Statistics Primer

to accompany

Life: The Science of Biology, Eighth Edition

Lauren Ancel Meyers

Why Do We Do Statistics?

Almost everything varies. We live in a variable world, but we believe that there are predictable patterns and we use science to find these patterns. Consider any group of common things in nature … all women aged 22, all the cells in your liver, or all the blades of grass in your yard. While they will have many similar characteristics, they will also have important differences. Men aged 22 tend to be taller than women aged 22, but of course, not every man will be taller than every woman in this age group.

Natural variation can make it difficult to find general patterns. For example, scientists have determined that smoking increases the risk of getting lung cancer. But we know that not all smokers will develop lung cancer and not all nonsmokers will remain cancer-free. If we compare just one smoker to just one nonsmoker, we may end up drawing the wrong conclusion. So how did scientists discover this general pattern? How many smokers and nonsmokers did they examine before they felt confident about the risk of smoking?

Statistics helps us to find general patterns, even when nature does not always follow those patterns.

Avoiding false positives and false negatives. When a woman takes a pregnancy test, there is some chance that it will be positive even if she is not pregnant, and there is some chance that it will be negative even if she is pregnant. We call these kinds of mistakes *false positives* and *false negatives*.

Doing science is a bit like taking a medical test. We observe patterns in the world, and we try to draw conclusions about how the world works from those observations. Sometimes our observations lead us to draw the wrong conclusions. We might conclude that a phenomenon occurs, when it actually does not; or we might conclude that a phenomenon does not occur, when it actually does.

For example, Earth has been warming recently. The average global air temperature near Earth's surface has increased an estimated 1.1°F over the last century (Houghton et al., 2001). Ecologists are interested in whether plant and animal populations have been affected by global warming. If we have long-term information about the locations of species and temperatures in certain areas, we can determine whether species movements coincide with temperature changes. Such information can, however, be very complicated. Without proper statistical methods, one may not be able to detect the true impact of temperature or, instead, may think a pattern exists when it does not.

Statistics helps us to avoid drawing the wrong conclusions.

How Does Statistics Help Us Understand the Natural World?

Statistics is essential to scientific discovery. Most biological studies involve five basic steps, each of which requires statistics:

Step 1: Experimental Design
Clearly define the scientific question and the methods necessary to tackle the question.

Step 2: Data Collection
Gather information about the natural world through experiments and field studies.

Step 3: Organize and Visualize the Data
Use tables, graphs, and other useful representations to gain intuition about the data.

Step 4: Summarize the Data
Summarize the data with a few key statistical calculations.

Step 5: Inferential Statistics
Use statistical methods to draw general conclusions from the data about the way the world works.

Step 1: Experimental Design

We conduct experiments to gain knowledge about the world. Scientists come up with scientific ideas based on prior research and their own observations. These ideas may take the form of a question like “Does smoking cause cancer?,” a hypothesis like “Smoking increases the risk of cancer,” or a prediction like “If a person smokes, he/she will increase his/her chances of developing cancer.” Experiments allow us to test such scientific ideas, but designing a good experiment can be quite challenging.

We use statistics to guide us in designing experiments so that we end up with the right kinds of data. Before embarking on an experiment, we use statistics to determine how much data will be required to test our idea, and to prevent extraneous factors from misleading us. For example, suppose we want to conduct a fertilization experiment to test the hypothesis that nitrogen increases plant growth. If we include too few plants, we will not be able to determine whether or not nitrogen has an effect on growth, and the experiment will be for naught. If we include too many plants, we will waste valuable time and resources. Furthermore, we should design the experiment so that we can detect differences that are actually caused by nitrogen fertilization rather than by variation, for example, in sunlight or precipitation experienced by the plants.

Step 2: Data Collection

Taking samples. When biologists gather information about the natural world, they typically collect a few representative pieces of information. For example, when evaluating the efficacy of a candidate drug for medulloblastoma brain cancer, scientists may test the drug on tens or hundreds of patients, and then draw conclusions about its efficacy for all patients with these tumors. Similarly, scientists studying the relationship between body weight and clutch size (number of eggs) for female spiders of the species *Holocnemus pluchei*, drew conclusions about the global population of these spiders based on a study of just 57 female spiders (Skow and Jakob 2003).

We use the expression "sampling from a population" to describe this general method of taking representative pieces of information from the system under investigation (Figure 1). The pieces of information in a **sample** are called **observations**. In the cancer therapy example, each observation was the change in a patient's tumor size six months after initiating treatment, and the population of interest was all individuals with medulloblastoma tumors. In the spider example, each observation was a pair of measurements—body size and clutch size—for a single female spider, and the population of interest was all female spiders of this species.

Sampling is a matter of necessity, not laziness. We cannot hope (and would not want) to collect *all* of the female *H. pluchei* spiders on Earth. Instead, we use statistics to determine how many spiders we must collect in order to confidently infer something about the general population and then use statistics again to make such inferences.

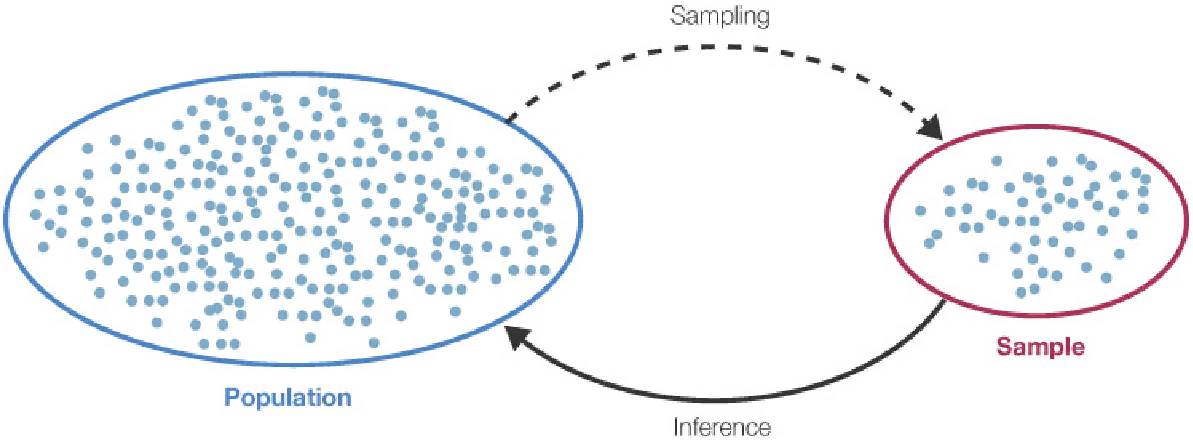


Figure 1 Sampling from a population. Biologists take representative samples from a population, use descriptive statistics to characterize their samples, and then use inferential statistics to draw conclusions about the original population.

