be deemed continuous. Then, discuss how the concept of continuity is used in applied social science research.

2.43. Discuss the nature of differential calculus. In a sentence or two, describe the goal and purpose of differential calculus.

2.44. What is a derivative in calculus, and state one popular use of derivatives in calculus and in science in general.

2.45. What is an integral in calculus, and state one popular use of integrals in calculus and in science in general.

2.46. Define what is meant by a vector in mathematics. What is a matrix? Why and how are these useful in applied statistics and methodology?

Further Discussion and Activities

2.47. The *base-rate fallacy* generally refers to individuals' failure to incorporate base rate information into probability estimates. Read Bar-Hillel, M. (1980) and discuss the nature of the problem and some of the causes of the phenomenon.

2.48. Calculus originated largely in the context of physical applications and was marveled mostly for its utility in addressing problems in areas of physics, astronomy, etc. It was only after calculus was deemed a success did mathematicians begin to investigate the logical foundations on which it was based, and to their alarm, discovered such a basis to be lacking. It was mathematicians in the 1800s who undertook the job of putting calculus on a *rigorous* footing. The new field became known as *analysis*. Concepts in this field are essentially deeper looks at the logical basis on which calculus is built. A brief overview of the origins of so-called "rigorous calculus" is given by historian of science Grabiner (1983). Read this paper and briefly discuss and summarize many of the features that distinguish calculus from *analysis*.

2.49. In the chapter we briefly discussed the difficulties in arriving at probability estimates. If a meteorologist reports that the probability of rain is equal to 0.70, the question you must ask is: *How is such an estimate computed?* Is it a rationally-derived probability? Is it enough to feel assured that it is being computed by an "expert" in the field and thus must be correct? A more interesting example is the probability of the so-called "Big One" earthquake in California. Apparently, experts predict that California has a 0.997 probability of a magnitude 6.7 earthquake or higher in the next 30 years or so. You can read the media report on the Web site of the *Southern California Earthquake Center* (http://www.scec.org/ucerf2/). The precise probability estimates are > 0.99 for a 6.7 magnitude quake, 0.94 for a 7.0, 0.46 for a 7.5, and 0.04 for a 8.0 magnitude quake. You may not be an expert in geography or seismology, but you should be a critical consumer of scientific reports such as this. What kinds of questions might you ask about these probability estimates? How do you think they were derived? Could the manner in which they were derived make them more or less accurate and/or rigorous?