Erich L. Lehmann

Fisher, Neyman, and the Creation of Classical Statistics

Springer

Fisher, Neyman, and the Creation of Classical Statistics

Erich L. Lehmann

Fisher, Neyman, and the Creation of Classical Statistics

Springer

Erich L. Lehmann

ISBN 978-1-4419-9499-8 e-ISBN 978-1-4419-9500-1
DOI 10.1007/978-1-4419-9500-1
Springer New York Dordrecht Heidelberg London

Library of Congress Control Number: 2011930669

Printed on acid-free paper

Springer is part of Springer Science+Business Media (www.springer.com)

Preface

Classical statistical theory – hypothesis testing, estimation, and the design of experiments and sample surveys – is mainly the creation of two men: R. A. Fisher (1890–1962) and J. Neyman (1894–1981). Their contributions sometimes complemented each other, sometimes occurred in parallel, and, particularly at later stages, often were in strong opposition. The two men would not be pleased to see their names linked in this way, since throughout most of their working lives they detested each other. Nevertheless, they worked on the same problems, and through their combined efforts created a new discipline.

Fisher's collected papers have been published in five volumes, in which the ones excluding Genetics are numbered 1–291. The complete bibliography, including all his books, and pdf files of all papers are now publicly available at http://digital.library.adelaide.edu.au/coll/special/fisher. The list of books and the numbered statistical bibliography are included as an Appendix in the present book. All Fisher references in the Appendix will be cited by date and will include the bibliography number in square brackets if necessary for clarity.

Even more influential than the papers were Fisher's two great statistical books. The first, "Statistical Methods for Research Workers" (SMRW), was published in 1925, with new editions appearing every few years up to the fourteenth edition, which was published posthumously in 1973. This was followed in 1935 by the "The Design of Experiments" (DOE), which went through eight editions, the last dated 1966.

It is these two books that established Fisher as the creator of a new statistical methodology, and accordingly we shall here present his work largely through a detailed consideration of these volumes.

Neyman's contributions to this enterprise were contained principally in five papers published between 1928 and 1937, three of them jointly with Egon Pearson. Neyman too provided a summary statement, his mimeographed "Lectures and Conferences on Mathematical Statistics" of 1938, later expanded into a book (1952). However, they did not have the impact of Fisher's books, and in this case we shall instead study Neyman's original papers. Biographies have been written of both men: Fisher's by his daughter Joan Fisher Box, "R. A. Fisher—the Life of a

Scientist" (1978), and Neyman's ("Neyman—from Life" (1982)) by Constance Reid, who had previously published biographies of the mathematicians Hilbert and Courant. We shall here focus nearly exclusively on the work, but shall provide a chronology of some of the relevant life events as a framework in the next section.

Before proceeding, I should perhaps explain my own relation to, and involvement with, the work of Neyman and Pearson. In 1942, I became a student of Neyman, and his teaching of course reflected his own point of view. Fisher's name was hardly ever mentioned. But neither did Neyman point out that most of what he was presenting was his own work.

I got to know about this fact only when in 1977 Constance Reid asked me to read the Neyman-Pearson correspondence and to summarize it for her, since she had no background in statistics.

Later, I became interested in the general recent history of statistics, and in this context began to acquaint myself with Fisher's writings. It did not take me long to become aware of his dominating influence and to become an admirer of his genius. I gradually came to realize that these two men, Fisher and Neyman, so different in background, personality and approach, and so antagonistic to each other in person, between them were largely responsible for creating the field of classical statistics. The present account is the result of this realization.

Berkeley, CA Erich L. Lehmann

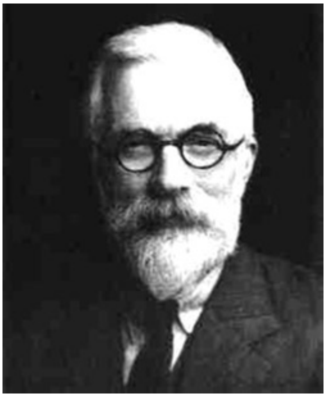
Contents

Chapter 1
Introduction

1.1 The Lives of Fisher and Neyman

As background to a study of their work, this section briefly sketches some of the main features of the lives of these two men.

1.1.1 Fisher



Ronald Aylmer Fisher was born in 1890, and at an early age showed a special ability for mathematics and science. After completing high school, he went to Cambridge in 1909 on a fellowship, and in 1912 graduated as a wrangler.[1] He remained at Cambridge for another year, during which he studied the theory of errors, statistical mechanics, and quantum theory.

[1] The students doing best in the examinations were designated as wranglers.

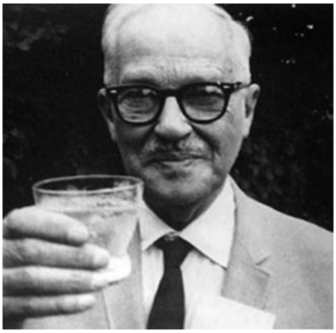
E.L. Lehmann, *Fisher, Neyman, and the Creation of Classical Statistics*,
DOI 10.1007/978-1-4419-9500-1_1, © Springer Science+Business Media, LLC 2011

When war broke out in 1914, Fisher volunteered for military service, but was rejected because of his poor eyesight and spent the next five years as a high school teacher. In 1919, he accepted a position as statistician at Rothamsted Experimental Station.[2] At first it was a temporary assignment, but after a year it became permanent, and Fisher stayed at Rothamsted until 1933. He began by analyzing records that had accumulated over many years, but then got involved with ongoing experiments and found methods to improve them.

The year 1933 brought the retirement of Karl Pearson as head of the Department of Applied Statistics at University College, London, which he had founded. The department was then split into the Department of Eugenics headed by Fisher as Galton Professor, and the Department of Statistics with Karl Pearson's son Egon as the head. It was an uncomfortable arrangement in which Fisher was barred from teaching statistics.

When World War II broke out in 1939, Fisher's department was evacuated and gradually dispersed. He did not find another position until in 1943 he was appointed to the Arthur Balfour Chair of Genetics at Cambridge. He remained in this position until his retirement in 1957. The remaining years of his life he spent in Adelaide, Australia where he died in 1962.[3]

1.1.2 Neyman



Neyman's life falls into two quite distinct periods of almost equal length: a European period (1894–1938) and an American one (1938–1981). He was born and raised in Russia to parents of Polish ancestry. In 1912, he entered the University of Kharkov, majoring first in physics and then mathematics. One of his mentors at Kharkov was Serge Bernstein,[4] who introduced him to probability theory and statistics, neither of which he found very interesting.

[2] Rothamstead is an agricultural experiment station located at Harpenden, about 25 miles north of London. An account of its history is provided in Box (1978, Chap. 4).

[3] For a retrospective look at Fisher's life and work with discussion by many statisticians, see Savage (1976).

[4] S. N. Bernstein (1880–1968). For an account of the life and work of this teacher of Neyman, see Heyde et al. (2001), pp. 339–342.

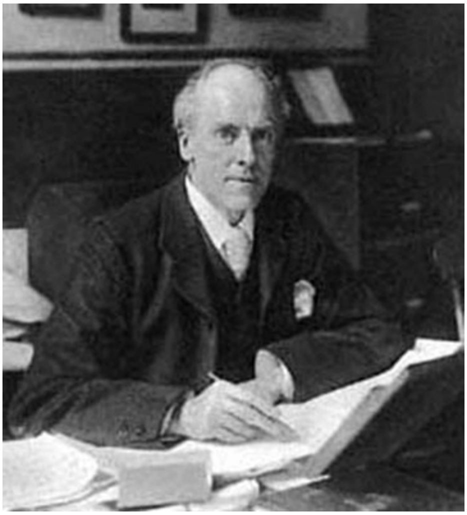
In 1921, Neyman went to Warsaw, and there found positions as statistician, first at the Agricultural Institute in Bydgoszcz, then at the State Meteorological Institute, and finally as assistant at the University of Warsaw and lecturer at the University of Krakow. In 1924, he obtained his doctorate with a thesis based on his work at Bydgoszcz.

Since no one in Poland was able to evaluate his work, he was given a fellowship to work with Karl Pearson, whose department in London was the center of the statistical world. The most important result of this year was the connection Neyman established with Pearson's son Egon. Impressed with Neyman's superior mathematical ability, Egon suggested that they might collaborate on some statistical problems about which he was thinking.

Neyman had obtained a fellowship for another year, which he was spending in Paris to work with Borel, Lebesgue, and Hadamard. In the fall of that year (1926), Egon sent him some notes on the proposed project, and this initiated their collaboration. After Paris, Neyman returned to Warsaw, where he eked out a precarious living as head of a small statistical laboratory and with various consulting jobs. His work with Egon was carried out by correspondence and occasional rare visits back and forth.

The situation improved greatly when Egon, now head of his own department, was able in 1934 to offer Neyman a position. By 1935, the position had become permanent and Neyman could now devote himself to teaching and research without the daily struggle for existence. Then, in the fall of 1937, out of the blue, came a most surprising letter. Its author, Griffith Evans, Chair of the mathematics department at the University of California at Berkeley, was offering Neyman a professorship in his department. After some hesitation, Neyman decided to accept. So, starting in the fall of 1938, he became a member of the Berkeley faculty and over the years built up a first-rate statistics department. He died in Berkeley in 1981.

1.2 Karl Pearson (1857–1936)



Fisher and Neyman may be considered the architects of a new discipline, but no one starts from scratch or works in isolation. In the present section we shall consider the contributions of the person who more than any other prepared the ground for their work, Karl Pearson.

Two contributions of Pearson that strongly influenced Fisher were the introduction of what are now known as Pearson curves (1895) and the later proposal to estimate their parameters by the method of moments (1902). This system of curves furnished an example of a family of distributions characterized by a few parameters (a term not used by Pearson), the generalization of which Fisher used as a fundamental concept. The method of moments provided Fisher with an opportunity to show the superiority of his maximum likelihood estimators.

Fisher (1922) [18] attests to the influence of Pearson's work when he refers to:

> The development by Pearson of a very extensive system of skew curves, the elaboration of a method of calculating their parameters, and the preparation of the necessary tables, a body of work which has enormously extended the power of modern statistical practice, and which has been, by pertinacity and inspiration alike, practically the work of a single man.

Pearson's most influential statistical contribution was the 1900 paper in which he proposed his χ^2-test for goodness of fit. A problem of great interest in the nineteenth century was to determine whether a given series of observations could be assumed to come from a normal distribution. The fit was made by eye and, as Pearson wrote: "The comparison in general amounts to a remark – based on no quantitive criterion – of how well practice and theory do fit!" This quantification (by means of his χ^2-statistic) was Pearson's crucial innovation.

Concerning it, Fisher, in the passage quoted above, continues: "Of even greater importance is the introduction of an objective criterion of goodness of fit. For empirical as the specification of the hypothetical population may be, this empiricism is cleared of its dangers if we can apply a rigorous and objective test of the adequacy with which the proposed population represents the whole of the available facts."

A very different but also highly influential publication was Pearson's 1892 book, "The Grammar of Science." From it, for example, Neyman (1957) learned a central aspect of his view of science, "that scientific theories are no more than models of natural phenomena."

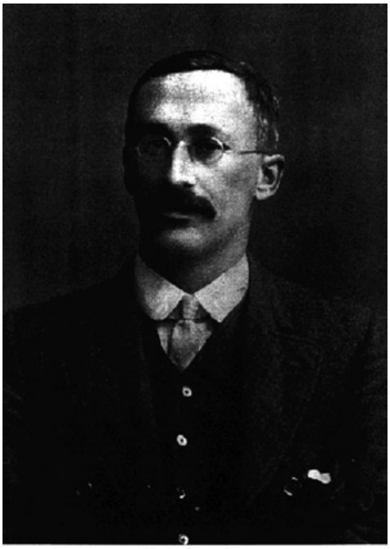
However, Pearson's contribution to statistics extended far beyond these specific innovations. He turned himself into the first professional statistician, and his department and laboratory became the first such center for statistical research and instruction. Of great importance also was the journal *Biometrika*, which he cofounded in 1901 and edited for the next 35 years.

A contemporary document attesting to Pearson's central role is a letter of June 1933 (Bennett, 1990, pp. 318–319), written by Major Greenwood on Fisher's appointment to the Galton Professorship at the University of London. Greenwood was a professor at the same university,[5] and welcomed him as a future colleague. He wrote:

[5] M. Greenwood (1889–1949) was Professor of epidemiology and Vital Statistics at the London School of Hygiene and Tropical Medicine from 1928 to 1945.

> There is, in my opinion, no other man alive worthy to sit in old K. P.’s chair; his mantle has descended on your shoulders and a double portion of his spirit is yours. Like all of us, you owe him much. If K. P. had never lived, you would have assuredly been one of our foremost men of science but very likely your field of work would not have been statistics, but perhaps, pure mathematics. …Although K. P. has often enough wounded my feelings and insulted my friends, I still love the man and venerate his genius; I should have been sad if his kingdom had fallen into the hands of a second-rater.

1.3 William Sealy Gosset (“Student”) (1876–1937)



“Student” in 1908

A second person who exerted a crucial influence on Fisher and – indirectly – on Neyman was, surprisingly, not a statistician but a brewer. W. S. Gosset studied chemistry at Oxford, and then in 1899 took a job with the brewery Arthur Guinness Son and Co., for which he worked until his early death at age 61. He soon found himself confronted with a flow of statistical problems he did not know how to handle. To educate himself, he read books on the theory of errors and least squares. In 1905, he consulted Karl Pearson and later wrote (E.S. Pearson, 1990) that

> He was able in about half an hour to put me in the way of learning the practice of nearly all the methods then in use.

In 1908, Gosset published, under the pseudonym “Student,” a paper (1908a) which initiated a new paradigm. For testing the value of a population mean, it had been customary to use a statistic equivalent to what today is called Student’s *t*, and to refer it to the normal distribution. For large samples, this provided a good approximation.

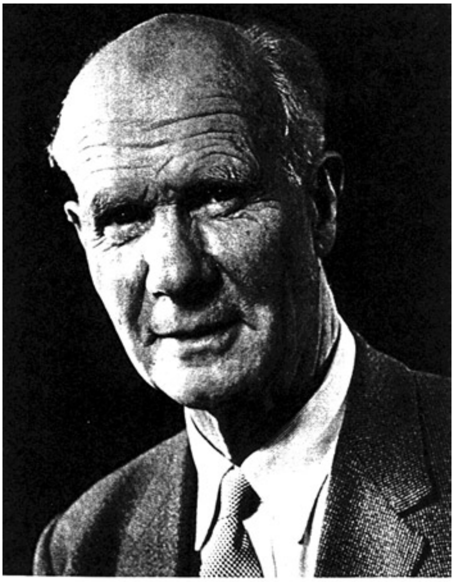
However, Gosset soon realized that for the small samples with which he had to work, the approximation was inadequate. He then had the crucial insight that exact results could be obtained by making an additional assumption, namely that the form of the distribution of the observations is known. Gosset undertook to determine it for the case that the underlying distribution is normal, and he obtained the correct result, although he was not able to give a rigorous proof.

The first proof was obtained (although not published) by Fisher in 1912. His proof was finally published in 1915 [4], together with the corresponding proof for the correlation coefficient that Student had conjectured in a second paper of 1908(b). Fisher followed this in 1921 [14] with a derivation of the distribution of the intraclass correlation coefficient. And then, as a result of constant prodding and urging by Gosset, he found a number of additional small-sample distributions, and in 1925 presented the totality of these results in his book, “Statistical Methods for Research Workers.”

So far we have mentioned two contributions by Gosset: his path-breaking paper of 1908, which was the starting point of Fisher’s new small-sample methodology; and his role in getting Fisher to develop this methodology much further than he (Fisher) had originally intended. A third equally crucial contribution – this one to the Neyman-Pearson theory – will be considered in the next section.

In his 1939 [165] obituary of “Student,” Fisher calls him “one of the most original minds in contemporary science,” and he refers to Gosset’s 1908(a) paper as providing “a fundamentally new approach to the classical problem of the theory of errors, the consequences of which are still only gradually coming to be appreciated in the many fields of work to which it is applicable.” Indeed, Student’s *t*-test continues, even today, to be one of the most widely used statistical procedures.

1.4 Egon S. Pearson (1895–1980)



EGON SHARPE PEARSON

Egon Pearson, the only son of Karl Pearson (who also had two daughters), obtained his degree in mathematics from Cambridge University in 1919, and continued with graduate studies in astronomy. In 1921, he joined his father's Department of Applied Statistics in a role akin to today's teaching assistant. His duties gradually increased and in 1926 he began to teach on his own. When his father retired in 1933, Egon succeeded him as chair of the Department of Statistics.

After having been trained by Karl Pearson, Egon, particularly after the publication in 1925 of Fisher's "Statistical Methods" book, became aware of the new small-sample approach of Gosset and Fisher. As he later wrote (1990, p. 77):

> In 1925-6, I was in a state of puzzlement, and realized that, if I was to continue an academic career as a mathematical statistician, I must construct for myself what might be termed a statistical philosophy, which would have to combine what I accepted from K. P.'s large-sample tradition with the newer ideas of Fisher.

He decided to write to Gosset, and in his 1939 obituary of Gosset recalled that, "I had been trying to discover some principle beyond that of practical expediency which could justify the use of 'Student's' *z*." (This is now called Student's *t*.) In his reply, Gosset pointed out:

> Even if the chance is very small, say .00001, that doesn't in itself necessarily prove that the sample is not drawn randomly from the population [specified by the hypothesis]; what it does is to show that if there is any alternative hypothesis which will explain the occurrence of the sample with a more reasonable probability, say .05 (such as that it belongs to a different population or that the sample wasn't random or whatever will do the trick), you will be very much more inclined to consider that the original hypothesis is not true." (E. S. Pearson, 1939.)

In his obituary of Gosset, Pearson continues,

> Gosset's reply had a tremendous influence on the direction of my subsequent work, for the first paragraph contains the germ of that idea which has formed the basis of all the later joint researches of Neyman and myself. It is the simple suggestion that the only valid reason for rejecting a statistical hypothesis is that some alternative explains the observed events with a greater degree of probability.

So Gosset had done it again. His 1908(a) paper had initiated a whole new way of thinking which was realized in Fisher's small-sample methodology. Now again he had supplied the needed key insight – the result was the Neyman-Pearson theory of hypothesis testing.

Inspired by Gosset's suggestion, Pearson started to work out the consequences, and soon proposed to Neyman to collaborate on the problem. He thought Neyman would be helpful because of his greater mathematical ability. But there may have been another reason. Egon realized that he had to break free of his father's dominance, that Karl Pearson would neither approve of nor sympathize with the new direction that Egon was taking. Neyman would be able to give him moral support in this difficult undertaking. Neyman was especially suitable for this role because as a foreigner he was independent of Karl Pearson while Egon's young colleagues were all students of "the professor."

Thus, in 1926 started a collaboration that was to last for slightly more than a decade and which gave us the Neyman-Pearson theory of hypothesis testing.

Neyman followed this work with his theories of survey sampling and confidence intervals, and later with a new and very influential philosophy of statistics. However, in these projects Pearson did not participate.

1.5 Fisher's Foundations Paper

We continue this introduction with a discussion of Fisher's 1922 [18] paper, "On the mathematical foundations of theoretical statistics." The reason for considering this paper here instead of with Fisher's other work is its fundamental role in establishing the framework for all the later work of Fisher and Neyman.

Stigler (2005) calls it "arguably the most influential article on that subject [theoretical statistics] in the twentieth century," and describes it as "an astonishing work: It announces and sketches out a new science of statistics, with new definitions, a new conceptual framework and enough hard mathematical analysis to confirm the potential and richness of this new structure." And Hald (1998, p. 713), with regard to it, states that, "For the first time in the history of statistics, a framework for a frequency-based general theory of parametric statistical inference was clearly formulated."

The paper opens with a list of 15 definitions in alphabetical order, of which perhaps the most important are: consistency, efficiency, likelihood, optimum, specification, and sufficiency. These definitions establish a vocabulary for the new discipline Fisher was creating.

This list is followed by a brief historical section which Fisher uses to inveigh against inverse probability (i.e., the Bayes approach).[6] He dismisses it as "a mistake (perhaps the only mistake to which the mathematical world has so deeply committed itself)."

The body of the paper begins with section 2, "The purpose of statistical method." It states that the task of statistics is to replace the full data by a "relative few quantities, which shall adequately represent the whole." The next sentences are crucial; they set up the framework within which statistics was to function for a long time:

> This object is accomplished by constructing a hypothetical infinite population, of which the actual data are regarded as constituting a random sample. The law of distribution of this hypothetical population is specified by relatively few parameters which are sufficient to describe it exhaustively in respect of all qualities under discussion.

Such parametric families or models constituted the home for classical statistical inference, as it was developed by Fisher and his successors during the next decades. The most useful parametric families for this purpose turned out to be not the family of Pearson curves, but two families defined by Fisher in 1934 [108]. In modern terminology, they are the exponential and transformation (or group) families.

[6] The role of inverse probability in the nineteenth century is discussed, for example, in Stigler (1986) and Hald (1998).

This section also contains Fisher's definition of probability. He says of it that

> It is a parameter which specifies a simple dichotomy in an infinite hypothetical population, and it represents neither more nor less than the frequency ratio which we imagine such a population to exhibit.

He illustrates the concept with the probability of throwing a five with a die. "When we say that [this] probably is one-sixth," he writes, "we must not be taken to mean that of any six throws with that die one and only one will necessarily be a five; or that of any six million throws, exactly one million will be fives; but that of a hypothetical population of an infinite number of throws, with the die in its original condition, exactly one-sixth will be fives." We shall return to Fisher's view of probability in section 6.2 in the context of fiducial probability.

The third section of Fisher's paper is entitled, "The problems of statistics," and it lists three such problems: specification (i.e., of the model), estimation, and distribution.

Of the problems of estimation, Fisher explains that "they involve the choice of methods of calculating from a sample…statistics, which are designed to estimate the values of the parameters of the hypothetical population."

The problems of distribution include "discussions of the distribution of statistics derived from samples, or in general any functions of quantities whose distribution is known."

Today's readers will ask with surprise: What about hypothesis testing? Reading on, they will find that Pearson's χ^2-test and Student's work are mentioned as examples of problems of distribution. To explain why this is the case and how Fisher's view of this issue shifted requires more general considerations and will be postponed to the next section.

Having defined the problem of statistics to be the estimation of parameters, Fisher in Section 4 (Criteria of estimation) states the properties that he desires for his estimators. They are consistency, efficiency, and sufficiency. His concept of consistency differs from that now associated with this term (convergence in probability to the true value as the sample size tends to infinity). Instead he requires:

> that when applied to the whole population the derived statistic should be equal to the parameter.

(This is now called Fisher consistency.)

Since many statistics will satisfy this criterion, Fisher supplements it by requiring efficiency, namely:

> that in large samples, when the distribution of the statistics tend to normality, that statistic is to be chosen which has the least probable error.

Both of these criteria are asymptotic, and two statistics, both of which are consistent and efficient, may differ for all finite samples. So Fisher adds the third condition of sufficiency:

> that the statistic chosen should summarize the whole of the relevant information supplied by the sample.

This leaves the statistician with the task of determining an estimator satisfying these criteria. After a section of examples, Fisher takes up this challenge in section 6, entitled, "Formal solution of the problems of estimation."

He begins the section by discussing inverse probability and Bayes' Theorem and then considers the concept of likelihood in the context of a binomial distribution with unknown parameter *p*. He reiterates that about the frequencies of different values of *p*, "we know nothing whatever," and continues:

> We must return to the actual fact that one value of *p*, of the frequency of which we know nothing, would yield the observed result three times as frequently as would another value of *p*. If we need a word to characterize this relative property of different values of *p*, I suggest that we may speak without confusion of the *likelihood* of one value of *p* being thrice the likelihood of another, bearing always in mind that likelihood is not here used loosely as a synonym of probability, but simply to express the relative frequencies with which such values of the hypothetical quantity *p* would in fact yield the observed sample.

He then proposes what he had already suggested earlier in the section on the solution of the estimation problem, the method of maximum likelihood, which "consists, then, simply of choosing such values of these parameters as have the maximum likelihood."

Fisher believes that this method satisfies his three criteria, in particular that it satisfied the criterion of sufficiency, although he states that he "is not satisfied as to the mathematical rigor of any proof which I can put forward to that effect." He also claims that sufficiency implies efficiency.

The paper continues with several sections of examples, elaborations, and comparisons with other approaches, but this is the gist of it. Not everything turned out to be correct, but basically maximum likelihood estimators are asymptotically efficient. Thus, in this paper Fisher has not only formulated the general problem of optimal estimation, but he has also provided a solution. It is a stunning achievement.

1.6 Estimation and Testing

Let us now return to the question raised in the preceding section of why the only statistical inference considered in Fisher's foundations paper was estimation and included no explicit mention of testing, while only a few years later testing became his primary focus. To this end, we shall examine certain aspects of two series of papers written by Fisher between 1913 and 1925, the year of publication of his book, "Statistical Methods for Research Workers," which we shall consider in some detail in Chap. 2.

The two series consist respectively of papers 1, 12, 18, and 42; and 4, 14, 20, 30, 35, 36, and 43, where the numbers are those assigned to these papers in Fisher's five-volume Collected Papers (see the Appendix).

These two series are of very different nature. The first is concerned with the theory of estimation: the choice of an appropriate estimate of an unknown parameter and the comparison of different estimates. This series consisted of Fisher's student paper [1] of 1913, in which he proposed maximum likelihood estimation, although not under this name and unaware that he was not the first to suggest this method.[7] This was followed by [12] in 1920 with a comparison of two estimates of a normal variance that led to Fisher's discovery of sufficiency. This discovery was central to his foundational paper [18] of 1922, which was refined and elaborated on in 1925 with a paper [42] simply entitled, "Theory of statistical estimation." At this point Fisher felt that he had now "cleared up the main outstanding difficulties" and had achieved "a theory of statistical estimation with some approach to logical completeness."

The other series is concerned with what Fisher in [18] had called "problems of distribution," i.e., with the determination of the small-sample ("exact") distribution of a number of increasingly complex statistics. These were all derived under the assumption that the underlying observations are normally distributed.

The point of view in these papers is quite different from that in the estimation series. While there the central problem was that of determining an appropriate estimate, here the statistic is given and the difficulty is that of determining its distribution. Another basic difference is that in all but one of the estimation papers Fisher is attempting to provide a general theory; on the other hand each of the small-sample distribution papers deals with one or more quite specific situations. The statistical (inferential) use to be made of these distributions is peripheral in these papers, and Fisher's view on this use seems to have shifted over time.

In the first paper (1915) [4] of the series, he wrote:

> "Student," if I don't mistake his intentions, desiring primarily a just estimate of the accuracy to be ascribed to the mean of a sample…

This despite the fact that Student stated in the introduction to his paper:

> The aim of the present paper is to determine the point at which we may use the tables of the probability integral [i.e., the large-sample normal approximation] in judging the significance of the mean…

In the body of the second (1921) [14] paper, Fisher was still wedded to the estimation point of view when he explains:

> Thus, in calculating the correlation from a sample, we are making an estimate of the [population] correlation. …We wish to make the best possible estimate and to know as accurately as possible how far the estimate may be relied upon.

Only at the end of the paper, in an example, is a question of significance raised. A p-value is calculated under the hypothesis that the population correlation $\rho = 0.18$ and the difference of the sample value from the hypothetical one is declared to be "now much more significantly apparent." This is essentially a significance test.

The phrase "testing for significance" occurs for the first time in the last (sixth) section of a 1922 paper [20]. This section obtains the distribution of regression coefficients. Here, referring to tables of Student's distribution, he states that, "These tables are in a suitable form for testing the observed significance of an observed regression coefficient." From then on, exact significance tests rather than estimation seem to have been the principal focus of Fisher's small-sample work.

[7] For discussion of and historical background on this approach, see Stigler (2007).

For example, in the survey paper [36] of 1924, he mentions that the small-sample distribution of many important statistics is far from normal and that then

> tests of *Significance* based upon the calculation of a "probable error" or "standard error" are inadequate and may be very misleading. In addition to tests of significance, tests of goodness of fit also require accurate error functions; both types of tests are constantly required in practical research work; the test of goodness of fit may be regarded as a kind of generalized test of significance…

To today's reader, it seems strange that Fisher restricted the term "test of significance" to the testing of hypotheses that specify the value of one or more parameters. But in any case, there clearly was a shift in his point of view concerning the principal use to be made of the small-sample distributions he was deriving.

We can only speculate as to what caused this change. A plausible explanation seems to be the new position Fisher took in 1919. From 1914 to 1919, he had been teaching high school physics and mathematics. Then in 1919 he accepted an appointment as statistician (later chief statistician) at Rothamsted Experiment Station, with the initial assignment of studying the massive agricultural data that had been accumulated there. In this work he may have found significance testing a very useful approach.

This explanation receives some support from Fisher's first paper [15] of 1921 on these data, "Studies in crop variation." Table II of this paper lists a number of *p*-values and discusses the resulting significance implications, and section 8 is entitled, "The significance of an observed term."

Our discussion so far has been concerned with the shift in Fisher's focus from estimation to testing. Let us now briefly consider why originally he focused so exclusively on estimation.

In his undergraduate study at Cambridge, Fisher took a single course in statistics and that was on the theory of errors. The standard textbooks on this subject were primarily concerned with least squares estimation and the calculation of the probable error as a measure of accuracy of the estimates. Thus, significance testing is not likely to have played much of a role in this course.

In addition, it would seem that from a naïve, intuitive point of view, estimation is a natural first task for statistics. In comparison, testing appears more artificial and convoluted. Thus, estimation was a natural starting point for Fisher.

Chapter 2
Fisher's Testing Methodology

2.1 The Small-Sample and χ^2 Papers

We saw in Sect. 1.3 that Student in 1908a brought a new point of view to statistical inference by determining the small-sample (exact) distribution of what he called z, now called t, under the assumption of normality. Student found the correct form of this distribution but was not able to prove it.

In 1912 Ronald Fisher, then still a Cambridge undergraduate, sent Gosset a proof using n-dimensional geometry. Gosset, stating that he did not "feel at home in more than three dimensions," sent it on to Karl Pearson, with the suggestion to publish it as a note in *Biometrika*.

However, Pearson replied that, "I do not follow Mr. Fisher's proof and it is not the kind of proof which appeals to me. … Of course, if Mr. Fisher will write a proof in which each line flows from the preceding one and define his terms, I will gladly consider its publication. Of the present proof I can make no sense."