Data come in all shapes and sizes. In statistics, we use the word *variable* to mean a measurable characteristic of an individual or a system. Some variables are on a numerical scale, like the daily high temperature (a numerical value constrained by the precision of our thermometer), or the clutch size of a spider (a whole number: 0, 1, 2, 3,…). We call these **quantitative variables**. Quantitative variables that only take on whole number values are called **discrete variables**, whereas variables that can also take on any fractional value are called **continuous variables**.

Other variables take categories as values, like a human blood type (A, B, AB, or O) or an ant caste (queen, worker, or male). We call these **categorical variables**. Categorical variables with a natural ordering, like a final grade in Biology 101 (A, B, C, D, or F), are called **ordinal variables.**

Each class of variables comes with its own set of statistical methods, as depicted in Figure 2.

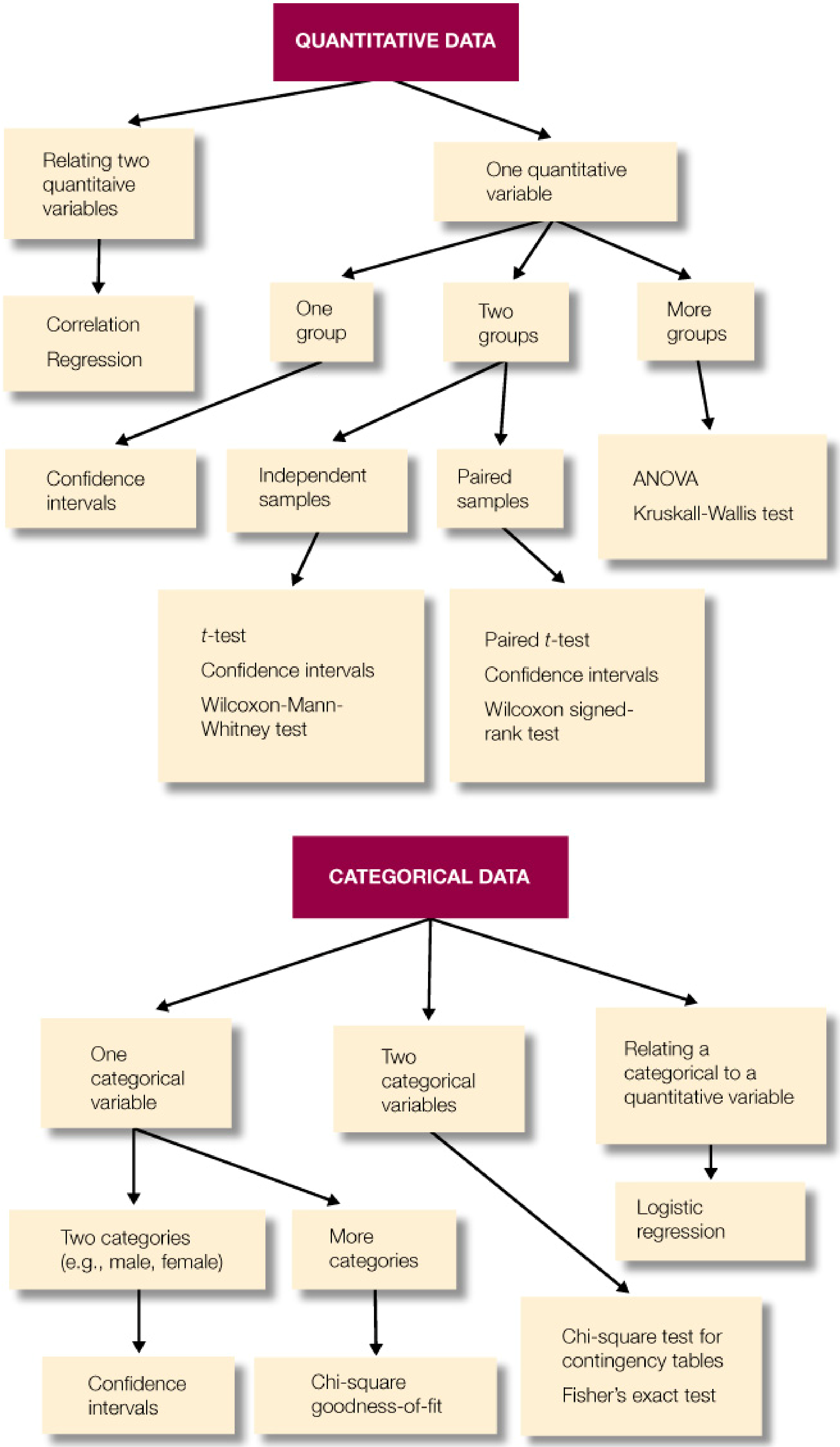


Figure 2 Statistical roadmap. This flowchart shows some of the commonly used methods of statistical inference for different combinations of data. Detailed descriptions of these methods can be found in most introductory biostatistics textbooks.

Step 3: Organize and Visualize the Data

Tables and graphs can help you gain intuition about your data, design appropriate statistical tests, and anticipate the outcome of your analysis. A **frequency distribution** lists all possible values and the number of occurrences of each value in the sample.

Table 1 shows a frequency distribution of the colors of 182 poinsettia plants (red, pink, or white) resulting from an experimental cross between two parent plants (Stewart and Arisum 1966). For categorical data like this, we can visualize the frequency distribution by constructing a **bar chart**. The heights of the bars indicate the number of observations in each category (Figure 3).

