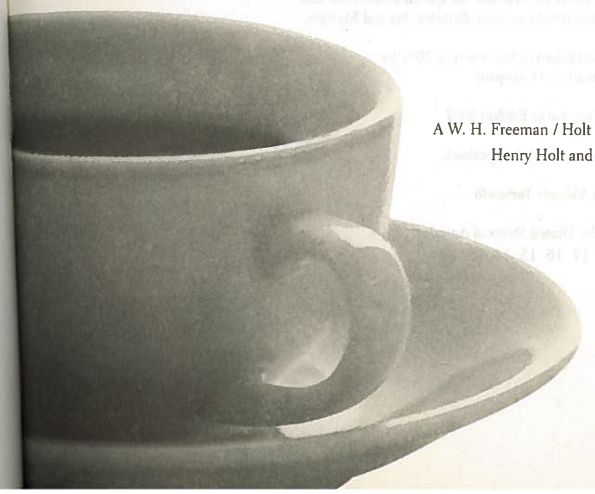
# The Lady Tasting Tea

## How Statistics Revolutionized Science in the Twentieth Century

David Salsburg



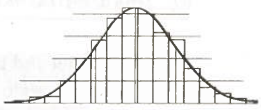
A W. H. Freeman / Holt Paperback
Henry Holt and Company
New York

has to be specific and enable the scientist to determine the difference in outcome that is due to weather versus the difference that is due to the use of different fertilizers. In particular, it is necessary to include all the treatments being compared in the same experiment, something that came to be called "controls."

In his book, *The Design of Experiments*, Fisher provided a few examples of good experimental designs, and derived general rules for good designs. However, the mathematics involved in Fisher's methods were very complicated, and most scientists were unable to generate their own designs unless they followed the pattern of one of the designs Fisher derived in his book.

Agricultural scientists recognized the great value of Fisher's work on experimental design, and Fisherian methods were soon dominating schools of agriculture in most of the English-speaking world. Taking off from Fisher's initial work, an entire body of scientific literature has developed to describe different experimental designs. These designs have been applied to fields other than agriculture, including medicine, chemistry, and industrial quality control. In many cases, the mathematics involved are deep and complicated. But, for the moment, let us stop with the idea that the scientist cannot just go off and "experiment." It takes some long and careful thought—and often a strong dose of difficult mathematics.

And the lady tasting tea, what happened to her? Fisher does not describe the outcome of the experiment that sunny summer afternoon in Cambridge. But Professor Smith told me that the lady identified every single one of the cups correctly.



CHAPTER

2

# THE SKEW DISTRIBUTIONS

As with many revolutions in human thought, it is difficult to find the exact moment when the idea of a statistical model became part of science. One can find possible specific examples of it in the work of the German and French mathematicians of the early nineteenth century, and there is even a hint of it in the papers of Johannes Kepler, the great seventeenth-century astronomer. As indicated in the preface to this book, Laplace invented what he called the error function to account for statistical problems in astronomy. I would prefer to date the statistical revolution to the work of Karl Pearson in the 1890s. Charles Darwin recognized biological variation as a fundamental aspect of life and made it the basis of his theory of the survival of the fittest. But it was his fellow Englishman Karl Pearson who first recognized the underlying nature of statistical models and how they offered something different from the deterministic view of nineteenth-century science.

When I began the study of mathematical statistics in the 1960s, Pearson was seldom mentioned in my

classes. As I met and talked with the major figures in the field, I heard no references to Pearson or his work. He was either ignored or treated as a minor figure whose activities had long since been outmoded. Churchill Eisenhart, from the U.S. National Bureau of Standards, for instance, was studying at University College, London, during the final years of Karl Pearson's life. He remembered Pearson as a dispirited old man. The pace of statistical research had swept him by, dashing him and most of his work into the dustbin of the past. The bright young students at University College were flocking to study at the feet of the newer great men, one of them Karl Pearson's own son, but no one was coming to see old Karl in his lonely office far from the bustle of new, exciting research.

It wasn't always like this. In the 1870s, young Carl [sic] Pearson had left England to pursue his graduate studies in political science in Germany. There he became enamored of the work of Karl Marx. In tribute to Marx, he changed the spelling of his own first name to Karl. He returned to London with a doctorate in political science, having written two respectable books in the field. In the very heart of stuffy Victorian England, he had the audacity to organize a Young Men's and Women's Discussion Club. At the club, young men and women gathered together (unchaperoned), in an equality of the sexes modeled after the salons of upper-class German and French society. There they discussed the great political and philosophical problems of the world. The fact that Pearson met his wife in this environment suggests that there may have been more than one motive for founding the club. This little social venture provides some insight into Karl Pearson's original mind and his utter disregard for established tradition.