Despite this negative reaction, three years later Pearson did accept a paper in which Fisher used his geometric method to derive the small-sample distribution of the correlation coefficient from a bivariate normal distribution (and in passing also that of Student's t) (1915) [4]. For the case of zero correlation, this distribution had been correctly conjectured (but again not proved) by Student in a second *Biometrika* paper, also in 1908 (Student, 1908b).

After 1915, for a number of years Fisher did no further work on such distributional problems, but he was pulled back to them when investigating the difference between the inter- and intraclass correlation coefficients. The distribution of the latter was still missing and Fisher derived it by the same geometric method he had used previously (Fisher, 1921) [14].

An idea of Fisher's thinking at the time can be gathered from his 1922 [18] paper discussed in Sect. 1.5, where he mentions the "absence of investigation of other important statistics, such as regression coefficients, multiple correlations and the correlation ratio." He refers to these problems in the summary of the paper as "still affording a field for valuable enquiry by highly trained mathematicians."

E.L. Lehmann, *Fisher, Neyman, and the Creation of Classical Statistics*,
DOI 10.1007/978-1-4419-9500-1_2, © Springer Science+Business Media, LLC 2011

This passage suggests that Fisher thought these problems to be difficult, and that he had no plans to work on them himself. However, in April 1922 [20] he received two letters from Gosset that apparently changed his mind. We are lucky that these and subsequent letters from Gosset to Fisher (but unfortunately few of Fisher's replies) were preserved and in 1970 were privately published by Gosset's employer, the firm Arthur Guinness Son and Co. (Gosset, 1970).

In the first of the letters mentioned, Gosset pleaded: "But seriously, I want to know what is the frequency distribution of $r\ \sigma_x/\sigma_y$ for small samples, in my work I want that more than the r distribution now happily solved."

Fisher's solution to this problem (together with that of the two-sample problem) appeared in the *Journal of the Royal Statistical Society* in 1922 [20]. The paper is primarily concerned with a different problem, that of testing the goodness of fit of regression lines. At the end, Fisher appends a section which in view of the dates must have been added at the last moment. Fisher acknowledges in this section that:

> an exact solution of the distribution of regression coefficients…has been outstanding for many years; but the need for its solution was recently brought home to the writer by correspondence with "Student," whose brilliant researches in 1908 form the basis of the exact solution.

The solution to Gosset's problem turned out to be surprisingly easy. It consisted essentially in showing that the regression coefficients had the same structure as Student's *t*, and hence were distributed according to Student's distribution, however with the degrees of freedom in the denominator reduced by one in the case of simple linear regression, and by more if more than one variable is involved. The argument was so simple that Fisher was able to send it to Gosset by return mail.

Gosset's second letter followed within a few days. It was a short note, saying:

> I forgot to put another problem to you in my last letter, that of the probable error of the partial $\left\{\frac{\text{Correlation}}{\text{Regression}}\right\}$ coefficients in small samples.

Using once more the geometric method, Fisher was able to reduce the partial correlation to the total correlation based on a smaller sample. He sent this solution to Gosset without much delay and published it in a short note in 1924 [35].

In the same year, Fisher gave a lecture[1] at the International Congress of Mathematics in Toronto, "On a distribution yielding the error functions of several well-known statistics" [36]. The distribution in question, which Fisher called the general z-distribution, is equivalent to that of the ratio of two independent χ^2-based estimates of a common variance, required for the analysis of variance. This talk and a 1925 paper on "Applications of 'Student's' distribution" outlined the new testing methodology Fisher had developed during the preceding decade and of which he wrote an account in a book that will be discussed in the next sections.

The small-sample papers reviewed in the present section constitute the second series mentioned in Sect.1.6. It should be pointed out that between 1922 and 1924,

[1] The Proceedings were not published until 1928.

Fisher also published a third series consisting of papers 19, 31, and 34 on χ^2-tests of goodness of fit. In the first of these he corrected Pearson’s error regarding degrees of freedom[2] and noted that his proof applies to χ^2-tests in “all cases in which the frequencies observed are connected with those expected by a number of relations, beyond their restriction to the same total frequency.”

The next paper is concerned with the special case of a 2×2 table where Fisher shows that when the margins are fixed so that the whole table is determined by the entry in any one of the four cells, the χ^2-statistic for testing independence has a limiting χ^2-distribution with one degree of freedom, rather than three as Pearson had claimed.

Finally, paper 34 of 1924 deals with goodness of fit tests in which some parameters have to be estimated. Here, Fisher finds that the χ^2-distribution with the appropriate number of degrees of freedom is valid provided the parameters are estimated efficiently, and hence in particular when the estimators used are either maximum likelihood or minimum chi-squared.

2.2 “Statistical Methods for Research Workers” I

The impulse to write a book on the statistical methodology he had developed came not from Fisher himself but from D. Ward Cutler, one of the two editors of a series of “Biological Monographs and Manuals” being published by Oliver and Boyd. We owe this information to a letter Fisher wrote in 1950 to Robert Grant, then a director of Oliver and Boyd (Bennett, 1990, pp. 317–318.) In this letter, he also points out that writing the book did not require new research but, “I only had to select and work out in expository detail the examples of the different methods proposed. It was often quite a job to find a good example.”

The book came out in 1925 with the title, “Statistical Methods for Research Workers” (SMRW), and it revolutionized the practice of statistics. The opening introductory chapter lays out Fisher’s vision of statistics. He states that

> statistics may be regarded as (i) the study of populations, (ii) as the study of variation, (iii) as the study of methods of the reduction of data.

In the latter category, he again mentions, as he had done in his 1922 [18] foundational paper, that the problems arising in this reduction can be divided into three types: problems of specification, estimation, and distribution.

He goes on to discuss the role of probability in this work, and to attack the method of “inverse probability” (what today we would call the Bayesian approach), stating his “personal conviction…that the theory of inverse probability is founded upon an error, and must be wholly rejected.” As an alternative, he suggests that “the

[2] A careful analysis of Pearson’s evolving progress on this issue is given in Chap. 19, “Karl Pearson and degrees of freedom,” of Stigler (1999).

mathematical quantity which appears to be appropriate for measuring our preference among different possible populations" is that of likelihood.

Next, he defines a number of concepts of the theory of estimation, such as consistency, efficiency, sufficiency, and maximum likelihood.

It comes as somewhat of a surprise when the next section of the introduction entitled "the scope of this book" states that

> the prime object of this book is to put into the hands of research workers…the means of applying statistical tests accurately to numerical data accumulated in their own laboratories…

and later refers to

> the exact distributions with the use of which this book is chiefly concerned…

Thus, the book does not primarily deal with estimation but with significance testing. In fact, estimation is never again mentioned.

Under these circumstances, one wonders why the introduction mentions distributions as the second principal concern of statistics rather than testing. As was already indicated in Sect. 1.6, this is actually quite reasonable. Unlike estimation, where the choice of an appropriate estimate was the central problem, the corresponding issue for testing did not arise for Fisher. Assuming normality, the choice of test statistic was intuitively obvious to him, and once its distribution had been obtained, the test consisted simply of declaring significance for observations in the appropriate tail of the distribution. Thus, the only serious problem was that of distribution, i.e., of determining the distribution of the test statistic.

The introduction concludes with a brief discussion of the tables "which form a part essential to the use of the book."

The introductory chapter is followed by two preparatory chapters on diagrams and distributions respectively. The latter introduces the reader to various aspects of three basic distributions: the normal, Poisson, and binomial.

Chapter IV deals with χ^2-tests of goodness of fit, independence, and homogeneity. After reminding the reader that such tests were already used in the preceding chapter to test the goodness of fit of Poisson and binomial distributions, Fisher states that the general class of problems for which it is appropriate is "the comparison of the number actually observed to fall into any number of classes with the numbers which upon some hypothesis are expected." The proposed test statistic, denoted by χ^2, is the sum over all the classes of x^2/m where "m is the number expected and $m + x$ the number observed in any class."

Fisher points out that "the more closely the observed numbers agree with those expected the smaller χ^2 will be; in order to utilize the table it is necessary to know the value of n with which the table is to be entered. The rule for finding n is that n is equal to the number of degrees of freedom in which the observed series may differ from the hypothetical; in other words, it is equal to the number of classes the frequencies in which may be filled up arbitrarily." This definition, which an uninitiated reader may have found somewhat cryptic, is made clearer through a number of examples.

There still remains an important piece of business (relating to the table) to deal with before the tests are illustrated by means of seven examples covering twelve pages. A table of the χ^2-distribution had been published in 1902 in *Biometrika*, but it was protected by copyright and Fisher could not get permission to reprint it. So instead of giving a table of the probability[3]

$$P = P(\chi^2 \geq x),$$

he tabled x as a function of P for selected values of P between 0.99 and 0.01. He comments that:

> In preparing this table we have borne in mind that in practice we do not want to know the exact value of P for any observed χ^2, but, in the first place, whether or not the observed value is open to suspicion. If P is between .1 and .9 there is certainly no reason to suspect the hypothesis tested. If it is below .02 it is strongly indicated that the hypothesis fails to account for the whole of the facts. We shall not often be astray if we draw a conventional line at .05 and consider that higher values of χ^2 indicate a real discrepancy.

This statement has been quoted in full because of its great influence. Fisher's recommendation of 5% as a fixed standard took hold and, for good or ill, has permeated statistical practice. Its advantages and disadvantages have been much discussed and will be compared with a more flexible alternative approach in Chap. 3.

Fisher gives further support to the proposal in the examples which illustrate the χ^2-test in a number of representative situations. In the first of these, in particular, he finds a p-value between 0.01 and 0.02 and concludes: "If we take $P = 0.05$ as the limit of significant deviation, we shall say that in this case the deviations from expectation are significant."

The χ^2-chapter concludes with a section on the partitioning of χ^2 into components "to test the separate components of a discrepancy." This foreshadows the analysis of variance treated in the last chapter of the book.

2.3 "Statistical Methods for Research Workers" II

The four chapters described so far make up nearly half of SMRW. The second half is concerned with the "exact" small-sample tests which are the avowed purpose of the book. Chapter V in particular deals with the problems for which the t-distribution provides the solution: tests of significance of means, differences of means, and regression coefficients.

Fisher, however, starts the chapter with a brief discussion of the large-sample tests for a mean or a difference of two means, pointing out that these procedures are what today we call distribution-free, i.e., valid for all distributions with finite variance.

Only then does he consider the small-sample t-test, stating that

> The distribution of t for random samples of a normal population about zero as mean, is given in the table of t.

[3] Today such p-values are usually denoted by a lower-case rather than capital p.

After illustrating the procedure on a sample of ten patients that had already been used by Student in 1908, he next discusses the "significance of difference of means of small samples." This case is of particular interest since it is one of the few in which today's approach differs from that presented by Fisher.

The heading of Fisher's treatment of the two-sample problem reads:

(*) To test whether two samples belong to the same population, or differ significantly in their means.

The test he proposes for this problem is the two-sample *t*-test.

However, the dichotomy (*) is clearly incomplete. Realizing this, Fisher in the second (1928) edition of the book adds the following paragraph:

> It will be noted…that a difference in variance between the populations from which the samples are drawn will tend somewhat to enhance the value of *t* obtained.[4] The test, therefore, is decisive, if the value of *t* is significant, in showing that the samples could not have been drawn from the same population; but it might conceivably be claimed that the difference indicated lay in the variances and not in the means. The theoretical possibility, that a significant value of *t* should be produced by a difference between the variances only, seems to be unimportant…; as a supplementary test, however, the significance of the difference between the variances may be tested directly by the method of § 41.

In still later editions, Fisher adds yet another paragraph:

> It has been repeatedly stated, perhaps through a misreading of the last paragraph, that our method involves the "assumption" that the two variances are equal. This is an incorrect form of statement; the equality of variances is a necessary part of the hypothesis to be tested, namely that the two samples are drawn from the same normal distribution. The validity of the *t*-test is therefore absolute, and requires no assumption whatever.

Fisher concludes this later discussion by pointing out that one could of course ask the question: "Might these samples have been drawn from different normal populations having the same mean?" He states that this problem has been solved (it is what is today known as the Behrens-Fisher problem), but that "the question seems somewhat academic."

Despite Fisher's protest, the modern view considers the testing of the difference of two means in two versions:

(i) Assuming nothing about the two variances (the Behrens-Fisher problem) or
(ii) Assuming the two variances to be equal, in which case the *t*-test is the appropriate (and by a number of criteria the best possible) procedure.

This disagreement serves to emphasize the fact that for most of the procedures set out in SMRW, Fisher's way of seeing and doing things still holds sway today. It also illustrates that in an argument Fisher rarely gave an inch. Those holding views different from his own had "misread" him and their statements were "incorrect." We shall see this attitude repeated in later controversies.

The remainder of Chapter V is concerned with regression. It is shown how the *t*-test also applies to the testing of regression coefficients in linear regression. Fisher then

[4] This statement is correct for some balanced designs but is not correct in general; see, for example, Scheffé (1959, p. 353).

turns to the fitting of "curved regression lines" for the case "when the variability of the independent variate is the same for all values of the dependent variate, and is normal for each such value." The chapter concludes with the testing of partial regression coefficients, where the distribution of the test statistic is again found to be Student's *t*.

Chapter VI has a somewhat narrower focus: correlation coefficients. It defines the correlation coefficient ρ in a bivariate normal distribution and discusses its estimation, and from there progresses to partial correlation. It next takes up the significance of an observed correlation coefficient *r* and states that when $\rho=0$, then

$$t = r\sqrt{n-2}\Big/\sqrt{1-r^2},$$

has a *t*-distribution with $n-2$ degrees of freedom, and extends this result to partial correlation coefficients.

Having thus provided the means for testing that these coefficients are zero, Fisher turns to the problem of testing $\rho=\rho_0$ for $\rho_0\neq 0$. For this purpose, he proposes the transformation

$$z = \tfrac{1}{2}\log\,[(1+r)/(1-r)]$$

and states that z is approximately normally distributed with mean ρ and standard error $1/\sqrt{n-3}$. He cautions that

> The distribution of *z* is not strictly normal, but it tends to normality rapidly as the sample is increased, whatever may be the value of the correlation.

After considerable further discussion, Fisher shows in two examples that the approximation is reasonably accurate. As an additional application, he points out that this transformation also makes it possible to test the difference between two observed correlations.

Chapter VII extends these considerations to the case of intraclass correlations, but it does so apologetically. It starts out with the admission that the data to be analyzed in this chapter by means of correlations "may be more usefully and accurately treated by the analysis of variance, that is by the separation of the variance ascribable to one group of causes, from the variance ascribed to other groups."

And again, after discussing and illustrating a generalized *z*-transformation of the intraclass correlation coefficient, Fisher states that "a very great simplification is introduced into questions involving intraclass correlation when we recognize that in such cases the correlation merely measures the relative importance of two groups of factors causing variation."

Still later, he points out that "the data provide us with independent estimates of two variances; if these variances are equal the correlation is zero;… . If, however, they are significantly different, we may if we choose express the fact in terms of a correlation."

The final sections of the chapter then at last deal directly with tests which arise in the analysis of variance. "The test of significance of intraclass correlations is thus simply," Fisher writes, "an example of the much wider class of tests of significance which arise in the analysis of variance. These tests are all reducible to the single problem of testing whether one estimate of variance derived from n_1 degrees of freedom is significantly greater than a second such estimate derived from n_2 degrees of freedom. This problem is reduced to its simplest form by calculating z equal to half the difference of the logarithms of the estimates of the variance." Fisher provides a table for $P = 0.05$ and selected values of n_1 and n_2.

The remainder of the chapter illustrates the use of the table and contains a brief discussion of "analysis of variance into more than two portions."

The last chapter (Chapter VIII) consists of applications of the analysis of variance to a few important situations. The first half of the chapter is concerned with testing the structure of regression lines; this includes a discussion of the multiple correlation coefficient, and is an extension of earlier material.

The second half strikes out in a new direction, resulting from Fisher's work at Rothamsted. As he writes, "The statistical procedure of the analysis of variance is essential to an understanding of the principles underlying modern methods of arranging field experiments. The first requirement which governs all well-planned experiments is that the experiment should yield not only a comparison of different manures, treatments, varieties, etc., but also a means of testing the significance of such differences as are observed."

In the last pages of the book, he illustrates such analyses first on the case that the agricultural plots are assigned to the different treatments (with replication) completely at random, then into randomized blocks, and finally into Latin squares.

The expansion of this program was to be one of Fisher's most important tasks during the next decade.

2.4 The Reception of SMRW

Fisher was greatly disappointed in the reviews of his book. The most important British journal of statistics publishing book reviews was the *Journal of the Royal Statistical Society (JRSS)*, and its January issue of 1926 carried a review of Fisher's book signed L. I. The reviewer was Leon Isserlis, a statistician at the Chamber of Shipping, with a strong Cambridge mathematics background and graduate studies under Karl Pearson.

Isserlis' review was not very enthusiastic: "We have presented [here]," he wrote, "a very full account of the statistical methods favored by the author and of the conclusions he has reached on topics, some of which are still in the controversial stage. Much is lacking if the book is to be regarded as an authoritative record of achievement in statistical method apart from Mr. Fisher's own contributions."

After briefly describing the content of the book, the review reaches the conclusion that

> The book will undoubtedly prove of great value to research workers whose statistical series necessarily consist of small samples, but will prove a hard nut to crack for biologists who attempt to use it as a first introduction to statistical method."

If Isserlis was not overly enthusiastic, Fisher himself must bear part of the blame. In the author's preface, he asserted:

> Little experience is sufficient to show that the traditional machinery of statistical processes is wholly unsuited to the needs of practical research. Not only does it take a cannon to shoot a sparrow, but it misses the sparrow.

Such wholesale and insulting dismissal of traditional procedures can hardly have endeared him to the British establishment.

However, even reviewers who were not offended by Fisher's attack on traditional methods found much to criticize. In particular, they complained about Fisher's dogmatism, the lack of proofs, the emphasis on small samples, and the difficulty of the book.

The only review that did justice to Fisher's achievement came not from Great Britain but from America. The reviewer was Harold Hotelling (1927), who had obtained his Ph.D. in 1924 from Princeton in topology, and who then took up a position as research associate at the Food Research Institute at Stanford University. He was so impressed with the book that he submitted a review to the *Journal of the American Statistical Association* without having been asked to do so.

He opens the review by stating that

> Most books on statistics consist of pedagogic rehashes of identical material. This comfortably orthodox subject matter is absent from the volume under review, which summarizes for the mathematical reader the author's independent codification of statistical theory and some of his brilliant contributions to the subject, not all of which have previously been published.

After some discussion of the content of the book, the review concludes with:

> The author's work is of revolutionary importance and should be far better known in this country.

The first edition was soon sold out, and in 1928 Fisher brought out a second edition. He was pleased with this success and considered it a vindication. He gave expression to his feelings in the preface to the new edition, when he wrote:

> The early demand for a new edition has more than justified the author's hope that use could be made of a book which, without entering into the mathematical theory of statistical methods, should embody the latest results of that theory, presenting them in the form of practical procedures appropriate to those types of data with which research workers are actually concerned.

He goes on to defend his much-criticized decision not to include proofs by pointing out that, "The practical application of general theorems is a different art from their establishment by mathematical proof, and one useful to many to whom the other is unnecessary."

The remainder of the preface outlines the changes from the first edition. The most important of these is the addition of a new chapter on "The Principles of Statistical Estimation." The presentation is very unusual. Fisher illustrates the concepts of consistency, efficiency, relative efficiency, and so on by means of a single example: a contingency table arising in genetics, in which all cell probabilities are expressible in terms of a single parameter θ.

Before discussing the estimation of θ, Fisher notes that, "It is a useful preliminary before making a statistical estimate…to test if there is anything to justify estimation at all." He therefore tests for independence (which would specify the value of θ) and rejects the hypothesis. He then points out that, "Nothing is easier than to invent methods of estimation," and proposes to consider four estimates of θ (T_1 to T_4), the fourth being the maximum likelihood estimate.

The first two are linear functions of the cell frequencies, and Fisher shows how to calculate their variances. The other two are more complicated, so he provides large-sample approximations to their variances. These asymptotic variances for T_3 and T_4 turn out to be equal, and of this common value he writes that "it is of great importance for our problem, for it has been proved that no statistic can have a smaller variance, in the theory of large samples, than has the solution of the equation of maximum likelihood. This group of statistics…are therefore of particular value, and are designated *efficient* statistics."

He goes on to consider the reciprocal of this minimum asymptotic variance "as a numerical measure of the total amount of information, relevant to the value of θ, which the sample contains." He further states that "the actual fraction utilized by inefficient statistics in large samples is obtained by expressing the random sampling variance of efficient statistics as a fraction of that of the statistic in question," and exemplifies the calculation on the estimates of T_1 and T_2.

The review of the second edition of SMRW in the *Journal of the Royal Statistical Society* was considerably more favorable than that of the first, and refers to Fisher's contributions as having "already had, and are still more likely to have in the future, a far-reaching influence on the subject." The review must, however, be read with some caution for possible bias, since the reviewer, J.O. Irwin (1929), although a student of Karl Pearson, was at the time of the review a member of Fisher's department at Rothamsted.

As he had done for the first edition, Hotelling also volunteered a review of the second edition (in fact for each of the first seven editions). By that time he was an associate professor in the Stanford Mathematics Department.

He refers to the book as "this unique work" and concludes with the advice that

> An American using statistics will do well to work through the book, translate it into his own tongue, and change his habits accordingly.

A third review by Pearson (1929) led to a controversy which will be considered in the next section.

2.5 The Assumption of Normality

Egon Pearson reviewed the second edition of Fisher's book in the 8 June 1929 issue of *Nature*. The review was more positive than an earlier review he had written of the first edition, but it also contained the following critical paragraph:

> There is one criticism, however, which must be made from the statistical point of view. A large number of the tests developed are based…on the assumption that the

> population sampled is of the "normal" form. That this is the case may be gathered from a careful reading of the text, but the point is not sufficiently emphasized. It does not appear reasonable to lay stress on the "exactness" of the tests when no means whatever are given of appreciating how rapidly they become inexact as the population sampled diverges from normality. That the tests, for example, connected with the analysis of variance are far more dependent on normality than those involving Student's z (or t) distribution is almost certain, but no clear indication of the need for caution in their application is given.

Fisher was deeply offended; he felt that his honesty had been impugned, and he wrote a blistering reply to *Nature* which has not been preserved. The editor sent it on to Pearson for comment, and Pearson drafted a reply which he showed to Gosset. In defense of the criticism in his review, Pearson quoted from a 1929 paper by the American economist Howard R. Tolley which showed that at least one reader had been misled by Fisher's book. In this paper, Tolley claimed that

> Recently the English school of statisticians has developed formulas and probability tables to accompany them which, they state, are applicable regardless of the form of the frequency distribution. These formulas are given, most of them without proof, in Fisher's book (1925). … If we accept the statements of those who have developed those newer formulas [i.e., Student and Fisher], skew frequency functions and small samples need cause us no further difficulty as far as measurement of error is concerned.…

Gosset offered to write to Fisher in an effort to mediate the dispute and in his letter referred Fisher to Tolley's paper. Fisher, somewhat shaken by Tolley's complete misunderstanding, gave a very conciliatory reply. After some more correspondence back and forth, Fisher, in a letter of 27 June 1929 [80], suggested that Gosset should write to *Nature* in his stead and that he (Fisher) would withdraw his letter.

Fisher's letter to Gosset is unusually revealing and the following paragraphs shed light on Fisher's attitude toward his book:

> In rereading the review, you must have noticed that there was a criticism which must be made on statistical grounds. I take this (and so do you) to mean that there was something wrong with it as statistics; what was wrong? The claim of exactness for the solutions and tests given was wrong, although a careful reader would find that I had kept within the letter of the law by hidden allusions to normality.
>
> …
>
> However important the question of normality may be, it is certainly irrelevant to the book he [E. S. Pearson] was reviewing. The reviewer takes up the position that he could have admitted this if I had mentioned normality more often or in larger type, or if I had made the meaningless claim that the solutions are approximate to those of certain undefined problems.
>
> About my examples I think you are right, they are nearly all very imperfect from the point of view of a purely theoretical treatise. In a practical treatise, I submit that the only question a critic has a right to raise is: – Are the right conclusions drawn, and by the right methods?

Gosset did write, as Fisher had suggested, and his letter was published in the 20 July (1929) issue of *Nature*. After quoting Tolley and absolving Fisher from subscribing to Tolley's views, he continues:

> The question of the applicability of normal theory to non-normal material is, however, of considerable importance. … I have always believed…that in point of fact "Student's" distribution will be found to be very little affected by the sort of small departures from normality which obtain in most biological and experimental work. … We should, however, be grateful

> to Dr. Fisher if he would show us elsewhere on theoretical grounds what sort of modification we require to make when the samples…are drawn from populations which are neither symmetrical nor mesokurtic.

In his reply of August 17, Fisher rejects Gosset's suggestion that he should give some guidance on how to modify the *t*-test for data from non-normal populations for a variety of reasons, but he hints at the existence of distribution-free tests and at taking higher moments into account. He concludes his letter by reiterating that

> On the practical side there is little enough room for anxiety, especially among biologists. … I have never known difficulty to arise in biological work from imperfect normality.

Thus, in his letter of 27 June 1929 [80], Fisher admits that his claim of exactness of various tests was wrong, but he says that he "had kept within the letter of the law by hidden allusions to normality." This suggests a conflicted attitude towards the presentation of this material. On the one hand, he wanted to teach his intended readers, mostly biologists, who would know little mathematics, how to use his methods. For this purpose, the presentation should be as simple and straightforward as possible, avoiding complications that would cause confusion and raise doubts.

On the other hand, he was a scientist and it bothered him to put down statements that were false, even if they held "for all practical purposes." So, to salve his conscience, he would sometimes add a "hidden allusion" that would prove his honesty without disturbing his readers.

A striking example of this strategy can be found in Chapter VI (Section 35). After introducing the transform z of the correlation coefficient, Fisher states:

> The standard error of z is simpler in form
>
> $$\sigma_z = 1/\sqrt{n-3} \quad (n \text{ is the sample size}),$$
>
> and is practically independent of the value of the correlation in the population from which the sample is drawn.

The verb "is" before σ_z seems to claim unambiguously that the formula for the standard error is exact. But if so, σ_z is completely independent of the population correlation, not only "practically independent." The added phrase is a hidden admission of the fact that the formula is only an approximation. (It is noteworthy that in later editions Fisher added the word "approximately" before the formula for σ_z).

As another example of the same strategy, consider Fisher's presentation of the distribution of Pearson's χ^2-statistic for goodness of fit in Chap. 4 (Sect. 4.1). After defining the statistic, Fisher states

> The form of the distribution of χ^2 was established by Pearson in 1900; it is therefore possible to calculate…

This statement gives the impression that χ^2 is one of the exact tests with which the book is chiefly concerned. Fisher does not mention that it is in fact a large-sample test based on the normal approximation to the multinomial distribution.

However, in Example 9 in this section, when testing the goodness of fit of a Poisson distribution, he casually mentions that:

> In applying the χ^2-test to such a series, it is desirable that the number expected should in no group be less than 5, since the calculated distribution of χ^2 is not very closely realized for very small classes.

What is the reader to make of the vague statement, "not very closely realized," when earlier it had been stated unequivocally that the distribution in question was a χ^2–distribution? It is again one of those hidden corrections, this time tucked away in an example separated by several pages from the original statement.

Unlike the later correction regarding the formula for σ_z, Fisher kept the presentation of the χ^2–distribution unchanged throughout all the editions.

2.6 The Impact of Fisher's SMRW

Fisher's SMRW brought together the two testing strands whose starting points were discussed in Chap. 1. Both Pearson's χ^2 and Student's small-sample tests had been extensively developed by Fisher in the intervening years, and the book thus made a wealth of new methods available to laboratory workers. The result was a complete change in statistical methodology.

Although reviewers were slow in recognizing this revolutionary development, the book made its way. The first edition of 1,050 copies was sold out after three years, and the second edition of 1,250 copies in another two. Every two to three years necessitated a new edition, which usually contained some improvements and often additions. The size of the editions steadily increased and the eleventh edition of 1950 ran to 7,500 copies. The last edition, the fourteenth, was published posthumously in 1970 from notes Fisher had prepared before his death in 1962.

In 1972, to commemorate Fisher's death ten years earlier, Oxford University Press brought out a one-volume edition of Fisher's three statistical books, SMRW, "The Design of Experiments," and "Statistical Methods and Scientific Inference," with an introduction by Fisher's coworker and later successor at Rothamsted, Frank Yates. Thus, SMRW is still in print today, 83 years later, a testament to its enduring value.

The enormous impact this book had on the field of statistics is indicated by a number of later developments, of which I shall mention two.

On 10 May 1950, Fisher received a letter from W. Allen Wallis of the University of Chicago[5]:

> This year marks the twenty-fifth anniversary of the publication of *Statistical Methods for Research Workers*. The *Journal of the American Statistical Association*, of which I have

[5] Published in Bennett (1990).

> recently become editor, hopes to mark this event by one or two articles on the character and consequences of that volume. In that connection, I would appreciate your help on two points.
>
> First, can you suggest two or three persons whom I might invite to prepare articles? … The two names that come first to my mind are Hotelling and Cochran.[6]…
> Second, would it be possible for you to prepare a paper for us on the history of SMRW? …

Fisher replied that

> I am sure that the choice of Hotelling and Cochran is an excellent one … . They are, however, both professional mathematicians, and the book was, of course, written not for mathematicians but for practitioners, who, insofar as they understand their fields of application, are good judges of the kind of statistics which aids them in their work. They constitute, I believe, the real and ultimate judges of such a book as mine, though, of course, the majority of them are rather inarticulate in mathematical circles.
>
> I should suggest, therefore…that others from one or more fields of application should be invited to show what the book or the ideas which it was intended to express have done for their own subjects.

In the event, the March 1951 issue of the Journal carried four articles on SMRW. Their authors were Frank Yates from Rothamsted; the mathematician, statistician, and economist Harold Hotelling; the chemist Jack Youden; and the biologist Kenneth Mather. Together, they provided the comprehensive view that Fisher had hoped for. Wallis's second suggestion of a paper by Fisher himself unfortunately was not realized.

A second example of the stature the book had attained can be found in a book published in 2005, "Landmark Writings in Western Mathematics, 1640–1940," edited by Grattan-Guinness. Each of its 77 chapters is devoted to a pathbreaking book or paper in pure or applied mathematics. The subject of Chapter 67 by the Cambridge Statistician A. W. F. Edwards is Fisher's SMRW.

