5 Inferential Statistics

The Controlled Experiment, Hypothesis Testing, and the z Distribution

Chapter 5 Goals

- Understand hypothesis testing in controlled experiments
- Understand why the null hypothesis is usually a conservative beginning
- Understand nondirectional and directional alternative hypotheses and their advantages and disadvantages
- Learn the four possible outcomes in hypothesis testing
- Learn the difference between significant and nonsignificant statistical findings
- Learn the fine art of baloney detection
- Learn again why experimental designs are more important than the statistical analyses
- Learn the basics of probability theory, some theorems, and probability distributions

Recently, when I was shopping at the grocery store, I became aware that music was softly playing throughout the store (in this case, the ancient rock group Strawberry Alarm Clock's "Incense and Peppermint"). In a curious mood, I asked the store manager, "Why music?" and "why this type of music?" In a very serious manner, he told me that "studies" had shown people buy more groceries listening to this type of music. Perhaps more

businesses would stay in business if they were more skeptical and fell less for scams that promise "a buying atmosphere." In this chapter on inferential statistics, you will learn how to test hypotheses such as "music makes people buy more," or "HIV does not cause AIDS," or "moving one's eyes back and forth helps to forget traumatic events."

Inferential statistics is concerned with making conclusions about populations from smaller samples drawn from the population. In descriptive statistics, we were primarily concerned with simple descriptions of numbers by graphs, tables, and parameters that summarized sets of numbers such as the mean and standard deviation. In inferential statistics, our primary concern will be testing hypotheses on samples and hoping that these hypotheses, if true of the sample, will be true and generalize to the population. Remember that a population is defined as the mostly hypothetical group to whom we wish to generalize. The population is hypothetical for two reasons: First, we will rarely, if ever, have the time or money, or it will not be feasible to test everyone in the population. Second, we will attempt to generalize from a current sample to future members of the population. For example, if we were able to determine a complete cure for AIDS, we would hope that the cure would not only work for the current population of AIDS patients in the world but also any future AIDS patients.

The most common research designs in inferential statistics are actually very simple: We will test whether two different variables are related to each other (through correlation and the chi-square test) or whether two or more groups treated differently will have different means on a response (or outcome) variable (through *t* tests and analyses of variance). Examples of whether two different variables are related to each other are plentiful throughout science and its many disciplines. We may wish to know whether cigarettes are related to cancer, whether violent crime rates are related to crack cocaine use, whether breast implants are related to immunodeficiency disease, whether twins' IQs are more highly related than siblings' IQs, and so on. Note that finding a relationship between two variables does not mean that the two variables are causally related. However, sometimes determining whether relationships exist between two variables, such as smoking and rates of lung cancer, may give up clues that allow us to set up controlled experiments where causality may be determined. Controlled experiments, typically with two or more groups treated differently, are the most powerful experimental designs in all of statistics. Whereas correlational designs, which determine whether two variables are related, are very common and useful, they pale in comparison to the power of a well-designed experiment with two or more groups.

It is perhaps unfortunate (maybe statisticians should hire a public relations firm) that the most powerful experimental design is simply called a controlled experiment. There are other theories in science that have much better names, such as the big bang theory, which attempts to explain the origins of the universe. Nonetheless, for the present, we are stuck with the name

controlled experiment. While some statisticians might argue which statistical tests best evaluate the outcome of some types of controlled experiments, few statisticians would argue about the powerful nature of the classic two-group controlled experiment.

The **controlled experiment** is a two-group experiment, consisting typically of an experimental group and a control group. In this experimental design, which allows the determination of causality, the independent variable is the factor that the experimenter is manipulating. For example, in a drug effectiveness experiment, the independent variable is the drug itself. One group receives the drug, and the other group receives a placebo. The experimenter then determines whether the drug has an effect on some outcome measure, response variable, or, as it is also known, the dependent variable. The dependent variable is the behavior, which is measured or observed to change as a function of the two levels of the independent variable (drug group and the placebo group). The experimenter wants to see if the independent variable changes the dependent variable. Some statisticians have compared this experimental process to signal detection theory. If a treatment really works, then the two groups treated differently should score differently on the dependent variable or response variable, and this difference is the *signal*. If the treatment does not work at all, then the two groups should score similarly on the response variable. However, due to random errors or pure chance, it is highly unlikely that the two groups (even if the drug does not work any differently than a placebo) will have exactly the same means on the response variable. If the independent variable or treatment does not work, the two groups' means should be close but not exactly equal. This difference between the two means is attributed to chance or random error, and it is called *noise*. Thus, inferential statistics can be likened to the **signal-to-noise ratio.** If the independent variable really works, then the signal should be much greater than the noise. If the independent variable does not work, then the signal will not exceed the background noise.

Hypothesis Testing in the Controlled Experiment

A hypothesis is an educated guess about some state of affairs. In the scientific world, researchers are usually conservative about their results, and they assume that nothing has been demonstrated unless the results (signal) can be clearly distinguished from chance or random error (noise). Usually, experiments are conducted with a research idea or hunch, which is typically called the **research hypothesis.** The research hypothesis is usually what the experimenter believes to be true. Despite this belief, however, in theory, all experiments are begun with a statement called the **null hypothesis** (abbreviated H_0), which states that there is no relationship between the independent variable and the dependent or response variable in the population.

In the drug effectiveness experiment, the null hypothesis would be that the drug has no effect on the dependent variable. Thus, frequently, the null

hypothesis will be the opposite of what the scientist believes or hopes to be true. The prior research hunch or belief about what is true is called the **alternative hypothesis** (abbreviated H_a).

As noted earlier in the book, science must work slowly and conservatively. The repercussions of poorly performed science are deadly or even worse. Thus, the null hypothesis is usually a safe, conservative position, which says that there is no relationship between the variables or, in the case of the drug experiment, that the drug does not affect the experimental group differently on the dependent variable compared to the control group.

Hypothesis Testing: The Big Decision

All experiments begin with the statement of the null and alternative hypotheses (at least in the experimenter's mind, but not usually in the published article). However, the null hypothesis is like a default position: We will retain the null hypothesis (or we will fail to reject the null hypothesis) unless our statistical test tells us to do otherwise. If there is no statistical difference between the two means, then the null hypothesis is retained. If the statistical test determines that there is a difference between the means (beyond just chance differences), then the null hypothesis will be rejected.

In summary, when a statistical test is employed, one of two possible decisions must be made: (a) retain the null hypothesis, which means that there are no differences between the two means other than chance differences, or (b) reject the null hypothesis, which means that the means are different from each other well beyond what would be expected by chance.

How the Big Decision Is Made: Back to the *z* Distribution

A statistical test of the classic two-group experiment will analyze the difference between the two means to determine whether the observed difference could have occurred by chance alone. The *z* distribution, or a similar distribution, will be used to make the decision to retain or reject the null hypothesis.

To appreciate how this occurs, imagine a large vat of 10,000 ping-pong balls (see Figure 5.1).

Let us suppose that each ping-pong ball has a *z* score written on it. Each *z* score on a ball occurs with the same frequency as in the *z* distribution. Remember that the *z* distribution reveals that exactly 68.26% of the 10,000 balls will fall within ±1 standard deviation of the mean *z* score of 0.00. This means that 6,826 of the 10,000 ping-pong balls will have numbers ranging from −1.00 to +1.00. Also, 95.44% of all the balls will fall within ±2 standard deviations of the mean. Therefore, 9,544 of the ping-pong balls will range between −2.00 and +2.00. Finally, we know that 9,974 ping-pong balls will be numbered from −3.00 to +3.00.