Although his doctorate was in political science, Pearson's main interests were in the philosophy of science and the nature of mathematical modeling. In the 1880s, he published *The Grammar of Science*, which went through a number of editions. For much of the period prior to World War I, this was considered one of the great books on the nature of science and mathematics. It is filled



*Karl Pearson, 1857–1936, with a bust of Raphael Weldon in the background*

with brilliant, original insights, which make it an important work in the philosophy of science. It was also written in a smooth, simple style that makes it accessible to anyone. You don't have to know mathematics to read and understand *The Grammar of Science*. Although, at this writing, the book is over a hundred years old, the insights and the ideas found in it are pertinent to much mathematical research of the twenty-first century and provide an understanding of the nature of science that holds true even today.

## THE GALTON BIOMETRICAL LABORATORY

At this point in his life, Pearson fell under the influence of the English scientist Sir Francis Galton. Most people who have heard

of Galton know him as the "discoverer" of fingerprints. The realization that fingerprints are unique to each individual and the methods usually used to classify and identify them are Galton's. The uniqueness of fingerprints lies in the occurrence of irregular marks and cuts in the finger patterns, which are called "Galton Marks." Galton did far more. Independently wealthy, he was a dilettante scientist who sought to bring mathematical rigor into the science of biology through the study of patterns of numbers. One of his first investigations involved the inheritance of genius. He collected information about pairs of fathers and sons who were reputed to be highly intelligent. He found the problem very difficult, however, because there was no good measure of intelligence at the time. He decided to look at the inheritance of traits that were more easily measured, like height.

Galton set up a biometrical laboratory (*bio* for biology, *metric* for measurement) in London and advertised for families to come and be measured. At the biometrical laboratory, he collected heights, weights, measurements of specific bones, and other characteristics of family members. He and his assistants tabulated these data and examined and reexamined them. He was looking for some way to predict measures from parents to children. It was obvious, for instance, that tall parents tended to have tall children, but was there some mathematical formula that would predict how tall the children would be, using only the heights of the parents?

## Correlation and Regression

In this way, Galton discovered a phenomenon he called "regression to the mean." It turned out that sons of very tall fathers tended to be shorter than their fathers and sons of very short fathers tended to be taller than their fathers. It was as if some mysterious force were causing human heights to move away from the extremes and toward the mean or average of all humans. The phenomenon of regression to the mean holds for more than human heights. Almost



all scientific observations are bedeviled by regression to the mean. We shall see in chapters 5 and 7 how R. A. Fisher was able to turn Galton's regression to the mean into statistical models that now dominate economics, medical research, and much of engineering.

Galton thought about his remarkable finding and then realized that it had to be true, that it could have been predicted before making all his observations. Suppose, he said, that regression to the mean did not occur. Then, on the average, the sons of tall fathers would be as tall as their fathers. In this case, some of the sons would have to be taller than their fathers (in order to average out the ones who are shorter). The sons of this generation of taller men would then average their heights, so some sons would be even taller. It would go on, generation after generation. Similarly, there would be some sons shorter than their fathers, and some grandsons even shorter, and so on. After not too many generations, the human race would consist of ever taller people at one end and ever shorter ones at the other.

This does not happen. The heights of humans tend to remain stable, on the average. This can only happen if the sons of very tall fathers average shorter heights and the sons of very short fathers average greater heights. Regression to the mean is a phenomenon that maintains stability and keeps a given species pretty much the "same" from generation to generation.

Galton discovered a mathematical measure of this relationship. He called it the "coefficient of correlation." Galton gave a specific formula for computing this number from the type of data he collected at the biometrical laboratory. It is a highly specific formula for measuring one aspect of regression to the mean, but tells us nothing whatsoever about the cause of that phenomenon. Galton first used the word *correlation* in this sense. It has since moved into popular language. Correlation is often used to mean something much more vague than Galton's specific coefficient of correlation. It has a scientific ring to its sound, and nonscientists often bandy the word around as if it described how two things are related. But

unless you are referring to Galton's mathematical measure, you are not being very precise or scientific if you use the word correlation, which Galton used for this specific purpose.

## DISTRIBUTIONS AND PARAMETERS

With the formula for correlation, Galton was getting very close to this new revolutionary idea that was to modify almost all science in the twentieth century. But it was his disciple, Karl Pearson, who first formulated the idea in its most complete form.