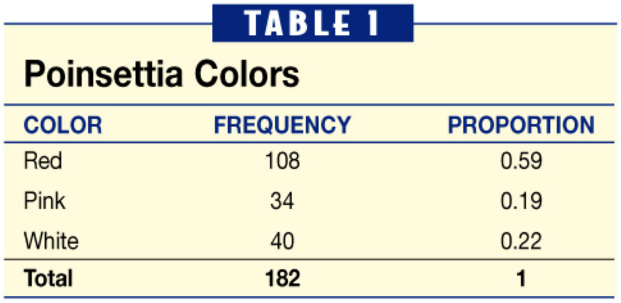


TABLE 1

Poinsettia Colors

COLOR	FREQUENCY	PROPORTION
Red	108	0.59
Pink	34	0.19
White	40	0.22
Total	**182**	**1**

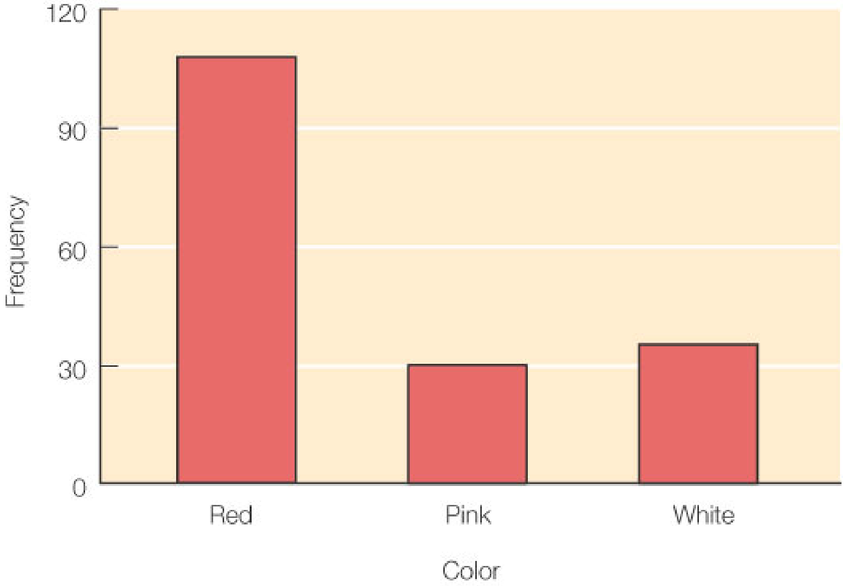


Figure 3 Bar chart of poinsettia colors.

For quantitative data, it is often useful to condense your data by grouping (or binning) it into **classes.** In Table 2, we see a grouped frequency distribution of fish weights for a sample of 34 fish (*Abramis brama*) caught in Lake Laengelmavesi in Finland (Brofeldt 1917). The second column (*Frequency*) gives the number of observations in each class and the third column (*Relative frequency*) gives the overall proportion of observations falling into each class.

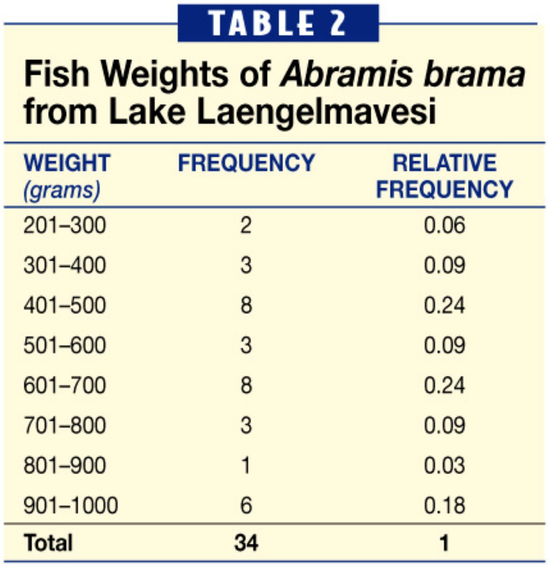


TABLE 2

Fish Weights of *Abramis brama* from Lake Laengelmavesi

WEIGHT *(grams)*	FREQUENCY	RELATIVE FREQUENCY
201–300	2	0.06
301–400	3	0.09
401–500	8	0.24
501–600	3	0.09
601–700	8	0.24
701–800	3	0.09
801–900	1	0.03
901–1000	6	0.18
Total	**34**	**1**

Histograms depict frequency distributions for quantitative data. The histogram in Figure 4 shows the relative frequencies of each weight class in this study. When grouping quantitative data, it is necessary to decide how many classes to include. It is often useful to look at multiple histograms before deciding which grouping offers the best representation of the data.

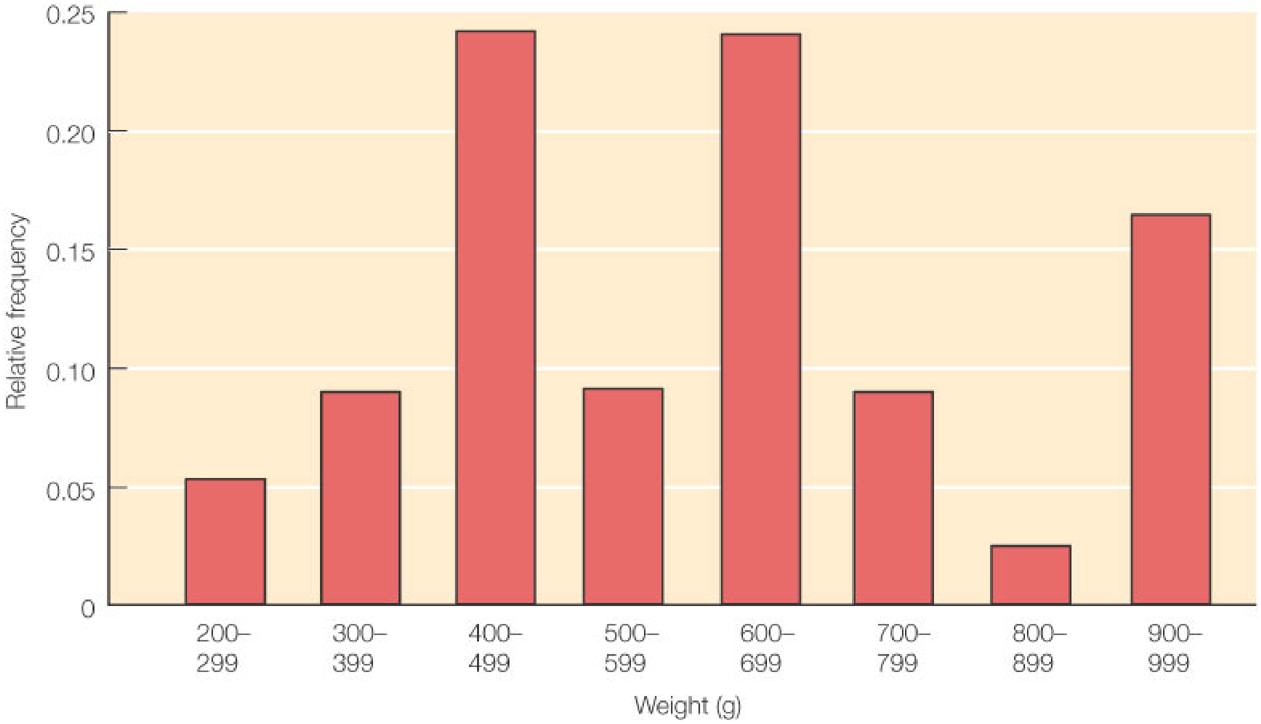


Figure 4 Histogram of *Abramis brama* weights.