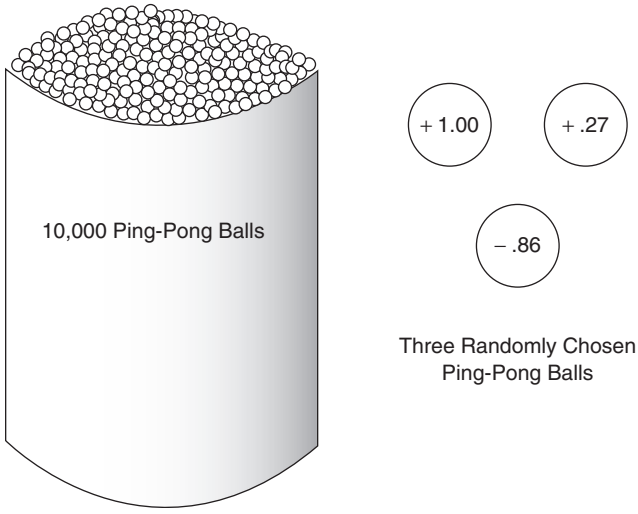

Figure 5.1 A Vat of 10,000 Ping-Pong Balls, Each With a Single Value of *z*, Occurring With the Same Frequency as in the *z* Distribution

Now, let us play a game of chance. If blindfolded and I dig into the vat of balls and pull out one ball in random fashion, what is the probability that it will be a number between –1.00 and +1.00? If I bet you $20 that the number would be greater than +1.00 or less than –1.00, would you take my bet? You should take my bet because the probability that the ball has a number between –1.00 and +1.00 is 68.26%. Therefore, you would roughly have a 68% chance of winning, and I would only have a 32% chance of winning.

How about if we up the stakes? I will bet you $100 that a *z* score on a randomly chosen ball is greater than +2.00 or less than –2.00. Would you take this bet? You should (and quickly) because now there is a 95.44% you would win and less than a 5% chance that I would win.

What would happen if we finally decided to play the game officially, and I bet a randomly chosen ball is greater than +3.00 or less than –3.00? You put your money next to my money. A fair and neutral party is chosen to select a ball and is blindfolded. What would be your conclusion if the resulting ball had a +3.80 on it?

There are two possibilities: Either we both have witnessed an extremely unlikely event (only 1 ball out of 10,000 has a +3.80 on it), or something is happening beyond what would be expected by chance alone (namely, that the game is rigged and I am cheating in some unseen way).

Now, let us use this knowledge to understand the big decision (retain or reject the null hypothesis). The decision to retain or reject the null hypothesis will be tied to the *z* distribution. Each of the individual subject's scores in the two-group experiment will be cast into a large and complicated formula, and a single *z*-like number will result. In part, the size of this single *z*-like number will be based on the difference between the two groups' means.

If the two means are far apart, then the *z*-like number will most likely be large. If the two means are very close together (nothing but noise), then the *z*-like number will more likely be small. In other words, the data will be converted to a single number in a distribution similar to the *z* distribution. If this *z*-like value is a large positive or negative value (such as +3.80 or –3.80), then it will be concluded that this is a low-probability event. It is unlikely that what has happened was simply due to chance. In this case, the signal is much greater than the noise. Therefore, we will make the decision to reject the null hypothesis. If the formula yields a value between +1.96 and –1.96, then the null hypothesis will be retained because there is exactly a 95.00% probability by chance alone that the formula will yield a value in that range. See Figure 5.2 for a graphic representation of the *z* distribution and these decisions.

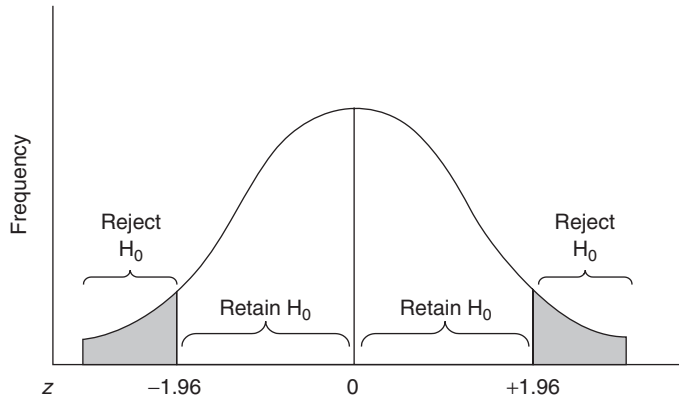

Figure 5.2 *z* Distribution With Retention and Rejection Regions for H_0

The Parameter of Major Interest in Hypothesis Testing: The Mean

In the classic two-group experiment, the means of the two groups are compared on the dependent variable. The null hypothesis states that there is no difference between the two populations' means:

$$H_0: \mu_1 = \mu_2$$

where μ_1 represents the mean for the first population, and μ_2 represents the mean for the second population. Because we will not be using the actual population means, we are going to be making inferences from our sample means, $\bar{x}_1$ and $\bar{x}_2$, to their respective population means. We hope that what we have concluded about the sample is true of the populations.

Another way of thinking about the two means is whether they were both drawn from the same population distribution (in other words, the treatment did not work to make one sample different from another) or whether the two means came from different populations (because the treatment did work on one group and made its mean much larger or much smaller than the other group's mean).

The alternative hypothesis is often what we hope is true in our experiment. The alternative hypothesis is most often stated as

$$H_a: \mu_1 \neq \mu_2$$

Note that the alternative hypothesis is stated as "Mean 1 does not equal Mean 2." This is its most common form, and it is called a **nondirectional alternative hypothesis.** Logically, the "does not equal" sign allows for two possibilities. One possibility is that Mean 1 is greater than Mean 2, and the other is Mean 1 can be less than Mean 2.

Because the controlled experiment involves making inferences about populations, the analysis of the experiment involves inferential statistics. Thus, the mean is an essential parameter in both descriptive and inferential statistics.

Nondirectional and Directional Alternative Hypotheses

An experimenter has a choice between two types of alternative hypotheses when hypothesis testing, a nondirectional or a directional alternative hypothesis. A **directional alternative hypothesis,** in the two-group experiment, states the explicit results of the difference between the two means. For example, one alternative hypothesis could be

$$H_a: \mu_1 > \mu_2$$

Here, the experimenter predicts that the mean for Group 1 will be higher than the mean for Group 2. Another possibility is that the experimenter predicts

$$H_a: \mu_1 < \mu_2$$

Here, the experimenter predicts that Mean 1 will be less than Mean 2. In practice, however, most statisticians choose a nondirectional alternative hypothesis. One of the reasons for this is that the nondirectional alternative hypothesis is less influenced by chance. Directional alternative hypotheses, however, are not all bad. They are more sensitive to small but real differences between the two groups' means. Most statisticians agree that the directional alternative hypothesis should be reserved for situations where the

experimenter is relatively certain of the outcome. It is legitimate to wonder, however, why the experimenter was conducting the experiment in the first place, if he or she was so certain of the outcome.

A Debate: Retain the Null Hypothesis or Fail to Reject the Null Hypothesis

Remember that the classic two-group experiment begins with the statement of the null and the alternative hypotheses. Some statisticians are concerned about the wording of the decision that is to be made. Some say, "The null hypothesis was retained." Others insist that it should be worded, "The null hypothesis was not rejected." Although it may seem to be a trivial point, it has important implications for the entire meaning of the experiment.

After an experiment has been performed and statistically analyzed, and the null hypothesis was retained (or we failed to reject it), what is the overall conclusion? Does it really mean that your chosen independent variable has no effect whatsoever on your chosen dependent variable? Under any circumstances? With any kind of subjects? No! The conclusion is really limited to this particular sample of subjects. Perhaps the null hypothesis was retained because your sample of subjects (although it was randomly chosen) acted differently from another or larger sample of subjects.

There are other possibilities for why the null hypothesis might have been retained besides the sample of subjects. Suppose that your chosen independent variable does affect your subjects but you chose the wrong dependent variable. One famous example of this type of error was in studies of the effectiveness of Vitamin C against the common cold. Initial studies chose the dependent variable to be the number of new colds per time period (e.g., per year). In this case, the null hypothesis was retained. Does this mean that Vitamin C has no effect on the common cold? No! When the dependent variable was the number of days sick within a given time period, the null hypothesis was rejected, and it was preliminarily concluded that Vitamin C appears to reduce the number of days that people are sick with a cold.

It is important to remember that just because you do not find an effect does not mean it does not exist. You might be looking in the wrong place (using the wrong subjects, using the wrong experimental design) and/or you might be using the wrong dependent variable to measure the effect.

Thus, some statisticians recommend that it be stated, "The null hypothesis was not rejected." This variation of the statement has the connotation that there still may be a significant effect somewhere, but it just was not found this time. More important, it has the connotation that, although the null hypothesis was retained, it is not necessarily being endorsed as true. Again, this reflects the conservative nature of most statisticians.