To understand this revolutionary idea, you have to cast aside all preconceived notions about science. Science, we are often taught, is measurement. We make careful measurements and use them to find mathematical formulas that describe nature. In high school physics, we are taught that the distance a falling body will travel versus time is given by a formula involving a symbol g, where g is the constant of acceleration. We are taught that experiments can be used to determine the value of g. Yet, when the high school student runs a sequence of experiments to determine the value of g, rolling small weights along an inclined plane and measuring how long it takes them to get to different places on the ramp, what happens? It seldom comes out right. The more times the student runs the experiment, the more confusion occurs, as different values of g emerge from different experiments. The teacher looks down from his superior knowledge and assures the students that they are not getting the right answer not because they are sloppy or being careless or have copied incorrect numbers.

What he does not tell them is that all experiments are sloppy and that very seldom does even the most careful scientist get the number right. Little unforeseen and unobservable glitches occur in every experiment. The air in the room might be too warm and the sliding weight might stick for a microsecond before it begins to slide. A slight breeze from a passing butterfly might have an effect. What one really gets out of an experiment is a scatter of numbers,



not one of which is right but all of which can be used to get a close estimate of the correct value.

Armed with Pearson's revolutionary idea, we do not look upon experimental results as carefully measured numbers in their own right. Instead, they are examples of a scatter of numbers, a *distribution* of numbers, to use the more accepted term. This distribution of numbers can be written as a mathematical formula that tells us the probability that an observed number will be a given value. What value that number actually takes in a specific experiment is unpredictable. We can only talk about probabilities of values and not about certainties of values. The results of individual experiments are random, in the sense that they are unpredictable. The statistical models of distributions, however, enable us to describe the mathematical nature of that randomness.

It took some time for science to realize the inherent randomness of observations. In the eighteenth and nineteenth centuries, astronomers and physicists created mathematical formulas that described their observations to a degree of accuracy that was acceptable. Deviations between observed and predicted values were expected because of the basic imprecision of the measuring instruments, and were ignored. Planets and other astronomical bodies were assumed to follow precise paths determined by the fundamental equations of motion. Uncertainty was due to poor instrumentation. It was not inherent in nature.

With the development of ever more precise measuring instruments in physics, and with attempts to extend this science of measurement to biology and sociology, the inherent randomness of nature became more and more clear. How could this be handled? One way was to keep the precise mathematical formulas and treat the deviations between the observed values and the predicted values as a small, unimportant error. In fact, as early as 1820, mathematical papers by Laplace describe the first probability distribution, the error distribution, that is a mathematical formulation of the probabilities associated with these small, unimportant errors.

This error distribution has entered popular parlance as the "bell-shaped curve," or the normal distribution.[1]

It took Pearson to go one step beyond the normal, or error, distribution. Looking at the data accumulated in biology, Pearson conceived of the measurements themselves, rather than errors in the measurement, as having a probability distribution. Whatever we measure is really part of a random scatter, whose probabilities are described by a mathematical function, the distribution function. Pearson discovered a family of distribution functions that he called the "skew distributions" and that, he claimed, would describe any type of scatter a scientist might see in data. Each of the distributions in this family is identified by four numbers.

The numbers that identify the distribution function are not the same type of "number" as the measurements. These numbers can never be observed but can be inferred from the way in which the measurements scatter. These numbers were later to be called parameters—from the Greek for "almost measurements." The four parameters that completely describe a member of the Pearson System are called

1. the mean—the central value about which the measurements scatter,
2. the standard deviation—how far most of the measurements scatter about the mean,
3. symmetry—the degree to which the measurements pile up on only one side of the mean,
4. kurtosis—how far rare measurements scatter from the mean.

[1] It is sometimes called the Gaussian distribution, in honor of the man once believed to have first formulated it, except that it was not Carl Friedrich Gauss but an earlier mathematician named Abraham de Moivre who first wrote down the formula for the distribution. There is good reason to believe Daniel Bernoulli came across the formula before this. All of this is an example of what Stephen Stigler, a contemporary historian of science, calls the law of misonomy, that nothing in mathematics is ever named after the person who discovered it.



There is a subtle shift in thinking with Pearson's system of skew distributions. Before Pearson, the "things" that science dealt with were real and palpable. Kepler attempted to discover the mathematical laws that described how the planets moved in space. William Harvey's experiments tried to determine how blood moved through the veins and arteries of a specific animal. Chemistry dealt with elements and compounds made up of elements. However, the "planets" that Kepler tried to tame were really a set of numbers identifying the positions in the sky where shimmering lights were seen by observers on earth. The exact course of blood through the veins of a single horse was different from what might have been seen with a different horse, or with a specific human being. No one was able to produce a pure sample of iron, although it was known to be an element.

