THE GREAT COURSES®

Topic
Science
& Mathematics

Subtopic
Mathematics

Meaning from Data: Statistics Made Clear

Course Guidebook

Professor Michael Starbird
The University of Texas at Austin

PUBLISHED BY:

THE GREAT COURSES
Corporate Headquarters
4840 Westfields Boulevard, Suite 500
Chantilly, Virginia 20151-2299
Phone: 1-800-832-2412
Fax: 703-378-3819
www.thegreatcourses.com

Michael Starbird, Ph.D.

University Distinguished Teaching Professor
of Mathematics
The University of Texas at Austin

Professor Michael Starbird is a professor of mathematics and a University Distinguished Teaching Professor at The University of Texas at Austin. He received his B.A. degree from Pomona College in 1970 and his Ph.D. in mathematics from the University of Wisconsin, Madison, in 1974. That same year, he joined the faculty of the Department of Mathematics of The University of Texas at Austin, where he has stayed except for leaves as a Visiting Member of the Institute for Advanced Study in Princeton, New Jersey; a Visiting Associate Professor at the University of California, San Diego; and a member of the technical staff at the Jet Propulsion Laboratory in Pasadena, California.

Professor Starbird served as Associate Dean in the College of Natural Sciences at The University of Texas at Austin from 1989 to 1997. He is a member of the Academy of Distinguished Teachers at UT and chairs its steering committee. He has won many teaching awards, including a Minnie Stevens Piper Professorship, which is awarded each year to 10 professors from any subject at any college or university in the state of Texas; the inaugural award of the Dad's Association Centennial Teaching Fellowship; the Excellence Award from the Eyes of Texas, twice; the President's Associates Teaching Excellence Award; the Jean Holloway Award for Teaching Excellence, which is the oldest teaching award at UT and is presented to one professor each year; the Chad Oliver Plan II Teaching Award, which is student-selected and awarded each year to one professor in the Plan II liberal arts honors program; and the Friar Society Centennial Teaching Fellowship, which is awarded to one professor at UT annually and includes the largest monetary teaching prize given at UT. Also, in 1989, Professor Starbird was the Recreational Sports Super Racquets Champion.

The professor's mathematical research is in the field of topology. He recently served as a member-at-large of the Council of the American Mathematical Society and currently serves on the national education committees of both the American Mathematical Society and the Mathematical Association of America.

Professor Starbird is interested in bringing authentic understanding of significant ideas in mathematics to people who are not necessarily mathematically oriented. He has developed and taught an acclaimed class that presents higher-level mathematics to liberal arts students. He wrote, with co-author Edward B. Burger, *The Heart of Mathematics: An invitation to effective thinking*, which won a 2001 Robert W. Hamilton Book Award. Professors Burger and Starbird have written a book that brings intriguing mathematical ideas to the public, entitled *Coincidence, Chaos, and All That Math Jazz: Making Light of Weighty Ideas*, published by W.W. Norton, 2005. Professor Starbird has produced two previous courses for The Teaching Company, one entitled *Change and Motion: Calculus Made Clear* and a second one with collaborator Edward Burger entitled *The Joy of Thinking: The Beauty and Power of Classical Mathematical Ideas*. Professor Starbird loves to see real people find the intrigue and fascination that mathematics can bring. ■

Table of Contents

Table of Contents

Table of Contents

Acknowledgments

These lectures were prepared in collaboration with Thomas Starbird, Ph.D., a principal member of the technical staff at the Jet Propulsion Laboratory, Pasadena, California, and Jennifer Kaplan, a Ph.D. student in statistics education at The University of Texas at Austin. We three worked together on the concept, design, and details of the entire course. I would also like to thank the following people for discussions and ideas: Joe Gallian, Terry Kahn, Mary Parker, Tony Petrosino, James Scott, and Ann Watkins. Thanks to Lucinda Robb, Noreen Nelson, Pamela Greer, Alisha Reay, and others from The Great Courses, not only for providing excellent professional work during the production of this series of lectures but also for creating a supportive and enjoyable atmosphere in which to work. Finally, thanks to my family, Roberta, Talley, and Bryn, for their special encouragement. ■

Meaning from Data: Statistics Made Clear

Scope:

A statistical fact: On average, each American has one testicle and one ovary. Should we take cholesterol-lowering medication? Evidence for and against is presented to us in the form of data and statistical conclusions. Should we buy stocks or sell? Much of the information we use to make the decision is based on numerical data. Will it rain tomorrow? Will the real estate market rise or fall? How good a student will Mr. Jones be, if admitted? Should we buy lottery tickets when the jackpot gets really big? Should a coach leave a player in the game when he's in a slump? How can we tell if gender discrimination influenced college admissions procedures? Trying to understand the economy, the weather, school systems, grading, the quality of products, risk, measurements of everything, social trends, marketing, science, and most practical aspects of our world fundamentally involves coming to grips with data.

The trouble with data is that data do not arrive with meaning. Data are value-free and useless or actually misleading until we learn to interpret their meaning appropriately. Statistics provides the conceptual and procedural tools for drawing meaning from data.

Analyzing data correctly is one of the most powerful tools that we have for understanding our world. But it is a two-edged sword. Mark Twain attributed to Benjamin Disraeli perhaps the most famous quip about statistics: "There are three kinds of lies: lies, damned lies, and statistics." But an apt rejoinder is: "It is easy to lie with statistics, but it is easier to lie without them." In this course, we will see the two sides of data—their uses and their misuses.

We will learn basic principles and ideas of statistics and understand how they can bring meaning to data. We will learn about probability and the central role it plays in understanding the meaning of statistics. One of the great ideas of modern quantitative analysis of our world is that the uncertain and the unknown can be described quantitatively. Random events show global

trends in the aggregate, and probability and statistics can help us describe and measure those trends.

We present statistics by isolating two major challenges: (1) How can we describe and draw meaning from a collection of data when we know all the pertinent data? (2) How can we infer information about the whole population when we know data about only some of the population (a sample)? These two questions form the structural backbone of our approach.

The challenge of describing a collection of data when we know all the data arises, for example, when we have complete records of all students who have ever attended a given university. We know the incoming Scholastic Aptitude Test (SAT) scores and high school class rank of all students, and we know their grade point averages (GPAs) on graduation. We can ask and answer many questions regarding those data. Perhaps we would like to know some summary information, for example, the mean GPA or the range of SAT scores. Maybe we would like to describe how well the SAT scores and high school rank-in-class predict the students' future performance. Describing income data, age data, sports statistics, and a myriad of other examples all present us with the challenge of taking a mound of figures and assembling them in a fashion from which we can glean meaning.

The second challenge is the challenge of statistical inference. Suppose we take a poll of 1000 voters before an election to find out how they will vote. We really want to know how the 100 million voters will vote in the next election. How confident can we be that the opinions of the 1000 voters we ask really do reflect the opinions of the 100 million voters who will vote in the election? That is one of the challenges of statistical inference. Predicting the future weather given information about past weather, deducing whether a new drug is efficacious, guessing the future performance of the stock market, and doing scientific experiments on a few mice and drawing conclusions about all animals are all examples of the statistical challenge of inferring conclusions about the whole population when we have information about only a sample of the population.

Lectures 1 through 12 introduce the concepts of statistics. Typically, several different application areas are used to illustrate each statistical concept. Lectures 13 through 24 are organized by application area. Typically, several different statistical concepts are introduced and used in each application area. Both parts of the course are full of interesting and entertaining examples from all corners of our world—business and economics, medicine, education, sports, social science, and many more areas. For example, we will see how statistics was used to estimate the number of German tanks in World War II from the serial numbers of captured tanks, and we will see how a statistical analysis provides strong evidence concerning the disputed authorship of 12 of *The Federalist Papers*. Statistics offers unrivaled scope for connecting inherently intriguing mathematical ideas to the real world.

A typical statistics course in college emphasizes various technical tests. Students emerge with the impression that statistics amounts to plugging data into a formula. Although we will introduce important statistical formulas, this course emphasizes the logical foundations and underlying strategies of statistical reasoning. We describe why randomness lies at the heart of statistical reasoning. We explain what it means when the headlines blare, "Candidate A to get 59% of the vote with a + or – 3% margin of error." We differentiate between *statistically significant* and *significant*.

Our goal is to convey an authentic understanding of one of the most useful, powerful, and pervasive modes of reasoning employed in the world today. We will see why statistics will become increasingly important as technological advances continue to bring larger data sets and more detailed techniques of analysis within the range of practicality.

Note: Although the data used in this booklet are often real, some have been created to illustrate particular statistical concepts. ■

Describing Data and Inferring Meaning

Lecture 1

You look at articles about politics, elections, world conflict, certainly economics, business, and all of these centrally involve data and interpretation of data.

Trying to understand the economy, the weather, education, politics, risk, measurement, society, marketing, science, sports, medicine, and nearly every other aspect of our world fundamentally involves our ability to work with data in a meaningful way. The statistical analysis of data has become an integral part of how we describe our world. We expect data and statistical analysis to support opinions and decisions in almost every aspect of public and private life. The fundamental challenge for statistics is to assemble data and to interpret them to provide meaning. The statistical study of data can be viewed as dealing with two fundamental questions: (1) How can we describe and understand a situation when we have all the pertinent data about it? (2) How can we infer features of all the data when we know only some of the data? The goal of statistical perspectives and methods is to allow us to draw meaning from data.

The trouble with data is that they can be misleading or meaningless. Data by themselves do not have meaning. Perhaps the most famous quotation about statistics is from Mark Twain, which he attributed to Benjamin Disraeli, namely, "There are three kinds of lies: lies, damned lies, and statistics." There is an apt rejoinder to Twain's quote, namely, "It is easy to lie with statistics, but it is easier to lie without them."

The two most fundamental words in the study of statistics, *statistics* and *data*, share a grammatical issue. Are they singular or plural? *Data* is the plural of datum, a single piece of information. *Statistics* can be singular or plural, depending on the meaning. For example, statistics *is* the study of data, but statistics *are* bits of information.

Many people, in considering statistics, think about summarizing very complicated situations in one or two words. But that is far too simplistic a

view of statistical analysis. Instead, we need to develop tools and vocabulary for describing a complicated collection of information in ways that do not lose too much detail, yet in which the summary descriptions are sufficiently simple that they convey meaning to us.

This course is divided into two parts. The first part, Lectures 1 through 12, is organized to present a logical, conceptual development of the study of statistics. The second part, Lectures 13 through 24, considers different application areas and presents examples of the use of statistics in those various areas. Lectures 2 and 3 together form an introduction to the whole of analysis of statistics.

Lectures 4 through 7 present the basic conceptual tools for organizing, describing, and summarizing data. We introduce the basic concept of the *distribution* of data, which refers to how the whole collection of data is arrayed. We also introduce various shapes, such as the bell-shaped curve, that model many data sets.

Lecture 8 introduces the concept of *probability*. Probability is the glue that connects our methods for describing data sets to the idea of statistical inference. Probability is the study of measuring random behavior. Knowing what to expect from random events then allows us to compare data with randomness to make inferences.

In Lecture 9, we introduce one of the basic concepts of inference, namely, a *sample*. A sample consists of just some of the members of a population. We obtain data about them. A typical sample is seen when a pollster asks perhaps 1200 voters how they will vote.

In Lecture 10, we introduce a fundamental statistical strategy for inferring information about the world, called a *hypothesis test*. The strategy is to do some experiment from which we gather data, then investigate whether the results accord with a view of the world that we are evaluating. The logic of hypothesis testing lies at the heart of most statistical inferences we read about.

In Lecture 11, we introduce the concept of a *confidence interval*. We have all seen headlines that say, "Candidate A will receive 59% of the vote with a margin of error of ±3%." We will see what this phrase means. The headline is not complete as written, and we will see what is missing in order to truly understand the meaning of this confidence interval.

Lecture 12 describes the design of experiments. In order to gather data from which sound deductions can be drawn, it is important to think ahead about how those data will be used. Poorly designed experiments can result in a great deal of data from which almost nothing can be learned.

The second half of the course concerns application areas. Lecture 13 deals with applications of statistics in the law. You will be on the jury. Lectures 14 and 15 both concern democracy—voting and elections—and present interesting and paradoxical features of the simple-sounding idea of following the will of the people. Lecture 16 concerns sports, an arena full of statistically interesting issues. Lecture 17 discusses risk in war and insurance. Lecture 18 is a case study using real estate. We demonstrate how data about features of houses can be used to predict their sale prices. Lecture 19 is full of interesting and amusing examples of how statistical information can, unintentionally or otherwise, be presented in misleading ways. Lecture 20 concerns social sciences. Lecture 21 addresses statistics in health matters. Lecture 22 explores trends in economics, including the consumer price index and the stock market. Lecture 23 is about statistics in science. Finally, Lecture 24 argues not only that the prominence of statistics is enormous now but also that its influence and importance will increase a great deal over the next decades. ■

"There are three kinds of lies: lies, damned lies, and statistics."

—Mark Twain

Suggested Reading

Norman L. Johnson and Samuel Kotz, eds., *Leading Personalities in Statistical Sciences: From the Seventeenth Century to the Present*.

David S. Moore, *Statistics: Concepts and Controversies*, 5th ed.

Questions to Consider

1. Do the prevalence of data and the universal use of statistical arguments produce better decisions in our society?

2. Select essentially any topic, and explore the Internet to find data and statistical information that deepen your understanding of the issue.

Data and Distributions—Getting the Picture

Lecture 2

Data. Data. Data. Statistics starts with data. Governments, businesses, universities, sports fans—they amass mounds of data about people, about products, academics, win/loss records. You talk about the age of people, salary, gender, position, price, measurements, graduation rates, world records. They're all assembled into massive tables. Our question is, How can we understand this ocean of numbers?

The first three rules of statistics are draw a picture, draw a picture, draw a picture. A visual representation of data reveals patterns and relationships in the data, for example, the distribution of one variable or an association between two variables. A graph may show important features of the data, such as the center or spread, or unexpected values, such as outliers. Graphical representations can be used to tell others the story embodied in the data. Typically, we describe distributions of data by characterizing the general shape of the distribution (for example, bell shaped), finding where the distribution of the data is centered, then measuring how spread out or how concentrated the data are from the center.

The world is full of tables of data. This lecture describes how we can make sense of such lists of numbers. The challenge is to organize, describe, and summarize a set of data. Organizing a set of data consists of listing the data in useful order and grouping data effectively, such as with a histogram, which we will describe and define in this lecture. More imaginative graphical representations of data can sometimes help us to see patterns, as in Florence Nightingale's rose charts of 1857, showing the causes of mortality of soldiers during the Crimean War.

Our first example looks at data on the distribution of the income of associate professors of statistics. Putting the numbers in order gives some structure and information, such as the largest ($105,550) and smallest ($52,290) number. It is common to summarize a collection of data with a single number. The *mean* is one summary of the data. The mean is obtained by adding up all the numbers and dividing by the number of data points. In this case, the mean

is $68,500. The mean may not be a good summary of the data. Knowing how many data items there are, we can find the number in the middle of the ordered list. That number is called the *median*. In this example, the median is $65,600.

We can also see how much income marks the dividing line between the 25% of the people that earn the least and the other 75% of the people. This value is called the *first quartile* and is $63,480. Similarly, the *third quartile* is the value three-quarters of the way through the list, in this case, $75,350.

Five values—minimum, first quartile, median, third quartile, and maximum—give a *five-number summary* of the data. This information can be displayed graphically by drawing a *box plot* with whiskers. Data items that lie far outside the values between the first and third quartile are called *outliers*.

Drawing a histogram can give a more detailed view of how the data are arrayed. A *histogram* is created by dividing the possible values of the data (such as personal incomes) into disjoint groups and counting how many data items lie in each group. A histogram gives a sense of the shape of the data. For instance, the data may be *skewed* to the right. The *distribution* of a set of data is defined as the values that a variable takes and how often it takes them.

More imaginative graphical representations of data can sometimes help us see patterns, as in Florence Nightingale's rose charts of 1857, showing the causes of mortality of soldiers during the Crimean War.

Another example is Scholastic Aptitude Test (SAT) scores of students at a university. The histogram has a center point at about 1030 and is somewhat symmetrical about that value. Another example is the heights of men in the United States. We can see a symmetrical shape from the histogram, based on 2-inch increments.

SAT scores of students taking the test at a K–12 private school illustrate another feature of a distribution. This is a *bimodal distribution*, meaning that it has two distinct peaks. The reason for the two peaks is that all 7th graders

and all 11th graders take the test. Consider now the distribution of the heights of women. The shape is about the same as the shape for men, but it is shifted to the left; that is, the center is a smaller height. Two other examples are baseball players' batting averages in the year 1920 and in the year 2000. The center point is the same (about .265), but the spread is different. These examples illustrate the three main aspects of a distribution: *shape*, *center*, and *spread*.

Another fundamental aspect of getting meaning from data is the relationship between two varying aspects of the same individual in a population. An example is a student's SAT score and grade point average (GPA) in college. Each individual is represented by a dot on the graph, using the person's SAT score to determine how far to the right to place the dot and using the person's GPA to determine how far up to place the dot. This graph is called a *scatter plot*. In our onscreen example, we see that the two quantities, SAT score and GPA, appear to be somewhat related to one another. The precise meaning of *correlation* as used in statistics will be given in Lecture 7, where we will learn how to quantify the correlation.

We can add to the graph a line that seems to go in the same direction as the associated data seem to be going and see the extent to which that straight-line relationship between the data is reflected from the data. More than two varying aspects of the same individual can be related, for example, a person's SAT score, high school GPA, and college GPA.

In sum, our goal is to organize, describe, and summarize collections of data. The basic way to get meaning from data involves the concept of distribution. The main aspects of a distribution are its shape, center, and spread. Histograms and box plots are useful graphical aids. Such quantities as the mean, median, and quartiles are often useful measures to summarize the data set, but they do not preserve all the information. Data come in different shapes, some that we will run into often. Associated data can be visualized by graphing a scatter plot and can sometimes be approximated by a straight line or a plane. All of these concepts will be quantified in future lectures. ■

Suggested Reading

David S. Moore and George P. McCabe, *Introduction to the Practice of Statistics*, 5th ed.

Edward R. Tufte, *The Visual Display of Quantitative Information*.

Howard Wainer, *A Trout in the Milk and Other Visual Adventures*.

———, *Visual Revelations: Graphical Tales of Fate and Deception from Napoleon Bonaparte to Ross Perot*.

Questions to Consider

1. What is meant by the *shape* of a set of data?

2. Would you expect the mean and the median to be close to the same value or quite different when looking at the points scored by all basketball players in the NBA? Or looking at all the lifetimes of similar light bulbs in a test of longevity?

Inference—How Close? How Confident?

Lecture 3

In this lecture, we're going to introduce basic concepts and principles of statistical inference, namely, how can we use information about just some members of a population to infer that information about the whole population?

When an election is impending, pollsters ask perhaps 1000 people how they will vote, but what we want to know is how the whole population will vote. The challenge of statistical inference is to infer from the information about some people what the best guess is about the whole population. The logic of statistical inference is to compare data that we collect to expectations about what the data would be if the world were random in some particular respect. Analyses of randomness and probability allow us to quantify our confidence in extrapolations from some of the data to the whole population. Randomness and probability are the cornerstones of all methods for testing hypotheses. Here we introduce the rather subtle logic by which statistical inferences flow.

The goal of statistical inference is to answer the questions "How *close*?" and "How *confident*?" How *close* are the shape, center, and spread of just some of the population to the shape, center, and spread of the whole population? How *confident* are we of that?

Suppose we know the heights of only some adult men, but we want to know an accurate description of the distribution of heights of all adult men. We might just choose a few men whose heights somehow mirror the heights of all men, but we do not know the whole population's distribution to start with.

On the other hand, we know that representative samples do exist, but how can we find them without knowing what the whole population looks like in advance? Suppose we choose one adult male at random and measure him. Our random man's height is not likely to be extremely short relative to others or extremely tall; a random choice would generally find someone closer to the middle. The more men in the sample, however, the more likely

it is that their mean will be an increasingly better approximation to the population's mean.

One crucially important principle is that random choice in samples can lead us to information about the whole population. How can we estimate how close our sample's evidence is to the truth in the whole population? This is a question for probability theory. Probability gives us a quantitative measure for how likely a random event is.

Randomness is a central idea to the whole of statistical inference.

In order to make a useful conclusion, we have to make a reasonable guess about how *close* our random sample is to the population's values and how *confident* we are of being that close. Political polls are common examples in which statistical inference is used. Suppose the reality is that 60% of voters favor Candidate A and 40% favor Candidate B. Suppose we ask 100 random future voters how they will vote in the upcoming presidential election. Some number of those will favor Candidate A. Or we can run 10 simulations in which we ask 100 voters each time, with each simulation giving us a number for Candidate A.

It turns out that 95% of samples of size 100 will give a value between 50% and 70% in favor of Candidate A. We can be quite confident that the true percentage of all voters for Candidate A lies within ±10% of that percentage in a random sample of size 100. In this case, the answer to "How close?" is 10%, and the answer to "How confident?" is 95%. Surprisingly, 95% of the time, a random sample of 1200 people will give us an estimate that lies within ±3% of the actual truth in the whole population.

We have seen two important principles:

- Randomness is involved in statistical inference.
- The conclusion of a statistical inference is an estimate, together with a pair of numbers that tells us how close to the true population value the estimate is and how confident we are that the estimate is that close.

Another kind of challenge for statistical inference is how to tell whether a coin is fairly balanced and equally likely to come up heads as tails. In practice, we put the coin "on trial." We flip the coin 100 times and count how often it is heads and tails; we are getting random evidence about the frequency with which heads or tails arise.

Statistical inference figures prominently in tests of medications. Suppose we want to establish that a medicine works. A new medication is given to 1000 people who have a particular ailment, and a placebo is given to another 1000 people with the same ailment. If a much higher proportion of people get well taking the new medicine than taking the placebo, then that is strong evidence that the medication has a beneficial effect. If a somewhat higher proportion of people get well taking the new medicine, then we need to do some mathematical analysis to determine how persuasive the evidence is that the medication actually has a beneficial effect. To measure the strength of the evidence, we need to understand how much variation in results we should expect from the random process of some people getting well spontaneously.

If we want to see if a deck of cards is complete, we could devise an experiment. In our experiment, we choose a random card from the deck 3000 times and record the card each time. The histogram of our findings reveals whether cards are missing or there are multiples of any cards in the deck.

The logic of statistical inference is to compare data that we collect to expectations about what the data would be if the world were random in some particular respect. Analyses of randomness and probability allow us to quantify our confidence in extrapolations from some of the data to the whole population. Randomness and probability are the cornerstones of all methods for testing hypotheses. ■

Suggested Reading

Donald A. Berry, *Statistics: A Bayesian Perspective*.

David S. Moore and George P. McCabe, *Introduction to the Practice of Statistics*, 5th ed.

Questions to Consider

1. If you read that 53% of people polled favor Candidate A in an upcoming election, what additional information would you want to know to determine what you can conclude about Candidate A's chances of being elected?

2. Suppose a medicine is subjected to a test against a placebo, and 60% of the patients taking the medicine get better, while only 40% of those taking the placebo get better. What additional information would you need to know about the test to determine whether you could conclude that the medicine is effective?

Describing Dispersion or Measuring Spread

Lecture 4

To describe a set of data, we have to confront the challenge of taking a list of numbers and putting some structure on them through which we can garner meaning.

In this lecture, we begin a four-lecture series concerning principles and methods for organizing, describing, and summarizing data when we have all the data in the population. This lecture explores measures of center and measures of dispersion, or spread, of the data.

The mean and the median are both measures of central tendency. The median (the middle number in an ordered list) is not affected by outliers. The mean (the sum of the data divided by the number of data items) has a physical interpretation: It is the value around which the data items balance. The mean is also affected significantly by outliers.

Here is an example that shows why the mean does not tell us all we would like to know about a data set: The batting averages of baseball players in 1920 form a distribution whose mean is about the same as the mean of the batting averages in 2000, but the dispersions are different. The challenge of describing how widely a data set is dispersed, or spread out, leads us to develop several measures of dispersion.