The chapter begins with some details of the publication history of the book, including a list of its translations into French, German, Italian, Spanish, Japanese, and Russian. It then provides sections on "The Author" and the "Writing of Statistical Methods." The third section summarizes the contents of the first edition, and this is followed with a fairly detailed account of the later editions. The final section (except for a brief epilog) deals with the impact of the book. It begins by stating

That Fisher is the father of modern statistics, no one will dispute…

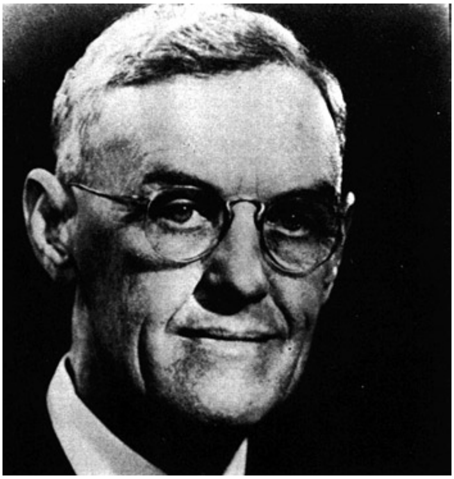
and then imagines how statistics would have developed if Fisher had not written this book.

The author concludes that

> This would have made little difference to the dissemination of Fisher's more theoretical work, which would have still formed the backbone of 20th-Century mathematical statistics; but the effect of that work on applied statistical practice would have been felt more slowly, particularly in biology, including genetics, medicine and the design of agricultural experiments. If, as seems only just, one includes *The Design of Experiments* in the assessment, the impact of *Statistical Methods* throughout the biological sciences was profound and permanent, and from biology the influence spread rapidly into the social sciences.

[6] William Cochran.

2.7 Snedecor's "Statistical Methods": (George Snedecor)



The dissemination of Fisher's ideas set forth in his "Statistical Methods for Research Workers" was greatly furthered in an unexpected way, through a book with the shorter title, "Statistical Methods" (and the subtitle, "Applied to Experiments in Agriculture and Biology"), by a friend and admirer, George Snedecor (1881–1974).

Snedecor, who in 1933 had become the founding director of a statistical laboratory at Iowa State, in 1936 wrote to Fisher that he was working on an elementary text "designed to lead the beginner to an appreciation of your books." Snedecor's text came out in 1937 and covered roughly the same material as Fisher's SMRW: χ^2-tests, the normal one- and two-sample problem, regression, correlation, and the analysis of variance and of covariance. In addition, it included some of the ideas of Fisher's 1935 book on design of experiments, such as factorial experiments, randomized blocks, Latin squares, and confounding.

What distinguished it from Fisher's book was its style. The mode of presentation was much easier, much more user-friendly than Fisher's very terse way of writing.

Much of Snedecor's book is written as a conversation between author and reader. He constantly asks questions, points out to the reader how a new topic is related to ideas covered earlier, and illustrates each idea with numerous examples and exercises. It is also helpful that the material is divided into bite-size sections of usually a page or two.

As an example of the style, consider the beginning of Section 2.2, entitled, "The experiment described. The mean."

> Imagine a newly discovered apple, attractive in appearance, delicious in flavor, having apparently all the qualities of success. It has been christened "King." Only its yielding capacity in various localities is yet to be tested. The following procedure is decided upon. King is planted adjacent to Standard in 15 orchards scattered about the region suitable for their production. Years later, when the trees have matured, the yields are measured and recorded in Table 2.1.

This table, it is pointed out, shows that King is superior in every orchard and that the average difference over the 15 orchards is 6 bushels.

After some preliminary material that makes up Chap. 2, "An experiment designed to compare measurements of individuals," Chap. 3 discusses "Sampling from a normal distribution," and in Sect. 3.8 introduces the *t*-distribution. In this material, the apple example is occasionally used to illustrate the concepts. It resumes its central role in the next section which brings the *t*-test and states:

> At last we have reached the goal of this chapter. The aim is to test the significance of the mean difference, 6 bushels, between the yields of 15 pairs of apple trees.

The difference is found to be significant at the 1% level, and the section ends by asking:

> How great is the difference between the two population yields? We do not know. The best estimate we have is 6 bushels. In addition, we have the convincing circumstantial evidence of significance that King is the better yielder. Most people are satisfied with that information. If you are not, we shall try to give you something to think about in a section to follow (3.11).

It reads like a suspense story and keeps the reader interested.

The book was an enormous success. It has sold more than 200,000 copies in many editions, and was for some years one of the most cited books in the Science Citation Index. Fisher's own books were too difficult for mass appeal. As Fisher's daughter Joan Fisher Box wrote in the biography of her father (p. 313), "It was George W. Snedecor … who was to act as midwife in delivering the new statistics in the United States".

Chapter 3
The Neyman-Pearson Theory

3.1 Start of the Collaboration

How the Neyman-Pearson collaboration came about was told in Sects. 1.1 and 1.4. We are lucky to have several sources that enable us to follow the development of their joint work in some detail. They are a paper by Pearson (1966), "The Neyman-Pearson Story," Neyman's account reported in Constance Reid's (1982) Neyman biography, and Neyman's letters to Pearson.

The first step was Pearson's conversation in the early fall of 1926. Concerning it, Pearson (1966) recounts that after receiving Gosset's letter of 11 May 1926,

> From then on a number of new ideas must have begun to take shape. The possibility of getting a mathematical entry into the problem by specifying a class of alternative hypotheses which should be accepted as "admissible" for formal treatment; the difference between what we later distinguished as "simple" and "composite" hypotheses; the "rejection region" in the sample space; the "two sources of error." These were points which we must have discussed during the autumn of 1926.

Pearson continues,

> It was still necessary to find a principle for determining the choice among possible contours in the sample space in such a way that the hypothesis tested became "less likely" and alternatives "more likely" as a sample point moved outward across them. From rough notes, it seems that the idea of using the likelihood ratio criterion as a method of determining these contours took form in November 1926.

It was then that Pearson wrote up some notes on these ideas and sent them to Neyman. While these notes and Pearson's letters to Neyman have not survived, Neyman's letters, beginning with his reply to Pearson's notes, were preserved by Pearson, and a copy brought to California by Constance Reid for use in her book on Neyman.[1]

[1] Neyman's letters are now in the Manuscripts Collection of the Bancroft Library, University of California at Berkeley: Constance Reid Research Materials, BANC MSS 2008/250, Box 1, Folders # 1–2, 1–3, 1–9.

E.L. Lehmann, *Fisher, Neyman, and the Creation of Classical Statistics*,
DOI 10.1007/978-1-4419-9500-1_3, © Springer Science+Business Media, LLC 2011

Neyman's reaction to Pearson's communication, dated 9 December 1926, shows a complete lack of understanding. He begins his letter by stating:

> I think to have the possibility of testing it is necessary to adopt a principle such as Student's, but it seems to me also that this principle is equivalent to the principle leading to inverse probabilities but it must be generalized, and I think we must have curage[2] enough to adopt this principle, because without it every test seems to be impossible.

And a few sentences later, referring to Pearson's likelihood ratio test for the mean of a normal distribution, he comments:

> What you have done can be expressed in words: wishing to test the probability of hypothesis *A* we have to assume that all hypotheses are a priori equally probable and to calculate the probability a posteriori of *A*.

It is clear from this letter and Pearson's recollections that the notes must have included the principal ideas underlying what was to be the first joint paper: consideration of the alternatives (suggested by Gosset), the two kinds of error, the sample space, rejection regions, and the likelihood ratio test (based on Fisher's maximum likelihood estimate).

Of these, the representation of the data as a point in an n-dimenstional space was really not new since Fisher had used it with great success in the derivation of his small-sample distributions. However, Fisher had applied this idea only in specific situations. In that sense, Pearson's quite general, abstract formulation was an innovation. In fact, Neyman had considerable difficulty with this representation. In his fourth letter (of January 5, 1927), he objected:

> there is some inconvenience in considering a sample as equivalent to the point of the n-dimensional space. The observation of a single point on the surface $\Sigma x_i^2 = \chi^2$ if χ^2 is reasonable, means nothing astonishing, but if this point has the coordinates $x_1 = x_2 = \cdots = x_n$ and if n is considerable – it means certainly some thing, which escapes when we consider only points.

And on the same issue, two days later:

> It seems to me that the idea of the fundamental space and the identity between the sample and a single point, which is very like a single observation is rather unfortunate.

However, after this, he appears to have accepted the geometric representation and, starting with the next letter of January 25, uses it freely. In addition, as Pearson recalls,

> It was not long, I think, before Jerzy accepted the utility of the [likelihood ratio] principle, but in our first joint paper (1928) we agreed to keep the door open by tackling problems in a variety of alternative ways, one of which was based on an inverse probability approach.

[2] At the time, Neyman still had trouble with English, and here and in later quotes I retain his spelling even when it is incorrect.

3.2 Likelihood Ratio Tests

The first joint paper of Neyman and Pearson, "On the use and interpretation of certain test criteria," appeared in two parts of 66 and 32 pages, respectively, in the 1928 volume of *Biometrika*.

Part I begins with a general discussion of the problem of testing the hypothesis A "that the population from which the sample Σ has been drawn is that specified, namely Π." Then it mentions "two distinct methods of approach, one to start from the population Π, and to ask what is the probability that a sample such as Σ has been drawn from it, and the other the inverse method of starting from Σ and seeking the probability that Π is the population sampled." Next, the necessity of considering alternatives to Π is mentioned and the two approaches are compared, with the conclusion:

> the inverse method may be considered by some the more logical of the two; we shall consider first, however, the other solution.

The introduction continues by mentioning the representation of Σ "by a point in a hyperspace whose dimension will depend upon the particular problem considered; and to associate the criteria for acceptance or rejection with a system of contours in this space, so chosen that in moving out from contour to contour hypothesis A becomes less and less probable."

This sounds like inverse probability, but the statement is modified by a somewhat cryptic footnote:

> Here and later the term "probability" used in connection with hypothesis A must be taken in a very wide sense. It cannot necessarily be described by a single numerical measure of inverse probability; as the hypothesis becomes less "probable," our confidence in it decreases, and the reason for this lies in the meaning of the particular contour system that has been chosen.

Finally, a last topic is introduced: the two kinds of error.

(1) Sometimes, when hypothesis A is rejected, Σ will in fact have been drawn from Π.
(2) More often, in accepting hypothesis A, Σ will really have been drawn from [some alternative population] Π'.

After the introduction comes a chapter in which the general ideas discussed so far are applied to the problem of testing the mean of a normal distribution. The authors consider two cases:

(A) The problem of testing that the mean and standard deviation are both known against the alternative that both are unknown.
(B) Testing that the mean is known and the standard deviation unknown against the alternative that both are unknown.

They then introduce the "criterion of likelihood" λ, but only for the case that the hypothesis completely specifies the distribution Π:

$$\lambda = \frac{\text{Likelihood of } \Pi}{\text{Likelihood of } \Pi' \text{ (max)}},$$

where Π' denotes any alternative to Π. This is followed by deriving the λ-test for hypothesis A, of which the authors write:

> Without claiming that this method is necessarily the "best" to adopt, we suggest that the use of this contour system…provides at any rate one clearly defined method of discriminating between samples for which hypothesis A is more probable and those for which it is less probable (sic!). In the appendix they provide tables for carrying out the test, for sample sizes 3 to 50.

Having completed their discussion of hypothesis A, the authors next turn to hypothesis B concerning the subuniverse of normal populations, $M(\Pi)$, with known means a and varying standard deviations. However, here they run into a difficulty:

> B is really a multiple hypothesis. It only becomes precise upon definition of the manner in which σ is distributed, that is to say, upon defining the a priori probability distribution of σ.

They then discuss the difficulty of specifying such a prior distribution, and conclude that "under such conditions the criterion of likelihood will probably be of service. – Corresponding to any [sample point] Σ, we can find the population Π out of $M(\Pi)$ for which the likelihood is a maximum." They thus maximize the likelihood not only under the alternatives as proposed in their λ-test but also under the hypothesis. The result is what they would later call the likelihood ratio test. For hypothesis B, they find it to coincide with Student's t-test.

This treatment of the problem is followed by two others, of which we shall only mention the one labeled "solutions by the inverse method." As the authors say:

> The ordinary method of inverse probability consists in postulating some function $\phi\,(a, \sigma)$ to represent the probability a priori that the sampled population is Π [i.e., the normal distribution with mean a and standard deviation σ]….

However, they then draw back and state,

> The difficulty with this procedure in any practical problem lies in the fact that it is almost impossible to express ϕ in exact terms. We prefer therefore to follow a line of argument which, while really equivalent to the above with ϕ assumed constant, makes use of the principle of likelihood rather than the somewhat vague concept of a posteriori probability.

Formally, the likelihood and the density with constant prior for (a, σ) are of course equal, so the difference is just a matter of interpretation. However, the likelihood interpretation runs into its own difficulty:

> Likelihood as defined by Fisher is a quantity which cannot be integrated. With this in the strict sense we agree…

They integrate anyway, since without this they cannot calculate the probabilities of interest. The resulting procedure is exactly the one a statistician would use who believed in the constant prior density, but the authors refuse to admit that this is what they are doing. Instead, they give it a rather convoluted likelihood interpretation which contradicts the heading of the section: "Solutions by the inverse method."

This rather strange presentation reflects, I believe, a basic disagreement between the two authors. Pearson thought that in the likelihood ratio test, he had found a generally satisfactory approach to the testing problem. On the other hand, Neyman, up to this time, had not abandoned his belief in the inverse method, an approach with which Pearson (perhaps for political reasons) did not want to be identified. The attempt to reconcile these different points of view resulted in a confused and confusing account.

After considering a number of examples, the authors tackle the normal two-sample problem. They mention a number of possible tests, but then settle on their own procedure:

> We may approach the problem by making use of the principle of likelihood and reaching the test given by Fisher. Suppose that we have reason to believe that two samples have been drawn from normal populations with the same standard deviation σ, but that it is necessary to compare the relative probability of two hypotheses (1) that the samples come from identical populations with mean a, and (2) that while identical as to variability, these are two different means a_1 and a_2.
>
> They then derive the likelihood ratio test by maximizing the likelihood both under (1) and (2), and find that the resulting test is a t-test with $n_1 + n_2 - 1$ degrees of freedom.

Two comments on this passage may be in order: (i) although the authors quote Fisher, they deviate from his approach (discussed above in Sect. 2.3) by assuming the two standard deviations to be equal. This is a consequence of their need to clearly state the alternatives; and (ii) they still speak of the relative probability of the two hypotheses, contrary to what they are in fact doing.

The remainder of Part I of this first joint paper extends the likelihood ratio test and the authors' interpretation of the inverse method from the normal to rectangular and exponential distributions. A conclusion section briefly discusses the problem of the robustness of the tests against deviations from the assumed distributional form.

Part II of the paper opens with a section entitled, "An extension of the definition of likelihood." It defines the concepts of simple and composite hypothesis, and then extends the original definition of likelihood ratio, which assumed the hypothesis to be simple, to the composite case as the ratio of the maximum of the likelihood under the hypothesis to that under the alternatives. This is followed by a reminder that in Part I this ratio had been obtained for the normal case and found to be Student's t.

Regarding this extended definition, the authors comment:

> The value of this criterion in the case of testing composite hypotheses will perhaps be questioned. It may be argued that it is impossible to estimate the probability of such a hypothesis without a knowledge of the relative a priori probabilities of the constituent simple hypotheses. But in general it is quite impossible even to attempt to express our a priori knowledge in exact terms.

They conclude that,

> we are inclined to think that this ratio of maximum chances or frequencies of occurrence provides perhaps the best numerical measure that can be found under the circumstances to guide our judgment.

The rest of Part II is concerned with testing goodness of fit for a multinomial distribution. The authors derive the likelihood ratio test and show that when higher-order terms are neglected it agrees with Pearson's χ^2 test.

This long two-part paper is a great achievement. It introduces the consideration of alternatives, the two kinds of error, and the distinction between simple and composite hypotheses. In addition, of course, it proposes the likelihood ratio test. This test is intuitively appealing, and Neyman and Pearson show that in a number of important cases it leads to very satisfactory solutions. It has become the standard approach to new testing problems.

3.3 A New Approach

Let us now return to the Neyman-Pearson correspondence. From the first letter in 1926 through January 1931, these letters were primarily concerned with likelihood ratio tests and the resulting publications, the 1928 paper and a number of follow-up papers published between 1929 and 1931.

Both Neyman and Pearson believed that this approach was the right answer. However, while Pearson felt that its intuitive appeal, bolstered by its success in the situations in which they had examined it, was enough of a justification, Neyman thought that something was lacking. If these tests were really the best, it seemed to him, it should be possible to prove this.

The first time he raised this issue was in a letter to Pearson of 1 February 1930. After some news about his living arrangements and the proposal of a joint book, he breaks in with:

> I have now a rather exciting point concerning the two tests z and t[3] (supposing I am not wrong and they are both normally distributed). It seems that we can have an experimental proof [his underlining, here and below] that the principle of likelihood "est fait pour quelque chose."
>
> (1) If the two sampled populations have the common variance and we test the hypothesis that $a_1 = a_2$, the appropriate test is the t-test.
>
> (2) If we do not know anything about the variances, the appropriate test of the same hypothesis is the z-test.
>
> As both of them control the errors of rejecting a true hypothesis, the difference in results of their application will consist in different frequency of accepting a false one. And this can be experimentally proved.

He next describes the sampling experiment he proposes, and concludes the paragraph with: "In any case, the sampling experiment is interesting, isn't it?"

And then he adds an afterthought which does not really follow from what has gone before:

> If we show that the frequency of accepting a false hypothesis is minimum when we use λ tests, I think it will be quite a thing!"

[3] These are the tests of hypotheses *A* and *B* of their 1928 paper, discussed in Section 3.2 above.

What happened next Neyman described to Constance Reid (Reid, 1982, p. 92), who reports as follows:

> The first step…came suddenly and unexpectedly in a moment which Neyman has never forgotten. Late one evening in the winter of 1930, he was pondering the difficulty in his little office. Everyone else had gone home, the building was locked. He was supposed to go to a movie with Lola [his wife] and some of their friends, and about eight o'clock he heard them outside calling for him to come. It was at that moment that he suddenly understood.

Neyman's letter to Pearson (dated 20 February 1930) concerning this insight is more prosaic. He writes:

> At present I am working on a variation calculus problem connected with the likelihood method. The results already obtained are a vigorous argument in favour of the likelihood method. I considerably forgot the variation calculus and until the present time I have only results for samples of two. But in all cases considered I have found the following.

We shall give this statement in full since it is the first formulation of what was to become an elaborate and influential theory of hypothesis testing.

> We test a simple hypothesis H concerning the value of some character $a = a_0$, and wish to find a contour $\varphi(x_1, \dots, x_n) = c$ such that
>
> (1) the probability $P(\varphi_0^a)$ of a sample point lying inside the contour (which probability is determined by the hypothesis H) is equal
>
> $$P\left(\varphi_0^a\right)=\varepsilon,$$
>
> where ε is a certain fixed value, say 0.01. (This is for controlling the errors in rejecting a true hypothesis) and
>
> (2) that the probability $P(\varphi_1^a)$ determined by some other hypothesis H' that $a = a_1 \neq a_0$ of sample lying inside the same contour be maximum.
>
> Using such contours and rejecting H when Σ [the sample point] is inside the contour, we are sure that the true hypothesis is rejected with a frequency less than ε, and that if H (the hypothesis tested) is false and the true hypothesis is, say, H', then most often the observed sample will be inside φ = const. and hence the hypothesis will be rejected.
>
> I feel you start to be angry as you think I am attacking the likelihood method! Be quiet! In all cases I have considered the φ = const. contours are the λ contours!

In the next letter, dated March 8, Neyman suggests that he and Egon must "fix a certain plan, as we have lot of problems already started and then left in the wood." He lists several such problems, among them:

> to finish what I have started to do with the variation calculus. You will understand it in a moment. To reduce for a given level the errors of rejecting a true hypothesis, we may use any test. Now we want to find a test which would 1) reduce the probability of rejecting a true hypothesis to the level $\leq \varepsilon$ and 2) such that the probability of accepting a false hypothesis should be minimum. – We find that if such a test exists, then it is the λ-test. I am now shure [sic] that in a few days I shall be ready. This will show that the "λ principle" is not only a principle but that there are arguments to prove that it is really "the best test."

He did indeed complete the argument soon thereafter, and on March 24 sent Pearson the proof of what now is known as the Neyman-Pearson Fundamental Lemma. However, instead of continuing with the correspondence, we shall now turn to the paper in which the authors present their account of these results.

3.4 The Core Results

The paper in question, published in 1933, was entitled, "On the Problem of the Most Efficient Tests of Statistical Hypotheses." After a brief historical introduction, the authors announce a very novel point of view: "Without hoping to know whether each separate hypothesis is true or false," they write,

> we may search for rules to govern our behavior with regard to them, in following which we insure that, in the long run of experience, we shall not be too often wrong.

This behavioral philosophy later became very influential. It also led to a rancorous controversy with Fisher, which we shall consider in Sect. 4.5.

Toward the end of the introduction, the authors comment on the relation of the present approach to their earlier work. They note that,

> In earlier papers we have suggested that the criterion appropriate for testing a given hypothesis could be obtained by applying the principle of likelihood. This principle was put forward on intuitive grounds after the consideration of a variety of simple cases. It was subsequently found to link together a great number of statistical tests already in use, besides suggesting certain new methods of attack. It was clear, however, in using it that we were still handling a tool not fully understood, and it is the purpose of the present investigation to widen, and we believe simplify, certain of the conceptions previously introduced.

The next section is called, "Outline of a General Theory," and defines a number of new concepts, among them simple and composite hypotheses, alternatives to the hypothesis, and the two kinds of error. The authors point out that "it is very easy to control errors of the first kind," that is, the probability of false rejection, and then formulate the basic criterion for selecting the best rejection region w.

This criterion is [their italics]:

> *To pick out from all possible* w_0 *regions for which* $P_0(w) = \varepsilon$, *that region,* w_0, *for which* $P_1(w_0)$ *is a maximum…*; this region (or regions if more than one satisfy the condition) we shall term the Best Critical Region for H_0 with regard to H_1.

A last important comment concerns the fact that

> the best critical region with regard to H_1 will not [necessarily] be the best critical region with regard to H_2.

"But it will be shown below, "the authors write,

> that in certain problems there is a common family of best critical regions for H_0 with regard to the whole class of admissible alternative hypotheses. In these problems we have found that the regions are also those given by using the principle of likelihood, although a general proof of this result has not so far been obtained when H_0 is composite.

> In the problems where there is a different best critical region for H_0 with regard to each of the alternatives…, some further principle must be introduced. …We have found here that the region picked out by the likelihood method is the envelope of the best critical regions with regard to the individual hypotheses of the set. This region appears to satisfy our intuitive requirements for a good critical region, but we are not clear that it has the unique status of the common best critical region of the former case.

This is an enormously ambitious program, a good part of which is carried out in the remainder of the paper. But the importance of the paper goes beyond the specific results obtained here. For the first time it states the aim of statistical theory to be the systematic search for optimal procedures. Much of the theory developed during the next decades was directed toward this end.

After outlining the general theory, the paper in the next section deals with the case of simple hypotheses and brings the statement and proof of the basic result, now known as the Neyman-Pearson Fundamental Lemma. It states that for testing a simple hypothesis against a simple alternative, the test that at a given level maximizes the probability of rejection is the likelihood ratio test at that level.

The result is illustrated by a number of examples. In particular, it is shown that for testing the hypothesis that the mean of a normal distribution is a_0 when the standard deviation is known, there exists a common best critical region against the alternatives $a > a_0$ (namely to reject when the sample mean is too large) and a different best critical region against the alternatives $a < a_0$. A corresponding result is seen to hold for testing of the standard deviation when a is known. The section concludes with consideration of the corresponding problems for rectangular distributions.

In the next section, the authors turn to the testing of composite hypotheses and here introduce a new condition. They mention that consideration must be given again to the two sources of error, and then state that,

In the first place it is evident that a necessary condition for a critical region, w, suitable for testing [a composite hypothesis] $H_0{}'$ is that

$$P_o(w) = \text{constant} = \varepsilon.$$

For every simple hypothesis [contained in $H_0{}'$]. …If this condition is satisfied, we shall speak of w as a region of "size" ε, similar to W [the sample space] with regard to the [nuisance parameters].

The remainder of the section is concerned with characterizing the totality of similar regions and the one that maximizes the rejection probability against a particular alternative. These derivations have since then been simplified and will not be discussed here.

The results are illustrated in a number of examples. In particular, it is shown that for testing the mean of a normal distribution when the variance is unknown, Student's t-test is the best similar region against one-sided alternatives. An analogous result is obtained for testing the variance when the mean is unknown. Finally, these results are extended to the two-sample case.[4]

[4] It should be noted that the terms "power," "uniformly more powerful," and "uniformly most powerful" are not used in this paper. They were only introduced in the author's following paper, also of 1933.

At this point, there is still a large gap in their work: what to do if there is no common best critical region? Neyman raises the question in a letter of 17 August 1931. He labels this portion of his letter "A" and writes:

> I am considering the question when there is no best critical region with regard to a given class of admissible hypotheses. What region should we then choose? I take the most simple case when the whole set of admissible hypotheses can be divided into two classes such that to each of them corresponds a "best critical region."

He then considers the problem of testing that the variance of a normal distribution σ^2 is 1 when the mean is 0, and points out that against the alternatives $\sigma^2 > 1$, there exists a best critical region

$$\Sigma x_i^2 > \chi_1^2,$$

while against alternatives $\sigma^2 < 1$, it is of the form

$$\Sigma x_i^2 < \chi_2^2.$$

But, if the hypothesis is to be tested against the whole set of alternatives $\sigma^2 \neq 1$, how should the limits χ_1^2 and χ_2^2 be determined, he wonders. As one possibility, he mentions to choose these constants so that the probability in each tail is $\varepsilon/2$, and then he breaks off and excitedly calls out:

> I think I have got the point! [His underlining.] Suppose we have to test a simple hypothesis H_0 with regard to a class of alternatives C with no common *B*. C. R. It would be no good to use a critical region w, having the following property: the class of alternatives contains a hypothesis, say H_1, such that ε_1 [the probability of rejection under H_1] is $< \varepsilon$. In fact, doing so we shall accept H_0 with larger frequency when it is false (and the true hypothesis is H_1) than when it is true. …
>
> Therefore, the good critical region w_0 should be chosen in such a way that [the probability of rejection under H_1 is ≥ than that under H_0].

He then determines χ_1^2 and χ_2^2 satisfying this condition for the problem *A* of testing $\sigma^2 = 1$ and notes that it "gives the maximum likelihood solution."

As a second problem, he considers testing the values of both mean *a* and the variance σ^2 of a normal distribution, i.e., that $a = 0$, $\sigma = 1$, and to his dismay he finds:

> it seems that if we use λ contours [i.e., the likelihood ratio test] to test the hypothesis H_0: $a = 0$, $\sigma = 1$, it shall accept it more often when $a = 0$, $\sigma = 1.1$, say, than when H_0 is true. With other words: the true hypothesis will be rejected more often than some of the false ones. … I should like to be wrong somewhere. If it is all correct, then the principle of maximum likelihood seems to lose its generality. Exceedingly interesting to know what are the conditions of its applicability?

This new development was not published until 1936 in a paper entitled, "Contributions to the theory of testing statistical hypotheses." By now the authors freely use the terms "power of a test," "power function" and "uniformly most powerful test," and they define a test to be unbiassed [sic] if its power against all admissible alternatives is ≥ its value under the hypothesis. They hope that there might exist a uniformly most powerful unbiassed test but begin with the simpler problem of finding what they call an unbiassed critical region of type *A*, namely which maximizes the power locally. Since the power function of an unbiased test of the hypothesis $\theta = \theta_0$ has a minimum at θ_0 and therefore a first derivative = 0, the type A test is obtained by maximizing the second derivative at θ_0.

The authors find a general method for determining the type A tests and illustrate it with the testing of the normal mean, and of the normal variance when the other parameter is known, against two-sided alternatives.

Finally, they define a test as being of type A_1 if the derivative of its power function is zero at $\theta = \theta_0$ and, subject to this condition, the test maximizes the power against all admissible alternatives. They point out that this property is close to but not identical with being uniformly most powerful unbiased, and show that the type A region for testing the variance of a normal distribution is also of type A_1.

Two years later (1938), this paper was followed by a continuation containing parts II and III under the same title. It was to be Neyman's and Pearson's last joint paper. In part II, they continued the investigation of the existence and structure of type A and A_1 tests, and in part III considered testing hypotheses specifying more than one parameter. This latter investigation, begun in a 1937 paper by Neyman, introduced tests of type C, which again concerned unbiased tests satisfying certain local desiderata.

Since Neyman and Pearson introduced tests of type A and C, one naturally wonders whether they also had a concept labeled type B. Such tests were in fact introduced, but in a 1935 paper of which Neyman was the sole author. It extended the type A property to situations in which H: $\theta = \theta_0$ was being tested in the presence of additional parameters.

However, neither Neyman alone nor the two authors jointly extended this typology to type B_1 tests, i.e., uniformly most powerful unbiased tests of $H{:}\theta = \theta_0$ in the presence of nuisance parameters. This extension was carried out by Henry Scheffé in 1942. Finally, type C_1 tests were never mentioned for the simple reason that uniformly most powerful unbiased tests of hypotheses specifying more than one parameter do not exist. As Neyman wrote in one of his late letters, such hypotheses require a new principle. Such a principle – the restriction to invariant tests – was later introduced by other authors, but that is no longer part of the Neyman-Pearson story.

3.5 The Process of Joint Work

So far this chapter has been concerned with the ideas and results of Neyman's and Pearson's joint work. In the present section, we shall consider the circumstances and the process of this collaboration, which started in 1926 and ended in 1938 when Neyman left England for California. It resulted in a total of ten papers, the most important five of which were discussed in the preceding sections.

The collaboration falls into two quite distinct parts. In the early stages, the important ideas, including in particular that of the likelihood ratio principle, all come from Pearson. In fact, Neyman frequency misunderstands them, and continually tries to interpret them in terms of inverse probability.

On the other hand, Pearson is sold on the likelihood ratio principle, which is intuitively appealing and which seems to give reasonable solutions in the cases on which they try it out. But for Neyman, as he is gradually catching on, intuitive appeal is not enough. If the principle is really as good as it appears to be, there ought to be logical justification.

