Leafcutter ant

1 Statistics and samples

1.1 What is statistics?

Biologists study the properties of living things. Measuring these properties is a challenge, though, because no two individuals from the same biological population are ever exactly alike. We can't measure everyone in the population, either, so we are constrained by time and funding to limit our measurements to a *sample* of individuals drawn from the population. Sampling brings uncertainty to the project because, by chance, properties of the sample are *not* the same as the true values in the population. Thus, measurements made from a sample are affected by who happened to get sampled and who did not.

Statistics is the study of methods to describe and measure aspects of nature from samples. Crucially, statistics gives us tools to *quantify the uncertainty* of these measures—that is, statistics makes it possible to determine the likely magnitude of their departure from the truth.

Statistics is about **estimation**, the process of inferring an unknown quantity of a target population using sample data. Properly applied, the tools for estimation allow us to approximate almost everything about populations using only samples. Examples range from the average flying speed of bumblebees, to the risks of exposure to cell phones, to the variation in beak size of finches on a remote Galápagos island. We can estimate the proportion of people with a particular disease who die per year and the fraction who recover when treated.

Most importantly, we can assess differences between groups and relationships between variables. For example, we can estimate the effects of different drugs on the possibility of recovery, we can measure the association between the lengths of horns on male antelopes and their success at attracting mates, and we can determine by how much the survival of women and children during shipwrecks differs from that of men.

Estimation is the process of inferring an unknown quantity of a population using sample data.

All of these quantities describing populations—namely, averages, proportions, measures of variation, and measures of relationship—are called **parameters**. Statistical methods tell us how best to estimate these parameters using our measurements of a sample. The parameter is the truth, and the estimate (also known as the **statistic**) is an approximation of the truth, subject to error. If we were able to measure every possible member of the population, we could know the parameter without error, but this is rarely possible. Instead, we use estimates based on incomplete data to approximate this true value. With the right statistical tools, we can determine just how good our approximations are.

A *parameter* is a quantity describing a population, whereas an *estimate* or *statistic* is a related quantity calculated from a sample.

Statistics is also about **hypothesis testing**. A **statistical hypothesis** is a specific claim regarding a population parameter. Hypothesis testing uses data to evaluate evidence for or against statistical hypotheses. Examples are "The mean effect of this new drug is not different from that of its predecessor," and "Inhibition of the *Tbx4* gene changes the rate of limb development in chick embryos." Biological data usually get more interesting and informative if they can resolve competing claims about a population parameter.

Statistical methods have become essential in almost every area of biology—as indispensable as the PCR machine, calipers, binoculars, and the microscope. This book presents the ideas and methods needed to use statistics effectively, so that we can improve our understanding of nature.

Chapter 1 begins with an overview of samples—how they should be gathered and the conclusions that can be drawn from them. We also discuss the types of variables that can be measured from samples, introducing terms that will be used throughout the book.

1.2 Sampling populations

Our ability to obtain reliable measures of population characteristics—and to assess the uncertainty of these measures—depends critically on how we sample populations. It is often at this early step in an investigation that the fate of a study is sealed, for better or worse, as Example 1.2 demonstrates.

EXAMPLE 1.2 Raining cats

In an article published in the *Journal of the American Veterinary Medical Association,* Whitney and Mehlhaff (1987) presented results on the injury rates of cats that had plummeted from buildings in New York City, according to the number of floors they had fallen. Fear not: no experimental scientist tossed cats from different altitudes to obtain the data for this study. Rather, the cats had fallen (or jumped) of their own accord. The researchers were merely recording the fates of the sample of cats that ended up at a veterinary hospital for repair. The damage caused by such falls was dubbed Feline High-Rise Syndrome, or FHRS.[1]

Not surprisingly, cats that fell five floors fared worse than those dropping only two, and those falling seven or eight floors tended to suffer even more (see Figure 1.2-1). But the astonishing result was that things got better after that. On average, the number of injuries was reduced in cats that fell more than nine floors. This was true in every injury category. Their injury rates approached that of cats that had fallen only two floors! One cat fell 32 floors and walked away with only a chipped tooth.

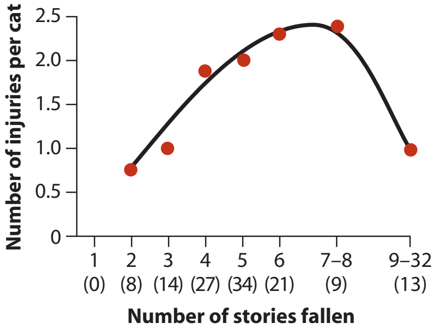


FIGURE 1.2-1
A graph plotting the average number of injuries sustained per cat according to the number of stories fallen. Numbers in parentheses indicate number of cats. Modified from Diamond (1988).

[1] "The diagnosis of high-rise syndrome is not difficult. Typically, the cat is found outdoors, several stories below, and a nearby window or patio door is open." (Ruben 2006).

This effect cannot be attributed to the ability of cats to right themselves so as to land on their feet—a cat needs less than one story to do that. The authors of the article put forth a more surprising explanation. They proposed that after a cat attains terminal velocity, which happens after it has dropped six or seven floors, the falling cat relaxes, and this change to its muscles cushions the impact when the cat finally meets the pavement.

Remarkable as these results seem, aspects of the sampling procedure raise questions. A clue to the problem is provided by the number of cats that fell a particular number of floors, indicated along the horizontal axis of Figure 1.2-1. No cats fell just one floor, and the number of cats falling increases with each floor from the second floor to the fifth. Yet, surely, every building in New York that has a fifth floor has a fourth floor, too, with open windows no less inviting. What can explain this curious trend?

To answer this, keep in mind that the data are a *sample* of cats. The study was not carried out on the whole population of cats that fell from New York buildings. Our strong suspicion is that the sample is biased. Not all fallen cats were taken to the vet, and the chance of a cat making it to the vet might have been affected by the number of stories it had fallen. Perhaps most cats that tumble out of a first- or second-floor window suffer only indignity, which is untreatable. Any cat appearing to suffer no physical damage from a fall of even a few stories may likewise skip a trip to the vet. At the other extreme, a cat fatally plunging 20 stories might also avoid a trip to the vet, heading to the nearest pet cemetery instead.