Knowing the maximum, minimum, first and third quartiles, and median gives some indication of the spread of the data. This five-number summary does not give a refined sense of where all the data lie. However, a histogram contains a great deal of information about a data set and gives us a picture of the dispersion.

One natural measure of the spread of data is to look at how far each datum is away from the mean and to take an average of those values. We must be careful to take the absolute value of the distance of each datum from the mean. The average distance from the mean is a potentially useful measure of dispersion, but it is not the most commonly used measure.

The most common measure of dispersion, or spread, of data is the *standard deviation*. This value will be larger if the data set is more widely spread and smaller if the data are close to each other. The standard deviation can be computed as follows:

- Take the mean of the data.
- Subtract each datum from the mean, and square the result (multiply the result times itself), getting a positive number (or zero).
- Add those values, and divide by the number of items. This gives a number called the *variance*.
- Take the square root. This gives the *standard deviation*.

The standard deviation is basically the square root of the average squared distance from data points to the mean.

The mean is often represented by $\bar{x}$ or μ. The standard deviation measures how much variation there is in the data. In compact form, the standard deviation of the set of data $\{x_i\}_{i=1,\ldots,n}$ is $\sqrt{\dfrac{\sum_{i=1}^{n}(x_i-\mu)^2}{n}}$, where $\mu=\dfrac{\sum_{i=1}^{n}x_i}{n}$.

The variance is usually denoted s^2 or σ^2 (the Greek letter sigma, squared). In compact form, the variance is $\dfrac{\sum_{i=1}^{n}(x_i-\mu)^2}{n}$.

In some cases, it is more appropriate to divide by $n - 1$ in the definition of the variance and the standard deviation, rather than dividing by n. If we are taking the standard deviation over the whole data set, then it is appropriate to divide by n as we have written; however, if we are taking the standard

deviation of a sample of the population, then it is usually more appropriate to divide by $n - 1$.

The variance and standard deviation involve squaring the differences from the mean. But just looking at the absolute value of the difference from the mean seems more straightforward. You may ask why the variance and standard deviation are more commonly used.

One reason for squaring the difference rather than using just the absolute value of the difference is that the population mean then plays a special role; namely, it is the unique value that minimizes the sum of the squared differences. This can be shown using calculus. Just like the mean, the standard deviation is affected significantly by outliers.

The standard deviation gives us the opportunity to understand some differences in the world that are captured by differences in dispersion of a set of data. For example, a comparison between salaries in large corporations in the United States versus Japan reveals differing strategies of incentive and compensation. Notice that inferring meaning from the statistical evidence is a step of reasoning requiring an understanding of the context of the situation. The statistics themselves are a useful tool, but further logic is required to garner meaning from the data.

One reason for squaring the difference rather than using just the absolute value of the difference is that the population mean then plays a special role; namely, it is the unique value that minimizes the sum of the squared differences.

In this lecture, we have discussed the concept of dispersion of data and introduced the standard deviation, which is a common measure of dispersion. The most basic idea from this lecture is that the mean and the median, though useful summaries of a collection of data, do not tell us anything about how widely spread out the data are. Histograms give us a good visual sense of the distribution, including how spread out the data are. The five-number summary (minimum, maximum, first and third quartiles,

and median) and associated box plot give some sense of how the data are spread out. The standard deviation is a numerical measure of roughly how far the data are on average from the mean. ■

Suggested Reading

Stephen J. Gould, *Full House: The Spread of Excellence from Plato to Darwin*.

David S. Moore and George P. McCabe, *Introduction to the Practice of Statistics*, 5th ed.

Questions to Consider

1. What aspects of the description of data do quartiles and standard deviation measure? Which measure is more sensitive to being influenced by outliers? Why?

2. Under what circumstances is it important to consider the distribution of the data rather than just relying on the mean or median as a summary? For example, would it be important to know the distribution of recovery rates in guessing how long you will be away from work for a serious illness, or would the average recovery rate be sufficient? Why?

Models of Distributions—Shapely Families

Lecture 5

In this lecture, we are going to aim for discussion of the shapes of the data. How do we go about describing the shape of the data? ... Remember that the big picture of how to describe a distribution was to describe three things: the shape, the center, and the spread.

Any shaped curve at all can correspond to, or can model, a data set. To be useful, however, we seek model shapes that are easily described, both in words and mathematically, and that we expect to correspond to the phenomena we are describing. Consider the salaries of employees at a large corporation. Those salaries would have a concentration of values near the low end with a few salaries that are much higher, corresponding to the top management. Salaries at a corporation or household incomes in the United States are examples of skewed distributions.

Consider the intervals of time between arrivals of cars at a tollbooth. Or consider the time for an atom to decay in a radioactively decaying substance. Those sets of times will be collections of numbers whose distributions have a specific, characteristic shape. In this lecture, we will introduce terms (*skewed*, *bimodal*) that are applicable to some shapes; we will also describe several different characteristically shaped classes of distributions, including *exponential* and *Poisson*. Each naturally arises in specific settings.

The simplest shape that a distribution can have is a flat line. These distributions are called *uniform distributions*.

Our strategy in this lecture will be, first, to identify general types or categories of shapes that arise commonly. Then, we will describe some collections of graphs that are specifically defined by mathematical formulas. These specific shapes provide useful models that can give quantitatively valuable information about data sets that

they approximate. Different types of processes that generate the data typically lead to data sets that are well modeled by the appropriate mathematical formula.

Here are some generic shapes or characteristics of shapes of the histograms of various data sets:

- Some collections of data have a single-peaked histogram.
- Some collections may be *symmetric* (about the mean).
- Some collections are *skewed*, with a tail on one side of the center.
- Some collections are *bimodal*, having two peaks.

If the shape of the data can be described by a specific mathematical formula, then the mathematical description can often give valuable information about the data. The simplest shape that a distribution can have is a flat line. These distributions are called *uniform distributions*. A uniform distribution has a simple mathematical formula: $f(x) = c$, a constant.

The uniform distribution is a mathematical model that approximates real data about the results of throwing a die many times. The approximating function does not fit the data exactly. In the example in the lecture, we expected the uniform distribution to be a good model because of our knowledge or assumptions we made about throwing a fair die.

Suppose that we observe the times that cars arrive at an intersection. The distributions of data that we expect will have similar shapes. If we plot in a histogram the number of one-minute time intervals in which 0, 1, 2, etc., cars arrived at an intersection, we typically obtain a characteristic shape called a *Poisson distribution*. The shape of the ideal Poisson histogram is skewed right.

There is a mathematical formula that gives this ideal shape. In our example, the proportion of intervals that have k cars equals $e^{-6}6^k / k!$ The formula gives proportions. To get the actual number of intervals, we would multiply by a constant coming from the example. The general formula for a Poisson distribution is $e^{-\lambda}\lambda^k / k!$. Lambda (λ) is a parameter. Each different value of lambda gives a different member of the Poisson family of distributions. This distribution family has one parameter, lambda, which is the mean of the distribution. That is, if some arrival process has this distribution, then on average, there will be lambda arrivals per minute. In summary, certain assumptions about a physical phenomenon imply that there is a family of mathematical functions (in this example, the Poisson distributions) that applies to summarize the distribution of data from the phenomenon.

The numbers of deaths each year in each corps of soldiers in the Prussian army from kicks by army mules gives another example of a Poisson distribution, because the death by a mule kick does not change the likelihood of a similar death elsewhere. Another example of a Poisson distribution is from a study by Lewis W. Richardson of the number of outbreaks of war in each of the years from 1500 to 1931. The graphs of these examples will tend to look the same, just with different widths.

Another useful distribution is the *exponential distribution*, which is related to the Poisson distribution. An example of an exponential distribution is given by radioactive decay. A radioactive material, such as strontium 89, has the property called its *half-life*, which refers to a length of time during which half of the radioactive material will change to its nonradioactive state. It will take just as long again for half of the remaining material to decay. Viewing each radioactive atom individually in the next interval of, say, 10 seconds, the atom has a certain probability of decaying. No matter how long you wait, if you then consider an atom that is still radioactive, the probability of its decaying in the next 10 seconds is the same. The possible lifetimes of an individual radioactive atom form an exponential distribution.

The *binomial distributions* constitute another family of distributions. About two-thirds of students who enter high school in the United States graduate. Suppose we choose every possible group of 50 students, and for each such set of 50, we count how many graduated. We can plot a histogram showing how many groups of 50 students have 0 graduates, 1 graduate, 2 graduates, … , up to 50 graduates. That histogram is a binomial distribution. It has one peak, centered at about 33 (two-thirds of 50). It looks basically symmetrical, with a bell shape. There are many situations in which we would expect the data to have a binomial distribution.

Whenever we look at all possible samples of a given size from a population and record how many of the samples have each possible number of a given attribute, then we will have a binomial distribution. Examples include the following: Consider all possible ways that a coin flipped 15 times can land, such as H H T T T H T H H H T T T H T T, and count how many heads come up in each possible sequence of 15 H's and T's. Consider every set of 1600 people, and for each set, find out how many will vote yes on Proposition B in the next election.

The fact that these examples of samples of a given size from a population give rise to a binomial distribution will be critical in our later analysis of statistical inference. The binomial distribution has a specific mathematical formula, the general form of which is $\frac{n!}{k!(n-k)!}p^k(1-p)^{n-k}$.

Here, n represents the number of items selected (the size of the sample), and p is the proportion of the whole population that has the property. The formula gives the proportion of the sets of size n that have a count of k. When n is large, the binomial distribution is somewhat bell shaped.

In this lecture, we explored ways to describe the shape of a distribution of data. Generic descriptions of shape included the following:

- Skewed
- Bimodal
- Symmetric

Specific families of distributions were modeled by formulas, including the following:

- Uniform distributions
- Poisson distributions
- Exponential distributions
- Binomial distributions

The basic strategy for describing the shape of a set of data is to find a mathematical model that approximates the histogram of the data we have. ■

Suggested Reading

Educational Testing Service, http://www.ets.org/Media/Education_Topics/pdf/onethird.pdf.

David S. Moore and George P. McCabe, *Introduction to the Practice of Statistics*, 5th ed.

Questions to Consider

1. Why would the numbers of years with 0, 1, 2, 3,… holes in one in Professional Golfers' Association (PGA) major tournaments resemble the shape of the numbers of one-minute intervals of time during which 0, 1, 2, 3,… cars arrived at a tollbooth? Similar generating principles lead to similarly shaped distributions.

2. Suppose you have a large jar with billions of marbles, 60% white and 40% black. If you consider every possible collection of 100 marbles and make a histogram counting how many of those collections have 0 white, 1 white, 2 white, 3 white,…, up to 100 white marbles, you will have a binominal distribution. If you do the same thing except using all subsets of 100,000 marbles, you will get a different binomial distribution. How would you describe the difference in shape between the 100-marble binomial distribution compared to the 100,000-marble binomial distribution?

The Bell Curve

Lecture 6

Well, the most famous shape of distributions, ever, is the bell-shaped curve, which is called the *Gaussian* or the *normal distribution*. So, the family of curves—the normal curves or the Gaussian curves—is the topic of this lecture.

The bell curve is symmetrical and has a specific shape and a specific mathematical formula that describes it. It arises frequently for several reasons, physical and mathematical. In this lecture, we will introduce the bell-shaped curve, explore its properties, and discuss the reasons why it arises so frequently. In the previous lectures, we examined data sets of various shapes. One method for using one data set to get another is to take all the samples of a given size and take the mean of each. That collection of data (the sample means) gives a new data collection. Basically, no matter what the shape of the population data with which we start, the distribution of the sample means will converge to a normal curve. That observation, known as the *central limit theorem*, is one of the core insights on which statistical inference is based. Here are some of the examples we will discuss in this lecture:

- Heights of men in the United States.
- Heights of women in the United States.
- Major league batting averages in 1920.
- Major league batting averages in 2000.
- Average value of "poker" hands.
- Binomial distributions from the last lecture.

The term *Gaussian* celebrates the famous German mathematician Carl Friedrich Gauss (1777−1855). In 1801, Gauss was able to predict a future position of the asteroid Ceres based on past measured positions. The known

observations (as with all measurements) included errors. Gauss fit a curve to the observed data to minimize the error between the curve and the known observations. This process involved looking at the distribution of errors. An old name for the Gaussian distribution is the *error distribution*.

The Gaussian distribution arises not just as the distribution of errors in measurements. Lambert Adolphe Jacques Quetelet (1796−1874) was the first to apply the normal distribution in a setting other than errors (in *Sur l'homme et le développement de ses facultés, essai d'une physique sociale*, 1835), Quetelet introduced the concept of the average man (l'homme moyen). For many characteristics of people, such as height, the histogram resembles a bell-shaped curve. Quetelet also introduced the Body Mass Index (BMI).

One reason that the Gaussian distribution arises so frequently is that many times, the value we are measuring is the result of many small influences that randomly increase or decrease the final value. But the normal distribution is bell shaped.

One reason that the Gaussian distribution arises so frequently is that many times, the value we are measuring is the result of many small influences that randomly increase or decrease the final value.

Two Gaussian distributions can differ in their means (centers). For example, in the histograms of the heights of men and the heights of women, the means of the two just shifted over. Two Gaussian distributions can differ in their spread, as well. For example, the batting averages in 1920 and 2000 are both approximately normal curves with approximately the same mean but differing in their spread or standard deviation.

The normal distribution has a specific mathematical formula. The normal distribution's formula has a formidable and complicated look: $\frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$.

One interesting feature is that it contains both of the most famous constants in mathematics, π and e. It has two parameters, traditionally labeled by the Greek letters μ (mu) and σ (sigma). Any particular choices for μ and σ ($\sigma > 0$) determine a specific distribution; that is, we would know its shape exactly. Setting $\mu = 0$ and $\sigma = 1$ gives the standard normal curve. The number μ is the mean of the distribution. The number σ is the standard deviation, which measures the spread.

If we know the standard deviation in a normal distribution, we know exactly how all the data are spread around the mean. The proportion of the population whose values differ from the mean value by less than σ is 68%. This is true for a Gaussian distribution that is very tight (σ is small) or for one that is spread out (σ is larger). The proportion of the population whose values differ from the mean value by less than 2 times σ is 95%. The 3-σ proportion is 99.7%.

If we have a physical reason for expecting a certain kind of distribution, then our challenge of getting a good summary of the data is reduced to knowing the values of the parameters. For example, let's consider the heights of men. From experience with such collections of data, we expect such heights to be distributed normally. Thus, knowing the values of the mean μ and the standard deviation σ gives us information about the population. We can compare items in two different normal collections in a meaningful way by measuring how many standard deviations each is away from its respective mean; that number of standard deviations away from the mean called the *z-score*. Every binomial distribution with n at least 5 is bell shaped and is well approximated by a normal curve.

Approximations of the normal distribution arise when we start with any shaped distribution and look at the distribution of averages of samples of a certain size. We will illustrate this idea with an example using a standard deck of cards. Assign a value to each card: Ace =1, 2 = 2, 3 = 3,…, 10 = 10, Jack = 11, Queen = 12, and King = 13. For each possible five-card hand in the whole deck, take the average value of the cards in the hand.

Plot a histogram of all the possible average values. Notice that it looks rather bell shaped. The histogram is centered at the mean of the whole deck, which is 7.

The central limit theorem says that starting with (almost) any distribution (such as a Poisson distribution, or a binomial distribution, or a uniform distribution), if we take many samples of size n, the distribution of the average values of the samples will be approximately a Gaussian distribution (assuming n is large). The standard deviation of the sample population will be the original population's standard deviation divided by the square root of n. This fact will be useful in doing statistical inference. The approximation gets better and better as the sample size n gets larger and larger.

In this lecture, we introduced the famous bell-shaped curve. This normal, or Gaussian, distribution arises in many settings, including heights of men, scores on tests, measurements, and sample means. If we start with any reasonable distribution and take the means of samples of a certain size, the distribution of those sample means will be well approximated by a normal distribution. If we know the mean and standard deviation of a normal distribution, then we know that about 68% of the data will be within one standard deviation from the mean; 95%, within two standard deviations from the mean; and 99.7%, within three standard deviations from the mean. Talking about how many standard deviations a value is away from the mean allows us to meaningfully compare different populations. ■

Suggested Reading

Donald A. Berry and Bernard W. Lindgren, *Statistics: Theory and Methods*, 2nd ed.

David S. Moore and George P. McCabe, *Introduction to the Practice of Statistics*, 5th ed.

Questions to Consider

1. Why would we expect to see a normal distribution of data if we looked at 50 measurements of the speed of light in careful experiments?

2. Suppose you have a data set that tells you the heights of men and you are selling pants. Suppose you want to stock pants for a certain percentage of men up to a given height. You consider a business plan calling for accommodating men up to the 70^{th} percentile, the 80^{th} percentile, and the 90^{th} percentile. Would you need more sizes to move you from the 70^{th} percentile to the 80^{th} percentile or from the 80^{th} percentile to the 90^{th} percentile?

Correlation and Regression—Moving Together

Lecture 7

In everyday life, social situations, you have lots of examples of cause and effect. If a student studies more, that student may possibly do better on the next test. The income of a person is related to the amount of education that they have. But you could certainly ask the question of whether the increase in the income is a result of the knowledge or skills they gained from education.

Describing relationships and cause and effect among variables is a basic strategy for understanding the world. A statistical challenge is to describe and measure how two variables, such as incoming SAT scores and college GPAs, are related. Higher SAT scores correspond with better GPAs in college; however, the SAT scores do not cause higher grades. In this lecture, we will introduce the ideas of *scatter plots*, which give a visual sense of the relationships between two variables; *correlation*, which gives a quantitative measure of the strength of the linear relationship between two varying quantities; and *linear regression*, which produces a straight-line approximation for a set of paired variables.

We will also touch on the idea that more than one variable might be involved in describing another variable. For example, perhaps SAT score and high school GPA might together form a better predictor for college success than either variable alone might do. Such a situation introduces the idea of *multiple regression*.

Everyday life and social situations have examples of cause and effect. A person's income seems related to the number of years of education. However, the education may or may not cause the extra income. If a cause and effect relationship is operating, the manifestation is a list of pairs of numbers.

In this lecture, we discuss how statistics can help determine whether in a list of pairs of numbers, there is a relation between the attribute measured by the first numbers and the attribute measured by the second numbers. If the two move together, they are correlated. The question of causation is outside the

realm of statistics. The first example we discuss involves SAT scores and college GPA.

The first step in determining whether two quantities are correlated is visual. Let's make a graph, called a *scatter plot*, in which the horizontal axis represents one quantity and the vertical axis, the other. For example, the horizontal axis could be SAT scores and the vertical axis could be GPAs after the first year of college. For each student, we place a dot on the graph, horizontally positioned as far to the right as the student's SAT score and vertically up as far as his or her GPA. The cloud of dots generally rises as it goes to the right because, in general, a higher SAT score is associated with a higher GPA.

If one quantity increases and the other quantity tends to increase, the correlation is called *positive*, and the opposite is *negative*.

We can quantify the extent to which two variables are related by giving a number defining the *correlation* that summarizes the relationship. If one quantity increases and the other quantity tends to increase, the correlation is called *positive*, and the opposite is *negative*. Unlike the scatter plot of SAT scores and GPAs, which demonstrates a positive correlation, the scatter plot of two negatively correlated quantities goes down as it goes to the right.

To understand the formula for correlation, we consider the two quantities individually at first. For example, we can compute the mean of the SAT scores and the standard deviation of those scores and perform similar computations for the GPAs. We say that a data set is perfectly correlated if the distance from the mean (measured in standard deviations) of one variable corresponds exactly to that distance (again measured in standard deviations) from the mean of the other variable.

The correlation is denoted r (or ρ [the Greek letter rho]). It is computed as follows: For each member of the population (e.g., student), take the distance of the first variable (SAT score) from the first variable's mean in units of standard deviation (that is, take the z-score of the SAT score) multiplied by

the distance of the second variable (GPA) from its mean, again measured in units of standard deviation (that is, multiply by the z-score of the GPA). Essentially, take the mean of those products over all the students. (Actually, we divide the sum by one less than the number of pairs.) The formula for correlation, then, is:

$$r = \frac{\sum_{i=1}^{n} \left(\frac{x_i - \overline{x}}{s_x} \right) \left(\frac{y_i - \overline{y}}{s_y} \right)}{n-1}.$$

The correlation is a number between 1 and –1. If the number of standard deviations from the mean of one variable corresponds exactly to the same number for the other variable, then the correlation will have value +1. In this case, the scatter plot will not be a cloud, but instead, all the dots will lie on a straight line. The correlation value is closer to zero if the dots are not as close to lying on a single straight line.

The statistical measure of correlation is not quite what we might think. Consider the data set {(1,1), (2,2), (3,3), (4,4), (5,5), (0,7)}. The statistical correlation of the set of all six members has value 0. Although five members of the data set are perfectly correlated, the sixth member (0,7) is not at all aligned with the others.

Measuring correlation can reveal some interesting associations, such as in the 1969 Vietnam War draft lottery. Each of the 366 capsules containing possible birthdays was drawn out of a container in a manner intended to be random. We can make a scatter plot of the day of the year (numbered 1 for January 1, 2 for January 2, and so on) versus the order drawn for the actual 1969 drawing. The scatter plot looks random, but computing the statistical correlation gives a value that is statistically unlikely to have occurred by chance alone. Perhaps the 366 capsules containing the birth dates were not mixed thoroughly enough when they were placed in the container from which they were drawn.

We often summarize a scatter plot by a straight line. If the data set has a high correlation, the straight line will lie near most of the points. The straight line is a kind of summary of the data. The line is called a *linear regression line*. It is a line that approximates the data.

We can measure how well the straight line approximates the data set. For each data point, its vertical distance from the line is called its *residual*. Squaring each residual and adding them up gives a measure of how far the data points are from the line. The line that minimizes the sum of squares of the differences between the actual and predicted vertical values is called the *least squares regression line*.

Similar mechanisms are used when there are more than two variables. Several *explanatory variables* can often make better predictions than any single variable alone. When two or more variables are used to predict another variable, we use *multiple regression* to summarize the situation.

In this lecture, we've dealt with a population that has more than one quantity (variable) associated with each member of the population. We looked at the statistical correlation of any two variables. We looked at regression, which is a way of summarizing data by computing the best-fit line, plane, or higher dimensional analog of a line or plane showing the predictive dependency of one variable on the others. The statistical concept of correlation helps us identify and quantify the extent to which paired qualities of members of a population vary together. Statistical correlation indicates an association but does not prove that there is a causal relationship between the variables. One of the misuses of statistical information is to mistakenly infer cause and effect from correlation. ■

Suggested Reading

B. Bowerman, R. O'Connell, and A. Koehler, *Forecasting, Time Series, and Regression: An Applied Approach*, 4th ed., part II.

David S. Moore and George P. McCabe, *Introduction to the Practice of Statistics*, 5th ed.

Questions to Consider

1. Think of several examples of data sets that you would expect to be correlated but for which cause and effect would not be a reasonable explanation.

2. The least squares regression line takes the x-coordinate of every data point, takes the difference between the y-coordinate of the data point and the y-coordinate on the regression line with that x-coordinate, squares those differences, and adds them up. Suppose that we started with the y-coordinate instead of the x-coordinate. Would we get a different sum of squares? Does it matter which we choose? Why?

Probability—Workhorse for Inference

Lecture 8

The foundation of statistical inference is comparing the results that we find from actual data that we collect with what we should expect from random processes. But our intuition about what we are supposed to expect from a random process is often far from accurate.

Probability is the bridge between the two ideas of (1) describing data when we know all the data and (2) inferring characteristics of the whole population from a sample. Probability is the study of randomness. Probability accomplishes the seemingly impossible feat of putting a useful numerical value on the likelihood of random events. Games of chance, such as dice and cards, form the historical background and give a clear introduction to the idea of probability. Our intuition about what to expect from randomness is often far from accurate. We will see several startling examples in which our intuition and reality are quite divergent. For the purposes of applications to statistics, it is vital that we know with quantitative accuracy what probabilities we should expect from randomness because the basis of making inferences about statistical data usually amounts to comparing two sets of data—what we would expect to get in a random situation versus what we actually measure.

