PEARSON'S GOODNESS OF FIT TEST

This problem was one that Karl Pearson recognized early in his career. One of Pearson's great achievements was the creation of the first "goodness of fit test." By comparing the observed to the predicted values, Pearson was able to produce a statistic that tested the goodness of the fit. He called his test statistic a "chi square goodness of fit test." He used the Greek letter chi (χ), since the distribution of this test statistic belonged to a group of his skew distributions that he had designated the chi family. Actually, the test statistic behaved like the square of a chi, thus the name "chi squared." Since this is a statistic in Fisher's sense, it has a probability distribution. Pearson proved that the chi square goodness of fit test has a distribution that is the same, regardless of the type of data used. That is, he could tabulate the probability distribution of this statistic and use that same set of tables for every test. The chi square goodness of fit test has a single parameter, which Fisher was to call the "degrees of freedom." In the 1922 paper in which he first criticized Pearson's work, Fisher showed that, for the case of comparing two proportions, Pearson had gotten the value of that parameter wrong.

But just because he made a mistake in one small aspect of his theory is no reason to denigrate Pearson's great achievement. Pearson's goodness of fit test was the forerunner of a major component of modern statistical analysis. This component is called "hypothesis testing," or "significance testing." It allows the analyst to propose

two or more competing mathematical models for reality and use the data to reject one of them. Hypothesis testing is so widely used that many scientists think of it as the only statistical procedure available to them. The use of hypothesis testing, as we shall see in later chapters, involves some serious philosophical problems.

Testing Whether the Lady Can Taste a Difference in the Tea

Suppose we wish to test whether the lady can detect the difference between a cup of tea into which the milk has been poured into the tea versus a cup of tea wherein the tea has been poured into the milk. We present her with two cups, telling her that one of the cups is tea into milk and the other is milk into tea. She tastes and identifies the cups correctly. She could have done this by guessing; she had a 50:50 chance of guessing correctly. We present her with another pair of the same type. Again, she identifies them correctly. If she were just guessing, the chance of this happening twice in a row is 1/4. We present her with a third pair of cups, and again she identifies them correctly. The chance that this has happened as a result of pure guesswork is 1/8. We present her with more pairs, and she keeps identifying the cups correctly. At some point, we have to be convinced that she can tell the difference. Suppose she was wrong with one pair. Suppose further that this was the twenty-fourth pair and she was correct on all the others. Can we still conclude that she is able to detect a difference? Suppose she was wrong in four out of the twenty-four? Five of the twenty-four?

Hypothesis, or significance, testing is a formal statistical procedure that calculates the probability of what we have observed, assuming that the hypothesis to be tested is true. When the observed probability is very low, we conclude that the hypothesis is not true. One important point is that hypothesis testing provides a tool for rejecting a hypothesis. In the case above, this is the hypothesis

that the lady is only guessing. It does not allow us to accept a hypothesis, even if the probability associated with that hypothesis is very high.

Somewhere early in the development of this general idea, the word *significant* came to be used to indicate that the probability was low enough for rejection. Data became significant if they could be used to reject a proposed distribution. The word was used in its late-nineteenth-century English meaning, which is simply that the computation signified or showed something. As the English language entered the twentieth century, the word *significant* began to take on other meanings, until it developed its current meaning, implying something very important. Statistical analysis still uses the word significant to indicate a very low probability computed under the hypothesis being tested. In that context, the word has an exact mathematical meaning. Unfortunately, those who use statistical analysis often treat a significant test statistic as implying something much closer to the modern meaning of the word.

FISHER'S USE OF P-VALUES

R. A. Fisher developed most of the significance testing methods now in general use. He referred to the probability that allows one to declare significance as the "p-value." He had no doubts about its meaning or usefulness. Much of *Statistical Methods for Research Workers* is devoted to showing how to calculate p-values. As I noted earlier, this was a book designed for nonmathematicians who want to use statistical methods. In it, Fisher does not describe how these tests were derived, and he never indicates exactly what p-value one might call significant. Instead, he displays examples of calculations and notes whether the result is significant or not. In one example, he shows that the p-value is less than .01 and states: "Only one value in a hundred will exceed [the calculated test statistic] by chance, so that the difference between the results is clearly significant."

