IDEAS IN CONTEXT

Edited by Wolf Lepenies, Richard Rorty, J. B. Schneewind and Quentin Skinner

The books in this series will discuss the emergence of intellectual traditions and of related new disciplines. The procedures, aims and vocabularies that were generated will be set in the context of the alternatives available within the contemporary frameworks of ideas and institutions. Through detailed studies of the evolution of such traditions, and their modification by different audiences, it is hoped that a new picture will form of the development of ideas in their concrete contexts. By this means, artificial distinctions between the history of philosophy, of the various sciences, of society and politics, and of literature may be seen to dissolve.

Titles published in the series:
Richard Rorty, J. B. Schneewind and Quentin Skinner (eds.), *Philosophy in History*
J. G. A. Pocock, *Virtue, Commerce and History*
M. M. Goldsmith, *Private Vices, Public Benefits: Bernard Mandeville's Social and Political Thought*
A. Pagden, *The Languages of Political Theory in Early-Modern Europe*
D. Summers, *The Judgment of Sense*
L. Dickey, *Hegel: Religion, Economics and the Politics of Spirit, 1770–1807*
Margo Todd, *Christian Humanism and the Puritan Social Order*
Edmund Leites (ed.), *Conscience and Casuistry in Early Modern Europe*
Lynn S. Joy, *Gassendi the Atomist: Advocate of History in an Age of Science*
Terence Ball, James Farr and Russell Hanson (eds.), *Political Innovation and Conceptual Change*
Wolf Lepenies, *Between Literature and Science: The Rise of Sociology*
Peter Novick, *That Noble Dream: The 'Objectivity Question' and the American Historical Profession*

This series is published with the support of the Exxon Education Foundation

The Empire of Chance

How probability changed science and everyday life

GERD GIGERENZER, ZENO SWIJTINK, THEODORE PORTER, LORRAINE DASTON, JOHN BEATTY, LORENZ KRÜGER

CAMBRIDGE UNIVERSITY PRESS

highly troubling to some thoughtful observers. In psychology, for example, the idea of what an experiment is was strikingly narrowed to fit the Procrustean Bed of Fisherian statistics. Alternatives, such as the Gestalt psychologists' endeavor to use experiment simply to *demonstrate* a universal effect, have virtually been ruled out of the field (see chapter 6). The experimental psychology that relies on Fisherian experimental design is often associated with an abandonment of the search for universal psychological phenomena that apply equally to all people. It concentrates on establishing general causal statements, with their implicit reference to a population, that require only an increase in the incidence of the effect when the cause is present. Since these claims do not necessarily apply to the individual, they are of foremost interest to the state and its administrators (Danziger, 1987a). Within statistics, Fisher's views on experimental design were largely accepted, but there his analysis of statistical inference was sharply criticized by advocates of alternative programs. His controversy with Jerzy Neyman and Egon Pearson, the son of Karl Pearson, is discussed in the following section.

3.4 THE CONTROVERSY: FISHER VS. NEYMAN AND PEARSON

Since the beginning of the twentieth century, several distinct views have emerged about how to draw conclusions from statistical data. The very different approaches of R. A. Fisher, of Jerzy Neyman and Egon Pearson, and of the Bayesians, all involve a considerable advance over earlier views: they are more systematic and each can account for a wider range of practices. They are the successful outcome of a tremendous intellectual effort. In that sense one can speak of a breakthrough, although there has been no declared winner. The issues that distinguish the several schools go deep into the foundations and practice of statistical inference. The different schools often disagree fiercely about basic issues, and value-laden words from ordinary speech such as "efficient," "unbiased," and "coherent," have been enlisted as names of central concepts in the various theories. By implication, rival approaches are charged with inefficiency, bias, and incoherence.

In statistical reasoning it has not (yet) proven possible to come to an all-encompassing theory, of which the current positions are special cases, appropriate if certain conditions are satisfied (but see Barnard, 1980). In fact, some working statisticians advocate an ecumenism in which one should apply different approaches to the same set of data (G. E. P. Box,

1986). In other areas there has been some convergence between different schools. For instance, Bayesians were for a long time opposed to experimental randomization, but now recognize the importance of paying attention to the procedures by which observations are collected (Rubin, 1978; Swijtink, 1982). Similarly, statisticians within the Neyman–Pearson school have recognized the importance of conditional inference (Lehmann, 1986, chapter 10).

There is a remarkable line of cleavage between the fields of applications conquered by the various schools. Methods and concepts of the two "frequency" schools, Fisher and Neyman–Pearson, have penetrated the experimental sciences, whereas Bayesians, usually designated the "subjective" school, have not. But Bayesians have made inroads in economics and have recovered traditional eighteenth-century applications of probability such as legal judgment and human rationality. We will discuss some of the recent applications of Bayes' theorem when we turn to the experimental study of thinking in chapter 6 and to applications in everyday life in chapter 7. There we will see that modern Bayesians often have fewer reservations about the range of applicability of Bayes' theorem than Bayes himself seems to have had.

In the following we will emphasize the conflicting views of the school founded by R. A. Fisher and of the Neyman–Pearson school. We will concentrate on the analysis of significance testing. Bayesian thinking will be discussed only insofar as it is relevant to understanding the conflicting viewpoints. These two are the dominant points of view, at least in the sciences discussed in this book. The form of statistical inference used in the social sciences mixes elements from these two views.

Sir Ronald A. Fisher

At Cambridge University, Fisher studied physics, mathematics, and biology. With this background, he became a leader in the inference revolution and one of the great geneticists of his time. He helped reconcile the Mendelian and the biometric approaches to the study of evolution and inheritance, not least because of his abilities as a statistician (see chapter 4). Eugenics was indeed Fisher's driving motivation, and he judged social measures according to the effects they had on the biological inheritance of man. As an undergraduate he explained the rise and fall of societies in terms of the birth rate among those whose hereditary superiority enabled them to accumulate wealth. Before he embarked on an academic career, he tried subsistence farming, since farming was, in his view, a eugenic way

of life, in which not money, but personal qualities were the dominant factor (J. F. Box, 1978). It was his investigations of inheritance that led Fisher to the concept of variance, and to the technique of the analysis of variance components to separate the contributions of different causal factors to observed correlations (J. F. Box, 1980; Fisher, 1918; see also 4.4).

In 1919, Fisher accepted the newly created post of statistician at Rothamsted Experimental Station, refusing an offer to become the chief statistician under Karl Pearson at the Galton Laboratory. Rothamsted was established in the late 1830s to investigate the effects on the soil of different combinations of bone meal, burnt bones, and various types of mineral phosphate with sulphate or muriate of ammonia. Continuous field experiments, begun in 1843, provided a wealth of agricultural field data, still largely unanalyzed when Fisher joined the station. Indeed, when one looks through the second edition of *The Book of the Rothamsted Experiments*, published in 1917, it is striking how little had been done with the data of over sixty years of experimentation (Hall, 1917). For the most part, only average yield per annum was tabulated, and inferences were based on a judgmental comparison of means. Rothamsted presented Fisher with just the right challenge. It provided daily confrontations with inferential problems that arose in the work of the station itself, and in that of visitors attracted there by Fisher's rapidly growing fame. Fisher's first book on statistical methodology, *Statistical Methods for Research Workers* (1925), was successful in introducing biologists and agriculturalists to the new techniques of statistical analysis, with nearly 20,000 copies sold during the first 25 years of the book's existence, to an increasingly international audience (Yates, 1951). His third book, *The Design of Experiments* (1935), provided a systematic account of the principles of comparative experimentation: replication, blocking, randomization, the factorial design, and confounding. It was similarly successful, and had reached a seventh edition by the time of Fisher's death in 1962.

Fisher's basic belief was that we learn from experience, although our knowledge must always remain provisional. His efforts in statistical inference were directed toward developing concepts of statistical evidence – ways to measure and express the uncertainty of hypotheses in the light of data. Fisher was not, however, satisfied with the approach based on Bayes' theorem, in which the uncertainty of a hypothesis in the light of data is expressed by a posterior probability. The use of Bayes' theorem presupposes the availability of a prior probability distribution over the possible hypotheses. Since Fisher was a frequentist, he insisted that every probability judgment must theoretically be verifiable to any chosen degree of

approximation by sampling its reference set (Fisher, 1962). Bayes' theorem can, according to Fisher, only be used in those cases where there is *a priori* distributional information about the population being sampled, that is, the cases where we know that the population from which the observations are drawn has itself been drawn at random from a superpopulation of known specification. These cases are obviously very uncommon. Fisher also held that Bayes' theorem cannot be consistently applied to other cases. For where we are ignorant and have no *a priori* distributional information, there will exist more than one way to express that ignorance probabilistically. To allow different researchers mutually inconsistent prior probabilities to express the very same state of ignorance, would lead to an unacceptable subjectivism, where strength of evidence is just a matter of taste.

