Brit. J. Phil. Sci. **57** (2006), 69–91

Models and Statistical Inference: The Controversy between Fisher and Neyman–Pearson

Johannes Lenhard

ABSTRACT

The main thesis of the paper is that in the case of modern statistics, the differences between the various concepts of models were the key to its formative controversies. The mathematical theory of statistical inference was mainly developed by Ronald A. Fisher, Jerzy Neyman, and Egon S. Pearson. Fisher on the one side and Neyman–Pearson on the other were involved often in a polemic controversy. The common view is that Neyman and Pearson made Fisher's account more stringent mathematically. It is argued, however, that there is a profound theoretical basis for the controversy: both sides held conflicting views about the role of mathematical modelling. At the end, the influential programme of Exploratory Data Analysis is considered to be advocating another, more instrumental conception of models.

1 Introduction

The mathematical theory of statistical inference was developed during the 1920s and 1930s mainly by three scholars: Ronald A. Fisher (1890–1962), Jerzy Neyman (1894–1981), and Egon S. Pearson (1895–1980). The theory of testing hypotheses is 'one of the most widely used quantitative methodologies, and has found its way into nearly all areas of human endeavour' (Lehmann [1993], p. 1242). The present contribution will discuss the conflict between Fisher on the one side and Neyman and Pearson on the other over a 'controversy between classical theories of testing' (Hacking [1965], p. 89) that already began in the 1930s and was never settled. My main intention will be

doi:10.1093/bjps/axi152

Advance Access published on January 3, 2006

to reinterpret this controversy as a dispute about different conceptions of mathematical–statistical modelling.

The essay 'On the Mathematical Foundations of Theoretical Statistics' ([1922]) published by Fisher in 1922[1] can be considered to be one of the cornerstones of mathematically oriented statistics. It contains a wealth of new concepts such as the level of significance, a concept that is pervasive today, introduced by Fisher in order to establish a logic of statistical inference. The construction of infinite hypothetical populations plays a central role for this logic.[2] In more precise terms: for Fisher, the specification of an infinite population represents the essential step in establishing a (parametric) model. This is indeed where Fisher introduces the concept of the mathematical model to statistical inference.

About a decade later, Neyman and Pearson developed what is known today as the Neyman–Pearson theory of statistical testing. During this time, they cooperated closely, and they can be considered as a unit for our purposes here, even though they ended their cooperation a short time later. The founding date for their theory is most probably 1933, when they published 'On the Problem of the Most Efficient Tests of Statistical Hypotheses', ([1933a]) an essay that they referred to between themselves as 'the big paper'.[3] This is the location of the so-called *fundamental lemma* that makes a mathematical claim for the existence of a certain optimal test method. This lemma forms the backbone of the Neyman–Pearson theory, and was considered by its authors as a justification of Fisher's older approach. The Neyman–Pearson theory is closely linked to a class of models that serves as prerequisite for the mathematical reasoning, and hence it is no wonder that Neyman and Pearson admired Fisher's approach to modelling, considering the concept of model to be an eminent mathematical acquisition.

What caused the bitter controversy to spring up suddenly between Fisher and Neyman–Pearson? Their scientific rivalry over the 'correct' logic of inference was certainly fostered by their close proximity to each other. For some years, they resided in the same building after Karl Pearson retired and his post had been split into two: in 1934, Egon Pearson and Fisher accepted the chairs for statistics and eugenics, respectively. In the literature, you find remarks like 'because of Fisher's remarkable talent for polemic, the debate never lacked for overblown rhetoric' (Gigerenzer et al. [1989], p. 105). An impression of the heated and polemical atmosphere can be obtained from the

[1] His book on statistical methods (Fisher [1925]) that explains this approach further became a real bestseller.

[2] Just such infinite populations, or more exactly their constructive surplus over any observation, had prompted Fisher's conflict with Karl Pearson, the father of Egon Pearson and the director of the London Galton Laboratory (*cf.* Morrison [2002]).

[3] Other important joint works of the two authors are (Neyman and Pearson [1928], [1933b]).

Royal Statistical Society's records of debate. Following the presentation of Neyman's contribution 'Statistical Problems in Agricultural Experimentation' ([1935], p. 193), the record says: 'Professor R. A. Fisher, in opening the discussion, said he 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.' Upon which there followed prompt retribution.

> Dr Pearson said while he knew that there was a widespread belief in Professor Fisher's infallibility, he must, in the first place, beg leave to question the wisdom of accusing a fellow-worker of incompetence without, at the same time, showing that he had succeeded in mastering his argument (*op. cit.*, p. 202).

Where discussion runs like this, serious debate is hardly intended. The polemics most certainly played a role in the course of the controversy, and, moreover, it is widely recognized that the fiducial argument on which Fisher insisted, is erroneous—as incisively pointed out by Neyman—or at least obscure. On the whole, Neyman and Pearson believed their own approach to have smoothened, or filled up some weaknesses and gaps in Fisher's argumentation. According to accepted opinion in the literature, it is quite possible to reconcile the two approaches in terms of the deployed statistical methods.

I should like to argue that the controversy rests, besides all personal aspects, on a profound conceptual basis. There is a common nucleus to both approaches and this is where one can find the reason for the controversy. They differ about a fundamental *methodological* issue: Both sides held conflicting views about the function of mathematical models and about the role of modelling in statistical inference.

In what follows, I should like to look at the systematic standpoints in more detail, arguing finally in favour of the hypothesis that Neyman–Pearson intended to deal with behaviours in the framework of their theory, that is, to integrate a new kind of object, as it appears in their lemma. This can be viewed as an attempt at mathematization: mathematical functions, for instance, first served to describe dynamic relations, then themselves become objects of mathematical theory. In a quite similar way, hypothesis testing now becomes an object of mathematical theory. Neyman–Pearson saw hypothesis testing as a course of action in the frame of a reiterated process, like sampling for quality control. The *fundamental lemma* is about mathematical properties of 'courses of action,' and it establishes the existence of an optimal procedure—on the basis of certain assumptions on what possible procedures look like. Neyman–Pearson intended to establish an objective basis for the logic of inference that no longer depended on constructing hypothetical populations—a method too subjective in their eyes.

With regard to modelling, Fisher held quite a different view, stressing that a model had to mediate between questions of application to real problems and data on the one hand, and questions of mathematical properties and arguments on the other. According to him, Neyman–Pearson had founded their theory on too strong prerequisites, thus too closely restricting the possibilities of application. Even worse, they had hardly left any room for the basic mediating role of models, because their concept of testing rested on a more or less complete description of the situation in mathematical terms. This is a remarkably ironic point: What precisely Neyman–Pearson admired as Fisher's major achievement, his own concept of model, gave rise to their controversy with him!