The Null Hypothesis as a Nonconservative Beginning

The null hypothesis, however, is not always a conservative position. As in the case of side effects of many prescription drugs, the null hypothesis is that there are no side effects of the drugs! To correct for this unusual and nonconservative position, the experimenter might increase the regular dosage to exceptionally high levels. If no harmful side effects were observed at high levels, then it might be preliminarily concluded that the drug is safe. Perhaps you have heard that grilled steak fat contains known carcinogenic agents. The beef industry is quick to point out that the experiments to test this hypothesis use animals and levels of grilled fat that would be the equivalent of more than 100 steaks per day for a human being. However, in defense of statisticians and because of their conservative nature, they would be willing to conclude that steak fat does not cause cancer if the equivalent of 100 steaks per day did not cause cancer in experimental animals. Thus, if 100 grilled steaks per day did not seem to cause cancer, then it would be a relatively safe assumption that grilled steak fat is not carcinogenic.

Even in cases where high levels are shown to be safe, scientists are still conservative. Scientists will typically call for a **replication** of the experiment, which means that the experiment will be performed again by another experimenter in a different setting. As noted previously, in science, it is said that one cannot prove a hypothesis to be true or false. Even after high dosages are shown to have no side effects and after repeated experiments with other dosage levels, the drug's safety still has not been proven. Successful replication simply lends additional weight to the hypothesis that the drug is safe. It may still be found that the drug is not safe under other conditions or for other types of people. Thus, replication is important because it generally involves manipulations of other independent variables such as the types of subjects, their ages, and so on.

The Four Possible Outcomes in Hypothesis Testing

There are four possible outcomes in hypothesis testing: two correct decisions and two types of error.

1. Correct Decision: Retain H_0, When H_0 Is Actually True

In this case, we have made a correct decision. In an example involving the relationship between two variables, the H_0 would be that there is no

relationship between the two variables. A statistical test (such as correlation) is performed on the data from a sample, and it is concluded that any relationship that is observed is due to chance. In this case, we retain H_0 and infer that there is no relationship between these two variables in the population from which the sample was drawn. In reality, we do not know whether H_0 is true. However, if it is true for the population and we retain H_0 for the sample, then we have made a correct decision.

2. Type I Error: Reject H_0, When H_0 Is Actually True

The **Type I error** is considered to be the more dangerous of the two types of errors in hypothesis testing. When researchers commit a Type I error, they are claiming that their research hypothesis is true when it really is not true. This is considered to be a serious error because it misleads people. Imagine, for example, a new drug for the cure of AIDS. A researcher who commits a Type I error is claiming that the new drug works when it really does not work. People with AIDS are being given false hopes, and resources that should be spent on a drug that really works will be spent on this bogus drug. The probability of committing a Type I error should be less than 5 chances out of 100 or $p < .05$. The probability of committing a Type I error is also called **alpha (α)**.

3. Correct Decision: Reject H_0, When H_0 Is Actually False

In this case, we have concluded that there is a real relationship between the two variables, and it is probably not due to chance (or that there is a very small probability that our results may be attributed to chance). Therefore, we reject H_0 and assume that there is a relationship between these two variables in the population. If in the population there is a real relationship between the two variables, then by rejecting H_0, we have made the correct decision.

4. Type II Error: Retain H_0, When H_0 Is Actually False

A **Type II error** occurs when a researcher claims that a drug does not work when, in reality, it does work. This is not considered to be as serious an error as the Type I error. Researchers may not ever discover anything new or become famous if they frequently commit Type II errors, but at least they have not misled the public and other researchers. The probability of a Type II error is also called **beta (β)**.

A summary of these decisions appears in Table 5.1.

Table 5.1

Our Decision	*In Reality*	*The Result*
Retain H_0	H_0 is true	Correct decision
Reject H_0	H_0 is true	Type I error (alpha = α)
Reject H_0	H_0 is false	Correct decision
Retain H_0	H_0 is false	Type II error (beta = β)

Significance Levels

A test of **significance** is used to determine whether we retain or reject H_0. The significance test will result in a final test statistic or some single number. If this number is small, then it is more likely that our results are due to chance, and we will retain H_0. If this number is large, then we will reject H_0 and conclude that there is a very small probability that our results are due to chance. The minimum conventional level of significance is p or $\alpha = .05$. This final test statistic is compared to a distribution of numbers, which are called critical values. The test statistic must exceed the critical value in order to reject H_0.

Significant and Nonsignificant Findings

When significant findings have been reported in an experiment, it means that the null hypothesis has been rejected. The word **nonsignificant** is the opposite of significant. When the word *nonsignificant* appears, it means that the null hypothesis has been retained. Do not use the word **insignificant** to report nonsignificant statistical findings. Insignificant is a value judgment, and it has no place in the statistical analysis section of a paper.

In the results section of a research paper, significant findings are reported if the data meet an alpha level of .05 or less. If the findings are significant, it is a statistical convention to report them significant at the lowest alpha level possible. Thus, although H_0 is rejected at the .05 level (or less), researchers will check to see if their results are significant at the .01 or .001 alpha levels. It appears more impressive if a researcher can conclude that the probability that his or her findings are due to chance is $p < .01$ or $p < .001$. It is important to note that this does not mean that results with alphas at .01 or .001 are any more important or meaningful than results reported at the .05 level.

Some statisticians also object to reporting results that are "highly significant." By this, they mean that their findings were significant not only at $p < .05$ but also at $p < .001$. These statisticians would argue that the null hypothesis is rejected at .05, and thus one's job is simply to report the lowest significance possible (e.g., $p < .01$ or $p < .001$). They find it inappropriate, therefore, to use the word *highly* before the word *significant*.

Trends, and Does God Really Love the .05 Level of Significance More Than the .06 Level?

Sometimes, researchers will report "trends" in their data. This usually means that they did not reject H_0 but that they came close to doing so. For example, computers do many of the popular statistics, and they commonly print out the exact alpha levels associated with the test statistic. A **trend** in the data may mean that the test statistic did not exceed the critical value at the .05 level, but the findings may be associated with an alpha of .06 or .10. In these cases, a researcher might say, "The findings approached significance." However, the American Psychological Association publication manual officially discourages reports of trends (American Psychological Association, 2001). The manual claims that if results do not meet the .05 level of significance, then they are to be interpreted as chance findings.

The decision to reject or retain the null hypothesis has been called dichotomous significance testing. Apparently, the need for dichotomous significance testing grew out of the early history of statistics, which developed in agriculture. Many of the statistical questions that an agriculturist might ask would be dichotomous in nature, for example, "Is the manure effective?" It is easy to see in this example how a yes or no answer is appropriate and practical. However, it has been noted, particularly in psychology, that dichotomous significance testing has no clear or early theoretical basis. Thus, two contemporary theoreticians have said "that surely, God loves the .06 nearly as much as the .05. Can there be any doubt that God views the strength of evidence for or against the null as a fairly continuous function of the magnitude of *p*?" (Rosnow & Rosenthal, 1989).

Directional or Nondirectional Alternative Hypotheses: Advantages and Disadvantages

Most statisticians use nondirectional alternative hypotheses. The advantage of the nondirectional alternative hypothesis is that it is less sensitive to chance differences in the data. Thus, the null hypothesis is less likely to be rejected, and a Type I error is less likely to be committed if a nondirectional alternative hypothesis is used. Because the Type I error is considered more serious than the Type II error, the nondirectional alternative hypothesis is more attractive to statisticians.

However, the nondirectional alternative hypothesis has one major disadvantage: It is less sensitive to real differences in the data compared to the directional alternative hypothesis. For example, in testing to see whether two groups' means are significantly different from each other, if there is a real difference but not a great difference, the nondirectional alternative hypothesis is less sensitive to this small but real difference between means.

Thus, it also follows that the directional alternative hypothesis has the advantage that it is more sensitive to real differences in the data. In other words, if there is a real difference between two groups' means, it is more likely to be detected with a directional alternative hypothesis. However, its major disadvantage is that it is also more sensitive to just chance differences between two groups' means.

Did Nuclear Fusion Occur?