This example illustrates the kinds of questions of interpretation that arise if samples are biased. If the sample of cats delivered to the vet clinic is, as we suspect, a distorted subset of all the cats that fell, then the measures of injury rate and injury severity will also be distorted. We cannot say whether this bias is enough to cause the surprising downturn in injuries at a high number of stories fallen. At the very least, though, we can say that, if the chances of a cat making it to the vet depends on the number of stories fallen, the relationship between injury rate and number of floors fallen will be distorted.

Good samples are a foundation of good science. In the rest of this section we give an overview of the concept of sampling, what we are trying to accomplish when we take a sample, and the inferences that are possible when researchers get it right.

Populations and samples

The first step in collecting any biological data is to decide on the target population. A **population** is the entire collection of individual units that a researcher is interested in. Ordinarily, a population is composed of a large number of individuals—so many that it is not possible to measure them all. Examples of populations include

- all cats that have fallen from buildings in New York City,
- all the genes in the human genome,
- all individuals of voting age in Australia,

- all paradise flying snakes in Borneo, and
- all children in Vancouver, Canada, suffering from asthma.

A **sample** is a much smaller set of individuals selected from the population.[2] The researcher uses this sample to draw conclusions that, hopefully, apply to the whole population. Examples include

- the fallen cats brought to one veterinary clinic in New York City,
- a selection of 20 human genes,
- all voters in an Australian pub,
- eight paradise flying snakes caught by researchers in Borneo, and
- a selection of 50 children in Vancouver, Canada, suffering from asthma.

> A *population* is all the individual units of interest, whereas a *sample* is a subset of units taken from the population.

In most of the above examples, the basic unit of sampling is literally a single individual. However, in one example, the sampling unit was a single gene. Sometimes the basic unit of sampling is a *group* of individuals, in which case a sample consists of a set of such groups. Examples of units that are groups of individuals include a single family, a colony of microbes, a plot of ground in a field, an aquarium of fish, and a cage of mice. Scientists use several terms to indicate the sampling unit, such as *unit, individual, subject,* or *replicate*.

Properties of good samples

Estimates based on samples are doomed to depart somewhat from the true population characteristics simply by chance. This chance difference from the truth is called **sampling error**. The spread of estimates resulting from sampling error indicates the **precision** of an estimate. The lower the sampling error, the higher the precision. Larger samples are less affected by chance and so, all else being equal, larger samples will have lower sampling error and higher precision than smaller samples.

> *Sampling error* is the difference between an estimate and the population parameter being estimated caused by chance.

Ideally, our estimate is **accurate** (or **unbiased**), meaning that the average of estimates that we might obtain is centered on the true population value. If a sample is not properly taken, measurements made on it might systematically underestimate (or overestimate) the population parameter. This is a second kind of error called **bias**.

[2] In biology, a "blood sample" or a "tissue sample" might refer to a substance taken from a single individual. In statistics, we reserve the word *sample* to refer to a subset of individuals drawn from a population.

Bias is a systematic discrepancy between the estimates we would obtain, if we could sample a population again and again, and the true population characteristic.

The major goal of sampling is to minimize sampling error and bias in estimates. Figure 1.2-2 illustrates these goals by analogy with shooting at a target. Each point represents an estimate of the population bull's-eye (i.e., of the true characteristic). Multiple points represent different estimates that we might obtain if we could sample the population repeatedly. Ideally, all the estimates we might obtain are tightly grouped, indicating low sampling error, and they are centered on the bull's-eye, indicating low bias. Estimates are *precise* if the values we might obtain are tightly grouped and highly repeatable, with different samples giving similar answers. Estimates are *accurate* if they are centered on the bull's-eye. Estimates are imprecise, on the other hand, if they are spread out, and they are biased (inaccurate) if they are displaced systematically to one side of the bull's-eye. The shots (estimates) on the upper right-hand target in Figure 1.2-2 are widely spread out but centered on the bull's-eye, so we say that the estimates are accurate but imprecise. The shots on the lower left-hand target are tightly grouped but not near the bull's-eye, so we say that they are precise but inaccurate. Both precision and accuracy are important, because a lack of either means that an estimate is likely to differ greatly from the truth.

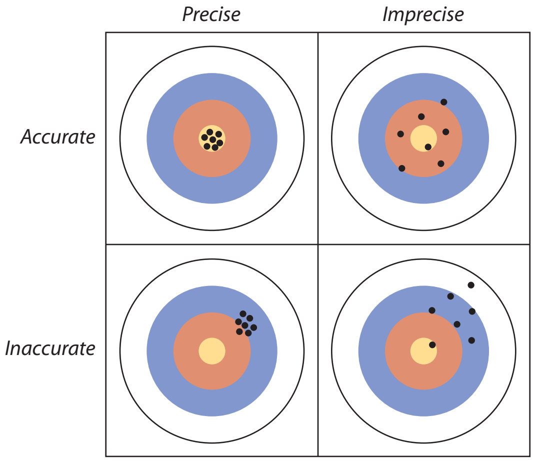


FIGURE 1.2-2
Analogy between estimation and target shooting. An accurate estimate is centered around the bull's-eye, whereas a precise estimate has low spread.

With sampling, we also want to be able to quantify the precision of an estimate. There are several quantities available to measure precision, which we discuss in Chapter 4.

The sample of cats in Example 1.2 falls short in achieving some of these goals. If uninjured and dead cats do not make it to the pet hospital, then estimates of injury rate are biased. Injury rates for cats falling only two or three floors are likely to be overestimated, whereas injury rates for cats falling many stories might be underestimated.

Random sampling

The common requirement of the methods presented in this book is that the data come from a **random sample**. A random sample is a sample from a population that fulfills two criteria.

First, every unit in the population must have an **equal chance** of being included in the sample. This is not as easy as it sounds. A botanist estimating plant growth might be more likely to find the taller individual plants or to collect those closer to the road. Some members of animal or human populations may be difficult to collect because they are shy of traps, never answer the phone, ignore questionnaires, or live at greater depths or distances than other members. These hard-to-sample individuals might differ in their characteristics from those of the rest of the population, so underrepresenting them in samples would lead to bias.