And then one day in early 1930, he sees the light. Since there are two sources of error, one of which is being controlled, the best test is the one minimizing the other one. And from then on, it is Neyman who has the new ideas and Pearson is the reluctant follower. Neyman formulates, and shortly thereafter proves, the Fundamental Lemma and realizes that in some special cases there exist what they later call uniformly most powerful tests. These turn out to coincide with the likelihood ratio tests.

These concepts and results were written up in the paper, "Most efficient tests…," but its publication caused the authors some difficulties. They seemed to have ruled out *Biometrika*, presumably because Karl Pearson had little sympathy or interest in small-sample theory. Neyman then suggested publishing it in Poland, but this would have condemned it to obscurity. Eventually, they decided to submit it to the *Philosophical Transactions of the Royal Society*, but with considerable doubt as to how it would be received. However, the *Transactions* did accept the paper and it appeared in 1933.

The year 1933 also saw the retirement of Karl Pearson. His position was then split: Egon was appointed to succeed him as Head of the Department of Statistics, while Fisher became Director of the Galton Laboratory. This affected Egon's work with Neyman in two ways. On the one hand, his new administrative duties left the younger Pearson much less time for research. On the other, the following year he was able to offer Neyman a position in his department.

Up to that time, after Neyman's two-year fellowship during 1925–1927, first in London and then in Paris, Neyman had held a number of different positions in Poland and much of his energy had been devoted to making enough money to support himself and his wife Lola. In June 1932, he wrote to Egon:

> I simply cannot work. The crisis and the struggle for existence [sic] takes all my time and energy. Besides, I am not sure that next year I shall not [be] obliged to take some job, I do not know where – in trade perhaps, selling coal or handkerchiefs.

With a regular position in England, this was all changed.

The difficulty of finding a suitable place to publish their 1933 paper, and Neyman's becoming a member of Egon's department, served to revive an old proposal of Neyman's. In a letter of December 1929, after suggesting that their paper be published in the *Proceedings of the Polish Academy* in Krakow, he had added:

> I have also another idea, but I do not want to express it in a definite form unless I shall find a firm support from you. In Poland every scientific laboratory is trying to publish its own if not journal then something of that sort – memoirs, etc. I think that it is not a good idea because there are so many different journals treating the same questions and that very many even valuable papers disappear – you cannot have or even know about all these journals.
>
> With mathematical statistics of course it is a quite different thing, as there are only very few journals consecrated to this science. Should I try? It could be not a regular edition – well,

> I think that perhaps some time – not at present – we want a propaganda – thus a journal having readers – all I have written about it is rather nonsense. Shall we accept the Academy?

The idea of a journal of their own was revived after the difficulties and uncertainties of the 1933 paper, and in 1936 led to a new publication:

> Statistical Research Memoirs
> Edited by
> J. Neyman and E.S. Pearson
> Volume I

It was published by the Department of Statistics, University of London, University College, and was dedicated to the memory of Karl Pearson, who had died earlier that year. The dedication read:

> To the memory of our Professor
> Karl Pearson
> 27 March 1857 – 27 April 1936
> Founder of mathematical statistics, originator of its various applications, teacher and inspirer of innumerable research workers of many nations and races.

In the Foreword the editors explained:

> It is widely felt that in spite of the existence of a large number of special problems for which perfect solutions exist, statistical theory in general in its present state is far from being completely satisfactory from the point of view of its accuracy. It is the ambition of the Department to contribute towards the establishment of a theory of statistics on a level of accuracy which is usual in other branches of mathematics.

The Foreword concludes with the statement that

> The Statistical Research Memoirs will contain only papers prepared in the Department of Statistics; while the series is not strictly periodical, it is hoped that a volume of over 150 pages will be issued about once a year.

The first volume ran to 161 pages and contained seven articles. The first of these was Part I of the "Contributions to the theory of testing statistical hypotheses," dealing with type A and type A_1 regions discussed in Sect. 3.4.

Parts II and III appeared in 1938 in the second volume of the Research Memoirs, which contained nine papers, including two by Pao-Lu Hsu and one by William Feller.

This second volume was also to be the last, and was in fact Neyman's and Pearson's last collaborative effort. In August 1938, Neyman left for California, and this marked the end of their joint work.

3.6 Retrospective

During the more-than-ten years of working together, the two authors' relation grew to a close friendship. This development can be seen in the way Neyman signed his letters. During the first two years, he signed them with "yours sincerely" and later "yours ever, J. Neyman."

In February 1929, they switched to first names and Neyman addressed his letters to "Dear Egon" and signed them "yours ever, Jurek." By the end of the year, Neyman signed off with "love from us both," once even "love and hearty kisses," and "yours 'very much' ever."

It thus may seem surprising that their collaboration did not survive their geographical separation. After all, during much of their early work Pearson had been in England and Neyman in Poland, and the collaboration was carried out mainly through correspondence. One circumstance that made collaboration more difficult this time was that Pearson had many administrative duties as head of his department and, since 1936, also as managing editor of *Biometrika*. But it turns out that there was a more basic reason, which Pearson himself spelled out in a letter to Neyman of 12 October 1976. This letter was in response to Neyman's sending him the draft of a paper "Frequentist Probability and Frequentist Statistics," a revised version of which was published in vol. 36 of *Synthèse*.

Pearson wrote:

> I received your letter of 24th with your Synthèse paper, and was so delighted that I have started devouring it, although only received about five hours ago! Shall I briefly tell you why? Will comment in detail shortly, but this is a quick reaction. There is a tale or fable about us. [This and the following are underlined in the original.]
>
> From 1926-36 we were working together in excited co-operation. My clumsily defined ideas, sharpened by your mathematical formulation, and we went on and on together, until by about 1936-37 we had solved between us what seemed the basic problems, and so found a statistical philosophy. But the time came when to find new mountains to scale, you were forced to tackle more and more mathematically complex problems – tests of "Type" A_1 or B_2 etc., etc., and I began to lose interest because I was always aiming at attacking types of problems with probability tools which seemed to get fairly simply into gear with the way which the human reason worked. And "Types" A_x, B_y, etc., etc., seemed to me stepping out of this field. Then you rightly went to the U.S. and war came. And I suppose we both turned our statistically trained minds to different kinds of jobs: bombs, A.A. shells and what not.

As we saw earlier, Pearson had initiated the collaboration by coopting Neyman into helping him with his program. Now he terminated it since he had "lost interest" in the direction Neyman was going. In fact, as we saw earlier, he had never really been fully sympathetic with Neyman's optimality program, but would have happily stopped with the likelihood ratio principle and its intuitive appeal.

Pearson's letter quoted above continues by recalling a visit by Neyman in 1950 to give a series of lectures to Pearson's group. The lectures were a great disappointment to Pearson. He had asked Neyman to talk about "where, after war-time experience and years in Berkeley, you had got to from the joint foundation of 1926–1936."

But instead, Neyman had talked about applied work he had been doing and of political turmoil at the University of California "so that," Pearson wrote, "I, let alone my students, got no insight into what you were after all those years thinking about N and P." What Pearson did not realize is that Neyman's interests had also shifted and that in Berkeley he was no longer working on the N.P. theory.

The London visit had also been a disappointment to Neyman. As he reported to Constance Reid (1982, p. 223), "the 'old resonance' between [him and Egon] had seemed to be lacking."

Now, after Pearson's receipt of Neyman's *Synthèse* paper, this resonance seems to have been reestablished. As Pearson writes toward the end of his October letter:

> Well now, this paper I have from you today seems to be providing just what I want, the present form of the old J.N. who, though in different style, may have moved on to roughly where the old E.S.P. has moved!

The letter is followed by eight pages of notes in which Pearson comments on Neyman's paper. The correspondence continues after Pearson receives a copy of the published version of the paper. Pearson thanks Neyman in a long letter of February 12, 1978, in which he includes some rough notes that he had made after receiving Neyman's first letter of 9 December 1926 discussed in Sect. 3.1 above. Neyman's reply is not available but was answered by Pearson in another letter of March 10, 1978. In this letter, Pearson reviews some of the work the authors did separately after their 1928 paper.

I shall here quote only the following comments on one of these papers by Neyman, which throws light on the attitudes of the two authors regarding the approach based on inverse probability. Pearson writes:

> The eighth paper in your Early Statistical Papers[5] series was of course the one you presented at the I.S.I. meeting in Warsaw. I have eight letters which you wrote to me during February and March 1929, trying to persuade me to put my name as a joint author. But you had introduced an a priori law of probability…, and I was not willing to start from this basis. True we had given the inverse probability as an alternative approach in our 1928 Part I paper, but I must in 1927-28 still have been ready to concede to your line of thought. However, by 1929 I had come down firmly to agree with Fisher that prior distributions should not be used, except in cases where they were based on real knowledge, e.g., in some Mendelian problems. You were disappointed, but accepted my decision; after all, the whole mathematical development in the paper was yours.

Eventually, Neyman too abandoned his interest in the inverse approach.

By 1935, he opens a paper, "Sur la vèrification de hypothèses statistiques composées" with the sentence [in my translation of the original French]:

> One knows that the problem of the verification of hypotheses has been treated since Thomas Bayes (1763). The solutions we obtained depended on probabilities *a priori*. Since these are generally unknown, one is obliged to make arbitrary hypotheses about them which make the results inapplicable to practical problems.

His conviction of the inapplicability of the inverse method had by then become a fundamental part of his statistical philosophy, from which he never wavered.

The 1928 and 1933 papers of Neyman and Pearson discussed in the present chapter, exerted enormous influence. The latter initiated the behavioral point of view and the associated optimization approach. It brought the Fundamental Lemma and exhibited its central role, and it provided a justification for nearly all the tests that Fisher had proposed on intuitive grounds.

[5] The paper in question was entitled, "Contributions to the theory of certain test criteria," and was reprinted in the volume, "A Selection of Early Statistical Papers of J. Neyman," Univ. of California Press (1967).

On the other hand, the applicability of the Neyman-Pearson optimality theory was severely limited. It turned out that optimum tests in their sense existed only if the underlying family of distributions was an exponential family (or, in later extensions, a transformation family). For more complex problems, the earlier Neyman-Pearson proposal of the likelihood ratio test offered a convenient and plausible solution. It continues even today to be the most commonly used approach.

Chapter 4
Fisher's Dissent

4.1 Fisher's Early Reaction

The preceding two chapters describe the two stages of the founding of the field of hypothesis testing. The first stage consisted of Fisher's development of a cohesive methodology of basic tests under the assumption of normality. It was followed by the Neyman-Pearson theory of optimal tests, which as its major application provided a justification of the tests Fisher had proposed on intuitive grounds. What, one wonders, was Fisher's reaction to this new perspective.

Neyman approached Fisher about his and Pearson's work on optimal tests in a letter of February 9, 1932 (published in Bennett (1990), p. 189):

> … I do not know whether you remember what I said, when being in Harpenden in January 1931, about our attempts to build the theory of "the best tests." Now it is done more or less. It follows from the theory that the "Student's" test, the "*t*-test" for two samples and some tests arising from the analysis of variance are the "best tests," that is to say, that they guarantee the minimum frequency of errors both in rejecting a true hypothesis and in accepting a false one. Certainly they must be properly applied.
>
> Presently Dr. Pearson is putting all the results in order. They will form a paper of considerable size. We would very much like to have them published in the Philosophical Transactions, but we do not know whether anybody will be willing to examine a large paper and eventually present it for being printed. The paper contains much of mathematics and not all the statisticians will like it just because of this circumstance.
>
> We think that the most proper critic are you, but we don't know whether you will be inclined to spend your time reading the paper… .

To this appeal, Fisher replied on February 12:

> I should be very much interested to see your paper on "the best tests," as the whole question of tests of significance seems to me to be of immense philosophical importance, and the work you showed me was surely of great promise. It is quite probable that if the work is submitted to the Royal Society, I might be asked to act as referee, and in that case I shall certainly not refuse.

Fisher not only read the paper, but read it sufficiently carefully to find a small mathematical error (the omission of a condition required for the validity of a statement).

E.L. Lehmann, *Fisher, Neyman, and the Creation of Classical Statistics*,
DOI 10.1007/978-1-4419-9500-1_4, © Springer Science+Business Media, LLC 2011

When the paper appeared in 1933, the omission was corrected, and a footnote acknowledged that, "We are indebted to Dr. R. A. Fisher – for kindly calling our attention to the fact that we had originally omitted to refer to this restriction."

A year after this publication, Fisher published a remarkable paper, "Two new properties of mathematical likelihood," which – in addition to other results – made an important contribution to the Neyman-Pearson theory.

The paper falls into two quite different parts. The first part is concerned with sufficient statistics, and derives the factorization criterion characterizing them. It also argues that the existence of a real-valued sufficient statistic implies that the underlying distributions form what is now known as a one-parameter exponential family. (Rigorous statements and proofs by later authors confirmed that this claim is essentially correct).

Fisher then goes on to discuss the implications of these results for the Neyman-Pearson theory. It is, I believe, the only time that he took this theory seriously and even used its terminology. He wrote:

> The property that where a sufficient statistic exists, the likelihood, apart from a factor parameter to be estimated, is a function only of the parameter and the sufficient statistic, explains the principal result obtained by Neyman and Pearson in discussing the efficacy of tests of significance. Neyman and Pearson introduce the notion that any chosen test of a hypothesis H_0 is more powerful than any other equivalent test, with regard to an alternative hypothesis H_1, when it rejects H_0 in a set of samples having an assigned aggregate frequency ε when H_0 is true, and the greatest possible aggregate frequency when H_1 is true.

Fisher continues with a verbal statement of the Fundamental Lemma, and then reminds the reader:

> The test of significance is termed uniformly most powerful with regard to a class of alternative hypotheses if this property [i.e., maximum power] holds with respect to all of them.

He shows that tests with this property require the existence of a real-valued sufficient statistic. And although he does not point this out, it now follows from his earlier results that uniformly most powerful tests exist only when the distributions constitute a one-parameter exponential family.

It is difficult from this discussion to assess how Fisher felt about the Neyman-Pearson theory. The tone is neutral, uncommitted. On the one hand, it is not as positive as in his earlier letter to Neyman when he asserted that the work was of "great promise." On the other hand, it is not as negative as, for example, in a 1951 letter to William Hick (Bennett 1990, p. 144), in which he referred to the same work as "that unnecessarily portentous approach to testing of significance represented by the Neyman and Pearson critical regions, etc.," to which he adds, "In fact, I and my pupils throughout the world would never think of using them."

It appears that Fisher's 1934 paper was written just at the point at which his attitude was beginning to change.

The second part of the paper is not directly concerned with hypothesis testing (although it has important applications to that problem too), and will be mentioned here only briefly. This section is entitled, "A second class of parameters for which estimation need involve no loss of information." In the opening paragraph, this class is described as occurring when,

> although the sets of observations which provide the same estimate differ in their likelihood functions, and therefore in the nature and quantity of the information they supply, yet when samples alike in the information they convey exist for all values of the estimate and occur with the same frequency for corresponding values of the parameter.

This definition is not easy to understand but becomes clear in light of the two examples discussed by Fisher: location families and location-scale families.

Fisher is considering situations that are invariant under a group of transformations.

The samples that are alike in the information they convey are sample points that can be transformed into each by a member of the group (i.e., lie on the same orbit) and the corresponding values of the parameter are the orbits under the group induced in the parameter space.

It is a striking example of Fisher's genius that at this early stage he discovered the two situations, exponential and transformation families, which permit reduction of the data to a space of fixed low dimensions, independent of the sample size. These two situations later constituted the basis for optimality theories of both testing and estimation.

4.2 The Fourfold Table

Fisher disagreed with the Neyman-Pearson theory not only theoretically but his approach led to different tests in a number of cases, which caused much controversy. One of these concerned the testing for independence in a 2×2 table. Although his avowed purpose in writing "Statistical Methods" (SMRW) in 1925 was to provide users with an accessible account of the new small-sample methods for significance testing, it fell short of this goal in one important respect. For the problems of independence and homogeneity in contingency tables, it presented the Pearsonian large-sample χ^2 methods (although as corrected by Fisher). The reason was that at the time, he had not yet developed an exact small-sample approach for these problems.

Fisher filled this gap in the fourth edition (of 1932) of SMRW, with a short section on, "The exact treatment of a 2×2 table." He considers the case of two independent series of respectively $a+b$ and $c+d$ binomial trials and the hypothesis that the success probability in the two series has a common value p. Since p is unknown, he restricts attention to samples in which not only the marginal totals $a+b$ and $c+d$ are fixed, but also the totals $a+c$ and $b+d$. He now points out the crucial fact that the conditional probability of a sample (a, b, c, d) given these fixed marginal totals is independent of p and is given in fact by

$$(a+b)!(c+d)!(a+c)!(b+d)!/[n!a!b!c!d!]$$

where $n = a+b+c+d$. This known distribution can then be used as a basis for a test, which has been named Fisher's exact test. (The test was about simultaneously proposed also by Irwin (1935) and Yates (1934). The latter, however, acknowledges that the idea of carrying out the test conditionally on the marginal totals was suggested to him by Fisher).

Fisher returns to the problem in Example 1 of a paper, "The logic of inductive inference" (1935). He motivates the restriction to the conditional distribution by the following argument:

> To the many methods of treatment hitherto suggested for the 2 × 2 table, the concept of ancillary information [introduced by Fisher in 1925 [42] suggests this new one. Let us blot out the contents of the table, leaving only the marginal frequencies. If it be admitted that these marginal frequencies by themselves supply no information[1] on the point at issue, namely, as to the proportionality of the frequencies in the body of the table, we may recognize the information they supply as wholly ancillary; and therefore recognize that we are concerned only with the relative probabilities of occurrence of the different ways in which the table can be filled in, subject to these marginal frequencies… .

Not everyone agreed with Fisher's conditional approach. For example, George Barnard proposed an alternative unconditional test in which the level was controlled by considering the maximum probability of an outcome, maximized with respect to the unknown common value under the hypothesis. (Barnard sent Fisher a copy in 1945 (Bennett 1990, pp. 2–4). See also Barnard 1947.) The paper claimed that,

> Other things being equal, the "power" of a test, in the sense of Neyman and Pearson, will increase with the "volume" of the rejection region chosen. In this sense we can say, roughly, that the maximum condition [i.e., Barnard's test] secures that our test should be as powerful as possible, consistent with validity [i.e., control of the level].

In his reply of the same year, Fisher refers "to the case considered by Barnard [this was actually not in Barnard's paper; see Morris De Groot's 1988 interview with Barnard], in which, using three experimental and three control animals, the experimental animals all die and the control animals all survive." Fisher notes that there are only four tables with the same marginals and their conditional probabilities "are in the ratio 1:9:9:1; or, in other words, the probability of obtaining the most successful outcome by chance is always 1 in 20."

He then points out that with repeated sampling, outcomes with other marginals "will often occur. If it were legitimate to judge the level of significance from the proportion of judgments in the whole series of 'repeated sampling from the same population' [i.e., unconditionally], these cases would be brought in to inflate the denominator." He shows that the maximum probability of the most successful outcome is then 1/64 rather than 1/20. At the 5% level, it would then be possible to include additional points in the rejection region and thereby increase the power of the test. However, Fisher argues:

> In my view, the notion of defining the level of significance by "repeated sampling of the same population" is misleading in the theory of small samples just because it allows of the uncritical inclusion in the denominator of material irrelevant to a critical judgment of what has been observed. In 2 of the 64 cases enumerated above, all animals die or all survive. The fact that such an unhelpful outcome as these might occur, or must occur with a certain probability, is surely no reason for enhancing our judgment of significance in cases where it has not occurred… .

[1] This statement is in fact not completely true, though very nearly so. See Plackett (1977).

Fisher repeats this explanation with only minor changes in a paper of 1948 [222] and again in his book, "Statistical Methods and Scientific Inference" (1956).

Let us now consider the same problem from a Neyman-Pearson point of view. If a test is performed at a conditional level α for all values of the conditioning variables, its level also will be α unconditionally. However, because of the discreteness of the conditional distribution in the case of the fourfold table, the conditional level of Fisher's exact test will vary with α, although it will never exceed α. The resulting unconditional test will be conservative and not have the property of similarity. It was pointed out by K. D. Tocher (1950) that this difficulty could be overcome as follows. Consider all the tables having a given set of marginals and form a conditional critical region by including them, one by one, according to their desirability for rejection of the hypothesis. Continue this process until a level $\alpha_{-} < \alpha$ is reached such that inclusion of the next table would result in a level $\alpha_{+} > \alpha$. Instead of definitely including this last table, include it only with (known) probability $(\alpha - \alpha_{-})/(\alpha_{+} - \alpha_{-})$ and exclude it with probability $(\alpha_{+} - \alpha)/(\alpha_{+} - \alpha_{-})$. The resulting conditional test will have exact level α. It this is done for every possible set of fixed marginals, the resulting unconditional test will then have the property of similarity. Tocher proved that among all similar tests of level α, the test so constructed is uniformly most powerful.

The difficulty with Tocher's test is that the extraneous randomization involved in splitting the last critical table is unacceptable to most statisticians.

Since no reasonable similar test exists, the solution from a Neyman-Pearson point of view is to find a test that is close to being similar by balancing conditional levels that are sometimes smaller and sometimes larger than the nominal level. A large number of such tests have been proposed, and their performance investigated. An excellent account of these proposals is provided by Upton (1982).

4.3 The Behrens-Fisher Problem

The Neyman-Pearson theory was successful in showing that the *t*-, χ^2- and *F*-tests for normal means and variances are uniformly most powerful against one-sided alternatives among all similar regions. There was, however, one important exception: The theory was unable to find a reasonable similar region for the problem of testing the equality of two normal means when nothing was known about the corresponding two variances. This was another case in which Fisher proposed a solution that was unacceptable from a Neyman-Pearson point of view.

The problem (now called the Behrens-Fisher problem) was considered by Fisher in 1935 [125]. The paper provided an exposition of his theory of fiducial inference which he had first put forward in 1930 [84]. The illustrations included a fiducial treatment of the Behrens-Fisher problem.

We shall discuss fiducial inference in Chap. 6, but here briefly sketch Fisher's proposal for testing the difference $\mu' - \mu$ of two normal means with both variances unknown. Considering a sample of n from the first distribution, Fisher points out that

$$\mu = \overline{x} + st,$$

where $\overline{x}$ is the sample mean, s^2 the usual estimate of the variance, and t has Student's t-distribution with $n-1$ degrees of freedom. Similarly, for the second sample

$$\mu' = \overline{x}' + s't',$$

where t' is distributed independently of t according to Student's distribution with $n'-1$ degrees of freedom. Hence

$$(\mu' - \mu) - (\overline{x}' - \overline{x}) = s't' - st \qquad (*)$$

and Fisher writes, "Since s' and s are known, the quantity represented on the right has a known distribution, though not one that has been fully tabled." [The distribution which was first suggested by Behrens (1929) has since been extensively tabled. See, for example, Johnson et al. (1995).]

Considering $\overline{x}$, $\overline{x}'$, s and s' as fixed known constants, this fiducial distribution of $\mu'-\mu$ can now be used to obtain a fiducial test of the hypothesis H: $\mu'=\mu$. On the other hand, a frequency point of view considers μ, μ', σ, and σ' as unknown constants and $\overline{x}$ and $\overline{x}'$ as random. Then even when H is true, the distribution of (*) depends on the unknown variances σ^2 and $(\sigma')^2$, and it is neither a similar test, nor can its level equal the nominal (fiducial) level. This fact was first pointed out by Bartlett (1936). Fisher replied briefly in 1937 and more fully in 1939 in a paper entitled, "The comparison of samples with possibly unequal variances." The debate went on and on without being very productive.

From a Neyman-Pearson point of view, the Behrens-Fisher test was unsatisfactory because it was not a similar region. But had this theory anything better to offer? Bartlett (1936) pointed out in a special case that similar regions did exist, and others later generalized his argument. Suppose, for example, that the sample sizes are equal, and pair the observations at random, so that $Z_i=X_i'-X_i$ $(i=1, \ldots, n)$ are independently normally distributed with mean $\zeta=\mu'-\mu$; then the hypothesis $\mu'=\mu$ is equivalent to $\zeta=0$ and can be tested by Student's test. However, again, tests involving such extraneous randomization are not considered satisfactory. And without randomization, it was later proved by Linnik (1968) that no reasonable similar tests exist.

A more satisfactory approach was initiated by B.L. Welch in a paper of 1938. He studies what he calls the validity (i.e., the control of the level for first kind error) of three tests. The first test statistic, denoted by u, is the t-statistic appropriate for testing $\mu'=\mu$ when the two variances are assumed to be equal. It divides $\overline{x}' - \overline{x}$ by the square root of the pooled estimate of the common variance.

In the second statistic, denoted by v, the pooled estimate of variance in the denominator is replaced by the sum of the estimates of the two separate variances from the two separate samples. The third, finally, is the Behrens-Fisher test.

Welch finds that none of the three statistics do a good job at controlling the first kind error at the nominal level. Continuing his investigation in a number of publications, he proposes in 1947 to refer the statistic v to the t-distribution with a random number of degrees of freedom, depending on the two estimated variances. For this test, the probability of the first kind error is very close to the nominal level. The Welch test was expanded by Aspin (1948), and the critical values tabled in Pearson and Hartley's *Biometrika Tables* of 1954. Chernoff (1949) developed a very general theory of the process of asymptotic Studentization that Welch had used to derive his test.

Fisher's earlier mocking of the objection "to a test of significance that it does not satisfy conditions which cannot possibly be satisfied" can no longer be upheld. However, Fisher now looks at this condition or requirement from a conditional point of view that also motivated his fiducial argument, but this time with a frequentist calculation.

In a note of 1956, "On a test of significance in Pearson's *Biometrika Tables* (No. 11)," he shows for the case of two samples of size seven and significance level 0.1, that the conditional probability of rejection of Welch's test when $s'/s = 1$ is ≥ 0.108 for all values of σ'/σ (of course under the assumption that $\mu' = \mu$). He thus shows that for this test, the subset $s'/s = 1$ of the sample space is what was later called a negatively biased relevant subset, i.e., a subset for which the probability of rejection under the hypothesis exceeds the level α by c for some $c > 0$, for all values of the nuisance parameters. (Later, Robinson (1976) proved that no such subsets exist for the Behrens-Fisher test).

The situation parallels that of the 2×2 table in that: (1) Fisher proposes a test that is far from being similar; and (2) no reasonable similar test exists; the Neyman-Pearson approach prefers approximately similar tests to those proposed by Fisher.

The difficult problem of appropriate reference set for frequentist calculations (or equivalently of conditional inference) was ignored by the Neyman-Pearson approach. Fisher recognized it and struggled with it (for example with his concept of ancillary statistics), but even he was not able to come up with a definitive answer.

4.4 Fixed Significance Levels vs. p-values

As we have seen, a crucial difference between Fisher's and the Neyman-Pearson approach to hypothesis testing was that Fisher considered the problem conditionally (for example, by conditioning on an ancillary statistic) while Neyman and Pearson evaluated tests unconditionally. This difference led to their disagreement about similar regions. Another more theoretical distinction was Fisher's frequently stated opposition to the consideration of alternatives and power (although his concept of the sensitivity of a test was a qualitative version of the notion of power).

A frequently claimed third distinction is that of Fisher's championing p-values, in contrast to the use of fixed significance levels by Neyman and Pearson. In the present and next sections, we shall consider the views of these authors on this issue in some detail.

Fisher discussed the matter early in his 1925 book SMRW. We already pointed out in Sect. 2.2 earlier that Fisher's position was influenced by his inability to reprint

Pearson's χ^2-tables, and how he justified the new kind of table he published instead. It is necessary now to repeat this justification because it is central to the present discussion. Fisher wrote,

> In preparing this table we have borne in mind that in practice we do not want to know the exact value of P for any observed value of χ^2, but, in the first place, whether or not the observed value is open to suspicion.

Whatever may have been his motive for doing so, he here clearly expresses a lack of interest in the p-value. He continues:

> If P is between .1 and .9 there is certainly no reason to suspect the hypothesis tested. If it is below .02 it is strongly indicated that the hypothesis fails to account for the whole of the facts. We shall not often be astray if we draw a conventional line at .05 and consider that higher values of χ^2 indicate a real discrepancy.

He spells out his position in more detail in a paper of 1926 [48], in which he states:

> It is convenient to draw the line at about the level at which we can say: "Either there is something in the treatment, or a coincidence has occurred such as does not occur more than once in twenty trials." This level, which we may call the 5 per cent. point, would be indicated, though very roughly, by the greatest chance deviation observed in twenty successive trials … .
>
> If one in twenty does not seem high enough, we may, if we prefer, draw the line at one in fifty (the 2 per cent. point), or one in a hundred (the 1 per cent. point). Personally, the writer prefers to set a low standard at the 5 per cent. point, and ignore entirely all results which fail this level.

Fisher kept the wording of the first Edition of SMRW up to the twelfth Edition of 1954. However, in the thirteenth Edition of 1958 he changed the last sentence to:

> The actual value of P obtainable from the table by interpolation indicates the strength of the evidence against the hypothesis. A value of χ^2 exceeding the 5 per cent. point is seldom to be disregarded.

This interest in the p-value suggests that, fairly late in life, Fisher's attitude had changed.

Let us consider what additional information on this question can be obtained from some of Fisher's other writing. In his second statistical book, "The Design of Experiments," which will be discussed in the next chapter, Fisher includes only a brief discussion of tests of significance. There, he states that:

> It is open to the experimenter to be more or less exacting in respect to the smallness of the probability he would require before he would be willing to admit that his observations have demonstrated a positive result… . It is usual and convenient for experimenters to take 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 introduced into their experimental results.

Unlike the changes he made in the 1958 edition of SMRW, Fisher left this statement unchanged even in the latest (seventh) Edition of his "Design" book, which was published in 1959.

Besides his theoretical discussions, we have another source of information regarding Fisher's attitude on this issue, namely his treatment of applied problems in the many examples given in SMRW. These examples provide real data and the testing of some hypothesis concerning them. And here his position is clear and did not change throughout all 14 editions. His interest was not in p-values, but in deciding whether or not the results were significant, where in nearly all cases he drew the line at 5%. The following quotations provide some illustrations. The example numbers and text are taken from the fourteenth Edition, and only the relevant sentence is cited:

> *Example 8* … hence $n=3$, $\chi^2=10.87$, the chance of exceeding which value is between .01 and .02; if we take $P=.05$ as the limit of significant deviation, we shall say that in this case the deviations from expectation are clearly significant.
>
> The value of n is now 2; … the value of χ^2, however, is so much reduced that P exceed*s* .2, and the departure from expectation is no longer significant.