In 1989, two chemists claimed that they produced nuclear fusion in a laboratory under "cold" conditions; that is, they claimed to have produced a vast amount of energy by fusing atoms and without having to provide large amounts of energy to do so. Their claims can still be analyzed in the hypothesis testing situation, although it is not absolutely known whether they did or did not produce fusion. However, most subsequent replications of their work were unsuccessful (see Park, 2000, for a fascinating discussion of the controversy).

The null and alternative hypotheses in this situation are as follows:

H_0: Fusion has not been produced.

H_a: Fusion has been produced.

Situation 1. If subsequent research supports their claims, then the two chemists made the correct decision to reject H_0. Thus, they will probably receive the Nobel Prize, and their names will be immortalized.

Situation 2. If subsequent research shows that they did not really produce fusion, then they rejected H_0 when H_0 was true, and thus they committed the grievous Type I error. Why is this a serious error? They may have misled thousands of researchers, and millions of dollars may have been wasted. The money and resources might have been better spent pursuing other lines of research to demonstrate cold fusion (because physicists claim cold fusion is theoretically possible) rather than these chemists' mistake.

What about the quiet researcher who actually did demonstrate a small but real amount of fusion in the laboratory but used a nondirectional alternative hypothesis? The researcher failed to reject H_0 when H_a was true, and thus the researcher committed a Type II error. What was the researcher's name? We do not know. Fame will elude a researcher if there is a continual commission of Type II errors because of an inordinate fear of a Type I error! Remember, sometimes scientists must dare to be wrong.

Baloney Detection

The late astronomer Carl Sagan, in his 1996 book *The Demon-Haunted World: Science as a Candle in the Dark,* proposed a baloney detection kit. The purpose of the kit was to evaluate new ideas. The primary tool in the kit was simply skeptical thinking, that is, to understand an argument and to recognize when it may be fallacious or fraudulent. The baloney detection kit would be exceptionally useful in all aspects of our lives, especially in regards to our health, where sometimes the quest for profit may outweigh the dangers of a product or when the product is an outright fraud. In the traditional natural sciences, the baloney detection kit can help draw boundaries between real science and pseudoscience. Michael Shermer, publisher of *Skeptic* magazine (www.skeptic.com), has modified Sagan's baloney detection kit. Let's use some of Sagan's and Shermer's suggestions to investigate three claims: (a) magician David Copperfield's recent announcement that he predicted Germany's national lottery numbers 7 months before the drawing; (b) mangosteen, a South Asian fruit, cures cancer, diabetes, and a plethora of other diseases and illnesses, and it works as well or better than more than 50 prescription drugs; and (c) therapeutic touch (TT), a therapy in which a medical patient is not actually touched but the patient's negative energy aura is manipulated by a trained TT therapist in order to relieve pain.

How Reliable Is the Source of the Claim?

A corollary of this criterion would be, Does the claimant have a financial (or fame) interest in the outcome? Pseudoscientists may, on the surface, appear to be reliable, but when we examine their facts and figures, they are often distorted, taken out of context, or even fabricated. Often, the claims are merely based on a desire for money and/or fame. Copperfield is a professional magician. He specializes in illusions such as making large jet planes disappear. How reliable is his claim to have predicted lottery numbers in advance? Not very. Would his claim advance his fame (and fortune)? Of course!

The chief promoter of mangosteen is identified as a prominent medical doctor and medical researcher. In reality, the doctor is a Georgia family physician who has not published even a single clinical study in any medical journal. He has written a self-published book on mangosteen, touting near-miraculous cures for a variety of diseases with his patients. We noted earlier in Chapter 1 that books, particularly self-published and those published by commercial presses, have no scientific standards to meet; therefore, they often fail to supply us with any acceptable scientific evidence *whatsoever!* Claiming something is true or saying something is true does not make it so. Mangosteen is being marketed for $37 a bottle. Distributorships are being sold. Mangosteen's proponents are clearly interested in financial gain. The latter is not a heinous crime, but it becomes one if its proponents know there are no clinical studies with humans that support their outlandish claims.

TT was developed by a nursing professor and a "fifth-generation sensitive" in the 1970s. While the nursing professor's academic credentials are credible, a self-proclaimed "sensitive" is someone who can perceive "energies" not normally perceived by other people. Thus, TT claims are initially tainted by their association with a less than scientific credible source. This, of course, does not automatically make all TT claims false, but the claims would have been more scientifically credible had they been made by another medical practitioner, scientist, or professor. In addition, those who now control the training of TT therapists do so for profit. They have a strong financial interest in maintaining their claims about TT.

Does This Source Often Make Similar Claims?

Although this may be the first time Copperfield claims to have predicted the winning numbers in a lottery, he has made many similar claims in the past. It becomes very difficult to believe in his present boast when his entire life consists of making magic and creating illusions. One of the current leading sellers of mangosteen was previously involved in the selling (multilevel marketing) and excessive hype of another "miracle" juice, which was supposedly from Tahiti. With regard to TT therapy, Dora Kunz had already claimed to be a "sensitive." Thus, she had already claimed to have extrasensory powers. Such previous claims seriously diminish her stature as a reliable expert with regard to TT.

Have the Claims Been Verified by Another Source?

Pseudoscientists often do not have their claims verified by anyone other than themselves or close followers. With regard to Copperfield, he said he "predicted the numbers 7 months ago." It would have been a much more impressive demonstration if he had produced the winning ticket 7 months earlier. In this case, Copperfield simply announced *after the lottery* that he had predicted these numbers 7 months earlier. He offered no tangible proof. His only "proof" was his claim! In addition, it is probably dangerous for any scientist or statistician without training in illusions or magic to investigate Copperfield's claim. Magicians and illusionists are typically very good at what they do, and a scientist or statistician, no matter how sophisticated their scientific knowledge, could be as easily fooled as anyone.

The literature regarding TT follows a similar course. Successful demonstrations of TT (despite severe methodological design flaws) cite other successful demonstrations of TT. Meta-analyses (a summary study of other studies) of TT cannot take into account design flaws. Interestingly, there are meta-analytic studies that have given some credence to TT claims. However, all of the positive TT studies have failed to control for the placebo effect, and thus meta-analytic studies will include these flawed studies in their overall analysis.

If TT therapists were serious about the scientific establishment of TT, they would employ acceptable scientific standards in their research. They would show that the results of TT are not due to the placebo effect. They would demonstrate scientifically that trained TT therapists can detect energy fields. To date, only one published TT study has attempted to determine whether TT therapists can detect energy auras better than chance. That study was published by a 9-year-old girl as a fourth-grade science fair project, and she found that experienced TT therapists could do no better than chance in detecting which hand she held over one of the therapist's hands (when the TT therapists could not see their hands). TT proponents tend to seek out other proponents. They cite research with positive outcomes. They ignore or deny claims to the contrary.

How Does the Claim Fit With Known Natural Scientific Laws?

A corollary of this criterion would be, Does the finding seem too good to be true? Copperfield claims his lottery prediction was not a trick. He said it was more like an experiment or a mental exercise. If it was, how does it fit into any known or replicated scientific principle? It simply does not. We would have to create a new principle to explain his mind/matter experiment or use an old one that is without any scientific merit (such as clairvoyance). There is no accepted scientific principle that explains how one would predict lottery numbers in advance. That is simply too good to be true.

The health claims for mangosteen actually pass this criterion but not its corollary. The fruit does appear to contain known antioxidants called xanthones. Xanthones from mangosteen do appear to have some antibacterial and antiviral properties *in test tubes only!* Where mangosteen fails to live up to its excessive hype is that there has not been one human clinical study to date that has demonstrated that the xanthones in mangosteen have helped or cured a disease.

TT proponents propose that humans have an energy field, which can be detected by other "trained" humans. They propose that imbalances in the patient's energy field cause disease and pain. TT therapists claim they can restore these imbalances by sweeping their hands about 3 inches over the patients' bodies and in order to get rid of their excess negative energy. Does this fit with any scientifically supported natural laws? No. Does it seem too good to be true? Yes.