The closest he came to defining a specific p-value that would be significant in all circumstances occurred in an article printed in the *Proceedings of the Society for Psychical Research* in 1929. Psychical research refers to attempts to show, via scientific methods, the existence of clairvoyance. Psychical researchers make extensive use of statistical significance tests to show that their results are improbable in terms of the hypothesis that the results are due to purely random guesses by the subjects. In this article, Fisher condemns some writers for failing to use significance tests properly. He then states:

> In the investigation of living beings by biological methods, statistical tests of significance are essential. Their function is to prevent us being deceived by accidental occurrences, due not to the causes we wish to study, or are trying to detect, but to a combination of many other circumstances which we cannot control. An observation is judged significant, if it would rarely have been produced, in the absence of a real cause of the kind we are seeking. It is a common practice to judge a result significant, if it is of such a magnitude that it would have been produced by chance not more frequently than once in twenty trials. This is an arbitrary, but convenient, level of significance for the practical investigator, but it does not mean that he allows himself to be deceived once in every twenty experiments. The test of significance only tells him what to ignore, namely all experiments in which significant results are not obtained. He should only claim that a phenomenon is experimentally demonstrable when he knows how to design an experiment so that it will rarely fail to give a significant result. Consequently, isolated significant results which he does not know how to reproduce are left in suspense pending further investigation.

Note the expression "knows how to design an experiment . . . that . . . will rarely fail to give a significant result." This lies at the heart of Fisher's use of significance tests. To Fisher, the significance test makes sense only in the context of a sequence of experiments, all aimed at elucidating the effects of specific treatments. Reading through Fisher's applied papers, one is led to believe that he used significance tests to come to one of three possible conclusions. If the p-value is very small (usually less than .01), he declares that an effect has been shown. If the p-value is large (usually greater than .20), he declares that, if there is an effect, it is so small that no experiment of this size will be able to detect it. If the p-value lies in between, he discusses how the next experiment should be designed to get a better idea of the effect. Except for the above statement, Fisher was never explicit about how the scientist should interpret a p-value. What seemed to be intuitively clear to Fisher may not be clear to the reader.

We will come back to reexamine Fisher's attitude toward significance testing in chapter 18. It lies at the heart of one of Fisher's great blunders, his insistence that smoking had not been shown to be harmful to health. But let us leave Fisher's trenchant analysis of the evidence involving smoking and health for later and turn to 35-year-old Jerzy Neyman in the year 1928.

JERZY NEYMAN'S MATHEMATICAL EDUCATION

Jerzy Neyman was a promising mathematics student when World War I erupted across his homeland in Eastern Europe. He was driven into Russia, where he studied at the University of Kharkov, a provincial outpost of mathematical activity. Lacking teachers who were up to date in their knowledge, and forced to miss semesters of schooling because of the war, he took the elementary mathematics he was taught at Kharkov and built upon it by seeking out

articles in the mathematics journals available to him. Neyman thus received a formal mathematics education similar to that taught to students of the nineteenth century, and then educated himself into twentieth-century mathematics.

The journal articles available to Neyman were limited to what he could find in the libraries of the University of Kharkov and later at provincial Polish schools. By chance, he came across a series of articles by Henri Lebesgue of France. Lebesgue (1875–1941) had created many of the fundamental ideas of modern mathematical analysis in the early years of the twentieth century, but his papers are difficult to read. Lebesgue integration, the Lebesgue convergence theorem, and other creations of this great mathematician have all been simplified and organized in more understandable forms by later mathematicians. Nowadays, no one reads Lebesgue in the original. Students all learn about his ideas through these later versions.