Probability was introduced in the 17th century. In 1654, the French gambler Antoine Gombaud, the chevalier de Méré, posed a question about a gambling game to Blaise Pascal, who asked Pierre Fermat. Pascal and Fermat introduced the basic concepts of probability.

When every possible outcome of an experiment is equally likely, it is easy to calculate the probability of an outcome. If a coin is equally likely to land heads as tails when flipped, we say the probability of heads is 50% (or 0.5). For a fair die, the probability is 1/6 of rolling a 4, for example. For cases where possible outcomes are not equally likely, the *probability* of a possible outcome means the fraction of the time that the outcome will occur.

Randomness becomes more subtle the more we think about it. When flipping a coin, in what sense is the outcome a random event? Once we let the coin go, physics determines what the outcome will be. Interestingly, in quantum physics, the basic model of particles is inherently probabilistic.

We humans have bad intuition about probabilities. The following is an example that is surprising to most people. Suppose there are 50 random people in a room. What is the probability that two of them will have the same birthday? The surprising answer is that there is a 97% chance that two of them will have the same birthday.

It is easier to compute the opposite possibility, namely, that no two of the 50 people have the same birthday. Imagine asking the people one by one for their birth dates. To compute the probability that all the people have different birthdays, you would multiply as follows:

$$\frac{365}{366} \times \frac{364}{366} \times \frac{363}{366} \times \ldots \times \frac{317}{366} = 0.03 .$$

The product of all the fractions is about 0.03. Thus, the probability that no two people have the same birthday is only about 3%. Hence, the chance that two people *do* have the same birthday is about 97%.

When two fair dice are thrown and summed, not all possible sums are equally likely. Suppose you have a red die and a black die, and each is a fair die. There are six ways of getting the sum 7: red 1, black 6; red 2, black 5; and so on. We can draw a histogram showing how many ways there are to get each of the numbers 2 through 12. The histogram has a peak over the value 7. When three fair dice are thrown and summed, the histogram of possible sums is more peaked than the histogram for two dice.

To compare these histograms more easily, we rescale the histograms. Think of a gas station with a peculiar way of setting the price for a gallon of gas:

- You'll pay between $1 and $6 dollars per gallon, determined by throwing dice.

- You throw the dice and take the average of the values showing on the dice. That is your price per gallon.

- It is not the sum of the dice that we will now deal with, but rather, we'll deal with their average.

We will draw the histograms not over the possible values of the sum but over the possible values of the average (the sum of the dice divided by the number of dice). We rescale the vertical axis of the histogram to make the area of the histogram 1 (i.e., 100%). If we had four dice, we could likewise draw the histogram showing what proportion of the possible ways four dice could land given each of the possible average values of the dice. This histogram is more peaked than the earlier histograms. Likewise, the histogram for five dice is even more peaked. If we do this process with 100 dice, the histogram would be even more peaked.

When flipping a coin, in what sense is the outcome a random event? Once we let the coin go, physics determines what the outcome will be. Interestingly, in quantum physics, the basic model of particles is inherently probabilistic.

Saying that the histogram is more peaked is saying that the standard deviation is smaller. In fact, the standard deviation for the case of 100 dice is 1/10 of the standard deviation for the case of a single die. If we used 1 million dice, we would get an extremely peaked histogram: The standard deviation would be only 1/1000 of the single-die case. The standard deviation decreases proportionately as the square root of the number of dice increases. The fact that the histogram becomes more peaked as the number of thrown dice increases is really an illustration of the central limit theorem.

If we consider any distribution, the distribution of the sample means will be peaked (at least for fairly large sample sizes). The peak will be near the mean of the population. As the sample size gets larger, the standard deviation of the distribution of the average value of the sample gets smaller. The histogram gets more peaked. The central limit theorem implies that for a fairly large sample size n, the distribution of the sample means is close to the normal distribution centered at the same mean as the population mean and with a standard deviation approximately equal to the standard deviation of the original distribution divided by the square root of n. The central limit theorem is one of the workhorses of statistical inference.

In this lecture, we saw that probability gave us a number between 0 and 1 that measures the likelihood of a random event. Probability can be effectively measured for many random events. The most straightforward cases occur when there are equally likely events, such as in rolling a fair die. Especially important is that we can measure the probability distribution of sample means; that is, we think about taking all samples of a certain size and taking the averages of each sample. These sample means have a distribution that is approximated by a normal distribution.

In the next lectures, we will see how our ability to understand probabilities lies at the heart of the logic of statistical inference. ■

Suggested Reading

E. T. Jaynes and G. Larry Bretthorst, eds., *Probability Theory: The Logic of Science*.

David S. Moore, *Statistics: Concepts and Controversies*, 5th ed.

Questions to Consider

1. Suppose someone flips two coins, looks at the coins, and announces, "At least one of the coins landed heads up." What is the probability that both coins are heads? You can perform experiments to confirm that your answer is correct.

2. Suppose there is a 50% probability of rain on each of Saturday and Sunday. What is the probability that it will rain this weekend?

Samples—The Few, The Chosen

Lecture 9

In this lecture, we're going to come to grips with the question of what only part of the data really means. So, this lecture is about the concept of a sample.

With this lecture, we begin the investigation of how to infer features of the whole population from information about just some of the members of the population. A common scenario in which this plays out occurs when a poll is taken to estimate what proportion of voters favor which candidate. The poll asks a few hundred or a few thousand people which candidate they prefer. That information is used to deduce what percentage of the whole population will vote for one or another. The backbone of this statistical analysis entails understanding the extent to which the sample accurately represents the opinions of the whole population. Many interesting and potentially problematic issues arise in taking and using samples. We will see several examples of potential sampling pitfalls, including bias, sample sizes that are too small, and receiving dishonest responses to questions that may be controversial or sensitive. Randomness is a key component of obtaining a representative sample.

Recall that the structure of our course is to view statistical analysis as having two basic parts: (1) how to describe, summarize, and organize a collection of data if we know all the data and (2) how to infer information about the whole population if we know only part of the data. The term *population* refers to the whole collection of people or things being considered. A *sample* is a subset of the total population, whether people, auto parts, or anything else we are investigating. We want to infer information about the whole population from information about the sample.

We might consider several techniques of drawing inferences about the whole population. We would like to know how the total population feels about something. Because we cannot afford to ask everyone, we need to be content with asking a subset of people. We might think that in deciding exactly whom to ask, we should carefully pick people with different specific traits. However,

the central characteristic of good sampling involves randomness rather than intent. If we choose a sample randomly, we are likely to get opinions that represent the whole population. If we intentionally choose certain groups, that choice may reflect biases we have about what sort of people are likely to have what opinions, resulting in a sample that is not representative of the whole population. We would like the proportion of people in the sample who have a certain opinion to be the same as the proportion of people in the whole population who have that opinion.

George Gallup used a poll of 50,000 people in the 1936 election, making a correct prediction of the election. Randomness was a key feature in his sampling technique.

One pitfall in sampling is bias. A famous example concerns the 1936 presidential election. The *Literary Digest* took a major poll to determine who would win. The *Literary Digest* sent out 10 million surveys and received 2.4 million replies. Based on the surveys, the *Literary Digest* predicted that Landon would win by a landslide, 370 electoral votes to Roosevelt's 161. The election was a landslide. However, Roosevelt won, not Landon—523 electoral votes for Roosevelt to 8 for Landon. Roosevelt received 62% of the popular vote in the election! Obviously, the *Literary Digest's* sample was not representative of the population. The *Literary Digest* used various lists of people, including their own subscribers and owners of cars and telephones. Their poll was biased toward wealthy people, whose opinions were not representative of the population at large.

Meanwhile, George Gallup used a poll of 50,000 people in the 1936 election, making a correct prediction of the election. Randomness was a key feature in his sampling technique. Randomness is a basic ingredient of essentially all standard statistical techniques.

Another pitfall of the *Literary Digest* survey was that it was a voluntary response survey; in other words, only those people who sent back the surveys had their opinions counted. This pitfall is illustrated by a decidedly unscientific survey undertaken by Ann Landers. In response to a letter

she received, she asked people whether, if they had to do it over again, they would have children; 70% said no. Subsequent surveys have shown that this result was completely inaccurate. A vast majority of people with children—a statically valid survey estimated 91%—would have them again. The collection of people who had read the Ann Landers question and went to the trouble of replying was strongly biased toward those who had a poor experience with child rearing.

Voluntary response polls can lead to results that present a totally inaccurate view of the whole population. The simplest way to get a sample that is likely to be representative of the whole population is to just randomly pick a subset of the population. This is called a *simple random sample* (SRS).

Another potential pitfall of surveys is that people may lie, particularly if honesty would be embarrassing or worse. Surprisingly, there is a way to get accurate results that can be done in a public room. This method allows us to estimate the percentage of students who cheat without revealing any particular student's history of cheating. Here is the technique:

- Ask every student to secretly flip a coin.
- Ask everyone to raise their hand if they either threw a head or they cheat.
- Suppose we do this experiment with 1000 students, and 800 raise their hands.
- We can estimate that about 60% of the students cheat because if 500 threw heads, then 300 of the 500 who threw tails cheated—about 60% (300/500).

In taking samples, our goal is to obtain samples from which we can accurately infer information about the whole population. That is, we want our samples to be representative of the whole population. Pitfalls to avoid in taking samples include avoiding biased samples and being wary of results obtained from voluntary response samples. We can obtain statistical information concerning sensitive information by using clever techniques in which no

individual needs to divulge sensitive secrets. The basic purpose of getting data from a sample is to infer information about the whole population. How that inference is made is the subject of the next lectures. ■

Suggested Reading

David S. Moore and George P. McCabe, *Introduction to the Practice of Statistics*, 5th ed.

Ann E. Watkins, Richard L. Scheaffer, and George W. Cobb, *Statistics in Action: Understanding a World of Data*.

Questions to Consider

1. Why is randomness an important feature for selecting a good sample? Wouldn't it be better to try to make sure that the sample is well balanced by choosing a balanced profile? What are the pros and cons of that approach?

2. Devise a technique using two coins for determining an estimate of the fraction of students who cheat without an observer being able to deduce of any particular student whether he or she cheats.

Hypothesis Testing—Innocent Until

Lecture 10

Hypothesis testing is one of the workhorses of statistical inference. The goal of this lecture is crystal clear: to illustrate and explain the logic of hypothesis testing. Basically, that's the core of the idea of statistical inference in general.

Hypothesis testing is one of the most common statistical techniques used in education, psychology, social sciences, natural sciences, medicine, law, and every area to which statistics is applied. The strategy is best illustrated with an example. Suppose we have a promising new medicine that might be effective in curing athlete's foot. We gather 200 itchy people and treat 100 with the new ointment and 100 with a placebo. If the new cream had no beneficial effect (the null hypothesis), we would expect about the same number of people to be cured by the ointment as by the placebo.

Suppose 80 cream-treated people are itch-free in a week and 60 placebo-treated people are itch-free. Can we conclude that the cream is effective? To measure the strength of the evidence, we compare the data obtained from our experiment to the data that are predicted by an appropriate sampling distribution. Using results from probability, we can quantify the likelihood that the data we collected were the result of luck alone. In this lecture, we explain this rather subtle strategy that lies at the heart of all statistical inference. If the result of the experiment would be very unlikely if the world were as hypothesized, we conclude that we have gathered strong evidence that the world is not as hypothesized, and we reject the null hypothesis.

We'll first describe how hypothesis testing works in an example. Our null hypothesis is that if we spin a penny, it has a .5 probability of landing heads up. At the end of the whole hypothesis testing process, the crux of the decision of whether to reject the null hypothesis or not depends on quantifying how rare different possible outcomes of the experiment are.

Thus, we need to investigate the possible outcomes of the coin experiment and how likely or unlikely each possible outcome is. We will spin the penny 100 times. Under our hypothesis, all possible outcomes are equally likely because each spin's landing is independent of that of the other spins. We can make a graph in which the x-axis (the horizontal axis) tells how many heads are in the outcome and the y-axis tells how many of all the possible outcomes have exactly that many heads. The graph has a bell shape. The curve is an approximation to a normal distribution centered at 50.

Next, we do the experiment and analyze the results. We spin the coin 100 times. The result was 39 heads. We can look at the graph to see where the value 39 fits. The height of the curve over the value 39 and more extreme values is, visually speaking, essentially 0. The probability of getting that value or a value further from the mean is called the *p value*. It is the probability that under the assumption of the null hypothesis, an outcome as rare as (or rarer than) the actual outcome would happen. The smaller the p value, the more extreme the outcome. Our intuition that the result of the experiment is strong evidence against the null hypothesis is quantified and confirmed by the fact that the p value is very small.

The phrase *reject the null hypothesis* is the way of saying that the hypothesis about the world is unlikely to be true.

Under the assumption of the hypothesis, the computations show that the probability of getting the experimental results is very low; thus, the experimental results are strong evidence for rejecting the null hypothesis. The phrase *reject the null hypothesis* is the way of saying that the hypothesis about the world is unlikely to be true.

Another example of hypothesis testing occurs if we are trying to determine whether or not a particular medication works by giving some people the medication and some a placebo and seeing whether there is a significant difference in the number who recover. For our experiment, we'll take 100 people, all with athlete's foot, and have them use the medication for a week. After the week, we count how many people are free from athlete's foot.

Assume that we know that 40% of the time, a person's athlete's foot will disappear in a week without any medication. Thus, our null hypothesis is that each person has a 40% probability of cure with or without the medication (i.e., the medication has no effect). We'll take as the alternative hypothesis that the probability of cure for people using the medication is different from 40%. We also decide how extreme a result of the experiment will have to be for us to reject the null hypothesis.

Our required significance level is 0.05, meaning that if the results of the experiment are so extreme that under the assumption of the null hypothesis, only 5 out of 100 experiments would yield a result that extreme or more extreme, we will reject the null hypothesis. In other words, if the p value is less than the significance level, we will reject the null hypothesis. We call such a result *statistically significant*. In the field of statistics, *significant* does not mean "important." Rather, it simply means that the significant item or event signifies something.

Suppose that 51 of the 100 people who used the medication are free from athlete's foot. In this case, the calculation gives the p value of 0.032, quite small. We conclude that it is very unlikely that the medication had no effect. Because the p value of 0.032 is less than the significance level of 0.05 that we specified before we did the experiment, we reject the null hypothesis.

As another example, consider the question of whether the typical American male adult consumes an average of 2400 calories each day. We ask 25 American adult males, and they average 2500.

Unlike the previous examples, knowing just the sample mean is not enough information. We also need to know something about how spread out the population's eating habits are, that is, the standard deviation of the population. If there is a huge variation from person to person in number of calories consumed, then the difference between the sample mean of 2500 and the null hypothesis of 2400 is not as significant. The sample responses themselves give some information about the standard deviation of the population. Let's suppose that the sample standard deviation (i.e., the spread in the 25 responses) is 270 calories.

There is a distribution, called the *Student's t distribution*, that applies to the case with which we are dealing in this example. The t distribution applies when the population distribution is normal or when the sample size is large, but where we don't know the standard deviation. This time, we compare our experimental results to expectations about the t distribution.

The general idea is exactly the same as in the previous examples. We do a calculation, based on the t distribution, that determines how extreme our sample is, assuming the null hypothesis is true. It turns out that a sample of mean 2500 calories and standard deviation 270 calories is not extreme, assuming the null hypothesis of 2400 calories and requiring a 0.05 level of significance. The p value is 0.085, greater than 0.05. We do not reject the null hypothesis. That is, the data would not warrant a conclusion that the average number of calories consumed by an American adult male is different from 2400.

This lecture has been on hypothesis testing, describing the strategy and some terminology. The steps are as follows:

- State a hypothesis about the way the world is. The hypothesis is called the null hypothesis and is often the opposite of what we actually think might be true.

- State an alternative hypothesis. In the case where the null hypothesis is that the probability of something is a certain value, three standard possible alternative hypotheses exist:

    - The probability is different from the value stated in the null hypothesis.

    - The probability is greater than the value stated in the null hypothesis.

    - The probability is less than the value stated in the null hypothesis.

- Declare what level of significance we require to reject the null hypothesis.

- Do the experiment, gather the results, and compute the probability that, under the assumption of the null hypothesis, the results would be as extreme as (or more extreme than) the result we actually obtained. This gives a p value.

- If the result is rare (p value smaller than our required level of significance), then we assert that the results are statistically significant, and we reject the null hypothesis:

    - Rejecting the null hypothesis means concluding that it is false.

    - Rejecting the null hypothesis also means concluding that the alternative hypothesis is true.

    - If the result is not rare (p value larger than our required level of significance), then we do not reject the null hypothesis.

This style of reasoning, comparing a hypothesized state of the world with the experimental data that we gather, is a fundamental strategy of statistical inference. ■

Suggested Reading

Vic Barnett, *Comparative Statistical Inference*.

E. T. Jaynes and G. Larry Bretthorst, eds., *Probability Theory: The Logic of Science*.

David S. Moore and George P. McCabe, *Introduction to the Practice of Statistics*, 5th ed.

Questions to Consider

1. Suppose in reality a medication cures 1% more people than a placebo does. Could a hypothesis test ever find statistically significant results that determine that such a medication works?

2. In a hypothesis test, it is customary to use a $p < 0.05$ level for statistical significance. Why wouldn't we instead use a $p < 0.5$ level, arguing that such a result would imply that the null hypothesis is more apt to be wrong than right?

Confidence Intervals—How Close? How Sure?

Lecture 11

Today we're going to talk about one of the most common statistical statements that we read in the newspapers, which is the headline that we see every election, which will say something like this: "Candidate A will receive 59% of the vote with a margin of error of + or – 3%."

In this lecture, we will see what such a statement means and why it is incomplete as written. The actual meaning uses the same subtle reasoning that we encountered in hypothesis testing in the previous lecture. The meaning boils down to trying to ascertain how confident we are that the data about our sample (the poll) do, in fact, accurately reflect the facts about the whole population. A key to both confidence intervals and hypothesis testing comes from the central limit theorem, which tells us how likely it is that the mean of the sample is close to the mean of the population.

Suppose there is an election coming up between two candidates, A and B. Imagine that we find out how 1000 random people will vote. What we really want to know is how the population of perhaps 100 million voters will vote. The topics of this lecture are the concepts of *confidence intervals* and *margin of error*. Namely, what can we conclude from the information we learn from the 1000 people about the 100 million people? We have actually done the substance of this whole analysis before. This repetition is intentional because this idea is so central to statistical inference.

Now, let's start our analysis. Suppose we choose 1000 people at random. The 1000 people we choose may, by luck, have a larger or smaller percentage who will vote for A than the percentage in the population. This lecture gets at the question of what the probability is that the percentage of people for A in the sample will be "close to" the percentage of the population for A and what "close to" means. If we ask 1000 people and 59% of them will vote for A, then we know that 59% of the people in the sample will vote for A. But how confident can we be that the whole population percentage that will vote for Candidate A is close to 59%?

To get across the ideas, we'll consider the exaggerated situation in which we ask only 10 voters instead of 1000. We can make a graph, putting on the x-axis (the horizontal axis) the numbers 0, 0.1, 0.2, …, 1.0, each being the percentage of the 10-person sample that is for A, and for each such makeup of A supporters, draw a point whose height is the percentage of the possible groups of 10 people that would have that makeup. From these calculations, we see that there is a significant chance that a group of 10 people chosen at random will have a makeup (that is, a certain percentage for Candidate A) significantly different from the general population's percentage of 60%.

Now, let's do the same analysis with larger samples. As we choose larger samples, 100, then 1000, then 10,000, the graphs become increasing more peaked, always centered close to the population's true mean, which we will assume is 60%. For a sample of size 1000, the vast majority of samples of size 1000 have a percentage for A that is very close to the population's 60%.

Mathematical analysis can tell us how big a sample needs to be. Recall that the central limit theorem tells us that the standard deviation (which is related to the width of the bell-shaped curve) declines by the square root of the sample size.

We can now explain what is meant by the report: "The poll shows that Candidate A will receive 59% + or – 3% of the votes." It means that the poll has taken a sample of sufficient size so that we are 95% confident that the actual population percentage of voters for A lies in the range 56% to 62%. The 95% is called the *confidence level*. It is traditional that if no confidence level is explicitly mentioned in reporting a poll's result (or any other result of some measurement) as a value plus or minus a margin of error, then 95% is understood. The range reported is called a *95% confidence interval*. If we want to be more confident than 95% that the interval we report contains the population's true value, we need to report a larger interval. Another way to be more confident is to use a bigger sample size.

Mathematical analysis can tell us how big a sample needs to be. Recall that the central limit theorem tells us that the standard deviation (which is related to the width of the bell-shaped curve) declines by the square root of the sample size. Taking four times as large a sample makes the standard deviation half as large. Larger samples give a tighter range. It turns out that to get a margin of error of 3% (i.e., to report a range of percentages +/– 3%) at a confidence level of 95% requires a sample size of about 1,200.

Notice that the required sample size does not depend significantly on the size of the whole population if the population is large. To do a poll in a municipal election with 100,000 voters requires the same number of people in the sample as in a national election with 100 million voters to get the same size range with the same level of confidence. A random sample of size 1,200 in a poll of two candidates is large enough for us to report a value to within 3 percentage points with 95% confidence, no matter how large the population is.

In the case of estimating the mean of a population with unknown variance, the meaning of the confidence interval and its level of confidence is related to the ideas of hypothesis testing discussed in the previous lecture.

In this lecture, we have learned why a headline of the form "Poll shows Candidate A to receive xx% +/– yy%" is incomplete. It has an unstated level of confidence of 95%. It means that the method used to calculate the stated range will come up with a range that contains the true population mean for 95% of the possible samples used in the poll. ■

Suggested Reading

Vic Barnett, *Comparative Statistical Inference*.

E. T. Jaynes and G. Larry Bretthorst, eds., *Probability Theory: The Logic of Science*.

David S. Moore and George P. McCabe, *Introduction to the Practice of Statistics*, 5th ed.

Questions to Consider

1. What is the relationship between making a histogram of all possible ways of choosing 1000 marbles from a collection of 100 million marbles, of which 60% are white and 40% are black, and the probability of choosing various percentages of supporters of Candidate A versus Candidate B if 60% of the voters are for Candidate A, if we take a random sample of size 1000 from a voting population of size 100 million?

2. Suppose 100 polls are conducted, each of which has a confidence interval of + or – 3%. How many of those polls would you expect to fail to contain the actual voters' preferences?

Design of Experiments—Thinking Ahead

Lecture 12

That's the challenge of experimental design, to make sure that we gather the data in such a way that we are able to draw the meaning from the data, that we are able to use the techniques of hypothesis testing and confidence intervals, to make the mathematical kind of deductions that make logical sense and that allow us to actually infer from the data the ideas of interest.

We have seen in the previous several lectures how we can interpret data about samples to support or refute hypotheses about features of a population. Often, experiments are undertaken for the express purpose of obtaining a sample to analyze. The design of such experiments is crucial to being able to make confident conclusions. One goal of good experimental design is to be able to separate influences of factors that are not of interest from those that are. Other goals are that the sample represented by the subjects of the experiment be representative of the population intended for inference and that the amount of data obtained allows us to make confident inferences. Practical considerations, such as cost, add to the challenge of experimental design. This lecture introduces such strategies as double-blind experiments as ways to meet these goals and explores where randomness has a role in experimental design.

Ronald Fisher pioneered experimental design in the early part of the 20th century. A famous case involving Fisher is told of a tea party where a lady claimed to be able to taste a cup of tea and tell whether the milk had been poured in first, then the tea added or whether the tea had been poured in first, then the milk added. Fisher took 20 cups and paired them up. In the first cup, out of sight of the lady, he randomly determined (perhaps by flipping a coin) whether to pour tea first, then milk or milk, then tea, and he used the opposite order for the second cup. He presented the two cups to the lady, asking her to determine which of the two cups had the tea poured first. He recorded whether she was right or wrong. He then repeated this experiment using the second pair of cups and so on for a total of 10 pairs (20 cups). This resulted in 10 data points (10 instances of "right" or "wrong"). The lady

was reportedly correct each time—convincing evidence that she could tell a difference.