> *Example 11.* The value of P is between .02 and .05, so that the sex difference … is probably significant.

> *Example 12.* For $n=9$, the value of χ^2 shows that P is less than .01, and therefore the departures from proportionality are not fortuitous.

> *Example 27.* For $n=18$, this shows that P is less than .01, and the correlation is therefore significant.

> *Example 28.* We find $t=2.719$, whence it appears from the table that P lies between .02 and .01. The correlation is therefore significant.

> *Example 35.* The difference does not exceed twice the standard error, and cannot therefore be judged significant.

> *Example 37.* The difference in variability, though suggestive of a real effect, cannot be judged significant on these data.

> *Example 4*3. The value of z is thus 1.3217 while the 1% point is about .714, showing that the multiple correlation is clearly significant. The actual value of the multiple correlation may easily be calculated from the above table, … but this step is not necessary in testing the significance.

These examples, to which many others could be added, show that Fisher considered the purpose of testing to be to established whether or not the results were significant at the 5% level, and that he was not particularly interested in the p-values per se. However, his approach differed from that of Neyman and Pearson also in other ways; not only with respect to similar regions, but also with respect to interpretation. This difference will be considered in the next section.

4.5 Fisher's "Statistical Methods and Scientific Inference"

Three years after his "Design" book, Fisher (jointly with Yates) published an enormously useful and influential book, "Statistical Tables for Biological, Agricultural, and Medical Research." These tables greatly facilitated the use of the tests that Fisher had developed, and at the same time illustrated the large variety of situations to which they were applicable. We shall not discuss this book here since it does not bear on the issues with which this chapter is concerned.

Instead we shall turn to Fisher's fourth (and last) book, published in 1956 when Fisher was 66. Its purpose was quite different from that of its predecessors. It was not to introduce readers to new statistical methods, but "an attempt to consolidate the specifically logical gains of the past half-century."

The book in fact brings very little that is new, but sets out Fisher's positions on the methods and concepts he had initiated, mostly in the 1920s and 1930s. In addition, it attempts to demolish alternative approaches that had been developed by others, particularly by Neyman and Pearson.

In the present section, we shall consider what the book has to say about significance testing.

In the Foreword, Fisher mentions Gosset's "inaugurating the first stage of the process by which statistical methods attained sufficient refinement to be of real assistance in the interpretation of data." Here, of course, he is referring to his own work on small-sample tests. He continues:

> As a result of his work, problems of distribution were one after another given exact solutions; by about 1930 all statistical problems which were thought to deserve careful treatment were being discussed in terms of mathematically exact distributions, and the tests of significance based upon them. The logical basis of these scientific applications was the elementary one of excluding, at an assigned level of significance, hypotheses, or views of the causal background, which could only by a more or less implausible coincidence have led to what had been observed. The "Theory of Testing Hypotheses" was a later attempt, by authors who had taken no part in the development of these tests, or in their scientific application, to reinterpret them in terms of an imagined process of acceptance sampling, …

It is worth noting that here Fisher avoids naming either himself or Neyman and Pearson, although he was not always so reticent. The quoted passage is fairly typical for the tone of the whole book.

It should finally be pointed out that Fisher does not mention *p*-values, but only "an assigned level of significance."

The Foreword is followed by an account of, "The Early Attempts and their Difficulties," which discusses the work of Bayes, Boole, Venn, and other early writers. The next chapter is entitled, "Forms of Quantitative Inference," and begins with a section on, "The simple test of significance."

Much of this section is concerned with refuting a Bayesian approach that would assign a posterior probability to the hypothesis being true. Fisher also restates, and expands on, his own position, and contrasts it with an alternative view:

> The attempts that have been made to explain the cogency of tests of significance in scientific research, by reference to hypothetical frequencies of possible statements, based on them being right or wrong, thus seems to miss the essential nature of such tests. A man who "rejects" a hypothesis provisionally, as a matter of habitual practice, when the significance is at the 1% level or higher, will certainly be mistaken in not more than 1% of such decisions… . However, the calculation is absurdly academic, for in fact no scientific worker has a fixed level of significance at which from year to year, and in all circumstances, he rejects hypotheses; he rather gives his mind to each particular case in the light of the evidence and his ideas.

The passage clearly is intended as a criticism of Neyman and Pearson, although again their names are not mentioned. However, these authors never recommended a fixed level of significance that would be used in all cases. On the contrary, in their first discussion (in 1928) of the two kinds of error, they write:

> These two sources of error can rarely be eliminated completely; in some cases it will be more important to avoid the first, in others the second… . From the point of view of mathematical theory, all that we can do is to show how the risk of the errors may be controlled and minimized. The use of these statistical tools in any given case, in determining just how the balance should be struck, must be left to the investigator.

This suggests that the level should not be fixed at a constant value such as the 5% recommended by Fisher, but should take into account also the power that can be achieved against the alternatives of interest. Thus, Fisher rather incongruously appears to be attacking his own past position rather than that of Neyman and Pearson.

Much of the remainder of the chapter is concerned with introducing Fisher's fiducial theory.

The next chapter is called, "Some Misapprehensions about Tests of significance." The first section, entitled, "Tests of significance and acceptance decisions," contrasts Fisher's view with that of Neyman, Pearson and Abraham Wald. It is followed by a section interpreting Student's *t*-test from a fiducial point of view. The argument is then extended to the case of linear regression. Finally, Fisher takes up the two-by-two table and the Behrens-Fisher problem in a way very similar to that discussed in Sects. 4.2 and 4.3 above. The book then turns to issues other than testing, which we shall take up later.

4.6 Inductive Inference vs. Inductive Behavior

In their 1933a paper, Neyman and Pearson suggested an idea which was to have repercussions far beyond hypothesis testing. The idea (already mentioned in Sect. 3.4 above) was that the statistician's task is to "search for rules to govern our behavior with regard to them."

Fisher stated his own different philosophy of inference in a far-ranging 1935 paper to which he gave the title: "The logic of inductive inference." There, he proposed that:

> Everyone who does habitually attempt the difficult task of making sense of figures is, in fact, essaying a logical process of the kind we call inductive, in that he is attempting to draw inferences from the particular to the general; or, as we may usually say in statistics, from the sample to the population.

He returns to this point again toward the end of the paper, when he asserts:

> In inductive reasoning, we are forming part of the process by which new knowledge is created…. The study of inductive reasoning is the study of the embryology of knowledge, of the processes by means of which truth is extracted from its native ore in which it is fused with much error.

In its main body, the paper discusses fairly new concepts that had been introduced earlier by Fisher such as likelihood, amount of information, and fiducial inference. It was a discussion paper of the *Royal Statistical Society*, and in the discussion it was received with incomprehension and hostility.

The most sympathetic of the discussants was Neyman, but even he stated that,

> I can't help thinking: "What an interesting way of asking and answering questions, but can't I do it differently,"

and later he set forth his own approach:

> Now, a system of the theory of statistics, if it is to be built "differently," must differ from that of Professor Fisher in something fundamental, that is to say, in the principle of choice… .
>
> Now what could be considered as sufficiency simple in statistical work? I think the basic conception here is the conception of the frequency of errors in judgment.

Three years later Neyman (1938a), without naming him, more explicitly attacks Fisher's concept of inductive reasoning. The following excerpts are my translations from the original French.

> Some authors attach to statistical methods the term "inductive reasoning." If, after having made the observations and calculated the limits [$\underline{\lambda} = 8.04$ and $\overline{\lambda} = 11.96$], the physician decides to assert that

$$8.04 \le \lambda' \le 11.96, \tag{*}$$

> it seems to us that the process that led to this assertion cannot be called inductive reasoning… . To see more clearly, let us distinguish between what we can know and what we believe. It seems to me that we can know or be sure of only (1) the results of experiences that have already taken place, and (2) the consequences of some definitions and postulates under which these consequences have been proved. Our attitude toward any other kind of assertion can only be described by the terms "belief" or "uncertainty."…
>
> To decide to "assert" means neither "to know nor to believe." It is an act of will preceded by some experiences and some deductive reasoning… . Consequently, it seems to me that the term "inductive reasoning" does not correspond to the nature of the process… . If one wants a special term to describe these methods and in particular to describe the decision to assert that the inequalities (*) are valid, one may perhaps propose "inductive behavior."

Neyman returned to the subject in 1947, 1955, and 1957, without adding much that was new. It may be worth noting that Pearson did not participate in the discussion. Responding in 1955 to Fisher's attacks on the Neyman-Pearson theory, he concluded his paper with: "Professor Fisher's final criticism concerns the use of the term 'inductive behavior'; this is Professor Neyman's field rather than mine."

Fisher strongly disagreed with Neyman's position in a 1955 paper, "Statistical methods and scientific induction" [261], and attacked it more stridently in his 1960 paper, "Scientific thought and the refinement of human reasoning." There, he wrote:

> There seems to be in the United States many converts to the opinion of J. Neyman, who so recently as 1938 averred that inductive reason positively did not exist, and that no process deserving the name of reasoning could be applied to the data of science… .

> Neyman's doctrine challenged not only the rapidly developing statistical science of the 20th Century, but its foundations in the 19th and 18th Centuries. On the contrary, it will be my thesis that the continuous development of mathematical thought … has come to fruition in our own time, by cross-fertilization with the natural sciences, in supplying just such a model of the correct use of inductive reasoning, as was supplied by Euclid for deductive logic. [In the Foreword to his 1956 book, Fisher expressed himself more strongly by writing that, "There is something horrifying in the ideological movement represented by the doctrine that reasoning, properly speaking, cannot be applied to empirical data to lead to inferences valid in the real world."]

After some discussion of the nature and role of probability, the 1960 paper turned to the subject of tests of significance and became so vituperative that the editors of the journal in which it appeared invited Neyman to reply. He did so in a paper of 1961, to which he gave the title, "Silver jubilee of my dispute with Fisher."

In section 3, "The essence of the dispute," Neyman stated that,

> The subject of the dispute may be symbolized by the opposing terms "inductive reasoning" and "inductive behavior." Professor Fisher is a known proponent of inductive reasoning. After a conscientious effort to find the exact meaning of this term, I came to the conclusion that, at least in the sense of Fisher, the term is empty except perhaps for a dogmatic use of certain measures of "rational belief" such as the likelihood function and the fiducial probability.

This and some of the other controversies between Fisher and Neyman were the results of different temperaments and world views. Fisher relied on his intuition, while Neyman strove for logical clarity. In addition, their dispute was fueled by aspects that were personal rather than scientific. They too are part of the history, and we shall discuss them in the next sections.

4.7 Personal Aspects I

As mentioned in Sect. 4.1, the early relationship of Fisher and Neyman was very friendly. Fisher responded quite positively to a draft of the Neyman and Pearson paper of 1933a, and he was in fact the referee who recommended publication when the paper was submitted to the *Transactions of the Royal Society*. Neyman was grateful and thanked Fisher in a letter of October 16, 1932 (Bennett 1990, pp. 189–190):

> E. S. Pearson writes that you have recommended our paper for publication. Although it may be considered ridiculous to thank a judge, I have intense feeling of gratefulness, which I hope you will kindly accept… .

Fisher replies, "It was a great pleasure to hear from you again." He also mentions a small technical error (discussed in Sect. 4.1), which leads to further correspondence.

In a letter acknowledging Fisher's criticism, Neyman writes that,

> I am often thinking that it would be very useful for me to work with you. Unfortunately, this requires considerable amount of money – without speaking of your consent – of course.

Fisher answers, "You may be sure of my consent," and in the next letter, "I like hearing from Poland. Best wishes for a Merry Christmas."

Up to this point, Neyman's letters from Poland were addressed to Fisher at Rothamstead. But this was about to change as a result of Karl Pearson's retirement in the summer of 1933, as can be seen from the following letter of Neyman's, followed by Fisher's response:

> Dr. Pearson writes me that soon you will be Galton Professor at the University College, London. Very probably this means a general reorganization of the Department of Applied Statistics and possibly new people will be needed. I know that there are many statisticians in England and that many of them would be willing to work under you. But improbable things do happen sometimes and you may have a vacant position in your laboratory. In that case please consider whether I can be of any use.

Fisher's reply, though cordial, was not encouraging:

> Many thanks for your letter of congratulation. You will be interested to hear that the Dept. of Statistics has now been separated officially from the Galton Laboratory. I think Egon Pearson is designated as Reader in Statistics. This arrangement will be much laughed at, but it will be rather a poor joke, I fancy, for both Pearson and myself. I think, however, we will make the best of it. I shall not lecture on statistics, but probably on "the logic of experimentation," so that my lectures will not be troubled by students who cannot see through a wire fence. I wish I had a fine place for you, but it will be long before my new department can be given any sort of unity and coherence, and you will be head of a faculty before I shall be able to get much done. If in England, do not fail to see me at University College.

It must indeed have been bitter for Fisher, the creator of modern statistics, that under this arrangement he was not permitted to teach statistics.

However, Neyman did get an appointment at University College after all – not in Fisher's department, but in that of his friend and collaborator Egon Pearson: first on a 3-month visiting appointment and at the end of the term (in the summer of 1934) on a regular appointment as a lecturer.

In June of that year, Neyman presented a paper to the *Royal Statistical Society*. The main body of the paper was concerned with the theory of survey sampling, but in an appendix he also set forth for the first time his theory of confidence intervals and compared it with Fisher's fiducial limits.

One of the discussants was Fisher, who referred to the "luminous account Dr. Neyman had given of the sampling technique." In discussing the relationship of confidence and fiducial limits, he stated that Neyman "had every reason to be proud of the arguments he had developed for its perfect clarity." Thus, though he did not quite agree with Neyman on the respective merits of the two approaches, he was quite complimentary, and his tone was friendly.

Another paper that Neyman presented in March 1935 did not fare so well. It was entitled, "Statistical problems in agricultural experimentation," and discussed various aspects of randomized blocks and Latin squares. In it, he found fault with some of Fisher's work on Latin squares. Fisher, who was the first discussant [128], was furious, and his remarks scathing.

He opened the discussion by saying that:

> I had hoped that Dr. Neyman's paper would be on a subject with which the author was fully acquainted, and on which he could speak with authority, as in the case of his address to the

> Society delivered last summer. Since seeing the paper, I have come to the conclusion that Dr. Neyman had been somewhat unwise in his choice of topics.

And Fisher concluded with the sentence,

> Were it not for the persistent effort which Dr. Neyman and Dr. Pearson had made to treat what they speak of as problems of estimation, by means merely of tests of significance, I have no doubt that Dr. Neyman would not have been in any danger of falling into the series of misunderstandings which his paper revealed.

The explanation for their difference of opinion is that they were considering different hypotheses. As Neyman stated in his reply:

> The problem I considered is stated on p. 162 as follows: "Our purpose in the field experiment is to compare the *average* time yields which our objects are able to give when applied to the whole field! It is seen that this problem is essentially different from what Professor Fisher suggested. So long as the *average* yields of any treat are identical, the question as to whether these treats affect *separate* yields on a *single* plot seems to be uninteresting and academic, and certainly I did not consider methods for its solution.

At this point, Fisher interrupted with

> It may be foolish, but that is what the $\underline{z}$ test was designed for, and the only purpose for which it has been used.

Neyman later (Reid 1982, p. 126) recalls that a week after this meeting, Fisher stopped by his room at University College:

> And he said to me that he and I are in the same building... . That, as I know, he had published a book – and that's *Statistical Methods for Research Workers* – and he is upstairs from me so he knows something about my lectures – that from time to time I mention his ideas, this and that – and that this would be quite appropriate if I were not here in the College but, say, in California – but if I am going to be at University College, then this is not acceptable to him. And then I said, "Do you mean that if I am here, I should just lecture using your book?" And then he gave an affirmative answer. And I said, "Sorry, no. I cannot promise that." And then he said, "Well, if so, then from now on I shall oppose you in all my capacities."

Reid also reports (p. 124) that,

> After the Royal Statistical Society meeting of March 28, relations between workers on the two floors of K. P.'s old preserve became openly hostile. One evening, late that spring, Neyman and Pearson returned to their department after dinner to do some work. Entering, they were startled to find strewn on the floor the wooden models which Neyman had used to illustrate his talk on the relative advantages of randomized blocks and Latin squares. They were regularly kept in a cupboard in the laboratory. Both Neyman and Pearson always believed that the models were removed by Fisher in a fit of anger.

4.8 Personal Aspects II

In his publications after 1935, Fisher for many years ignored Neyman (except for some references to the Neyman-Pearson theory in connection with the Behrens-Fisher problem). For example, there is not a single mention of Neyman in Fisher's 1935 book on the design of experiments.

In part, this may have been due to the fact that in 1938 Neyman left England for California so the irritation of daily presence was removed. Also, before long, war broke out. All of Neyman's energy went into war work. On the other hand, to Fisher's great disappointment, the British government was not willing to make use of his statistical expertise for this purpose. So most of his efforts during the war years, first as Galton Professor of Eugenics at University College, London, and after 1943 as Balfour Professor of Genetics at Cambridge, were devoted to problems in genetics rather than statistics.

Finally, another reason for Fisher's ignoring Neyman is stated in a 1951 letter to Horace Gray, who had brought to Fisher's attention Neyman's review of Fisher's Selected Papers, for the Scientific Monthly (Vol. LXII, June 1951).

After first praising Fisher's early work (while describing him condescendingly as "a very able 'manipulative' mathematician"), Neyman had written:

> In his long scholarly work, Fisher also made frequent attempts at the conceptual side of mathematical statistics, and these efforts are duly reflected in this volume. In particular, three major concepts were introduced by Fisher and consistently propagandized by him in a number of publications. These are mathematical likelihood as a measure of the confidence in a hypothesis, sufficient statistics, and fiducial probability. Unfortunately, in conceptual mathematical statistics, Fisher was much less successful than in manipulatory, and of the three above concepts only one, that of a sufficient statistic, continues to be of substantial interest. The other two proved to be either futile or self-congratulatory, and have been more or less generally abandoned.

So Neyman had done it again. As in 1935, he had attacked Fisher publicly and without provocation. Fisher's reaction is understandable. In thanking Gray, he wrote:

> Neyman is, judging by my own experience, a malicious mischief maker. Probably by now this is sufficiently realized in California. I would not suggest to anyone to engage in scientific controversy with him, for I think that discussion is only profitable when good faith can be assumed in the common aim of getting at the truth.

However, Fisher was mistaken about Neyman's situation and reputation in California, where Neyman was building a first-rate department and his symposia were drawing the best people worldwide. At those meetings he was a gracious, enterprising, and much-admired host. Of course, Neyman in his review was also in error when he announced the early demise of the likelihood concept. Likelihood has continued to be considered of first-rate importance.

Fisher did not keep to his resolve not to engage in controversy with Neyman. In a 1955 paper, "Statistical methods and scientific induction" [261], he contrasted his own approach with that of Neyman (and Wald, whose book on decision theory had appeared in 1950 and who had died that same year). His summary at the beginning of the paper characterizes the difference as follows:

> The attempt to reinterpret the common tests of significance used in scientific research as though they constituted some kind of acceptance procedure and led to "decisions" in Wald's sense, originated in several misapprehensions and has led, apparently, to several more.
>
> The three phrases examined here, with a view to elucidating the fallacies they embody, are: (i) "repeated sampling from the same population," (ii) errors of the "second kind," (iii) "inductive behavior."

> Mathematicians without personal contact with the natural sciences have often been misled by such phrases. The errors to which they lead are not always only numerical.

The distinction between scientific work and acceptance procedures (the latter Fisher's interpretation of Neyman's behavioristic view which Wald had also embraced) constitutes the main point of the paper. After an introductory section, Fisher discusses his objection to the evaluation of statistical procedures in terms of, "Repeated sampling from the same population." He illustrates the issue with the debate about the 2×2 table (discussed in Sect. 4.2 above). Referring to Barnard's suggestion (and later withdrawal[2]) of an alternative to Fisher's test, Fisher writes that,

> Professor Barnard has a keen and highly trained mathematical mind, and the fact that he was misled into much wasted effort and disappointment should be a warning that the theory of testing hypotheses set out by Neyman and Pearson has missed at least some of the essentials of the problem, and will mislead others who accept it uncritically.

However, this and similar references to Neyman and Pearson later in the paper are mild compared to those in Fisher's 1960 paper, "Scientific thought and the refinement of human reasoning" [282]. Here, referring to the Neyman-Pearson theory, Fisher states that,

> A whole series of erroneous tests of significance were incorporated into statistical teaching, and although one by one they have been exposed and discredited, and have seldom gained a place among the tests used in genuine research, they have ensured that many young men now entering employment in research, or industry, or administration have been partly incapacitated by the crooked reasoning with which they have been indoctrinated.

Why, after his long silence, did Fisher late in life (in 1960 he was 70) lash out so violently against his opponents?

Fisher was justifiably proud of what he had achieved. The development of the methodology of small-sample tests (including the analysis of variance) alone would have constituted a notable lifetime achievement. But he had followed this with the creation of a science of experimental design, and of a theory of estimation with many new concepts such as likelihood, sufficiency, amount of information, and fiducial inference.

And then Neyman, Pearson, and more recently Wald had deformed and, as he felt, besmirched his life's work with ideas that seemed to him completely inappropriate. And these ideas were being accepted as important innovations by a new generation of statisticians.

The situation is described by Fisher's daughter Joan Fisher Box in her biography, "R. A. Fisher – The Life of a Scientist." "If he had decided to move to the United States in 1937," she writes,

> his contact with statistical developments in America would not have been broken during these critical years when new influences made themselves effective. J. Neyman emigrated to America in 1939 and A. Wald in 1940[3]; in teaching, theoretical and mathematical aspects

[2] In a 1949 paper, Barnard stated that "further meditation has led me to think that Professor Fisher was right after all."

[3] Actually, both moved to the U.S. in 1938.

> of statistics received increasing emphasis. Changes took place in the mood of American statisticians which extinguished the enthusiasm that had so charmed Fisher in 1936. When he returned to the United States in 1946, he was welcomed by the younger statisticians as a great originator and authority certainly, but also as a foreigner whose ways were not always their ways, nor his thoughts their thoughts.

This must have been extremely galling to Fisher, and it resulted in his giving expression to his anger and frustration rather intemperately.

Chapter 5
The Design of Experiments and Sample Surveys

5.1 Lady Tasting Tea

As mentioned toward the end of Sect. 2.3, the last pages of SMRW dealt not with significance testing but with a new subject: the planning of experiments. Fisher extended this brief sketch in a 1926 survey paper, "The arrangement of field experiments." His interest in experimental design grew out of his work at Rothamsted and resulted in many papers dealing with the new ideas and techniques he developed in this context. In 1935, he gave a systematic account of the new science of experimental design he had created in a book, "The Design of Experiments" (DOE). Instead of following the details of his work on the subject, the present and following sections will provide a discussion of this enormously influential book.

The book begins with an introductory chapter that is partly historical. It sets out the book's purpose, and at its center reviews and rejects the approach through inverse probability. It argues that,

> advocates of inverse probability seem forced to regard mathematical probability, not as an objective quantity measured by observed frequencies, but as measuring merely psychological tendencies, theorems respecting which are useless for scientific purposes.

Fisher goes on to claim that,

> In fact, in the course of this book, I propose to consider a number of different types of experimentation, with especial reference to their logical structure, and to show that when the appropriate precautions are taken to make this structure complete, entirely valid inferences may be drawn from them, without using the disputed [i.e., Bayesian] axiom.

The first experiment Fisher discusses is described in Chap. II, "The principles of experimentation illustrated by a psycho-physical experiment." This very simple, though somewhat artificial, example has acquired much fame under the title, "The lady tasting tea." It is based on an event that really happened one afternoon at Rothamsted.[1]

[1] See Box (1978, p. 134).

E.L. Lehmann, *Fisher, Neyman, and the Creation of Classical Statistics*,
DOI 10.1007/978-1-4419-9500-1_5, © Springer Science+Business Media, LLC 2011

As Fisher describes the experiment:

> A lady declares that by tasting a cup of tea with milk she can discriminate whether the milk or the tea infusion was first added to the cup… . Our experiment consists in mixing eight cups of tea, four in one way and four in the other, and presenting them to the subject for judgment in random order. The subject has been told in advance of what the test will consist, namely that she will be asked to taste eight cups, that these shall be four of each kind, and that they shall be presented to her in random order… . Her task is to divide the 8 cups into two sets of four, agreeing, if possible, with the treatments received.

He next points out that there are $\binom{8}{4}=70$ possible outcomes, so that "a subject without faculty of discrimination would in fact divide the 8 cups correctly into two sets of 4 in one trial out of 70." This completes the description of the experiment, to which Fisher adds a section of explanation entitled, "Interpretation and its reasoned basis." This is followed by a section on "The test of significance," which summarizes material from his earlier book SMRW. The next section bears the title, "The null hypothesis." Fisher explains this term by stating that the hypothesis to be tested is,

> In this case, the hypothesis that the judgments given are in no way influenced by the order in which the ingredients have been added… . This hypothesis, which may or may not be impugned by the result of an experiment, is again characteristic of all experimentation. Much confusion would be avoided if it were explicitly formulated when the experiment is designed. In relation to any experiment, we may speak of this hypothesis as the "null hypothesis," and it should be noted that the null hypothesis is never proved or established, but it is possibly disproved, in the course of experimentation.

The paragraph ends with a sentence that has often been quoted: "Every experiment may be said to exist only in order to give the facts a chance of disproving the null hypothesis."

A central concern of the book is the need for randomization, and Fisher devotes the next two sections to this topic. He argues that any experiment without it is flawed, and that on the other hand, it "will suffice to guarantee the validity of the test of significance, by which the result of the experiment is to be judged." Randomization is a subject to which the book will return again and again.

The last topic addressed in this chapter is how to improve the sensitiveness of an experiment. Fisher states that,

> By increasing the size of the experiment, we can render it more sensitive, meaning by this that it will allow of the detection of a lower degree of sensory discrimination… . Since in every case the experiment is capable of disproving, but never of proving this hypothesis, we may say that the value of the experiment is increased whenever it permits the null hypothesis to be more readily disproved.

Fisher's sensitiveness is essentially the same as the Neyman-Pearson concept of power. However, unlike the latter, it does not provide a quantitative measure. As we know, Fisher was familiar with the Neyman-Pearson theory; in his 1934 paper, he had even used the terms uniformly more, and most powerful. By not utilizing the idea of power, Fisher deprives himself of the ability to resolve one of the most important issues of experimental design, the determination of sample size.

5.2 An Experiment of Darwin's

The concepts and techniques introduced in Fisher's Chap. II are extended in the following chapter to the continuous case. It again deals with a single example, an experiment carried out by Charles Darwin and reported in his book of 1876. The example concerns a comparison of crossed and self-fertilized plants. Three plants of each kind were grown in four pots, and Darwin then asked Francis Galton to analyze the measurements he had made on their sizes.

Fisher proceeds to describe the method used by Darwin, the method of paired comparisons. The discussion of various aspects of this design is the principal purpose of the chapter. Fisher points out that,

> The method of pairing … illustrates well the way in which an appropriate experimental design is able to reconcile two desiderata, which sometimes appear to be in conflict. On the one hand we require the utmost uniformity in the biological material, which is the subject of experiment, in order to increase the sensitiveness of each individual observation; and, on the other, we require to multiply the observations so as to demonstrate as far as possible the reliability and consistency of the results… .
>
> However, …, there is no real dilemma. Uniformity is only requisite between the objects whose response is to be contrasted (that is, objects treated differently). It is not requisite that all the parallel plots under the same treatment shall be sown on the same day, but only that each such plot shall be sown as far as possible simultaneously with the differently treated plot or plots with which it is to be compared. If, therefore, only two kinds of treatments are under examination, pairs of plots may be chosen, one plot for each treatment; and the precision of the experiment will be given its highest value if the members of each pair are treated closely alike, but will gain nothing from similarity of treatment applied to different pairs, nor lose anything if the conditions in these are somewhat varied.

Having described the experimental setup, Fisher next discusses the *t*-test for testing the null hypothesis of no difference under the assumption of normality which, he says,

> By wide experience, has been found to be appropriate to the metrical characters of experimental material in biology.

The first step consists in subtracting,

> From the height of each cross-fertilised plant the height of the self-fertilised plant belonging to the same pair. With respect to [the resulting] differences, our null hypothesis asserts that they are normally distributed about a mean value at zero, and we have to test whether our 15 observed differences are compatible with the supposition that they are a sample from such a population.

The section ends with the details of carrying out the *t*-test, and finds that "the observed value of *t*, 2.148, just exceeds the 5% point, and the experimental result may be judged significant, though barely so."

After some further discussion, Fisher comes to a crucial aspect of the experimental design. He states that,

> Having decided that, when the structure of the experiment consists in a number of independent comparisons between pairs, our estimate of the error of the average difference must be

> based upon the discrepancies between the differences actually observed, we must next enquire what precautions are needed in the practical conduct of the experiment to guarantee that such an estimate shall be a valid one; … .
>
> In the experiment under consideration, apart from chance differences in the selection of seeds, the sole source of the experimental error in the average of our fifteen differences lies in the difference in soil fertility, illumination, evaporation, etc., which makes the site of each crossed plant more or less favourable to growth than the site assigned to the corresponding self-fertilised plant… . If now, when the fifteen pairs of sites have been chosen, …, we then assign at random, as by tossing a coin, which site shall be occupied by the crossed and which by the self-fertilised plant, we shall be assigning by the same act whether this particular ingredient of error shall appear in our average with a positive or negative sign.
>
> …
>
> Randomisation properly carried out … ensures that the estimates of error will take proper care of all such causes of different growth rates, and relieves the experimenter from the anxiety of considering and estimating the magnitude of the innumerable causes by which his data may be disturbed. The one flaw in Darwin's procedure was the absence of randomisation.

The last section of the chapter brings a remarkable and surprising insight. Entitled, "Test of a wider hypothesis," the section considers the hypothesis "which merely asserts that the two series are drawn from the same population without specifying that it is normally distributed."

Fisher writes:

> It seems to have escaped recognition that the physical act of randomisation which, as has been shown, is necessary for the validity of any test of significance, affords the means, in respect of any particular body of data, of examining the wider hypothesis in which no normality of distribution is implied.