Second, the selection of units must be **independent**. In other words, the selection of any one member of the population must in no way influence the selection of any other member.[3] This, too, is not easy to ensure. Imagine, for example, that a sample of adults is chosen for a survey of consumer preferences. Because of the effort required to contact and visit each household to conduct an interview, the lazy researcher is tempted to record the preferences of multiple adults in each household and add their responses to those of other adults in the sample. This approach violates the criterion of independence, because the selection of one individual has increased the probability that another individual from the same household will also be selected. This will distort the sampling error in the data if individuals from the same household have preferences more similar to one another than would individuals randomly chosen from the population at large. With non-independent sampling, our sample size is effectively smaller than we think. This, in turn, will cause us to miscalculate the precision of the estimates.

In a *random sample,* each member of a population has an equal and independent chance of being selected.

In general, the surest way to minimize bias and allow sampling error to be quantified is to obtain a random sample.[4]

Random sampling minimizes bias and makes it possible to measure the amount of sampling error.

[3] Other than by the removal from the population of those individuals already selected, which prevents them from being sampled again.

[4] Methods are available for more complicated sampling designs incorporating non-random sampling, but we don't discuss them in this book.

How to take a random sample

Obtaining a random sample is easy in principle but can be challenging in practice. A random sample can be obtained by using the following procedure:

1. Create a list of every unit in the population of interest, and give each unit a number between one and the total population size.
2. Decide on the number of units to be sampled (call this number n).
3. Using a random-number generator,[5] generate n random integers between one and the total number of units in the population.
4. Sample the units whose numbers match those produced by the random-number generator.

An example of this process is shown in Figure 1.2-3. In both panels of the figure, we've drawn the locations of all 5699 trees present in 2001 in a carefully mapped tract of Harvard Forest in Massachussets, USA (Barker-Plotkin et al. 2006). Every tree in this population has a unique number between 1 and 5699 to identify it. We used a computerized random-number generator to pick $n = 20$ random integers between 1 and 5699, where 20 is the desired sample size. The 20 random integers, after sorting, are as follows:

156, 167, 232, 246, 826, 1106, 1476, 1968, 2084, 2222, 2223, 2284, 2790, 2898, 3103, 3739, 4315, 4978, 5258, 5500

These 20 randomly chosen trees are identified by red dots in the left panel of Figure 1.2-3.

How realistic are the requirements of a random sample? Creating a numbered list of every individual member of a population might be feasible for patients recorded in a hospital database, for children registered in an elementary-school system, or for

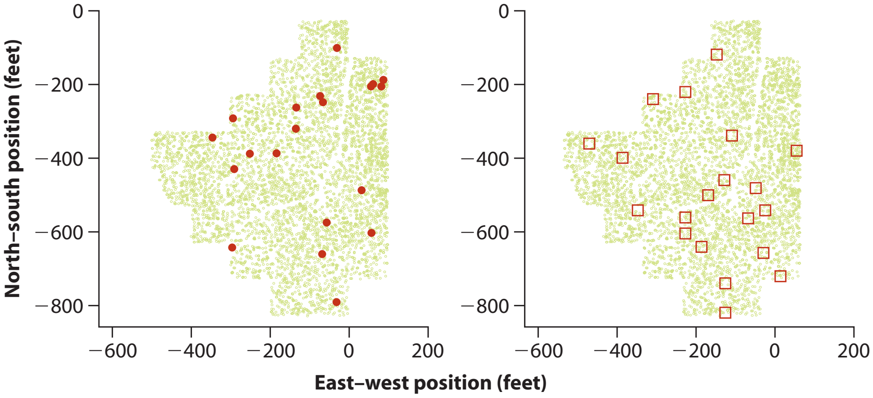


FIGURE 1.2-3 The locations of all 5699 trees present in the Prospect Hill Tract of Harvard Forest in 2001 (green circles). The red dots in the left panel are a random sample of 20 trees. The squares in the right panel are a random sample of 20 quadrats (each 20 feet on a side).

[5] For example, one is available at www.random.org.

some other populations for which a registry has been built. The feat is impractical for most plant populations, however, and unimaginable for most populations of animals or microbes. What can be done in such cases?

One answer is that the basic unit of sampling doesn't have to be a single individual—it can be a group, instead. For example, it is easier to use a map to divide a forest tract into many equal-sized blocks or plots and then to create a numbered list of these plots than it is to produce a numbered list of every tree in the forest. To illustrate this second approach, we divided the Harvard Forest tract into 836 plots of 400 square feet each. With the aid of a random-number generator, we then identified a random sample of 20 plots, which are identified by the squares in the right panel of Figure 1.2-3.

The trees contained within a random sample of plots do *not* constitute a random sample of trees, for the same reason that all of the adults inhabiting a random sample of households do not constitute a random sample of adults. Trees in the same plot are not sampled independently; this can cause problems if trees growing next to one another in the same plot are more similar (or more different) in their traits than trees chosen randomly from the forest. The data in this case must be handled carefully. A simple technique is to take the average of the measurements of all of the individuals within a unit as the single independent observation for that unit.

Random numbers should always be generated with the aid of a computer. Haphazard numbers made up by the researcher are not likely to be random (see Example 19.1). Most spreadsheet programs and statistical software packages on the computer include random-number generators.

The sample of convenience

One undesirable alternative to the random sample is the **sample of convenience**, a sample based on individuals that are easily available to the researcher. The researchers must assume (i.e., dream) that a sample of convenience is unbiased and independent like a random sample, but there is no way to guarantee it.

> A *sample of convenience* is a collection of individuals that are easily available to the researcher.

The main problem with the sample of convenience is bias, as the following examples illustrate:

- If the only cats measured are those brought to a veterinary clinic, then the injury rate of cats that have fallen from high-rise buildings is likely to be underestimated. Uninjured and fatally injured cats are less likely to make it to the vet and into the sample.
- The spectacular collapse of the North Atlantic cod fishery in the last century was caused in part by overestimating cod densities in the sea, which led to excessive allowable catches by fishing boats (Walters and Maguire 1996).

> Density estimates were too high because they relied heavily on the rates at which the fishing boats were observed to capture cod. However, the fishing boats tended to concentrate in the few remaining areas where cod were still numerous, and they did not randomly sample the entire fishing area (Rose and Kulka 1999).

A sample of convenience might also violate the assumption of independence if individuals in the sample are more similar to one another in their characteristics than individuals chosen randomly from the whole population. This is likely if, for example, the sample includes a disproportionate number of individuals who are friends or who are related to one another.