No one, that is, except Jerzy Neyman, who had only Lebesgue's original articles, who struggled through them, and who emerged seeing the brilliant light of these great new (to him) creations. For years afterward, Neyman idolized Lebesgue, and, in the late 1930s, finally got to meet him at a mathematics conference in France. According to Neyman, Henri Lebesgue turned out to be a gruff, impolite man, who responded to Neyman's enthusiasm with a few mutterings and turned and walked away in the midst of Neyman's talking to him.

Neyman was deeply hurt by this rebuff, and perhaps with this as an object lesson, always went out of his way to be polite and kind to young students, to listen carefully to what they said, and to engage them in their enthusiasms. That was Jerzy Neyman, the man. All who knew him remember him for his kindness and caring manners. He was gracious and thoughtful and dealt with people with genuine pleasure. When I met him, he was in his early eighties, a small, dignified, well-groomed man with a neat white

moustache. His blue eyes sparkled as he listened to others and engaged in intensive conversation, giving the same personal attention to everyone, no matter who they were.

In the early years of his career, Jerzy Neyman managed to find a position as a junior member of the faculty of the University of Warsaw. At that time, the newly independent nation of Poland had little money to support academic research, and positions for mathematicians were scarce. In 1928, he spent a summer at the biometrical laboratory in London and there came to know Egon S. Pearson and his wife, Eileen, and their two daughters. Egon Pearson was Karl Pearson's son, and a more striking contrast in personalities is hard to find. Where Karl Pearson was driving and dominating, Egon Pearson was shy and self-effacing. Karl Pearson rushed through new ideas, often publishing an article with the mathematics vaguely sketched in or even with some errors. Egon Pearson was extremely careful, worrying over the details of each calculation.

The friendship between Egon Pearson and Jerzy Neyman is preserved in their exchange of letters between 1928 and 1933. These letters provide a wonderful insight into the sociology of science, showing how two original minds grapple with a problem, each one proposing ideas or criticizing the ideas of the other. Pearson's self-effacing comes to the forefront as he hesitantly suggests that perhaps something Neyman had proposed might not work out. Neyman's great originality comes out as he cuts through complicated problems to find the essential nature of each difficulty. For someone who wants to understand why mathematical research is so often a cooperative venture, I recommend the Neyman-Pearson letters.

What was the problem that Egon Pearson first proposed to Neyman? Recall Karl Pearson's chi square goodness of fit test. He developed it to test whether observed data fit a theoretical distribution. There really is no such thing as *the* chi square goodness of fit test. The analyst has available an infinite number of ways to apply

the test to a given set of data. There appeared to be no criterion on how "best" to pick among those many choices. Every time the test is applied, the analyst must make arbitrary choices. Egon Pearson posed the following question to Jerzy Neyman:

> If I have applied a chi square goodness of fit test to a set of data versus the normal distribution, and if I have failed to get a significant p-value, how do I know that the data really fit a normal distribution? That is, how do I know that another version of the chi square test or another goodness of fit test as yet undiscovered might not have produced a significant p-value and allowed me to reject the normal distribution as fitting the data?

NEYMAN'S STYLE OF MATHEMATICS

Neyman took this question back to Warsaw with him, and the exchange of letters began. Both Neyman and young Pearson were impressed with Fisher's concept of estimation based on the likelihood function. They began their investigation by looking at the likelihood associated with a goodness of fit test. The first of their joint papers describes the results of those investigations. It is the most difficult of the three classic papers they produced, which were to revolutionize the whole idea of significance testing. As they continued looking at the question, Neyman's great clarity of vision kept distilling the problem down to its essential elements, and their work became clearer and easier to understand.

Although the reader may not believe it, literary style plays an important role in mathematical research. Some mathematical writers seem unable to produce articles that are easy to understand. Others seem to get a perverse pleasure out of generating many lines of symbolic notation so filled with detail that the general idea is lost in the picayune. But some authors have the ability to display complicated ideas with such force and simplicity that the development

appears to be obvious in their exposition. Only upon reviewing what has been learned does the reader realize the great power of the results. Such an author was Jerzy Neyman. It is a pleasure to read his papers. The ideas evolve naturally, the notation is deceptively simple, and the conclusions appear to be so natural that you find it hard to see why no one produced these results long before.