The context of the experiment and our beliefs and understanding of possible causes or lack of causes influence how we interpret experimental results. If we can think of no physical or chemical reason for a difference, then we might be very skeptical of the result. For example, we might think that the lady saw how Fisher filled each cup, or that someone else told her, or that the "witness" wanted to make a good story. But thinking further, one may be able to imagine a physical reason that makes the difference a reality. For example, perhaps adding milk to hot tea scalds the milk in a way that could be tasted. If we can think of no possible physical difference that is potentially detectable by the woman, then we might remain quite skeptical of her abilities even after the experiment.

There are two types of ways we can go wrong in inferring a conclusion from an experiment:

- A *type I error* is where we reject the null hypothesis when, in fact, it is true.

- A *type II error* is where we do not reject the null hypothesis when, in fact, it is false.

Fisher began as a statistician at an agricultural research center. When he arrived, he found there was lots of information about crop yields, rainfall, fertilizer, and so forth, but the way the information had been collected made it difficult to draw useful conclusions from the data. Many of the variables involved were confounded with each other, meaning several different aspects of the growing conditions were changing at once, so it was difficult to tell which feature was causing what result. The kind of fertilizer, the kind of soil, and the amount of water were some of the variables. To confidently assign an effect to a kind of fertilizer, for example, we need to be able to disentangle its possible effects from the effects of the other variables.

When focusing on one variable, one basic experimental design is to fix all other variables. We can compare contrasting experiments involving two

kinds of fertilizer, for example, to see which one yields information from which we can glean meaning. One problem with reality is that variables are not fixed. Maybe soil varies from place to place in the same field. Trying to control all the variables may not be possible.

When more than one variable is of interest, one approach to experimental design is to deal with all possible combinations. Suppose we're interested in three variables: kind of fertilizer, type of corn, and soil type (hill or valley). A good experiment, proposed by Fisher, is to alternate corn types row by row, in both the valley and the hill, and alternate the kind of fertilizer by pairs of rows. The effect on an individual factor can be more easily seen. For example, if all the patches using one kind of fertilizer did better, regardless of the soil type or corn type, then the fertilizer is probably the feature that made the difference in the level of growth. In designing an experiment, we want to avoid being in the position of not being able to disentangle possible causes. We want to avoid confounded variables.

In designing an experiment, we want to avoid being in the position of not being able to disentangle possible causes. We want to avoid confounded variables.

In medical experiments, real effects often come from no physical cause, and those effects must be taken into account when testing new medications; that is, we need to consider the *placebo effect*. The fact of a placebo effect makes testing of drugs more difficult. If we give some people a new drug for, say, curing the common cold, and they get over their colds more quickly than people who did not take a drug, was it the drug or the placebo effect that shortened the cold?

Double-blind experiments avoid the placebo effect. The patient, the people who directly contact the patient or administer the drug, and the people who make judgments on the progress of the patient are all blind as to whether a real drug or a placebo is being used for that patient. Any possible psychological bias, intentional or not, is removed. Double-blind experiments are the gold standard, although the other qualities of design are still important, such as random selection from the population of who will be given the drug.

The *Hawthorne effect* is another famous example of difficulty in experimental design. Studies were done of the Hawthorne Plant of the Western Electric Company in Cicero, Illinois, between 1927 and 1932 to determine the effect of lighting, humidity, and other factors on worker productivity. It was discovered that the mere fact of doing the experiments, independent of which environmental factors were being changed, had a significant positive effect on the workers' ability to perform their tasks. Perhaps workers liked the extra attention.

Lurking variables are another challenge to good experimental designs. Lurking variables are variables that aren't being studied or controlled for but that have an effect on experimental results. For example, suppose that data show that married people in the workforce typically have a higher income than unmarried people in the workforce. We might deduce that being married is a cause of higher income. But thinking further, we realize that married people are, on average, older; hence, they have been in the workforce longer. In this case, age is a lurking variable. Placebo effect and Hawthorne effect are also examples of lurking variables.

The hallmarks of experimental design are as follows:

- *Control*: Try to control all of the variables other than the one of interest or have a fixed number of variables operate independently of each other so that the results can be disentangled.

- *Randomization*: Do not introduce your own bias during sample selection.

- *Replication*: Ensure that the experiment is replicable.

This lecture concludes the first part of this course. In the second part, we will look at application areas of statistics, starting with Lecture 13 on the law. ■

Suggested Reading

William S. Peters, *Counting for Something: Statistical Principles and Personalities*.

David Salsburg, *The Lady Tasting Tea: How Statistics Revolutionized Science in the Twentieth Century*.

Questions to Consider

1. In the lady tasting tea experiment, the lady was given 10 pairs of tea cups; each pair had one cup into which the tea was poured first and the other cup into which the milk was poured first. Do you think it would have been more or less persuasive an experiment if she had just been presented with 20 cups of tea, half milk first, half tea first, and asked to decide each one?

2. What are some of the reasons that experiments about the effects of child-rearing techniques and the nature versus nurture debate are so difficult to conduct?

Law—You're the Jury

Lecture 13

[C]ertainly, in the law, there are many places where statistical data and inferences are used to help in making legal decisions. In this lecture, we're going to look at two cases where you will be the jury.

We'll look at two main examples of courtroom drama where we imagine ourselves on the jury. We'll present the facts, and it will be up to us to determine what conclusions are appropriate to draw. We begin with a hit-and-run accident. The star witness gives crucial testimony with a high degree of credible confidence. But our interpretation of the significance of the testimony requires a little more thought. By viewing the witness's observations as a sample, we can understand the significance of the evidence in a quantitative way. The second courtroom example involves a gender-discrimination case. Do the data support an allegation of illegal discrimination on the basis of gender or do the data suggest a mere random association with gender? This example illustrates a surprising statistical anomaly known as *Simpson's Paradox*. We will present the evidence; you determine the verdict.

- Mr. Jones witnessed a hit-and-run accident involving a cab.
- Mr. Jones stated that he thought the cab was blue.
- The jury has come to the conclusion that a guilty verdict hangs on whether or not Mr. Jones's testimony implies that the probability that the cab was blue is over 50%.
- We did some experiments to determine how accurate Mr. Jones's vision was. He was able to correctly identify the blue cabs as blue 80% of the time and the green cabs as green 80% of the time. The prosecutor summarized the situation by saying that Mr. Jones is 80% sure the cab was blue; therefore, the jury should convict.

Some additional information was presented:

- There are exactly 100 cabs in the city.
- 90 are green cabs and 10 are blue.

Let's do the following thought experiment: Suppose instead of having one accident, all 100 cabs in the city simulated this accident. Each one did it once. How many times would Mr. Jones have testified that the cab was blue? Following the logic, we see that he would testify "blue" 26 times, but only 8 of those times would the cab actually be blue.

Another case that demonstrates the same statistical issue about testing for rare events arises when considering random drug testing in a company. Suppose the company has 280,000 employees, of whom 500 employees actually use the illegal drugs that are the target of this policy. Suppose the drug test will correctly read positive for 95% of people who actually use those drugs; thus, 475 employees who use drugs would receive a positive test result. Now suppose that the test gives a false positive 1% of the time; of the 279,500 employees who do not use drugs, 1% will get a false-positive result. That is, the drug test will read positive for 2795 employees who do not, in fact, use drugs.

The law provides many opportunities in which statistics plays a significant role in the dispensation of justice.

The total number of employees whose drug tests are positive will be 475 + 2795 = 3270 employees. Therefore, if an employee gets a positive test result, his or her chance of actually using drugs is 475/3270, or less than 15%. Thus, there is a great danger of inappropriate firings or accusations based on positive drug testing.

We now turn to gender discrimination. To determine discrimination by gender, race, religion, or age, it is natural to look at data about treatment of specific groups. Let's consider a case of admissions to a program at a

university. For the purposes of this example, let's assume that 1000 men applied and 1000 women applied to the program and that all 2000 applicants had exactly the same qualifications. Here are the facts:

There was a 70% acceptance rate for the men and only a 40% acceptance rate for the women. It appears that the women were clearly discriminated against, and of course, we are outraged. The chi-square test can identify the acceptance rate you would expect from random chance alone.

But let's look at the case more deeply. Suppose the total program to which the 2000 applied actually had two subprograms. One subprogram was an Excellent Program to which 200 men applied and 800 women applied. The other subprogram was a Mediocre Program to which 800 men applied and 200 women applied. Of the 200 men who applied to the Excellent Program, 20% were accepted. Of the 800 women who applied to the Excellent Program, 25% were accepted.

In the Mediocre Program, 800 men applied and 82.5% were accepted, while 200 women applied to the Mediocre Program and 100% were accepted. In each subprogram, a higher percentage of women were accepted. This situation is an anomaly, because overall, it appeared that the women were being discriminated against; however, looking at the subprograms we see a different picture.

This scenario is an illustration of a phenomenon known as *Simpson's Paradox*, that is, a situation where both subprograms indicate that the women are being treated better, yet overall, the men appear to be treated better. Given that the women actually had a higher acceptance rate in each subprogram, let's think about how we would decide whether the differences in acceptance rates are serious enough to be viewed as clear discrimination or whether the differences could be reasonably accounted for as the result of simple random luck.

Let's look at a table that records the data:

Excellent Program	Accept	Reject	Total	Accept Rate
Men	40	160	200	20%
Women	200	600	800	25%
Total	240	760	1000	24%

Mediocre Program	Accept	Reject	Total	Accept Rate
Men	660	140	800	82.5%
Women	200	0	200	100%
Total	860	140	1000	86%

How rare an event would it be to have one acceptance rate as much as 5% less than the other? To determine how surprised we should be at getting a 5% difference in acceptance rates, we can look at a normal probability curve that tells us the probability of different deviations.

The gender discrimination case illustrated a statistical anomaly known as *Simpson's Paradox*. Simpson's Paradox is an example of a possible effect of a lurking variable. In that case, the lurking variable was the existence of the subprograms.

Two other examples of legal issues arose during the O. J. Simpson trial. After evidence had been presented that O. J. had been guilty of wife beating, his lawyer, Johnnie Cochran, presented evidence that only 1 in 1000 wife beaters went on to kill their wives. Given that O.J. beat his wife, Cochran argued, there is only a 1 in 1000 chance that he went on to commit the murder. There are two fallacies here. The first regards the relative frequency of murders by wife-beaters compared to murders by non-wife-beaters. The fact that 1 in

1000 wife-beaters go on to murder their wives needs to be compared with the rate of *non*-wife-beaters who murder their wives. The rate of murder among non-wife-beaters is much smaller than among wife-beaters; thus, evidence of wife-beating increases the likelihood of murder relative to others in the population.

The second fallacy in Cochran's argument is that a wife was actually murdered. The relevant question would be, if a wife is murdered, what is the probability that she had previously been beaten? The idea of using wife beating as exculpatory evidence is ridiculous. Statistics never proves ridiculous conclusions.

Another statistical issue about trials is the following: DNA evidence is not in itself damning if the DNA was used to find the culprit. Suppose everyone's DNA were on file. A crime is committed, and the perpetrator's DNA is found on the scene. One out of a million has a specific matching DNA type. The data bank is combed, and someone matching the DNA type is arrested. At the trial, the prosecutor says, "There is only a one in a million chance that the DNA type would match." But that is bad reasoning, because DNA type was used to arrest the person. Instead, if there are 10 in the large city with that DNA type, then the probability of guilt would be only 1 in 10. The law provides many opportunities in which statistics plays a significant role in the dispensation of justice.

In this lecture, we saw how a witness's testimony was not as compelling as it at first appeared. We saw that universal or random drug tests present problems because of false positives. In the gender-discrimination case, we saw an example of Simpson's Paradox; we also saw how the principles of hypothesis testing helped us to interpret evidence as indicative of discriminatory practices or not.

In the next two lectures, we turn to statistical anomalies involved with voting. ■

Suggested Reading

Donald A. Berry, *Statistics: A Bayesian Perspective*.

David S. Moore and George P. McCabe, *Introduction to the Practice of Statistics*, 5th ed.

Questions to Consider

1. Suppose two eyewitnesses had identified the cab as blue in the car accident case. What would be the probability that the cab was actually blue? Would you convict in that case?

2. Is it possible to devise an example in which a higher percentage of men than women are admitted to a program, but upon looking at two subprograms, a higher percentage of women are accepted than men in both subprograms, and upon looking at two *sub*-subprograms in each of the subprograms, in all four sub-subprograms, the men have a higher acceptance rate than the women?

Democracy and Arrow's Impossibility Theorem

Lecture 14

In this lecture, we're going to face the big challenge of democracy, namely that the government responds to the will of the people. That's the basic concept of democracy. But the question is, "What do we mean by the 'will of the people'?"

Usually, the will of the people is ascertained through voting. An election takes the individual opinions of each voter and assembles those many opinions into one societal decision, the election winner—the will of the people. In this lecture, we will consider an unfortunate and counterintuitive reality about the election process. An election's outcome may have less to do with the voters' preferences about the candidates than with the voting method employed. A method by which the voters' preferences are combined to determine the winner is a means of making a statistical summary of data. We will see that such summaries are fraught with peril. Arrow's Impossibility Theorem proves that every election method has undesirable features.

The most fundamental idea of democracy is that the government responds to the will of the people, but what do we mean by "the will of the people"? An election takes the individual opinions of each voter and assembles those many opinions into one societal decision—the will of the people. We will see that that idea has some serious inhe`rent difficulties. From a statistical point of view, an election is a process of summarizing a set of data.

In this lecture, we will consider two counterintuitive realities in the election process. First, the voters' preferences about the candidates may have less to do with an election's outcome than the actual voting method employed. Second, every voting method is seriously flawed, and in some sense, the only self-consistent voting method is a dictatorship. These results about the election process provide a cautionary tale about difficulties associated with summaries of data.

Elections take the choices of each member of the population and return one societal choice. At first, it seems there is nothing to discuss—an election is held; whoever gets the most votes wins. We will soon see that this method of voting, called *plurality voting*, works great when there are two candidates. When a third candidate wades into the election, problems arise.

Let's consider a municipal election among three candidates—two Republicans and one Democrat. In this city, let's assume that Republicans prefer any Republican candidate over any Democratic candidate and Democrats prefer any Democratic candidate over any Republican candidate. Let's suppose that there are a few more Republicans than Democrats in the city. In fact, to ground our discussion, let's assume that there are only 22 voters total—10 Democrats and 12 Republicans. Suppose an important election is held and there are three candidates: Ron Republican, Rick Republican, and Dan Democrat. The voters' preference orders are recorded in the following table:

Rank	**8 Republicans prefer**	**4 Republicans prefer**	**6 Democrats prefer**	**4 Democrats prefer**
1st	Ron	Rick	Dan	Dan
2nd	Rick	Ron	Rick	Ron
3rd	Dan	Dan	Ron	Rick

Who should be declared the winner? There are several reasonable methods for determining the winner of an election. Let's introduce three of them in this situation. Simply count the number of first-place votes. .First-place votes are, of course, the votes a person would get if each voter just got to vote for one candidate, which is the usual method. This method of counting the votes is called *plurality voting*.

Another method would take into account the second-place preferences, as well. We could let each person vote for two candidates and declare the winner to be the person who gets the most votes. This vote-for-two method avoids electing someone who is viewed as the last-place candidate by a lot

of people. The third method we will consider weights the preferences of the voters. That is, each voter gives 2 points to his or her first-place candidate, 1 point to the second-place candidate, and 0 points to the third-place candidate. This method is called a *Borda Count*. Jean-Charles de Borda was a French scientist who was one of the pioneers in the study of voting methods.

Arrow's Impossibility Theorem makes us realize that the concept of democratic choice is an intrinsically problematic issue.

Let's see how the various candidates fare under these different possible election methods and record the outcomes in the following table. With plurality voting, we simply read across the top row of the preference table. Dan is the winner using plurality voting. With the vote-for-two method, we count how many voters put each candidate in one of the top two rows. Rick wins using the vote-for-two method. To compute the Borda Count, we give 2 points whenever a candidate appears in first place, 1 point for a second-place vote, and 0 for a third-place vote. Ron wins using the Borda Count method.

Voting Method	Ron	Rick	Dan	Winner
Plurality	8	4	10	Dan
Vote-for-Two	16	18	10	Rick
Borda Count	24	22	20	Ron

The voting method determined different winners even though the voters' opinions did not change.

We have not considered all methods. Let's now think about the possibility of a run-off. We can construct an example in which getting more support causes a winning candidate to become a losing candidate. Better is not necessarily better in a run-off method of voting.

Yet another method of voting is called *pair-wise sequential voting*. The idea is that the candidates are put in some order, Candidate 1, Candidate 2, Candidate 3, Candidate 4, and so on. Then, an election is held between Candidate 1 and Candidate 2. The winner then goes head-to-head against Candidate 3. That winner then goes head-to-head against Candidate 4, and so on, until we are through all the candidates. This voting method can elect someone when every single voter prefers a specific alternative candidate. This method fails to go along with the consensus of all the voters.

Here are three desirable characteristics that we would like to see in a voting system:

- Go Along with Consensus (*Pareto condition*): If everyone prefers one candidate to another, the lower ranked one should not win. The pair-wise sequential voting system fails this condition.

- Better Is Better: If more people vote for a winner, that person shouldn't lose. The run-off method fails to have this feature.

- Irrelevant Is Irrelevant: Suppose a candidate wins the election, then some losing candidate is eliminated; the winner should not then become a loser. Plurality fails this condition.

Arrow's Impossibility Theorem proves that no voting method is possible that satisfies the three desirable qualities:

- Go Along with Consensus

- Better Is Better

- Irrelevant Is Irrelevant

Arrow's Impossibility Theorem makes us realize that the concept of democratic choice is an intrinsically problematic issue. In the next lecture, we will see that the situation, if possible, gets worse still. ■

Suggested Reading

Donald G. Saari, *Chaotic Elections! A Mathematician Looks at Voting*.

Questions to Consider

1. Track meets often score overall team winners by awarding perhaps 5 points for first-place finishes, 3 points for second, and 1 point for third. In other words, a Borda Count method is used, in which different weights are given to the different places. Can you devise a list of voter preferences among three candidates so that under the 2, 1, 0 weighting of votes for first, second, and third choice, Candidate A wins, whereas with the 5, 3, 1 weighting, Candidate B wins?

2. Arrow's Impossibility Theorem is often portrayed as a limitation on democracy. How fundamental an issue do you find it with respect to the concept of a society responding to the will of the people?

Election Problems and Engine Failure

Lecture 15

[I]n this lecture, we're going to start out with a case study, where we look at the preferences of the people and ask the question of "Who is the people's choice?" This example actually is a real election—it was real data from an important election—so it will be interesting for you to see it and make your own decision.

This lecture begins as a continuation of the previous lecture by looking at some famous real elections in which we might wonder whether the will of the people prevailed. The challenge of choosing an election winner can be thought of as taking voters' rank orderings of the candidates and returning a societal rank ordering. An analogous and mathematically similar situation occurs in a totally different setting. Suppose we are trying to determine which type of engine lasts longest among several competing versions. One statistical strategy for making such a selection is to run several of each type of engine until they expire, then to put the experimental results in order of longevity. Of course, all is well if one type of engine always lasts longer than the others. However, in reality, the correspondence of lifetime-to-type might be less consistent. Combining those data into one choice of engine with longest expected life incurs the same paradoxical difficulties that we previously encountered in our analysis of elections.

Many people feel that if there is a specific candidate who would beat every other candidate in a head-to-head contest, then that candidate, the *Condorcet winner*, should be declared the victor.

Here is a chart that summarizes the voting preferences of the population, giving for each candidate the percentage of the population that ranked the candidate first, second, and so on.

Candidate	% of 1st	% of 2nd	% of 1 + 2	% of 3	% of 1 + 2 + 3
A	40	14	54	16	70
B	13	46	59	33	92
C	18	18	36	3	39
D	29	22	51	48	99
Total %	100	100	200	100	300

What follows is a summary of how well each candidate would have done under each voting scheme.

- In plurality voting, A would win.
- In the vote-for-two scheme, B would win.
- Using a run-off, D would win.
- Using the Borda Count method, D would win.

What candidate do you think best represents the will of the people? Many people feel that Candidate D is the best choice. Before telling you who actually won this election, let me tell you that this election was a presidential election; thus, there is one more column to be added to the table, namely, the Electoral College votes. In the Electoral College, Candidate A won handily with 180 electoral votes. Interestingly Candidate C, who was not a contender in any of the other schemes presented, actually received the second highest number of electoral votes.

Perhaps now is the time to tell you what election this was: It was the presidential election of 1860. Candidate A is Abraham Lincoln, B is Bell, C is Breckinridge, and D is Douglas. The percentages of people who rated the candidates in second and third places are estimates obtained from historians

of the Civil War. It is intriguing to think about the consequences of the statistical issue of how to summarize the data of the voters' opinions.

Candidate	Plurality	Vote-for-Two	Run-Off	Borda Count	Electoral College
A	40%	27%	46%	164	180
B	13%	30%		165	39
C	18%	18%		92	72
D	29%	25%	54%	179	12
Winner	A	B	D	D	A

We have seen lots of bad news about election problems. Now, if possible, it gets worse. Next, we will discuss the *Condorcet Paradox*. Let's look at an example of an election among three candidates, A, B, and C. The following chart summarizes the views of the 30 voters about these candidates.

Rank	10 Voters	10 Voters	10 Voters
1st	A	B	C
2nd	B	C	A
3rd	C	A	B

For every candidate, two-thirds of the people have a specific alternative candidate whom they prefer. We can produce even worse cases—with 10 candidates, for example, when no matter who is declared the winner, there is a specific alternative candidate who is preferred by 90% of the voters.

Many people feel that if there is a specific candidate who would beat every other candidate in a head-to-head contest, then that candidate, the *Condorcet winner*, should be declared the victor. Marie Jean Antoine Nicolas Caritat,

marquis de Condorcet, wanted to point out a weakness of the Borda Count method. He did so by producing the following famous voters' profile:

Rank	30	10	10	1	29	1
1st	A	B	C	A	B	C
2nd	B	C	A	C	A	B
3rd	C	A	B	B	C	A

B wins with the Borda Count. But A is the Condorcet winner. Thus, the Borda Count method does not always select the Condorcet winner. However, in a sort of voting theory double-reverse, recently, Donald Saari has pointed out an interesting further analysis of this old example. If we first erased collections of voters whose votes canceled one another, then B should win. Perhaps the Borda Count method chose the right winner.

20	0	0	0	28	0
A	B	C	A	B	C
B	C	A	C	A	B
C	A	B	B	C	A

This example shows again the subtleties of summarizing data meaningfully.

The voting method was decisive in choosing the location of the 2000 Olympic Games. The method used was plurality voting, in which the bottom-ranked choice is systematically eliminated and the remaining cities are voted upon.

Here's what happened:

- Starting with five cities, in the first three contests, Beijing was the victor.

- In the final vote, Sydney won.

Sydney, not Beijing, hosted the 2000 Olympics.

City	1st Vote	2nd Vote	3rd Vote	4th Vote
Beijing	32	37	40	43
Sydney	30	30	37	45–**Win**
Manchester	11	13	11	
Berlin	9	9		
Istanbul	7			

These voting paradoxes are examples of trying to summarize a set of data that has reflections beyond voting.

Another example: How can we tell which of three engine brands lasts longest? There is variability among the engines produced by each of the three contenders, of course, so we can't just run one engine from each company and see which lasts longest.

Here is a method called the *Kruskal-Wallis test*.

- We take several engines, say five for illustrative purposes, from each company.

- We run all the engines until they fail.

- We score the engines 1, 2, 3, …, 15 based on how long they lasted, with the longest lasting scored 1.

- We add up the scores for each company's engines.
- The lowest number wins.

	Eng. 1	Eng. 2	Eng. 3	Eng. 4	Eng. 5
A's Time to Failure	1137	993	472	256	207
B's Time to Failure	1088	659	493	259	238
C's Time to Failure	756	669	372	240	202

	Eng. 1	Eng. 2	Eng. 3	Eng. 4	Eng. 5	Total
A's Failure Order	1	3	8	11	14	37
B's Failure Order	2	6	7	10	13	38
C's Failure Order	4	5	9	12	15	45

The Kruskal-Wallis technique has defects similar to those we saw in voting methods. In this example, suppose we eliminate C's engines. Then, the recomputed Kruskal-Wallis test would indicate that B's engines are superior.