Volunteer bias

Human studies in particular must deal with the possibility of **volunteer bias**, which is a bias resulting from a systematic difference between the pool of volunteers (the **volunteer sample**) and the population to which they belong. The problem arises when the behavior of the subjects affects their chance of being sampled.

In a large experiment to test the benefits of a polio vaccine, for example, participating schoolchildren were randomly chosen to receive either the vaccine or a saline solution (serving as the control). The vaccine proved effective, but the rate at which children in the saline group contracted polio was found to be higher than in the general population. Perhaps parents of children who had not been exposed to polio prior to the study, and therefore had no immunity, were more likely to volunteer their children for the study than parents of kids who had been exposed (Brownlee 1955, Bland 2000).

Compared with the rest of the population, volunteers might be

- more health conscious and more proactive;
- low-income (if volunteers are paid);
- more ill, particularly if the therapy involves risk, because individuals who are dying anyway might try anything;
- more likely to have time on their hands (e.g., retirees and the unemployed are more likely to answer telephone surveys);
- more angry, because people who are upset are sometimes more likely to speak up; or
- less prudish, because people with liberal opinions about sex are more likely to speak to surveyors about sex.

Such differences can cause substantial bias in the results of studies. Bias can be minimized, however, by careful handling of the volunteer sample, but the resulting sample is still inferior to a random sample.

Data in the real world

In this book we use real data, hard-won from observational or experimental studies in the lab and field and published in the literature. Do the samples on which the studies are based conform to the ideals outlined above? Alas, the answer is often no. Random samples are desired but often not achieved by researchers working in the trenches. Sometimes, the only data available are a sample of convenience or a volunteer sample, as the falling cats in Example 1.2 demonstrate.

Scientists deal with this problem by taking every possible step to obtain random samples. If random sampling is impossible, then it is important to acknowledge that the problem exists and to point out where biases might arise in their studies.[6] Ultimately, further studies should be carried out that attempt to control for any sampling problems evident in earlier work.

1.3 Types of data and variables

With a sample in hand, we can begin to measure variables. A **variable** is any characteristic or measurement that differs from individual to individual. Examples include running speed, reproductive rate, and genotype. Estimates (e.g., average running speed of a random sample of 10 lizards) are also variables, because they differ by chance from sample to sample. **Data** are the measurements of one or more variables made on a sample of individuals.

> *Variables* are characteristics that differ among individuals.

Categorical and numerical variables

Variables can be categorical or numerical. **Categorical variables** describe membership in a category or group. They describe qualitative characteristics of individuals that do not correspond to a degree of difference on a numerical scale. Categorical variables are also called attribute or qualitative variables. Examples of categorical variables include

- survival (alive or dead),
- sex chromosome genotype (e.g., XX, XY, XO, XXY, or XYY),
- method of disease transmission (e.g., water, air, animal vector, or direct contact),
- predominant language spoken (e.g., English, Mandarin, Spanish, Indonesian, etc.),

[6] We biologists are generally happier to find such flaws in other researchers' data than in our own.

- life stage (e.g., egg, larva, juvenile, subadult, or adult),
- snakebite severity score (e.g., minimal severity, moderate severity, or very severe), and
- size class (e.g., small, medium, or large).

A categorical variable is *nominal* if the different categories have no inherent order. Nominal means “name.” Sex chromosome genotype, method of disease transmission, and predominant language spoken are nominal variables. In contrast, the values of an *ordinal* categorical variable can be ordered. Unlike numerical data, the magnitude of the difference between consecutive values is not known. Ordinal means “having an order.” Life stage, snakebite severity score, and size class are ordinal categorical variables.

Categorical data are qualitative characteristics of individuals that do not have magnitude on a numerical scale.

A variable is **numerical** when measurements of individuals are quantitative and have magnitude. These variables are numbers. Measurements that are counts, dimensions, angles, rates, and percentages are numerical. Examples of numerical variables include

- core body temperature (e.g., degrees Celsius, °C),
- territory size (e.g., hectares),
- cigarette consumption rate (e.g., average number per day),
- age at death (e.g., years),
- number of mates, and
- number of amino acids in a protein.

Numerical data are quantitative measurements that have magnitude on a numerical scale.

Numerical data are either continuous or discrete. *Continuous* numerical data can take on any real-number value within some range. Between any two values of a continuous variable, an infinite number of other values are possible. In practice, continuous data are rounded to a predetermined number of digits, set for convenience or by the limitations of the instrument used to take the measurements. Core body temperature, territory size, and cigarette consumption rate are continuous variables.

In contrast, *discrete* numerical data come in indivisible units. Number of amino acids in a protein and numerical rating of a statistics professor in a student evaluation are discrete numerical measurements. Number of cigarettes consumed on a specific day is a discrete variable, but the rate of cigarette consumption is a continuous variable when calculated as an average number per day over a large number of days. In practice, discrete numerical variables are often analyzed as though they were continuous, if there is a large number of possible values.

Just because a variable is indexed by a number does not mean it is a numerical variable. Numbers might also be used to name categories (e.g., family 1, family 2, etc.). Numerical data can be reduced to categorical data by grouping, though the result contains less information (e.g., "above average" and "below average").

Explanatory and response variables

One major use of statistics is to relate one variable to another, by examining associations between variables and differences between groups. Measuring an association is equivalent to measuring a difference, statistically speaking. Showing that "the proportion of survivors *differs* between treatment categories" is the same as showing that the variables "survival" and "treatment" are *associated*.

Often when association between two variables is investigated, a goal is to assess how well one of the variables, deemed the **explanatory variable**, predicts or affects the other variable, called the **response variable**. When conducting an experiment, the treatment variable (the one manipulated by the researcher) is the explanatory variable, and the measured effect of the treatment is the response variable. For example, the administered dose of a toxin in a toxicology experiment would be the explanatory variable, and organism survival would be the response variable. When neither variable is manipulated by the researcher, their association might nevertheless be described by the "effect" of one of the variables (the explanatory) on the other (the response), even though the association itself is not direct evidence for causation. For example, when exploring the possibility that high blood pressure affects the risk of stroke in a sample of people, blood pressure is the explanatory variable and incidence of strokes is the response variable. When natural groups of organisms, such as populations or species, are compared in some measurement, such as body mass, the group variable (population or species) is typically the explanatory variable and the measurement is the response variable. In more complicated studies involving more than two variables, there may be more than one explanatory or response variable.