The basis of this test is the fact that random assignments of the sites within each pair implies that the 15 observed differences "would each have occurred with equal frequency [i.e., with probability ½] with a positive or negative sign. It is now possible to calculate the p-value of t within this population of 2^{15} equally likely cases and this gives a value of 5.255%, very close to the 5% obtained from Student's t-distribution."

Although he shows the closeness of the two p-values only in this one example, Fisher implicitly suggests that the conclusion holds quite generally. These randomization tests, as Fisher points out, are very tedious to carry out, and as a result they only came into their own after computers made the calculations easy. A proof that (at least asymptotically) Fisher's conjecture is correct was provided by Hoeffding in 1952.

5.3 Randomized Blocks

After the thorough discussion of randomized paired comparisons, it was easy for Fisher to extend the idea to the comparison of more than two treatments, and he does so in the next chapter entitled, "An agricultural experiment in randomised blocks." He takes the example of testing "the relative productivity, or yield, of five different varieties of farm crop."

"To this end," Fisher writes,

> we shall suppose that the experimental area is divided into eight compact, or approximately square, blocks, and that each of these is divided into five plots, running from end to end of the block, and lying side by side, making forty plots in all. Apart from the differences in variety to be used, the whole area is to have uniform agricultural treatment.
>
> In each block the five plots are assigned one to each of the five varieties at random.

He then discusses how to carry out these random assignments.

Having described the experiment, he goes on to consider the "statistical analysis of the observations." The procedure, which Fisher calls an analysis of variance [ANOVA], begins with a determination of "the numbers of degrees of freedom, or independent comparisons, which can be made between the plots, or relevant groups of plots." Fisher states that,

> Between 40 plots, 39 independent comparisons can be made, and so the total number of degrees of freedom will be 39. This number will be divided into three parts representing the numbers of independent comparisons (a) between varieties, (b) between blocks, and (c) representing the discrepancies between the relative performances of different varieties in different blocks, which discrepancies provide a basis for the estimation of error.

He then gives a table showing these degrees of freedom as being

Varieties	4
Blocks	7
Error	28
Total	39

Realizing that these numbers need some justification, he explains:

> It is easy to see that the number of degrees of freedom for any group of simple comparisons, such as those between varieties or between blocks, must be one less than the number of items to be compared. In the present instance, in which the plots are assigned within the blocks wholly at random, the whole of the remaining 28 degrees of freedom are due simply to differences in fertility between different plots within the same block, and are therefore available for providing the estimate of error… . The form we have set out is appropriate to the question whether the yields given by the different varieties show, as a whole, greater differences than would ordinarily be found, had only a single variety been sown on the same land. It is appropriate to test the null hypothesis that our 5 varieties give in fact the same yields.

Fisher goes on to discuss "the completion of the analysis of variance… . It consists in the partition of a quantity known as *the sum of squares* (i.e., of deviations from the mean) into the same three parts as those into which we have already divided the degrees of freedom."

After going into details of the calculation of these three parts, he states that,

> In this way, the total sum of squares, representing the total amount of variation due to all causes between the 40 yields of the experiment, is divided into the 3 portions relevant to its interpretation, measuring respectively the amount of variation between varieties, the amount of variation between blocks, and the amount of discrepancy between the performances of the different varieties in the different blocks.

The section concludes by showing how to test the null hypothesis of no difference in the yields of the five varieties. Under this hypothesis, the mean squares for varieties and for error are independent estimates of the same quantity, the variance of the errors. Their ratio therefore has a known distribution [now known as the F-distribution] and can be used to test the null hypothesis.

Finally, Fisher points out regarding this test that "as with the t-test, its appropriateness to any particular body of data may be verified arithmetically [i.e., by means of a randomization test which does not require the assumption of normality]."

The next section is concerned with the steps to be taken when the null hypothesis is rejected. Fisher suggests that one might want then to test each difference of yields. Each of these differences (taking account of their signs) can then be tested by means of a t-test. However, he warns,

> much caution should be used before claiming significance for special comparisons… . Comparisons suggested by a scrutiny of the results themselves are open to suspicion; for if the variants are numerous, a comparison of the highest with the lowest observed value will often appear to be significant, even from undifferentiated material.

This is the problem of multiplicity, and Fisher suggests a remedy: to use a level of 5% divided by $m(m-1)/2$, where m is the total number of varieties. In the comparison of 20 varieties, for example, 0.05/190 would be used instead of 0.05. Although he does not mention it, he was presumably aware of the fact that this procedure guarantees that the probability of at least one false rejection is controlled at the 5% level.

Much of the rest of the chapter is concerned with issues specific to agriculture, and we shall turn directly to the next chapter.

5.4 Latin Squares

The Latin Square design,[2] which forms the topic of Chap. V, is concerned with controlling the effects of two different sources of heterogeneity. Fisher again discusses the procedure in terms of an example. He considers the comparison of six agricultural treatments on an experimental area which is divided into 36 equal plots lying in 6 rows and 6 columns. He states that "it is then a combinatorial fact that we can assign plots to the six treatments such that for each treatment, one plot lies in each row and one in each column of the square."

For a particular field, he writes, "it is not known whether heterogeneity [of the soil] will be more pronounced in the one or the other direction in which the field is ordinarily cultivated… . The effects are sufficiently widespread to make apparent the importance of eliminating the major effects of soil heterogeneity not only in one direction across the field, but at the same time in the direction at right angles to it."

[2] Interest in Latin Squares, though not the use Fisher makes of them, goes back at least to the eighteenth century; see, for example, Dénes and Keedwell (1974).

To obtain an effective randomization, Fisher considers sets of Latin Squares that can be transformed into each other by various types of transformations. He then proposes to choose at random (i.e., with equal probabilities) one member of such a set.

He next turns to an analysis of variance similar to that used for randomized blocks. He points out that, "The 35 independent comparisons among 36 yields give 35 degrees of freedom. Of these, five are ascribable to differences between rows, and five to differences between columns… . Of the remaining 25 degrees of freedom, 5 represent differences between the treatments tested, and 20 represent components of error which have not been eliminated, but which have been carefully randomized so as to ensure that they shall contribute no more and no less than their share to the errors."

Finally, he proposes to divide "what we have called the total sum of squares" into components corresponding to the various degrees of freedom.

The next two sections deal with common mistakes made in applications of the Latin Square design. The first is concerned with errors in the ANOVA, the second with the replacement of randomly chosen squares with squares chosen systematically in an effort to achieve good balance. The last topic treated in the chapter is a special kind of Latin Square, called Graeco-Latin Square, which allows the control of not only two, but three, different sources of heterogeneity.

5.5 Factorial Designs

Chapters II–V of DOE are closely related, with randomized blocks (Chap. IV) being a generalization of randomized pairing (Chaps. II and III), and Latin Squares (Chap. V) a refinement of randomized blocks. The next three chapters deal with a new issue. Fisher begins Chap. VI by setting forth the commonly held philosophy of experimentation and his disagreement with it:

> In expositions of the scientific use of experimentation it is frequent to find an excessive stress laid on the importance of varying the essential conditions *only one at a time*. The experimenter interested in the causes which contribute to a certain effect is supposed, by a process of abstraction, to isolate these causes into a number of elementary ingredients, or factors, and it is often supposed, at least for purposes of exposition, that to establish controlled conditions in which all of these factors except one can be held constant, and then to study the effects of this single factor, is the essentially scientific approach to an experimental investigation.

Fisher disagrees with this view because, as he writes,

> We are usually ignorant which, out of innumerable possible factors, may prove ultimately to be the most important, though we may have strong suppositions that some few of them are particularly worthy of study. We have usually no knowledge that any one factor will exert its effects independently of all others that can be varied, or that its effects are particularly simply related to variations in these other factors.

To illustrate his alternative approach, Fisher considers a study of "the effects of variations in composition of a mixture containing four active ingredients."

He assumes that “16 different mixtures are made up in the 16 combinations possible by combining either a larger or smaller quantity of each of the 4 ingredients to be tested.”

He next supposes that six tests are made with each mixture, 96 in all, and considers two designs: (1) assign all 96 cases to the experimental units (agricultural plots, patients, etc.) completely at random; (2) a randomized block design, with six blocks “supposedly more homogeneous than the whole,” with the 16 members of each assigned at random.

Before discussing the ANOVA for these two cases, Fisher points out the advantages of such a factorial design. He states that, “The first factor contributing to the efficiency of experiments designed on the factorial system is that every trial supplies information upon each of the main questions, which the experiment is designed to examine.”

He continues by asserting that,

> The advantage of the factorial experiment over a series of experiments, each designed to test a single factor, is, however, much greater than this. For with separate experiments we should obtain no light whatever on the possible interactions of the different ingredients, … .”

He then goes on to explain first and higher order interactions, and this leads him to conclude that:

> the 15 independent comparisons of 16 different mixtures … may be resolved into 15 intelligible components, as shown in the following table:

Effects of single ingredients	4
Interactions of 2 ingredients	6
Interactions of 3 ingredients	4
Interactions of 4 ingredients	1
Total	15

It now only remains to lay out the ANOVA. For the completely randomized design, the 95 degrees of freedom are assigned, 15 to treatments and the remaining 80 to error. In the randomized block design, the division is: 5 to blocks, 15 to treatments, and 75 to error.

In the next section, entitled, “The basis of inductive inference,” Fisher adds a third advantage of factorial experiments to those mentioned earlier:

> This is that any conclusion … has a wider inductive basis when inferred from an experiment in which the quantities of other ingredients have been varied than it would have from any amount of experimentation in which these had been kept strictly consistent. The exact standardization of experimental conditions, which is often thoughtlessly advocated as a panacea, always carries with it the real disadvantage that a highly standardized experiment supplies direct information only in respect to the narrow range of conditions achieved by standardisation.

A related advantage is the possibility of including in the experiment what Fisher calls subsidiary factors with little loss of precision in the inferences about the main factors. In a final section, he points out that in experiments with a very large number of factors, the size of the experiment may make replication undesirable. There will then be no degrees of freedom available for the estimate of error. However, he writes, “there will be numerous interactions the apparent effects of which are principally due to error, and these may be used to provide a measure of the precision of the more important comparisons.”

5.6 More Complex Designs

The first six chapters of DOE are easy to read, and the important concepts they introduce: randomization, blocking, and factorial experimentation, are presented with utmost clarity. However, the topics treated in Chaps. VII–XI: confounding, partial confounding, and concomitant measurements, are more complex, and as a result these chapters are more difficult to read. We shall not discuss them in detail but, to give at least a flavor, shall consider Fisher’s introduction to confounding.

The issue arises in factorial experiments in which the heterogeneity within blocks is still thought to be too high. In order to reduce it, the blocks are divided into two or more parts, which then form blocks with fewer plots than the number of treatments.

Fisher illustrates the situation with the case of three factors, each of which has two levels, say level 1 and level 2 (e.g., treatment and control). There are then eight treatment combinations and the factorial design described above would require blocks with eight plots each. In order to reduce heterogeneity, each block is divided into two more homogeneous blocks of size 4. To one of these smaller blocks are assigned the treatment combinations in which all three treatments are at level 1, and those in which one is at level 1 and the other two at level 2, four in all. To the other smaller blocks are assigned the combinations in which two of the treatments are at level 1 (and the remaining at level 2), and that in which all three treatments are at level 2.

With this design, Fisher points out, the third order interaction turns out to be the simple contrast between the sums of the treatments in the two blocks of size 4. This contrast has therefore been confounded with, i.e., is indistinguishable from, the third order interaction. The description of this design is followed by the ANOVA for this case, and then by consideration of a second, larger example. The chapter ends with an introduction to partial confounding which is discussed more fully in the following chapter. Finally, Chap. XI deals with “the increase of precision by concomitant measurements. Statistical Control.”

The book ends with two additional chapters, which are, however, less concerned with experimental design than with the theory of estimation, a topic we shall take up in the next chapter.

5.7 Further Developments

As “Statistical Methods” had done, “The Design of Experiments” established an entirely new subject. The present section sketches some of the discussions and further considerations that followed Fisher’s presentation.

5.7.1 Randomization

This concept which stood at the center of Fisher's approach to experimental design did not find favor with everyone. In particular, Fisher's friend Gosset argued that balanced systematic designs were preferable. Some references to this disagreement can be found in their correspondence in the 1930s. Gosset's ("Student's") first public statement on the issue occurred in a paper of 1936, which led to a series of alternating publications of the two authors, culminating in a posthumous paper of 1938 by Student, who had died before completing his revisions.

In this paper, Student summarizes the dispute as follows:

> He [Fisher] holds that balanced arrangements may or may not lead to biased means according to the lie of the ground, but that in any case the value obtained for the error is so misleading that conclusions drawn are not valid, while I maintain that these arrangements tend to reduce the bias due to soil heterogeneity and that so far from the conclusions not being valid, they are actually less likely to be erroneous than those drawn from artificially randomized arrangements.

In this paper, Student also raises an issue which later writers have considered the main drawback of randomization, namely that,

by the luck of the draw [the treatments may] come to be arranged in a very unbalanced manner.

This difficulty can of course be avoided by what is called restricted randomization, i.e., by eliminating from consideration the most unbalanced arrangements and randomizing among the rest. Such an approach was in fact suggested by Fisher in the only comment of his on the matter that seems to have been recorded. In Savage et al. (1962), cited in Fienberg and Hinkley (1980, p. 45), Savage reports that in 1952, he asked Fisher:

> What would you do if, drawing a Latin Square at random for an experiment, you happened to draw a Knut Vik Square? Sir Ronald said he thought he would draw again and that, ideally, a theory explicitly excluding regular squares should be developed.

5.7.2 Analysis of Variance (ANOVA)

Regarding Fisher's ANOVA, there also emerged some difficulties, but they were of a very different kind from the problems encountered with randomization. Nobody found fault with the ANOVA procedures, everyone agreed that they were enormously useful and one of Fisher's greatest achievements. The problem was that no one was quite sure of what really constituted an ANOVA.

The difficulty in some way originated with Fisher, who never gave a general definition and who in the chapter of SMRW (fourth Edition) entitled, "Further applications of the ANOVA," wrote:

> It is impossible in a short space to give examples of all the different applications which may be made of this method; we shall therefore limit ourselves to those of the most immediate practical importance.

These examples were systematized in Scheffé's 1959 book, "The Analysis of Variance," which took as its basic framework linear models with either fixed or random effects. Whether Fisher would have agreed with this formulation is hard to tell. In DOE, he wrote:

> The arithmetical discussion by which the experiment is to be interpreted is known as the analysis of variance. This is a simple arithmetical procedure by means of which the results may be arranged and presented in a single compact table, which shows both the structure of the experiment and the relevant results in such a way as to facilitate the necessary tests of their significance.

The description at first sounds as if ANOVA were a data-analytic procedure not requiring a model. This is the view taken by some later writers, but it is to some extent vitiated by the last part of Fisher's statement, although even this is ambiguous since it only requires that the procedure "facilitates" the test, but does not actually make the test part of the procedure.

Attempts were later made to extend ANOVA to other than the linear models with normal errors that constituted the subject of Scheffé's book. When the normality assumption is dropped, sums of squares which were central to Fisher's concept may no longer be relevant. A different approach was reviewed by Terry Speed in a 1987 paper entitled, "What is an analysis of variance?" The paper was followed by comments of 11 discussants, none of whom agreed with either Speed's definition or that of any of the other discussants.

In practice, ANOVA seems to be used mainly in linear models and their extension to generalized linear models (McCullagh and Nelder 1983).

5.7.3 Optimality

Like the test discussed in SMRW, the designs of DOE were presented by Fisher without much justification, based entirely on his intuitive understanding of what the situation demanded. But again later writers found justifications by showing that Fisher's procedures possessed certain optimality properties. A theory of optimal designs was initiated by Wald (1943) and Wald's student Ehrenfeld (1953). However, the person most responsible for such a fully developed theory was Kiefer (1985), who between 1958 and his early death in 1981, published more than 40 papers on the subject. These papers by Jack Kiefer complemented and to some extent completed Fisher's work on experimental design as the Neyman-Pearson theory had done for Fisher's testing methodology.

5.8 Design of Sample Surveys

In all three areas of Fisher's major statistical work: small-sample (exact) tests, experimental design, and theory of estimation, Neyman played an important complementary role. In testing, the Neyman-Pearson theory provided justification for the normal-theory tests that Fisher had proposed on intuitive grounds. (In a few cases, it later turned out that the two approaches led to different solutions). In design, as will be discussed in the present section, Fisher and Neyman developed, about simultaneously but independently, theories that had important parallels, but in quite different areas of application. Finally, in estimation they tackled a common problem but arrived at very different solutions. Estimation will be considered in the next chapter.

The purpose of sampling is to estimate one or more characteristics of a population Π by means of a sample from the population. Neyman's principal contribution to the design of statistical studies was his 1934 paper, "On the two different aspects of the representative method: the method of stratified sampling and the method of purposive selection." This paper has been described by Fienberg and Tanur (1996) as "Neyman's watershed paper, in which he was able to capture the essential ingredients of the problem of sampling."

The paper starts with an introduction in which Neyman reviews the recent history of the representative method of sampling, and emphasizes that,

> there are two different aspects of the representative method. One of them is called the method of random sampling and the other the method of purposive selection… . The two kinds of method were discussed by A. L. Bowley in his book … as being equally to be recommended.

He goes on to cite a 1926 report by a commission appointed by the International Statistical Institute, which took the same attitude, and to describe an application of the purposive method.

It may be worth noting that the distinction between these two approaches corresponds roughly to that between randomized and systematic designs discussed by Fisher.

Neyman finally points out that "the theory of purposive selection seems to have been extensively presented only in the two papers mentioned, while that of random sampling has been discussed probably by more than a hundred authors." He therefore concludes that "it seems justifiable to consider carefully the basic assumptions underlying the former. This is what I intend to do in the present paper."

The introduction is followed by a chapter entitled, "Mathematical theories underlying the representative method," and which is divided into two sections. The first bears the title, "The theory of probabilities a posteriori and the work of R. A. Fisher." It mentions the difficulties attending the first of these [i.e., the inverse method] but notes that "an approach to problems of this type has been suggested by R. A. Fisher which removes the difficulties involved in the lack of knowledge of the a priori probability law." This is Fisher's fiducial inference, and Neyman states that in Appendix I he has "described the main lines in a way somewhat different from that followed by Fisher." He mentions that "the form of this solution consists in determining certain intervals, which I propose to call the confidence intervals." We shall postpone discussion of this concept to the next chapter.

Section 2 deals with "The choice of the estimates." Assume some characteristic θ of a known population Π is to be estimated based on a sample from the population. Neyman considers only linear estimates of θ. If the sample is obtained by some random method, Neyman makes use of Markoff's theory of least squares to find a best linear unbiased estimate. If the number of observations is not too small, he claims, the estimate is approximately normally distributed and, suitably Studentized, is distributed according to Student's *t*. Since this is a known and tabled distribution, it can be used to obtain confidence intervals for the parameter being estimated.

To apply this approach to the problem of comparing randomized and purposive sampling, Neyman now assumes that the population Π is comprised of *k* subpopulations $\Pi_1,\ldots, \Pi_{k,}$ from which samples of sizes $n_1,\ldots,n_k$ are to be drawn. The means $A_1,\ldots,A_k$ of the populations are unknown, and of the variances it is supposed that $\sigma_i^2 = \sigma_0^2/P_i$, where σ_0^2 is unknown but the divisors P_i are known. If the samples have been drawn at random, Neyman then applies the Markoff theory to obtain confidence intervals for one or more linear functions of the means A_i.

In the next chapter, Neyman compares what he calls the two aspects of the representative method, first the method of random sampling, then the method of purposive sampling. If samples are drawn non-randomly, estimates of means and confidence intervals may be biased.

He begins his study of random sampling by noting that there are several types.

(a) The sample Σ is obtained by taking at random single individuals from the population Π. This may be done with or without replacement. (This is now called a simple random sample).

(b) The population Π is divided into several subpopulations now called strata $\Pi_1,\ldots, \Pi_k$ and the sample is obtained by drawing random samples $\Sigma_1,\ldots,\Sigma_k$ from the strata. The sizes of the strata are denoted by $M'_1,\ldots,M'_k$ and those of the samples by $m'_1,\ldots,m'_k$. (This is now called a stratified random sample).

Before considering a third type of sampling, Neyman points out that in practice the members of the population are typically grouped (as an example he mentions households), and that it is much easier to take a random sample of such groups than of individuals. He therefore proposes a third method:

(c) Assuming that the population Π of M' individuals consists of M_0 groups, he considers the population Π whose elements are the groups. A random sample is now taken from Π. Again, the sampling can be unrestricted or stratified. In view of the advantages of the latter, Neyman says he will consider only stratified sampling from Π. (This is now called a cluster or stratified cluster sample).

Neyman next notes that random sampling is not supported by everyone. In particular, he refers to work on purposive sampling by Bowley and by the Italian statisticians Corrado Gini and Luigi Galvani. He describes their method and introduces a probabilistic structure that allows him to describe the assumptions, based on his previous analysis that would make the estimate from this type of purposive sampling satisfactory. He then considers the consequences for the method of purposive sampling when these assumptions are not satisfied.

A major conclusion is that if these hypotheses are not satisfied, then with purposive selection, "we are not able to appreciate the accuracy of the results obtained… ." We are very much in the position of a gambler, betting at 1 time £100. Hence: "the final conclusion which both the theoretical considerations and the above examples suggest is that the only method which can be advised for general use is the method of stratified random sampling. If the conditions of the practical work allow, then the elements of the sampling units should be individuals. Otherwise we may sample groups, which however should be as small as possible."

The above is a very inadequate summary of this highly original paper, which Neyman published in three versions: first in a 1933 Polish pamphlet of 123 pages; then in the presentation to the *Royal Statistical Society* discussed above; and finally in the 1937 lectures and conferences delivered at, and published by, the Graduate School of the U. S. Department of Agriculture in 1938.

The most important contributions of the paper are:

1. The extension of stratified random sampling to the case that the elements being sampled are groups rather than individuals.
2. The development of a probabilistic model for purposive selection that made it possible to assess its accuracy.
3. The realization that purposive selection consists of samples of very few, very large, sampling units, while on the contrary good accuracy requires many small units.

A fourth contribution, discussed mainly in the appendix, is Neyman's theory of confidence intervals. This will be considered in the next chapter.

Finally, it may be worth noting certain parallels between Fisher's theory of design and Neyman's theory of sampling. Most importantly, both insist on randomization providing the only reliable basis for dependable statistical inference. The other parallel is that between blocking (in design) and stratification (in sampling). Of course, the settings and details in the two fields of application are quite different. A detailed comparison and prehistory of Fisher's and Neyman's work in these two areas is provided by Fienberg and Tanur (1996).

Chapter 6
Estimation

6.1 Point Estimation

Throughout the nineteenth century, the most commonly used statistical procedure was estimation by means of least squares. In 1894, Karl Pearson broke new ground by proposing an alternative approach: the method of moments. Of this method, Fisher, in his fundamental paper of 1922 [18] (discussed in Sect. 1.5), wrote that it is "without question of great practical utility." On the other hand, he points out that it requires the population moments involved to be finite, and "that it has not been shown, except in the case of a normal curve, that the best values will be obtained by the method of moments." And he asserts that "a more adequate criterion is required." For this purpose, he proposes the method of maximum likelihood.

In several sections, Fisher then compares the efficiency of the method of moments relative to that of maximum likelihood for a variety of classes of distributions, including those of the Pearson curves. For the latter, he finds that the relative efficiency of the method of moments exceeds 80% only when the distribution is close to the normal. The comparison shows clearly how very superior maximum likelihood is to Pearson's approach.

On the general subject of point estimation, Fisher, in his book, "Statistical Methods and Scientific Inference," points out that estimates are of little value without an idea of their precision. He writes that,

> A distinction without a difference has been introduced by certain writers who distinguish "point estimation," meaning some process of arriving at an estimate without regard to its precision, from "interval estimation," in which the precision of the estimate is to some extent taken into account. "Point estimation" in this sense has never been practiced either by myself, or by my predecessor Karl Pearson, who did consider the problem of estimation in some of its aspects, or by his predecessor Gauss of nearly one hundred years earlier, who laid the foundations of the subject.

Despite this disclaimer, Fisher's early writing suggests that originally he viewed estimation as point estimation, much as we do now. Thus, in his basic paper of 1922 he defines the problems of estimation as follows:

E.L. Lehmann, *Fisher, Neyman, and the Creation of Classical Statistics*,
DOI 10.1007/978-1-4419-9500-1_6, © Springer Science+Business Media, LLC 2011

> These involve the choice of methods of calculating from a sample statistical deviates, or as we shall call them statistics, which are designed to estimate the values of the parameters of a hypothetical population.

Fisher returns to the problem of estimation three years later in a paper entitled "Theory of statistical estimation." He now states,

> Problems of estimation arise when we know, or are willing to assume, the form of the frequency distribution of the population, as a mathematical function involving one or more unknown parameters and wish to estimate the values of these parameters by means of the observational record available. A statistic may be defined as a function of the observations designed as an estimate of any such parameter.

In both papers, he thus calls "statistics" what today we would consider to be point estimates. He then proceeds to study properties of these statistics such as consistency, sufficiency, efficiency, and so on. We shall not follow this work here, but turn to the tangled and complicated story of Fisher's and Neyman's approaches to estimation by interval or distribution.

6.2 Fiducial Inference

Fisher's ultimate solution to the problem of estimation (and to statistical inference in general) was what he called fiducial inference. His first publication on this new approach to inference was a 1930 paper "Inverse probability." The paper begins with a critique of the inverse (Bayesian) method. This section ends with Fisher's asking:

> If, then, we follow writers like Boole, Venn and Chrystal in rejecting the inverse argument as devoid of foundation and incapable even of consistent application, how are we to avoid the staggering falsity of saying that however extensive our knowledge of the values of x may be, yet we know nothing and can know nothing about the values of θ?

As an answer to this question, Fisher proposes the concept of likelihood and in particular estimation by maximum likelihood. After some discussion of these ideas which stresses the fact that likelihood is not probability, Fisher turns to another answer by stating that "there are, however, certain cases in which statements in terms of probability can be made with respect to the parameters of the population."

Although he restricts the argument to maximum likelihood estimators T, it is in fact much more general. Fisher assumes T to have a continuous distribution depending on a single parameter θ. He then considers the probability

$$P_\theta(T \leq c) = F(T, \theta) = P$$

and now argues,

> If we give to P any particular value such as .95, we have a relationship between T and θ such that T is the 95 per cent value corresponding to a given θ, and this relationship implies the perfectly objective fact that in 5 per cent of samples T will exceed the 95 per cent value corresponding to the actual value of θ… .

And now comes the crucial step:

> To any value of T there will moreover be usually a particular value of θ to which it bears this relationship; we may call this "the fiducial 5 per cent value of θ" corresponding to given T. If, as usually if not always happens, T increases with θ for all possible values, we may express the relationship by saying that the true value of θ will be less than the fiducial 5 per cent value corresponding to the observed value of T in exactly 5 trials in 100… . This then is a definite probability statement about the unknown parameter θ.

Fisher then illustrates this argument with an example and concludes by saying that,

> Generally the fiducial distribution of a parameter θ for given statistic T may be expressed as
>
> $$\mathrm{d}f = -\frac{\partial}{\mathrm{d}\theta} F(T,\theta)\mathrm{d}\theta$$
>
> while the distribution of the statistic T for a given value of the parameter is
>
> $$\mathrm{d}f = -\frac{\partial}{\mathrm{d}T} F(T,\theta)\mathrm{d}T.$$

The paper concludes with a comparison of this fiducial approach with that of inverse probability.

The passage referred to above as "the crucial step" caused great difficulty to later readers. The argument starts with T being a random variable and θ an unknown parameter and ends with T a fixed value and "a probability statement about the unknown parameter θ." This phrase suggests that θ has now become a random variable, but it could also be just a rhetorical flourish. An argument for the latter interpretation is that at the time Fisher and Neyman thought that their approaches were the same.

If this is what Fisher had thought in 1930, he certainly changed his mind later. In his 1956 book SMSI, he stated unequivocally (p. 51),

> The fiducial argument uses the observations to change the logical status of the parameter from one in which nothing is known of it, and no probability statement about it can be made, to the status of a random variable having a well-defined distribution.

After his 1930 paper, Fisher next mentions fiducial inference in a 1935 paper on "The logic of inductive inference," some aspects of which were discussed in Sect. 4.6 above. The new feature is the identification of the fiducial limits with the set of parameters θ_0 for which the hypothesis $\theta = \theta_0$ is accepted at the given level. This interpretation had already been suggested by Neyman in the Appendix to his 1934 paper.

In 1935, Fisher also published a more extensive treatment of "the fiducial argument in statistical inference," in which he emphasizes an aspect that had been implicit in the use of maximum likelihood estimates in his 1930 paper. This is the requirement that "the statistics used contain the whole of relevant information which the sample provides." As an example, he mentions a statistic t', in which the usual denominator of t has been replaced by one derived from the mean error. He claims that this would be "logically equivalent to rejecting arbitrarily a portion of the observational data."

He continues by saying that,

> Dr. J. Neyman has unfortunately attempted to develop the argument of fiducial probability in a way which ignores the results from the theory of estimation, in the light of which it was originally put forward. His proofs, therefore, purport to establish the validity of a host of probability statements many of which are mutually inconsistent.

Fisher neglects to point out that Neyman too strives to obtain a preferred solution, namely the one leading to the intervals that are shortest in some suitable sense.

In the last sections of the paper, Fisher extends the fiducial argument to several parameters and then applies it to the Behrens–Fisher problem.

As discussed in Sect. 4.3 above, this is the problem of testing that the means μ and μ' of two normal distributions with unknown variances are equal. As we have seen, conditionally given the two samples, Fisher treats μ and μ' as two independent random variables with known distributions from which he derives the distribution of $\mu' - \mu$.

We shall return to the subject of fiducial probability in Sect. 6.5.[1]

6.3 Confidence Intervals

After having sketched his theory of confidence intervals in the appendix of his 1934 sampling paper, Neyman provides a full account in a 1937 paper to which he gives the title "Outline of a theory of statistical estimation based on the classical theory of probability." He motivates his approach by noting that "the practical statistician required some measure of the accuracy of the estimate T. The generally accepted method of describing this accuracy consists in calculating the estimate, say S_T^2, of the variance V_T of T and in writing the result of all the calculations in the form $T \pm S_T \ldots$. This shows that what the statisticians really had in mind in problems of estimation is not the idea of a unique estimate, but that of two estimates, having the form, say,

$$\underline{\theta} = T - k_1 S_T \quad \text{and} \quad \overline{\theta} = T + k_2 S_T,$$

where k_1 and k_2 are certain constants, indicating the limits between which the true value of θ presumably falls."