Fisher's research program in statistical inference should thus be understood in the light of his highly nuanced objections to the use of Bayes' rule. The Bayesians, he thought, are wrong to assume that all uncertainties can be expressed in terms of probabilities. There are, in fact, different ways to represent uncertainty that are appropriate in different situations. In comparative experiments, when one does not have a good idea about what is going on, one can make a significance test. A significance test is a weak argument and can only suggest that a hypothetical model (the null hypothesis) is implausible in the light of the data, assuming that the experiment was performed properly. A significance test does not permit one to assign any specific degree of probability to the hypothesis. When past experience and theoretical considerations make one confident in accepting a "full parametric model," Fisher proposed other methods to calculate and represent uncertainty, such as a likelihood function of the parameters in the model given the data. Only in certain special situations, where his so-called fiducial argument applies, the uncertainty of hypotheses, Fisher believed, can be expressed in terms of probability. It is especially here that many have questioned the consistency of Fisher's adherence to frequentism, since it is not clear with respect to what reference class a particular hypothesis has a frequency of being correct.

We will deal here only with significance testing. This does not mean that the other tools caused less controversy. For instance, Richard von Mises, himself a major proponent of the frequentist point of view, agreed with Fisher's analysis of a frequentist use of Bayes' theorem. But he believed that the route taken by Bayes was the only way to express uncertainty of hypotheses in the light of data. To draw meaningful conclusions from a small number of observations without using Bayes' theorem, as

Fisher wanted, meant getting too much from nothing, and he proclaimed in 1951 that "the heyday of small sample theory . . . is already past" (von Mises, [1928] 1957, p. 159). Only large samples, he believed, could form the basis of objective inference, since here the influence of prior probability assumptions on the posterior distribution vanishes. He confessed not to understand "the many beautiful words used by Fisher and his followers in support of the likelihood theory" (von Mises, [1928] 1957, p. 158). Jerzy Neyman, still less cautiously, held: "the theory of fiducial inference is simply non-existent in the same way as, for example, a theory of numbers defined by mutually contradictory definitions" (Neyman, 1941, p. 149).

For the purposes of this exposition, the essential features of a test of significance can be summarized as follows. In a test of significance, such as the one given in section 3.2, one confronts a null hypothesis with observations to see whether the observations deviate enough from the hypothesis that one can conclude the hypothesis is implausible. There are thus three concepts here: the null hypothesis, an ordering of the possible observations as to their deviation from the null hypothesis, and a measure of how far a particular observation deviates from the null hypothesis. We will take up these three concepts in that order.

(1) The null hypothesis must allow the specification of a unique distribution function for the test statistic. For instance, in the agricultural example of section 3.2, the null hypothesis stated that each member of a pair was a random observation from the same population, in which the characteristic observed (yield of grain in bushels) has a normal distribution with unknown mean particular to the pair, and unknown variance the same for all pairs. The differences z_i are then random observations from a normal distribution with mean zero and unknown variance. The t-statistic will have a known distribution, independent of the unknown variance. It has to be emphasized here that, although one speaks here of "random observations from the same population," the population is not a real one that could in principle be sampled repeatedly. For instance, if we were to repeat the experiment on the same field, using the same design, a lack of rain might lead to a quite unrelated body of data that could not be considered as taken from the same population as the first body of data. Fisher called the population hypothetical, both since it concerns the possibly hypothetical situation that the treatment is ineffective, and since it refers to a hypothetical series of repetitions in which the same "causal matrix" is operative. We will return to the importance of this point later.

(2) The ordering of the possible observations should reflect their relative degree of deviation from the null hypothesis. But the observations

can deviate in different respects from a null hypothesis. It may be necessary to consider different orderings, reflecting different kinds of deviation, and thus different tests. For instance, if the null hypothesis is that a process behaves like independent coin-tossing, and one wants to test for independence, one may choose a test based on the number of runs, or a test based on the length of the longest run (Bradley, 1968). The choice of the test statistic, and of null hypotheses worth testing, remains, for Fisher, an art, and cannot be reduced to a mechanical process:

> It is, I believe, nothing but an illusion to think that this process can ever be reduced to a self-contained mathematical theory of tests of significance. Constructive imagination, together with much knowledge based on experience of data of the same kind, must be exercised before deciding on what hypotheses are worth testing, and in what respects. Only when this fundamental thinking has been accomplished can the problem be given a mathematical form (Fisher, 1939, p. 6).

(3) As a measure of how much a particular observation deviates from the null hypothesis, Fisher used the probability under the null hypothesis of the tail area of the test statistic s beyond its observed value s_{obs}, $p(s \geqslant s_{obs}; H_0)$. We noted earlier that this custom derived from the use of significance tests in the rejection of outliers. The original explanation Fisher gave of this was not very satisfactory, since it called something obvious that needed more motivation (Fisher, 1935, §7). Why should the discrepancy between an observation and a hypothesis depend on outcomes not actually observed? In his later book of 1956, *Statistical Methods and Scientific Inference*, Fisher conceded that this was a questionable feature of significance testing, "not very defensible save as an approximation" (Fisher, 1956, p. 66), and, indeed, it is at odds with his likelihood theory. All in all, the question remains whether the Fisherian significance level is a useful measure of the discrepancy of the data with respect to a hypothesized model. A lower significance level in the *same experiment* and with respect to the *same test statistic* will indeed indicate a greater discrepancy. So, if we had obtained, in the comparative experiment of section 3.2, observations with a significance level of, say, 1%, these observations would be more discordant with the hypothesis than the observations actually made. *Across experiments* or with *different test statistics* the situation is less clear. The question whether a significance level is a meaningful measure for discrepancy, or whether any other meaningful measure can be developed, remains unsettled (Martin-Löf, 1974; Seidenfeld, 1979; Berger and Sellke, 1987).

It should be recognized that, according to Fisher, rejecting the null

hypothesis is not equivalent to accepting the efficacy of the cause in question. The latter cannot be established on the basis of one single experiment, but requires obtaining more significant results when the experiment, or an improvement of it, is repeated at other laboratories or under other conditions. Therefore, not only significant, but also non-significant results should be published in order to let the literature correctly reflect the frequency with which a certain type of experiment has led to significant results. Already in his book *The Design of Experiments* of 1935 he wrote: "no isolated experiment, however significant in itself, can suffice for the experimental demonstration of any natural phenomenon. . . In relation to the test of significance, we may say that a phenomenon is experimentally demonstrable when we know how to conduct an experiment which will rarely fail to give us a statistically significant result" (Fisher, 1935, §7). In this passage, Fisher distinguished *significance testing* from *the demonstration of a natural phenomenon*. Careless writing on Fisher's part, combined with selective reading of his early writings has led to the identification of the two, and has encouraged the practice of demonstrating a phenomenon on the basis of a single statistically significant result. As we will show in section 3.5, this practice is part of what we will call the *hybrid theory* of statistical inference that mixes elements of Fisherian significance testing with ideas from the so-called Neyman–Pearson–Wald school of hypothesis testing.

Both in this and in his insistence that a null hypothesis can only be shown implausible, and can never be shown plausible, Fisher's *Design of Experiments* has the same message as another remarkable book published in the very same year, 1935: Karl Popper's *Logic of Scientific Discovery*. Popper gave the same characterization of the demonstration of a natural phenomenon: "[T]he scientifically significant *physical effect* may be defined as that which can be regularly reproduced by anyone who carries out the appropriate experiment in the way prescribed" (Popper, [1935] 1968, p. 45). And just as Fisher wrote: "Every experiment may be said to exist only in order to give the facts a chance of disproving the null hypothesis" (Fisher, 1935, p. 16), meaning that performing an experiment that cannot possibly disprove the null hypothesis is futile, Popper used falsifiability as a criterion of demarcation between science and non-science: "it must be possible for an empirical scientific system to be refuted by experience," for "theories are . . . *never* empirically verifiable" (Popper, [1935] 1968, pp. 40–1). Indeed, just as Popper needed a notion of "degree of corroboration" to express the degree to which unfalsified hypotheses had stood up to tests and "proved their mettle" (Popper, [1935] 1968, pp. 40–1),

Fisher also vacillated, and gave non-refutation some credit: "it is a fallacy, so well known as to be a standard example, to conclude from a test of significance that the null hypothesis is thereby established; at most it maybe said to be confirmed or strengthened" (Fisher, 1955, p. 73). Fisher never elaborated upon this remark.

The above is one reading of Fisher (Cox, 1977). But his writings are diverse, and not always transparent to even the most hermeneutic reader. For instance, Fisher sometimes formulated the null hypothesis as: "the treatment has no effect, period." One can call this the *substantive* null hypothesis. If "rejecting" this hyothesis is the same as "adopting the belief that it is false," one does indeed accept that the treatment has an effect. But, in actual fact, the hypothesis that is rejected is the *statistical* null hypothesis – that the two samples are drawn from the same distribution. The statistical null hypothesis is not equivalent to the substantive null hypothesis. For one thing, there may be systematic errors in the execution of the design that make the statistical null hypothesis false even if the treatment has no effect, or true while the treatment is effective. Efforts to replicate the result may thus still be required. Furthermore, Fisher did not mean by "rejecting or disproving a null hypothesis" a categorical adoption of the belief that it is false. As he added to the seventh edition of *The Design of Experiments*, "in learning by experience, . . . conclusions are always provisional and in the nature of progress reports, interpreting and embodying the evidence so far accrued" (Fisher, [1935] 1960, §12.1). In other words, inductive conclusions cannot be detached from their evidential basis.