Sometimes you will hear variables referred to as "independent" and "dependent." These are the same as explanatory and response variables, respectively. Strictly speaking, if one variable depends on the other, then neither is independent, so we prefer to use *explanatory* and *response* throughout this book.

1.4 Frequency distributions and probability distributions

Different individuals in a sample will have different measurements. We can see this variability with a frequency distribution. The **frequency** of a specific measurement in a sample is the number of observations having a particular value of the measurement.[7] The **frequency distribution** shows how often each value of the variable occurs in the sample.

The *frequency distribution* describes the number of times each value of a variable occurs in a sample.

Figure 1.4-1 shows the frequency distribution for the measured beak depths of a sample of 100 finches from a Galápagos island population.[8]

The large-beaked ground finch on the Galápagos Islands.

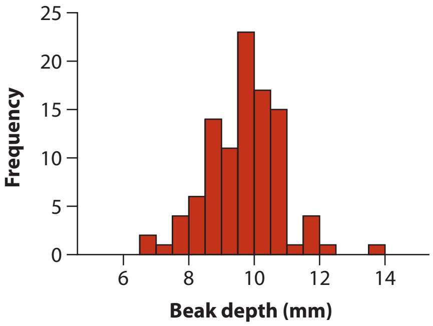


FIGURE 1.4-1 The frequency distribution of beak depths in a sample of 100 finches from a Galápagos island population (Boag and Grant 1984). The vertical axis indicates the frequency, the number of observations in each 0.5-mm interval of beak depth.

We use the frequency distribution of a sample to inform us about the distribution of the variable (beak depth) in the population from which it came. Looking at a frequency distribution gives us an intuitive understanding of the variable. For example, we can see which values of beak depth are common and which are rare, we can get an idea of the average beak depth, and we can start to understand how variable beak depth is among the finches living on the island.

The distribution of a variable in the whole population is called its **probability distribution**. The real probability distribution of a population in nature is almost never known. Researchers typically use theoretical probability distributions to approximate the real probability distribution. For a continuous variable like beak depth, the distribution in the population is often approximated by a theoretical probability distribution known as the **normal distribution**. The normal distribution drawn

[7] The **absolute frequency** is the number of times that a value is observed. The **relative frequency** is the proportion of individuals which have that value.

[8] Beak depth is the height of the beak near its base.

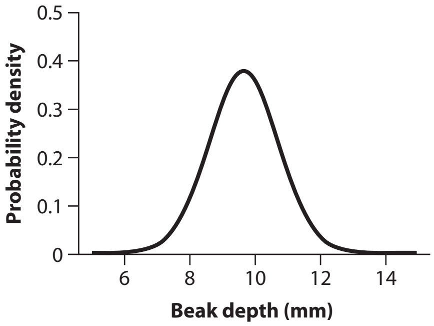


FIGURE 1.4-2
A normal distribution. This probability distribution is often used to approximate the distribution of a variable in the population from which a sample has been drawn.

in Figure 1.4-2, for example, approximates the probability distribution of beak depths in the finch population from which the sample of 100 birds was drawn.

The normal distribution is the familiar "bell curve." It is the most important probability distribution in all of statistics. You'll learn a lot more about it in Chapter 10. Most of the methods presented in this book for analyzing data depend on the normal distribution in some way.

1.5 Types of studies

Data in biology are obtained from either an experimental study or an observational study. In an **experimental study**, the researcher assigns different treatment groups or values of an explanatory variable randomly to the individual units of study. A classic example is the clinical trial, where different treatments are assigned randomly to patients in order to compare responses. In an **observational study**, on the other hand, the researcher has no control over which units fall into which groups.

Studies of the health consequences of cigarette smoking in humans are all observational studies, because it is ethically impossible to assign smoking and nonsmoking treatments to human beings to assess the effects of smoking. The individuals in the sample have made the decision themselves about whether to smoke. The only experimental studies of the health consequences of smoking have been carried out on nonhuman subjects, such as mice, where researchers can assign smoking and nonsmoking treatments randomly to individuals.

A study is *experimental* if the researcher assigns treatments randomly to individuals, whereas a study is *observational* if the assignment of treatments is *not* made by the researcher.

The distinction between experimental studies and observational studies is that experimental studies can determine cause-and-effect relationships between variables, whereas observational studies can only point to associations. An association between smoking and lung cancer might be due to the effects of smoking per se, or perhaps to an underlying predisposition to lung cancer in those individuals prone to smoking. It is difficult to distinguish these alternatives with observational studies alone. For this reason, experimental studies of the health hazards of smoking in nonhuman animals have helped make the case that cigarette smoking is dangerous to human health. But experimental studies are not always possible, even on animals. Smoking in humans, for example, is also associated with a higher suicide rate (Hemmingsson and Kriebel 2003). Is this association caused by the effects of smoking, or is it caused by the effects of some other variable?

Just because a study was carried out in the laboratory does not mean that the study is an experimental study in the sense described here. A complex laboratory apparatus and careful conditions may be necessary to obtain measurements of interest, but such a study is still observational if the assignment of treatments is out of the control of the researcher.

1.6 Summary

- Statistics is the study of methods for measuring aspects of populations from samples and for quantifying the uncertainty of the measurements.
- Much of statistics is about estimation, which infers an unknown quantity of a population using sample data.
- Statistics also allows hypothesis testing, a method to determine how well hypotheses about a population parameter fit the sample data.
- Sampling error is the chance difference between an estimate describing a sample and the corresponding parameter of the whole population. Bias is a systematic discrepancy between an estimate and the population quantity.
- The goals of sampling are to increase the accuracy and precision of estimates and to ensure that it is possible to quantify precision.
- In a random sample, every individual in a population has the same chance of being selected, and the selection of individuals is independent.
- A sample of convenience is a collection of individuals easily available to a researcher, but it is not usually a random sample.
- Volunteer bias is a systematic discrepancy in a quantity between the pool of volunteers and the population.
- Variables are measurements that differ among individuals.
- Variables are either categorical or numerical. A categorical variable describes which category an individual belongs to, whereas a numerical variable is expressed as a number.
- The frequency distribution describes the number of times each value of a variable occurs in a sample. A probability distribution describes the number of times each value occurs in a population. Probability distributions in populations can often be approximated by a normal distribution.