Sometimes we wish to compare two quantitative variables. For example, the researchers at Lake Laengelmavesi investigated the relationship between fish weight and length and thus also measured the length of each fish. We can visualize this relationship using a **scatter plot** in which the weight and length of each fish is represented as a single point (Figure 5). We say that these two variables have a **linear relationship** since the points in their scatter plot fall roughly on a straight line.

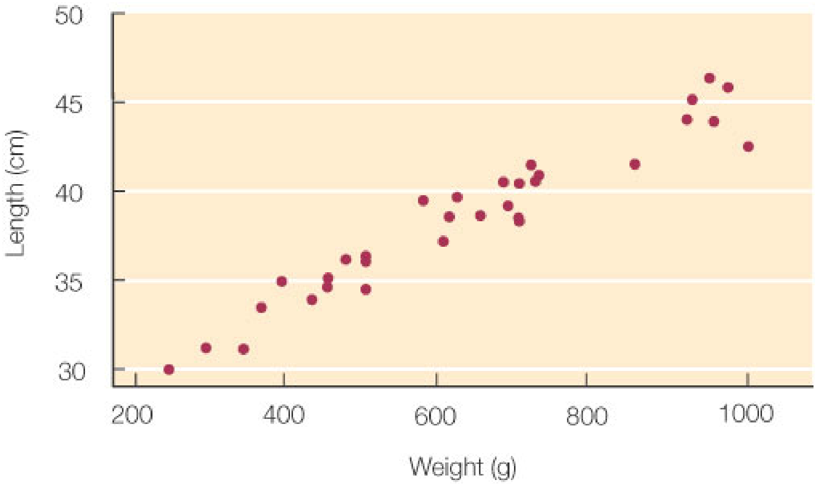


Figure 5 Scatter plot of *Abramis brama* weights and lengths (measured from nose to end of tail). These two variables have a linear relationship since the data points lie on close to a straight line.

Tables and graphs are critical to interpreting and communicating data, and thus should be as self-contained and comprehensible as possible. Their content should be easily understood simply by looking at them. Axes, captions, and units should be clearly labeled, statistical terms should be defined, and appropriate groupings should be used when tabulating or graphing quantitative data.

Step 4: Summarize the Data

A **statistic** is a numerical quantity calculated from data, while **descriptive statistics** are quantities that describe general patterns in data. Descriptive statistics allow us to make straightforward comparisons between different data sets and concisely communicate basic features of our data.

Describing categorical data. For categorical variables, we typically use proportions to describe our data. That is, we construct tables containing the proportions of observations in each category. For example, the third column in Table 1 provides the proportions of poinsettia plants in each color category.

Describing quantitative data. For quantitative data, we often start by calculating the average value or **mean** of our sample. This familiar quantity is simply the sum of all the values in the sample divided by the number of observations in our sample (Box 1). The mean is only one of several quantities that roughly tell us where the *center* of our data lies. We call these quantities **measures of center**. Other commonly used measures of center are the **median**—the value that literally lies in the middle of the sample—and the **mode**—the most frequent value in the sample.

Box 1 Descriptive Statistics for Quantitative Data

Notation

$x_1, x_2, \ldots, x_n$ are the n observations in the sample.

$\sum_{i=1}^{n} x_i = x_1 + x_2 + \cdots + x_n$ is the sum of all of the observations.

Equations

Mean: $\bar{x} = \dfrac{\sum_{i=1}^{n} x_i}{n}$

Standard deviation: $s = \sqrt{\dfrac{\sum (y_i - \bar{y})^2}{n-1}}$

Correlation coefficient: $r = \dfrac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}}$

Least-squares regression line: $Y = b_0 + b_1 X$

where $b_1 = \dfrac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2}$ and $b_0 = \bar{y} - b_1 \bar{x}$

It is often just as important to quantify the variation in the data as it is to calculate its center. There are several statistics that tell us how much the values differ from one another. We call these **measures of dispersion**. The easiest to one understand and calculate is the **range**, which is simply the largest value in the sample minus the smallest value. The most commonly used measure of dispersion is the **standard deviation**, which calculates the extent to which the data are spread out from the mean. A deviation is the difference between an observation and the mean of the sample, and the standard deviation is a number that summarizes all of the deviations. Two samples can have the same range, but very different standard deviations if one is clustered closer to the mean than the other. In Figure 6, for example, the left sample has a lower standard deviation (s = 2.6) than the right sample (s = 3.6), although the two samples have the same means and ranges.

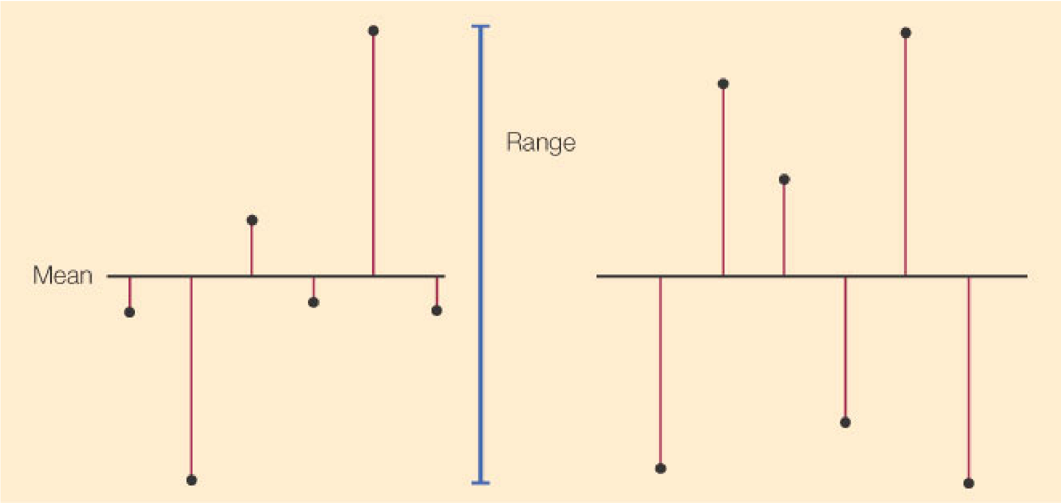


Figure 6 Measures of dispersion. Two samples with the same mean (black horizontal lines) and range (blue vertical line). Red lines show the deviations of each observation from the mean. Samples with large deviations have large standard deviations. The left sample has a smaller standard deviation than the right sample.