These nuances, however, were often not picked up by Fisher's readers. They found, both in his earlier book of 1925, *Statistical Methods for Research Workers*, and in *The Design of Experiments* a great emphasis on tests of significance, which were relatively simple in comparison with fiducial intervals. Partly in consequence, social scientists adopted a statistical practice that did not call for assessing the size of the effects they studied, but merely claimed to have established their existence by means of a single significant experimental result (Yates, 1951).

Fisher's contributions to statistics remain controversial. A. W. F. Edwards, who has done much to develop further Fisher's likelihood theory, hailed him as "an inventive genius of the highest order," but also wrote that "the significance tests he promoted I now think ill-founded, though they work most of the time, and have contributed greatly to scientific advance" (Edwards, 1972, p. 212).

J. Neyman and Egon S. Pearson

Egon S. Pearson was the son of Karl Pearson, and worked at his father's Galton Laboratory at University College, London. Father and son disagreed actively, however, about some of the most fundamental issues in statistics. In 1925, Jerzy Neyman, a young lecturer at the University of Warsaw and at the Central College of Agriculture in the same city, arrived at the Galton Laboratory in London. Over the next couple of years, Neyman and Egon Pearson formed a personal and intellectual friendship that led to a whole new school of inferential statistics. Both agreed with Fisher's criticism of the use of Bayes' theorem. Both were also dissatisfied with Karl Pearson's work, in part because it sometimes involved Bayesian assumptions, and in part because it seemed to them too eclectic. They were impressed by Fisher's new ideas, especially by his theory of estimation and his concept of a statistical model (Fisher, 1922b). But especially Neyman, who had a continental European attitude towards mathematical rigor (he used to say, "I am a student of Lebesque"), felt that Fisher lacked a unified point of view that was strictly deduced from first principles. He and E. S. Pearson tried to provide this in what later became known as the Neyman–Pearson theory of "statistical inference as inductive behavior."

Fisher never perceived the emerging Neyman–Pearson theory as correcting and improving his own work on tests of significance. Right up to his death in 1962 he rejected the key concepts of the Neyman–Pearson theory, such as "errors of the second kind," "repeated sampling from the same population," and "inductive behavior." His recurring reproach was that Neyman and Pearson were mere mathematicians without experience in the natural sciences, and that their work reflected this insulation from all living contact with real scientific problems. Fierce disagreement was not new to statistics. For many years, Karl Pearson had declined to publish Fisher's work in *Biometrika*, which he edited, possibly since Fisher had pointed out errors in Pearson's work. Fisher never forgave Pearson this slight; he held that: "the terrible weakness of his mathematical and scientific work flowed from his incapacity in self-criticism, and unwillingness to admit the possibility that he had anything to learn from others, even in biology, of which he knew very little" (Fisher, 1956, p. 3; see also 4.4). Neyman (1967) reported that he and E. S. Pearson tried to avoid getting involved in this feud, but they could not long stay above the fray, and their debate with Fisher was marked throughout by a bitter personal tone.

Already in the late 1920s, Neyman and Pearson began to argue that Fisher had no logical basis for his choice of test statistics (such as the t-statistic discussed in 3.2), or for choosing the tail area beyond the observed value of the test statistic as the significance level of the observations. But while Fisher became more and more dissatisfied with the latter feature of significance testing, Neyman and Pearson proposed to supply this logical basis by replacing Fisher's single null hypothesis with a set of rival hypotheses. They conceived of a statistical test as providing a means to choose among such alternatives. A mechanism of choice, they held, would be reasonable if it rarely led to an error. As they understood him, Fisher had defined only one kind of statistical error: rejecting the null hypothesis when it is in fact true. This they called an "error of the first kind." Their theory implied that one should also consider another kind of error, the "error of the second kind" – that is, accepting a hypothesis when it is false. In the simplest example of Neyman–Pearson hypothesis testing, two hypotheses are given, and it is assumed that one of these is true. The purpose of making an observation is to distribute, on the basis of observation, praise and blame over these two hypotheses, viz. to reject one and accept the other. If the two hypotheses are called H_1 and H_2, a test of H_1 against H_2 is defined by a set of observations, the so-called rejection region, say R. If one makes an observation in R, one rejects H_1 and accepts H_2, and if one makes an observation outside R, one accepts H_1 and rejects H_2. The probabilities $p(R; H_1)$ and $p(R; H_2)$ are called, respectively, the size and the power of the test, that is, of the rejection region R. These are conventionally indicated by α and $1-\beta$. α and β are thus, respectively, the probability of making an error of the first kind, or type-I error, and of making an error of the second kind, or type-II error. Given this specification of statistical acceptance/rejection strategies, a rational procedure could be defined. First, identify the more important of the two hypotheses, that is, that one for which one wants to keep the error of the first kind small. Next, search for a rejection region R of the desired small size that is most powerful. In the simple case considered here – of two so-called simple hypotheses that specify probability distributions for sets of observations – a fundamental lemma by Neyman and Pearson shows that such an R always exists. (If one of the hypotheses is composite, things become more complicated.)

Neyman and Pearson were able to explain both the traditional choices of test statistics and the use of the tail area to measure significance using their idea of a rejection region of a certain size and maximum power. For it turned out that both methods were mathematically equivalent. The tail

area is for the usual choices of alternative hypotheses, nothing more than the projection of a rejection region on the real line, and can be justified on the grounds of avoiding an error of the first kind. Furthermore, it turns out that many Fisherian choices of a test statistic are equivalent to a choice of an alternative hypothesis.

But let us now examine some of the major differences between Fisher's significance test and the Neyman–Pearson theory of testing statistical hypotheses using a typical application of the latter theory, quality control in industrial manufacturing. Imagine a manufacturer who produces metal plates that are used in medical instruments. It is important that the diameter of these plates should not exceed an optimal value, say 8 millimeters, by too much, since this would cause unreliability in the medical instruments. The manufacturer considers a certain diameter, say 10 millimeters, as definitely unacceptable. Every day she takes a random sample of n plates from production in order to decide between the two hypotheses that interest her, i.e., whether the diameter is 8 millimeters (H_1) or 10 millimeters (H_2). From past experience she knows that the random fluctuation of diameters is approximately normally distributed; furthermore she knows the standard deviation of these fluctuations, which is not dependent on the mean. This allows her to determine the sampling distributions of a sample statistic, such as the mean diameter for each of the two hypotheses. Based on this statistical model and the actual mean found in the sample, the manufacturer wants to make one of two decisions (with important practical consequences): either to accept H_1 and reject H_2, i.e. to place the whole production lot on the market; or to reject H_1 and accept H_2, i.e. to stop the production and look for the cause of the apparent malfunctioning.

Each of the two decisions involves a possible error, with very different consequences. If she accepts H_2 while H_1 is true, this will cause unnecessary delays in the production process. If she accepts H_1 although H_2 is true, the defective instruments may cause harm to some patients, and the firm's reputation may suffer. Since the latter seems to her the greater danger, she decides to make this the error of the first kind, and to set its probability α at 0.1%. An error of the first kind would thus be made if she accepted that the production run was faultless (H_1), when the run was in fact flawed with the plates having a diameter of about 10 millimeters (H_2). Now she has to choose a rejection region that minimizes the error of the second kind, false alarms. Since by varying the sample size, she has control over the error of the second kind, she decides to set β at 10%, and makes the calculations of the required sample size that will

give her a test of this size and power. The actual sample is taken after this initial phase of combined personal judgment (about the validity of the statistical model and about the respective costs of the possible errors) and mathematical calculation. From here on, the procedure is quite mechanical. If the sample falls into the rejection region, H_2 is rejected and the whole production lot is placed on the market; otherwise H_2 is accepted and the production is stopped.

As Neyman emphasized, to accept a hypothesis is not to regard it as true, or to believe it. At most it means to act as if it were true. Because the manufacturer has set $\beta = 10\%$, for example, she must expect in one out of ten days to produce a false alarm – to stop the production even though it is satisfactory. She will not necessarily believe that H_1 is false, but only proceed as if it were.