This is a common ploy in pseudoscience: Concoct exaggerated claims around a kernel of scientific truth. Some fruits (those containing Vitamin C) do appear to aid physical health. Some cancer drugs have been created from plants. But it is not scientifically ethical to claim that mangosteen prevents and cures cancer, as well as lowers cholesterol and prevents heart disease, without acceptable scientific proof, and theoretical proof (i.e., mangosteen

has xanthones, xanthones have antioxidant properties, and antioxidants are thought to aid physical health) is not sufficient. Its power to prevent and cure disease must be demonstrated in empirical studies with humans.

The same is true of TT. For example, there is some evidence that humans can interact with energy fields. For example, have you ever noticed that when straightening an antenna, you can sometimes get better reception when you are holding the antenna? However, it is a severe stretch (and pseudoscientific) to claim humans generate energy fields, that imbalances in these fields cause pain, and that restoring balance by eliminating negative energy is a skill that can be learned.

Sagan noted that we tell children about Santa Claus, the Easter Bunny, and the Tooth Fairy, but we retract these myths before they become adults. However, the desire to believe in something wonderful and magical remains in many adults. Wouldn't it be wonderful if there were super-intelligent, super-nice beings in spaceships visiting the Earth who might give us the secrets to curing cancer and Alzheimer's disease? Wouldn't it be great if we only had to drink 3 ounces of mangosteen twice a day to ward off nearly all diseases and illnesses? Wouldn't it be great if playing a classical CD to a baby boosted his or her IQ? Wouldn't it be amazing if a person could really relieve pain without touching someone else? But let us return to the essential tool in the baloney detection kit—skeptical thinking. If something seems too good to be true, we should probably be even more skeptical than usual. Perhaps we should demand even higher scientific standards of evidence than usual, especially if the claims appear to fall outside known natural laws. It has been said that extraordinary claims should require extraordinary evidence. An extraordinary claim, however, might not always have to provide extraordinary evidence if the evidence for the claim was *ordinary but plentiful*. A preponderance of ordinary evidence will suffice to support the scientific credibility of a theory. Thus, the theory of evolution has no single extraordinary piece of evidence. However, a plethora of studies and observations help to support it overall. I tell my students not to be disappointed when wonderful and magical claims are debunked. There are plenty of real wonders and magic in science yet to be discovered. We do not have to make them up. Francis Crick, Nobel Prize winner for unraveling DNA, reportedly told his mother when he was young that by the time he was older, everything will have been discovered. She is said to have replied, "There'll be plenty left, Ducky."

Can the Claim Be Disproven or Has Only Supportive Evidence Been Sought?

Remember, good scientists are highly skeptical. They would always fear a Type I error, that is, telling people something is true when it is not. Pseudoscientists are not typically skeptical. They believe in what they propose without any doubts that they are wrong. Pseudoscientists typically seek only

confirmatory evidence, and they ignore or even reject vehemently any contradictory evidence. They often seek out only people who support their theories and often castigate and demean those who do not. Good scientists check their claims, check their data, recheck their claims and data, verify their findings, seek replication, and encourage others to replicate their findings.

In the case of mangosteen, test tube studies are certainly the first step in establishing scientific credibility. However, before mangosteen can be touted as a cancer preventative or curative agent, human trials must be conducted.

In a recent study of TT's "qualitative" effectiveness, the results were reported of treating 605 patients who were experiencing discomfort. There was no control condition. Only 1 patient out of the 605 rated the TT treatment as poor. All of the others rated it from excellent (32%), very good (28%), good (28%), or fair (12%). However, the design of this study does not allow for the evaluation of the placebo effect. Are the patients simply responding to some special, caring time that a nurse spends with his or her patient regardless of the treatment? Studies such as these do not advance the scientific credibility of TT. These studies are simply seeking supportive evidence for their claims. Interestingly, the study also reported a "small sampling" of 11 written patient responses, and all were very positive. If the authors were truly interested, as they initially stated, in improving the quality of TT, shouldn't they have interviewed the ones who said it was only fair or poor? How does interviewing only the patients who said positive things about TT improve the quality of TT? It does not. The crux of the problem in TT is the lack of studies that demonstrate that it is working for reasons other than the placebo effect. "Qualitative studies," such as the one previously cited, are not interested in disconfirming evidence. They are only interested in confirming what they already believe.

Do the Claimants' Personal Beliefs and Biases Drive Their Conclusions or Vice Versa?

As Shermer noted, all scientists have personal beliefs and biases, but their conclusions should be driven by the results of their studies. It is hoped that their personal beliefs and biases are formed by the results of their studies. The title of the qualitative TT study was "Large Clinical Study Shows Value of Therapeutic Touch Program." Does it sound like the authors had a preconceived bias as to the outcome of their study? Perhaps the authors' findings helped form their personal beliefs in TT's effectiveness, but I doubt it. Even their initial paragraph cites only studies of the positive effects of TT. It cites no negative outcome studies or any studies that are critical of TT.

Recently, a highly influential religious leader (with a Ph.D. in psychology) claimed that one's sexual orientation was completely one's choice. When asked whether his conclusion was based on his religious beliefs or based on scientific evidence, he firmly stated it was definitely not based on his religious

beliefs but on the lack of even a "shred" of scientific evidence that sexual orientation is biologically determined. Because there is clear and increasing empirical evidence that sexual identity and sexual orientation are highly heritable and biologically based (e.g., Bailey, Pillard, Neale, & Agyei, 1993; Bailey et al., 1999; Bailey, Dunne, & Martin, 2000; Coolidge, Thede, & Young, 2002), it might be concluded that this religious leader is woefully ignorant of such studies, he is unconsciously unaware that his religious beliefs are driving his conclusions, or he is lying about his religious beliefs not biasing his conclusions.

Conclusions About Science and Pseudoscience

As noted earlier, skeptical thinking helps to clear a boundary between science and pseudoscience. As Shermer noted, it is the nature of science to be skeptical yet open-minded and flexible. Thus, sometimes science seems maddeningly slow and even contradictory. Good scientists may even offer potential flaws or findings that would disconfirm their own hypotheses! Good science involves guessing (hypothesizing), testing (experimentation), and retesting (replication). The latter may be the most critical element in the link. Can the results of a particular experiment be duplicated (replication) by other researchers in other locations? There may not always be a very clear boundary between science and pseudoscience, but the application of the skeptical thinking offered by the principles in the baloney detection kit may help to light the way.

The Most Critical Elements in the Detection of Baloney in Suspicious Studies and Fraudulent Claims

In my opinion, there are two most critical elements for detecting baloney in any experiment or claim. The first and most important in the social and medical sciences is, Has the placebo effect been adequately controlled for? For example, in a highly controversial psychotherapeutic technique, eye movement desensitization reprocessing (EMDR), intensively trained therapists teach their patients to move their eyes back and forth while discussing their traumatic experience (see Herbert et al., 2000, for a critical review). Despite calls for studies to control for the placebo effect—in this case, the therapist's very strong belief that the treatment works—there are few, if any, EMDR studies in which the placebo effect has been adequately controlled. In addition, there are obviously *demand characteristics* associated with the delivery of EMDR. **Demand characteristics** are the subtle hints and cues in human interactions (experiments, psychotherapy, etc.) that prompt participants to

act in ways consistent with the beliefs of the experimenter or therapist. Demand characteristics usually operate below one's level of awareness. Psychologist Martin Orne has repeatedly demonstrated that demand characteristics can be very powerful. For example, if a devoted EMDR therapist worked for an hour on your traumatic experience and then asked you how much it helped (with a very kind and expectant facial expression), would you not be at least slightly inclined to say "yes" or "a little bit" even if in reality it did not help at all because you do not wish to disappoint the EMDR therapist? Controlling for the gleam, glow, and religiosity of some devotees of new techniques and the demand characteristics of their methods can be experimentally difficult. However, as Sagan and Shermer have noted, often these questionable studies do not seek evidence to disconfirm their claims, and only supporting evidence is sought. As I have already stated, this is particularly true where strong placebo effects are suspected.

