but also *how* a complex observational record may be handled with intelligibility and precision.

The subject is a new one, and in many ways the most that the author can hope is to suggest possible lines of attack on the problems with which others are confronted. Progress in recent years has been rapid, and the few sections devoted to the subject in the author's *Statistical Methods for Research Workers*, first published in 1925, have, with each succeeding edition, come to appear more and more inadequate. On purely statistical questions the reader must be referred to that book; on logic, and the analysis of meaning, to *Statistical Methods and Scientific Inference*. The present volume is an attempt to do more thorough justice to the problems of planning and foresight with which the experimenter is confronted.

REFERENCES AND OTHER READING

T. Bayes (1763). An essay towards solving a problem in the doctrine of chances. Philosophical Transactions of the Royal Society, liii. 370.

A. de Morgan (1838). An essay on probabilities and on their application to life contingencies and insurance offices. Preface, vi. Longman & Co.

R. A. Fisher (1930). Inverse probability. Proceedings of the Cambridge Philosophical Society, xxvi. 528-535.

R. A. Fisher (1932). Inverse probability and the use of likelihood. Proceedings of the Cambridge Philosophical Society, xxviii. 257-261.

R. A. Fisher (1935). The logic of inductive inference. Journal Royal Statistical Society, xcviii. 39-54.

R. A. Fisher (1936). Uncertain inference. Proceedings of the American Academy of Arts and Sciences, 71. 245-258.

R. A. Fisher (1925-1963). Statistical methods for research workers. Oliver and Boyd Ltd., Edinburgh.

R. A. Fisher (1956, 1959) Statistical methods and scientific inference. Oliver and Boyd Ltd., Edinburgh.

II

THE PRINCIPLES OF EXPERIMENTATION, ILLUSTRATED BY A PSYCHO-PHYSICAL EXPERIMENT

5. Statement of Experiment

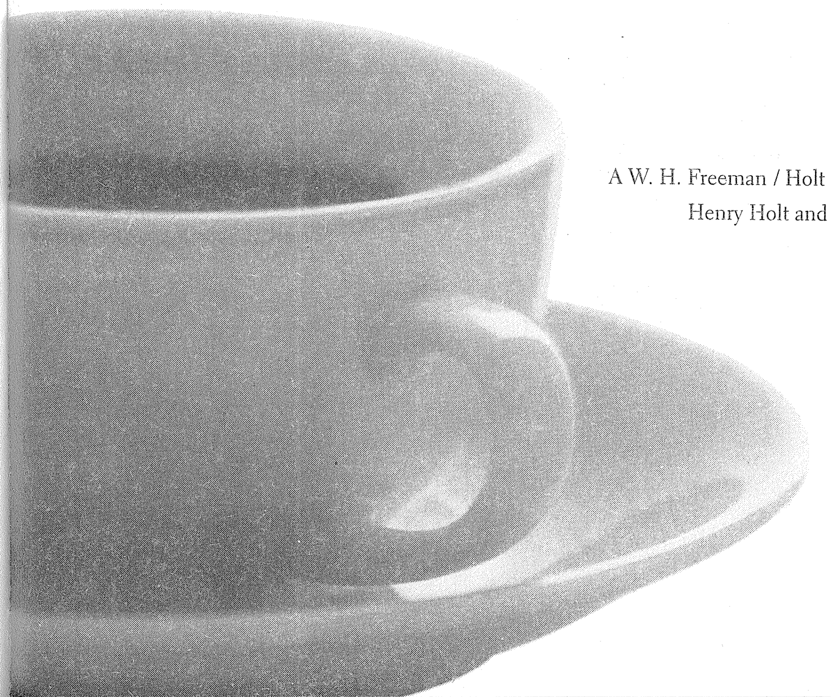
A lady declares that by tasting a cup of tea made with milk she can discriminate whether the milk or the tea infusion was first added to the cup. We will consider the problem of designing an experiment by means of which this assertion can be tested. For this purpose let us first lay down a simple form of experiment with a view to studying its limitations and its characteristics, both those which appear to be essential to the experimental method, when well developed, and those which are not essential but auxiliary.