In this example, it makes sense to give a behavioral interpretation to acceptance and rejection, and the relative severity of making the two kinds of error can be evaluated in terms of costs, thus providing a basis for choosing size and power. But these very features indicate that the Neyman–Pearson theory may be less suitable for scientific inference. To see this, we will return to the example of section 3.2. There we wondered, with the agricultural chemist Johnston, whether a certain proposed fertilizer would be causally effective in increasing grain yield. Here we take as one hypothesis, H_1, the null hypothesis of the Fisherian treatment. H_1 states that the treatment is ineffective, or, in statistical terms, that the z_i's are independent observations from a normal distribution with mean $\mu = 0$. To give a Neyman–Pearson treatment of that example, we need the explicit introduction of at least one other hypothesis. For this we take the hypothesis that our treatment has an (average) positive effect of 0.3 bushels of grain per treatment plot. Statistically the hypothesis H_2 states that the z_i's are independent observations from a normal distribution with positive mean $\mu = 0.3$. We want to act conservatively, and to keep the probability of rejecting H_1, when it is true, small. As α we take 0.05. Our task is now to define a rejection region R such that $p(R; H_1) = \alpha$, with maximum power of all possible rejection regions of that size. That is, $p(R; H_2) = 1-\beta$ should be made as large as possible. The fundamental lemma of Neyman and Pearson implies that such an R exists in this simple situation. Since we have nine independent observations, the rejection region is a region in a nine-dimensional space, which is hard to visualize. However, in this situation, the t-statistic projects this space on the real line in such a way that one-sided tail areas correspond to a Neyman–Pearson most powerful rejection region. If we choose R to be the set of

observations with a t-value larger than 1.86, we will have a region of approximately the right size. To calculate the power of this rejection region is complicated; it involves the so-called non-central t-distribution (Resnikoff and Lieberman, 1957). If the alternative hypothesis is that $\mu = 0.21$, the power of the test is about 0.90; if the alternative is $\mu = 0.33$, the power becomes about 0.99. If we want to opt for a smaller α, say $\alpha = 0.01$, the power of the test, when $\mu = 0.3$, becomes dangerously low: about 0.6. There would be a chance of one in three that we will fail to reject the null hypothesis, if it is in fact false – that is, if μ is about 0.3. We could alleviate this by increasing the number of plots in the experiment. Considerations of power are, thus, quite useful in the design of an experiment, and in a sense they make explicit what in Fisher's approach is called the sensitivity of an experimental design (Cox, 1958; Cohen, 1977).

However, in a scientific application we rarely want to assert the disjunction "$\mu = 0$ or $\mu = 0.21$," and the interpretation of the errors of the first and second kind loses its cogency when we cannot do that. Although this is partly a consequence of our simplistic treatment of the example as a disjunction, it remains true that in science we will often have used the wrong hypothesis, and thus need a way to measure the discordance between data and hypothesis. Similarly, it is not clear how a scientific context can provide the utility considerations that go into the choice of the size and power of a test. In the mixed case of an applied science, such as agriculture, the alternative may be defined by a break-even point, where the present costs of applying the fertilizer are just offset by the present market price of the increase in yield. But this may not be too helpful, since these are subject to sharp fluctuations. Thus even in a semi-utilitarian science such as agronomy, we need conclusions that are independent of the present market prices of fertilizers and bushels of grain.

There are three ways in which Neyman and Pearson believed they had made Fisher's theory of significance testing more complete and consistent. First is the introduction of a rival hypothesis, which allows one to look at testing as a choice between hypotheses. It has already been accepted that one of the hypotheses must be true, a procedure that traditionally has been called "induction by elimination." This makes it possible to talk about the power of a test, and to calculate the required sample size for the desired power, where Fisher had only informally talked about the sensitivity of an experimental design. The Neyman–Pearson theory thus gives a more complete framework for planning an experiment.

Second, the frequencies of the errors of the first and second kind are calculated on the basis of repeated sampling of the distributions in the original mathematical specification of the problem, and the probabilities have therefore a direct frequency interpretation (although, perhaps, still a hypothetical one: the manufacturer will never commit an error of the first kind if her production process always runs faultlessly). Recall Fisher's belief that, in scientific applications, the population of the appropriate statistical model for the analysis of experimental data cannot in any realistic sense be sampled repeatedly, and has "no objective reality, being exclusively the product of the statistician's imagination." Therefore, after the sample is in, certain features of the sample (ancillary statistics) may be used to discern a subpopulation with respect to which the more relevant probabilities can be calculated (i.e., a conditional analysis). This led Fisher to say that "the infrequency with which in particular circumstances, decisive evidence is obtained, should not be confused with the force, or cogency, of such evidence" (Fisher, 1956, p. 92), a remark that Oskar Kempthorne has called a frontal attack on the repeated sampling principle (Kempthorne, 1976).

Third, in place of what Neyman and Pearson saw as Fisher's quasi-Bayesian view that the exact level of significance somehow measures the discordancy of the data with the null hypothesis, their interpretation of statistical inference was a purely behavioristic one that refrained from any epistemic interpretation. The concepts of size and power apply to a test, whereas Fisher's significance level is a property of the sample. If "inductive inference" is inferring an evidential relation between a sample and a hypothesis that determines a certain mental attitude towards the hypothesis, as Fisher wanted to have it, inductive inference, according to Neyman, cannot exist and, therefore, science cannot depend on it. What had been thought to be inductive reasoning or inference, Neyman argued, is really better called *inductive behavior*. To accept or reject a hypothesis is "an act of will or a decision to take a particular action, perhaps to assume a particular attitude towards the various sets of hypotheses" (Neyman, 1957). Whether one calls this inference or behavior may be a matter of taste. Indeed, if inference is the assertion of sentences on the basis of assumptions, the Neyman–Pearson theory may well be looked at as a theory of inference (Hacking, 1980). Valid deductive inference is strictly truth-preserving: if the premises on which the inference is based are true, its conclusion is bound to be true. Inductive inference cannot guarantee that much. But the Neyman–Pearson theory can promise high frequency of getting it right. Suppose the decision one makes is to assert

one of the possible conclusions, here "H_1" or "H_2." If the assumption of "H_1 or H_2" is true, one will assert the true hypothesis with probability of at least the minimum of $1-\alpha$ and $1-\beta$. If both α is chosen small and β is made small, through experimental design and choice of rejection region, an inference rule with high frequency of asserting a true statement results, assuming the premises to be true. The latter qualification is important, and shows that, even if one accepts the Neyman–Pearson theory of hypothesis testing as a theory of inductive inference, significance testing still may have a role to play. For statistics may still need tests that are able to call into question the whole assumed model, the whole disjunction. If one makes an observation that is discordant with both H_1 and H_2, the Neyman–Pearson test will accept the hypothesis the data are least discordant with. In practice, we would reject the disjunction "H_1 or H_2," for it is also important to avoid an error of the third kind: giving the right answer to the wrong question, asked by the wrong model (Kimball, 1957). Its probability, however, is not well-defined.

In various publications (e.g. 1955, 1956), Fisher rejected each of these three "corrections" and "improvements." He believed Neyman had mistakenly reinterpreted his tests of significance in terms of acceptance procedures, an ideological point of view that valued expediency over truth. Indeed, Fisher likened Neyman to

> Russians [who] are made familiar with the ideal that research in pure science can and should be geared to technological performance, in the comprehensive organized effort of a five-year plan for the nation. . . [While] in the U.S. also the great importance of organized technology has I think made it easy to confuse the process appropriate for drawing correct conclusions, with those aimed rather at, let us say, speeding production, or saving money (Fisher, 1955, p. 70).

Neyman, for his part, said that some of Fisher's tests "are in a mathematically specifiable sense 'worse than useless,' " since their power is less than their size (see Hacking, 1965, p. 99). Although acceptance procedures and quality control require utility considerations such as costs of possible errors, Fisher argued, these play no role in and must not be confused with inductive inference in the sciences, which is what tests of significance are about. Fisher drew a bold line between his significance tests and the hypothesis tests of Neyman and Pearson, and ridiculed the latter as having only a very limited field of application because derived from "the phantasy of circles [i.e. mathematicians] rather remote from scientific research" (1956, p. 100). Proponents of the other camp, however, argued that there was no difference in fields of application, because they had simply made Fisher's theory more consistent. For instance, in a paper presented to a conference on the question "For what use are tests of hypotheses and tests of significance?" Neyman wrote: "The title of the present session involves an element that appears mysterious to me. This element is the apparent distinction between tests of statistical hypotheses, on the one hand, and tests of significance, on the other. If this is not a lapse of someone's pen, then I hope to learn the conceptual distinction" (Neyman, 1976b).

Because of Fisher's remarkable talent for polemic, the debate never lacked for overblown rhetoric. He branded Neyman's position as "childish" and "horrifying [for] intellectual freedom in the west." Both parties called up the heroes of the past, such as Laplace and Gauss, to be their witnesses. The authority of W. S. Gosset was claimed by both camps. Fisher described him as a man "actively concerned with research in the natural sciences" (Gosset worked for Guinness, the brewers), and claimed that Gosset used his test in the same spirit as had Fisher. In answer, Pearson published a letter from Gosset stating that a test "doesn't itself necessarily prove that the sample is not drawn randomly from the population even if the chance is very small, say .00001: what it does is to show that if there is any alternative hypothesis which will explain the occurrence of the sample with a reasonable probability, say .05, . . . you will be very much more inclined to consider that the original hypothesis is not true" (E. S. Pearson, 1938, p. 243). Here Gosset made a strong point for rival hypotheses, although not for cost-benefit considerations. As Hacking (1965, p. 83) put it, the man who first conceived one of the great tests was now urging "that it is not merely low likelihood which matters, but rather the ratio of the likelihoods."