Pearson proposed that these observable phenomena were only random reflections. What was real was the probability distribution. The real " things" of science were not things that we could observe and hold but mathematical functions that described the randomness of what we could observe. The four parameters of a distribution are what we really want to determine in a scientific investigation. In some sense, we can never really determine those four parameters. We can only estimate them from the data.

Pearson failed to recognize this last distinction. He believed that if we collected enough data the estimates of the parameters would provide us with true values of the parameters. It took his younger rival, Ronald Fisher, to show that many of Pearson's methods of estimation were less than optimal. In the late 1930s, as Karl Pearson was approaching the end of his long life, a brilliant young Polish mathematician, Jerzy Neyman, showed that Pearson's system of skew distributions did not cover the universe of possible distributions and that many important problems could not be solved using the Pearson system.

But let us leave the old, abandoned Karl Pearson of 1934 and return to the vigorous man in his late thirties, who was filled with

enthusiasm over his discovery of skew distributions. In 1897, he took over Galton's biometrical laboratory in London and marshaled legions of young women (called "calculators") to compute the parameters of distributions associated with the data Galton had been accumulating on human measurements. At the turn of the new century, Galton, Pearson, and Raphael Weldon combined their efforts to found a new scientific journal that would apply Pearson's ideas to biological data. Galton used his wealth to create a trust fund that would support this new journal. In the first issue, the editors set forth an ambitious plan.

## THE PLAN OF *BIOMETRIKA*

Galton, Pearson, and Weldon were part of an exciting cadre of British scientists who were exploiting the insights of one of their most prominent members, Charles Darwin. Darwin's theories of evolution postulated that life forms change in response to environmental stress. He proposed that changing environments gave a slight advantage to those random changes that fit better into the new environment. Gradually, as the environment changed and life forms continued to have random mutations, a new species would emerge that was better fit to live and procreate in the new environment. This idea was given the shorthand designation "survival of the fittest." It had an unfortunate effect on society when arrogant political scientists adapted it to social life, declaring that those who emerged triumphant from the economic battle over riches were more fit than those who plunged into poverty. Survival of the fittest became a justification for rampant capitalism in which the rich were given the moral authority to ignore the poor.

In the biological sciences, Darwin's ideas seemed to have great validity. Darwin could point to the resemblances among related species as suggesting a previous species out of which these modern ones had emerged. Darwin showed how small birds of slightly different species and living on isolated islands had many anatomical commonalities. He pointed to the similarities among embryos of different species, including the human embryo, which starts with a tail.



The one thing Darwin was unable to show was an example of a new species actually emerging within the time frame of human history. Darwin postulated that new species emerge because of the survival of the fittest, but there was no proof of this. All he had to display were modern species that appeared to "fit" well within their environment. Darwin's proposals seemed to account for what was known, and they had an attractive logical structure to them. But, to translate an old Yiddish expression, "For instance is no proof."

Pearson, Galton, and Weldon set out in their new journal to rectify this. In Pearson's view of reality as probability distributions, Darwin's finches (an important example he used in his book) were not the objects of scientific investigation. The random distribution of all finches of a species was the object. If one could measure the beak lengths of all the finches in a given species, the distribution function of those beak lengths would have its own four parameters, and those four parameters would *be* the beak length of the species.

Suppose, Pearson said, that there was an environmental force changing a given species by providing superior survivorship to certain specific random mutations. We might not be able to live long enough to see a new species emerge, but we might be able to see a change in the four parameters of the distribution. In their first issue, the three editors declared that their new journal would collect data from all over the world and determine the parameters of their distributions, with the eventual hope of showing examples of shifts in parameters associated with environmental change.

They named their new journal *Biometrika*. It was funded lavishly by the Biometrika Trust that Galton set up, and was so well funded that it was the first journal to publish full color photographs and foldout glassine sheets with intricate drawings. It was printed on high-quality rag paper, and the most complicated mathematical formulas were displayed, even if they meant extremely complicated and expensive typesetting.

For the next twenty-five years, *Biometrika* printed data from correspondents who plunged into the jungles of Africa to measure tibia and fibula of the natives; sent in beak lengths of exotic tropical birds caught in the rain forests of Central America; or raided ancient cemeteries to uncover human skulls, into which they poured buckshot to measure cranial capacity. In 1910, the journal published several sheets of full color photographs of flaccid penises of pygmy men, laid on a flat surface against measuring sticks.