The disputes over the activity of mathematical modelling, and about how to deal with problems of application, however, concern a permanent problem of applied science. I think one cannot beg the question by simply assenting either to Fisher or to Neyman–Pearson whose views were characterized by *mediation* versus *integration*, respectively. They form indispensable, complementary aspects of applied mathematics. The balance of these aspects is decisive as was shown by an interpretation of John W. Tukey's *Exploratory Data Analysis* (EDA), which was elaborated in the 1960s and 1970s. While it represents an influential current in applied statistics, it is nevertheless frequently ignored in philosophical reflections about statistics. It started as a critical assessment of the use of models in statistics and in the end, it is argued, EDA provides a new systematic role for models. In my view, an adequate reconstruction of the controversies characterizing modern statistics of the 20th century can be attained only by distinguishing between the various conceptions of models and of modelling.

2 Models in statistics—'of what population is this a random sample?'

Until the 1920s, Karl Pearson, director of the London Galton Laboratory, assumed a leading, even dominant position in statistics. He founded the biometric school, represented by the renowned journal *Biometrika*, whose incisive dispute with the so-called Mendelians, namely Bateson, on the correct way of formulating population genetics acquired some fame. As Margaret Morrison has recently pointed out ([2002]), the conflict, in retrospect, appears completely superfluous, because Fisher had found a way of representing a kind of synthesis between the two positions. Fisher, as a declared Mendelian convinced of the fundamental importance of the stochastic laws of hereditary transmission, at the same time used mathematical models and sophisticated mathematical techniques.

Whereas infinite populations have an important role in Fisher's model approaches, they do not, of course, correspond directly to reality. It was only the inferences from the idealizing dynamic within the model, e.g. diffusion limits, that could be compared with observations; and to deviate like this by way of a mathematical model world that acquired independent status[4] led Fisher to strongly oppose Karl Pearson who advocated a more limited use of mathematical methods and models. However fruitful these approaches were for population genetics (which was to a large part initiated by them, cf. Provine [1986]), Pearson categorically rejected something like infinite hypothetical populations, in complete agreement with his positivist 'bible,' 'The Grammar of Science' (Pearson [1892]). Fisher, who was also personally hurt by Pearson's verdict against mathematical modelling, found himself compelled to accept the newly created position of a statistician at the Rothamsted Experimental Station specializing in agricultural science in order to avoid becoming head statistician at the Galton Laboratories under Pearson as director. He initially considered himself to have been banished from the academic environment, but, in retrospect, his daily confrontation with practical problems prompted him to conceive of an effective statistical method. Apart from that, Fisher loved in later years to use his practical experience as an argument against overly academic positions. The following description can also be understood as a continuation of the story told by Morrison: While introducing idealizing mathematical models brought about the breakthrough for population genetics, grafting mathematical modelling onto the method of inference led to conflicts within theoretical statistics.

In Rothamsted, Fisher elaborated his own approach to a comprehensive 'logic of inductive inference', as he called it, the leading role again being assigned to constructing mathematical models. What is more, the cornerstone of Fisher's conception of inference logic, his presumably philosophically fundamental innovation, consists in precisely describing what is to be understood by a model, and how models are to be imbedded in the logic of inference.

If one wished to give a year for the emergence of inference logic and of testing hypotheses one might choose 1922, when Fisher published his seminal contribution 'On the Mathematical Foundations of Theoretical Statistics' ([1922]). Fisher himself later described it as 'the first large-scale attack on the problem of estimation' ([1971], p. 277), and this is where we find a number of influential new concepts, among them the level of significance and the parametric model,[5] whose systematic role within statistical inference was

[4] Winsberg ([2003]) uses the term *semi-autonomous* to specify Morrison's ([1999]) somewhat stronger, but basically correct characterization of 'models as autonomous mediators.'

[5] For the purposes here, it is adequate to use models and parametric models synonymously.

elaborated for the first time. Fisher describes the general goal of statistics as follows:

> In order to arrive at a distinct formulation of statistical problems, it is necessary to define the task which the statistician sets himself: briefly, and in its most concrete form, the object of statistical methods is the reduction of data. A quantity of data, which usually by its mere bulk is incapable of entering the mind, is to be replaced by relatively few quantities which shall adequately represent the whole, or which, in other words, shall contain as much as possible, ideally the whole, of the relevant information contained in the original data ([1922], p. 311).

At first glance, it may seem that Fisher's concern is merely a technical question of the reduction of data. This, however, is not the case, for the problem of whether certain standard quantities 'adequately represent' the entirety of data cannot be solved on the basis of the data themselves. The same holds for 'relevant information'—whether it is still contained in a data-reducing statistic will have to be measured according to further criteria. Fisher continues:

> 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 ([1922], p. 311).

Fisher explicitly mentions the constructive character of this undertaking, which consists in conceiving of the data observed as of an instance of the underlying model-type population. The merit of this is that such a population, i.e. its distribution law, is exhaustively ('in respect of all qualities under discussion,' i.e. with regard to a concrete question of application) described by a small number of parameters. Of course, everything will depend on whether a model population appropriate to the respective application situation has been selected. The mathematical properties of this model population play a decisive role in this, for it is these that permit a reduction of data. It is this ideal, constructed, mathematical world of hypothetical populations that lends itself to be specified by 'relatively few parameters.' The birthday of mathematical statistics and the introduction of parametric models occur at the same time. Fisher subdivided the general task of statistics into three types of problems:

1. *Problems of specification*. 'These arise in the choice of the mathematical form of the population' ([1922], p. 366). This step thus consists in forming a model, and it cannot be derived, but requires deliberations, like those on the basis of practical experience gained with similar situations. Fisher attributes this step to the logic of inference. It could even be considered typical for his *inductive inference*, for, after all, it is the transition from concrete data to mathematical models that makes this step 'inductive.' He was also aware of

the fact that choosing an adequate model depended on the mathematical structures available and treatable. Progress in these matters would instantly change the reasoning about modelling: '[. . .] problems of Specification are found to be dominated by considerations which may change rapidly during the progress of Statistical Science ([1922], p. 366).[6]

2. *Problems of estimation* whose formulation already requires framing by a mathematical-statistical model. For this second type of problem, he saw a general solution in his 'logic of inductive inference' of 1922: 'The principal purpose of this paper is to put forward a general solution of problems of Estimation' ([1922], p. 366). This is where mathematical theory formation in the strict sense takes place, for instance the introduction of the concept of the sufficiency of a statistic and the derivation of mathematical propositions connected with it.