These, then, were vigorous controversies, and they have not ended. Disputes no less heated have characterized the relationship between Bayesians and frequentists. So it is especially remarkable that all of these unresolved controversial issues, conceptual ambiguities, and personal insults have been more or less completely suppressed from the textbooks that have taught significance testing to the customer – the experimenter in the sciences. The need for personal judgment – for Fisher in the choice of model and test statistic; for Neyman and Pearson in the choice of a class of hypotheses and a rejection region; for the Bayesians in the choice of a prior probability – as well as the existence of alternative statistical conceptions, were ignored by most textbooks. As a consequence, scientific

researchers in many fields learned to apply statistical tests in a quasi-mechanical way, without giving adequate attention to what questions these numerical procedures really answer.

3.5 HYBRIDIZATION: THE SILENT SOLUTION

The intellectual effort of statisticians to provide a mathematical foundation for hypothesis testing has had a tremendous impact on the sciences, especially on biology and the social sciences. In sociology and psychology, significance testing has become practically the only statistical tool, and other developments such as confidence intervals, the likelihood function, or Bayesian inference have been for the most part ignored by experimenters. In part, this is probably due to the stress Fisher put on significance testing in the first edition of his 1925 book, and to the theory of experimental design he provided, together with significance testing, in his 1935 *Design of Experiments*. Although the debate continues among statisticians, it was silently resolved in the "cookbooks" written in the 1940s to the 1960s, largely by non-statisticians, to teach students in the social sciences the "rules of statistics." Fisher's theory of significance testing, which was historically first, was merged with concepts from the Neyman–Pearson theory and taught as "statistics" *per se*. We call this compromise the "hybrid theory" of statistical inference, and it goes without saying that neither Fisher nor Neyman and Pearson would have looked with favor on this offspring of their forced marriage.

The creation of the hybrid can be understood on three levels – mathematical statisticians, textbook writers, and experimenters. On the first level, there was a tendency to resolve the controversial issues separating the three major schools by distinguishing between theory and application, and by saying that practical-minded people need not be bothered by these mainly theoretical issues (noted in Hogben, 1957). To users of statistics, this seemed perfectly acceptable, since often the same formulae were used and the same numerical results obtained. The great differences in conceptual interpretation were overlooked in the plug-in-and-crank-through use of statistical rules.

But, on the second level, writers of textbooks for education, psychology, sociology, and so on, commenced peace negotiations and created a hybrid theory, to which shelves and shelves in research libraries now pay tribute. The hybrid theory combines concepts from the Fisherian and the Neyman–Pearson framework. It is presented anonymously as statistical method, while unresolved controversial issues and alternative approaches

to scientific inference are completely ignored. Key concepts from the Neyman–Pearson theory such as power are introduced along with Fisher's significance testing, without mentioning that both parties viewed these ideas as irreconcilable. For instance, checking (without random sampling) thirty books on statistics for psychology, education, and sociology that were readily available, we found that the names of Neyman and E. S. Pearson were not even mentioned in twenty-five of them, although some of their ideas were presented. None even hinted at the existence of controversy, much less spelled out the issues in dispute. The crucial concepts were not identified with their creators – which is very unusual in fields like psychology, where textbooks list competing theories and the researchers who proposed them for almost every phenomenon discussed. Statistics is treated as abstract truth, the monolithic logic of inductive inference.

The hybrid theory comes with a list of prescriptions that are held to constitute what is "scientific" and "objective." The researcher must specify the level of significance before conducting the experiment (following Neyman and Pearson rather than Fisher); he must not draw conclusions from a non-significant result (following Fisher's writings, but not Neyman–Pearson); and so on. Neyman's behavioristic interpretation did not become part of the hybrid, and the type-I and type-II errors are given an epistemic interpretation. This has led to an enormous confusion about the meaning of a significance level. For instance, in practice (contrary to prescript), experimenters often will note, when inspecting the data, at what most stringent conventional level the data are significant with respect to the null hypothesis. They then report that the null hypothesis is, say, "rejected at the 0.01 level," an expression that occurs neither in Fisher nor in the writings of Neyman and Pearson.

The hybrid theory was institutionalized by editors of major journals and in the university curricula, in what has been called the "inference revolution" (Gigerenzer and Murray, 1987). By the mid-1950s, the use of significance tests and of conventional rejection levels was well established in sociological research, and researchers not using significance tests felt the pressure to explain and defend their deviant behavior (Morrison and Henkel, 1970). In the hands of the experimenters and editors, the hybrid theory often degenerated into a mechanical ritual, although Fisher, Gosset, Neyman, and Pearson had all warned against drawing inferences from tests without judgment. In some fields, a strikingly narrow understanding of statistical significance made a significant result seem to be the ultimate purpose of research, and non-significance the sign of a badly

conducted experiment – hence with almost no chance of publication (see 6.3). This practice of neglecting non-significant results may have been derived directly from an ill-considered passage in Fisher's *The Design of Experiments*: "It is usual and convenient for experimenters to take the 5 per cent. as a standard level of significance, in the sense that they are prepared to ignore all results which fail to reach this standard, and, by this means, to eliminate from further discussion the greater part of the fluctuations which chance causes have thrown into their experimental results" (1935, §7). Here Fisher seems clearly to sanction the practice of taking no notice of statistically insignificant results. But this reading is in direct contradiction to his analysis of the relation between significance testing and the demonstration of a natural phenomenon, presented in the very same section. Since he thought that experimental demonstration of, say, an interesting psychological phenomenon required confirmation by similar experiments in other laboratories, it would follow that non-significant results should also be published, so that the literature will correctly reflect the frequency with which a type of experiment has led to significant results. This implication of Fisher's writings was not heeded in the social sciences. Negative results submitted for publication are often rejected with a facile declaration that "the sensitivity of the experiment [is] substandard for the type of investigation in question" (Melton, 1962, p. 554).

As an apparently non-controversial body of statistical knowledge, the hybrid theory has survived all attacks since its inception in the 1940s. If only for practical reasons, it has easily defeated ecumenism (Box, 1986), in which one applies the different approaches to the same data, acknowledging that the different approaches are conceptually unlike. It has survived attacks from proponents of the Neyman–Pearson school, and the Bayesians (Edwards, Lindman, and Savage, 1963), and Popperians (Meehl, 1978). Its dominance permits the suppression of the hard questions. What, if any, is the relation between statistical significance and substantial importance within the scientific discipline? To what aspects of the scientific enterprise do the ideas of Fisher, and of Neyman and Pearson, appeal, and how can these be combined? Are the experimental designs developed by statisticians in agriculture and biology really a good model for all experimentation in the social sciences?

What is most remarkable is the confidence within each social-science discipline that the standards of scientific demonstration have now been objectively and universally defined. In fact, the standardization of statistical methods becomes much less complete if one looks across disciplines. In

econometrics, to take the most striking contrast, experiment is comparatively rare, and the standard statistical tool is regression analysis. It has often been applied by economists with a lack of imagination that matches the psychologists' use of hypothesis testing (McCloskey, 1985). Graduate students within the social and biological sciences have routinely been taught to view their statistical tools as canonical, given by logic and mathematics. The methods of statistical inference could be seen by practitioners uncomfortable with higher mathematics as someone else's concern, the province of statistical specialists.

3.6 THE STATISTICAL PROFESSION: INTELLECTUAL AUTONOMY

Statistical inference, and the accompanying mathematics, have become the basis for an expertise that extends to an enormous range of disciplines and practical problems and that supports a whole profession of statisticians. The abstracting journal *Statistical Theory and Method Abstracts*, a publication of the International Statistical Institute (ISI), lists over 130 journals mainly devoted to statistics, from the Indian *Aligarh Journal of Statistics* to the West-German *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete*. Beside the ISI (founded 1885), there are now many other international statistical organizations, such as the Bernoulli Society for Mathematical Statistics and Probability. Many industrialized countries have their own organization for theoretical statistics, one of the oldest being the Royal Statistical Society, founded in 1834 as the Statistical Society of London. Universities now often have a separate department for statistics, or even for biostatistics. Inferential statisticians work in many other departments, such as psychology, economics, and archeology. Statisticians are consultants in science, industry, and government. More and more we find statisticians acting as expert witnesses in the courts, and it is debated whether the way evidence should be combined in the courts can be modeled using ideas from probability theory (see 7.4; also Eggleston, 1978; DeGroot *et al.*, 1986).

This professionalization of statistics and of statisticians has several aspects. Two of them we call autonomy and influence. In this section we ask how statistics became a discipline *per se*, where it earlier had been an appendage of other disciplines like sociology, or biology. That is, how did statistics become autonomous? In section 3.7, we consider the institutions of statistics, and ask how statisticians were able to reach out to, and affect, so many other disciplines and other social institutions.

Whence this tremendous influence? What were its channels and what needs did it satisfy? How did it change the disciplines involved and how was it changed itself in the process? Without trying to answer these questions in depth, we will indicate some of what we think are the key issues.