- In studies of association between two variables, one variable is typically used to predict the value of another variable and is designated as the explanatory variable. The other variable is designated as the response variable.
- In experimental studies, the researcher is able to assign subjects randomly to different treatments or groups. In observational studies, the assignment of individuals to treatments is not controlled by the researcher.

PRACTICE PROBLEMS

Answers to the practice problems are provided at the end of the book, starting on page 747.

1. Which of the following numerical variables are continuous? Which are discrete?
   a. Number of injuries sustained in a fall
   b. Fraction of birds in a large sample infected with avian flu virus
   c. Number of crimes committed by a randomly sampled individual
   d. Logarithm of body mass

2. The peppered moth (*Biston betularia*) occurs in two types: peppered (speckled black and white) and melanic (black). A researcher wished to measure the proportion of melanic individuals in the peppered moth population in England, to examine how this proportion changed from year to year in the past. To accomplish this, she photographed all the peppered moth specimens available in museums and large private collections and grouped them by the year in which they had been collected. Based on this sample, she calculated the proportion of melanic individuals in every year. The people who collected the specimens, she knew, would prefer to collect whichever type was rarest in any given year, since those would be the most valuable.
   a. Can the specimens from any given year be considered a random sample from the moth population?
   b. If not a random sample, what type of sample is it?
   c. What type of error might be introduced by the sampling method when estimating the proportion of melanic moths?

3. What feature of an estimate—precision or accuracy—is most strongly affected when individuals differing in the variable of interest do not have an equal chance of being selected?

4. In a study of stress levels in U.S. army recruits stationed in Iraq, researchers obtained a complete list of the names of recruits in Iraq at the time of the study. They listed the recruits alphabetically and then numbered them consecutively. One hundred random numbers between one and the total number of recruits were then generated using a random-number generator on a computer. The 100 recruits whose numbers corresponded to those generated by the computer were interviewed for the study.
   a. What is the population of interest in this study?
   b. The 100 recruits interviewed were randomly sampled as described. Is the sample affected by sampling error? Explain.
   c. What are the main advantages of random sampling in this example?
   d. What effect would a larger sample size have had on sampling error?

5. An important quantity in conservation biology is the number of plant and animal species inhabiting a given area. To survey the community of small mammals inhabiting Kruger National Park in South Africa, a large series of live traps were placed randomly throughout the park for one week in the main dry season of 2004. Traps were set each evening and checked the following morning. Individuals caught were identified, tagged (so that new captures could be distinguished from recaptures), and released. At the end of the survey, the total number of small mammal species in the park was estimated by the total number of species captured in the survey.
   a. What is the parameter being estimated in the survey?
   b. Is the sample of individuals captured in the traps likely to be a random sample? Why or why not? In your answer, address the two criteria that define a sample as random.
   c. Is the number of species in the sample likely to be an unbiased estimate of the total number of small mammal species in the park? Why or why not?

6. In a recent study, researchers took electrophysiological measurements from the brains of two rhesus macaques (monkeys). Forty neurons were tested in each monkey, yielding a total of 80 measurements.
   a. Do the 80 neurons constitute a random sample? Why or why not?
   b. If the 80 measurements were analyzed as though they constituted a random sample, what consequences would this have for the estimate of the measurement in the monkey population?

7. Identify which of the following variables are discrete and which are continuous:
   a. Number of warts on a toad
   b. Survival time after poisoning
   c. Temperature of porridge
   d. Number of bread crumbs in 10 meters of trail
   e. Length of wolves' canines

8. A study was carried out in women to determine whether the psychological consequences of having an abortion differ from those experienced by women who have lost their fetuses by other causes at the same stage of pregnancy.
   a. Which is the explanatory variable in this study, and which is the response variable?
   b. Was this an observational study or an experimental study? Explain.

9. For each of the following studies, say which is the explanatory variable and which is the response variable. Also, say whether the study is observational or experimental.
   a. Forestry researchers wanted to compare the growth rates of trees growing at high altitude to that of trees growing at low altitude. They measured growth rates using the space between tree rings in a set of trees harvested from a natural forest.
   b. Researchers randomly assign diabetes patients to two groups. In the first group, the patients receive a new drug tasploglutide, whereas patients in the other group receive standard treatment without the new drug. The researchers compared the rate of insulin release in the two groups.
   c. Psychologists tested whether the frequency of illegal drug use differs between people suffering from schizophrenia and those not having the disease. They measured drug use in a group of schizophrenia patients and compared it with that in a similar sized group of randomly chosen people.
   d. Spinner Hansen et al. (2008) studied a species of spider whose females often eat males that are trying to mate with them. The researchers removed a leg from each male spider in one group (to make them weaker and more vulnerable to being eaten) and left the males in another group undamaged. They studied whether survival of males in the two groups differed during courtship.
   e. Bowen et al. (2012) studied the effects of advanced communication therapy for patients whose communication skills had been affected by previous strokes. They randomly assigned two therapies to stroke patients. One group received advanced communication therapy and the other received only social visits without formal therapy. Both groups otherwise received normal,

best-practice care. After six months, the communication ability (as measured by a standardized quantitative test score) was measured on all patients.

10. Each of the examples (a–e) in problem 9 involves estimating or testing an association between two variables. For each of the examples, list the two variables and state whether each is categorical or numerical.

11. A random sample of 500 households was identified in a major North American city using the municipal voter registration list. Five hundred questionnaires went out, directed at one adult in each household, and surveyed respondents about attitudes regarding the municipal recycling program. Eighty of the 500 surveys were filled out and returned to the researchers.

a. Can the 80 households that returned questionnaires be regarded as a random sample of households? Explain.

b. What type of bias might affect the survey outcome?

12. State whether the following represent cases of estimation or hypothesis testing.

a. A random sample of quadrats in Olympic National Forest is taken to determine the average density of *Ensatina* salamanders.

b. A study is carried out to determine whether there is extrasensory perception (ESP), by counting the number of cards guessed correctly by a subject isolated from a second person who is drawing cards randomly from a normal card deck. The number of correct guesses is compared with the number we would expect by chance if there were no such thing as ESP.

c. A trapping study measures the rate of fruit fall in forest clear-cuts.

d. An experiment is conducted to calculate the optimal number of calories per day to feed captive sugar gliders (*Petaurus breviceps*) to maintain normal body mass and good health.

e. A clinical trial is carried out to determine whether taking large doses of vitamin C benefits health of advanced cancer patients.

f. An observational study is carried out to determine whether hospital emergency room admissions increase during nights with a full moon compared with other nights.