3. *Problems of distribution*. The matter here is tractability—that abstract reasoning is profitable in estimation only if it eventually leads to concrete numerical results. The most beautiful model is good for nothing if it yields no distribution curves. Fisher continued:

> As regards problems of specification, these are entirely a matter for the practical statistician, for those cases where the qualitative nature of the hypothetical population is known do not involve any problems of this type. In other cases we may know by experience what forms are likely to be suitable, and the adequacy of our choice may be tested a posteriori. We must confine ourselves to those forms which we know how to handle, or for which any tables which may be necessary have been constructed ([1922], p. 314).[7]

At this point, it is appropriate to remark on the terminology and then to provide an example of its application. For Fisher, a *model* is an entire class of hypotheses, and he terms the process of selecting one hypothesis from this class (among those possible according to the model) *specification*. The crucial aspect in this is that both modelling, i.e. framing the subsequent mathematical analysis, and specification belong to the process of inference.

> A very important feature of inductive inference, unknown in the field of deductive inference, is the framing of the hypothesis in terms of which the

[6] Fisher saw clearly that the constructive step of building a model can be justified by a comparison of the implications of the model with the empirical observations: 'the adequacy of our choice may be tested a posteriori. . . . 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' ([1922], p. 314).

[7] The availability of numerical methods has radically changed the situation: even in quite complicated models the distribution functions can be determined quasi-empirically by simulation. This has made a greatly extended class of models applicable whose analysis is in no way dependent on pre-existing tables and quantiles.

> data are to be interpreted. The hypothesis is sometimes called a model, but I should suggest that the word model should only be used for aspects of the hypothesis between which the data cannot discriminate (Fisher [1955], p. 75).

A model may assume a certain family of distributions whose parameters have to be specified by estimation from the data. Hence the data cannot discriminate between the assumed family of distributions and another one. A simple, admittedly very simplified, example may explain the terminology: During his work in Rothamsted, Fisher was intensely engaged in agro-science experiments such as estimating the effect of a certain fertilizer. A model could look as follows:

Consider n lots of land $1, \ldots, n$. Let E_i be the yield of the lot i, and let E_i be normally distributed for all i with mean m and variance σ^2. This is to say that the yield of the various acreages is equally distributed, that is, normally distributed to the two parameters (m, σ^2). This establishes essential assumptions of the *model*. The effect of the fertilizer, it is further assumed, will only change the parameter m. In other words, the yield of a fertilized acreage is normally distributed to a mean m'. A typical question regarding the statistical inference to be drawn from the data $(E_1, \ldots, E_n)$ would then be: which effect is produced by treating with the fertilizer? The null hypothesis H_0 would be that the fertilizer has no effect at all; that is, that the means are equal, and all differences observed are random:

$$H_0 : m = m'.$$

What is at issue here is not to derive an answer from some data, the method being explained in every textbook of statistics, but only of the use of terms: A hypothesis is *specified* by m and σ^2 being given; all information contained in the data not concerning these parameters is irrelevant (under the model's assumptions). Given the model, the specification is achieved by assigning the values of the two parameters: It is a mathematical fact that the normal distribution is characterized by mean and variance. In Fisher's terms, the normal distribution is part of the model while assigning concrete values to the parameters specifies a hypothesis.

Thus, in the case of the normal distribution, the probability of an observation falling in the range dx, is

$$\frac{1}{\sigma\sqrt{2\pi}} \mathrm{e}^{-(x-m)^2/2\sigma^2} \mathrm{d}x$$

in which expression x is the value of the variate, while m, the mean, and σ, the standard deviation, are the two parameters by which the hypothetical population is specified. If a sample of n be taken from such a population, the data comprise n independent facts. The statistical process of the reduction of these data is designed to extract from them all relevant information respecting the

values of m and σ, and to reject all other information as irrelevant (Fisher [1922], p. 312).

Hence the assumption of a model made it possible to speak of 'relevant information' contained in the data and, furthermore, offered a way to discard a hypothesis in light of the data on the basis of the famous 'level of significance' that guaranteed Fisher's approach seminal influence. This 1922 paper also saw the first use of what are now quite familiar terms of a statistic's efficiency and sufficiency that also require argumentation in the frame of a model. Fisher defines:

> Sufficiency.—A statistic satisfies the criterion of sufficiency when no other statistic which can be calculated from the same sample provides any additional information as to the parameter to be estimated ([1922], p. 310).

It needs to be noted that the judgement whether a statistic encompasses all the relevant information must be based on the assumption of a model. There is an entire cosmos of concepts formed by Fisher, but not all are of interest in our context here.[8] I should again like to emphasize two aspects of Fisher's approach:

1. Modelling is an essential part of inference logic. A model forms the frame for formulating testable hypotheses, and some kind of platform for further mathematical argumentation.
2. The constructive and mathematically idealizing character of modelling is emphasized. Assuming a model will always make one of several possibilities 'real' and exclude others, the applied mathematician's (or statistician's, or researcher's) power of judgement taking an important role in this.

> In short: the framing by means of a model is located at the beginning of the statistical treatment of a problem of application: 'The postulate of randomness thus resolves itself into the question, "Of what population is this a random sample?" which must frequently be asked by every practical statistician' (Fisher [1922], p. 312–3).

3 The fundamental lemma

During the following decade, Jerzy Neyman and Egon Pearson elaborated the theory of statistical inference that bears their names. Egon Pearson was the son of Karl Pearson, and he worked in London at the Galton Laboratory. For several years, he cooperated closely with Jerzy Neyman, a Polish

[8] I should like to stress that his curious concept of fiducial probability does not contribute anything to the theory of models to be treated here—although it did give rise to controversies, see Hacking ([1965]). The experimental design (*cf.* Fisher [1935]) will not be treated either—while the production of data already has to do with models, this aspect plays no role in the controversy between Fisher and Neyman–Pearson.

mathematician and statistician. This is why Egon's personal reminiscences on the occasion of a festschrift for Neyman is titled 'The Neyman–Pearson Story' ([1966]). Jerzy Neyman had come to London as a postdoc to stay with Karl Pearson in 1925, and it soon turned out that Neyman fitted into the theoretical–mathematical line of development represented by Fisher, which Karl Pearson disliked and was dismissive of. Revealingly, Gosset, the inventor of the *t*-test who published under the pseudonym of *Student*, prepared Fisher in Rothamsted for Neyman's arrival by writing: 'He is fonder of algebra than correlation tables and is the only person except yourself I have heard talk about maximum likelyhood as if he enjoyed it' (Reid [1982], p. 59).