To demonstrate these descriptive statistics, we return to the Lake Laengelmavesi study. The researchers also caught and recorded the weights of six fish in the species *Leusiscus idus*: 270, 270, 306, 540, 800, and 1000 grams. The mean weight in this sample is:

$$\bar{x} = \frac{(270+270+306+540+800+1000)}{6} = 531\,.$$

Since there is an even number of observations in the sample, then the median weight is the value halfway between the two middle values: $\frac{306+540}{2} = 423$. The mode of the sample is 270, the only value that appears more than once. The standard deviation is:

$$s = \sqrt{\frac{(270-531)^2 + (270-531)^2 + (306-531)^2 + (540-531)^2 + (800-531)^2\,(1000-531)^2}{5}} = 309.6$$

and the range is 1000 – 270 = 730.

Describing the relationship between two quantitative variables. Biologists are often interested in understanding the relationship between two different quantitative variables: How does the height of an organism relate to its weight? How does air pollution relate to the prevalence of asthma? How does biodiversity relate to temperature? Recall that scatter plots visually represent such relationships.

We can quantify the strength of the relationship between two quantitative variables using a single value called the Pearson product–moment **correlation coefficient** (see Box 1). This statistic ranges between –1 and 1, and tells us how closely the points in a scatter plot conform to a straight line. A negative correlation coefficient indicates that one variable decreases as the other increases; a positive correlation coefficient indicates that the two variables increase together, and a correlation coefficient of zero indicates that there is no linear relationship between the two variables (Figure 7).

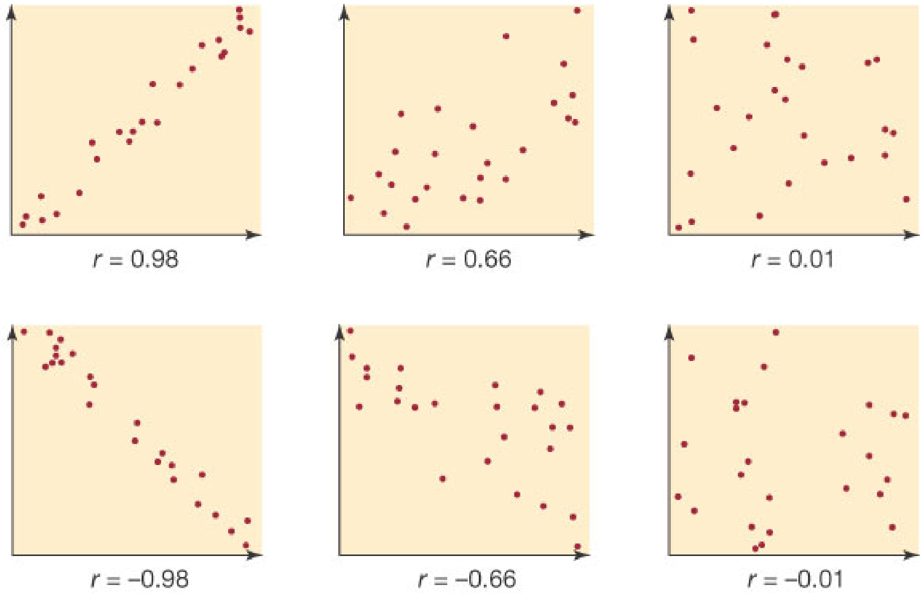


Figure 7 Correlation coefficients. The correlation coefficient (*r*) indicates both the strength and the direction of the relationship.

One must always keep in mind that *correlation does not mean causation*. Two variables can be closely related without one causing the other. For example, the number of cavities in a child's mouth correlates positively with the size of their feet. Clearly cavities do not enhance foot growth; nor does foot growth cause tooth decay. Instead the correlation exists because both quantities tend to increase with age.

Intuitively, the straight line that tracks the cluster of points on a scatter plot tells us something about the *typical* relationship between the two variables. Statisticians do not, however, simply eyeball the data and draw a line by hand. They often use a method called least-squares **linear regression** to fit a straight line to the data (see Box 1). This method calculates the line that minimizes the overall vertical distances between the points in the scatter plot and the line itself. These distances are called **residuals** (Figure 8).

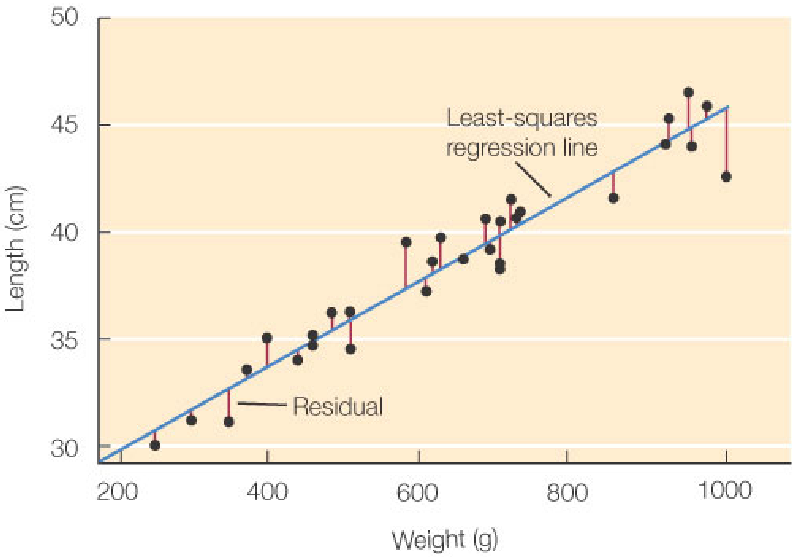


Figure 8 Linear-least squares regression line for *Abramis brama* weights and lengths (measured from nose to end of tail). The regression line (blue line) is given by the equation $y = 26.1 + 0.02x$. It is the line that minimizes the sum of the squares of the residuals (red lines).

Step 5. Inferential Statistics

Data analysis often culminates with statistical inference—an attempt to draw general conclusions about the system under investigation. As depicted in Figure 1, the primary reason we collect data is to gain insight into the larger system from which the data are collected. When we test a new medulloblastoma brain cancer drug on ten patients, we do not simply want to know the fate of those ten individuals; rather, we hope to predict its efficacy on the much larger group of all medulloblastoma patients.