Specialized knowledge

A scientific discipline is characterized by a body of specialized knowledge and skills, and by a complex of institutions, formal and informal, that guide its development and workings. The specialized knowledge of statisticians consists in methods to determine how data should be gathered, to analyze and summarize data, to make inferences on the basis of data, and to propose decisions on the basis of theories, data, and goals. Their skills include the tacit knowledge needed in the application of this knowledge, since it often involves a degree of subjective judgment. Those skills are developed by working as a consultant for fellow scientists and for clients outside the academic environment, in government, industry, and the like.

The methods the statistician uses are mathematical and abstract. They do not result from one single idea, but form a network of interrelated ideas. Most, but not all, use the concept of *probability*. Many, but not even most, use the concept of a *statistical model*. Still, these two concepts, probability and model, are central to the network of knowledge of the statistician, since even the methods that do not use them, such as exploratory data analysis (Tukey, 1977a) or distribution-free statistical tests (Bradley, 1968), are often partly characterized by the very fact that they do not use them! The two concepts are highly abstract, and because they are so abstract they can be applied to, and recognized in, many different situations. The historical events through which these two abstract concepts were defined were seminal for the development of statistics as a scientific discipline.

The concept of probability was defined in the early 1930s by the Soviet mathematician A. N. Kolmogorov, who further developed an axiomatization of comparative probability due to S. N. Bernstein by incorporating ideas from set theory and the theory of functions (Maistrov, [1964] 1974). Of course, long before Kolmogorov, people had referred to and used the concept of probability. This whole book is a testimony to that. But their use was often tied to what in hindsight appears to be only a limited application. And their calculations sometimes seem incoherent to us when they implicitly assume probability to have properties we do not attribute to it (Shafer, 1978). In his 1933 paper "Grundbegriffe der

Wahrscheinlichkeitsrechnung," Kolmogorov laid down axiomatically what properties the concept of probability should have. Probability is defined as a set-function. It assigns to each set in a "field of sets" its "probability," a real number between zero and one. If two sets have no elements in common, the probability of their union is equal to the sum of their probabilities. The probability of a certain basic set, the set E of all elementary events, is equal to one. All other sets in the field are subsets of E. The notion of a random variable is defined as a function from the set E into the real numbers; prior to Kolmogorov this was taken to be a primitive notion, not defined in terms of other more basic notions. With these definitions, Kolmogorov showed that there are striking analogies between the notion of the measure of a set and the probability of an event, between the integral and mathematical expectation, and between orthogonality of functions and the independence of random variables. In this way he was able to systematize on the basis of first principles many results on the law of large numbers obtained by Khinchin, Borel, Cantelli, and Hausdorff. The work of the Russian school of probabilists, including Chebyshev, Markov, Lyapunov, Bernstein, Khinchin, and Kolmogorov, reestablished probabilty theory as a serious mathematical discipline, which it had not been since new standards of rigor in mathematics were introduced early in the nineteenth century by Cauchy and others (see Schneider, 1987). Its influence reached far beyond the borders of the Soviet Union: first over Central Europe and then, through the diaspora of the 1930s, all over the world. Many of the now retired probabilists in the United States have their roots in this Central European tradition (Gani, 1982).

The concept of a statistical model was introduced by Ronald A. Fisher in 1922, in his fundamental paper "On the Mathematical Foundations of Theoretical Statistics." Fisher did not formally define the concept; he was not interested in abstract mathematics, but used an intuitive, conceptual approach. In fact, it was a citizen of one of the smaller European countries, the Swede Harald Cramér, who, between 1930 and 1950, mediated between the British and American science of statistics – which was based mainly in the empirical and experimental tradition – and the mathematical rigor of the Continental European work in probability theory. He tried to unify the two traditions in his 1946 book *Mathematical Methods of Statistics*. This unification has never been complete, and one still finds probability theory typically worked on in departments of mathematics, while statisticians, who apply results in probability theory to statistical inference, have their own organizational units. Characteristic of this is the

split, in 1973, of the *Annals in Mathematical Statistics*, into the *Annals of Statistics* and the *Annals of Probability Theory*.

Still, since Fisher consciously pursued conceptual clarifications in his 1922 paper, we can look at it as introducing the abstract concept of a statistical model. In his paper, Fisher emphasized the distinction between a sample and the population from which the sample is drawn. The population may be an actually existing one, as in a survey of the farming community of a country, but more generally it may be a hypothetical and even infinite population, as in the set of all possible coin tosses with a certain coin (where it is assumed that the same type of toss is being performed and that the coin does not wear out), or the set of all possible repetitions of a comparative agricultural experiment (where it is assumed that the same experimental procedure is followed and that the soil does not become impoverished). The tosses actually performed and the trials actually made are then considered as a random sample from this conceptual population. The distinction between population and sample had not been sufficiently heeded by his predecessors, Fisher felt. For instance, they often talked about a *mean* indiscriminately, as the average of a sample or as the average of the population from which the sample is drawn. Fisher used small Greek letters, like σ, for the characteristics of the population, called parameters, and small Roman letters like s for characteristics of the sample $(x_1, \ldots, x_n)$. These characteristics he then called *statistics*. The parameters may be partially unknown and the sample can give information about these unknown parameters. It is the task of statistical inference, Fisher stated, to find summaries of the data, that is to obtain statistics of the data, that contain as much as possible of the relevant information the data provide about the population and its parameter values. A statistic that contains the same information as the full data about the population Fisher called a *sufficient statistic*.

In abstract terms, a *parametric statistical model* M consists of a specification of an observable variable X, a parameter Θ, and for each value of Θ a probability function $p(x; \theta)$, that gives the probability of making the observation $X = x$ when $\Theta = \theta$ (Dawid, 1983). Fisher's idea is that an observation x is informative about Θ when $p(x; \theta)$ is not the same for all values θ of Θ. Suppose it is observed that $X = x$ and that $p(x; \theta_1)$ is larger than $p(x; \theta_2)$; then it is said that the likelihood of θ_1 is larger than the likelihood of θ_2, and that, on the basis of the observation that $X = x$ alone, θ_1 is more likely than θ_2.

A statistical model is a very abstract and flexible concept of wide applicability. For instance, the nineteenth-century problem of measurement and measuring error can be conceptualized in terms of a statistical model. A measurement procedure and the object to be measured determine a possible measurement result x. Assuming there is no constant error in the measurements, the precision σ of the procedure determines how close the measurements are grouped around the true value μ to be measured. Assuming that there is a lot of experience with the procedure, σ may be known (say 1), and the only unknown parameter is μ. In the traditional error theory $p(x; \mu)$ is then often taken to be a normal distribution with variance 1 and mean μ. If we take twenty measurements, X is the vector $(X_1, \ldots, X_{20})$ and $p(x; \mu)$ is the product $p(x_1; \mu) \ldots p(x_{20}; \mu)$, since in error theory repeated measurements are considered to be independent, and, ideally, experimental safeguards are used to guarantee this. The average of the sample $\bar{x} = \Sigma x_i / n$ is a sufficient statistic, in the sense that $p(x \mid \bar{x}; \mu) = p(x; \mu)/p(\bar{x}; \mu)$ is the same for different values of μ, and thus contains no information about μ. That is, given we know the value of the average of the measurements, to know more about the individual measurements will not give more information about μ. A scientist is therefore justified, assuming that the precision of his procedures is known, in communicating to his colleagues only the average value of his measurements plus the number of individual measurements taken (the size of the sample). The average is at the same time the maximum likelihood estimate of μ: if we set μ_0 equal to $\bar{x}$ then $p(\bar{x}; \mu_0)$ is maximized; assuming any other value for μ will assign a smaller probability to the observation x actually made. It is assumed throughout that M is a "correct" statistical model, in the sense that it contains the true probability distribution for the observed quantities X_i.

Karl Pearson had sought to move statistics away from the tendency to assume that data are distributed according to the normal curve, and to this end he defined a whole family of curves, of which the normal was only a special case. He generally fit these curves to observational, not experimental, data. In determining their "frequency constants," or "quaesita," as the parameters were often called before Fisher, he did not aim to identify entities with causal powers, but merely to summarize the observations. The biometricians, therefore, provided a biological theory without causes, completely in line with Pearson's *Grammar of Science*. Fisher's parametric statistical models, in contrast, were closely tied to experiment. His parameters gave estimates of the causal power of a fertilizer or drug under test. The parametric families of distributions Fisher used in his statistical models were usually simpler than Pearson's, and one may argue that the success of Fisher's new concepts, like sufficient statistics, was

bought by limiting himself to these more restricted families of distributions (Stigler, 1976). Ironically, Fisher's work was in this respect a return to nineteenth-century ideas, since he often assumed that the observations, or some known function of the observations, were normally distributed, thus contributing to the "myth of normality." Recently, concern about how good statistical methods are under mild deviations from normality or other classical distribution functions has led to a study of the "robustness" of these methods (Huber, 1981). Similar concerns have fueled interest in so-called distribution-free tests (Bradley, 1968).