Pfizer Central Research, where I worked for twenty-seven years, sponsors a yearly colloquium at the University of Connecticut. The statistics department of the university invites a major figure in biostatistical research to come for a day, meet with students, and then present a talk in the late afternoon. Since I was involved in setting up the grant for this series, I had the honor of meeting some of the great men of statistics through them. Jerzy Neyman was one such invitee. He asked that his talk have a particular form. He wanted to present a paper and then have a panel of discussants who would criticize his paper. Since this was the renowned Jerzy Neyman, the organizers of the symposium contacted well-known senior statisticians in the New England area to constitute the panel. At the last minute, one of the panelists was unable to come, and I was asked to substitute for him.

Neyman had sent us a copy of the paper he planned to present. It was an exciting development, wherein he applied work he had done in 1939 to a problem in astronomy. I knew that 1939 paper; I had discovered it years before while still a graduate student, and I had been impressed by it. The paper dealt with a new class of distributions Neyman had discovered, which he called the "contagious distributions." The problem posed in the paper began with trying to model the appearance of insect grubs in soil. The female insect flew about the field, laden with eggs, then chose a spot at random to lay the eggs. Once the eggs were laid, the grubs hatched and crawled outward from that spot. A sample of soil is taken from the field. What is the probability distribution of the number of grubs found in that sample?

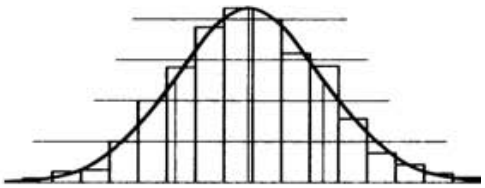
The contagious distributions describe such situations. They are derived in the 1939 paper with an apparently simple series of equa-

tions. This derivation seems obvious and natural. It is clear, when the reader gets to the end of the paper, that there is no other way to approach it, but this is clear only after reading Neyman. Since that 1939 paper, Neyman's contagious distributions have been found to fit a great many situations in medical research, in metallurgy, in meteorology, in toxicology, and (as described by Neyman in his Pfizer Colloquium paper) in dealing with the distribution of galaxies in the universe.

After he finished his talk, Neyman sat back to listen to the panel of discussants. All the other members of the panel were prominent statisticians who had been too busy to read his paper in advance. They looked upon the Pfizer Colloquium as a recognition of honor for Neyman. Their "discussions" consisted of comments about Neyman's career and past accomplishments. I had come onto the panel as a last-minute replacement and could not refer to my (nonexistent) previous experiences with Neyman. My comments were directed to Neyman's presentation that day, as he had asked. In particular, I told how I had discovered the 1939 paper years before and revisited it in anticipation of this session. I described the paper, as best I could, showing enthusiasm when I came to the clever way Neyman had developed the meaning of the parameters of the distribution.

Neyman was clearly delighted with my comments. Afterward, we had an exciting discussion about the contagious distributions and their uses. A few weeks later, a large package arrived in the mail. It was a copy of *A Selection of Early Statistical Papers of J. Neyman*, published by the University of California Press. On the inside cover was the inscription: "To Dr. David Salsburg, with hearty thanks for his interesting comments on my presentation of April 30, 1974. J. Neyman."

I treasure this book both for the inscription and the set of beautiful, well-written papers in it. I have since had the opportunity to talk with many of Neyman's students and coworkers. The friendly, interesting, and interested man I met in 1974 was the man that they knew and admired.