The second most important element of baloney detection for your author is Sagan's and Shermer's fourth principle: How does the claim fit with known natural scientific laws? In the case of EMDR, its rationale relies on physiological and neurological processes such as information processing and eye movements somehow related to rapid eye movement (REM) sleep. Certainly, there is good support for cognitive behavioral models of therapy, and there is a wealth of evidence for REM sleep. However, the direct connection between information-processing models, cognitive behavior techniques, and eye movements in the relief of psychological distress arising from traumatic experiences has not been demonstrated. In each case where I hear of a new technique that seems too good to be true, I find that the scientific or natural explanation for why the technique works is unstated, vague, or questionable. In the case of new psychotherapeutic techniques, I always think that the absence of clear scientific explanations with specifically clear sequences for how the therapy works makes me wonder about how strong the placebo effect plays in the therapy's outcome. In their defense, I will state that it is sometimes difficult to explain how some traditionally accepted therapies, such as psychoanalysis, work. However, that defense is no excuse for not searching for reasonable and scientific explanations for how a new therapy works. *It is also absolutely imperative, in cases where scientific explanations for the therapeutic mechanism are somewhat difficult to demonstrate, that the placebo effects are completely and adequately controlled for and that disconfirming evidence has been sincerely and actively sought.*

Can Statistics Solve Every Problem?

Of course not! In fact, I must warn you that sometimes statistics may even muddle a problem. It has been pointed out that it is not statistics that lie; it is people who lie. Often, when I drive across the country, I listen to "talk radio." I've heard many extremely sophisticated statistical arguments from both sides of the gun control issue. I am always impressed at the way each

side manages to have statistics that absolutely show that guns are safe or guns are dangerous. As a statistician myself, if I get muddled by these statistical arguments, I imagine it must be difficult for almost anyone to see his or her way through them. Thus, I firmly believe that statistics can help us solve most of our problems in most aspects of our lives. However, in some extremely contentious issues (e.g., religious beliefs, gun control, abortion rights, etc.), both proponents and opponents of these issues prepare themselves so well with statistical studies that it will become virtually impossible to make a decision based on statistical evidence. In these situations, I recommend that you vote with your heart, beliefs, or intuition. If you believe that you and your children are safer in your house with a gun, then by all means, keep a gun in your house. If you believe your children are safer without a loaded gun in your house, then do not have one.

Probability

The Lady Tasting Tea

An all-time classic statistics book, *The Design of Experiments,* was written by Sir Ronald A. Fisher (1890–1962) and first published in 1935. It became a standard textbook and reference book for statisticians and their students, and it was continuously published until 1970. Curiously, Chapter 2 begins with the following:

> A Lady declares that by tasting a cup of tea made with milk she can discriminate whether the milk or the tea was first added to the cup. (p. 11)

In his book, Fisher proceeded to design an experimental test of the null hypothesis that the "Lady" cannot discriminate between the two types of tea/milk infusion, and he also demonstrated how to apply a test of significance to the **probability** associated with the null hypothesis. Although Fisher's example appears to be hypothetical, a statistician, Professor H. Fairfield Smith of the University of Connecticut, revealed that Fisher's example was not hypothetical but actually occurred in the late 1920s at an afternoon tea party in Cambridge, England. Apparently, after the Lady voiced her opinion, there was a consensus from the professors in attendance that it was impossible to distinguish whether tea had been added to the milk or the milk was added to the tea. Professor Smith said that he had attended the tea party, and at the moment the Lady made her declaration and the professors voiced their august and oppositional opinions, Ronald Fisher (in his 30s at the time) came forward to suggest there was an applied mathematical analysis (called statistics) and an associated experimental method to discern whether she could or could not taste the difference in the two tea infusions. Fisher proposed that she be told she would taste a series of cups of tea and

voice her opinion after each. In his 1935 book, he gave the example of a series of eight cups of tea, four with milk added to tea and four with tea added to milk. He proposed that the order of the presentation of the infusions be randomized and that the Lady be told that half of the eight cups would contain tea added to milk and half would be milk added to tea.

Probability is the science of predicting future events. The probability of a specified event is defined as the ratio of the number of ways the event can occur to the total number of equally likely events that can occur. Probabilities are usually stated numerically and in decimal fashion, where 1 means that the event will certainly happen, 0 means the event will not happen, and a probability of .5 means that the event is likely to happen once in every two outcomes.

As your intuition might already tell you, if only one cup of tea was presented to the Lady for testing, then the probability of her being right would be ½ or $p = .50$. The probability of her being wrong is also $p = .50$, and the total probability sums to 1.00; that is, after she tastes the tea, the probability of her being right or wrong is 1.00. If we label the tea tasting as an *event* and the Lady's opinion as an *outcome,* then we have established that an event with two equally likely outcomes will have each of those outcomes equal to $p = .50$.

The Definition of the Probability of an Event

If an event has n equally likely outcomes, then each individual outcome will have a probability of $1/n$ or (p_i), where n is the number of equally occurring outcomes. Thus, if a rat has a choice of entering three equally appearing doors in search of food, the probability of one door being chosen is $p_i = .33$ because $p_i = 1/n$ or $p_i = 1/3 = .33$. Also, the overall probability is $p_1 + p_2 + p_3 = 1.00$. For a larger number of events (also called mutually exclusive events), it holds that

$$p_1 + p_2 + \ldots + p_n = 1.00$$

where p_n is the final event.

The Multiplication Theorem of Probability

If there are two independent events (i.e., the outcome of one event has no effect on the next event) and the first event has m possible outcomes, and the second event has n possible outcomes, then there are $m \times n$ total possible outcomes. Now, in the case of the Lady tasting tea, if there are eight events (eight cups of tea in succession), and each event has two outcomes, then her probability of choosing correctly for all eight cups of tea is $1/2^8$ or $½ \times ½ \times ½ \times ½ \times ½ \times ½ \times ½ \times ½ = (½)^8 = p = 1/256$ or $p = .0039$. The latter holds because there are two outcomes (right or wrong) for each event, and there are

eight events (the eight cups of tea); thus, there are 256 total outcomes (2^8), but only one can be correct (identifying all eight cups of tea correctly).

However, remember Fisher said that the Lady should be informed that half of the eight cups would have tea added to milk and half would have milk added to the tea. Thus, the Lady knows that if she is forced to guess, her best strategy would be to make half her guesses one way (e.g., tea added to milk) and half her guesses the other way (e.g., milk added to tea). This reduces her overall probability of being wrong on at least one guess and increases her overall probability of being right on all 8 guesses to 1/70 or $p = .014$ because of the combinations theorem of probability.

Combinations Theorem of Probability

This theorem states that if a selection of objects is to be selected from a group of n objects and the order of the arrangement selected is not important, then the equation for obtaining the number of ways of selecting r objects from n objects is

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

where

C = the total number of combinations of objects,

$n!$ = (is read n factorial or $n(n-1)(n-2)(n-3)$ etc.) so $3! = 3 \times 2 \times 1 = 6$.

Example 1. Thus, for the Lady tasting tea, there were 70 possible outcomes because

$$C = \frac{8!}{4!(8-4)!}$$
$$C = \frac{40320}{24(24)}$$
$$C = 70$$

Fisher had already established the concept of significance and p values in his 1925 book. He began the practice of accepting an experiment's outcome as significant if it could not be produced by chance more frequently than once in 20 trials or $p = .05$. Importantly, he noted in his 1935 book that it was the experimenter's choice to be more or less demanding of the ***p* level** acceptable for demonstrating significance. However, he cautioned that even for the Lady tasting tea, who might correctly identify every tea infusion at a probability of 1 in 70 or .014 (well below the conventional level of significance of $p = .05$),

> no isolated experiment, however significant in itself, can suffice for the experimental demonstration of any natural phenomenon; for the "one chance in a million" will undoubtedly occur, with no less and no more than its appropriate frequency, however surprised we may be that it should occur to *us*. (Fisher, 1935, pp. 13–14)

Example 2. Suppose it is suspected that a combination of two drugs might interact to improve attention and reduce hyperactivity in children with attention-deficit hyperactivity disorder (ADHD). Let us propose that there are four available drugs, and label them A, B, C, and D. How many combinations are there?

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

where

$r = 2$ drug objects from $n = 4$ drugs,
C = the total number of combinations of 2 drugs.

$$C = \frac{4!}{2!(4-2)!}$$
$$C = \frac{4!}{2!(2)!}$$
$$C = \frac{24}{4}$$
$$C = 6$$

Intuitively, you can see they would be AB, AC, AD, BC, BD, and CD.