The seminal essay 'On the Problem of the Most Efficient Tests of Statistical Hypotheses' ([1933]) can be considered to be the most important contribution, and perhaps the founding document—an essay referred to by the authors as 'the big paper.'[9] The theoretical backbone of the Neyman–Pearson theory is formed by the so-called *fundamental lemma* proved in this essay. It is an expression of the two authors' systematic approach.[10]

What is this lemma about? The two authors had begun by finding deficits in the logic of testing that mainly concerned two points. First, Neyman–Pearson criticized the asymmetrical treatment of the null-hypothesis as a deficit of Fisher's logic of testing. Fisher started with the null hypothesis that no effect could be observed, and a test might lead to accepting another hypothesis, thereby rejecting the null hypothesis. The name alone already testified the asymmetrical conception. Neyman–Pearson insisted at an early stage, that is, prior to their 'big paper,' that the situation must in a certain sense be conceived of as symmetrical. This was to say that a model should consist of two competing hypotheses ('hypothesis' versus 'alternative'), and observing the data should lead to the decision on which hypothesis was to be preferred.

To confront two hypotheses certainly represents a conceptual progress, inasmuch as accepting or rejecting a hypothesis will always tell us something about the alternative. On the mathematical level, this view is reflected in the emphasis on the so-called likelihood quotient. This quotient relates the plausibility values of the data observed under the various hypotheses to one another. Neyman–Pearson indeed focussed on this quantity, and not on the

[9] Actually, the intense cooperation between Neyman and Egon Pearson ended soon after that, and the latter distanced himself from some positions taken by Neyman in later years. For the purposes of the present argument, however, it is permissible to treat the two as a unity, that of Neyman–Pearson.

[10] Like Fisher, Neyman and Pearson initiated innovations in diverse directions. In particular, they coined the concept of the confidence interval, which is indubitably of fundamental importance in modern statistics. Ian Hacking, for instance, has gone as far as to speak of an independent 'confidence-interval approach' ([1980]) that he contrasts with Fisher's 'logicist' approach. For the controversy on which we are focusing here, it is sufficient, however, to examine the fundamental lemma and the approach to the logic of testing it implies.

likelihood function itself, which assigns a value to data observed under the assumption of a hypothesis.

Above all, however, Neyman–Pearson introduced the 'error of the second kind', thus finding a name for the problem. Choosing one of two competing hypotheses, of course, can be wrong every time, just as choosing the other one can be. One can thus commit errors of the first kind (incorrectly accepting a false hypothesis), and errors of the second kind (wrongly rejecting a true hypothesis), and one should therefore make the relative assessment of the two an object of the method as well. In a conference at the Royal Statistical Society, Neyman described this as follows:

> I have met several cases while considering questions of practical experimentation, in which the level of significance $\alpha = 0.01$ proved definitely too stringent. It is the business of the experimenter to choose a proper level in any particular case, remembering that the fewer the errors of one kind, the more there are of the other ([1935], p. 180).[11]

The second focus of the Neyman–Pearson theory was to justify using the likelihood principle, that is, the prescription to choose the value of the likelihood function, or the likelihood quotient derived from it, as the quantity to be maximized. It was one thing to establish a symmetry of situation with regard to two hypotheses, but quite another thing to select a standard quantity for the purpose of weighing the two hypotheses, a second step that had to be conceptually separated from the first. Neyman–Pearson were pursuing some kind of justificatory logic, that would have to consist in a *mathematical* argument in favour of the likelihood principle whose validity, incidentally, they did not doubt. 'If we show that the frequency of accepting a false hypothesis is minimum when we use (likelihood) tests, I think it will be quite a thing' (Neyman in a letter to Pearson, cited according to Reid [1982], p. 92).

The fundamental lemma yielded success for both goals: proving the lemma represented a mathematical achievement, but what was of at least equal importance was to formulate the problems adequately so that the lemma's assertion offers a solution. The following is a succinct description by Hacking that does without mathematical formalism:

> According to this theory, there should be very little chance of mistakenly rejecting a true hypothesis. Thus, if R is the rejection class, the chance of

[11] An interesting detail is that Fisher was among those listening to Neyman in the Royal Society's auditorium, and he must have felt a bit challenged. For Neyman examined methods introduced by Fisher like that of 'Latin squares' as to their consistent use regarding the weighing of hypotheses, a problem which Fisher had failed to see. Thus, Neyman–Pearson repeatedly pointed out that the tests of significance suggested by Fisher could be 'worse than useless,' as one could reject a badly supported alternative while insisting on the original hypothesis that may be supported even less. Some short quotes from the subsequent dispute have been given in the Introduction above.

> observing a result in R, if the hypothesis under test is true, should be as small as possible. This chance is called the *size* of the test; the size used to be called the significance level of the test.
>
> In addition to small size, says the theory, there should be a good chance of rejecting false hypotheses. Suppose simple h is being tested against simple i. Then, for given size, the test should be so designed that the chance of rejecting h, if i is true, should be as great as possible. This chance is called the *power* of h against i (Hacking [1965], p. 92).

From their analysis of two types of statistical error, Neyman–Pearson had derived the concepts of *size* and of *power.* In this, size corresponds to the level of significance, whereas power corresponds to the analogous quantity for the error of the second kind. The mode in which the problem is posed here is quite crucial: When two hypotheses confront one another, the first thing to do is to fix the *size* of a test, and the second is to optimize its *power*. These reflections are transformed into a theory by a mathematical proposition: The *Fundamental Lemma of Neyman and Pearson*: In the case of a simple dichotomy of hypotheses, there exists, for any possible size, a uniquely most powerful test of that size.[12]

The proof of the lemma had been seen, as is proper for mathematical theorems, by Neyman in a sudden flash of intelligence in 1930, as he vividly recounts in his talks with Reid (Reid [1982]). Neyman–Pearson chose to embed a test within a certain logic of method that does not consider the individual case, but rather what happens if one proceeds in accordance with such a rule (as described by Hacking above). Framed in this way, the possible courses of action can be treated as mathematical objects that form risk sets with topological properties, namely they are convex. The problem whether a *most powerful test* exists could then be solved by variational calculus. There is a unique element with minimal distance (maximal power) to the point specified by size. The crucial step is the *outline of the problem* (iterated procedure, two alternatives, first determine size, then maximize power) that yielded, roughly speaking, convexity. Another conceptualization of the problem would not have allowed that abstract-geometrical argument.