Statistical hypotheses. When it comes to inferring something about the real world from our data, we often have a "*Whether or not*" question in mind. For example, we would like to know whether or not global warming impacts biodiversity; whether or not the clutch size of a spider increases with body size; or whether or not soil nitrogen increases the growth of a particular plant species.

Before making statistical inferences from data, we must formalize our "*Whether or not*" question into a pair of opposing hypotheses—a **null hypothesis** (denoted H_0) and an **alternative hypothesis** (denoted H_A). The alternative hypothesis is the *"Whether"*—it is formulated to describe the effect that we expect our data to support; the null hypothesis is the "*or not*"—it is formulated to represent the absence of the effect. In other words, we typically conduct our experiment seeking to demonstrate something new (the alternative hypothesis) and thereby reject idea that it does not occur (the null hypothesis).

Suppose, for example, we would like to know *whether or not* a new vaccine is more effective than an existing vaccine at immunizing children against *Haemophilus influenzae* type b (Hib). Our hypotheses would be as follows:

H_0: The new vaccine is not more effective than the old vaccine.
H_A: The new vaccine is more effective than the old vaccine.

If we would like to know whether radiation increases the mutation rate in the bacteria *Escherichia coli*, we would set up the following hypotheses:

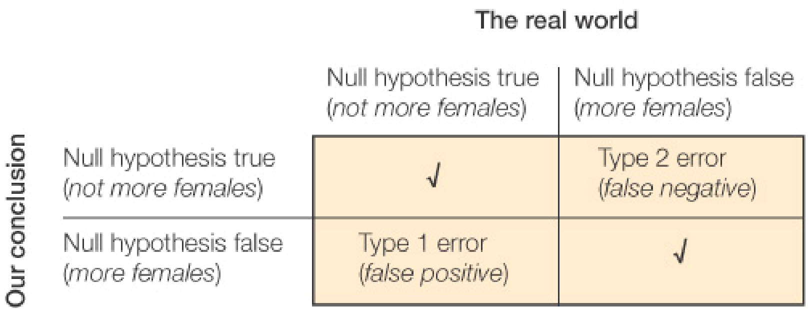
H_0: Radiation does not increase the mutation rate of *E. coli*.
H_A: Radiation does increase the mutation rate of *E. coli*.

Statistical burden of proof. In the U.S. justice system, people are innocent until proven guilty. In statistics, the world is *null until proven alternative*. Statistics requires overwhelming proof in favor of the alternative hypothesis before rejecting the null hypothesis. In other words, scientists favor existing ideas and resist adopting new ideas until compelling evidence suggests otherwise. This is based on a philosophy that it is worse to accept new claims when they are false than to miss out on discovering some true facts about world.

When testing a new Hib vaccine, the burden of proof is on the new vaccine. Suppose we were to vaccinate three children with the new vaccine (Group A), three with the old vaccine (Group B) and leave three children unvaccinated (Group C). If no children from Group A, one child from Group B, and one child from Group C became infected, would we have enough evidence to conclude that the new vaccine is superior to the old vaccine? No, we would not. If the study were enlarged, and two out of 100 children in group A, seven out of 100 children in group B, and 22 out of 100 children in group C become infected, would we then have sufficient evidence to choose the new vaccine? Perhaps, but we need to use statistics to be sure.

This is the traditional burden of proof in biology and science in general. As a consequence, scientists are more likely to miss out on discovering something new (and true) about the world than they are to make a false discovery. In recent years, scientists have begun to question this approach and develop an alternative statistical approach, called **Bayesian inference**, which makes it easier to favor new hypotheses. In this primer, we discuss only traditional statistical methods, often called **frequentist statistics**.

Jumping to the wrong conclusions. There are two ways that a statistical test can go wrong (Figure 9). We can reject the null hypothesis when it is actually true (**Type I error**) or we can accept the null hypothesis when it is actually false (**Type II error**). These kinds of errors are analogous to false positives and false negatives in medical testing, respectively. If we mistakenly reject the null hypothesis when it is actually true, then we falsely endorse the incorrect hypothesis. If we are unable to reject the null hypothesis when it is actually false, then we fail to realize a yet undiscovered truth.



		The real world	
		Null hypothesis true *(not more females)*	Null hypothesis false *(more females)*
Our conclusion	Null hypothesis true *(not more females)*	√	Type 2 error *(false negative)*
	Null hypothesis false *(more females)*	Type 1 error *(false positive)*	√

Figure 9 Possible outcomes of a statistical test. Statistical inference can result in correct and incorrect conclusions about the population of interest.

Suppose we would like to know whether there are more females than males in a population of 10,000 individuals. To determine the makeup of the population, we choose 20 individuals randomly and record their sex. Our null hypothesis is that there are *not* more females than males; and our alternative hypothesis is that there are. The following scenarios illustrate the possible mistakes we might make:

Scenario 1: The population actually has 40% females and 60% males. While our random sample of 20 people is likely to be dominated by males, it is certainly possible that, by chance, we will end up choosing more females than males. If this occurs, and we mistakenly reject the null hypothesis (that there are *not* more females than males), then we make a Type I error.

Scenario 2: The population actually has 60% females and 40% males. If, by chance, we end up with a majority of males in our sample and thus fail reject the null hypothesis, then we make a Type II error.

Fortunately, statistics has been developed precisely to avoid these kinds of errors and inform us about the reliability of our conclusions. The methods are based on calculating the **probabilities** of different possible outcomes. Although you may have heard or even used the word "probability" on multiple occasions, it is important that you understand its mathematical meaning. A probability is a numerical quantity that expresses the likelihood of some event. It ranges between zero and one; zero means that there is no chance the event will occur and one means that the event is guaranteed to occur. This only makes sense if there is an element of chance, that is, if it is possible the event will occur and possible that it will not occur. For example, when we flip a fair coin, it will land on heads with probability 0.5 and land on tails with probability 0.5. When we select individuals randomly from a population with 60% females and 40% males, we will encounter a female with probability 0.6 and a male with probability 0.4.

Probability plays a very important role in statistics. To draw conclusions about the real world (the population) from our sample, we first calculate the probability of obtaining our sample if the null hypothesis is true. Specifically, statistical inference is based on answering the following question:

Suppose the null hypothesis is true. What is the probability that a random sample would, by chance, differ from the null hypothesis as much as our sample differs from the null hypothesis?

If our sample is highly improbable under the null hypothesis, then we rule it out in favor of our alternative hypothesis. If, instead, our sample has a reasonable probability of occurring under the null hypothesis, then we conclude that our data are consistent with the null hypothesis and we do not reject it.