Permutations Theorem of Probability

This theorem states that if the order of selection of r objects is important from a group of n objects, then

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

where

P = the total number of permutations of objects,
$n!$ = is the factorial number of objects,
r = the number of objects selected.

Note: It is not possible to give a relevant example for the Lady tasting tea because the order in which she gets the teacups correctly identified is not relevant here.

Example 1. Suppose it is suspected that a combination of two drugs might interact to help children with ADHD, but the order of presentation of the drugs might be important. Thus, taking Drug A in the morning and taking Drug B in the afternoon might produce a different response than taking Drug B in the morning and Drug A in the afternoon. If there are four available drugs (A, B, C, and D), how many different drug permutations are there?

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

where

P = the total number of permutations of objects,

$n!$ = is the factorial number of objects,

r = the number of objects.

$$P = \frac{4!}{(4-2)!}$$
$$P = \frac{24}{2}$$
$$P = 12$$

Because there are a limited number of possibilities, we can actually list them all: AB, BA, AC, CA, AD, DA, BC, CB, BD, DB, CD, and DC.

Let us revisit these rules briefly with another intuitive example. A couple wishes to have three children. Assuming that the probability for this couple is the same for having a girl (G) as having a boy (B) (*Note:* This is probably not true for all couples), then there are k^n different outcomes or $2 \times 2 \times 2 = 8$ permutations (GGG, GGB, GBB, GBG, BBB, BBG, BGG, BGB) and four combinations corresponding to the total number of girls, irrespective of order (three girls, two girls, one girl, no girls). Now, let us summarize the probability of these outcomes in a table and a graph (Table 5.2 and Figure 5.3), where order is not important (a summary of combinations and not permutations). The results will be known as a probability distribution.

Table 5.2 Probability Distribution Table

Event	*Fractional Probability*	*Probability*
Three girls	1/8	.125
Two girls, one boy	3/8	.375
One girl, two boys	3/8	.375
Three boys	1/8	.125
Totals	8/8	1.000

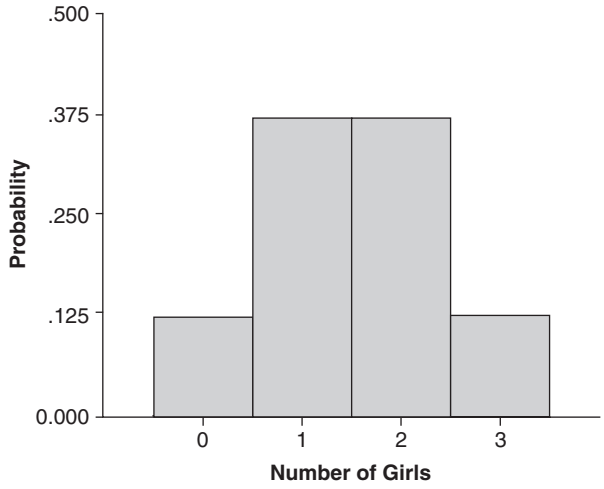

Figure 5.3 Probability Distribution Graph

Probability distributions are the heart of statistical tests of significance. One of the more popular and useful probability distributions is the *binomial distribution.* In this type of distribution, the outcome has only two choices (it is said to be dichotomous), such as boy or girl, heads or tails, or at risk or not at risk. Probability distributions may appear as tables or graphs, as previously noted.

As a final example of the application of probability theory to statistics, who among us has not dreamed of winning a lottery? Certainly, enough states and countries have discovered that lotteries are a consistently powerful means of generating income. Many state lotteries quote that the odds of winning the big grand prize is said to be about 1 in 5 million. How are these probabilities generated?

In my state lottery (Colorado), there are six different balls drawn from numbers 1 to 42. If there was only 1 number drawn (from 1 to 42), then it is easy to see that my odds of winning would be 1 in 42 or about $p = .0238$. If only two balls were drawn, my odds of picking one of the two winning balls with my first number would be 2 in 42 or about $p = .0476$, and now there are only 41 balls left to be drawn, so my odds of picking each of the two balls is the product of the two individual probabilities, that is, $2/42 \times 1/41$, which is $p = .0012$. Now, we can see that my odds of winning have dropped dramatically from about 2 chances in 100 to a little more than 1 chance in 1000. In order to pick six correct numbers (where order does not matter), we will use the following formula for combinations:

$$C = \frac{n!}{r!(n-r)!}$$
$$C = \frac{42!}{6!(42-6)!}$$
$$C = \frac{42!}{6!(36!)}$$
$$C = \frac{42 \cdot 41 \cdot 40 \cdot 39 \cdot 38 \cdot 37 \cdot 36 \cdot 35 \cdot \ldots 2 \cdot 1}{6 \cdot 5 \cdot 4 \cdot 3 \cdot 2 \cdot 1 \cdot 36 \cdot 35 \cdot 34 \cdot 33 \cdot \ldots 2 \cdot 1}$$
$$C = \frac{42 \cdot 41 \cdot 40 \cdot 39 \cdot 38 \cdot 37}{6 \cdot 5 \cdot 4 \cdot 3 \cdot 2 \cdot 1}$$
$$C = \frac{3776965920}{720}$$
$$C = 5245786$$

Thus, my chances of winning are about 1 in 5 million! Of course, I also once got a fortune from a fortune cookie that read, "Without a dream, no dream comes true," so despite these odds, I continue to support Colorado parks (which receive half the losing lottery money each week).

Gambler's Fallacy

A word of caution must be issued. Humans have an innate ability to search for meaning and understanding when an unusual or improbable event takes place. People who win lotteries often attribute their win to special numbers, a higher authority, or being born lucky. They are making a causal connection where none exists. If 10 million or 20 million lottery tickets are sold in Colorado every few days and the odds of a winning ticket are about 1 in 5 million, then eventually we expect that at least one ticket will win. Although the winning ticket is a low-probability event, it will eventually occur due to the law of very large numbers (which, simply stated, means that with a very large sample, any outrageous thing is likely to occur). Gambler's fallacy occurs when people attribute a causal relationship to truly independent events. For example, in tossing a fair coin, if I obtain two heads in a row, some might believe that tails are slightly more likely on the next toss because they are "due." However, the odds of tossing heads or tails on the next toss are equally likely events. The coin has no memory for the previous two tosses. Also, remember obtaining three heads in a row, although it has a probability of only 1/8, is still likely to occur sometime, if I engage in long strings of coin tossing. Probability is about predicting outcomes in a long series of future events.

Coda

The Lady Tasting Tea is also the name of a wonderful book about how statistics revolutionized science in the 20th century (Salsburg, 2001). It is in this book that the story of the Lady tasting tea is recounted by Professor Smith. Fisher did not reveal the outcome of his example in his 1935 book because he presented it as hypothetical, but Professor Smith reported that the Lady correctly identified every single cup of tea!

History Trivia

Egon Pearson to Karl Pearson

Egon S. Pearson (1895–1980) was the son of the famous statistician Karl Pearson. Egon was born and reared in England, and he studied under his father at University College in London in the Department of Applied Statistics. In 1922, Egon began to publish articles that would establish him as an important theoretical figure in modern statistics. In the 1920s, Egon Pearson also began an important collaboration with Jerzy Neyman (1894–1981), who would help shape one of Egon's most important contributions, statistical hypothesis testing.

Neyman grew up in the Ukraine. He studied mathematics early in his college training, and he moved to Poland in 1921 and lectured in mathematics and statistics. He received his doctorate in 1924 at the University of Warsaw. In 1925, he received a fellowship to study statistics in England at the University College with Karl Pearson. Here, Neyman would meet William S. Gosset, who would ultimately make his own unique contributions to statistics, the development of statistical analysis with small samples. Gosset introduced Neyman to Ronald Fisher (who, among his other contributions, developed the analysis of variance). Neyman was also to meet Egon Pearson, who worked as an assistant in his father's statistics laboratory.

Egon Pearson, with the mathematical help of Neyman, further developed the notion of hypothesis testing, including the testing of a simple hypothesis against an alternative; developed the idea of two kinds of errors in hypothesis testing; and proposed the likelihood ratio criterion, which can be used to choose between two alternative hypotheses by comparing probabilities. With respect to the interactions between Gosset, Egon Pearson, and Neyman, it has been said that Gosset asked the question, Egon Pearson put the question into statistical language, and Neyman solved it mathematically.