Formulating the problem in this way led to a fundamental shifting of emphasis, for Neyman and Pearson started by searching for a mathematical justification for the likelihood principle. They were successful, because (in the simplest case) the likelihood principle coincides with the most powerful test. Later, however, the *most powerful test* became the theoretically more fundamental concept, i.e. the likelihood principle was subsumed as a special case.

[12] The lemma speaks of a simple hypothesis and a simple alternative, that is, a simple dichotomy. Treating more complicated cases proved to be a research program over several decades.

Neyman and Pearson considered the double task of establishing symmetry between competing hypotheses and of finding a justification for the likelihood principle to be basically achieved. Introducing the conception of a *most powerful test*, in their eyes, yielded the justification sought. The determination of possible rational methods (two hypotheses, first determine *size*, then optimize *power*) had been the clue to treat the possible courses of testing as mathematical objects. This approach implies that a model has to fit into the methodological framework that is conceived of as more fundamental, or prior, to modelling. The feasibility of Neyman–Pearson's solution obviously depends on how the problem is formulated. This approach led to a violent controversy with Fisher that was never settled.

4 Controversy about models

Whereas Neyman–Pearson, notwithstanding the bitterness of the debate, saw their own contribution as an important supplement to Fisher's position, making it mathematically consistent, Fisher took a contradictory view, rejecting this 'improvement' in a roundabout way—and doing this frequently in a polemic style that seems to have been his speciality. Fisher judges the Neyman–Pearson theory as follows:

> [The unfounded presumptions] would scarcely have been possible without that insulation from all living contact with the natural sciences, which is a disconcerting feature of many mathematical departments ([1955], p. 70).

Fisher states that his own approaches had been reinterpreted in a way that could not claim to have developed it further:

> There is no difference to matter in the field of mathematical analysis, though different numerical results are arrived at, but there is a clear difference in logical point of view, ... this difference in point of view originated when Neyman, thinking that he was correcting and improving my own early work on tests of significance, as a means to the "improvement of natural knowledge", in fact reinterpreted them in terms of that technological and commercial apparatus which is known as an acceptance procedure ([1955], p. 69).

That a decisive change had occurred here would certainly be admitted by Neyman and Pearson. Whereas Fisher considers his own logic of inference distorted, Neyman–Pearson mainly see a mathematical rounding off and improvement of Fisher's approaches. In the literature, this controversy has repeatedly been treated both under mathematical and under philosophical aspects (*cf.* Braithwaite [1953], Hacking [1965], Kyburg [1974], Seidenfeld [1979], Gigerenzer et al. [1989], Lehmann [1993]). According to the common interpretation, the reason for the bitter controversy lasting several decades lies chiefly in the polemics, mainly from Fisher's side, that heated up the

atmosphere. In retrospect, however, the viewpoints do not look so incompatible at all, and there have been attempts at 'reconciliation' (e.g. Lehman [1993]). Some state that the reconciliation would consist in fusing the best parts of the two approaches, while others hold that Neyman and Pearson were actually right in their surmise of having mainly filled in the gaps in Fisher's conception. The latter view is indeed so widespread that Hacking was led to write: 'The mature theory of Neyman and Pearson is very nearly the received theory on testing statistical hypotheses' ([1965], p. 92).

The present contribution, however, follows another direction, for the matter is not to systematically evaluate the justification of certain statistical procedures, but rather to analyse the controversy between Fisher and Neyman–Pearson, elaborating its underlying fundamental philosophical differences about the role and function of mathematical models. I shall like to argue that from the models based approach it becomes clear, or at least reasonable, why the controversy was not resolved—the function of models and the process of building models were conceived of in a different and incompatible manner by both sides.

It is revealing that Neyman–Pearson criticized Fisher for having introduced *infinite populations* as constructed entities. Although they certainly did not repeat the positivist critique in Karl Pearson's vein, they considered the construction and selection from several possible hypothetical populations a far too subjective method. Consequently, they designed a competing approach in order to escape the manifold interpretability of possible hypothetical populations. By the mathematization described above, they intended to attain a more complete and more formal description of the application process.[13]

The key to analysing the controversy is the concept of model, emphatically assigned a role by both parties. Neyman circumscribes the concept as follows:

> A model is a set of invented assumptions regarding invented entities such that, if one treats these invented entities as representations of appropriate elements of the phenomena studied, the consequences of the hypotheses constituting the model are expected to agree with observations. ([1957], p. 8)

[13] Quite in line with these intentions, Neyman introduced his famous concept of *inductive behaviour* as a counter-design to *inductive inference* in order to emphasize the reference to a *course of action* instead of to a process of construction (*cf*. Neyman [1957]). Subsequent to Neyman's 1938 move to the United States and his entrenchment at Berkeley, where he created an institute for statistics that formed a school of its own, the so-called *decision theory* was elaborated, a theory also called Neyman–Pearson–Wald theory, for instance, *cf*. Wald's volume 'Statistical Decision Functions' ([1950]). Connected to this is the emergence of *operations research*, a field combining statistical decision theory and economics. The mathematical treatment of decisions also drew many hints from game theory that combines strategic decisions and maximizing benefits. Gahmari-Tabrizi ([2000]) speaks in another connection of the 'quintessential behaviouralist thesis of the 1950' in the United States, a pronouncement that might well be extended to mathematical statistics.

This is where essential components of how Fisher conceived of the relation between mathematics and its application in the real world resurface. Whereas mathematical entities and models must be distinguished from the world of the 'phenomena studied,' they are necessary for mathematical arguments to be applied. In this respect, statistical models are similar to mathematical models in physics whose methodological importance for modern natural science has often been emphasized, *cf*. for recent examples of that time-honoured debate of Cartwright ([1999]), Giere ([1999]), or Morrison ([1999]). Models hence remain central concepts of the mathematical approaches developed by both parties of the controversy.[14] There is, however, a profound difference in the views about both the function of models and the scope and meaning of model construction that marks the controversy between Fisher and the Neyman–Pearson schools.