	Eng. 1	Eng. 2	Eng. 3	Eng. 4	Eng. 5	Total
A's Failure Order	1	3	6	8	10	28
B's Failure Order	2	4	5	7	9	27

All the examples in the last two lectures suggest that summaries of complex situations require contextual arguments to decide among them. Statistical and logical analysis can help a great deal in choosing which arguments to find most persuasive and which systems to use in which settings. Voting theory is an intriguing topic for further study. ■

Suggested Reading

Donald G. Saari, *Chaotic Elections! A Mathematician Looks at Voting*.

Questions to Consider

1. Strategic voting is encouraged when some people are better off voting for someone they don't really want in order to elect the ones they do want. In which of these voting methods—the Borda Count, run-offs, plurality, vote-for-two—is strategic voting encouraged? Are some methods more susceptible to strategic voting than others?

2. A voting method we did not discuss much is approval voting, in which each voter can vote for as many or as few candidates as the voter finds acceptable. Many people find this an attractive system. What are the pros and cons of this system?

Sports—Who's Best of All Time?

Lecture 16

Today we're going to take on the statistical challenge of analyzing sports statistics, which is a sport of its own.

Analyzing sports statistics is a sport of its own. We record statistics about the performances of individuals and teams, then use those data to bolster our arguments about sports prowess. In this lecture, we will examine a couple of statistical questions that illustrate principles applicable well beyond their sporting origins. We will discuss the question: "Who is the best hitter in baseball history?" This question immediately presents statistical challenges that concern comparisons of performance in different eras and different circumstances. Next, we will consider the question of streaks. Do athletes enter "the zone" and have a "hot hand" for periods of time? What is the correct interpretation of slumps and streaks? Such questions force us to confront our understanding—or misunderstanding—of what to expect from randomness. The question, "What is random?" lies at the heart of the issue of streaks and slumps.

Who was the best hitter in the history of major league baseball? This question forces us to clarify the relationship between what is measured and what quality we are trying to describe. The batting average is basically defined as the number of times the batter gets a hit divided by the number of times he is at bat. We'll also assume that we mean the batting average over a single season. And we'll ignore those players who didn't have many at-bats over the course of the season. All of the 18 highest batting averages in a season in the history of professional major league baseball occurred before 1942. Why?

We might suspect, for example, that batting averages in general, that is, the averages of all major league players, were higher in the earlier years of baseball than in recent years. Comparing the 1920 histogram to the histogram of batting averages in 2000, we find that the center, the mean, of the two is about the same (about .265), but the 1920 histogram is more spread out. The standard deviation of the 1920 data set is .050, larger than the standard

deviation of the 2000 data set, which is .038. Recall that we expect 68% of the data to be within 1 standard deviation of the mean and 95%, within 2 standard deviations of the mean.

Comparing standard deviations away from the mean is a method of normalizing the comparisons over the different eras. In a sense, it measures how well a person performed relative to his contemporaries. The number of standard deviations that a batting average is away from the mean is not necessarily an integer. Every batting average in any given year could be described by how many standard deviations it is above or below the mean. Recall that the number of standard deviations away from the mean is the z-score.

Statistics about sports are fun. They also help us to understand sports and appreciate and evaluate the success of individuals and teams.

One way to measure across eras would be to measure how many standard deviations above the mean a batter's average is. For example, given that Joe Jackson of the Chicago White Sox had a batting average in 1920 that is 2.36 standard deviations above the mean for that year and that Moises Alou of the Houston Astros had a batting average in 2000 that is 2.31 standard deviations above the mean for the year 2000, we might consider those two players about equally good batters. They are about equally extreme outliers.

We could list the 10 batters whose batting averages were the greatest number of standard deviations above the mean for their years and declare a winner on that basis. Stephen Jay Gould opines that pitching, fielding, and batting have all gotten better over the years, and their approach to human limits of perfection accounts for the lower standard deviation. Making an interpretation such as Gould's can be helpful in understanding the data.

Another complication to the question of who is the best hitter in baseball history is the fact that doubles, triples, and homeruns are more valuable than singles. Likewise, walks are not recognized, although extremely important. Other measures of offensive prowess, such as slugging percentage and

on-base percentage, can be used. People come up with various formulas combining these sorts of raw statistics, attempting to get a measure that is highly correlated with helping the team to win games.

Our second statistical issue from sports is the question of the "hot hand." A commentator is often heard to say, "This player is on a streak. He can't seem to miss." Is there really such a thing as a "hot hand," meaning that the player is better for a period of time? Or are the streaks (which are real) accounted for by random luck alone?

Suppose I flip a coin. If I flip lots of coins over and again, there will be, from time to time, long streaks of heads. We would not ascribe the streaks of H's and T's in the flipped coins to some property that has changed in the coin for that time. Likewise, if we have an NBA player whose lifetime percentage of making a shot is, say, 0.4, we would expect him, just by randomness, to have some intervals when he makes a fairly large number of baskets in a row.

The question is whether the streaks that are seen for real basketball players are explainable by randomness alone. One possible way to analyze the question of whether streaks are explainable by randomness alone is this: Suppose that when a player makes a basket, his probability of making a basket on the next shot is higher than his average. Most real data of this sort do not indicate the reality of a hot hand. As usual, our strategy is to compare the data that we find with distributions of data that we would expect to arise from randomness alone, that is, that would arise under the assumption of no hot hand. If the data are so extreme that they and their more extreme versions would happen only rarely given the assumption of no hot hand, we would take the data as evidence that there is a hot hand phenomenon. The statistical test used to measure rarity in this example is called a *chi-square test*. In the example considered in the lecture, the data were not sufficiently extreme to reject the assumption of no hot hand. Thus, the data do not warrant the conclusion that there is a hot hand.

Trying to distinguish randomness from some other cause is difficult. There are many ways to look at a string of data and ask whether the string is explained best as a random process or as a process that is influenced by

something internal to it. There are many ways of looking for patterns. There is room for interpretation.

These two examples of sports analysis have brought up many statistical issues. Trying to find the greatest hitter in baseball history brought up questions about the relationship between what we measure, what we want to know, and how to compare performances in different eras or under different circumstances. The hot hand issue brought up fundamental questions about the nature, meaning, and measure of randomness. ■

Suggested Reading

Jim Albert and Jay Bennett, *Curve Ball: Baseball, Statistics, and the Role of Chance in the Game*.

Stephen J. Gould, *Full House: The Spread of Excellence from Plato to Darwin*.

Michael Lewis, *Moneyball: The Art of Winning an Unfair Game*.

Questions to Consider

1. What method would you use to select the greatest athlete of the 20th century? (I choose the 20th century rather than all time because we would not have much reliable data on earlier athletes.)

2. If randomness with a certain probability really accounts for sports performance, does that lessen the interest in watching sports? Would it change the approach to sports psychology and the training strategies?

Risk—War and Insurance

Lecture 17

Today's lecture concerns two risky businesses: war and insurance.

In World War II, the serial numbers on captured Mark V German tanks were used to deduce the number of Mark V tanks produced altogether. We will use that scenario to introduce a variety of methods of inference and to analyze how different plausible methods might be compared for expected accuracy. Risk closer to home occurs when we deal with insurance. Insurance is an industry based on probability. In determining whether buying an insurance premium or extended warranty on a product is a good investment, we deal with the statistical ideas of expected value and the distribution models of product lifetimes.

In World War II, one of the challenges of the Allied intelligence officers was to estimate the strength of the German fighting machine. In particular, one wanted to estimate the number of tanks that the Germans had manufactured and were using in battle. During World War II, when a German tank was captured, analysts noticed that the tanks had serial numbers and it appeared that the serial numbers were consecutive, starting with 1 and increasing as each new tank was built.

Statisticians approached the situation as a statistics question. We know some information about part of the population, and we wish to infer information about the whole population. We assume there are a certain number of tanks in the German army, numbered from 1 to N. We assume that the tanks captured are a random sample from the whole population of tanks. We would like to estimate the total number of tanks. We'll look at possible *estimators*, that is, methods or strategies for calculating an estimate.

Let's do a specific example. Suppose we've captured tanks whose numbers are {68, 35, 38, 107, 52}. What estimate for the number of total tanks would we make? One idea might be to take the mean of the five numbers, double the result, and subtract 1—giving an estimate of 119. Another method for estimating the midpoint of the numbers 1 to N would be to take the median

value of the sample. Then, we would double that, giving an estimate of 104. This estimate is clearly too small, because it is less than the number on a tank we actually captured. In fact, both of these strategies can produce an estimate that is less than the highest numbered tank we have actually captured.

Another strategy for guessing the midpoint of the total tank numbers would be to add the biggest and smallest numbers that we've captured (107 + 35) and take their average (71). We would double that average to get our estimate (142). This method always produces a number bigger than the largest in our sample. Thus, this estimator doesn't suffer the previous method's flaw. These various estimators each have some intuition behind them, but there are actually other strategies that are superior.

Another desirable quality of an estimator is to have as small a variance as possible, because that would mean that the estimator is, on average, close to the true value.

We'll discuss what qualities different estimators can have in order to guide our decision about which method, among reasonable-sounding methods, to use. One feature that a particular method of generating estimates (an estimator) can have is that the method maximizes the probability of choosing the sample we actually got. $N = 107$, the maximum number on a captured tank, would maximize the chance of capturing our collection. However, our intuition tells us that a good estimator would estimate a larger number of tanks than the largest number we have actually captured. In this case, the *maximum likelihood estimator* (as such an estimator is called) doesn't seem to be reasonable.

One property that we might want a method to have is that in performing the method many times and taking the average of the estimates the method produced, we would, on average, get the true number of tanks. This average of the estimates is called the *expected value* of the estimator. An estimator whose expected value is the correct value is called an *unbiased estimator*. For example, the 2 × sample mean – 1 estimator is unbiased because the sample mean is an unbiased estimator of the population mean. But we saw

that this estimator had other drawbacks (namely, producing values that are definitely wrong).

Can we think of a strategy (an estimator) that is unbiased and doesn't give us answers that are definitely wrong? Yes, we can.

Estimate = $\frac{k+1}{k}\max(x_1, x_2, x_3, ..., x_k) - 1$.

In our example, $\frac{5+1}{5}\max(68, 35, 38, 107, 52) - 1 = \frac{6}{5}107 - 1 = 127.4$.

This method is unbiased. That is, if we use this computation on every possible sample, then the average of those estimates will be *N*. The expected value of the estimator is *N*, the quantity we are trying to estimate. That is, we can't be sure that the estimator will give the correct value, but on average, it will give the correct value.

Another desirable quality of an estimator is to have as small a variance as possible, because that would mean that the estimator is, on average, close to the true value. It turns out that the estimator previously defined (that multiplies the maximum of the sample by $(k + 1)/k$ and subtracts 1) is the *minimum variance unbiased estimator*. This study of tanks brought up the idea of expected value, which is a central idea in the risky business of buying and selling insurance.

One of the practical ways of tempering the vagaries of risky life is through insurance. The whole concept of insurance is based on statistics. We can view insurance as a game of chance. Our decisions on whether to buy an extended warranty, health insurance, or other insurance are best based on understanding the distributions of the foreseen calamities that the insurance is aimed to mitigate. But most people are generally not good at gauging large numbers or rare events.

We can view an insurance company that sells insurance to many people as playing the same game with many people. The company needs to consider the distribution of possible payouts. To illustrate, we'll consider the following

game: We shuffle a deck of 52 cards. You draw a card. If it is the queen of spades, the insurance company pays you $100. If the game is played by 1000 customers, the distribution of the number of payouts is binomial, with $p = 1/52$ and $n = 1000$. Thus, we would expect about 20 people out of 1000 to be winners. For that distribution, 98.7% of the time, the number of winners (payouts) will be between 0 and 29. If the company has enough money on hand to settle 29 claims, it will be 98.7% sure that it won't run out of money.

Now, let's take the example of extended warranties on electronic items. The distribution of the time to failure on a new electronic item is not just a smooth distribution that declines over time. Instead, it may be more like a bimodal distribution. Electronic items that are going to fail sooner than average will tend to fail almost immediately because they were never made properly. But if they are working after a few months, then they are likely to continue to work until much later, when they come to the end of their expected life.

Thus, the distribution is bimodal, with one peak near the beginning of use and another after some prescribed length of service. The extended warranty really only covers the period between the end of the manufacturer's warranty and the time when the manufacturer thinks that second peak will occur. That tends to indicate that those kinds of insurance policies may be particularly poor values.

From risks in war to risks in insurance, statistical analyses pay good dividends. ■

Suggested Reading

David S. Moore and George P. McCabe, *Introduction to the Practice of Statistics*, 5th ed.

Questions to Consider

1. Suppose you captured tanks numbered 25, 64, 253, 135, and 85. Assuming that you were selecting tanks randomly from ones numbered sequentially, what would be your best guess for the number of tanks that exist altogether?

2. None of the methods we talked about concerning guessing the number of tanks made use of the order of the numbers of the captured tanks. Under the conditions of the question, that is, that the tanks are captured randomly, could the order in which the tanks were captured be a relevant factor?

Real Estate—Accounting for Value

Lecture 18

In this lecture, we're going to confront a real-life problem and see what statistical techniques are involved in solving it. The problem that we're going to think about is producing assessments of the market values of houses in a city.

Tax authorities often need to set valuations for each house in the tax district. Because some of the houses have sold during the year, their market values are known; however, most houses were not sold. The challenge is to use the data about the sold houses to assess the values of all the houses. This situation is a classic example of statistical inference. Using multiple linear regression, we can find a formula that predicts the selling price of a house based on measurable quantities, such as square footage, number of bathrooms, distance from the city center, and so on. A case study illustrates how such multiple linear regressions are done.

The goal of this lecture is to give some sense of the types of issues that we confront when actually doing a real-life problem. Probably you will not feel the need to follow every detail, but you will get a sense of the rhythm of a multiple regression analysis. In this lecture, we confront the real-life problem of producing assessments of the market values of houses in a city. In most of our lectures so far, we have dealt with one or, at most, two varying quantities. Our goals here are to organize, describe, and summarize data when multiple variables are involved. We wish to know which quantities affect, explain, or are related to which others (and to what degree). Square footage, lot size, number of bedrooms, number of bathrooms, distance from city center, and other factors all influence the market value of a house.

The square footage seems likely to be the most influential single variable. Our data consist of the square footages and the selling prices of a collection of 113 houses sold during the last year in our example city. We first describe a *linear regression* using square footage as the *explanatory variable* and selling price as the *response variable*.

One step is to look at the two variables independently. Specifically, we can draw a histogram and compute the mean and the standard deviation of the square footages and perform similar calculations for the sales prices. The resulting graphs look similar and both have right skew.

We can visualize the relationship between these two variables by making a scatter plot of the two variables. The cloud of points looks roughly linear going up to the right. We can approximate the scatter plot by the least squares regression line. The difference between the regression line's second coordinate and the data pair's second coordinate is the *residual*. Squaring each such difference and adding them up gives a sum of squares. That is, we are taking the sum of the squares of the residuals.

The *least squares regression line* is the line for which the sum of the squares of the residuals is least. Software can compute a formula for this line. In our example, the equation is: Price = \$161 × square footage – \$63,600.

The slope of the regression line tells us how much change in the y variable is expected from each unit change in the x variable. In other words, the slope tells us how much more we would pay with each additional square foot. In our example, that slope is \$161 per square foot.

The view of how we are thinking of the data is summarized by: Data = Model + Residuals. How well does that summary capture the actual data set? We know that the values of the second coordinate (in our example, the house prices) vary. The variance (the square of the standard deviation) of the house prices is a measure of how spread out those prices are. Recall that correlation measures how closely two quantities move together. In this example, the correlation between square footage and house price is .835. Now we see how the variance of the house prices compares to the variance of the amounts that the house prices differ from the values predicted by the regression line. The square of the correlation is equal to the fraction of the variation in the prices that is explained by the square footage.

We now turn our attention to what we do when there are more variables being used to explain a variable, in this case the selling price of the house. Suppose for a collection of houses we know:

- Age of the house in years
- Number of bedrooms
- Number of bathrooms
- Distance from city center in miles
- Number of garage parking spaces
- Size of the lot in acres
- Number of floors of living space in the house
- One response variable, the selling price of the house

How can we deal with this more complicated situation? We do a multiple linear regression. Here's what that means. *Multiple linear regression* is a technique by which we can approximate or summarize a situation where there are several explanatory variables influencing the response variable. We will use the concepts that we developed for the case of paired variables, such as square footage and price, and follow the same pattern of analysis for several variables. We know already that square footage is correlated with house price; however, if we did not already know that one or more of our variables had predictive value, we would do an analysis of variance (ANOVA), which determines such predictive value.

The idea of multiple regression is that we find coefficients for each of the explanatory variables so that they combine to predict the house price. The output of running a multiple regression program gives us the least squares coefficients for each of the variables. Each coefficient can be interpreted as the expected amount of difference in the price of the house from increasing

the explanatory variable by one. For example, one more acre raises the house price, on average, by $49,200. On the other hand, adding a bedroom appears to decrease the value of the house, perhaps because in given houses with the same square footage, one with fewer bedrooms has larger rooms, a feature generally associated with higher-quality houses. The point is that the multiple linear regression produces a way to predict house prices if we are given values of the explanatory variables.

The output of a multiple regression program typically also provides additional information. Among other information, the output includes an R^2 value, which tells us what fraction of the variation in the house prices is captured by this model. In our case, 77.7% is captured by the model, which means that 77.7% of the variation in house prices is explained by our predictive model that used square footage, lot size, distance from center of town, and number of bedrooms.

Multiple linear regression **is a technique by which we can approximate or summarize a situation where there are several explanatory variables influencing the response variable.**

The strategy of doing a multiple regression analysis is that we find a model that predicts the response variable as a combination of the explanatory variables. We measure how well the model fits by measuring how much difference there is between the predicted values and the actual data. We can determine what percentage of the variation of the actual value is explained by each variable or by any set of variables. Having established such a model for house prices based on all the houses that were actually sold during a year, the model might be used by the tax department to produce market valuations of all houses in the city. ■

Suggested Reading

B. Bowerman, R. O'Connell, and A. Koehler, *Forecasting, Time Series, and Regression: An Applied Approach*, 4th ed., part II.

R. Dennis Cook and Sanford Weisberg, *Applied Regression Including Computing and Graphics.*

David S. Moore and George P. McCabe, *Introduction to the Practice of Statistics*, 5th ed.

Questions to Consider

1. In our example of multiple regression, the constant term was negative. That seems to imply that if the house had 0 square feet, 0 bathrooms, etc., then it would cost some negative amount. Does that feature imply that the model is wrong? What is an interpretation of it?

2. In linear regression, the scatter plot is approximated by a line. Can we interpret the multiple regression approximation in some geometrically meaningful way?

Misleading, Distorting, and Lying

Lecture 19

You may remember Mark Twain's quote that he attributed to Disraeli, that, "There are three kinds of lies: lies, damned lies, and statistics." In this lecture, we're going to embrace that quote, and talk about how to distort, mislead, and lie with statistics.

In this lecture, we will learn some effective ways to lie with statistics. Lying with statistics means one of several things. We might, of course, simply present false data. But more interesting methods involve taking perfectly valid data and distorting their meaning by using misleading presentations or by drawing improper inferences.

Here's one example of several we'll explore: A large college wishes to advertise that it has small classes, so it creates 99 one-student classes, then makes one class contain the remaining 901 students. Because the college has 100 classes and 1000 students, it advertises that its average class size is 10. But 901 of the 1000 students experience a class size of 901. The mean often does not suggest a meaningful story. By examining several misleading uses of statistics we learn to recognize the inadvertent or purposeful misuse of statistics.

Mark Twain attributes the following quotation to Disraeli: "There are three kinds of lies: lies, damned lies, and statistics." Techniques of analyzing and presenting statistical data can be misused, intentionally or unintentionally, to give distorted views of the world.

Here's a misleading statistical fact: The average American has one testicle and one ovary. The statement is correct but completely misleading. There is a lurking variable: sex. Bringing the lurking variable to light gives quite a different view of the data. For many cases of existing data, we don't know whether or not there is a lurking variable.

Here's an example showing that outliers can have so large an effect on the mean that merely stating the mean gives a distorted view of the actual income distribution. In a recent year, the mean increase in net worth of graduates of Lakeside High School in Seattle was more than $2,000,000. But the reason is not that a lot of the graduates make millions of dollars a year. The reason, instead, is that the average included Bill Gates and Paul Allen, graduates of Lakeside High School and the founders of Microsoft Corporation—extraordinarily wealthy outliers whose net worth greatly distorted the mean. A much better measure of the center than the mean would be the median, which is not affected by outliers.

Even something as simple as class size in a university is a bit tricky to summarize. Let's do an extreme example, making the unrealistic assumption that each student at the university takes just one class. Suppose there are 1000 students, with 901 in one class and with each of the remaining 99 students individually tutored. The mean class size is 10, an accurate but misleading summary. The median class size is 1, which also is misleading. A better summary value for expressing the average experience of class size by a student is to add the class size experienced by each student ($901 \times 901 + 99 \times 1$) and divide by 1000. The result is 811.9, a better summary number. The median "class size as experienced by the students" is 901, also a good summary.

Here's a misleading statistical fact: The average American has one testicle and one ovary. The statement is correct but completely misleading.

Biased samples, some intentional, some not, are common. Using a sample that is not representative of the whole population gives a distorted view. In an earlier lecture, we saw the example of the *Literary Digest*'s biased sample. As a source of understanding our world, our friends form a biased sample. The people we know, on average, tend to be like us. We can easily believe that everyone thinks like us because every time we ask our friends about something, they tend to agree with us more or less.

The wording of questions in a survey can influence results. Surveys can either intentionally or unintentionally have questions worded in ways that affect the way people answer them. Consider this question: "Would you rather have: the very risk-taking Smith, or Jones, who is likely to save us from desolation?"

Virtually any news source is biased, in the sense that its contents are chosen for interest. Frequently, the interest in a story comes from being rare and being bad. Television news is biased toward stories with visual content. In a recent year, the rate of death from terrorist attacks in Israel was .00038, which is one-third the rate of death from traffic accidents in the United States. But terrorist attacks are better news stories. Any news source is biased, but we must realize that an unbiased news reporting system would be dull. Through selective reporting, statistics can be manipulated to persuade people that a particular drug is effective, even though it is not, or to predict falsely the future of stock market prices.

People can answer surveys with wrong answers. Sometimes people give the answer to a question that they think the questioner would like to hear. Many studies showed that child molesters were frequently molested as children. Some recent studies put those data into question. Child molesters may gain some advantages in the legal system by claiming to have been molested as children.

Graphs or phrasings can be distorted. To make a change that is small percentage-wise look large in a graph, omit from the graph most of the possible range of the quantity. Another way to have accurate graphs that are misleading is to draw them so that the height of the graphical symbol accurately represents the issue, but it is drawn to look like a three-dimensional thing; we intuitively think of the thing's volume, which of course, increases much faster than its height does.

Consider this statement: "People who eat a particular food have a 30% higher chance of contracting a certain rare disease." The rate may go up from 1 in 10,000,000 people getting the disease to 1.3 in 10,000,000. Yes, this is a 30% increase, but the increase may be insignificant or it may not. The simple statement "30% increase" doesn't tell the story. Likewise, suppose we

read that Company X has increased its profits by 50% over the last quarter. This impressive statistic is not necessarily so impressive if the profit in the preceding quarter was $1000.

Extrapolating trends mindlessly can give ridiculous conclusions. Some of the scare stories we hear are the result of extrapolating trends. Trends in economic growth or population growth are a common subject of inappropriate extrapolation. In sports, world-record running times cannot continue a linear trend because there are physical limits to how fast people can possibly run.

Confusing correlation with causation is a big source of the misleading use of statistical information. People often start with a true correlation but then derive a false causal relationship from it. Lurking variables often underlie such misconceptions.

Statistics can be an incredibly useful tool; however, we must be cautious as consumers of statistics to avoid being taken in by the pitfalls intentionally or unintentionally included in the presentation or interpretation of statistical information. ■

Suggested Reading

Darrell Huff, *How to Lie with Statistics*.