He then considers the probability that the true parameter value lies between these limits. He points out that it is legitimate to consider this probability since $\underline{\theta}$ and $\overline{\theta}$ are random variables, and only the unknown parameter value θ is a constant. He then requires that the probability of

$$\underline{\theta} \le \theta \le \overline{\theta} \tag{*}$$

[1] For more on the fiducial approach, see Zabell (1992).

be equal to a preassigned level α, no matter what is the true value of θ, or, if they are present, of any nuisance parameters. He calls $\underline{\theta}$ and $\overline{\theta}$ the lower and upper confidence limits and α the confidence coefficient. After calculating the values of $\underline{\theta}$ and $\overline{\theta}$, Neyman writes, the practical statistician must state that (*) holds. This is justified he asserts, because when following this rule, "in the long run he will be correct in 99% (the assumed value of α) of all cases." He then adds two very important comments.

First he points out that "for this conclusion to be true, it is not necessary that the problem should be the same in all the cases. For instance during a period of time the statistician may deal with a thousand problems of estimation and in each case the parameter θ to be estimated and the probability law of the x's may be different."

As long as the functions $\underline{\theta}$ and $\overline{\theta}$ are properly calculated, Neyman states, and correspond to the same value of α, "the probability of their resulting in a correct statement will be the same, α. Hence the frequency of actually correct statements will approach α." This very important point has frequently been overlooked by later commentators on Neyman's work.

On the second aspect, Neyman writes that,

> It will be noticed that in the above description the probability statements refer to the problems of estimation with which the statistician will be concerned in the future.

He then raises the question of what can be concluded once the sample has been drawn.

Suppose $\underline{\theta} = 1$ and $\overline{\theta} = 2$. Can we then say "that in this particular case the probability of the true value θ falling between 1 and 2 is equal to α?" And he replies:

> The answer is obviously in the negative. The parameter θ is an unknown constant and no probability statement concerning its value may be made, that is except for the hypothetical and trivial ones that the probability of $1 \le \theta \le 2$ equals 1 if θ lies between these limits, and 0 if it doesn't.

The next section concerns the fact that a confidence set for θ can be viewed as the totality of parameter values θ_0 for which the hypothesis H: $\theta = \theta_0$ is accepted at the given level. This equivalence provides a method for constructing confidence sets, which then is illustrated on a number of examples.

Finally, the question is taken up of what are "best" confidence intervals. Neyman calls confidence intervals shortest if they minimize the probability of covering false values of θ_0. He points out that they are uniformly shortest (i.e., minimize the probability of covering all false values) if the corresponding hypothesis tests are uniformly most powerful.

6.4 Some Priority Considerations

Fisher first published his fiducial concept in 1930 [84]; Neyman's first publication in English of his theory of confidence intervals occurred in 1934 in the appendix of his paper on sampling. There, he says of his confidence limits $\theta_1(x)$ and $\theta_2(x)$ that

> they are what R. A. Fisher calls the fiducial limits of θ. Since the word 'fiducial' has been associated with the concept of 'fiducial probability,' which has caused the misunderstandings I have already referred to, and which in reality cannot be distinguished from the ordinary concept of probability, I prefer to avoid the term and call the intervals $[\theta_1(x), \theta_2(x)]$ the confidence intervals corresponding to the confidence coefficient ε.

After describing how to construct these intervals, Neyman adds the following footnote:

> The theory developed by Fisher runs on somewhat different lines. It applies only to the case just described when we know the distribution of x depending on only one unknown character [i.e., parameter]. The method I am using seems to have the advantage that it allows an easy generalization to the case where there are many unknown parameters… .

Fisher was one of the discussants. Talking about inference without inverse probability, he said

> the particular aspect of this work, of which Dr. Neyman's paper was a notable illustration, was the deduction of what Dr. Fisher had called fiducial probability. Dr. Neyman had not used this term, which he suggested had been misunderstood, but he used instead the term 'confidence coefficient'… . When Dr. Neyman said 'it really cannot be distinguished from the ordinary concept of probability', Dr. Fisher agreed with him; and that seemed to him a reason for calling it a probability rather than a coefficient. He qualified it from the first with the word *fiducial* to show that it was probability by the fiducial method of reasoning, then unfamiliar and not by the classical method of *inverse* probability. Dr. Neyman qualified it with the word *confidence*. The meaning was evidently the same… .

Two facts stand out from this interchange:

1. Both authors believe they are talking about the same concept, though derived in different ways.
2. Neyman acknowledges Fisher's priority.

It therefore comes as a surprise that in his basic 1937a paper on confidence intervals, Neyman does not at all refer to Fisher's earlier work on the subject. It seems likely that he justified this in that by then he no longer felt that they were dealing with the same concept.

However, Fisher bitterly complained about this absence of any acknowledgment of his earlier work. In a letter of November 9, 1942 to Maurice G. Kendall (Bennett 1990, p. 181), he wrote:

> The particular innovation of reasoning which you specified in the words, "The essential concept of the theory of confidence intervals, if I have understood it correctly, is the replacement of these specified limits by … functions of the sample numbers," is to be found quite explicitly in the paper to which I referred you (Fisher's 1930 paper on "inverse probability"). I may say it took a lot of private coaching by me to get this simple modification into Neyman's head. When he had grasped it, he proceeded, with Pearson's help, to expound it to the world with the minimum of reference to its origin.

Later, Neyman discovered, and acknowledged in the 1952 second edition of his "Lectures and Conferences," that his theory of confidence intervals had important forerunners.

On page 221, he states that "we shall make a very brief review of early papers of several authors in which one can discern the germs of the theory of confidence intervals."

He first mentions that,

> The idea of estimation by confidence intervals is very clearly and faultlessly stated in a few last sentences of a paper by Hotelling published in 1931. However, the statement of this idea was not followed by an attempt to develop a systematic theory. The relevant passage is very brief and its brevity must have contributed to its being overlooked by many readers, including myself.

Neyman then quotes the passage in question. Actually, although Neyman apparently did not know this, Hotelling, in a joint paper with Holbrook Working, had already, in 1929, obtained confidence bands for regression curves (Working and Hotelling 1929).

Next, Neyman writes, "it would be necessary to refer to papers by R. A. Fisher concerned with the so-called 'fiducial argument.' The early papers of Fisher given to this subject definitely suggest the idea of confidence intervals. Later on, however, there appeared to be a substantial difference between the two theories." (In a later section, Neyman discusses the history and connection of fiducial and confidence limits in more detail.)

Finally, Neyman mentions that "before either Hotelling or Fisher, the idea of confidence intervals is found in papers by Edwin B. Wilson and Stanislas Millot. Both authors are concerned with estimating the probability *p* ..." [in a binomial distribution].

What is clear from these references is that the idea of confidence intervals was in the air. Neyman's great achievement was to elevate the discussion from the consideration of isolated examples to a general theory. In addition, this theory included a definition of optimal or "shortest" intervals that minimized the probability of covering false values.

6.5 Fisher's Concept of Probability

One way of trying to understand Fisher's fiducial probability is to study what he meant by probability.

Fisher formulated his idea of probability as early as 1922 in his fundamental paper "On the mathematical foundations of theoretical statistics." There, he defined the task of the statistician as the reduction of data, and stated that "this object is accomplished by constructing a hypothetical infinite population, of which the actual data are regarded as constituting a random sample."

He then explains the meaning of probability:

> When we speak of the probability of a certain object fulfilling a certain condition, we imagine all such objects to be divided into two classes, according as they do or do not fulfill the condition. This is the only characteristic in them of which we take cognisance.

He goes on to say that probability

> is a parameter which specifies a simple dichotomy in an infinite hypothetical population, and it represents neither more nor less than the frequency ratio which we imagine such a population to exhibit.

Fisher illustrates his definition with the statement that the probability of throwing a five with a die is one-sixth. By this we mean, he says, that in

> a hypothetical population of an infinite number of throws, with the die in its original condition, exactly one-sixth will be fives. Our statement will not then contain any false assumption about the actual die, as that it will not wear out with continued use, or any notion of approximation, as in estimating the probability from a finite sample, although this notion may be logically developed once the meaning of probability is apprehended.

To see in more detail how Fisher thought about probability, we shall take a closer look at one of the several papers he wrote on the subject. It is a late paper (1959) with the title "Mathematical probability in the natural sciences."

The first section contains the following statement which is central to his view. Of probability statements, he writes that all they

> assert is that the exact nature and degree of uncertainty is just *as if* we know [the subject] to have been chosen at random from [a population in the real world]. The subject of probability statements … is singular and unique; we have some degree of uncertainty about its value, and it so happens that we can specify the exact nature and extent of our uncertainty by means of the concept of mathematical probability as developed by the great minds of the 17th Century Fermat, Pascal, Leibnitz, Bernoulli and their immediate followers.

In the next paragraph, he illustrates:

> The type of uncertainty which the new concept was capable of specifying with exactitude was that which confronts a gambler in a game played fairly, and with perfect apparatus. The probability statements refer to the particular throw or to the particular shuffling of the cards, on which the gambler lays his stake. The state of rational uncertainty in which he is placed may be *equated* to that of the different situation which can be imagined in which his throw is chosen at random out of an aggregate of throws, or of shuffling, which might equally well have occurred, though such aggregates exist only in the imagination.

The most striking feature of this explanation is that Fisher's probability concept refers to singular events and is a state of mind (rational uncertainty), what today would be called an epistemic probability.

In the next section, Fisher starts over again and considers the problem from another angle. He states that "the requirements of a correct statement are only three".

The first is "a conceptual reference set, which may be *conceived* as a population of possibilities of which an exactly known fraction possess some chosen characteristic." As an example, he mentions the set of possible throws with a die, [which] may be conceived to have exactly 1/6 aces.

The second condition is that "it must be possible to assert that the subject of the probability statement belongs to this set." This requirement, he states, "puts our probability statement into the real world."

Finally, Fisher requires that "no subset can be *recognized* having a different probability. Such subsets must always exist; it is required that no one of them shall be recognizable. This is a postulate of ignorance."

Against this background, Fisher devotes the next section to explain his concept of fiducial probability, of which he states that it is a probability in the sense he has explained this notion, and differs only in the way it was derived "by the particular argument to which the term fiducial has been applied."

As in his other expositions of this controversial idea, Fisher provides no general account but discusses it in terms of an example. In this case, he uses the estimation of the mean μ of a normal distribution based on the sample mean $\overline{X}$ and estimate S^2 of the variance, so that – as he points out – the quantity

$$T = (\mu - \overline{X}) / S$$

has a distribution which is independent both of the unknown parameter μ and of the unknown variance. "Therefore the value t_P can be tabulated such that for any random sample from any normal distribution," he writes,

$$\text{“}P\,[t < t_P] = P$$

for all values of P from 0 to 1."

Given N observations, one can then calculate $\overline{x}$ and s, "and substitute their numerical values in the expression for t so obtaining the probability statement

$$P[\mu < \overline{x} + st_P] = P.\text{”}$$

He continues, "The subject of these probability statements is the unknown μ, a property of the real world to be determined by observation and experiment like an atomic weight."

Having made this bald and unorthodox statement, Fisher then asks: "on what conditions do a system of statements such as I have inferred by the fiducial argument represent genuine statements of probability in the real world?" And he explains:

> As regards the Reference Set we may recognize that the triad of values (μ, $\overline{x}$, s) must exist for every sample from every normal population, that for some of these samples but not for all, the inequality
>
> $$\mu < \overline{x} + st_P$$
>
> is satisfied, and whatever may be the true values of the mean and variance of the population sampled, the proportion of random samples which satisfy the inequality is exactly P.

He goes on to say that "the second requirement, that our data really belong to this reference set, depends … on the adequacy of [the] experimental design."

Finally, Fisher argues that no subset can be "recognized within the general set." His proof is unconvincing and the assertion was in fact later shown to be false.

Unfortunately, Fisher does not complete this discussion of fiducial probability by connecting it with the interpretation of single event probability he had set out in the first section of the paper. Had he done so, he might have stated that given the particular values of μ and the observed values of $\overline{x}$ and s, the probability p that $\mu < \overline{x} + st_P$ measures the resulting "state of [the] observer's rational uncertainty" concerning this inequality.

This is at least one view of fiducial probability that is consistent with the ideas expressed in this paper. Whether either then or at other times Fisher really thought of it in this way seems impossible to tell.

It should be pointed out that in any case such an interpretation does not justify treating μ (conditionally, given $\overline{x}$ and s) as a random variable with a known distribution as Fisher did, for example, in his solution of the Behrens–Fisher problem.

It is a curious and somewhat ironic fact that the interpretation of the probability of a random event once it has been observed as epistemic, suggested in the first part of Fisher's paper, is the way many users think about confidence intervals. As pointed out in Sect. 6.3 above, for Neyman probability referred to the frequency of future events. Probability as a state of mind was a concept he strongly opposed.

Chapter 7
Epilog

7.1 A Review

As stated in the Preface, it is the aim of this book to trace the creation of classical statistics, and to show that it was principally the work of two men, Fisher and Neyman. Since the main story is somewhat lost in the details, let us now review their contributions to hypothesis testing, estimation, design, and the philosophy of statistics.

7.1.1 Hypothesis Testing

After a small-sample ("exact") approach to testing was initiated by Gosset ("Student") in 1908 with his t-test, Fisher in the 1920s, under frequent prodding by Gosset, developed a battery of such tests, all based on the assumption of normality. These tests (based on the t, F, and χ^2 distributions) today still constitute the bread and butter of much of statistical practice.

Fisher's tests were based solely on his intuition. The right choice of test statistics was obvious to him. A theory that would justify his choices was developed by Neyman and Pearson in their papers of 1928 and 1933.

Their basic idea was to fix the level of the test at a predetermined value α, independent of any nuisance parameters, and – subject to this condition – to maximize the power of the test, i.e., its probability of rejecting the hypothesis when it is false. It turned out that in all the initial simple cases considered by Fisher, the tests he had proposed were optimal in this sense.

The contributions by Fisher on the one hand and by Neyman-Pearson on the other, thus fitted seamlessly together to provide an enormously useful methodology with a persuasive theoretical underpinning.

More important perhaps than the optimality properties were the practical uses of the concept of power. It provided a measure of the effectiveness of the test and a way

E.L. Lehmann, *Fisher, Neyman, and the Creation of Classical Statistics*,
DOI 10.1007/978-1-4419-9500-1_7, © Springer Science+Business Media, LLC 2011

to determine what sample size was needed to make the test effective against the alternatives of interest.

Fisher's first reaction to the Neyman-Pearson theory was favorable but later he changed his mind, calling it in a letter of October 8, 1951 (Bennett 1990, p. 144) "that unnecessarily portentous approach to tests of significance represented by the Neyman and Pearson critical regions, etc." He continues to say that, "In fact I and my pupils throughout the world would never think of using them." In particular, Fisher never acknowledged that the concept of power might have any use.

Fisher's opposition to the Neyman-Pearson theory stemmed partly from the fact that for some more complicated problems such as the Behrens-Fisher problem, Fisher's fiducial theory led to a very different solution than the Neyman-Pearson approach.

A final issue, which turned out to be of great practical importance, concerned the choice of the significance level α. As pointed out in Sect. 2.2, Fisher, in his 1925 book SMRW, recommended 5% as "a conventional line," and throughout in his work used 5% or 1% if a more stringent requirement was desired.

In contrast, Neyman and Pearson, in their basic paper of 1933, denote the level by ε but never suggest a particular value for it. The reason for this can be seen in their general discussion (in Part II of their paper) of the two kinds of error: false rejection and false acceptance. They state that,

> From the point of view of mathematical theory all that we can do is to show how the risk of the errors may be controlled and minimized. The use of these statistical tools in any given case, in determining just how the balance should be struck, must be left to the investigator.

This suggests that they felt in determining ε, consideration should be given to the power that could be achieved with this value.

In his late, 1956, book SMSI, Fisher protested that "no scientific worker has a fixed level of significance at which from year to year, and in all circumstances, he rejects hypotheses; he rather gives his mind to each particular case, and his ideas." However, his early recommendation and life-long practice prevailed. The desire for standardization trumped the advantages of considering each case on its own merits.

It is a historical curiosity that already Jacob Bernoulli, in his Ars Conjectandi, published posthumously in 1713, wrote,

> It would be useful if definite limits for moral certainty were established by the authority of the magistry. For instance, it might be determined whether 99/100 of certainty suffices or whether 999/1000 is required. Then a judge would not be able to favor one side, but would have a reference point to keep constantly in mind in pronouncing a judgment.

Fisher's 5% has now supplied such a reference point.

7.1.2 Design

The contributions of Fisher and Neyman-Pearson to the construction of a fairly cohesive subject of hypothesis testing were complementary. Fisher laid the groundwork by developing a methodology of testing. Neyman and Pearson then created a theory to support it.

When it came to design, the contributions of Fisher and Neyman were made in parallel, putting forward very similar ideas, such as the need for randomization, and the desirability of randomizing within fairly uniform strata. However, they developed these ideas in different contexts, Fisher in agricultural experimentation, Neyman in the sampling of human populations. Despite the similarity of those two lines of work, many special features kept them apart, and they are considered two separate fields.

7.1.3 Interval Estimation

Here again, although both authors contributed, clearly the origin of their work was Fisher's 1930 paper, "Inverse Probability." Neyman, in his 1934 paper, was intending to generalize Fisher's approach and to free it from what seemed to Neyman unnecessary restrictions. At first, Fisher and Neyman agreed that they were talking essentially about the same thing, although from slightly different points of view. However, starting in 1935, Fisher's ideas about fiducial inference veered off in a different direction and it became clear to both authors that fiducial inference and confidence intervals were two very different concepts. While Neyman's theory was very clear and has become a standard part of classical statistics, few were able to follow Fisher's fiducial argument.

And yet this does not tell the whole story. For Neyman, probability was long-run frequency. Hence, once the numerical values of the confidence limits were known, the only probability statements about the parameter still possible were 1 if the interval covered the true value, 0 if it did not. However, though Neyman repeatedly stressed this point, many users did not accept it. For them, the predata probability of coverage became their postdata degree of certainty that the (now fixed and numerically known) interval contained the unknown parameter. This was exactly Fisher's interpretation expressed at the beginning of his 1959 paper (quoted in Sect. 6.3 above). Thus, while Neyman's confidence intervals became the accepted solution of the estimation problem, in practice they were often interpreted in a way that was much closer to Fisher's view than Neyman's.

7.1.4 The Role of Statistics in the Philosophy of Science

Both Fisher and Neyman believed that they had made important contributions to the philosophy of science, but each felt that the other's views were completely wrongheaded.

Fisher felt that with his concepts of likelihood and fiducial inference, he had made great progress in the age-old problem of induction.

"In inductive reasoning," he writes in 1932,

> we attempt to argue from the particular, which is typically a body of observational material, to the general, which is typically a theory applicable to future experience. In statistical

> language we are attempting to argue from the sample to the population, from which it was drawn. Since recent statistical work [i.e., his own] has shown that this type of argument can be carried out with exactitude in a usefully large class of cases by means somewhat different from those of the classical theory of probability, it may be useful briefly to restate the logical and mathematical distinctions which have to be drawn.

The importance Fisher attaches to these ideas can be seen from the fact that in 1934 he made a major presentation to the *Royal Statistical Society* entitled, "The Logic of Inductive Inference." Here, he states that,

> It would not be surprising or exceptional to find mathematicians of this class [i.e., who have been trained … almost exclusively in the technique of deductive reasoning] ready to deny at first sight that rigorous inferences from the particular to the general were even possible… . It will be sufficient here to note that the denial implies, qualitatively, that the process of learning by observation, as experiment, must always lack real cogency.

Inductive inference was central to Fisher's thinking and he probably took more pride in his having legitimized induction and, as he believed, put it on a firm foundation, than in any of his other contributions.

Neyman, for the first time, in 1938, in a French presentation of his theory of confidence sets, states his objections to the idea of inductive reasoning. He asserts that (my translation):

> If, after having made the observations and calculated the [confidence] limits, the physician claims that [the unknown parameter lies within these limits], the procedure which has led to this claim can not be called inductive reasoning.

He then explains the difference between what we can know and what we can believe:

> It seems to me that we can know only (1) the results of experiments that have already been carried out, and (2) the consequences of some definitions and postulates, provided such consequences have been proved. Our attitude toward any assertion other than (1) and (2) can only be described by the terms "belief" or "doubt."

Neyman goes on to state his own opposing point of view:

> The only reason for the physician to decide to assert that … [the parameter] lies between the confidence limits is that in a long series of empirical results the interval will cover the parameter in about $\alpha = .95$ of cases.
>
> But deciding to assert does not mean either "knowing" or "believing." It is a voluntary act preceded by some experience and some deductive arguments… . Consequently it seems to me that the term "inductive reasoning" does not correspond to the nature of the procedure which begins by certain postulates concerning the observable variables [i.e., by postulating a model], and ends up by an assertion. More generally I doubt that there exist cases in which the application of the term to a statistical method is later justified. If one wants a special term to describe these methods and in particular to describe the decision that the confidence limits are valid, one could perhaps propose "inductive behavior" ["comportement inductif"].

On one subject, Fisher and Neyman agreed. Fisher, after 1922, and Neyman, after 1937, were united in their strong opposition to the use of prior distributions (unless they were based on substantial empirical evidence).

7.2 Further Developments

The strength of a new discipline can be judged in part by its usefulness to the work of later generations. From this point of view, the new discipline of statistics was enormously successful. Fisher's t- and F- tests, Neyman's confidence intervals, and the ideas of randomization and blocking, are being used the world over, not only in agriculture and biology, but in nearly every area of human activity.

However, the vitality of a new discipline must also be measured by its ability to grow, to extend to new situations, and to develop beyond its original formulation. In the present section we shall briefly discuss some areas of such further developments.

7.2.1 Multivariate Analysis

An obvious extension of Fisher's tests for univariate normal samples is to the case of samples from multivariate normal distributions. Fisher himself initiated such an effort with his early work on correlation coefficients and, in 1931, Hotelling generalized Student's t-test to the multivariate case. New issues arose such as testing for independence, principal components, and canonical correlations. Multivariate analysis became a flourishing subject of its own.

7.2.2 Bayesian Inference

It seems ironic that one of the most significant developments after Fisher and Neyman had established their foundations was to rejuvenate an approach they both had strongly opposed and thought to have vanquished: inverse probability. The story is complicated and we shall only mention a few highlights.

The nineteenth century approach to inverse probability, championed particularly by Laplace, considered the prior distribution to represent complete ignorance. This concept, now called objective Bayes, was taken up and improved by the Cambridge geophysicist Harold Jeffreys, culminating in his 1939 book, "Theory of Probability."

A different Bayesian approach, called subjective, was proposed by Ramsey (1926) and Bruno de Finetti in the 1930s.[1] It considered probability as a measure of a person's subjective degree of uncertainty about a situation. This view came into its own with the publication in 1954 of L. J. Savage's book, "Foundations of Statistics," in which he derives the existence of such subjective probabilities from a few, quite plausible, axioms.

[1] For references to early work of Ramsey and de Finetti, see Forcina (1982).

7.2.3 Statistical Decision Theory

Both Fisher and Neyman thought of testing, estimation, and design as three separate theories, which together constitute the field of statistical inference. Starting in the late 1930s and culminating in his 1950 book, "Statistical Decision Functions," Abraham Wald conceived of a very general framework, of which these theories were special cases. His formulation was so abstract that one might not expect it to produce any substantive results. However, one would be wrong. Wald proved that any reasonable statistical procedure is a Bayes procedure corresponding to some prior distribution (or a suitable limit of such procedures). In addition, he found a characterization of minimax procedures (i.e., of procedures that minimize the maximum risk). He showed that they are Bayes solutions corresponding to a least favorable prior or at least favorable sequence of priors.

Mimimax theory turned out to be an influential unifying approach, but the impact of Wald's first result mentioned above was even greater. It suggested a way to generate good statistical procedures: Find the Bayes solution for some reasonable prior, and then examine its frequentist properties to see whether it was satisfactory. The difficulty with this proposal is the choice of a reasonable prior. The search for such priors, now called reference priors, became an important topic for Bayesians. Ultimately, this approach bridges the divide between Bayesian and frequentist methodology, although it is, of course, unsatisfactory to subjective Bayesians.

7.2.4 Nonparametric Inference

A limitation of Fisher's small-sample tests was the assumption of the normality of the underlying distributions. Fisher himself had pointed to randomization tests as a way out of this problem, but they were difficult to use. An alternative, to replace the observations by their ranks, was proposed for testing independence by Hotelling and Pabst in 1936. The idea did not get much traction until 1945, when Frank Wilcoxon published his one- and two-sample rank tests as alternatives to the corresponding t-tests. Because of their great simplicity, they quickly became popular, particularly in the social sciences. There was, however, a concern: that these tests were very inefficient because replacing the original observations by their ranks discarded much useful information.

To measure this loss of efficiency, Pitman in 1947 introduced the concept of asymptotic relative efficiency (ARE), and obtained the surprising result that in the normal shift model, the ARE of Wilcoxon to t is $3/\pi=0.955$. Thus, in the situation for which the t-test was designed and is optimal, the efficiency loss is quite small. Later, it was shown that for shift models with heavy-tailed distributions, Wilcoxon tends to be much more efficient than t, while for no distributions is it much less efficient.

These results turned rank-based versions of the normal theory t- and F-tests into serious competitors. In addition, point and interval estimates derived from these rank tests shared their efficiencies. Thus, at least for simple problems, a nonparametric methodology became a useful alternative to the parametric methodology Fisher and Neyman had developed. (For a detailed comparison of the advantages and disadvantages of the two approaches, see Lehmann (2009).)[2]

7.2.5 Robustness

The classical Fisher-Neyman theory of statistical inference assumed that the observations were normally distributed or that their distribution belonged to some other given parametric family. Nonparametrics, in contrast, assumed that essentially nothing was known about the form of the underlying distribution(s). In 1964, Peter Huber proposed an intermediate formulation. He suggested that it is frequently realistic to assume that the underlying distribution while not exactly normal is at least approximately so or, more generally, that it lies in a suitable neighborhood of some given distribution. A procedure is robust in Huber's sense if small deviations from the model result in small changes in the performance of the procedure.

In this first 1964 paper, Huber developed a theory of robust estimation of a location parameter. He later extended it to other situations, among them the problem of testing the neighborhood of one distribution against that of another. For this case, Huber proved a beautiful generalization of the Neyman-Pearson Lemma. An important contribution to robustness theory was Frank Hampel's influence function.

Gradually, robustness became a subject in its own right, defined by Huber in his book, "Robust Statistics" (Huber 1981; Huber and Ronchetti 2009). The interest in the topic can be seen, for example, in the publication of several later books, such as Hampel, Ronchetti, Rousseeuw, and Stahel: "Robust Statistics: The Approach Based on Influence" (1986), Staudte and Sheather, "Robust Estimation and Testing" (1990), and Jureckova and Sen, "Robust Statistical Procedures" (1996).

These developments during the 1940s, 1950s, and 1960s, to which others could be added, give an indication of the vigor and success of the discipline that Fisher and Neyman had created.

[2] Recent advances in computer technology make randomization tests based on the actual values practically feasible, although not necessarily more efficient.

Appendix

Bibliography of the Publications of R.A. Fisher

This bibliography includes some publications which were unfortunately omitted from the bibliography given in the *Collected Papers*. These are indicated by the symbol # preceding the relevant citations. Several of these publications are displayed in full within this website and in these cases they have been given new CP numbers (e.g., CP92A).

(Adapted from Volume 1 of the *Collected Papers of R.A. Fisher*, with several changes and additions prefixed by #).

I Books

**Statistical Methods for Research Workers*. Edinburgh: Oliver & Boyd, 1925, 1928, 1930, 1932, 1934, 1936, 1938, 1941, 1944, 1946, 1950, 1954, 1958, 1970. Also published in French, German, Italian, Japanese, Spanish and Russian.

E.L. Lehmann, *Fisher, Neyman, and the Creation of Classical Statistics*,
DOI 10.1007/978-1-4419-9500-1, © Springer Science+Business Media, LLC 2011

The Genetical Theory of Natural Selection. Oxford: Clarendon Press, 1930; New York: Dover Pubns., 1958.
A Complete Variorum Edition. (With a foreword and notes by J.H. Bennett) Oxford: Oxford University Press, 1999.
see http:\\www.oup.co.uk/isbn/0-19-850440-3

**The Design of Experiments.* Edinburgh: Oliver & Boyd, 1935, 1937, 1942, 1947, 1949, 1951, 1960, 1966. Also published in Italian, Japanese and Spanish.

Statistical Tables for Biological, Agricultural and Medical Research. (With F. Yates) Edinburgh: Oliver & Boyd, 1938, 1943, 1948, 1953, 1957, 1963. Also published in Spanish and Portuguese.

The Theory of Inbreeding. Edinburgh: Oliver & Boyd, 1949, 1965.

Contributions to Mathematical Statistics. New York: Wiley, 1950.

**Statistical Methods and Scientific Inference.* Edinburgh: Oliver & Boyd, 1956, 1959 ; New York: Hafner, 1973.

> *Re-issued as *Statistical Methods, Experimental Design and Scientific Inference,* with a Foreword by F. Yates (edited by J.H. Bennett). Oxford: Oxford University Press, 1990.