That Fisher and Neyman–Pearson started from quite different basic coordinates in their controversy is nicely illustrated by how the two parties related to W. S. Gosset. Both referred to his undisputed reputation as an experienced statistician for confirmation of their views (see also the description in Gigerenzer et al. [1989], p. 105). In his eulogy for Gosset in the *Annals of Eugenics* ([1939]), Fisher stresses that Student (Gosset's pen name), too, had always assumed a plurality of possible models that had to be adequately specified for the purposes of an applied problem. Fisher uses this indication to confirm his own view that mathematical models do not 'fit' real applications without further ado, and that mediating between the two was the applied statistician's (or mathematician's) task—and, above all, his view that this problem will arise ever anew whenever the questions or data vary.

Fisher's opponents quoted Gosset's view with regard to the symmetry between hypothesis and alternative. As late as 1968, Pearson still published letters from Gosset in *Biometrica* (Pearson [1968]) in which the latter emphatically argued with him that using the likelihood quotient, which stands for the comparison of two hypotheses, was much more convincing than using the likelihood function. This statement represented an implicit criticism of Fisher; and, moreover, the effort to find a mathematical justification for using the likelihood quotient had been the initial ignition for the Neyman–Pearson theory, as we have seen. This is where Gosset bears witness to the incompleteness of Fisher's conception.

Both parties to the controversy saw modelling as a fundamental step in applying mathematical statistics. For both grasping the mathematical–statistical model conceptually was the decisive step forward that made a logic

[14] There are systematic reasons for that as well: the frequentistic interpretation of probability, which was more or less advocated by all of the participants in the conflict, refers to a mathematical limit of frequencies, that is, an idealized model. Fisher had shown how effective statistical inference was possible without using Bayesian knowledge about prior distributions. Realizing this, Neyman–Pearson thought Bayesianism obsolete.

of inference possible. The 'problems of estimation' (Fisher's name for the second type of inference problems) were seen quite similarly by everybody concerned, and both approaches relied on solving these problems on the basis of mathematical argumentation, which was only possible on the basis of models. How the two parties conceived of a model as a platform for further argumentation (*size*, *power*, *sufficiency*, etc.) thus seems to be quite comparable at first.

On the other hand, they held very divergent views on the function that should be assigned to models, or better, what is the task of building a model. Neyman and Pearson stand for the programme of mathematizing behaviour (courses of action). The paradigm they had in mind was the methods of quality control in industrial production in which the rationality of action is fixed to a large extent. This does not mean that an optimal behaviour is evident, but rather that the criteria of optimality are given, or can at least be formalized in principle. This, in turn, permits making this behaviour an object of mathematics, just as Neyman–Pearson did with their fundamental lemma.

In the frame of the Neyman–Pearson theory, the reiterated application of a procedure forms the basis for the statistical inferences (just think of an *acceptance procedure*, a method of taking random samples of shipments and of accepting or refusing the latter in accordance with the results). This particular conceptualization was the only way that Neyman–Pearson could provide an objective basis for the logic of inference, thereby excluding the merely hypothetical infinite populations as superfluous elements. It can be said that Neyman–Pearson rely on a concept of model that includes much more preconditions, according to which much of the statistician's method is already fixed. In short: the achievement of mathematization represented by developing the fundamental lemma that made *courses of action* into objects of mathematical arguments had caused a central component of inference logic to switch sides, whereas a statistician, according to Fisher, uses mathematical reasoning within the logic of inference, e.g. building and adjusting a model to the data at hand and to the questions under discussion. In Neyman–Pearson's theory, the reasoning of the statistician himself (e.g. finding an appropriate acceptance procedure) has become an object of the mathematical argument. Hence I should like to call it the *integration* view.

With this, however, they place themselves in strict opposition to Fisher. For him, modelling creates the objects one can argue about mathematically, whereas Neyman–Pearson shape the basic situation in which modelling takes place, requiring reiterated procedures and competing hypotheses. Fisher considered the applied mathematician's situation fraught in principle with many subjective components—working on an applied problem requires a high degree of 'judgement.' According to Fisher, reflecting this application situation and its non-mathematical components was an integral part of applied

mathematics or statistics. Modelling thus has the task of *mediating* between real-world problems and mathematics. Hence, Neyman–Pearson intended to get rid of precisely that constructive act of modelling that is the focus of Fisher's inference logic!

In Fisher's view, Neyman–Pearson simply erred in eliminating the fundamental step of modelling because they assumed the situation to be objective already: 'By ignoring this necessity a "theory of testing hypotheses" has been produced in which a primary requirement of any competent test has been overlooked' ([1955], p. 71). In Neyman–Pearson's view, in contrast, only their own theory provides the foundation and justification for a consistent frequentist and objective interpretation of probability. And this justification relied on the fundamental lemma, which in turn required stronger assumptions, assuming a class of models existing across an extended period while refraining from considering a new model for new data, as Fisher requires.

Just like any additional proven proposition of mathematics, Neyman–Pearson's lemma represents an unequivocal progress—one should think. But it is precisely this effort at extending and completing statistical logic that Fisher pokes fun at. For him, it was not a weakness, but a strength of the logic of inference that new data will create a new situation.

The two views could be summarized as follows: Neyman–Pearson wish to extend the mathematical argumentation in order to find a mathematical justification for the likelihood principle used by Fisher. They succeed in mathematizing decision processes. This achievement, however, comes at a cost: the requirements of the lemma. Fisher's criticism concerned the argument's applicability, not its consistency. Thus, the controversy between Fisher and Neyman–Pearson was about how far statistical methods, in the sense of a rational proposal for solving a problem of application, can be formulated, and thus also formalized, in the world of models. This presents a remarkably ironic point: It was precisely what Neyman–Pearson admired most in Fisher, his conception of a model, that caused their dispute with him!

Reflexive mathematization, that is, the mathematization of mathematics, in this case of statistics, is quite a typical feature of modern mathematics–statistics being no exception here. The dispute about the activity of mathematical modelling, and about how to deal with problems of application, however, concerns a permanent problem of applied science. I think one cannot beg the question by simply assenting either to Fisher or to Neyman–Pearson, taking the views of either mediation or integration, as I have called them. They form indispensable, complementary aspects of applied mathematics. The balance of these aspects is decisive as shown by the following interpretation of EDA. It started as a critical assessment of the use of models in statistics and in the end, it is argued, EDA provides a new systematic role for models. It does so in the frame of a conception of applied mathematics

which can be termed moderate and expansive at the same time; moderate because it is far distant from claims to rationality like those of Neyman–Pearson, and expansive as it also includes questions concerning the first steps in the process of modelling.