Questions to Consider

1. Many prisoners cannot read. Is it an important insight or a logical fallacy to argue that programs that teach reading to prisoners might reduce rates of recidivism?

2. To some extent the use of statistics should come with a "buyer beware" label. Where should the ethical and legal responsibility lie with regard to the presentation of statistics in a meaningful way? For example, should misleading (but technically correct) graphs be prohibited along with literally false advertising, or not?

Social Science—Parsing Personalities

Lecture 20

In this lecture, we'll discuss two statistical issues that arise from the social sciences.

Social policy and social sciences rely on the interpretation of statistical data. This lecture discusses two separate topics related to the application of statistics to social science. The first is a statistical technique, *factor analysis*, which can shed light on what quality several correlated, measured quantities all might be measuring. The technique seeks to identify underlying latent factors that explain correlation among a larger group of measured quantities. The other topic is possible limitations of hypothesis testing.

In this lecture, we will discuss two aspects of statistics that arise in the social sciences. The first is a technique called *factor analysis*. The second part of the lecture discusses an issue within the social sciences community concerning the over-reliance on hypothesis testing and its validity.

Factor analysis is a statistical technique that tries to find whether data comprising a number of variables can be summarized, or explained, by a smaller number of "factors." Charles Spearman, studying intelligence, is credited with inventing the technique of factor analysis about 100 years ago. He hypothesized that there is one underlying factor of general intelligence (called the *g factor*) that underlies results of various other measures of mathematical and verbal skills.

The assumption underlying factor analysis is that there is a small group of latent factors that accounts for the correlation among a larger group of observed variables. For example, a questionnaire might ask 50 questions about emotions, each with a numeric answer, such as, "How much fear are you feeling?" "How much control are you feeling?" and so on, yielding 50 observed variables. The factors found in factor analysis are chosen specifically so that they have no correlation. They represent independent

characteristics (of a person). A successful factor analysis would yield a small number of factors that explain much of the total variation in the original data.

We'll begin by looking at the famous Myers-Briggs Personality Type Indicator. After answering about 100 multiple-choice questions, the Myers-Briggs test presents us with a summary of our personality or preferences using four scales: (1) extraversion/introversion, (2) sensing/intuition, (3) thinking/feeling, and (4) judging/perceiving. The results show us where on each scale our answers put us. A technique called factor analysis, though not used historically in the development of the Myers-Briggs indicators, describes how the 100 questions give rise to four axes. The idea is that the answers to the 100 questions can be combined in specific ways to reveal a separate rating for extraversion/introversion; another combination gives the answer to thinking/feeling, and so on.

Factor analysis is a statistical technique that tries to find whether data comprising a number of variables can be summarized, or explained, by a smaller number of "factors."

Finding the combinations that give uncorrelated combinations is a mathematical procedure. The researcher calls each factor an evocative summary name, depending on the ingredient variables. For example, in a study of jealousy, 20 measures of qualities were reduced to 3 factors, called by the researcher Reactive Jealousy, Anxious Suspicion, and Interpersonal Insecurity. The correlations between each of these 3 factors and the 20 original variables served to divide the 20 into 3 groups. The variables in the first group are highly correlated to the first factor and usually have low correlation with the other two factors, and so on. Part of the intent is that something has been learned about jealousy, namely, that there are 3 principal ingredients that underlie the original 20 variables. Care must be taken in interpretation, but insight can be gained about the psychological or social issue being studied, as reflected in the factors that are a good mathematical model or summary of the data.

We now turn to some issues concerning hypothesis testing. Hypothesis testing is an example of a strategy for testing features of our world, but some social scientists and many others feel that there has traditionally been too heavy a reliance on hypothesis testing as a way of adducing and interpreting evidence.

Another model of progressing toward a clearer understanding of our world may be closer to the way we often proceed in real life. How do we come to evaluate people in our world? We have a sense of a range of what they might be like, but we don't know. Then, we update our view. This strategy underlies a statistical point of view known as *Bayesian statistics*.

Perhaps the best way to understand the distinction between this updating model and standard hypothesis testing is to examine the following three experiments—all hypothesis-testing experiments with statistically identical results. In the first experiment, a musicologist is presented with 10 pairs of sheet music, and he must determine if the composer is Mozart or Haydn. He is correct 10 times out of 10. We are already familiar with the second experiment of the lady tasting tea, in which a lady was able to taste 10 pairs of cups of tea with milk and tell correctly each time whether the milk or the tea had been poured in first. In the third experiment, a drunk claims he can tell whether a coin will land heads or tails every time it is flipped. And, indeed, he does guess correctly 10 times out of 10.

Most people would say that the Mozart/Haydn experiment was very persuasive and the lady tasting tea less so. With the drunk person, we retain a good deal of skepticism. We are updating the prior assessment of our view of reality. In the Bayesian point of view, we view our assessment of the world as a probability graph rather than a fixed number. It is an interesting philosophical perspective that takes our everyday experience and captures it on a mathematically sound footing. ■

Suggested Reading

Vic Barnett, *Comparative Statistical Inference*.

Donald A. Berry, *Statistics: A Bayesian Perspective*.

Donald A. Berry and Bernard W. Lindgren, *Statistics: Theory and Methods*, 2nd ed.

B. K. Gehl and D. Watson, *Defining the Structure of Jealousy through Factor Analysis* (available at http://www.psychology.uiowa.edu/students/gehl/definingjealousy.doc).

David S. Moore and George P. McCabe, *Introduction to the Practice of Statistics*, 5th ed.

Questions to Consider

1. What combination of environmental and genetic factors most influences a person's desire to learn? What data would you seek to specify such combinations?

2. Suppose you flip a coin and cover it up before seeing how it landed. Do you believe it is meaningful to say, "There is a 50% chance it is heads?" or do you feel that because it is either heads or it's not, an assertion about its "headedness" is not susceptible to a probabilistic analysis?

Quack Medicine, Good Hospitals, and Dieting

Lecture 21

Every day, we make decisions about our personal health; it's one of the most important things that we think about. Our state of health is often measured by data. The kind of data that we are interested in are things like our weight, our blood pressure, the levels of our cholesterol, and other various chemicals in our blood stream. These data present to us a picture of our well-being.

Note: References to specific diseases and treatments and all data and interpretations are used solely for the purpose of illustrating ideas of statistics and are in no way designed to be used as medical references for the diagnosis or treatment of medical illnesses or trauma.

We make decisions every day about our personal health. If our cholesterol is above some number, our doctor is likely to suggest a medication to supplement healthful eating and exercise. Such a recommendation is commonly based on statistical results of studies that show a correlation between high cholesterol and increased risk of heart attack. Aspects to consider in applying results of such a study to ourselves include how like us the people in the study are and the difference between correlation and causation. Another statistical concept, *regression to the mean*, explains why quack medicine may appear to work often.

Part of our decision-making about such questions as whether to take cholesterol-lowering medication is done by comparing our numbers to those reported in studies that have been conducted with a large number of people. We may need to evaluate studies to see if they reveal significant and important information or if they yield only statistically significant information, meaning a difference is detectable, but not important. As well, the studies frequently involve many people who in many ways, such as age, weight, inheritance, or gender, are not like us.

What we would really like is a study involving people who are as much like us as possible, because their experience with, say, a medication is more apt to be similar to ours. We'd like to *condition* the overall study data on several variables, looking at the subset of the study data that matches us with respect to those variables. In general, the concept of *conditioning* the data on some criteria means that we look at only the data with a certain feature. We prefer to respond to the statistical results conducted on people most similar to us so that we would have a sense that the studies pertained most specifically to us.

Part of our decision-making about such questions as whether to take cholesterol-lowering medication is done by comparing our numbers to those reported in studies that have been conducted with a large number of people.

An intriguing possibility is to use the computational capabilities of modern technology to find people who were in the scientific studies and whose characteristics are similar to our own. We'd then perform statistical analysis on that subpopulation for, say, the effect of a cholesterol-lowering drug. With that approach, we might get rather different statistical results than results based on the larger population.

Tailoring the statistics to our individual personal health situation could conceivably provide a greater increase in personal health than would improvements in treatments themselves. Of course, doctors are aware that different patients react differently, and to some extent, they try to tailor their treatments to the individual. Using their experience to modify the results of the studies can be good, but it can also be problematic, because an individual doctor sees many fewer patients than there are in some studies.

Another statistical issue arises when we need to decide which of two hospitals to go to for heart surgery. Suppose we have the following chart summarizing how successful each hospital is with each of three subcategories of patients: those entering in fair condition, in serious condition, and in critical condition.

Patient	Hospital A	Hospital B	Survivors from A/%	Survivors from B/%
Fair	700	100	600/86%	90/90%
Serious	200	200	100/50%	150/75%
Critical	100	700	10/10%	300/43%
Total	1000	1000	710/71%	540/54%

Looking at the data broken down in this way, we see that B has a higher success rate in all three categories of difficulties. When averaged all together, however, the impression is that hospital A is superior, with a 71% survival rate. But hospital B is superior in each of the three categories. Notice that mathematically, this is the same kind of example as one we saw in an earlier lecture on gender discrimination. This is Simpson's Paradox.

Another statistical question arises when trying to distinguish between a quack medicine and a real medicine. The reason that quack medicines sometimes seem to work is that usually people recover from a minor disease even without medicine. Quack medicines appear to work because of the phenomenon called *regression to the mean*: An ill person usually expects to return to his or her mean health situation.

A similar example arises in childrearing. Suppose (somewhat tongue-in-cheek perhaps) that what you say to your child has no effect on the child's behavior. Under that supposition, after praise or punishment, the child will usually return to his or her average behavior (this follows from the definition of *average*). It will appear that punishment works (because the child who was misbehaving will usually improve), but praise has the opposite of its intended effect (because the child who just did something extra good and got praise will usually return to his or her average behavior). This is similar to quack medicine; both are examples where something appears to have an effect, but the explanation is really regression to the mean. Other examples of regression to the mean are athletes doing poorly after appearing on the

cover of *Sports Illustrated*, tall people having children somewhat shorter than themselves, and short people having children somewhat taller than themselves.

A common issue in the realm of personal health management concerns dieting and weight. Many quantities, including weight, have an innate variability. If you weigh yourself daily and chart the results, you'll notice that the readings can differ by a pound or two, even if you didn't "gain" or "lose" weight. If you weigh yourself to the nearest half a pound, then make a histogram over the readings, you get a somewhat normal-shaped picture distributed around a central value. A weight chart is a *time series*, a value for different times. In the case of a person losing weight, the weight chart shows a downward trend, but it does not always go down uniformly each day. We can summarize the chart by drawing a straight line, the least squares regression line. The weight loss data illustrate how we use a mathematical model (the straight line) that captures the spirit and look of the data to summarize the data.

Several statistical issues are related to everyday health issues. We discussed the notion of conditioning on subpopulations that match a person of interest, which can give different results than analysis over the whole population. We illustrated Simpson's Paradox by the example of choosing a hospital. We illustrated how regression to the mean can be the explanation for quack medicine seeming to work and for punishment to work but praise not to work. We saw how to summarize a time series by a straight line. ■

Suggested Reading

David S. Moore and George P. McCabe, *Introduction to the Practice of Statistics*, 5th ed.

Ann E. Watkins, Richard L. Scheaffer, and George W. Cobb, *Statistics in Action: Understanding a World of Data*.

Questions to Consider

1. Suppose you have not gained or lost weight in many months. Suppose you notice that for a period of a week, your daily weight is always less than your traditional average. Can you conclude that you have lost weight? Suppose a month passes in which you are always below your traditional average. Where is the cut-off line at which you will assert that you actually weigh less?

2. What data about a medicine would you most like to know before taking the medicine? Your answer might include rates of spontaneous recovery, seriousness of the disease, measurement variation to determine that you have the disease, characteristics of the treatment studies, or other values.

Economics—"One" Way to Find Fraud

Lecture 22

In today's lecture, we're going to be talking about economics. Economics is certainly one of the most common arenas in which we think of statistical data as being centrally important, and it is.

Economics is one of the most common arenas in which statistical data are important. Data arise when we measure incomes and wealth, the balance of trade, the deficit, the stock market, the consumer price index, and employment levels. All these data are indications of the economic condition of the world and make real differences in our daily lives. The consumer price index influences such things as tax rates and Social Security benefits. Looking at historical trends is a suggestive method of getting a sense of how the world is changing. However, in some cases, looking at historical trends is a poor method for predicting the future. For example, it is a bad strategy to buy mutual funds based on last year's performance. A surprising feature of tables of data in economics is that the leading digits of numbers do not occur with equal frequency. This unexpected reality, whose full formation is called *Benford's Law*, gives us an unexpected statistical method for detecting fraud.

The national debt reflects one aspect of the country's financial situation. The Dow Jones Average reflects the business sector of the economy. We will look at trends of these indicators over time to get a historical perspective.

The CPI is an important economic indicator. The CPI measures the change over time in the price of the items that people buy in day-to-day living. The basic idea of the CPI is that it looks at how much various items cost from month to month. If the cost of items goes up, then the CPI records the amount more that we must pay to purchase the same items. The CPI has the goal of giving a sense about how much, more or less, it costs real people in the United States to buy a cross-section of goods and services. To produce the CPI, a specified collection of goods and services is used, called the *market basket*. Each month, people go out and see how much it would cost to buy that market basket. The cost from month to month is compared.

There are certainly many features that complicate the simple basic idea of the CPI. One change is that people buy different things as different products come on the market. It would not be sensible to compare only the items in the 1965 market basket with the costs of those same items today. Some items produced then would not be purchased today at all. Many items, such as computer-based goods, would not have been invented in 1965, yet form a significant part of consumer purchasing today. All these features lead to complicated strategies for adding and removing items from the market basket. If, over a certain number of years, it costs twice as much to buy the same collection of things, then we would be able to assert that inflation has doubled the prices. The CPI, then, gives us a sense of what a dollar is worth.

***Data mining* refers to the process of looking at an existing collection of data to find patterns or trends.**

When we say that, in 1970, a beginning teacher earned a particular amount and, today, a beginning teacher earns a different, undoubtedly higher, dollar amount, how do we know whether teachers' pay has increased or decreased? We would not know until we compared the value of a 1970 dollar to today's dollar. We would find that the CPI tells us that $1.00 in 1970 is equivalent, in some sense, to about $4.50 today. Thus we could find out if teachers' beginning salaries have, on average, decreased, increased, or remained about the same now as then.

Many social programs are legally influenced by the CPI. Social Security benefits for 50 million recipients are adjusted by a formula that uses the CPI. Federal civil service and military pension payments change based on the CPI. The Food Stamp program changes the payment to the more than 20 million Food Stamp recipients. The CPI changes the cost of school lunches. The CPI is used in the federal income tax to adjust tax brackets and the standard deduction. Many collective bargaining agreements involve the CPI to protect salaries and benefits from the effects of inflation.

If we want to understand how various parts of the economy are changing independent of inflation, we can use the CPI to put measures of wealth, income, and so forth in terms of constant dollars, thus allowing us to compare

economic conditions over time in a more meaningful way. A graph of the CPI over time tells us the value of a dollar at each time during the last 100 years. The national debt also shows trends that we can graph over time.

Now we can look at the Dow Jones average, which is the sum of the costs of a particular set of stocks. It shows a consistent increase, and even if we adjust for inflation, we still see a very sharp increase in the late 1990s. Looking at the Dow Jones average causes us to think about stocks and investing. Investing involves looking at data and trying to predict future performance.

Data mining refers to the process of looking at an existing collection of data to find patterns or trends. Data mining can be a very valuable strategy for identifying features of the world; however, there are dangers. In large sets of data, we expect there to be patterns that occur by random chance alone. We would also expect rare events to occur by chance. The appropriate use of data mining is to find patterns, then undertake new experiments to confirm or reject the hypothesis suggested by the mined data.

Suppose we use a data-mining technique as a means to choose a good investment opportunity. Our goal in choosing a mutual fund is to find one that is likely to go up. It is a natural strategy to simply look at which mutual fund increased in value most during the last year and buy it. Unfortunately, that reasonable-sounding strategy is a poor investment strategy. If we look at how well we would have done had we adopted the "buy the best of last year" strategy, we would see that our investments would actually lose money in many years and definitely be a poor investment overall.

An analogy that makes this point very clearly is the lottery. Suppose that, among investment strategies, we included buying lottery tickets. We would find that, among the possible investments, there was one $1 investment that earned $100 million. Following the logic of investing in the manner that worked best last year, we would invest in lottery tickets—perhaps we would choose to select the same numbers as the winning ticket. This investment decision is unlikely to be a good one.

Data mining leads us to an interesting phenomenon in economics: *Benford's Law*. Physicist Frank Benford did a study of some 20,000 various data

sets of numbers and discovered that approximately 30% of the numbers began with 1, rather than about 11%, as expected (1 out of 9). He formulated Benford's Law: The proportion of numbers beginning with 1 is $\log_{10}(1+\frac{1}{1}) = .301$ (around 30%); with 2 it is $\log_{10}(1+\frac{1}{2}) = .176$ (17.6%); with 3 it is $\log_{10}(1+\frac{1}{3}) = .125$ (12.5%); and so forth, until with 9, when the proportion is $\log_{10}(1+\frac{1}{9}) = .046$ (4.6%). If you doubt the veracity of this law, pick a random list of numbers and you will likely see that the number 1 appears disproportionately often as the lead number.

Let's take an example. Suppose you deposit $1.00 in a bank account that offers a 10%-per-year growth rate. In looking at your growing deposit through the years, you will find a preponderance of leading 1s at first, because when we start with a number that begins with 1, 10% more still has a leading 1, 10% more still does, and so forth. When you arrive at numbers starting with the digit 2, you start to make bigger jumps, so you have fewer numbers with leading 2s, and so forth until you arrive at the teens; at that point, you are back to leading 1s for quite some time as compared to the 20s, 30s, and so forth. If we begin with a number that starts with, say, 9, then 10% often pushes it beyond starting with 9 and definitely does so on the next occurrence.

People who might be called "forensic accountants" have used Benford's Law to detect fraud. When data are false, people tend to make up numbers that have many more leading 5s and 6s than would be expected under the distribution predicted by Benford's Law. Benford's Law is another example in which we can expect regularity in the aggregate that arises from randomness.

This lecture has presented some familiar economic indicators. Much of the way we measure our financial situation involves statistical presentations of the economic conditions of our lives. We must interpret data appropriately to know where we stand. Because dollars change in value over time, comparing a dollar from one time with a dollar from another does not capture the

meaning we seek. We must be careful to avoid the pitfalls of drawing inappropriate conclusions from data-mining methods. ■

Suggested Reading

B. Bowerman, R. O'Connell, and A. Koehler, *Forecasting, Time Series, and Regression: An Applied Approach*, 4th ed.

John A. Paulos, *A Mathematician Plays the Stock Market*.

Questions to Consider

1. There is considerable debate about whether having a large national debt is bad for the economic health of the country. What data would you gather and what statistical analysis would you undertake to inform your decision on this question?

2. One of the criteria used in stock management literature concerns modifying the risk of a portfolio by including a balance of stable and volatile stocks. What statistical indicators would you look for in data about a stock to assess where that stock fits on the stable-volatile spectrum?

Science—Mendel's Too-Good Peas

Lecture 23

In this lecture, we're going to be talking about science and applications of statistics to science. Certainly, statistics and the statistical analysis of data are obviously central players in all aspects of science.

Advances in empirical science depend on drawing deductions from data. In many cases, a scientific theory is tested by comparing experimental results to predictions of the theory. Randomness can enter in two ways. First, measurement error (noise) in experimental results adds randomness to otherwise definite predictions. Second, some theories, such as Mendel's theory of trait inheritance over generations of pea plants, are inherently probabilistic. In fact, reported results with too little fluctuation can be evidence of fraudulent data. On the other hand, a measurement much different from other measurements (an outlier) can indicate either that some gross error in measuring has occurred, in which case the measurement should be discounted, or that some fundamental assumption is incorrect. Study of the ozone layer in the atmosphere supplies a cautionary example.

Statistics is involved in essentially all scientific matters, from weather reports to quantum physics. During the last 400 years, we humans have fundamentally altered our conception of the universe and our position in it, in many cases as a result of some scientific advance that ultimately is based on the analysis of data. In this lecture, we'll look at several examples of scientific developments and the role of statistics in them.

The first example involves Johannes Kepler, the famous astronomer, working in about 1600 as assistant to the astronomer Tycho Brahe, who had amassed vast amounts of data about the locations of planets and stars. Kepler computed that the data fit a model of the solar system in which the planets revolve around the Sun following elliptical orbits. In devising his laws of planetary motion, Kepler used the statistical technique of creating a mathematical model to summarize the data. Later, Isaac Newton formulated

his universal law of gravitation, which implies that two masses will follow elliptical orbits about one another.

Science frequently progresses in this way. Observations are made that are well summarized by a mathematical equation or model, based on statistical curve-fitting techniques. Later, a more basic understanding of causes and effects can explain that physical model.

Hubble's observations about the red shift in spectra from receding stars form another example of statistics playing a prominent role in science. The pattern for a receding star is shifted toward longer wavelengths. The faster the star is receding, the greater the shift.

Another example in astronomy is the 3-degree radiation left over from the Big Bang. Researchers were trying to build a precise radio telescope and kept sensing background noise. After many attempts to fine-tune their instruments to avoid that "error," the researchers discovered that the background noise was a real phenomenon: the 3-degree Kelvin radiation left over from the Big Bang at the creation of the universe.

Randomness is at the heart of quantum physics. Modern theories of physics postulate the very unintuitive concept that a subatomic particle, such as an electron, is not in a precise location at a particular time. Instead, the location of an electron is a probability distribution. These theories put the statistical and probabilistic nature of existence in a fundamental position in our understanding of the world.

Measurements and interpretations of measurements are very basic to the scientific process. For example, measurements of the thickness of the ozone layer in the stratosphere or upper atmosphere illustrate another aspect of statistics. When data on this subject were collected by satellite in the 1970s, the values near the South Pole seemed surprisingly small. At first, these were deemed to be bad readings, reflecting some problem in the measuring process, and were omitted from the data summaries. With later measurements, however, it was discovered that the measurements were correctly reporting a real phenomenon, the ozone hole. When one has a lot of data and most of

the data are consistent, decisions must be made about what to do with the outliers. Science proceeds by developing models based on data, then testing the models by comparing experimental results to predictions of the model. In many cases, a scientific theory is tested by comparing experimental results to predictions of the theory.

Another example of statistics in science is Mendel's famous experiments with peas. Mendel noted statistical patterns in data concerning hereditary traits of pea plants. Yellow is the dominant gene. If homozygous yellow pea plants (those with two yellow genes) and homozygous green pea plants (those with two green genes) are crossed, the first-generation offspring all look yellow, being yellow heterozygous plants (those with one yellow gene and one green gene). The second-generation offspring, however, which are the result of crossing yellow heterozygous plants, are about one-quarter green, indicating homozygous green pea plants.

In many cases, a scientific theory is tested by comparing experimental results to predictions of the theory.

This was the fundamental observation that led to the concepts of genetic inheritance and dominant and recessive genes. Two genes, one from each parent plant, combine to form the genetic makeup governing the color of the offspring. Yellow is the dominant gene; thus, only if both genes in a plant are for green will the plant be green. Assuming that one of the genes is randomly selected from each parent plant, we would expect that *about*, but not exactly, one-quarter of the time, both contributions will be green.

To determine whether a yellow plant was heterozygous or homozygous, Mendel took the yellow plants and bred them with themselves 10 times. If the plant was homozygous, on each of those 10 times, he would always get a yellow plant. However, if he had a heterozygous plant, he reasoned that the chances were very good that in 10 breedings, 1 of the self-breedings would contribute both green genes, and the plant would come out green. If we performed many experiments, with 800 yellow plants in the second

generation, we would expect different numbers of homozygous yellow plants in different experiments, with the center of the distribution around 200 but with occasional outliers.