Our experiment consists in mixing eight cups of tea, four in one way and four in the other, and presenting them to the subject for judgment in a random order. The subject has been told in advance of what the test will consist, namely that she will be asked to taste eight cups, that these shall be four of each kind, and that they shall be presented to her in a random order, that is in an order not determined arbitrarily by human choice, but by the actual manipulation of the physical apparatus used in games of chance, cards, dice, roulettes, etc., or, more expeditiously, from a published collection of random sampling numbers purporting to give the actual results of such manipulation. Her task is to divide the 8 cups into two sets of 4, agreeing, if possible, with the treatments received.

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

CHAPTER

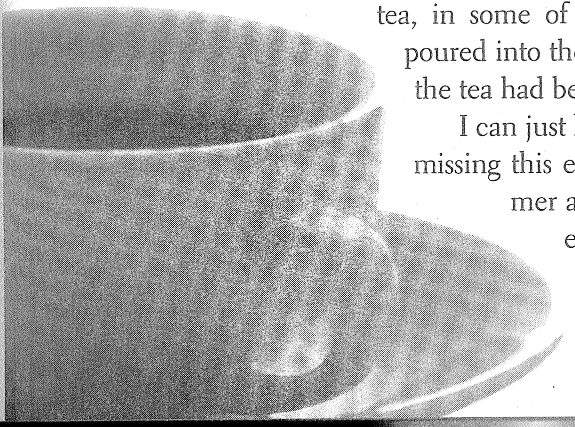
1

THE LADY TASTING TEA

It was a summer afternoon in Cambridge, England, in the late 1920s. A group of university dons, their wives, and some guests were sitting around an outdoor table for afternoon tea. One of the women was insisting that tea tasted different depending upon whether the tea was poured into the milk or whether the milk was poured into the tea. The scientific minds among the men scoffed at this as sheer nonsense. What could be the difference? They could not conceive of any difference in the chemistry of the mixtures that could exist. A thin, short man, with thick glasses and a Vandyke beard beginning to turn gray, pounced on the problem.

"Let us test the proposition," he said excitedly. He began to outline an experiment in which the lady who insisted there was a difference would be presented with a sequence of cups of tea, in some of which the milk had been poured into the tea and in others of which the tea had been poured into the milk.

I can just hear some of my readers dismissing this effort as a minor bit of summer afternoon fluff. "What difference does it make whether



the lady could tell one infusion from another?" they will ask. "There is nothing important or of great scientific merit in this problem," they will sneer. "These great minds should have been putting their immense brain power to something that would benefit mankind."

Unfortunately, whatever nonscientists may think about science and its importance, my experience has been that most scientists engage in their research because they are interested in the results and because they get intellectual excitement out of the work. Seldom do good scientists think about the eventual importance of their work. So it was that sunny summer afternoon in Cambridge. The lady might or might not have been correct about the tea infusion. The fun would be in finding a way to determine if she was right, and, under the direction of the man with the Vandyke beard, they began to discuss how they might make that determination.

Enthusiastically, many of them joined with him in setting up the experiment. Within a few minutes, they were pouring different patterns of infusion in a place where the lady could not see which cup was which. Then, with an air of finality, the man with the Vandyke beard presented her with her first cup. She sipped for a minute and declared that it was one where the milk had been poured into the tea. He noted her response without comment and presented her with the second cup....

THE COOPERATIVE NATURE OF SCIENCE

I heard this story in the late 1960s from a man who had been there that afternoon. He was Hugh Smith, but he published his scientific papers under the name H. Fairfield Smith. When I knew him, he was a professor of statistics at the University of Connecticut, in Storrs. I had received my Ph.D. in statistics from the University of Connecticut two years before. After teaching at the University of Pennsylvania, I had joined the clinical research department at Pfizer, Inc., a large pharmaceutical firm. Its research campus in Groton, Connecticut, was about an hour's drive from Storrs. I was dealing with many difficult mathematical problems at Pfizer. I was the only statistician there at that time, and I needed to talk over these problems and my "solutions" to them.