In 1933, when Karl Pearson retired, the Department of Applied Statistics was split in two. Ronald Fisher became the Galton Professor, succeeding Karl Pearson. Egon Pearson was appointed reader and, later, professor of statistics. In 1936, Egon took over the editorship of the famous journal *Biometrika* which was founded by his

father to study mathematical and statistical contributions to life sciences. He remained editor for 30 years until his own retirement in 1966.

According to Helen Walker (1929/1975), a statistician and statistics historian, the modern history of statistics could be summarized thusly. The first great wave of theoretical contributions came from Francis Galton and Karl Pearson. They promoted the idea that statistical analysis would provide important information, heretofore unknown, about people, plants, and animals. With their far-reaching direction and influence, even medicine and society would be positively changed by the science of statistics. Their contributions also included the invention of measures of association such as the correlation coefficient and chi-square analysis, as well as the construction and publication of tables of statistics that were needed by statisticians and biometricians.

The second wave was begun by Gosset and completed by Ronald Fisher. According to Walker, this period was characterized by the development of statistical methods with small samples, initial development of hypothesis testing, design of experiments, and the development of criteria to choose among statistical tests.

The third wave was led by Egon Pearson and Jerzy Neyman. During their 10-year collaboration, the science of statistics enjoyed an ever-increasing popularity and appreciation. Hypothesis testing was refined, and the logic of statistical inference was developed. The notion of confidence intervals was created, and ideas for dealing with small samples were further advanced and refined.

Although Karl Pearson held some controversial views, his contributions to the philosophy of science and statistics are tremendous. His idea that the scientific method and statistical analysis should be considered part of the "grammar of science" is a watershed in the history of statistics.

Key Terms, Symbols, and Definitions

Alpha (α)—The probability of committing the Type I error. In order to consider findings significant, the probability of alpha must be less than .05.

Alternative hypothesis (H_a)—Most frequently, what the experimenter thinks may be true or wishes to be true before he or she begins an experiment; also called the research hypothesis. It can also be considered the experimenter's hunch.

Beta (β)—The probability of committing the Type II error. A Type II error can occur only when the null hypothesis is false and the experimenter fails to reject the null hypothesis.

Controlled experiment—A two-group experiment, with one group designated as the experimental group and one as the control group. The parameter of statistical interest is the difference between the two groups' means.

Demand characteristics—Subtle hints and cues that guide participants to act in accordance with the experimenter's wishes or expectations.

Directional alternative hypothesis—Also called a one-tailed test of significance where the alternative hypothesis is specifically stated beforehand; for example, Group 1's mean is greater than Group 2's mean.

Insignificant—A value judgment, such as deciding between good and evil, worthless and valuable. It typically has no place in statistics.

Nondirectional alternative hypothesis—Also called a two-tailed test of significance where the null hypothesis will be rejected if either Group 1's mean exceeds Group 2's mean, or vice versa, or where the null hypothesis will be rejected if a relationship exists, regardless of its nature.

Nonsignificant—Findings are considered statistically nonsignificant if the probability that we are wrong is greater than .05. Nonsignificant findings indicate that the null hypothesis has been retained, and the results of the experiment are attributed to chance.

Null hypothesis (H_0)—The starting point in scientific research where the experimenter assumes there is no effect of the treatment or no relationship between two variables.

***p* level—**The probability of committing the Type I error; that is, rejecting H_0 when H_0 is true.

Probability—The science of predicting future events or the likelihood of any given event occurring.

Replication—A series of experiments after an initial study where the series of experiments varies from the initial study in types of subjects, experimental conditions, and so on. Replication should be conducted not only by the initial study's author but also by other scientists who do not have a conflict of interest with the eventual outcome.

Research hypothesis—Also called the alternative hypothesis. It is most frequently what the experimenter thinks may be true or wishes to be true before he or she begins an experiment. It can also be considered the experimenter's hunch.

Signal-to-noise ratio—Borrowed from signal detection theory, in which the effect of a treatment is considered the signal, and random variation in the numbers is considered the noise.

Significance—Findings are considered statistically significant if the probability that we are wrong (where we reject H_0 and H_0 is true) is less than .05. Significant findings indicate that the results of the experiment are substantial and not due to chance.

Trend—Frequently reported when the data do not reach the conventional level of statistical significance (.05) but come close (e.g., .06 or .07). The American Psychological Association's (2001) publication manual officially discourages reports of trends.

Type I error—When an experimenter incorrectly rejects the null hypothesis when it is true.

Type II error—When an experimenter incorrectly retains the null hypothesis when it is false.

Chapter 5 Practice Problems

1. State H_0 and H_a for the following problems:
   a. a new drug cures the AIDS virus
   b. the relationship between drinking milk and longevity
2. Be able to recognize and state the four possible outcomes in hypothesis testing.
3. Be able to state why statisticians prefer NOT to say "insignificant findings," "highly significant," and "the null was retained."

Problems 4–10. True/False

4. A Type II error is considered less serious than a Type I error.
5. The probability of making a Type I error is referred to as alpha.
6. In comparison to a directional alternative hypothesis, a nondirectional alternative hypothesis is more sensitive to real differences in data.
7. The null hypothesis always represents the conservative position.
8. The conventional level of significance is $p < .01$.
9. An experimenter should report the lowest alpha level possible.
10. The probability of a Type II error is called beta.

Chapter 5 Test Questions

1. Which of the following is true about correlational designs?
   a. Finding a relationship between two variables does not mean that one causes the other but could give clues to set up experiments where causality may be determined.
   b. Finding a relationship between two variables NEVER means that one causes the other.
   c. Correlational designs are not very common.
   d. Correlational designs are more powerful than a controlled experiment.
2. An independent variable in an experiment can be likened to the _____ in the signal-to-noise ratio.
   a. signal
   b. to
   c. noise
   d. ratio
3. The difference between two means in a controlled experiment that is just attributed to chance or random error is called _____ in the signal-to-noise ratio.
   a. signal
   b. to
   c. noise
   d. ratio
4. Which of the following is a Type I error?
   a. rejecting H_0 when H_0 is false
   b. rejecting H_0 when H_0 is true
   c. not rejecting H_0 when H_0 is true
   d. not rejecting H_0 when H_0 is false
5. Because it is said in science "one experiment does not prove anything," statisticians rely on _____ to test the usefulness of theories.
   a. a controlled experiment
   b. at least two experiments

c. parametric and nonparametric tests
d. replication

6. The probability of committing the Type I error is also called _____.

a. alpha
b. beta
c. delta
d. omega

7. The probability of committing the Type II error is also called _____.

a. alpha
b. beta
c. delta
d. omega

8. The minimum conventional level of statistical significance is _____.

a. .01
b. .05
c. .10
d. .50

9. When findings in an experiment do not reach the conventional level of statistical significance, they are reported to be _____.

a. nonsignificant
b. insignificant
c. significant
d. unworthy

10. A trend in the data means that the experimenter

a. rejected the null hypothesis at $p < .05$
b. rejected the null hypothesis at $p < .01$
c. did not reject the null hypothesis but came close to doing so
d. none of the above

11. In the section "Trends, and Does God Really Love the .05 Level of Significance More Than the .06 Level?" Rosnow and Rosenthal have argued that the strength of evidence for or against a null hypothesis should be a fairly continuous function of the size of

a. *p* (the significance level)
b. beta
c. the sample size
d. the size of the standard deviations for each group

12. If a magician claims that he can bend spoons with his mind, what did Carl Sagan propose to test the claim?

a. a lie detector test
b. the baloney detection kit
c. the salami detection kit
d. the Atkins detection kit

13. In the section "Can Statistics Solve Every Problem?" your course author argues
    a. yes
    b. no

14. In the classic two-group controlled experiment, what parameter is of central interest between the two groups?
    a. the means
    b. the standard deviations
    c. the sample size
    d. the median

15. Which of the following is more likely to end up being too sensitive to chance differences?
    a. a nondirectional alternative hypothesis
    b. a directional alternative hypothesis
    c. a nondirectional research hypothesis
    d. all of the above