13. A researcher dissected the retinas of 20 randomly sampled fish belonging to each of two subspecies of guppy in Venezuela. Using a sophisticated laboratory apparatus, he measured the two groups of fish to find the wavelengths of visible light to which the cones of the retina were most sensitive. The goal was to explore whether fish from the two subspecies differed in the wavelength of light that they were most sensitive to.

a. What are the two variables being associated in this study?

b. Which is the explanatory variable and which is the response variable?

c. Is this an experimental study or an observational study? Why?

ASSIGNMENT PROBLEMS

14. Identify whether the following variables are numerical or categorical. If numerical, state whether the variable is discrete or continuous. If categorical, state whether the categories have a natural order (ordinal) or not (nominal).

a. Number of sexual partners in a year

b. Petal area of rose flowers

c. Heartbeats per minute of a Tour de France cyclist, averaged over the duration of the race

d. Birth weight

e. Stage of fruit ripeness (e.g., underripe, ripe, or overripe)

f. Angle of flower orientation relative to position of the sun

g. Tree species

h. Year of birth

i. Gender

15. In the vermilion flycatcher, the males are brightly colored and sing frequently and prominently. Females are more dull-colored and make less sound. In a field study of this bird, a researcher attempted to estimate the fraction of individuals of each sex in the population. She based her estimate on the number of individuals of each sex detected while walking through suitable habitat. Is her sample of birds detected likely to be a random sample? Why or why not?

16. Not all telephone polls carried out to estimate voter or consumer preferences make calls to cell phones. One reason is that in the USA, automated calls ("robocalls") to cell phones are not permitted and interviews conducted by humans are more costly.

- **a.** How might the strategy of leaving out cell phones affect the goal of obtaining a random sample of voters or consumers?
- **b.** Which criterion of random sampling is most likely to be violated by the problems you identified in part (a): equal chance of being selected, or the independence of the selection of individuals?
- **c.** Which attribute of estimated consumer preference is most affected by the problem you identified in (a): accuracy or precision?

17. The average age of piñon pine trees in the coast ranges of California was investigated by placing 500 ten-hectare plots randomly on a distribution map of the species in California using a computer. Researchers then found the location of each random plot in the field, and then they measured the age of every piñon pine tree within each of the 10-hectare plots. The average age within the plot was used as the unit measurement. These unit measurements were then used to estimate the average age of California piñon pines.

- **a.** What is the population of interest in this study?
- **b.** Why did the researchers take an average of the ages of trees within each plot as their unit measurement, rather than combine into a single sample the ages of all the trees from all the plots?

18. Refer to problem 17.

- **a.** Is the estimate of age based on 500 plots influenced by sampling error? Why?
- **b.** How would the sampling error of the estimate of mean age change if the investigators had used a sample of only 100 plots?

19. In each of the following examples, indicate which variable is the explanatory variable and which is the response variable.

- **a.** The anticoagulant warfarin is often used as a pesticide against house mice, *Mus musculus*. Some populations of the house mouse have acquired a mutation in the *vkorc1* gene from hybridizing with the Algerian mouse, *M. spretus* (Song et al. 2011). In the Algerian mice, this gene confers resistance to warfarin. In a hypothetical follow-up study, researchers collected a sample of house mice to determine whether presence of the *vkorc1* mutation is associated with warfarin resistance in house mice as well. They fed warfarin to all the mice in a sample and compared survival between the individuals possessing the mutation and those not possessing the mutation.
- **b.** Cooley et al. (2009) randomly assigned either of two treatments, naturopathic care (diet counseling, breathing techniques, vitamins, and a herbal medicine) or standardized psychotherapy (psychotherapy with breathing techniques and a placebo added), to 81 individuals having moderate to severe anxiety. Anxiety scores decreased an average of 57% in the naturopathic group and 31% in the psychotherapy group.
- **c.** Individuals highly sensitive to rewards tend to experience more food cravings and are more likely to be overweight or develop eating disorders than other people. Beaver et al. (2006) recruited 14 healthy volunteers

and scored their reward sensitivity using a questionnaire (they were asked to answer yes or no to questions like: "I'm always willing to try something new if I think it will be fun"). The subjects were then presented with images of appetizing foods (e.g., chocolate cake, pizza) while activity of their fronto–striatal–amygdala–midbrain was measured using functional MRI. Reward sensitivity of subjects was found to correlate with brain activity in response to the images.

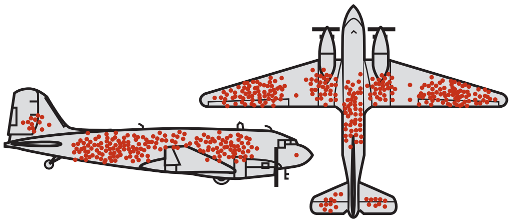
d. Endostatin is a naturally occurring protein in humans and mice that inhibits the growth of blood vessels. O'Reilly et al. (1997) investigated its effects on growth of cancer tumors, whose growth and spread requires blood vessel proliferation. Mice having lung carcinoma tumors were randomly divided into groups that were treated with doses of 0, 2.5, 10, or 20 mg/kg of endostatin injected once daily. They found that higher doses of endostatin led to inhibition of tumor growth.

20. For each of the studies presented in problem 19, indicate whether the study is an experimental or observational study.

21. In a study of heart rate in ocean-diving birds, researchers harnessed 10 randomly sampled, wild-caught cormorants to a laboratory contraption that monitored vital signs. Each cormorant was subjected to six artificial "dives" over the following week (one dive per day). A dive consisted of rapidly immersing the head of the bird in water by tipping the harness. In this way, a sample of 60 measurements of heart rate in diving birds was obtained. Do these 60 measurements represent a random sample? Why or why not?

22. Researchers sent out a survey to U.S. war veterans that asked a series of questions, including whether individuals surveyed were smokers or nonsmokers (Seltzer et al. 1974). They found that nonsmokers were 27% more likely than smokers to respond to a survey within 30 days (based on the much larger number of smokers and nonsmokers who eventually responded). Hypothetically, if the study had ended after 30 days, what effect would this have on the estimate of the proportion of veterans who smoke? (Use terminology you learned in this chapter to describe the effect.)