Still, Fisher's idea of a parametric statistical model is a powerful and unifying concept, and is not restricted to the special kind of models he studied himself and for which his concepts of statistical inference, such as the sufficient statistic, seem so appropriate. In fact one may say that the load carried in the Bayesian approach by the prior probability distribution, is borne by the model in this part of Fisher's analysis of statistical inference (Hotelling, 1951).

A particularly important kind of statistical model is a so-called stochastic process, which describes a system that changes over time according to probabilistic laws. The error-theory models of the nineteenth century assumed independence of the successive measurements (Lancaster, 1972), and even the correlational studies, in which measurements like the height of fathers and the height of sons were obviously not independent, did not have the dynamic character that one now associates with time series and stochastic processes. Interestingly, the probabilistic theory of stochastic processes is not an outgrowth of the empirical study of random phenomena, but of the Russian theoretical studies in mathematical probability theory that culminated in Kolmogorov's axiomatization (see, however, the discussion of stochasticity in physics in 5.7).

According to Bernoulli's theorem (see 1.7), frequencies of independent chance events must converge to the underlying probabilities. When, near the end of the nineteenth century, an explicit interest in the notion of dependent trials arose, it was still thought by many that this so-called "law of large numbers" – the term is Poisson's – is only true for independent trials. In the context of a general investigation by the Russian school of necessary and sufficient conditions for laws of large numbers, A. A. Markov showed in 1906 that the convergence holds even under conditions of weak dependency (Markov, 1906). What is now called a Markov chain is a mathematical model of a process without after-effects, which describes a physical system in which the probability of transition to another state depends only on the state of the system at the given time and not on the previous history of the process. Questions about dependence and the importance of the work of Lexis and Karl Pearson were further raised in a correspondence between the probabilist Markov and the statistician Chuprov, which became a starting point of a general theory of stochastic processes (Ondar, 1981). The Swedes, especially Harald Cramér and Herman Wold, were again central in combining the mathematical work of the Russians with the more observational approach of Karl Pearson and G. Udny Yule (Bartlett, 1959).

Stochastic processes are now a modeling tool for a wide variety of scientific disciplines: econometrics, meteorology, oceanography, sociology, epidemiology, plant and animal ecology, chemistry, physics, architecture, and cosmology, to mention just a few (Gani, 1986). For the social sciences they deliver, in part, what Adolphe Quetelet expected from "social physics" (Quetelet, 1869; see also 2.2), by making it possible to explain social trends and to make short-range predictions (Bartholomew, 1967). In the physical sciences they provide a way to show the lawlikeness of some natural phenomena that elude the more classical approaches, such as the regular shape of ripples on a beach (Barndorff-Nielsen, 1985).

3.7 THE STATISTICAL PROFESSION: INSTITUTIONS AND INFLUENCE

The discipline of inferential statistics is characterized not only by a coherent network of specialized knowledge, but also by a complex of institutions, formal and informal, that guide its development and workings. These include international and national professional organizations and the sections of universities where statisticians work.

The central international organization is the International Statistical Institute, which celebrated its centenary jubilee in 1985 (Atkinson and Fienberg, 1985). It is instructive to compare the ISI when it was founded in 1885 with what it is today. At its foundation the ISI was intended to be a continuation of the International Statistical Congresses, the first of which was organized, under the leadership of Quetelet, by the Central Statistical Commission of Belgium, and held in Brussels in 1853 (Neumann-Spallart, 1885). These congresses aimed to formulate uniform methods of classification and collection to promote international comparability of statistical data. Their members were mostly directors of official statistical bureaus. Some of the scientific members hoped, with Quetelet, that the amassing of careful and comparable statistical data would bring into the open statistical laws and regularities for a future

"social physics." Thus the congresses had also given attention to methods of statistical data analysis and data representation. However, the congresses slowly lost their initial zeal and became a victim of their double goal: to be a meeting place for government officials with the power of binding resolutions (which also exposed them to political turmoil, as in the Franco-Prussian war), and to provide an opportunity for private individuals to exchange ideas and arguments of a moral or scientific nature. The ISI was therefore proposed as a purely free association analogous to the *Institut de France*, where members were selected on the basis of their personal qualifications, but with the same goal as the congresses: to introduce uniformity in the compilation of statistical data and to promote and foster the knowledge of statistical science.

The proceedings of the ISI provide interesting source material for a comprehensive history of statistics, since on its pages we see the clash in styles and interest of statisticians from many different local traditions. One such debate concerned the very possibility of using samples to get knowledge about a population, such as the farming community in Bulgaria. In the nineteenth century, statisticians who collected data had relied more and more on what Georg von Mayr called the "erschöpfende Beobachtung der primären sozialen Masse," that is, on complete investigation of the population under study (Mayr, 1895). "Partial investigations" were considered imprecise and unscientific. It was again a statistician from one of the smaller European countries, Anders Kiaer, director of the Central Bureau of Statistics of Norway, who pressed for using samples, or what he called "the representative method" (Kiaer, 1898). One of his arguments was that the quality of the data in a sample would often be much better than if the whole population had been investigated, since more care by better trained interviewers could be exerted. Kiaer's "representative method" did not, however, make use of random sampling; it was a systematic search for a sample that agreed in important characteristics with the population at large. These characteristics had to be learned from a complete investigation, a census. Random sampling was only understood in the beginning of this century (Jensen, 1926). An early paper by Jerzy Neyman still battled against Kiaer's version of sampling and was important in getting the general idea of randomness in sampling accepted (Neyman, 1934).

In the interwar period, organizations appeared that competed with the ISI, either in its aim to collect statistics and to set standards for data gathering or in its scientific goals. Among them were the League of

Nations, the International Labor Organization, the International Institute of Agriculture, the Econometric Society, and the International Union for the Scientific Investigation of Population Problems (Zahn, 1934). But it was the Second World War that marked a sharp break in the history of the ISI, and led to fundamental changes in the organization, constitution, and aims of the Institute (Nixon, 1960). Numerous international agencies in the context of the United Nations took over the administrative functions of the ISI. The ISI was more narrowly defined as an "international statistical academy" – a voluntary and scientific rather than an official organization; a community of statistical experts who were to be judged exclusively on their professional merit, and not on what country or organization they represented (Rice, 1947). Its activity shifted towards theory and methodology. It thus became the international agency of professionalized mathematical statistics, and shed all association with semi-governmental activities. Its active members are now mostly university professors. Of the forty-one contributors to the *Centenary Volume* (Atkinson and Fienberg, 1985), thirty-three hold university positions. The ISI recently approved a "Declaration of Professional Ethics" (International Statistical Institute, 1986).

Departments of statistics as we now know them are successors to the so-called "statistical laboratory." The earliest of these influential laboratories was the Galton Laboratory, endowed by Francis Galton in 1904, whose first director was Karl Pearson. This laboratory took in advanced students from science and industry to learn statistical methods that could be applied to the problems of their own field. A few of these students, such as W. S. Gosset, a chemist by training, became important pioneers in the mathematics of statistics. Most of them contributed mainly by mastering existing techniques and applying them to new problems. Some of their work was published in *Biometrika*, the journal founded by Francis Galton, Karl Pearson, and the zoologist W. F. R. Weldon in 1901 to collect biological data of a statistical kind and to spread the statistical methods and perspective that would promote a biology based on the study of variation, as opposed to morphological understanding in terms of ideal types. Pearson's laboratory institutionalized the new intellectual structure of statistics, in which statistics was first of all a body of mathematical tools and formulations which could be applied to an almost unlimited domain of topics. The laboratory slowly began to attract an international group of postgraduates. In 1925, when Neyman came from Warsaw to study with Karl Pearson, only one of the eight students was an

Englishman. The others were all from the U.S., with the exception of a Japanese. The next year brought students from Spain, Canada, China, India, and Yugoslavia. By that time Pearson had managed to incorporate the laboratory into a Department of Applied Statistics.

Fisher's presence in Rothamsted, and then University College, London, played a similar fertilizing role. Harold Hotelling, who exerted a great impact on statistics and the way it was taught in the USA before the Second World War, was a volunteer on the farm during the academic year 1929–30 (Hotelling, 1940). In the period 1934–44, when Fisher was Galton Professor of Eugenics at University College in London (having succeeded Karl Pearson) more than fifty people from all over the world, and from a variety of experimental disciplines – chemistry, biology, medicine, agriculture, and social science – came to work with him (Youden, 1951).

The influence of Fisher's first book, *Statistical Methods for Research Workers*, was tremendous (Yates, 1951). It went through eleven editions in the first twenty-five years of its existence, with nearly 20,000 copies sold, and was translated into French, Italian, Spanish, German, and Japanese. In that period, analysis of variance found applications in agricultural trials, biological assays, industrial experimentation, quality control, and many experimental scientific fields.

One of the broadest channels for the flow of information from Europe to the United States was the Statistics Laboratory of Iowa State College, the first of the great academic statistical centers in the United States. George W. Snedecor, the Director of the Laboratory, arranged for Fisher to lecture at two summer sessions in 1931 and 1936. Snedecor himself, in his teaching and in his well-known book *Statistical Methods*, made Fisher's methods available to a host of workers in agronomy and animal husbandry (Youden, 1951). Other centers for the new enthusiasm in statistics were the University of North Carolina at Chapel Hill, the University of Michigan at Ann Arbor, and Columbia University in New York.