What I had discovered working at Pfizer was that very little scientific research can be done alone. It usually requires a combination of minds. This is because it is so easy to make mistakes. When I would propose a mathematical formula as a means of solving a problem, the model would sometimes be inappropriate, or I might have introduced an assumption about the situation that was not true, or the "solution" I found might have been derived from the wrong branch of an equation, or I might even have made a mistake in arithmetic.

Whenever I would visit the university at Storrs to talk things over with Professor Smith, or whenever I would sit around and discuss problems with the chemists or pharmacologists at Pfizer, the problems I brought out would usually be welcomed. They would greet these discussions with enthusiasm and interest. What makes most scientists interested in their work is usually the excitement of working on a problem. They look forward to the interactions with others as they examine a problem and try to understand it.

THE DESIGN OF EXPERIMENTS

And so it was that summer afternoon in Cambridge. The man with the Vandyke beard was Ronald Aylmer Fisher, who was in his late thirties at the time. He would later be knighted Sir Ronald Fisher. In 1935, he wrote a book entitled *The Design of Experiments*, and he described the experiment of the lady tasting tea in the second chapter of that book. In his book, Fisher discusses the lady and her belief as a hypothetical problem. He considers the various ways in which an experiment might be designed to determine if she could tell the difference. The problem in designing the experiment is that, if she is given a single cup of tea, she has a 50 percent chance of guessing correctly which infusion was used, even if she cannot tell the difference. If she is given two cups of tea, she still might guess correctly. In fact, if she knew that the two cups of tea were each made with a different infusion, one guess could be completely right (or completely wrong).

Similarly, even if she could tell the difference, there is some chance that she might have made a mistake, that one of the cups was not mixed as well or that the infusion was made when the tea was not hot enough. She might be presented with a series of ten cups and correctly identify only nine of them, even if she could tell the difference.

In his book, Fisher discusses the various possible outcomes of such an experiment. He describes how to decide how many cups should be presented and in what order and how much to tell the lady about the order of presentations. He works out the probabilities of different outcomes, depending upon whether the lady is or is not correct. Nowhere in this discussion does he indicate that such an experiment was ever run. Nor does he describe the outcome of an actual experiment.

The book on experimental design by Fisher was an important element in a revolution that swept through all fields of science in the first half of the twentieth century. Long before Fisher came on the scene, scientific experiments had been performed for hundreds of years. In the later part of the sixteenth century, the English physician William Harvey experimented with animals, blocking the flow of blood in different veins and arteries, trying to trace the circulation of blood as it flowed from the heart to the lungs, back to the heart, out to the body, and back to the heart again.

Fisher did not discover experimentation as a means of increasing knowledge. Until Fisher, experiments were idiosyncratic to each scientist. Good scientists would be able to construct experiments that produced new knowledge. Lesser scientists would often engage in "experimentation" that accumulated much data but was useless for increasing knowledge. An example of this can be seen in the many inconclusive attempts that were made during the late

nineteenth century to measure the speed of light. It was not until the American physicist Albert Michelson constructed a highly sophisticated series of experiments with light and mirrors that the first good estimates were made.

In the nineteenth century, scientists seldom published the results of their experiments. Instead, they described their conclusions and published data that "demonstrated" the truth of those conclusions. Gregor Mendel did not show the results of all his experiments in breeding peas. He described the sequence of experiments and then wrote: "The first ten members of both series of experiments may serve as an illustration. . . ." (In the 1940s, Ronald Fisher examined Mendel's "illustrations" of data and discovered that the data were too good to be true. They did not display the degree of randomness that should have occurred.)

Although science has been developed from careful thought, observations, and experiments, it was never quite clear how one should go about experimenting, nor were the complete results of experiments usually presented to the reader.