23. During World War II, the British Royal Air Force estimated the density of bullet holes on different sections of planes returning to base from aerial sorties. Their goal was to use this information to determine which plane sections most needed additional protective shields. (It was not possible to reinforce the whole plane, because it would weigh too much.) They found that the density of holes was highest on the wings and lowest on the engines and near the cockpit, where the pilot sits (their initial conclusion, that therefore the wings should be reinforced, was later shown to be mistaken). What is the main problem with the sample: bias or large sampling error? What part of the plane should have been reinforced?



24. In a study of diet preferences in leafcutter ants, a researcher presented 20 randomly chosen ant colonies with leaves from the two most common tree species in the surrounding forest. The leaves were placed in piles of 100, one pile for each tree species, close to colony entrances. Leaves were cut so that each was small enough to be carried by a single ant. After 24 hours, the researcher returned and counted the number of leaves remaining of the original 100 of each species. Some of the results are shown in the following table.

Tree species	Number of leaves removed
Spondius mombin	1561
Sapium thelocarpum	851
Total	2412

Using these results, the researcher estimated the proportion of *Spondius* leaves taken as 0.65 and concluded that the ants have a preference for leaves of this species.

a. Identify the two variables whose association is displayed in the table. Which is the explanatory variable and which is the response variable? Are they numeric or categorical?

b. Why do the 2412 leaves used in the calculation of the proportion not represent a random sample?

c. Would treating the 2412 leaves as a random sample most likely affect the accuracy of the estimate of diet preference or the precision of the estimate?

d. If not the leaves, what units were randomly sampled in the study?

25. Garaci et al. (2012) examined a sample of people with and without multiple sclerosis (MS) to test the controversial idea that the disease is caused by blood flow restriction resulting from a vein condition known as chronic cerebrospinal venous insufficiency (CCSVI). Of 39 randomly sampled patients with MS, 25 were found to have CCSVI and 14 were not. Of 26 healthy control subjects, 14 were found to have CCSVI and 12 were not. The researchers found an association between CCSVI and MS.

a. What is the explanatory variable and what is the response variable?

b. Is this an experimental study or an observational study?

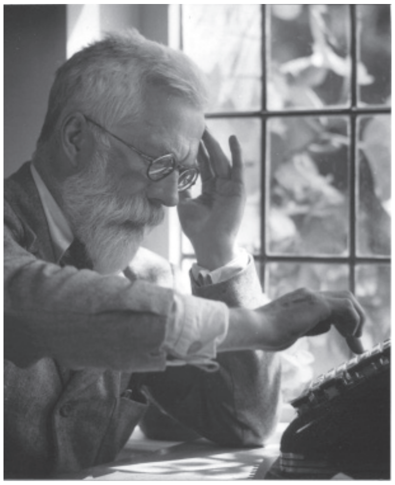
c. Where might hypothesis testing have been used in the study?

Biology and the history of statistics

The formal study of probability began in the mid-17th century when Blaise Pascal and Pierre de Fermat started to describe the mathematical rules for determining the best gambling strategies. Gambling and insurance continued to motivate the development of probability for the next couple of centuries.

The application of probability to data did not happen until much later. The importance of variation in the natural world, and by extension to samples from that world, was made obvious by Charles Darwin's theory of natural selection. Darwin highlighted the importance of variation in biological systems as the raw material of evolution. Early followers of Darwin saw the need for quantitative descriptions of variation and the need to incorporate the effects of sampling error in biology. This led to the development of modern statistics. In many ways, therefore, modern statistics was an offshoot of evolutionary biology.

One of the first pioneers in statistical data analysis was Francis Galton, who began to apply probability thinking to the study of all sorts of data. Galton was a real polymath, thinking about nearly everything and collecting and analyzing data at every chance. He said, "Whenever you can, count."[1] He invented the study of fingerprints, he tested whether prayer increased the life span of preachers compared with others of the middle class, and he quantified the heritable nature of many important traits. He once recorded his idea of the attractiveness of women seen from the window of a train headed from London to Glasgow, finding that "attractiveness" (at least according to Galton) declined as a function of distance from London. Galton is best known, though, for his twin interests in data analysis and evolution (he was the first cousin of Darwin). He invented the idea of regression, which we will learn more about in Chapter 17. Galton was also responsible for establishing a lab that brought more researchers into the study of both statistics and biology, including Karl Pearson and Ronald Fisher.



Sir Ronald Fisher

Karl Pearson, like Galton, was interested in many spheres of knowledge. Pearson was motivated by biological data—in particular, by heredity and evolution. Pearson was

[1] Quoted J. R. Newman, *The World of Mathematics* (New York: Simon & Schuster, 1956).

responsible for our most often-used measure of the correlation between numerical variables. In fact, the correlation coefficient that we will learn about in Chapter 16 is often referred to as Pearson's correlation coefficient. He also made many contributions to the study of regression analysis and invented the χ^2 contingency test (Chapter 9). He also invented the term *standard deviation* (Chapter 3).

Last, but definitely not least, Ronald Fisher was one of the great geniuses of the 20th century. Fisher is well known in evolutionary biology as one of the three founders of the field of theoretical population genetics. Among his many contributions are the demonstration that Mendelian inheritance is compatible with the continuous variation we see in many traits, the accepted theory for why most animals conceive equal numbers of male and female offspring, and a great deal of the mathematical machinery we use to describe the process of evolution. But his contributions did not end there. He is probably also the most important figure in the history of statistics. He developed the analysis of variance (Chapter 15), likelihood (Chapter 20), the *P*-value (Chapter 6), randomized experiments (Chapter 14), multiple regression, and many other tools used in data analysis. His mathematical knowledge was made practical by a lifelong association with biologists, particularly agricultural scientists at the Rothamsted Experimental Station in England. Fisher solved problems associated with the analysis of real data as he encountered them, and he generalized them for application to many other related questions. Moreover, Fisher developed experimental designs that would give more information than might otherwise be possible.

What this short hagiography is intended to demonstrate is that the early history of statistics is tightly bound up with biology. Biological questions motivated the development of most of the basic statistical tools, and biologists helped to develop those tools. Biology and statistics naturally go hand in hand.