Returning to the sex ratio example, we consider two new scenarios:

Scenario 3: Suppose we want to infer whether or not females constitute the majority of the population (our alternative hypothesis) based on a random sample containing 12 females and eight males. We would calculate the probability that a random sample of 20 people includes at least 12 females assuming that the population, in fact, has a 50:50 sex ratio (our null hypothesis). This probability is 0.13, which is too high to rule out the null hypothesis.

Scenario 4: Suppose now that our sample contains 17 females and three males. If our population is truly evenly divided, then this sample is much less likely than the sample in scenario 3. The probability of such an extreme sample is 0.0002, and would lead us to rule out the null hypothesis and conclude that there are more females than males.

This agrees with our intuition. When choosing 20 people randomly from an evenly divided population, we would be surprised if almost all of them were female, but would not be surprised at all if we ended up with a few more females than males (or a few more males than females). Exactly how many females do we need in our sample before we can confidently infer that they make up the majority of the population? And how confident are we when we reach that conclusion? Statistics allows us to answer these questions precisely.

Statistical significance: Avoiding false positives. Whenever we test hypotheses, we calculate the probability just discussed, and refer to this value as the ***p*-value** of our test. Specifically, the *p*-value is the probability of getting data as extreme as our data (just by chance) if the null hypothesis is, in fact, true. In other words, it is the likelihood that chance alone would produce data that differ from the null hypothesis as much as our data differ from the null hypothesis. How we measure the difference between our data and the null hypothesis depends on the kind of data in our sample (categorical or quantitative) and nature of the null hypothesis (assertions about proportions, single variables, multiple variables, differences between variables, correlations between variables, etc.).

For many statistical tests, *p*-values can be calculated mathematically. One option is to quantify the extent to which the data depart from the null hypothesis and then use look-up tables (available in most statistics textbooks) to find the probability that chance alone would produce a difference of that magnitude. Most scientists, however, find *p*-values primarily by using statistical software rather than hand calculations combined with look-up tables. Regardless of the technology, the most important steps of the statistical analysis are still left to the researcher: constructing appropriate null and alternative hypotheses, choosing the correct statistical test, and drawing correct conclusions.

After we calculate a *p*-value from our data, we have to decide whether it is small enough to conclude that our data are inconsistent with the null hypothesis. This is decided by comparing the *p*-value to a threshold called the **significance level**, which is often chosen even before making any calculations. We reject the null hypothesis only when the *p*-value is less than or equal to the significance level, denoted α. This ensures that, if the null hypothesis is true, we have at most a probability α of accidentally rejecting it. Therefore, the lower the value of α, the less likely you are to make a Type I error (lower left cell of Figure 9). The most commonly used significance level is $\alpha = 0.05$, which limits the probability of a Type I error to 5%.

If our statistical test yields a p-value that is less than our significance level α, then we conclude that the effect described by our alternative hypothesis is statistically significant at the level α and we reject the null hypothesis. If our *p*-value is greater than α, then we conclude that we are unable to reject the null hypothesis. In this case, we do not actually reject the alternative hypothesis, rather we conclude that we do not yet have enough evidence to support it.

Power: Avoiding false negatives. The **power** of a statistical test is the probability that we will correctly reject the null hypothesis when it is false (lower right cell of Figure 9). Therefore, the higher the power of the test, the less likely we are to make a Type II error (upper right cell of Figure 9). The power of a test can be calculated, and such calculations can be used to improve your methodology. Generally, there are several steps that can be taken to increase power and thereby avoid false negatives:

- **Decrease the significance level**, α. The higher the value of α, the harder it is to reject the null hypothesis, even if it is actually false.
- **Increase the sample size**. The more data one has, the more likely one is to find evidence against the null hypothesis, if it is actually false.
- **Decrease variability in the sample**. The more variation there is in the sample, the harder it is to discern a clear effect (the alternative hypothesis) when it actually exists.

It is always a good idea to design your experiment to reduce any variability that may obscure the pattern you seek to detect. After you have minimized such extraneous variation, you can use power calculations to choose the right combination of α and sample size to reduce the risks of Type I and Type II errors to desirable levels.

There is a trade-off between Type I and Type II errors: As α increases, the risk of a Type I decreases but the risk of a Type II error increases. As discussed above, scientists tend to be more concerned about Type I errors than Type II errors. That is, they believe that it is worse to mistakenly believe a false hypothesis then it is to fail to make a new discovery. Thus, they prefer to use low values of α. However, there are many real-world scenarios in which it would be worse to make a Type II error than a Type I error. For example, suppose a new cold medication is being tested for dangerous (life-threatening) side effects. The null hypothesis is that there are no such side effects. A Type II error might lead the FDA to approve a harmful medication that could cost human lives. In contrast, a Type I error would simply mean one less cold medication among the many that already line pharmacy shelves. In such cases, policymakers take steps to avoid a Type II error, even if, in doing so, they increase the risk of a Type I error.

Statistical inference with quantitative data. There are many forms of statistical inference for quantitative data. The flow chart in Figure 2 suggests some commonly used statistical tests for several basic experimental designs. When measuring a single quantitative variable, like birth weight in lambs, calcium concentration in the blood of pregnant women, or migration rate of birds, we often wish to infer the mean value of the population from which we drew the sample. However, the mean of a randomly chosen sample will not necessarily be the same or even close to the population mean. Suppose we wanted to know the average weight of newborn lambs on a particular farm. By chance, we may end up with a random sample that includes an excess of lightweight lambs and therefore a sample mean that is less than the overall mean in the population.

To infer the population mean from the sample data, we can calculate a **confidence interval for the mean**. This is a statistically derived range of values that is centered on the sample mean and is likely to include the population mean. For example, based on the sample of 34 *Abramis brama* weights from Lake Laengelmavesi (Table 2; Figure 5), the 95% confidence interval for the mean weight ranges from 554 grams to 698 grams. The true average weight for this species of fish is likely, but not guaranteed, to fall within this range.

Biologists frequently wish to compare the mean values in two or more groups; for example, newborn lamb weights on several different farms, calcium concentration in women in early and late stages of pregnancy, or migration rates in birds of different species. Based on the means and standard deviations calculated for each of the samples, they infer whether or not the means in the different populations are statistically different from one another. There are several statistical methods for this, and the correct method depends on the number of groups, the experimental design, and the nature of the data.