II Papers reprinted in the Collected Papers of R.A. Fisher (Adelaide: University of Adelaide, 1971–1974)

1 Statistical and Mathematical Theory and Applications (excluding Genetics)

Year	Paper number	Title
1912	1	On an absolute criterion for fitting frequency curves. *Messeng. Math.,* 41: 155–160
1913	2	Applications of vector analysis to geometry. *Messeng. Math.,* 42: 161–178
1915	4	Frequency distribution of the values of the correlation coefficient in samples from an indefinitely large population. *Biometrika,* 10: 507–521
1916	7	Biometrika. *Eugen. Rev.,* 8: 62–64
1920	12	A mathematical examination of the methods of determining the accuracy of an observation by the mean error, and by the mean square error. *Mon. Not. Roy. Ast. Soc.,* 80: 758–770
1921	14	On the "probable error" of a coefficient of correlation deduced from a small sample. *Metron,* 1: 3–32
	15	Studies in crop variation. I. An examination of the yield of dressed grain from Broadbalk. *J. Agric. Sci.,* 11: 107–135
	16	Some remarks on the methods formulated in a recent article on the quantitative analysis of plant growth. *Ann. Appl. Biol.,* 7: 367–372
1922	18	On the mathematical foundations of theoretical statistics. *Phil. Trans.,* A, 222: 309–368
	19	On the interpretation of X^2 from contingency tables, and the calculation of P. *J. Roy. Statist. Soc.*, 85: 87–94
	20	The goodness of fit of regression formulae, and the distribution of regression coefficients. *J. Roy. Statist. Soc.*, 85: 597–612

(continued)

Year	Paper number	Title
	21	(With W.A. Mackenzie). The correlation of weekly rainfall. *Quart. J. Roy. Met. Soc.*, 48; 234–242
	22	(With H.G. Thornton and W.A. Mackenzie). The accuracy of the plating method of estimating the density of bacterial populations. *Ann. Appl. Biol.*, 9: 325–359
	23	Statistical appendix to a paper by J. Davidson on biological studies of *Aphis rumicis. Ann. Appl. Biol.*, 9: 142–145
1923	30	Note on Dr. Burnside's recent paper on errors of observation. *Proc. Camb. Phil. Soc.*, 21: 655–658
	31	Statistical tests of agreement between observation and hypothesis. *Economica*, 3: 139–147
	32	(With W.A. Mackenzie). Studies in crop variation. II. The manurial response of different potato varieties. *J. Agric. Sci.*, 13: 311–320
1924	34	The conditions under which X^2 measures the discrepancy between observation and hypothesis. *J. Roy. Statist. Soc.*, 87: 442–450
	35	The distribution of the partial correlation coefficient. *Metron*, 3: 329–332
	36	On a distribution yielding the error functions of several well known statistics. *Proc. Int. Cong. Math., Toronto*, 2: 805–813
	37	The influence of rainfall on the yield of wheat at Rothamsted. *Phil. Trans.*, B, 213: 89–142
	38	A method of scoring coincidences in tests with playing cards. *Proc. Soc. Psych. Res.*, 34: 181–185
	39	(With S. Odén). The theory of the mechanical analysis of sediments by means of the automatic balance. *Proc. Roy. Soc. Edinb.*, 44: 98–115
1925	42	Theory of statistical estimation. *Proc. Camb. Phil. Soc.*, 22: 700–725
	43	Applications of "Student's" distribution. *Metron*, 5: 90–104
	44	Expansion of "Student's" integral in powers of n^{-1}. *Metron*, 5: 109–120
	45	(With P.R. Ansell). Note on the numerical evaluation of a Bessel function derivative. *Proc. Lond. Math. Soc.*, Series 2, 24: liv-lvi
	46	Sur la solution de l'équation intégrale de M.V. Romanovsky. *C.R. Acad. Sci., Paris*, 181: 88–89
1926	48	The arrangement of field experiments. *J. Min. Agric. G. Br.,* 33: 503–513
	49	Bayes' theorem and the fourfold table. *Eugen. Rev.*, 18; 32–33
	50	On the random sequence. *Quart. J. Roy. Met. Soc.*, 52: 250
	51	On the capillary forces in an ideal soil: correction of formulae given by W.B. Haines. *J. Agric. Sci.*, 16: 492–503
1927	56	(With J. Wishart). On the distribution of the error of an interpolated value, and on the construction of tables. *Proc. Camb. Phil. Soc.*, 23: 912–921
	57	(With T. Eden). Studies in crop variation. IV. The experimental determination of the value of top dressings with cereals. *J. Agric. Sci.*, 17: 548–562
	58	(With H.G. Thornton). On the existence of daily changes in the bacterial numbers in American soil. *Soil Sci.*, 23. 253–259
1928	61	The general sampling distribution of the multiple correlation coefficient. *Proc. Roy. Soc.*, A, 121: 654–673
	62	On a property connecting the X^2 measure of discrepancy with the method of maximum likelihood. *Atti Cong. Int. Mat., Bologna*. 6: 95–100

(continued)

Year	Paper number	Title
	63	(With L.H.C. Tippett). Limiting forms of the frequency distribution of the largest or smallest member of a sample. *Proc. Camb. Phil. Soc.*, 24: 180–190
	64	Further note on the capillary forces in an ideal soil. *J. Agric. Sci.*, 18: 406–410
	65	(With T.N. Hoblyn). Maximum- and minimum-correlation tables in comparative climatology. *Geogr. Ann.*, 10: 267–281
	66	Correlation coefficients in meteorology. *Nature*, 121: 712
	67	The effect of psychological card preferences. *Proc. Soc. Psych. Res.*, 38: 269–271
1929	74	Moments and product moments of sampling distributions. *Proc. Lond. Math. Soc.*, Series 2, 30: 199–238
	75	Tests of significance in harmonic analysis. *Proc. Roy. Soc.*, A, 125: 54–59
	76	The sieve of Eratosthenes. *Math. Gaz.*, 14: 564–566
	77	A preliminary note on the effect of sodium silicate in increasing the yield of barley. *J. Agric. Sci.*, 19: 132–139
	78	(With T. Eden). Studies in crop variation. VI. Experiments on the response of the potato to potash and nitrogen. *J. Agric. Sci.*, 19: 201–213
	79	The statistical method in psychical research. *Proc. Soc. Psych. Res.*, 39: 189–192
	80	Statistics and biological research. *Nature*, 124: 266–267
1930	83	The moments of the distribution for normal samples of measures of departure from normality. *Proc. Roy. Soc.*, A, 130: 16–28
	84	Inverse probability. *Proc. Camb. Phil. Soc.*, 26: 528–535
	85	(With J. Wishart). The arrangement of field experiments and the statistical reduction of the results. *Imp. Bur. Soil Sci. Tech. Comm.*, 10. 23pp
1931	90	(With J. Wishart). The derivation of the pattern formulae of two-way partitions from those of simpler patterns. *Proc. Lond. Math. Soc.*, Series 2, 33: 195–208
	91	The sampling error of estimated deviates, together with other illustrations of the properties and applications of the integrals and derivatives of the normal error function. *Brit. Assn. Math. Tab.*, 1. xxvi-xxxv (3rd ed., xxviii-xxxvii, 1951)
	92	(With S. Bartlett). Pasteurised and raw milk. *Nature*, 127: 591–592
	# 92A	Principles of plot experimentation in relation to the statistical interpretation of the results. *Report of a Conference on the Technique of Field Experiments,* Rothamsted, 7 May 1931, 11–13
1932	95	Inverse probability and the use of likelihood. *Proc. Camb. Phil. Soc.*, 28: 257–261
1933	102	The concepts of inverse probability and fiducial probability referring to unknown parameters. *Proc. Roy. Soc.*, A: 139: 343–348
	103	The contributions of Rothamsted to the development of the science of statistics. *Annual Report Rothamsted Experimental Station*, 43–50
1934	108	Two new properties of mathematical likelihood. *Proc. Roy. Soc.*, A, 144: 285–307
	109	Probability, likelihood and quantity of information in the logic of uncertain inference. *Proc. Roy. Soc.*, A, 146: 1–8

(continued)

Year	Paper number	Title
	110	(With F. Yates). The 6 x 6 Latin squares. *Proc. Camb. Phil. Soc.*, 30: 492–507
	111	Randomisation, and an old enigma of card play. *Math. Gaz.*, 18: 294–297
	112	Appendix to a paper by H.G. Thornton and P.H.H. Gray on the numbers of bacterial cells in field soils. *Proc. Roy. Soc.*, B, 115: 540–542
	# 112A	Contribution to a discussion of J. Neyman's paper on the two different aspects of the representative method. *J. Roy. Statist. Soc.*, 97: 614–619
1935	123	The mathematical distributions used in the common tests of significance. *Econometrica*, 3: 353–365
	124	The logic of inductive inference. *J. Roy. Statist. Soc.*, 98: 39–54
	125	The fiducial argument in statistical inference. *Ann. Eugen.*, 6: 391–398
	126	The case of zero survivors in probit assays. *Ann. Appl. Biol.*, 22: 164–165
	127	Statistical tests. *Nature*, 136: 474
	128	Contribution to a discussion of J. Neyman's paper on statistical problems in agricultural experimentation. *J. Roy. Statist. Soc., Suppl.*, 2: 154–157, 173
	129	Contribution to a discussion of F. Yates' paper on complex experiments. *J. Roy. Statist. Soc., Suppl.*, 2: 229–231
1936	137	Uncertain inference. *Proc. Amer. Acad. Arts Sci.*, 71: 245–258
	138	The use of multiple measurements in taxonomic problems. *Ann. Eugen.*, 7; 179–188
	139	(With S. Barbacki). A test of the supposed precision of systematic arrangements. *Ann. Eugen.*, 7: 189–193
	140	The half-drill strip system agricultural experiments. *Nature*, 138: 1101
	141	"The coefficient of racial likeness" and the future of craniometry. *J. Roy. Anthropol. Inst.*, 66. 57–63
1937	148	(With E.A. Cornish). Moments and cumulants in the specification of distributions. *Rev. Inst. Int. Statist.*, 5: 307–322
	149	Professor Karl Pearson and the method of moments. *Ann. Eugen.*, 7: 303–318
	150	(With B. Day). The comparison of variability in populations having unequal means. An example of the analysis of covariance with multiple dependent and independent variates. *Ann. Eugen.*, 7. 333–348
	151	On a point raised by M.S. Bartlett on fiducial probability. *Ann. Eugen.*, 7: 370–375
1938	155	The statistical utilization of multiple measurements. *Ann. Eugen.*, 8: 376–386
	156	Quelques remarques sur l'estimation en statistique. *Biotypologie*, 6: 153–158
	157	On the statistical treatment of the relation between sea-level characteristics and high-altitude acclimatization. *Proc.Roy. Soc.*, B, 126: 25–29
	158	The mathematics of experimentation. *Nature*, 142: 442–443
	159	Presidential address, Indian statistical conference. *Sankhyã*, 4: 14–17
1939	162	The comparison of samples with possibly unequal variances. *Ann. Eugen.*, 9: 174–180
	163	The sampling distribution of some statistics obtained from non-linear equations. *Ann. Eugen.*, 9: 238–249

(continued)

Year	Paper number	Title
	164	A note on fiducial inference. *Ann. Math. Stat.*, 10: 383–388
	165	“Student”. *Ann. Eugen.*, 9: 1–9
1940	173	On the similarity of the distributions found for the test of significance in harmonic analysis, and in Stevens’s problem in geometrical probability. Ann. Eugen., 10: 14–17
	174	An examination of the different possible solutions of a problem in incomplete blocks. *Ann. Eugen.*, 10: 52–75
	175	The precision of discriminant functions. *Ann. Eugen.*, 10: 422–429
1941	181	The asymptotic approach to Behrens’s integral, with further tables for the d test of significance. *Ann. Eugen.*, 11; 141–172
	182	The negative binomial distribution. *Ann. Eugen.*, 11; 182–187
	183	The interpretation of experimental four-fold tables. *Science*, 94: 210–211
1942	186	New cyclic solutions to problems in incomplete blocks. *Ann. Eugen.*, 11: 290–299
	187	Completely orthogonal 9 x 9 squares – a correction. *Ann. Eugen.*, 11: 402–403
	188	The likelihood solution of a problem in compounded probabilities. *Ann. Eugen.*, 11: 306–307
	189	The theory of confounding in factorial experiments in relation to the theory of groups. *Ann. Eugen.*, 11. 341–353
	190	Some combinatorial theorems and enumerations connected with the numbers of diagonal types of a Latin square. *Ann. Eugen.*, 11: 395–401
1943	193	A theoretical distribution for the apparent abundance of different species. *J. Anim. Ecol.*, 12: 54–58
	194	Note on Dr. Berkson’s criticism of tests of significance. *J. Amer. Statist. Assn.*, 38: 103–104
	195	(With W.R.G. Atkins). The therapeutic use of vitamin C. *J. Roy. Army Med. Corps*, 83: 251–252
1945	202	A system of confounding for factors with more than two alternatives, giving completely orthogonal cubes and higher powers. *Ann. Eugen.*, 12: 283–290
	203	The logical inversion of the notion of the random variable. *Sankhyã*, 7: 129–132
	204	Recent progress in experimental design. In *L’application du calcul des probabilités*, 19–31. *Proc. Int. Inst. Intell. Coop., Geneva*, (1939)
	205	A new test for 2 x 2 tables. *Nature*, 156: 388
1946	207	Testing the difference between two means of observations of unequal precision. *Nature*, 158: 713
1947	211	The analysis of covariance method for the relation between a part and the whole. *Biometrics*, 3: 65–68
	212	Development of the theory of experimental design. *Proc. Int. Statist. Conf.*, 3: 434–439
1948	222	Conclusions fiduciaires. *Ann. Inst. Henri Poincaré*, 10: 191–213
	223	(With D. Dugué). Un résultat assez inattendu d’arithmétique des lois de probabilité. *C.R. Acad. Sci., Paris*, 227: 1205–1206
	224	Biometry. *Biometrics*, 4: 217–219

(continued)

Year	Paper number	Title
	# 224A	Answer to Question 14 on combining independent tests of significance. *The American Statistician*, 2: 30
1949	230	A biological assay of tuberculins. *Biometrics*, 5: 300–316
1950	236	The significance of deviations from expectation in a Poisson series. *Biometrica*, 6: 17–24
1951	242	Statistics. In *Scientific Thought in the Twentieth Century*, (ed. A.E. Heath), 31–55. London: Watts
	# 242A	Answer to Query 91 on interaction of quantity and quality in agricultural field trials. *Biometrics,* 7: 433–434
1952	247	Sequential experimentation. *Biometrics*, 8: 183–187
1953	249	Dispersion on a sphere. *Proc. Roy. Soc.*, A, 217: 295–305
	250	Note on the efficient fitting of the negative binomial. *Biometrics*, 9: 197–199
	251	The expansion of statistics. *J. Roy. Statist. Soc.*, A, 116: 1–6; *Amer. Sci.*, 42: 275–282, 293
1954	256	The analysis of variance with various binomial transformations. *Biometrics*, 10: 130–139
	257	Contribution to a discussion of a paper on interval estimation by M.A. Creasy. *J. Roy. Statist. Soc.*, B, 16: 212–213
1955	261	Statistical methods and scientific induction. *J. Roy. Statist. Soc.*, B, 17: 69–78
	# 261A	Answer to Query 114 on the effect of errors of grouping in an analysis of variance. *Biometrics,* 11: 237
1956	264	On a test of significance in Pearson's *Biometrika Tables* (no. 11). *J. Roy. Statist. Soc.*, B, 18: 56–60
	265	(With M.J.R. Healy). New tables of Behrens' test of significance. *J. Roy. Statist. Soc.*, B, 18: 212–216
1957	267	The underworld of probability. *Sankhyã*, 18: 201–210
	268	Comment on the notes by Neyman, Bartlett and Welch in this Journal. (18, 288–302) *J. Roy. Statist. Soc.*, B, 19: 179
	269	Dangers of cigarette-smoking. *Brit. Med. J.*, 2: 43
	270	Dangers of cigarette-smoking. *Brit. Med. J.*, 297–298
1958	272	The nature of probability. *Centennial Rev*. 2: 261–274
	273	Mathematical probability in the natural sciences. *Proc. 18th Int. Congr. Pharmaceut. Sci.*; *Metrika*, 2:1–10; *Technometrics*, 1: 21–29; *La Scuola in Azione*, 20: 5–19
	274	Cigarettes, cancer and statistics. *Centennial Rev.*, 2: 151–166
	275	Lung cancer and cigarettes? *Nature*, 182: 108
	276	Cancer and smoking. *Nature*, 182: 596
1960	281	(With E.A. Cornish). The percentile points of distributions having known cumulants. *Technometrics*, 2: 209–225
	282	Scientific thought and the refinement of human reasoning. *J. Oper. Res. Soc. Japan*, 3: 1–10
	283	On some extensions of Bayesian inference proposed by Mr. Lindley. *J. Roy. Statist. Soc.*, B, 22: 99–301

(continued)

Year	Paper number	Title
1961	284	Sampling the reference set. *Sankhyã*, 23: 3–8
	285	The weighted mean of two normal samples with unknown variance ratio. *Sankhyã*, 23: 103–114
1962	288	The simultaneous distribution of correlation coefficients. *Sankhyã*, 24: 1–8
	289	Some examples of Bayes' method of the experimental determination of probabilities *a priori*. *J. Roy. Statist. Soc.*, B, 24. 118–124
	290	The place of the design of experiments in the logic of scientific inference. *Colloq. Int. Cent. Nat. Recherche Scientifique,* Paris, No. 110: 13–19; *La Scuola in Azione*, 9: 33–42 (in Italian)
	291	Confidence limits for a cross-product ratio. *Aust. J. Statist.*, 4: 41

References

Aspin, A. A. (1948). An examination and further development of a formula arising in the problem of comparing two mean values. *Biometrika* **35**, 88–96.

Barnard, G. A. (1945). A new test for 2 x 2 tables. *Nature* **156**, 177.

Barnard, G. A. (1947). Significance tests for 2 x 2 tables. *Biometrika* **34**, 123–138.

Barnard, G. (1949). Statistical Inference (with discussion). *Journal of the Royal Statistical Society (B)* **11**, 115–149.

Bartlett, M. S. (1936). The information available in small samples. *Proceedings of the Cambridge Philosophical Society* **32**, 560–566.

Behrens, W. (1929). Ein Beitrag Zur Fehler-Berechnung bei wenigen Beobachtungen. *Landwirtschaftliche Jahrbücher* **68**, 807–837.

Bennett, J. H. (Ed.) (1990). *Statistical Inference and Analysis: Selected Correspondence of R. A. Fisher.* Clarendon Press, Oxford.

Bernoulli, J. (1713). *Ars Conjectandi. Basel: Thurnisiorum.* (Translated into English as *The Art of Conjecturing* by Edith Dudley Sylla, Baltimore: The Johns Hopkins University Press.)

Box, J. F. (1978). *R. A. Fisher: The Life of a Scientist.* Wiley, New York.

Chernoff, H. (1949). Asymptotic studentization in testing of hypotheses. *Ann. Math. Statist.* **20**, 268–278.

Darwin, C. (1876). *The Effects of Cross- and Self-Fertilisation in the Vegetable Kingdom.* John Murray, London.

DeGroot, M. H. (1988). A conversation with George A. Barnard. *Statistical Science* **3**, 196–212.

Dénes, J. and Keedwell, A. D. (1974). *Latin Squares and their Applications.* English Universities Press, London.

Ehrenfeld, S. (1953). On the efficiency of experimental designs. *Annals of Mathematical Statistics* **26**, 247–255.

Fienberg, S. E. and Hinkley, D. V. (Eds.) (1980). *R. A. Fisher: An Appreciation.* Springer, New York.

Fienberg, S. E. and Tanur, J. M. (1996). Reconsidering the fundamental contributions of Fisher and Neyman on experimentation and sampling. *International Statistical Review* **64**, 237–253.

Fisher, R. A. *Collected Papers of R. A. Fisher 1890–1962.* http://digital.library.adelaide.edu.au/coll/special/fisher. For all other references to Fisher, see the Preface and Appendix.

Forcina, A. (1982). Gini's contributions to the theory of inference. *International Statistical Review* **50**, 65–70.

Gosset, W. S. Most references to Gosset are listed under "Student" in this bibliography.

Gosset, W. S. (1970). *Letters from W. S. Gosset to R. A. Fisher, 1915–1936 (with summaries by R. A. Fisher and a Foreword by L. McMullen).* Printed for private circulation by Arthur Guiness, Son and Co. (Park Royal, Ltd.)

Grattan-Guinness, I. (Ed.) (2005). *Landmark Writings in Western Mathematics, 1650–1940.* Elsevier, Oxford.

Hald, A. (1998). *A History of Mathematical Statistics from 1750–1930.* Wiley, New York.

Hampel, F. R.; Ronchetti, E. M.; Rousseeuw, P. J.; and Stahel, W. A. (1986). *Robust Statistics: The Approach Based on Influence Functions.* Wiley, New York.

Heyde, C. C., Seneta, E., Crepel, P., Fienberg, S. E., and Gani, J. (Eds.) (2001). *Statisticians of the Centuries.* Springer, New York.

Hoeffding, W. (1952). The large-sample power of tests based on permutations of observations. *Ann. Math. Statist.* **23**, 169–192.

Hotelling, H. (1927). Review [untitled]. *Journal of the American Statistical Association* **22**, 411–412.

Hotelling, H. (1931). The generalization of Student's ratio. *Annals of Mathematical Statistics* **2**, 360–378.

Hotelling, H. and Pabst, M. R. (1936). Rank correlation and tests of significance involving no assumption of normality. *Annals of Mathematical Statistics* **7**, 29–43.

Huber, P. J. (1964). Robust estimation of a location parameter. *Annals of Mathematical Statistics* **35**, 73–101.

Huber, P. J. (1981). *Robust Statistics.* Wiley, New York.

Huber, P. J. and Ronchetti, E. M. (2009). *Robust Statistics (2nd ed.).* Wiley, New York.

Irwin, J. O. (1929). Review [untitled]. *Journal of the Royal Statistical Society* **92**, 101–103.

Irwin, J. O. (1935). Tests of significance between percentages based on small numbers. *Metron* **12**, 83–94.

Isserlis, L. (1926). Review [untitled]. *Journal of the Royal Statistical Society* **89**, 144–145.

Jeffreys, H. (1939). *Theory of Probability.* Clarendon Press, Oxford.

Johnson, N. L.; Kotz, S.; and Balakrishnan, N. (1995). *Continuous Univariate Distributions, Vol. 2 (2nd ed.).* Wiley, New York.

Jureckova, J. and Sen, P. K. (1996). *Robust Statistical Procedures.* Wiley, New York.

Kiefer, J. C. (1985). *Collected Works, Vol. 3: Design of Experiments.* Springer, New York.

Lehmann, E. L. (2009). Parametrics vs. nonparametrics: Two alternative methodologies. *Journal of Nonparametric Statistics* **21**, 397–405.

Linnik, Y. V. (1968). Statistical problems with nuisance parameters. *American Mathematical Society Transl. Math. Monographs, Vol. 20.*

McCullagh, P. and Nelder, J. A. (1983). *Generalized Linear Models.* Chapman and Hall, London.

McCullagh, P. and Nelder, J. A. (1989). *Generalized Linear Models (2nd ed.).* Chapman and Hall, London.

Neyman, J. (1934). On the two different aspects of the representative method: the method of stratified sampling and the method of purposive selection. *Journal of the Royal Statistical Society* **97**, 558–625.

Neyman, J. (1935a). Sur la vérification des hypotheses statistiques composées. *Bull. Soc. Math. de France* **63**, 246–266.

Neyman, J. (with K. Iwaszkiewicz and St. Kolodziejczyk) (1935b). Statistical problems in agricultural experimentation. *Supplement to the Journal of the Royal Statistical Society* **2**, 107–180.

Neyman, J. (1937a). Outline of a theory of statistical estimation based on the classical theory of probability. *Philosophical Transactions of the Royal Society of London A*, **236**, 333–380.

Neyman, J. (1937b). 'Smooth' test for goodness of fit. *Skandinavisk Aktuarietidskrift* **20**, 149–199.

Neyman, J. (1938a). L'estimation statistique traitée comme un problème classique de probabilité. *Actualités Scientifiques et Industrielles* **No. 739**, 25–57.

Neyman, J. (1938b). *Lectures and Conferences on Mathematical Statistics.* U. S. Department of Agriculture, Washington.

Neyman, J. (1947). Raisonnement inductif ou comportement inductif? Les conceptions modernes de la statistique mathématique. *Proceedings of the International Statistical Conferences* **3**, 423–433.

Neyman, J. (1952). *Lectures and Conferences on Mathematical Statistics and Probability* (2nd ed. of 1938b). U. S. Department of Agriculture, Washington.

Neyman, J. (1955). Problem of inductive inference. *Communications on Pure and Applied Mathematics* **8**, 13–45.

Neyman, J. (1957). 'Inductive behavior' as a basic concept of philosophy of science. *International Statistical Review* **25**, 7–22.

Neyman, J. (1961). Silver Jubilee of my dispute with Fisher. *Journal of the Operations Research Society of Japan* **3**, 145–154.

Neyman, J. (1977). Frequentist probability and frequentist statistics. *Synthese* **36**, 97–131.

Neyman, J. and Pearson, E. S. (1928). On the use and interpretation of certain test criteria. *Biometrika* **20A**, 175–240, 263–295.

Neyman, J. and Pearson, E. S. (1933a). On the problem of the most efficient tests of statistical hypotheses. *Philosophical Transactions of the Royal Society (A)* **231**, 289–337.

Neyman, J. and Pearson, E. S. (1933b). On the testing of statistical hypothesis in relation to probabilities a priori. *Proceedings of the Cambridge Philosophical Society* **24**, 492–510.

Neyman, J. and Pearson, E. S. (1936). Contributions to the theory of testing statistical hypotheses, Part I. *Statistical Research Memoirs* **1**, 1–37.

Neyman, J. and Pearson, E. S. (1938). Contributions to the theory of testing statistical hypotheses, Parts II, III. *Statistical Research Memoirs* **2**, 25–57.

Pearson, E. S. (1929). Statistics and biological research (review). *Nature* **123**, 866–867; also **124**, 615.

Pearson, E. S. (1939). William Sealy Gosset: "Student" as a statistician. *Biometrika* **30**, 210–250.

Pearson, E. S. (1955). Statistical concepts in their relation to reality. *Journal of the Royal Statistical Society (B)* **17**, 204–207.

Pearson, E. S. (1966). The Neyman-Pearson story. In David, F. N. (Ed.) *Research Papers in Statistics (Festschrift for J. Neyman).* Wiley, London.

Pearson, E. S. (1990). *"Student": A Statistical Biography of William Sealy Gosset.* Clarendon Press, Oxford.

Pearson, E. S. and Hartley, H. O. (1954). *Biometrika Tables for Statisticians.* Cambridge University Press, Cambridge.

Pearson, K. (1892). *The Grammar of Science.* Charles Scribner's Sons, London.

Pearson, K. (1894). Contributions to the mathematical theory of evolution. *Philosophical Transactions of the Royal Society of London A*, **185**, 71–110.

Pearson, K. (1895). Contributions to the mathematical theory of evolution II: Skew variation in homogeneous material. *Philosophical Transactions of the Royal Society of London (A)* **186**, 343–414.

Pearson, K. (1900). On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. *Philosophical Magazine,* **Series V**, 157–175.

Pearson, K. (1902). On the systematic fitting of curves to observations and measurements. *Biometrika* **1**, 265–303, and **2**, 1–23.

Pitman, E. J. G. (1947). *Lecture Notes on Nonparametric Statistics* (Unpublished). University of North Carolina, Chapel Hill.

Plackett, R. L. (1977). The marginal totals of a 2 x 2 table. *Biometrika* **64**, 37–42.

Ramsey, F. P. (1926). Truth and probability. In Braithwaite, R. B. (ed.), *The Foundations of Mathematics and Other Logical Essays*, Ch. 7, 156–198. Harcourt, Brace.

Reid, C. (1982). *Neyman – from Life.* Springer, New York.

Robinson, G. (1976). Properties of Student's t and of the Behrens-Fisher solution to the two means problem. *Annals of Statistics* **4**, 963–971.

Savage, L. J. (1954). *The Foundations of Statistics.* Wiley, New York.

Savage, L. J. (1976). On rereading R. A. Fisher (with discussion). *Annals of Statistics* **4**, 441–500.

Savage, L. J. et al. (1962). On the foundations of statistical inference: Discussion. *Journal of the American Statistical Association* **57**, 307–326.

Scheffé, H. (1942). On the ratio of the variances of two normal populations. *Annals of Mathematical Statistics* **13**, 371–388.

Scheffé, H. (1959). *The Analysis of Variance*. Wiley, New York.

Snedecor, G. W. (1937). *Statistical Methods.* The Collegiate Press, Ames.

Speed, T. (1987). What is an analysis of variance? (with discussion.) *Annals of Statistics* **15**, 885–941.

Staudte, R. G. and Sheather, S. J. (1990). *Robust Estimation and Testing.* Wiley, New York.

Stigler, S. (1986). *The History of Statistics.* Harvard University Press, Cambridge.

Stigler, S. (1999). *Statistics on the Table*. Harvard University Press, Cambridge.

Stigler, S. (2005). Fisher in 1921. *Statistical Science* **20**, 32–49.

Stigler, S. M. (2007). The epic story of maximum likelihood. *Statistical Science* **22**, 598–620.

Student (1908a). The probable error of a mean. *Biometrika* **6**, 1–25.

Student (1908b). Probable error of a correlation coefficient. *Biometrika* **6**, 302–310.

Student (1929). Statistics in biological research [letter]. *Nature* **124**, 93.

Student (1936). Co-operation in large-scale experiments. *Supplement to the Journal of the Royal Statistical Society* **3**, No. 2, 115–136.

Student (1938). Comparison between balanced and random arrangements of field plots. *Biometrika* **29**, 363–379.

Tocher, K. D. (1950). Extension of the Neyman-Pearson theory of tests to discontinuous variates. *Biometrika* **37**, 130–144.

Tolley, H. R. (1929). Economic data from the sampling point of view. *Journal of the American Statistical Association* **24**, No. 165*, Supplement: Proceedings of the American Statistical Association* (Mar., 1929), 69–72.

Upton, G. (1982). A comparison of alternative tests for the 2 x 2 comparative trial. *Journal of the Royal Statistical Society (A)* **145**, 86–105.

Wald, A. (1943). On the efficient design of statistical investigations. *Annals of Mathematical Statistics* **14**, 134–140.

Wald, A. (1950). *Statistical Decision Functions.* Wiley, New York.

Welch, B. L. (1938). The significance of the difference between two means when the population variances are unequal. *Biometrika* **29**, 350–362.

Welch, B. L. (1947). The generalization of Student's problem when several different population variances are involved. *Biometrika* **34**, 28–35.

Wilcoxon, F. (1945). Individual comparisons by ranking methods. *Biometrics* **1**, 80–83.

Working, H. and Hotelling, H. (1929). Application of the theory of error to the interpretation of trends. *Journal of the American Statistical Association* **24**, no. 165: *Supplement: Proceedings of the American Statistical Association*, 73–85.

Yates, F. (1934). Contingency tables involving small numbers and the χ^2–test. *Supplement to the Journal of the Royal Statistical Society* **1**, 217–235.

Zabell, S. L. (1992). R. A. Fisher and the fiducial argument. *Statistical Science* **7**, 369–387.

Name Index

Subject Index