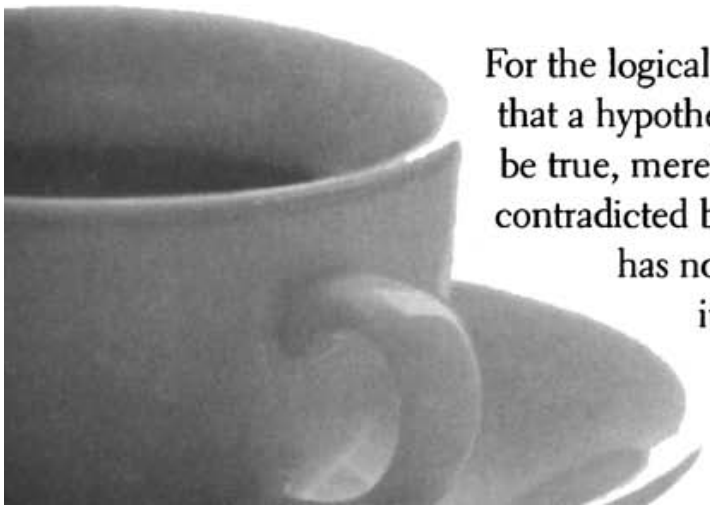
CHAPTER

11

HYPOTHESIS TESTING

At the start of their collaboration, Egon Pearson asked Jerzy Neyman how he could be sure that a set of data was normally distributed if he failed to find a significant p-value when testing for normality. Their collaboration started with this question, but Pearson's initial question opened the door to a much broader one. What does it mean to have a nonsignificant result in a significance test? Can we conclude that a hypothesis is true if we have failed to refute it?

R. A. Fisher had addressed that question in an indirect way. Fisher would take large p-values (and a failure to find significance) as indicating that the data were inadequate to decide. To Fisher, there was never a presumption that a failure to find significance meant that the tested hypothesis was true. To quote him:



> For the logical fallacy of believing that a hypothesis has been proved to be true, merely because it is not contradicted by the available facts, has no more right to insinuate itself in statistical than in

> other kinds of scientific reasoning. . . . It would, therefore, add greatly to the clarity with which the tests of significance are regarded if it were generally understood that tests of significance, when used accurately, are capable of rejecting or invalidating hypotheses, in so far as they are contradicted by the data: but that they are never capable of establishing them as certainly true. . . .

Karl Pearson had often used his chi square goodness of fit test to "prove" that data followed particular distributions. Fisher had introduced more rigor into mathematical statistics, and Karl Pearson's methods were no longer acceptable. The question still remained. It was necessary to assume that the data fit a particular distribution, in order to know which parameters to estimate and determine how those parameters relate to the scientific question at hand. The statisticians were frequently tempted to use significance tests to prove that.

In their correspondence, Egon Pearson and Jerzy Neyman explored several paradoxes that emerged from significance testing, cases where unthinking use of a significance test would reject a hypothesis that was obviously true. Fisher never fell into those paradoxes, because it would have been obvious to him that the significance tests were being applied incorrectly. Neyman asked what criteria were being used to decide when a significance test was applied correctly. Gradually, between their letters, with visits that Neyman made to England during the summers and Pearson's visits to Poland, the basic ideas of hypothesis testing emerged.[1]

A simplified version of the Neyman-Pearson formulation of hypothesis testing can now be found in all elementary statistics text-

[1]Throughout this chapter, I attribute the essential mathematical ideas to Neyman. This is because Neyman was responsible for the polished final formulation and for the careful mathematical development behind it. However, correspondence between Egon Pearson and William Sealy Gosset, which began six months before Pearson met Neyman, indicates that Pearson was already thinking about alternative hypotheses and different types of errors and that Gosset may have first suggested the idea. In spite of the fact that his was the initial input, Pearson acknowledged that Neyman provided the mathematical foundations for his own "loose ideas."

books. It has a simple structure. I have found that it is easy for most first-year students to understand. Since it has been codified, this version of the formulation is exact and didactic. This is how it must be done, the texts imply, and this is the only way it can be done. This rigid approach to hypothesis testing has been accepted by regulatory agencies like the U.S. Food and Drug Administration and the Environmental Protection Agency, and it is taught in medical schools to future medical researchers. It has also wormed its way into legal proceedings when dealing with certain types of discrimination cases.