Statisticians have looked at Mendel's reported results and have discovered that it would be very unusual, given the amount of his data, that all of the results would be in the narrow bounds he reported. In short, Fisher believed that the data that Mendel got were too good. Ronald Fisher, to whom we were introduced in Lecture 12, found that Mendel's results lie within 1 standard deviation of the mean much more often than the expected 68% of the time. Fisher also noted that Mendel used the method of cross-breeding yellow plants with themselves 10 times to identify whether they were homozygous or heterozygous. In an interesting twist, this method implies that we should have expected Mendel to misclassify a certain percentage of plants; however, Mendel's reported data are closer to the data expected if all the plants were classified correctly. Mendel's work illustrates all aspects of statistics, including design of experiments and interpretation of data, and may illustrate the possibility that the data were made to look somewhat better than they actually were.

Using carefully executed statistical capture-recapture methods, scientists can estimate quantities as diverse as the population of tigers in a jungle, the volume of water in a lake, and the size of a natural gas deposit in the ground. Statistical analysis of experimental data is key to validating or invalidating a scientific theory. ■

Suggested Reading

R. A. Fisher, "Has Mendel's work been rediscovered?" *Annals of Science* (available at http://www.library.adelaide.edu.au/digitised/fisher/144.pdf).

E. T. Jaynes and G. Larry Bretthorst, eds. *Probability Theory: The Logic of Science*.

Questions to Consider

1. Measurements are never exact. As instruments improve, would you expect the distributions of measurements of physical constants to have less variation, have a different mean, or both?

2. Suppose data are found that reject a scientific theory with a high level of statistical significance. Under what circumstances would you tend to reject the data rather than reject the theory? Is that ever a good idea?

Statistics Everywhere

Lecture 24

Often, statistical reasoning can contribute to decisive arguments in matters that seem very difficult to resolve, and even issues that don't appear to have any statistical component to them at all.

Statistics is a subject that permeates essentially every area of our lives and world. It is a powerful tool for seeing our world in a more detailed fashion and for making informed decisions, although its subtleties and potential misuses caution us to avoid thoughtless acceptance of statistical conclusions. The recent and expected future development of computer speed and capacity allow us to imagine using statistics with ever more scope and effect. How much information and understanding can we hope to gather from statistical data? How much more meaning would better statistical techniques allow us to find? Statistics is a tool with wide applicability. It has limits that need to be acknowledged and respected, but its potential for helping us find meaning in our data-driven world is enormous and growing.

Often, data can contribute decisive evidence in an otherwise difficult matter to resolve. For example, during the debate about ratification of the Constitution, Alexander Hamilton, James Madison, and some others anonymously wrote *The Federalist Papers*. People disputed the authorship of about a dozen of these essays. Arguments based on philosophy and style were not persuasive. Discriminant analysis, that is, statistical analysis concerning the frequency of the use of specific common words (for example, *on* instead of *upon*, where appropriate) provided powerful arguments for Madison's authorship.

This *Federalist Papers* example is satisfying in that it suggests that seeking data to find persuasive arguments is a valuable method for coming to conclusions. Sometimes it is not clear what data are pertinent. In the case of *The Federalist Papers* dispute, it would not be immediately obvious that counting the frequency of trivial words would be the road to decision on the authorship issue. Of course, the technique can be applied to other authorship questions, such as whether the works of Shakespeare were actually written by Marlowe or Bacon and whether Shakespeare wrote a newly discovered

poem attributed to him. In the latter instance, the new poem was discovered in 1985. Instead of looking for frequency of common words, as with *The Federalist Papers* controversy, statisticians looked for new, original words because Shakespeare was well known for inventing words. Because 9 new words appeared in this 429-word poem, these statisticians were able to deduce that, in fact, this poem was very possibly written by Shakespeare. The results in those cases, however, do not seem as clear-cut as in *The Federalist Papers* dispute. Data and appropriate interpretation are powerful arguments not easily refuted without further data.

Using data and statistical analyses will become an even more prominent part of our world in the future than it is now. The principal reason is the continuing development of computer technology.

The evidence for the determination that Madison wrote the anonymous articles would be evaluated differently by Bayesian statisticians versus frequentist statisticians. A Bayesian would be willing to say that, given the word usage in *The Federalist Papers*, there is a 99.9% chance that it is written by Madison. Frequentists would say that the articles are either written by Madison or not. Both camps would probably agree to a statement something like this: "The probability is only 2.4% of getting no *upon*s when randomly selecting a collection of 1000 words from a person's writing that generally has 6 *upon*s per 1000."

Using data and statistical analyses will become an even more prominent part of our world in the future than it is now. The principal reason is the continuing development of computer technology. With the computer, it is now possible to deal with large databases and use techniques that would have been computationally impossible previously. Some such techniques involve simulation as a means to understand a collection of data. Often, such methods are computationally intensive and, consequently, become increasingly valuable as computer power increases.

One such technique is called the *Monte Carlo method*, which involves using random processes by a computer to generate thousands of scenarios, enabling statistical techniques to derive the distribution of the behavior of a system. Today on a home computer, it is possible to do amazing statistical analyses essentially instantly. Early textbooks on statistics would emphasize methods for reducing computation. Now, with computers, those techniques are not so important.

Let us now consider some observations about the statistical enterprise altogether. First, there is often a difference between statistical knowledge and understanding based on deeper principles. Second, statistics is used more than it is understood, as evidenced in the blind application of statistical tests. Any attempt to reduce data to a formulaic adherence to following tests is likely to be misleading and can often produce nonsensical arguments.

Recall the mean levels of wealth of graduates at Lakeside High School in Seattle. Statistical reasoning is subtle and prone to counterintuitive examples; understanding the underlying logic is necessary in order to have confidence in the result. Recall the exercise of choosing the best hospital.

Hypothesis testing has issues of its own, such as the arbitrariness of the level of rarity that we deem statistically significant. The persuasive strength of a statistical argument requires a clear understanding of the statistical reasoning, the context of the situation, and the details about the study or data that allow us to interpret the meaning of the statistical data and arguments with conviction.

Unless we have high confidence that a survey was conducted appropriately, then the statistical result may not be as strong as reported. Recall the results of the *Literary Digest* poll. Good statistical results and inferences are far superior to anecdotal evidence on an issue, but we need to be critical consumers. Statistical knowledge is, by its very nature, an admission of ignorance. Dealing with statistics often means that we do not have the whole story. Statistics is a collection of profoundly powerful methods for understanding our world with more detail and more meaning.

We have seen statistics as having two basic parts:

- Organizing, describing, and summarizing a collection of data when we know all the data.

- Inferring information about the whole population when we have data about only a sample of the population.

When we make an estimate of the value of a feature of the whole population given data about a sample, our challenge is to describe how accurate our estimate is likely to be by answering the following questions:

- How close is our estimate to the correct value?

- How confident are we that our estimate is, in fact, that close?

Statistics is a powerful tool for understanding our world. We end the course with a statistic about the course: Among people who learn something about statistics, 100% appreciate our world with more clarity. ■

Suggested Reading

Norman L. Johnson, and Samuel Kotz, eds. *Leading Personalities in Statistical Sciences: From the Seventeenth Century to the Present*.

William S. Peters, *Counting for Something: Statistical Principles and Personalities*.

Theodore M. Porter, *The Rise of Statistical Thinking, 1820–1900*.

David Salsburg, *The Lady Tasting Tea: How Statistics Revolutionized Science in the Twentieth Century*.

Questions to Consider

1. As a practical matter, how can a more sophisticated understanding of statistical reasoning come into play on an everyday basis? Can we apply pressure on the media to make the details of studies more readily available so that we have a better chance of determining whether the statistical reasoning is sound?

2. Being attuned to statistical reasoning adds depth to our appreciation of the world around us. Choose an issue in which you have an interest and find data that are pertinent to your appreciation of it. Then apply the techniques of organizing, describing, and summarizing the data and inferring meaning from the data to understand that part of the world better.

Timeline

1532.....First weekly data collected on deaths in London.

1539.....Beginning of official data collection on baptism, marriages, and deaths in France.

1608.....Beginning of the Parish Registry in Sweden.

1662.....John Graunt publishes *Natural and Political Observations Mentioned in a Following Index and Made upon the Bills of Mortality*, which initiated the idea that vital statistics can be used to construct life and mortality tables for the relevant population.

1666.....First modern national demographic census (conducted in Canada).

1689 .....Jacob Bernoulli publishes the law of large numbers, a mathematical statement of the fact that when an experiment is repeated a large number of times, the relative frequency with which an event occurs will equal the probability of the event.

1693.....Edmund Halley publishes *Estimate of the Degrees of Mortality on Mankind*, which contained the mortality tables for the city of Breslau, Poland. It was one of the earliest works to relate mortality and age in a population and was highly influential in the future production of actuarial tables in life insurance.

1713.....Nicholas Bernoulli edits and publishes *Ars Conjectandi* (*The Art of Conjecture*), written by his uncle, Jacob Bernoulli, in which the work of others in the field of probability was reviewed and thoughts on what probability really is were presented.

1728.....Sir Isaac Newton publishes *The Chronology of Ancient Kingdoms Amended*, in which he gives a 65% confidence interval for the length of a king's reign.

1733.....Abraham de Moivre publishes an account of the normal approximation for the binomial distribution for a large number of trials. This work improves on Jacob Bernoulli's law of large numbers. The account will be included in the 1756 edition of de Moivre's *The Doctrine of Chances*, a treatise on probability first published in 1718.

1735.....The beginning of demographic data collection in Norway.

1746.....Publication in France of tables based on mortality data.

1749.....Sweden's first complete demographic census.

1753.....Austria's first complete demographic census.

1766.....First publication of Sweden's mortality tables.

1769.....Denmark and Norway's first complete demographic census.

1790.....Giammaria Veneziano Ortes publishes *Reflessioni sulla popolazione delle nazioni per rapporto all'economia nazionale* (*Reflections on the Population of Nations in Respect to National Economy*); America's first federal demographic census.

1801.....Britain's first complete demographic census; France's first complete demographic census.

1810.....Pierre-Simon Marquis de Laplace publishes a fairly general statement of the central limit theorem.

1825.....Benjamin Gompertz publishes *On the Nature of the Function Expressive of the Law of Human Mortality*, in which he uses logarithmic regression to show that the mortality rate increases exponentially as people age.

1827.....Pierre-Simon Marquis de Laplace publishes a paper on multiple regression analysis with applications to lunar tides and the atmosphere.

1829.....Belgium's first complete demographic census.

1834.....Establishment of the Statistical Society of London (later the Royal Statistical Society).

1835.....Adolphe Quetelet publishes *Sur l'homme et le développement de ses facultés, essai d'une physique sociale*, in which he presents his conception of the average man as the central value about which measurements of a human trait are grouped according to the normal distribution.

1837.....Simeon Denis Poisson publishes *Recherches sur la probabilité des jugements en matière criminelle et matière civile*, which introduces the expression *law of large numbers* and in which the Poisson distribution first appears; Public data collection of the demographic statistics in England. Establishment of the Registrar General Office.

1839.....Organization of the American Statistical Association.

1846.....Verhulst publishes a nonlinear differential equation describing the growth of a biological population, which he deduced from data. The equation predicts that population growth is limited by forces that increase with the square of the rate at which the population grows, rather than being unlimited exponential growth.

1853..... Augustin-Louis Cauchy presents an outline of the first rigorous proof of the central limit theorem; Adolphe Quetelet organizes the first international statistics conference.

1861.....Italy's first complete demographic census.

1867.....Pafnutii Lvovich Chebyshev publishes a paper, *On Mean Values*, which uses Bienaymé's inequality to give a generalized law of large numbers.

1869.....Establishment of the *Société de Statistique de Paris* (the Statistical Society of Paris).

1885.....Establishment of the International Statistical Institute in the Netherlands.

1887.....Pafnutii Lvovich Chebyshev publishes *On Two Theorems*, which gives the basis for applying the theory of probability to statistical data, generalizing the central limit theorem of de Moivre and Laplace.

1889.....Francis Galton publishes *Natural Inheritance*, in which he presents a summary of the work he had done on correlation and regression that included the idea of regression to the mean, discovered through his 1875 experiments with sweet peas.

1892.....I.V. Sleshinsky publishes the first complete rigorous proof of the central limit theorem, based on the outline by Cauchy.

1893.....Karl Pearson begins the publication of 18 papers entitled *Mathematical Contributions to the Theory of Evolution*, which contain his most valuable work in the form of contributions to regression analysis, the correlation coefficient, and the chi-square test of statistical significance. This work lasts through 1912.

1897.....George Udny Yule publishes *On the Theory of Correlation*, in which he begins the development of his approach to correlation via regression with a conceptually new use of least squares that would later dominate applications in the social sciences.

1901.....Publication of the first issue of *Biometrika*, a journal founded by Karl Pearson and Francis Galton.

1908.....William Sealy Gosset, under the pseudonym "Student," publishes the t-distribution as the sampling distribution of the mean when the population variance is unknown.

1921.....Ronald A. Fisher introduces the concept of maximum likelihood. In 1922, he would redefine statistics such that its purpose was the reduction of data. Fisher identified three fundamental problems: (1) specification of the kind of population from which the data came, (2) estimation, and (3) distribution.

1930.....Establishment of the Institute of Mathematical Statistics and the appearance of *Annals of Mathematical Statistics* in the United States.

1931.....Establishment of the Indian Statistical Institute.

1933.....Jerzy Neyman and Egon Pearson publish *On the Problem of the Most Efficient Tests of Statistical Hypotheses* and *The Testing of Statistical Hypotheses in Relation to Probabilities a Priori*, capping five productive years of research on statistical hypothesis testing.

1935.....Ronald A. Fisher publishes the first edition of *The Design of Experiments*, which revolutionizes the use of statistics in agriculture.

1937.....George W. Snedecor and William G. Cochran publish *Statistical Methods.*

1946.....Harald Cramer publishes *Mathematical Methods of Statistics*, which joins the science of statistical inference with the theory of classical probability and was reprinted as recently as 1999.

1947.....Ronald A. Fisher publishes *Statistical Tables.*

1966.....Foundation of the Working Party on Statistical Computing, which published guidelines for program development, descriptions, and code for statistical programs.

1967.....W. J. Hammerle publishes *Statistical Computations on a Digital Computer*, the first textbook devoted to statistical computing.

1972.....The American Statistical Association forms a section for statistical computing. In the following years, the use of computers in statistics will allow statisticians to generate, collect, organize, and analyze larger data sets and increase the complexity of the models fitted to the data. Displays become more impressive, using color and perspective. In the realm of hypothesis testing, permutation testing undergoes a resurgence. Computing power allows statisticians to develop theory using Monte Carlo simulation studies.

Glossary

analysis of variance (ANOVA): A procedure of statistical analysis by which differences in means of two or more groups can be assessed after eliminating variance that is due to other factors.

Bayesian statistics: The view in which probability is interpreted as a measure of degree of belief. In this view, the concept of probability distribution is applied to a feature of a population, such as the population mean, to indicate one's belief about possible values of that feature. The principal result of experiments is to update such a probability distribution, indicating a change in belief. The Bayesian viewpoint is in contrast to the frequentist view.

bias: The extent to which the statistical method used in a study does not estimate the quantity to be estimated or may not test the hypothesis to be tested.

binomial distribution: The probability distribution of the number of successes in n Bernoulli trials. For a series of events to be considered Bernoulli trials, they must satisfy three conditions: (1) the trials are independent of each other, (2) each trial has exactly two possible outcomes, and (3) the probability associated with each outcome is constant throughout all of the trials.

box plot: A graphical display for numerical data that shows the maximum and minimum values, the median, and the quartiles of the data.

central limit theorem: Statistical theorem that states the following: Starting with almost any distribution (such as a Poisson, binomial, or uniform distribution) with a finite standard deviation σ, if we take many samples of size n, the distribution of the average values of the samples will be approximately a Gaussian distribution (assuming n is large) with the same mean as the original distribution and with standard deviation $\frac{\sigma}{\sqrt{n}}$.

chi-square distributions: A family of distributions that take only positive values and are skewed to the right. Each chi-square distribution is specified by its degrees of freedom. The higher the degrees of freedom, the more skewed the distribution is. The chi-square family of distributions occurs often in hypothesis testing about categorical variables.

chi-square test for independence: A process used to test the hypothesis that two categorical variables have no relationship. The test statistic that is calculated has a chi-square distribution.

confidence interval: A range of values, constructed from information obtained from a sample of the population, that is believed, with a specified probability, to contain the value of the population parameter.

correlation coefficient: The quantification of the strength of linear association that exists between two numeric variables. The correlation coefficient takes values between –1 and 1, where negative correlations mean that as the value of one variable rises, the other falls, and positive correlations mean that the values of the two variables rise together. Values of the correlation coefficient near 1 or –1 indicate a strong linear relationship between the two variables. Values near 0 indicate no linear relationship between the two variables.

dispersion: The variation among values when the data values in a sample are not all the same.

estimator: A statistic, calculated based on the information from a sample, that is used to estimate the value of a parameter associated with the population from which the sample was selected.

event: An outcome or set of outcomes from a random process.

expected value: The average outcome that might be expected from a long run of trials of a probabilistic event.

experimental design: Procedures and planning used in an experimental study. In general, these procedures are designed to reduce bias, promote replication, use randomization in order to initiate study of causality, and ensure appropriate sample size.

extrapolation: The process of using the data to make estimates about values that lie beyond the range of the existing data.

factor analysis: A set of statistical procedures used to analyze multivariable data when many variables are known about the subjects. The underlying principle behind factor analysis is that variables that are highly correlated with each other are grouped together and separated from variables that are not highly correlated with the group. Each group represents a factor, thought to be a single underlying construct.

five-number summary: A numerical summary of data that includes the minimum and maximum values, the median, and the upper and lower quartiles. The five numbers divide the data into four groups, each containing the same number of data points. Often used to describe data that have skew.

frequentist statistics: The view in which probability is defined in terms of long-run frequency or proportion in outcomes of repeated experiments. The concept of probability is applied to outcomes of actual or hypothetical experiments because there is a random element to those. But in the frequentist view, probability is not used as a measure of knowledge or belief of the possible values of a quantity, such as the true population mean, that does not have a random element. The frequentist viewpoint is in contrast to the Bayesian view.

Gaussian distribution: See **normal (Gaussian) distribution**.

histogram: A graphical display for numerical data in which vertical bars show the number of observations that have a value between the values given on the x-axis at the base of the bar.

hypothesis test: The process of assessing whether observed data are consistent with some claim about the population in order to determine whether the claim might be false.

independent events: Two events are independent if knowing the outcome of one tells us nothing about the other. There is no relationship between the two events.

interquartile range (IQR): A measure of spread. The difference between the upper quartile and lower quartile. Also used in rules of thumb for identifying outliers.

lurking variable: A variable that has an important effect on the relationship among variables considered in a study but that is not, itself, considered in the study.

mean: A measure of the location of the center of numerical data. Also called the *arithmetic average*. It is computed by summing the values of the data and dividing by the number of data points. Conceptually, it is the balance point of the data when they are represented by a line plot. Because the mean is not particularly resistant to outliers, it is used mainly when the data have a roughly symmetric distribution.

median: A measure of the location of the center of numerical data. Once the data are ordered by their value, the median is the value taken by the data point that is in the middle, such that there are the same number of data points larger than the median as smaller than the median. If there is an even number of data points, then the median is the average of the values of the two in the middle. The median is also the 50^{th} percentile and the second quartile. Because the median is particularly robust to outliers, it is used when the data are skewed or contain outliers.

Monte Carlo method: A numerical modeling procedure that makes use of random numbers to simulate processes that involve an element of chance. In Monte Carlo simulation, a particular experiment is repeated many times with different randomly determined data to allow statistical conclusions to be drawn.

nonparametric test: Ill-defined term used generally to describe processes for inference that may be used either when the assumptions underlying parametric procedures, such as chi-square and one- and two-sample tests, are not met or when responses are difficult to quantify or contain rankings rather than meaningful numerical values.

normal (Gaussian) distribution: A family of single-peaked, symmetric probability distributions described as bell shaped. It is the distribution associated with errors in measurement, with heights and weights, and with standardized test scores, for example.

null hypothesis: A proposition or set of propositions to be tested.

observation: The value associated with one member of a sample.

one-sample test for means: A process for testing the hypothesis that the mean value of some quantitative aspect of a population has a particular value. The test statistic exhibits a roughly t-distribution when the standard deviation of the value in the population is not known.

one-sample test for proportions: A process for testing a hypothesis about the percent of members of a population who have a particular characteristic or opinion. The test statistic has a roughly normal distribution.

outlier: A data point with value that differs markedly from the rest of the values in the data set.

parameter: A numerical value about data that is calculated from the values of a population.

percentiles: The percentiles are the observations that divide the data into 100 groups, each with the same number of observations. For example, scoring in the 85^{th} percentile on the SAT means that one has outscored 85% of those tested.

Poisson distribution: A right-skewed probability distribution that describes the number of occurrences of an event in a given time period.

population: A population is any entire collection of people, animals, plants, or things from which we may collect data. It is the entire group in which we are interested and that we wish to describe or draw conclusions about.

power: The power of a hypothesis test is the ability of the test to accurately reject the null hypothesis when the null hypothesis is, indeed, false. One wants tests to have high power. However, as the power of a test increases, so does the probability of a type I error, that is, the rejection of the null hypothesis when it is actually true. The statistician must find a reasonable balance between power and the probability of a type I error.

probability distribution: A probability distribution is a table, function, or graph that assigns a probability to each possible outcome.

p-value: In a hypothesis test, the probability of obtaining the results that were obtained from a sample or results more unusual if the null hypothesis represents the truth about the population.

quartiles: The quartiles are the values of the data that divide the observations into four equal-sized groups. To find the quartiles, list the values of the data in order from smallest to largest. The second quartile (median) is the observation in the middle. The first quartile is the observation that divides the lower half of the data, between the minimum and the median, into two equal-sized groups. The third quartile is the observation that divides the top half of the data, between the maximum and the median, into two equal-sized groups.

regression analysis: A statistical process by which a model is created that predicts the value of a response variable through an equation using the values of one or more explanatory variables.

residual: The difference between the actual (observed) value of a response variable and that calculated from a regression equation.

sample: A subset of a population that is used to infer information about the population.

sample mean: The value of the mean of a sample.

sampling bias: Error that is introduced in a statistical study by the method of sampling. For example, the use of voluntary sampling, such as online polls, introduces bias because the respondents tend to be those who are passionate about the topic, rather than a random sample of people with all types of opinions.

sampling distribution: The theoretical distribution of the statistic calculated from a sample. The generation of this distribution is based on the calculation of the statistic from every possible sample from the population.

scatter plot: A two- or three-dimensional graph in which each axis represents one variable that is associated with an observation. Used in regression analysis as a visual display of patterns that may exist among variables in the data.

significance level: In a hypothesis test, a prespecified value at which the null hypothesis may be rejected. Sometimes used to describe the p-value of a hypothesis test, that is, the probability of obtaining the value that was obtained from a sample if the null hypothesis about the population from which the sample was selected is true.

simple random sample (SRS): A sample of a population that is chosen in such a way that each member of a population has an equal chance of being selected.

skewness: The lack of symmetry exhibited by a distribution. The direction of skew, left or right, tells the direction of the tail that causes the lack of symmetry.

standard deviation: The most commonly used measure of dispersion (spread) for numerical data. It is the square root of the variance. Like the variance and the mean, its calculation is not resistant to outliers and extreme skew.

standard error: The standard deviation of the sampling distribution of a statistic.

statistic: A numerical value about data that is calculated from the values of a sample.

stochastic: A synonym for random; the adjective applied to any phenomenon obeying the laws of probability.

stratified sample: A method of sampling by which the population is first divided into groups, or strata, based on common characteristics, such as gender or income. If a random sample is then selected from each group, the term *stratified random sample* may be used.

t-distribution: The t-distribution is the theoretical distribution of a sample mean calculated from a sample taken from a population whose standard deviation is not known. Its shape is roughly symmetric and similar to that of a normal distributed variable, but the tails are thicker.

two-sample test for means: A process for testing the hypothesis that two different populations have the same mean. The calculated test statistic is theorized to have a t-distribution.

two-sample test for proportions: The process for testing the hypothesis that two different populations share the same value for a binomial process (such as a yes-no question). The calculated test statistic has a roughly normal distribution.

type I error: Rejection of the null hypothesis when it is true.

type II error: Acceptance of the null hypothesis when it is false.

uniform distribution: A distribution in which every possible value is equally likely. The histogram of a uniform distribution has all of the bars the same height.

variance: A measure of dispersion (spread) for numerical data. It is roughly the average squared distance of the data values from the mean. It is calculated by summing the square of the differences between the data and the mean. To calculate a population variance, one divides by the number of elements in the population. To calculate the sample variance, one divides by one fewer than the number of observations in the sample.