Box 2 describes the steps of a *t*-test, a simple method for comparing the means in two different groups. To illustrate, we can apply a *t*-test to the Lake Laengelmavesi data to assess whether the two fish species *Abramis brama* and *Leusiscus idus* have significantly different mean weights. We begin by stating our hypotheses and choosing a significance level:

$\boldsymbol{H_0}$: *Abramis brama* and *Leusiscus idus* have the same mean weight.
$\boldsymbol{H_A}$: *Abramis brama* and *Leusiscus idus* have different mean weights.
$\alpha = 0.05$

The test statistic is calculated using the means, standard deviations, and sizes of the two samples: $t_s = \frac{626 - 531}{\sqrt{\frac{207^2}{34} + \frac{310^2}{6}}} = 0.724$. Using a statistical software package called R (Team 2004), we find the *p*-value to be $p = 0.497$. Since *p* is considerably greater than α, we fail to reject the null hypothesis and conclude that our study does not provide evidence that the two species have different mean weights.

Box 2 The *t*-test

What is the *t*-test? A standard method for assessing whether the means of two groups are statistically different from each another.

Step 1: State the null and alternative *hypotheses*:

H_0: The two populations have the same mean.
H_A: The two populations have different means.

Step 2: Choose a significance level, α, to limit the risk of a Type 1 error.

Step 3: Calculate the *test statistic*: $t_s = \dfrac{\bar{y}_1 - \bar{y}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$

Notation: $\bar{y}_1$ and $\bar{y}_2$ are the sample means; s_1 and s_2 are the sample standard deviations; and n_1 and n_2 are the sample sizes.

Step 4: Use the test statistic to assess whether the data are consistent with the null hypothesis:

Calculate the *p-value* (p) using statistical software or by hand using statistical tables.

Step 5: *Draw conclusions* from the test:

If $p \le \alpha$, then reject H_0, and conclude that the two means are significantly different.

If $p > \alpha$, then we do not have sufficient evidence to conclude that the means differ.

See an introductory statistics textbook to learn more about confidence intervals, *t*-tests, and the other statistical tests mentioned in Figure 2.

Statistical inference with categorical data. With categorical data, we often wish to infer the distribution of the different categories within the populations from which our samples are drawn. In the simplest case, we have a single categorical variable with two or more categories. If there are just two categories, we can construct a **confidence interval for the proportion** of the population that belongs to one of the two categories. This is a statistically derived range of values that is centered on the sample proportion and is likely to include the population proportion. If there are three or more categories, we can use a **chi-square goodness-of-fit** test to determine whether the distribution of the different categories in the population is consistent with a specific distribution.

Box 3 outlines the steps of a chi-square goodness-of-fit-test. As an example, consider the data described in Table 1. Many plant species have simple Mendelian genetic systems in which parent plants produce progeny with three different colors of flowers in a ratio of 2:1:1. However, a botanist believes that these particular poinsettia plants have a different genetic system that does not produce a 2:1:1 ratio of red, pink, and white plants. A chi-square goodness-of-fit can be used to assess whether or not the data are consistent with this ratio, and thus whether or not this simple genetic explanation is valid. We start by stating our hypotheses and significance level:

H

$\boldsymbol{H_0}$: The progeny of this type of cross have the following probabilities of each flower color: Pr{Red} = .50, Pr{Pink} = .25, Pr{White} = .25.
$\boldsymbol{H_A}$: At least one of the probabilities of H_0 is incorrect.
$\alpha = 0.05$

We next use the probabilities in H_0 and the sample size to calculate the expected frequencies.

	Red	Pink	White
Observed	108	34	40
Expected	(.50)(182) = 91	(.25)(182) = 45.5	(.25)(182) = 45.5

Based on these quantities, we calculate the chi-square test statistic:

$$\chi_s^2 = \sum_{i=1}^{C} \frac{(O_i - E_i)^2}{E_i} = \frac{(108-91)^2}{91} + \frac{(34-45.5)^2}{45.5} + \frac{(40-45.5)^2}{45.5} = 6.747$$

We find the *p*-value to be *p* = 0.0343 using the R statistical software package. Since *p* is less than α, we reject the null hypothesis and conclude that the botanist is correct: The plant color patterns cannot be explained by the simple Mendelian genetic model under consideration.

Box 3 The Chi-Square Goodness-of-Fit Test

What is the chi-square goodness-of-fit test? A standard method for assessing whether a sample came from a population with a specific distribution.

Step 1: State the null and alternative *hypotheses*:

H_0: The population has the specified distribution.
H_A: The population does not have the specified distribution.

Step 2: Choose a significance level, α, to limit the risk of a Type 1 error.

Step 3: Determine the *observed frequency* and *expected frequency* for each category:

The observed frequency of a category is simply the number of observations in the sample of that type.

The expected frequency of a category is the probability of the category specified in H_0 multiplied by the overall sample size.

Step 4: Calculate the *test statistic*: $\chi_s^2 = \sum_{i=1}^{C} \frac{(O_i - E_i)^2}{E_i}$

Notation: C is the total number of categories, O_i is the observed frequency of category i, and E_i is the expected frequency of category i.

Step 5: Use the test statistic to assess whether the data are consistent with the null hypothesis:

Calculate the *p-value* (*p*) using statistical software or by hand using statistical tables.

Step 6: *Draw conclusions* from the test:

If $p \leq \alpha$, then reject H_0, and conclude that the population distribution is significantly different than the distribution specified by H_0.

If $p > \alpha$, then we do not have sufficient evidence to conclude that population has a different distribution.

See an introductory statistics textbook to learn more about the statistical tests mentioned in Figure 2 and other inference methods for categorical data.

References

Brofeldt, P. 1917. Bidrag till kaennedom on fiskbestondet i vaara sjoear. Laengelmaevesi. In *Finlands Fiskeriet Band 4, Meddelanden utgivna av fiskerifoereningen i Finland*, T. H. Jaervi, (ed.), Helsingfors.

Houghton, J., et al. 2001. Climate change 2001: The scientific basis. *In Contributions of Working Group I to the Third Assessment Report of the Intergovernmental Panel on Climate Change*, Cambridge University Press.

R Development Core Team. 2008. *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. Vienna, Austria. <http://www.r-project.org>

Skow, C. and E. Jakob. 2003. Effects of maternal body size on clutch size and egg weight in a pholcid spider (*Holocnemus pluchei*). *The Journal of Arachnology* 31: 305–308.

Stewart, R. N. and T. Arisumi. 1966. Genetic and histogenic determination of pink bract color in poinsettia. *Journal of Heredity* 57: 217–220.