In 1921, a young female correspondent, Julia Bell, described the troubles she underwent when she tried to get anthropomorphic measurements of recruits for the Albanian army. She left Vienna for a remote outpost in Albania, assured that she would find German-speaking officers to help her. When she arrived, there was only a sergeant, who spoke three words of German. Undaunted, she took out her bronze measuring rods and got the young men to understand what she wanted by tickling them until they lifted their arms or legs as she desired.

For each of these data sets, Pearson and his calculators computed the four parameters of the distributions. The articles would display a graphical version of the best-fitting distribution and some comments about how this distribution differed from the distribution of other related data. In retrospect, it is difficult to see how all this activity helped prove Darwin's theories. Reading through these issues of *Biometrika*, I get the impression that it soon became an effort that was done for its own sake and had no real purpose other than estimating parameters for a given set of data.

Scattered throughout the journal are other articles. Some of them involve theoretical mathematics dealing with problems that arise from the development of probability distributions. In 1908, for instance, an unknown author, publishing under the pseudonym of "Student," produced a result that plays a role in almost all modern scientific work, "Student"'s t-test. We will meet this anonymous author in later chapters and discuss his unfortunate role in mediating between Karl Pearson and Ronald Fisher.



Galton died in 1911, and Weldon had died in a skiing accident in the Alps before that. This left Pearson as the sole editor of *Biometrika* and the sole dispenser of the trust's money. In the next twenty years, it was Pearson's personal journal, which published what Pearson thought was important and did not publish what Pearson thought was unimportant. It was filled with editorials written by Pearson, in which he let his fertile imagination range over all sorts of issues. Renovation of an ancient Irish church uncovered bones in the walls, and Pearson used involved mathematical reasoning and measurements made on those bones to determine whether they were, in fact, the bones of a particular medieval saint. A skull was found that was purported to be the skull of Oliver Cromwell. Pearson investigated this in a fascinating article that described the known fate of Cromwell's body, and then compared measurements made on pictures painted of Cromwell to measurements made on the skull.[2] In other articles, Pearson examined the lengths of reigns of kings and the decline of the patrician class in ancient Rome, and made other forays into sociology, political science, and botany, all of them with a complicated mathematical gloss.

Just before his death, Karl Pearson published a short article entitled "On Jewish–Gentile Relationships," in which he analyzed anthropomorphic data on Jews and Gentiles from various parts of the world. He concluded that the racial theories of the National Socialists, the official name of the Nazis, were sheer nonsense, that there was no such thing as a Jewish race or, for that matter, an Aryan race. This final paper was well within the clear, logical, carefully reasoned tradition of his previous work.

[2]After the restoration of the monarchy, following Cromwell's dictatorship, a truce between the two factions in the civil war in England meant that the new rulers could not prosecute any of the living followers of Cromwell. However, there was nothing in the truce about the dead. So the bodies of Cromwell and two of the judges who had ordered the execution of Charles I were dug up and tried for the crime of regicide. They were convicted, and their heads were chopped off and placed on pikes above Westminster Abbey. The three heads were left there for years and eventually disappeared. A head, supposedly that of Cromwell, showed up in a "museum" in London. It was that head which Pearson examined. He concluded that it was, indeed, the head of Oliver Cromwell.

Pearson used mathematics to investigate many areas of human thought that few would consider the normal business of science. To read through his editorials in *Biometrika* is to meet a man with a universal range of interests and a fascinating capacity to cut to the heart of any problem and find a mathematical model with which to attack it. To read through his editorials is also to meet a strong-willed, highly opinionated man, who viewed subordinates and students as extensions of his own will. I think I would have enjoyed spending a day with Karl Pearson—provided I did not have to disagree with him.

Did they prove Darwin's theory of evolution through survival of the fittest? Perhaps they did. By comparing the distributions of cranial capacity from skulls in ancient cemeteries to those of modern men and women, they managed to show that the human species has been remarkably stable across many thousands of years. By showing that anthropomorphic measurements on aborigines had the same distribution as measurements taken on Europeans, they disproved the claims of some Australians that the aborigines were not human. Out of this work, Pearson developed a basic statistical tool known as the "goodness of fit test," which is an indispensable tool for modern science. It enables the scientist to determine whether a given set of observations is appropriate to a particular mathematical distribution function. In chapter 10, we shall see how Pearson's own son used this goodness of fit test to undermine much of what his father had accomplished.