When the Neyman-Pearson formulation is taught in this rigid, simplified version of what Neyman developed, it distorts his discoveries by concentrating on the wrong aspects of the formulation. Neyman's major discovery was that significance testing made no sense unless there were at least two possible hypotheses. That is, you could not test whether data fit a normal distribution unless there was some other distribution or set of distributions that you believed it would fit. The choice of these alternative hypotheses dictates the way in which the significance test is run. The probability of detecting that alternative, if it is true, he called the "power" of the test. In mathematics, clarity of thought is developed by giving clear, well-defined names to specific concepts. In order to distinguish between the hypothesis being used to compute Fisher's p-value and the other possible hypothesis or hypotheses, Neyman and Pearson called the hypothesis being tested the "null hypothesis" and the other hypotheses the "alternative." In their formulation, the p-value is calculated for testing the null hypothesis but the power refers to how this p-value will behave if the alternative is, in fact, true.

This led Neyman to two conclusions. One was that the power of a test was a measure of how good the test was. The more powerful of two tests was the better one to use. The second conclusion was that the set of alternatives cannot be too large. The analyst cannot say that the data come from a normal distribution (the null hypothesis) or that they come from any other possible distribution. That is too wide a set of alternatives, and no test can be powerful against all possible alternatives.

In 1956, L. J. Savage and Raj Raghu Bahadur at the University of Chicago showed that the class of alternatives does not have to be very wide for hypothesis testing to fail. They constructed a relatively small set of alternative hypotheses against which no test had any power. During the 1950s, Neyman developed the idea of restricted hypothesis tests, where the set of alternative hypotheses is very narrowly defined. He showed that such tests are more powerful than ones dealing with more inclusive sets of hypotheses.

In many situations, hypothesis tests are used against a null hypothesis that is a straw man. For instance, when two drugs are being compared in a clinical trial, the null hypothesis to be tested is that the two drugs produce the same effect. However, if that were true, then the study would never have been run. The null hypothesis that the two treatments are the same is a straw man, meant to be knocked down by the results of the study. So, following Neyman, the design of the study should be aimed at maximizing the power of the resulting data to knock down that straw man and show how the drugs differ in effect.

WHAT IS PROBABILITY?

Unfortunately, to develop a mathematical approach to hypothesis testing that was internally consistent, Neyman had to deal with a problem that Fisher had swept under the rug. This is a problem that continues to plague hypothesis testing, in spite of Neyman's neat, purely mathematical solution. It is a problem in the application of statistical methods to science in general. In its more general form, it can be summed up in the question: What is meant by probability in real life?

The mathematical formulations of statistics can be used to compute probabilities. Those probabilities enable us to apply statistical methods to scientific problems. In terms of the mathematics used, probability is well defined. How does this abstract concept connect to reality? How is the scientist to interpret the probability statements

of statistical analyses when trying to decide what is true and what is not? In the final chapter of this book I shall examine the general problem and the attempts that have been made to answer these questions. For the moment, however, we will examine the specific circumstances that forced Neyman to find his version of an answer.

Recall that Fisher's use of a significance test produced a number Fisher called the p-value. This is a calculated probability, a probability associated with the observed data under the assumption that the null hypothesis is true. For instance, suppose we wish to test a new drug for the prevention of a recurrence of breast cancer in patients who have had mastectomies, comparing it to a placebo. The null hypothesis, the straw man, is that the drug is no better than the placebo. Suppose that after five years, 50 percent of the women on placebos have had a recurrence and none of the women on the new drug have. Does this prove that the new drug "works"? The answer, of course, depends upon how many patients that 50 percent represents.

If the study included only four women in each group, that means we had eight patients, two of whom had a recurrence. Suppose we take any group of eight people, tag two of them, and divide the eight at random into two groups of four. The probability that both of the tagged people will fall into one of the groups is around .30. If there were only four women in each group, the fact that all the recurrences fell in the placebo group is not significant. If the study included 500 women in each group, it would be highly unlikely that all 250 with recurrences were on the placebo, unless the drug was working. The probability that all 250 would fall in one group if the drug was no better than the placebo is the p-value, which calculates to be less than .0001.