5 Exploratory data analysis as a model-critical approach

EDA represents an influential current in modern statistics. It was initiated and propagated by John Wilder Tukey. The approach was already known among specialists in the 1960s, and Tukey's book with the programmatic title of 'Exploratory Data Analysis' appeared in 1977. Quite in contrast to its influence on the practice of statistics, EDA is often neglected in philosophically-oriented considerations. In the context of models, EDA is of great interest, because Tukey combined a strong critique of the concept and use of models with his programmatic design. What is data analysis about? The 'International Encyclopedia of Statistics' summarizes:

> Exploratory data analysis is the manipulation, summarization, and display of data to make them more comprehensible to human minds, thus uncovering underlying structure in the data and detecting important departures from that structure (Kruskal and Tanur [1978], p. 3).

Note the fine, but decisive difference from Fisher's account of the general goal in which 'reducing the data to relevant information' took the key role, a fact which again required reference to an underlying model. EDA, in contrast, concerns a process preceding the construction of a model, as it were, preparing the terrain for the modelling, and it does without strongly structuring elements like hypothetical populations or even hypotheses. Tukey conceived of EDA very consciously as a countermodel and as a supplement to *confirmatory data analysis* (CDA), as he called the Neyman–Pearson tradition. Fisher is taking something like an intermediate position here, as I am going to argue.

In data analysis we must look to a very heavy emphasis on judgment. At least three different sorts of judgement are likely to be involved in almost every instance:

> **a1.** judgement based upon the experience of the particular field of subject matter from which the data come,
>
> **a2.** judgement based upon a broad experience with how particular techniques of data analysis have worked out in a variety of fields of application,
>
> **a3.** judgement based upon abstract results about the properties of particular techniques, whether obtained by mathematical proofs or empirical sampling (Tukey [1962], p. 9).

In a certain sense, Tukey considered mathematical models in statistics to be a dangerous gift, as they suggested the applicability of rigorous mathematical arguments. Often, Tukey says, the complex difficulties arising from amorphous data are passed over too quickly. In other words, Tukey was convinced that application-oriented statistics must begin methodologically even *before* the data are inserted into the context of a model, or rather into the Procrustean bed of a model. For Tukey, mathematical, model-dependent arguments should enter at a late stage of the application process which would have to begin with exploring the data without bias by modelling assumptions. For instance, the judgement what part of the data are outliers and may therefore be ignored is often decided too quickly by reference to a model. For him, the very process of model building has to be guided by EDA—a position quite contrary to Neyman–Pearson's integrative effort.

EDA offers a complete toolbox of methods of representation, *techniques of data analysis*, that consistently do not involve a model, neither in Neyman–Pearson's nor in Fisher's sense, and have nothing to do with hypotheses either. The stem-and-leaf diagrams are one from a large number of examples. Tukey illustrated the relationship between explorative and confirmatory data analysis with the metaphor of the detective and the judge:

> Unless the detective finds the clues, judge or jury has nothing to consider. *Unless exploratory data analysis uncovers indications, usually quantitative ones, there is likely to be nothing for confirmatory data analysis to consider.* (Tukey [1977], p. 3).

Was that not the initial motivation of modelling as well? Modelling was indeed also one of the prerequisites for applying mathematical propositions to reality, by having models bring a practical situation into a sufficiently exact form. While Tukey does not challenge this, he insists on the systematic importance of the first preparatory steps in the process of modelling. His main issue is to clarify how the judgement necessary to construct an adequate mathematical–statistical model can itself depend on an investigation by means of mathematical tools. This extended frame of mathematical tools (far from deductive reasoning) then encompasses decidedly less precise concepts. In this connection, Tukey pleads in favour of vague concepts, a rather uncommon recommendation, at least in a mathematical context:

> Effective data analysis requires us to consider vague concepts, concepts that can be made definite in many ways. To help understand many definite concepts, we need to go back to more primitive and less definite concepts and then work our way forward (Mosteller and Tukey [1977], p. 17).

At the very outset of a problem of application, Tukey says, there are generally quite a number of possible ways to attain a more abstract, more rigorous, or more precise formulation of the problem. This view taken by Tukey recalls

Fisher's position that there are a multitude of possible infinite populations which come under consideration during the first steps of modelling.[15] Confirmatory data analysis assumes a class of models with the intention of extracting a testimony from the data, while EDA strives to have the data speak for themselves. To *assume* a model class (which is different from finally *aiming at* a model) is an 'epistemological restriction' of the confirmatory approach, as Biehler ([1982]) appropriately notes in his in-depth study of EDA. Fisher's and Tukey's conceptions do not contradict one another; rather, what becomes evident if one integrates the two is that the process of modelling is based on an interplay of data and models in the course of which *both* have to be considered variable. When Tukey and Wilks ([1970]) underline that using models to evaluate data is different from data to evaluate models, they do not intend to play down the use of models, but rather assign some autonomy to both approaches that then require mediation.[16] Models are an essential and central component, but are neither the alpha nor the omega of statistical inference.[17] Taking this view indeed places Tukey outside the logic of testing established by Neyman–Pearson, and rightfully opposes EDA to CDA. At the same time, Tukey takes up Fisher's approach again inasmuch as he takes into account the importance of the process of construction in which modelling consists (this is where Biehler is absolutely right).

This systematic role for models is connected with a conception of applied mathematics that can be termed moderate and expansive at the same time: moderate because it is far distant from claims to rationality like those of Neyman–Pearson; and expansive, because it also includes questions concerning the first steps in the process of modelling.

With EDA, Tukey begins by pursuing a model-critical programme, introducing a set of new tools like stem-and-leaf diagrams that are intended to make the explorative analysis of the data possible. EDA has introduced an entire class of new tools that are fundamentally based on the capacities of modern computers, in particular visualization. It seems quite plausible to me to conceive of models as mediating instruments; EDA therefore is based on a new concept of models in statistics. EDA may well be seen as herald of an instrument-driven and still ongoing multifaceted change in modern statistics that is triggered by the computer.

[15] The statistician Kimball has coined the term 'errors the third kind'. They occur by introducing mathematical models too hastily, suggesting an exact treatment of problems while placing it in an inadequate frame: 'A simple and almost ludicrous definition of the error of the third kind is *the error committed by giving the right answer to the wrong problem*' ([1957], p. 134).

[16] EDA starts from an unstructured set of data, different from Fisher's considerations about the design of experiments in ([1935]), in which he intends to *obtain* data in an effective way.

[17] Incidentally, this weakens Biehler's point of view (in [1982]), who saw a shift from the priority on models to that on methods. In my opinion, the point is mediation, and not priority.