Perhaps the most direct influence from Europe on the development of statistics in the United States was Neyman's acceptance, in 1938, of a faculty position at the University of California at Berkeley, where he remained the rest of his life. First within the Department of Mathematics, and later in a separate Department of Statistics, Neyman copied Pearson's statistical laboratory as he had seen it in London. From this base, Neyman collaborated fruitfully with a wide variety of scientists, including astronomers, biologists, and meteorologists.

Statistics goes to war

Neyman's arrival in the United States was, by chance, well timed. His philosophy of statistical inference, brought forward with particular explicitness in the expression he coined in 1938, "inductive behavior" (in contrast to "find reductions of data to communicate to fellow research workers," as statistical inference was understood by Fisher), fit well the mood of the time and the requirements dictated by the approaching war. "The only useful function of a statistician is to make predictions, and thus to provide a basis for action," wrote W. E. Deming of the War Department in 1942 (Wallis, 1980). And earlier the War Preparedness Committee of the Institute of Mathematical Statistics, an offspring of the less mathematical American Statistical Association, had stressed that statisticians were not just good in calculating averages and index numbers, but could also contribute to the National Defense Program in such areas as quality control, sample surveys, experimentation, personnel selection, gunnery and bombing, and weather forecasting (Eisenhart *et al.*, 1940).

In fact it was through their activities in the Second World War that statisticians were able to influence so many other disciplines and social institutions. Both the positive reception of statistics – by engineering and the social sciences, by industry and the military – and its departmental autonomy were furthered by the war effort (Fienberg, 1985; Barnard and Plackett, 1985). It led to new developments along the lines of Neyman's doctrine of inductive behavior, which was made more prudent in the theory of sequential analysis and more mathematical in the theory of statistical decision functions, both developed by Abraham Wald and published after the war (Wald, 1947, 1950). The theory of statistical decision functions was given a subjective twist in the personalistic or Bayesian decision theory of L. J. Savage (Savage, 1954).

In the United States during the Second World War, there were several major groups of statisticians working under contract of one of the branches of the armed forces, notably at Columbia (the Statistical Research Group, or S.R.G., under W. A. Wallis), at Princeton (under S. S. Wilks), and at Berkeley (under Neyman). Sequential analysis originated in the S.R.G. at Columbia through a suggestion of a Navy officer, an ordnance expert at the U.S. Naval Proving Ground in Dahlgren, Virginia. He argued that one could see after the first so many rounds that the experiment need not be completed, either because the new method was obviously superior or obviously inferior. This idea was picked up by Wallis and the economist Milton Friedman and transformed by Abraham

Wald into a full-blooded modification of the Neyman–Pearson theory, the theory of sequential analysis. Wald was a Romanian Jew who had worked for a short time at the Austrian Institute for Business Cycle Research, where Oskar Morgenstern was then director. He had been able to flee Austria after the *Anschluss* with Germany in 1938, and found refuge in the United States with the financial support of the Carnegie Foundation. Sequential analysis is the theory of sequences of Neyman–Pearson decisions in which, instead of choosing between two decisions (reject or accept), the statistician also has the option of deciding to make more observations. This means that the sample size is not fixed in advance. It turns out that sequential testing can lead to more powerful tests of the same size, in the Neyman–Pearson sense. Wald's theory of statistical decision functions interprets statistical problems as decision problems in a "game against nature." Using the theory of games and its formal utility theory developed by John von Neumann and Oskar Morgenstern, Wald was able to take into account the losses that one would incur when making a wrong decision as an error of the first kind.

It is clear that the Second World War influenced the specifics of statistical research. A large number of applied research projects that fit into the war effort got funded. After the war the same military funding sources, such as the Office of Navy Research, kept financing the basic research of the scientists and their students whom they had funded during the war (Old, 1961). Our claim here, however, is that it also strengthened the decision theoretic approach to statistical problems, and accustomed people to the idea that reasonable decisions can be made on the basis of formal, mechanized reasoning combined with measurements. A prime example of this is the role of statistics in psychology. Before 1940, psychologists used largely informal and unstandardized methods to assess their experimental and observational results (Gigerenzer and Murray, 1987). Through the impact of such war studies as Stouffer *et al.* (1949, 1950) on the social psychology of American soldiers during the Second World War, statistical techniques of the accept/reject variety strengthened their hold on the psychologists, a trend that was reinforced by the needs of educational administrators for an "objective" technique to guide them in their bureaucratic decisions on curriculum innovation (Danziger, 1987a).

3.8 CONCLUSION

The agricultural chemist Johnston found himself confronted with a problem concerning experimental design and the inference from observational data to causal hypotheses. These problems were not pursued by trying to analyze the laws that govern the subject matter under investigation – in Johnston's case, the physiology and chemistry of plants. Because of the variation displayed by natural objects in their natural environment, questions like this one have come to be studied in a quite different way. In fact, this agricultural example is typical of problems in a variety of disciplines that scientists are now accustomed to investigating according to a canon of research defined by the new abstract discipline of scientific inference, mathematical statistics. Bits and pieces of this discipline emerged here and there; they have accompanied the entire history of probability. Yet this combination into a single unified conceptual structure and methodological doctrine came late and is far from being completed. Alternative schools of inferential statistics compete, and doubtful compromises dominate in the practice of social scientists (see chapter 6). Nevertheless, impressive conceptual and institutional achievements, together with economic needs and political interests, have shaped a new profession of inference experts.

In the upbeat tone of his 1953 presidential address to the Royal Statistical Society, which provided the epigraph to this chapter, R. A. Fisher speculated that "hidden causes have been at work for much longer than the period of manifest efflorescence, preparing men's minds, and shaping the institutions through which they work, so that, quite suddenly when the academic tools had become sufficiently sharp and accurate, or perhaps, equally important, sufficiently realistic, there was no end to the number of applications impatiently awaiting methods which could, really, deliver the goods" (Fisher, 1953). The philosopher and the historian have to be more fastidious, and recognize that applications are equally often created because the tools are there to address them, and that existing problems come to be redefined in terms of the new concepts that accompany these tools. We have seen in this chapter some striking instances of this, such as the change in the ideal of a scientific experiment and in the meaning of causality. This is perhaps most clearly exemplified in psychology, which was transformed by statistics from a science asking for general psychological laws to a discipline dedicated to searching for causal factors that operate on the average in a population. We will return to this in chapter 6.

The changes brought by the new methods of statistical inference are ubiquitous and profound. In early modern science, knowledge had to be publicly demonstrated, and it was customary to have witnesses sign a document that the events reported had actually happened. But these witnesses were not required to be able to replicate the demonstrations, whose

proper execution still depended on the high priests of science. In modern bureaucratic and utilitarian societies, such reliance on private knowledge to reach conclusions important for social life has become ideologically unacceptable. In less democratic times and places, the judgment of a political and social elite had no need to be clothed in facts and figures. Scientists, having little public accountability, could be satisfied with eyeballing data and relying on intuition, shared by the members of their group. But experts in modern democracies, as we argue in chapter 7, must be armed with "objective" inferential tools and mechanized experimental setups. Numbers and the methods of manipulating them are crucial to their authority. Those methods are sanctioned by a new kind of professional, the statistical specialist.

Perhaps these developments, most decisively of all those described in this book, epitomize the deep transformation of our view of the world and our methods of dealing with it that was brought about by the advance of probability. And yet, as is characteristic of this story, the progress of ideas about statistical inference was closely tied to the development of methods in particular studies. During the last century, two such fields have played an especially important role. One of these, agriculture, has been given considerable attention in this chapter. The other, genetics, is treated in the next chapter.

What does chance ever do for us?
William Paley (1802)

4

Chance and life: controversies in modern biology

4.1 INTRODUCTION

Developments in probability theory and statistics have certainly had a great impact on biology. But the rise and role of probabilistic and statistical thinking in biology is no mere reflection of those developments. In the first place, as was discussed in chapter 2, biology itself has had a significant influence on statistical thought. In the second place, there are episodes having to do with the rise and role of probabilistic thinking in biology that are neither greatly illuminated by, nor shed much light on, the development of what we have come to call "probability theory" and "statistics" proper. These latter developments mainly have to do with problems internal to biology.

In continuing the discussion begun in previous chapters about attitudes of biologists toward chance, let us now shift our perspective and look at the topic from the point of view of biologists *qua* biologists with their specifically biological concerns in mind. A number of the episodes to be discussed involve the most recalcitrant biological controversies of modern times: controversies concerning vitalism, mechanism, teleology, essentialism, and levels of organization and explanation. The special senses of "chance" invoked in each of these controversies, and the various motivations for or against taking chance seriously in each case are bound up with the terms of the dispute in question. The senses of chance to be discussed here are thus quite varied, and do not all (though some do) fit neatly into conceptual frameworks designed to accommodate notions of chance in other areas of science.