The p-value is a probability, and this is how it is computed. Since it is used to show that the hypothesis under which it is calculated is false, what does it really mean? It is a theoretical probability associated with the observations under conditions that are most likely false. It has nothing to do with reality. It is an indirect

measurement of plausibility. It is not the probability that we would be wrong to say the drug works. It is not the probability of any kind of error. It is not the probability that a patient will do as well on the placebo as on the drug. But, to determine which tests are better than others, Neyman had to find a way to put hypothesis testing within a framework wherein probabilities associated with the decisions made from the test could be calculated. He needed to connect the p-values of the hypothesis test to real life.

THE FREQUENTIST DEFINITION OF PROBABILITY

In 1872, John Venn, the British philosopher, had proposed a formulation of mathematical probability that would make sense in real life. He turned a major theorem of probability on its head. This is the law of large numbers, which says that if some event has a given probability (like throwing a single die and having it land with the six side up) and if we run identical trials over and over again, the proportion of times that event occurs will get closer and closer to the probability.

Venn said the probability associated with a given event is the long-run proportion of times the event occurs. In Venn's proposal, the mathematical theory of probability did not imply the law of large numbers; the law of large numbers implied probability. This is the frequentist definition of probability. In 1921, John Maynard Keynes[2] demolished this as a useful or even meaningful interpretation, showing that it has fundamental inconsistencies that make

[2]There is a kind of misonomy involved with Keynes. Most people would think of him as an economist, the founder of the Keynesian school of economics, dealing with the ways in which government manipulation of monetary policy can influence the course of the economy. However, Keynes had his Ph.D. in philosophy; and his Ph.D. dissertation, published in 1921 as *A Treatise on Probability*, is a major monument in the development of the philosophical foundations behind the use of mathematical statistics. In future chapters, we will have occasion to quote Keynes. It will be from Keynes the probabilist, and not Keynes the economist, that we will be quoting.

it impossible to apply the frequentist definition in most cases where probability is invoked.

When it came to structuring hypothesis tests in a formal mathematical way, Neyman fell back upon Venn's frequentist definition. Neyman used this to justify his interpretation of the p-value in a hypothesis test. In the Neyman-Pearson formulation, the scientist sets a fixed number, such as .05, and rejects the null hypothesis whenever the significance test p-value is less than or equal to .05. This way, in the long run, the scientist will reject a true null hypothesis exactly 5 percent of the time. The way hypothesis testing is now taught, Neyman's invocation of the frequentist approach is emphasized. It is too easy to view the Neyman-Pearson formulation of hypothesis testing as a part of the frequentist approach to probability and to ignore the more important insights that Neyman provided about the need for a well-defined set of alternative hypotheses against which to test the straw man of the null hypothesis.

Fisher misunderstood Neyman's insights. He concentrated on the definition of significance level, missing the important ideas of power and the need to define the class of alternatives. In criticism of Neyman, he wrote:

> Neyman, thinking he was correcting and improving my own early work on tests of significance, as a means to the "improvement of natural knowledge," in fact reinterpreted them in terms of that technological and commercial apparatus which is known as an acceptance procedure. Now, acceptance procedures are of great importance in the modern world. When a large concern like the Royal Navy receives material from an engineering firm it is, I suppose, subjected to sufficiently careful inspection and testing to reduce the frequency of the acceptance of faulty or defective consignments. . . . But, the logical differences between such an operation and the work of scientific discovery by physical or

> biological experimentation seem to me so wide that the analogy between them is not helpful, and the identification of the two sorts of operations is decidedly misleading.

In spite of these distortions of Neyman's basic ideas, hypothesis testing has become the most widely used statistical tool in scientific research. The exquisite mathematics of Jerzy Neyman have now become an idée fixe in many parts of science. Most scientific journals require that the authors of articles include hypothesis testing in their data analyses. It has extended beyond the scientific journals. Drug regulatory authorities in the United States, Canada, and Europe require the use of hypothesis tests in submissions. Courts of law have accepted hypothesis testing as an appropriate method of proof and allow plaintiffs to use it to show employment discrimination. It permeates all branches of statistical science.