This was particularly true for agricultural research in the late nineteenth and early twentieth centuries. The Rothamsted Agricultural Experimental Station, where Fisher worked during the early years of the twentieth century, had been experimenting with different fertilizer components (called "artificial manures") for almost ninety years before he arrived. In a typical experiment, the workers would spread a mixture of phosphate and nitrogen salts over an entire field, plant grain, and measure the size of the harvest, along with the amount of rainfall during that summer. There were elaborate formulas used to "adjust" the output of one year or one field, in order to compare it to the output of another field or of the same field in another year. These were called "fertility indexes," and each agricultural experimental station had its own fertility index, which it believed was more accurate than any other.

The result of these ninety years of experimentation was a mess of confusion and vast troves of unpublished and useless data. It seemed as if some strains of wheat responded better than other strains to one fertilizer, but only in years when rainfall was excessive. Other experiments seemed to show that sulfate of potash one year, followed by sulfate of soda for the next year, produced an increase in some varieties of potatoes but not others. The most that could be said of these artificial manures was that some of them worked sometimes, perhaps, or maybe.

Fisher, a consummate mathematician, looked at the fertility index that the agricultural scientists at Rothamsted used to correct the results of experiments to account for differences due to the weather from year to year. He examined the competing indexes used by other agricultural experimental stations. When reduced to their elemental algebra, they were all versions of the same formula. In other words, two indexes, whose partisans were hotly contending, were really making exactly the same correction. In 1921, he published a paper in the leading agricultural journal, the *Annals of Applied Biology*, in which he showed that it did not make any difference what index was used. The article also showed that all these corrections were inadequate to adjust for differences in the fertility of different fields. This remarkable paper ended over twenty years of scientific dispute.

Fisher then examined the data on rainfall and crop production over the previous ninety years and showed that the effects of different weather from year to year were far greater than any effect of different fertilizers. To use a word Fisher developed later in his theory of experimental design, the year-to-year differences in weather and the year-to-year differences in artificial manures were "confounded." This means that there was no way to pull them apart using data from these experiments. Ninety years of experimentation and over twenty years of scientific dispute had been an almost useless waste of effort!

This set Fisher thinking about experiments and experimental design. He concluded that the scientist needs to start with a mathematical model of the outcome of the potential experiment. A mathematical model is a set of equations, in which some of the symbols stand for numbers that will be collected as data from the experiments and other symbols stand for the overall outcomes of the experiment. The scientist starts with the data from the experiment and computes outcomes appropriate to the scientific question being considered.

Consider a simple example from the experience of a teacher with a particular student. The teacher is interested in finding some measure of how much the child has learned. To this end, the teacher "experiments" by giving the child a group of tests. Each test is marked on a scale from 0 to 100. Any one test provides a poor estimate of how much the child knows. It may be that the child did not study the few things that were on that test but knows a great deal about things that were not on the test. The child may have had a headache the day she took a particular test. The child may have had an argument with parents the morning of a particular test. For many reasons, one test does not provide a good estimate of knowledge. So, the teacher gives a set of tests. The average score from all those tests is taken as a better estimate of how much the child knows. How much the child knows is the outcome. The scores on individual tests are the data.

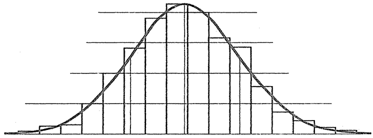
How should the teacher structure those tests? Should they be a sequence of tests that cover only the material taught over the past couple of days? Should they each involve something from all the material taught until now? Should the tests be given weekly, or daily, or at the end of each unit being taught? All of these are questions involved in the design of the experiment.

When the agricultural scientist wants to know the effect of a particular artificial fertilizer on the growth of wheat, an experiment has to be constructed that will provide data to estimate that effect. Fisher showed that the first step in the design of that experiment is to set up a group of mathematical equations describing the relationship between the data that will be collected and the outcomes that are being estimated. Then, any useful experiment has to be one that allows for estimation of those outcomes. The experiment 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