As the twentieth century advanced, more and more of the articles in *Biometrika* dealt with theoretical problems in mathematical statistics and fewer dealt with distributions of specific data. When Karl Pearson's son, Egon Pearson, took over as editor, the shift to theoretical mathematics was complete, and today *Biometrika* is a preeminent journal in that field.

But did they prove survival of the fittest? The closest they came to it occurred early in the twentieth century. Raphael Weldon conceived of a grand experiment. The development of china factories in southern England in the eighteenth century had caused some of the rivers to become silted with clay, so the harbors of Plymouth and Dartmouth had changed, with the interior regions more silted than those closer to the sea. Weldon took several hundred crabs from these harbors and put them into individual glass jars. In half the jars he used the silted water from the inner harbors. In the other half of the jars he used clearer water from the outer harbors. He then measured the carapaces of the crabs that survived after a period of time and determined the parameters of the two distributions of crabs: those that survived in clear water and those that survived in silted water.

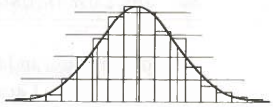


Just as Darwin had predicted, the crabs that survived in the silted jars showed a change in distribution parameters! Did this prove the theories of evolution? Unfortunately, Weldon died before he could write up the results of his experiment. Pearson described the experiment and its results in a preliminary analysis of the data, but a final analysis was never run. The British government, which had supplied the funds for the experiment, demanded a final report. It never came. Weldon was dead, and the experiment was ended.

Eventually, Darwin's theories were shown to be true for short-lived species like bacteria and fruit flies. Using these species, the scientist could experiment with thousands of generations in a short interval of time. Modern investigations of DNA, the building blocks of heredity, have provided even stronger evidence of the relationships among species. If we assume that the rate of mutation has been constant over the past ten million or more years, studies of DNA can be used to estimate the time frame of species emergence for primates and other mammals. At a minimum, it runs into the hundreds of thousands of years. Most scientists now accept Darwin's mechanism of evolution as correct. No other theoretical mechanism has been proposed that matches all known data so

well. Science is satisfied, and the idea that one needs to determine the shift in distribution parameters to show evolution on a short time scale has been dropped.

What remains of the Pearsonian revolution is the idea that the "things" of science are not the observables but the mathematical distribution functions that describe the probabilities associated with observations. Today, medical investigations use subtle mathematical models of distributions to determine the possible effects of treatments on long-term survival. Sociologists and economists use mathematical distributions to describe the behavior of human society. In the form of quantum mechanics, physicists use mathematical distributions to describe subatomic particles. No aspect of science has escaped the revolution. Some scientists claim that the use of probability distributions is a temporary stopgap and that, eventually, we will be able to find a way to return to the determinism of nineteenth-century science. Einstein's famous dictum that he did not believe that the Almighty plays dice with the universe is an example of that view. Others believe that nature is fundamentally random and that the only reality lies in distribution functions. Regardless of one's underlying philosophy, the fact remains that Pearson's ideas about distribution functions and parameters came to dominate twentieth-century science and stand triumphant on the threshold of the twenty-first century.

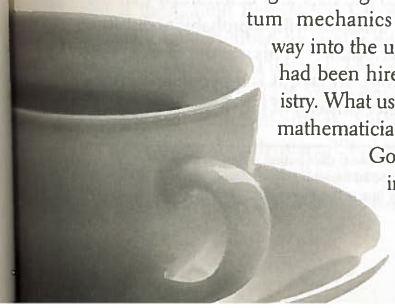


CHAPTER

3

# THAT DEAR MR. GOSSET

That ancient and honorable firm, the Guinness Brewing Company of Dublin, Ireland, began the twentieth century with an investment in science. Young Lord Guinness had recently inherited the enterprise, and he decided to introduce modern scientific techniques into the business by hiring the leading graduates in chemistry from Oxford and Cambridge Universities. In 1899, he recruited William Sealy Gosset, who had just graduated from Oxford at age twenty-three with a combined degree in chemistry and mathematics. Gosset's mathematical background was a traditional one of the time, including calculus, differential equations, astronomy, and other aspects of the clockwork universe view of science. The innovations of Karl Pearson and the first glimmerings of what was to become quantum mechanics had not yet made their way into the university curriculum. Gosset had been hired for his expertise in chemistry. What use would a brewery have for a mathematician?



Gosset turned out to be a good investment for Guinness. He showed himself to be a very