The analysis of modern statistics has led to several concepts of model. In my opinion, this should not tempt anybody to conceive of these concepts in an ever more general way in order to cover all possible modes of use. Such a comprehensive concept would not be desirable at all: In the case of modern statistics, the differences between the various concepts of model were the key to its formative controversies.

Acknowledgements

I would like to thank Margaret Morrison and an anonymous referee for their highly valuable comments on earlier versions of this paper.

University of Bielefeld
PB 100131, 33501 Bielefeld
Germany
johannes.lenhard@uni-bielefeld.de

References

Biehler, R. [1982]: *Explorative Datenanalyse—eine Untersuchung aus der Perspektive einer deskriptiv-empirischen Wissenschaftstheorie*, Bielefeld: Institut für Didaktik der Mathematik.

Braithwaite, R. B. [1953]: *Scientific Explanation*, Cambridge: Cambridge University Press.

Cartwright, N. [1999]: *The Dappled World. A Study of the Boundaries of Science*, Cambridge: Cambridge University Press.

Fisher, R. A. [1922]: 'On the Mathematical Foundations of Theoretical Statistics', *Philosophical Transactions of the Royal Society of London A*, **222**, pp. 309–68. Reprinted in Fisher [1971].

Fisher, R. A. [1925]: *Statistical Methods for Research Workers*, Edinburgh: Oliver and Boyd.

Fisher, R. A. [1935]: *The Design of Experiments*, Edinburgh: Oliver and Boyd.

Fisher, R. A. [1939]: 'Student', *Annals of Eugenics*, **9**, pp. 1–9.

Fisher, R. A. [1955]: 'Statistical Methods and Scientific Induction', *Philosophical Transactions of the Royal Society of London B*, **17**, pp. 69–78.

Fisher, R. A. [1971]: *Collected Papers of R. A. Fisher*, J. H. Bennett (*ed.*), Adelaide: University of Adelaide.

Ghamari-Tabrizi, S. [2000]: 'Simulating the Unthinkable: Gaming Future War in the 1950s and 1960s', *Social Studies of Science*, **30**, pp. 163–223.

Giere, R. N. [1999]: *Science Without Laws*, Chicago: University of Chicago Press.

Gigerenzer, G., Swijtink, Z. and Porter, T. [1989]: *The Empire of Chance*, Cambridge: Cambridge University Press.

Hacking, I. [1965]: *Logic of Statistical Inference*, Cambridge: Cambridge University Press.

Hacking, I. [1980]: 'The theory of probable inference: Neyman, Peirce and Braithwaite', in D. H. Mellor (*ed.*), *Science, Belief, and Behaviour*, Cambridge: Cambridge University Press, pp. 141–60.

Kimball, A. W. [1957]: 'Errors of the Third Kind in Statistical Consulting', *Journal of the American Statistical Association*, **52**, pp. 133–42.

Kruskal, W. H. and Tanur, J. M. (*eds*) [1978]: *International Encyclopedia of Statistics*, New York: Free Press.

Kyburg, H. E. J. [1974]: *The Logical Foundations of Statistical Inference*, Boston: D. Reidel.

Lehmann, E. L. [1993]: 'The Fisher, Neyman–Pearson Theories of Testing Hypotheses: One Theory or Two?' *Journal of the American Statistical Association*, **88**, pp. 1242–9.

Morrison, M. [1999]: 'Models as autonomous agents', in M. S. Morgan and M. Morrison (*eds*), *Models As Mediators. Perspectives on Natural and Social Science*, Cambridge: Cambridge University Press, pp. 38–65.

Morrison, M. [2002]: 'Modelling Populations: Pearson and Fisher on Mendelism and Biometry', *British Journal for the Philosophy of Science*, **53**, pp. 39–68.

Mosteller, F. and Tukey, J. W. [1977]: *Data Analysis and Regression: A Second Course in Statistics*, Reading, MA: Addison-Wesley.

Neyman, J. [1935]: 'Statistical Problems in Agricultural Experimentation', *Journal of the Royal Statistical Society*, **2** (Supplement), pp. 107–80.

Neyman, J. [1957]: ' "Inductive Behavior" as a Basic Concept of Philosophy of Science', *Revue d'Institute International de Statistique*, **25**, pp. 7–22.

Neyman, J. and Pearson, E. S. [1928]: 'On the Use and Interpretation of Certain Test Criteria for Purposes of Statistical Inference', *Biometrika*, **20A**, pp. 175–240, 263–94.

Neyman, J. and Pearson, E. S. [1933a]: 'On the Problem of the Most Efficient Tests of Statistical Hypotheses', *Philosophical Transactions of the Royal Society of London A*, **231**, pp. 289–337.

Neyman, J. and Pearson, E. S. [1933b]: 'The Testing of Statistical Hypotheses in Relation to Probabilities A Priori', *Proceedings of the Cambridge Philosophical Society*, **29**, pp. 492–510.

Pearson, E. S. [1966]: 'The Neyman–Pearson Story: 1926-34', in F. N. David (*ed.*), *Research Papers in Statistics. Festschrift for J. Neyman*, London: John Wiley and Sons, pp. 1–24.

Pearson, E. S. [1968]: 'Studies in the History of Probability and Statistics. XX. Some early correspondence between W. S. Gosset, R. A. Fisher and Karl Pearson, with notes and comments', *Biometrika*, **55**, pp. 445–57.

Pearson, K. [1892]: *The Grammar of Science*, London: Walter Scott.

Provine, W. B. [1986]: *Sewall Wright and Evolutionary Biology*, Chicago: University of Chicago Press.

Reid, C. [1982]: *Neyman from Life*, New York: Springer.

Seidenfeld, T. [1979]: *Philosophical Problems of Statistical Inference*, Boston: D. Reidel.

Tukey, J. W. [1962]: 'The Future of Data Analysis', *Annals of Mathematical Statistics*, **33**, pp. 1–67.

Tukey, J. W. [1977]: *Exploratory Data Analysis*, Reading, MA: Addison-Wesley.

Tukey, J. W. and Wilks, M. B. [1970]: 'Data Analysis and Statistics: Techniques and Approaches', in E. R. Tufte (*ed.*), *The Quantitative Analysis of Social Problems*, Reading, MA: Addison-Wesley, pp. 370–90.

Wald, A. [1950]: *Statistical Decision Functions*, New York: John Wiley and Sons.

Winsberg, E. [2003]: 'Simulated Experiments: Methodology for a Virtual World', *Philosophy of Science*, **70**, pp. 105–25.