Biographical Notes

Bayes, Thomas (1701–1761): British nonconformist minister. Little is known about Bayes's life save that he was the son of a nonconformist minister, educated at Edinburgh University, and a member of the Royal Society. His major contribution to the field of statistics was the work he did on the inverse probability problem. At the time, the calculation of the probability of a number of successes out of a given number of trials of a binomial event was well known. Bayes worked on the problem of estimating the probability of the individual outcome from a sample of outcomes and discovered the theorem for such a calculation that now bears his name.

Bernoulli, Jacques (often called Jacob or James) (1654–1705): Professor of mathematics at Basel and a student of Leibniz. He formulated the law of large numbers in probability theory and wrote an influential treatise on the subject.

Cauchy, Augustin-Louis (1789–1857): French mathematician and engineer. Professor in the Ecole Polytechnique and professor of mathematical physics at Turin. Cauchy worked in number theory, algebra, astronomy, mechanics, optics, and analysis. His contribution to statistics was the production of the outline of the first rigorous proof of the central limit theorem in 1853, in the course of a controversial debate during meetings of the Academy of Sciences and in the pages of its journal with Irenée-Jules Bienaymé (1796–1878). The debate started as a result of a critique made by Cauchy of the work of Laplace. Bienaymé, a student of Laplace, took exception to the criticism of his mentor, on whose work much of Bienaymé's was based, and a debate ensued. Although Cauchy only sketched his proof, I. V. Sleshinsky was able to fill in the details and missing steps. He produced a complete, rigorous proof of the central limit theorem based on the outline by Cauchy in 1892.

Chebyshev, Pafnutii Lvovich (1821–1894): Russian mathematician, founder of the St. Petersburg School of Mathematics. The culmination of his career of study in probability and statistics occurred in 1887, with his use of the method of moments to prove the first version of the central limit theorem for sums of independent but not identically distributed variables.

Cox, Gertrude Mary (1900–1978): American statistician and administrator. Cox's main contributions to the field of statistics were in the areas of experimental design and analysis of psychological data. In addition, in 1949 she became the first woman elected to the International Statistical Institute. Cox was the first head of the department of experimental statistics at North Carolina State University. She was a founding member of the Biometrics Society and editor of the journal *Biometrics* for 10 years.

Cramer, Harald (1893–1985): Swedish mathematician and statistician. Chair of the actuarial mathematics and mathematical statistics department and, later, president of Stockholm University, Cramer served as chancellor of the Swedish university system. He wrote several seminal books that expressed probability theory in a manner more useful in its application to statistical theory than had previously been articulated. Working as an actuary with the Svenka Life Insurance Company early in his career led Cramer to investigate stochastic processes as they related to insurance. His text, *Collective Risk Theory*, is concerned with the progress over time of monetary funds, with inputs, such as premiums and interest, and outputs, such as claims, as special cases of general stochastic processes.

Deming, W. Edwards (1900–1993): American statistician and quality-control expert. Trained as a physicist, Deming became interested in statistics while working at the U.S. Department of Agriculture. He then took a post at the U.S. Bureau of Census as the Head Mathematician and Advisor in Sampling. He is credited with importing the replicate subsampling method from India, which forms part of the national sampling plan used by the U.S. Bureau of Census and by polling corporations, such as Gallup. After World War II, Deming was assigned to General MacArthur's Supreme Command of the Allied Powers in Tokyo. While there, Deming undertook a systematic education of quality-control principles and techniques in the Japanese workforce. The Japanese attention to quality control as introduced to them by Deming is credited as the primary force behind that country's emergence as an industrial leader among nations.

Fisher, Ronald Aylmer (1890–1962): British statistician. Trained in mathematics and physics, Fisher is known as the father of modern statistical methods. Through correspondence with W. S. Gosset, Fisher was the first to

derive the general sampling distribution of the correlation coefficient. His major contributions to statistics were in the area of design of experiments. He introduced the concept of randomization and the process of analysis of variance (ANOVA) now widely used by statisticians. He was a fellow of the Royal Statistical Society and was elected to the American Academy of Arts and Sciences, the American Philosophical Society, the International Society of Haematology, the National Academy of Sciences of the United States, and the Deutsche Akademie der Naturforscher Leopoldina. He was awarded honorary degrees from many institutions, including Harvard University (1936), University of Calcutta (1938), University of London (1946), University of Glasgow (1947), University of Adelaide (1959), University of Leeds (1961), and the Indian Statistical Institute (1962). Fisher was knighted in 1952.

Galton, Francis (1822–1911): British explorer and anthropologist. Cousin to Charles Darwin. He was the first to calculate a quantitative value for correlation and a pioneer of the use of the variable *r* for correlation coefficient, although his calculation differs from that used by modern statisticians. He was the first person to document the phenomenon known as *regression to the mean*, which he discovered through experiments with sweet peas. His ideas strongly influenced the development of statistics, particularly his proof that a normal mixture of normal distributions is itself normal. Galton may be described as the founder of the study of eugenics. His principal contributions to science consisted of his anthropological inquiries, especially into the laws of heredity. In 1869, in *Hereditary Genius*, he endeavored to prove that genius is mainly a matter of ancestry via the application of statistical methods.

Gauss, Karl Friedrich (1777–1855): German mathematician and astronomer, nicknamed the "Prince of Mathematicians." His mathematical work included the concept of a distribution of errors that originally was known as the *error distribution* and later became known as the *Gaussian distribution*, or the *normal distribution*.

Gosset, William Sealy (1876–1937): British chemist who, while working at the Guinness Brewery in Dublin, Ireland, began a study of statistical methods as applied to small samples. Asked by the brewery to investigate the

relationship between the quality of materials, barley and hops, for example, and production conditions on the product, beer, the corporation required him to publish his results under a pseudonym to preserve the anonymity of the brewery. Gosset chose the name "Student" under which to publish his results about the derivation and use of a t-distribution in inference, leading to its being referred to as the *Student's t distribution*. In later work at the brewery, Gosset would come to support the use of a balanced design in agricultural applications, rather than either of the two available competing designs. Unfortunately, Gosset would pass away before this disagreement could be resolved.

Laplace, Pierre-Simon Marquis de (1749–1827): French mathematician and astronomer. Professor at Ecole Normale and Ecole Polytechnique. Primarily known for his contributions to calculus, analysis, and physics, toward the end of his life, he turned to research in statistics and obtained a fairly general statement of the central limit theorem in 1810. Between 1818 and his death, Laplace investigated multiple regression analysis as related to lunar tides and the atmosphere and published a comparison of absolute versus least-squares deviations and their use in regression analysis and the notion of a sufficient statistic.

Markov, Andre Andreevich (1856–1922): Russian mathematician. Member of the St. Petersburg Academy of Science. Markov was a student of Chebyshev and spent most of his career studying probability distributions, random variables, the weak law of large numbers, and the central limit theorem. Markov's significant contribution to probability theory was the introduction of the concept of a Markov chain as a model for studying the behavior of random variables. One example of a Markov chain is known as a simple random walk. In a random walk, each direction in which a "man" may step is assigned a probability. The paths that may occur and their assigned probabilities make up the "behavior" of this particular random variable. In modern statistical practices, Markov chains are used in conjunction with Monte Carlo methods to solve problems that are analytically complicated by generating suitable random numbers and observing the fraction of those numbers that obey some property or properties.

Mendel, Johann Gregor (1822–1884): Czech monk. Mendel was the first to apply statistical methods to biology in his calculation of ratios of genotypic structures. The application of his work to the field of genetics was not recognized until the 1930s, 50 years after his death. The validity of Mendel's most famous work, on the hybridization of peas, has come under question. Fisher initially concluded that Mendel's description of experimental design was correct but that the data lacked an appropriate amount of random variation and were likely fabricated or sanitized. Later authors have exonerated Mendel based on his use of sequential procedures and the inclusion of meteorological data.

Moivre, Abraham de (1667–1754): French-English mathematician. Born in France and educated at the Sorbonne in mathematics and physics, de Moivre, a Protestant, emigrated to London in 1688 to avoid further religious persecution. A future fellow of the Royal Society of London, de Moivre supported himself in England as a traveling mathematics teacher and by selling advice in coffee houses to gamblers, underwriters, and annuity brokers. De Moivre is recognized in statistics as the first to publish an account of the normal approximation to the binomial distribution. In fact, some of de Moivre's methods are so ingenious as to be shorter than modern demonstrations of solutions to the same problems.

Moore, David: Prolific and lucid author of statistics textbooks. His work was influential in redefining the common presentation of statistics in colleges, de-emphasizing the mathematical theory, and focusing on real data.

Newton, Sir Isaac (1642–1727): English mathematician and scientist known for the discovery of the law of gravity and as one of the fathers of calculus. Within the field of probability, Newton is known for his proof of the binomial theorem. There is also evidence that he gave thought to the variability of the sample mean, the basis for the central limit theorem. In his last work, *The Chronology of Ancient Kingdoms Amended*, published posthumously in 1728, Newton estimated the mean length of a king's reign to be between 18 and 20 years. In fact, the mean reign was 19.1 years, and the standard deviation of his sample was 1.01 years; thus, Newton's range of 18 to 20 years roughly corresponds to a 65% confidence interval (Johnson and Kotz).

Neyman, Jerzy (1894–1981): Polish statistician. Reader at University College in London and founder of the department of statistical sciences at the University of California at Berkeley. Elected to the International Statistical Institute and the U.S. National Academy of Sciences. Neyman's work in statistical inference leads some to call him the father of modern statistical methods. In a paper co-authored with Karl Pearson, Neyman explained the logical foundation and mathematical basis for the theory of hypothesis testing. Through his work, the theory of confidence intervals was developed from the theory of hypothesis testing. Neyman also contributed to innovative and precise use of statistics in fields ranging from agriculture and astronomy, biology, and social insurance to weather modification.

Nightingale, Florence (1820–1910): British nurse. Commonly known as the "Lady of the Lamp" for her work as a nurse for British troops during the Crimean War, Nightingale saved more soldiers through the use of statistics than medicine. Utilizing both statistical methods and innovative graphical techniques, she was able to convince the British army of the importance of hygiene and sanitation in hospitals, which led to widespread army hospital reform.

Pareto, Vilfredo Federigo Samaso (1848–1923): French-Italian engineer and economist. His contributions to statistics include work on interpolation and fitting curves to data and actuarial calculations in insurance and pensions. However, Pareto's most significant contribution to statistics was in his discovery of the first stable probability distribution other than the Gaussian (normal) distribution, named the *Pareto distribution* in his honor. Not only does the Pareto fit naturally arising situations, such as income distribution, but it also has theoretical applications. When conditions for the central limit theorem do not apply because the population distribution has heavy tails and, therefore, does not have finite variance, a modification of the central limit theorem may be applied if the behavior of the tail of the population distribution has roughly a Pareto distribution.

Pearson, Karl (1857–1936): British mathematician. Chair of the applied mathematics department at London's University College. Influenced by Galton (see above) and Walter Frank Raphael Weldon, a Darwinian zoologist who worked to make biology a more rigorous and quantitative science,

Pearson became interested in developing mathematical methods for studying heredity and evolution. Together, the three founded the journal *Biometrika*. Pearson worked out the mathematical properties of both the product-moment correlation coefficient and simple regression used to measure the relationship between two continuous variables. Later in his career, he explored relationships between two categorical variables and mixtures of categorical and continuous variables and developed the chi-square test.

Poisson, Simeon Denis (1781–1840): French mathematician. He published *Recherches sur la probabilité des jugements en matière criminelle et matière civile* in 1837, marking the first appearance of the Poisson distribution, originally found by de Moivre, which describes the probability that a random event will occur in a time or space interval under the conditions that the probability of the event's occurring is very small. Poisson also introduced the expression *law of large numbers*, by which he meant that for a larger number of trials, the proportion of successful outcomes exhibits statistical regularity even if the probability of success does not remain constant. Although we now rate this work as of great importance, it found little favor at the time, the exception being in Russia, where Chebyshev developed Poisson's ideas.

Quetelet, Lambert Adolphe Jacques (1796–1874): Belgian mathematician, astronomer, and meteorologist. In the first recorded attempt to summarize characteristics of the population, Quetelet coined the term *social physics* and used data collected from the national population to describe the average man, *l'homme moyen.* This led Quetelet to the notion that nature was attempting to create the average man as a prototype and that deviations from the average were errors. Working for the government, Quetelet collected and analyzed statistics on crime and mortality and devised improvements in census taking. Influenced by Laplace and Fourier, he was the first to use the normal curve other than as an error law. The distributions of measurements, such as chest circumferences of Scottish soldiers and heights of French conscripts, illuminated the appearance of normally distributed measures in nature and inspired work in fields as diverse as astronomy and physics. At an observatory in Brussels that he established in 1833 at the request of the Belgium government, Quetelet worked on statistical, geophysical, and meteorological data; studied meteor showers; and established methods for the comparison and evaluation of data. His studies of the numerical

consistency of crimes stimulated wide discussion of free will versus social determinism. His work produced great controversy among social scientists of the 19th century. Finally, the internationally used measure of obesity, the Body Mass Index (BMI), is derived from the Quetelet index.

Spearman, Charles Edward (1863–1945): British psychologist under whose leadership at University College emerged the "London School" of psychology, distinguished by its rigorous statistical and psychometric approach. Spearman formulated a two-factor theory of human intelligence, in which one factor is common to all mental activities and the other is task specific. He came to identify the first factor through the intercorrelations that existed between scores of subjects on various tests of intelligence. This quantifiable factor has come to be called *g* by cognitive psychologists. Spearman's model was based on a mathematical formulation that laid the groundwork for the statistical methods of factor analysis and contributed to research in test reliability.

Wilcoxon, Frank (1892–1965): American chemist who worked most of his career in industry researching fungicides and insecticides for the Boyce Thompson Institute, the Atlas Powder Company, and the American Cyanamid Company. He was a fellow of the American Statistical Association and the American Association for the Advancement of Science. Wilcoxon studied R. A. Fisher's *Statistical Methods for Research Workers*, which interested him in the application of statistics in experimentation, but through his research, he would seek statistical methods that were numerically simple and more easily understood and applied. Wilcoxon's main contribution to statistics was the development of nonparametric statistical processes, particularly the sign rank tests for two-sample and matched-pairs experiments and his method for multiple comparisons.

Yule, George Udny (1871–1951): Scottish statistician. Secretary, president, and fellow of the Royal Statistical Society. A student of Karl Pearson, Yule made fundamental contributions to the theory of regression and correlation, association between categorical variables, epidemiology, and times-series analysis.

Bibliography

Albert, Jim, and Jay Bennett. *Curve Ball: Baseball, Statistics, and the Role of Chance in the Game*. New York: Copernicus Books, 2003. Introduces the fundamental concepts of statistics through applications to historical baseball. Tackles in detail such issues as the best hitter and hitting streaks.

Barnett, Vic. *Comparative Statistical Inference*. London: John Wiley & Sons, 1973. This book discusses several different approaches to statistical inference and decision making. It compares Bayesian statistics and frequentist statistics and a decision-oriented approach.

Berry, Donald A. *Statistics: A Bayesian Perspective*. Belmont, CA: Duxbury Press at Wadsworth Publishing Company, 1996. An excellent elementary introduction to statistics, with many interesting real examples, most from medicine. The Bayesian approach is used.

Berry, Donald A., and Bernard W. Lindgren. *Statistics: Theory and Methods*, 2nd ed. Belmont, CA: Duxbury Press at Wadsworth Publishing Company, 1996. Introduces theory and application of modern statistics; designed for a year-long course in calculus-based statistics.

Bowerman, B., R. O'Connell, and A. Koehler. *Forecasting, Time Series, and Regression: An Applied Approach*, 4th ed., part II. Belmont, CA: Thomson, Brooks-Cole, 2005. Section of an undergraduate text that offers a good, concise explanation of regression and multiple-regression processes.

Cook, R. Dennis, and Sanford Weisberg. *Applied Regression Including Computing and Graphics*. New York: John Wiley & Sons, 1999. Text for a one-semester undergraduate/graduate course in regression techniques and graphics.

Gonick, Larry, and Woollcott Smith. *A Cartoon Guide to Statistics*. New York: Harper, 1993. A light but very informative view of serious statistics. It often goes right to the heart of a fundamental question.

Gould, Stephen J. *Full House: The Spread of Excellence from Plato to Darwin.* New York: Three Rivers Press, 1996. One of the most popular and prolific science writers of our time, Gould writes equally well about statistics of cancer and of .400 hitters. See part 2, chapter 4, and part 3, chapters 9 and 10.

Heyde, C. C., and E. Seneta, eds. *Statisticians of the Centuries*. New York: Springer-Verlag, 2001. This book contains short biographies of statisticians from the 16th to the 20th centuries.

Huff, Darrell. *How to Lie with Statistics*. New York: W.W. Norton, 1954. This charming little book has been in continuous publication since 1954. It is eminently readable and cheerfully describes methods to mislead with statistics.

Jaynes, E. T., and G. Larry Bretthorst, eds. *Probability Theory: The Logic of Science*. Cambridge, UK: Cambridge University Press, 2003. This book makes the case for the Bayesian approach to statistics, pointing out the difficulties in the orthodox approach. The book is technical.

Johnson, Norman L., and Samuel Kotz, eds. *Leading Personalities in Statistical Sciences: From the Seventeenth Century to the Present.* New York: Wiley, 1997. This book presents biographies of more than 100 statisticians and probabilists of the last four centuries.

Lewis, Michael. *Moneyball: The Art of Winning an Unfair Game*. New York: W.W. Norton, 2003. A delightful story of statistical power in baseball management.

Moore, David S. *Statistics: Concepts and Controversies*, 5th ed. New York: W.H. Freeman and Company, 2001. Informal introductory text on statistical ideas and reasoning; relates these to public policy, science, medicine, sociology, and daily life.

Moore, David S., and George P. McCabe. *Introduction to the Practice of Statistics*, 5th ed. New York: W.H. Freeman and Company, 2005. Collegiate-level introductory text focusing on data analysis, statistical reasoning, and the use of statistics in everyday life.

Paulos, John A. *A Mathematician Plays the Stock Market*. New York: Basic Books, 2003. Engaging, witty stories about what mathematical thinking can disclose about the stock market.

Peters, William S. *Counting for Something: Statistical Principles and Personalities*. New York: Springer-Verlag, 1987. This book illustrates both statistical topics and historical information using short chapters devoted to applications of statistical theory and the personalities who discovered them.

Porter, Theodore M. *The Rise of Statistical Thinking, 1820–1900*. Princeton, NJ: Princeton University Press, 1986. A fascinating historical description of the 19th-century background that led to the innovative development of modern statistics during the early 1900s.

Saari, Donald G. *Chaotic Elections! A Mathematician Looks at Voting*. Providence, RI: American Mathematical Society, 2001. An excellent introduction to the mathematics of voting, written for readers with high school mathematics. Includes an analysis of the American presidential voting scheme.

Salsburg, David. *The Lady Tasting Tea: How Statistics Revolutionized Science in the Twentieth Century*. New York: Henry Holt & Co., 2001. This book presents the history of statistics during the 20th century with humor and insight. Historical anecdotes bring the topics of statistics to life. Readable and enjoyable.

Schmitt, Samuel A. *Measuring Uncertainty: An Elementary Introduction to Bayesian Statistics.* Reading, MA: Addison-Wesley Publishing, 1969. An introduction to Bayesian statistics.

Tufte, Edward R. *The Visual Display of Quantitative Information*. Cheshire, CT: Graphics Press, 1983. Classic text of principles and theories of data graphics, including many historical examples of excellence in visual display.

Wainer, Howard. *A Trout in the Milk and Other Visual Adventures*. Princeton, NJ: Princeton University Press, 2005. Data graphics from a historical perspective in three parts. Part I is the story of the invention and early history of graphing data. Part II illustrates the power of the invention of data graphics, and Part III looks toward the future of visual representations of data.

———. *Visual Revelations: Graphical Tales of Fate and Deception from Napoleon Bonaparte to Ross Perot*. New York: Copernicus, Springer-Verlag, 1997. Descriptions of graphical failures and successes through history. That is, a discussion of the use of misleading graphs and poor execution of visual displays, contrasted with exemplary graphs that led to the discovery of a previously unknown relationship.

Watkins, Ann E., Richard L. Scheaffer, and George W. Cobb. *Statistics in Action: Understanding a World of Data*. Emeryville, CA: Key Curriculum Press, 2004. A textbook that emphasizes student activities.

Software:

Fathom: Dynamic Data Software, version 2. Emeryville, CA: Key Curriculum Press, Key College Publishing, 2005. Fathom is a user-friendly and powerful software system that allows the user to interactively bring ideas in statistics to life. Just by clicking and dragging, it is possible to manipulate data to explore many statistical ideas and illustrate them graphically.

Internet Resources:

Arc Software. University of Minnesota School of Statistics. Graphics can be produced in Arc, a free, downloadable statistical analysis tool for regression problems, as described in the book *Applied Regression Including Computing and Graphics* by Cook and Weisberg. http://www.stat.umn.edu/arc/software.html.

Barton, Paul E. "One-Third of a Nation." Educational Testing Service. This 46-page report discusses the fact that one-third of high school students in the United States do not graduate. http://www.ets.org/Media/Education_Topics/pdf/onethird.pdf.

The Baseball Archive. This website contains all sorts of data about baseball, including playing statistics, awards, records and feats, history, and salary and payrolls. www.baseball1.com.

Bureau of the Public Debt, U.S. Department of the Treasury. This website contains detailed information about the national debt. http://www.publicdebt.treas.gov/.

"Chance." This website is devoted to helping the teaching of probability or statistics courses. It contains lectures and other materials and a list of links to other sites. http://www.dartmouth.edu/~chance/.

Chance Magazine. This site has accessible articles about statistics and probability and their applications. http://www.amstat.org/publications/chance/index.html.

Consortium for the Advancement of Undergraduate Statistics Education (CAUSE) (homepage). This site has links to many resources about teaching and learning statistics. http://www.CAUSEweb.org.

Consumer Price Indexes (CPI). Bureau of Labor Statistics of the U.S. Department of Labor. This site contains a wealth of current and historical CPI data and a very complete description of the CPI and how it is computed. http://www.bls.gov/cpi/home.htm.

Fisher, R. A. "Has Mendel's work been rediscovered?" *Annals of Science*, 1 (1936):115–137. This 1936 article by one of the giants of statistics discusses the data in Mendel's famous experiments concerning inheritance of plant characteristics. Fisher's article argues that Mendel's data are too good to be unbiased reporting of real plant growth. http://www.library.adelaide.edu.au/digitised/fisher/144.pdf.

Gehl, B. K., and D. Watson. *Defining the Structure of Jealousy through Factor Analysis*. Los Angeles: Society for Personality and Social Psychology, February 2003. Poster presented at the Society for Personality and Social Psychology Annual Meeting. Available at http://www.psychology.uiowa.edu/students/gehl/definingjealousy.doc.

Index of Biographies. School of Mathematics and Statistics, University of St. Andrews, Scotland. This website gives biographical information about thousands of noted mathematicians. Both chronological and alphabetical indexes are presented, as well as such categories as female mathematicians, famous curves, history topics, and so forth. http://www-groups.dcs.st-andrews.ac.uk/~history/BiogIndex.html.

R Development Core Team (2004). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0. This website contains a downloadable statistical software tool for computing and graphing statistical calculations. http://www.R-project.org.

"StatCrunch, Statistical Software for Data Analysis on the Web." This award-winning data analysis package runs entirely in a web browser. StatCrunch includes most features that would be used in an introductory statistics course. http://www.statcrunch.com.