The climb of the Neyman-Pearson formulation to the pinnacle of statistics did not go unchallenged. Fisher attacked it from its inception and continued to attack it for the rest of his life. In 1955, he published a paper entitled "Statistical Methods and Scientific Induction" in the *Journal of the Royal Statistical Society*, and he expanded on this with his last book, *Statistical Methods and Scientific Inference*. In the late 1960s, David Cox, soon to be the editor of *Biometrika*, published a trenchant analysis of how hypothesis tests are actually used in science, showing that Neyman's frequentist interpretation was inappropriate to what is actually done. In the 1980s, W. Edwards Deming attacked the entire idea of hypothesis testing as nonsensical. (We shall come back to Deming's influence on statistics in chapter 24.) Year after year, articles continue to appear in the statistical literature that find new faults with the Neyman-Pearson formulation as frozen in the textbooks.

Neyman himself took no part in the canonization of the Neyman-Pearson formulation of hypothesis testing. As early as 1935, in an article he published (in French) in the *Bulletin de la Société Mathé-*

matique de France, he raised serious doubts about whether optimum hypothesis tests could be found. In his later papers, Neyman seldom made use of hypothesis tests directly. His statistical approaches usually involved deriving probability distributions from theoretical principles and then estimating the parameters from the data.

Others picked up the ideas behind the Neyman-Pearson formulation and developed them. During World War II, Abraham Wald expanded on Neyman's use of Venn's frequentist definitions to develop the field of statistical decision theory. Eric Lehmann produced alternative criteria for good tests and then, in 1959, wrote a definitive textbook on the subject of hypothesis testing, which remains the most complete description of Neyman-Pearson hypothesis testing in the literature.

Just before Hitler invaded Poland and dropped a curtain of evil upon continental Europe, Neyman came to the United States, where he started a statistics program at the University of California at Berkeley. He remained there until his death in 1981, having created one of the most important academic statistics departments in the world. He brought to his department some of the major figures in the field. He also drew from obscurity others who went on to great achievements. For example, David Blackwell was working alone at Howard University, isolated from other mathematical statisticians. Because of his race, he had been unable to get an appointment at "White" schools, in spite of his great potential; Neyman invited Blackwell to Berkeley. Neyman also brought in a graduate student who had come from an illiterate French peasant family; Lucien Le Cam went on to become one of the world's leading probabilists.

Neyman was always attentive to his students and fellow faculty members. They describe the pleasures of the afternoon departmental teas, which Neyman presided over with a courtly graciousness. He would gently prod someone, student or faculty, to describe some recent research and then genially work his way around the

room, getting comments and aiding the discussion. He would end many teas by lifting his cup and toasting, "To the ladies!" He was especially good to "the ladies," encouraging and furthering the careers of women. Prominent among his female protégées were Dr. Elizabeth Scott, who worked with Neyman and was a coauthor on papers ranging from astronomy to carcinogenesis to zoology, and Dr. Evelyn Fix, who made major contributions to epidemiology.

Until R. A. Fisher died in 1962, Neyman was under constant attack by this acerbic genius. Everything Neyman did was grist for Fisher's criticism. If Neyman succeeded in showing a proof of some obscure Fisherian statement, Fisher attacked him for misunderstanding what he had written. If Neyman expanded on a Fisherian idea, Fisher attacked him for taking the theory down a useless path. Neyman never responded in kind, either in print or, if we are to believe those who worked with him, in private.

In an interview toward the end of his life, Neyman described a time in the 1950s when he was about to present a paper in French at an international meeting. As he went to the podium, he realized that Fisher was in the audience. While presenting the paper, he steeled himself for the attacks he knew would come. He knew that Fisher would pounce upon some unimportant minor aspect of the paper and tear it and Neyman to pieces. Neyman finished and waited for questions from the audience. A few came. But Fisher never stirred, never said a word. Later, Neyman discovered that Fisher could not speak French.