Statistical Inference

Thomas Rahlf

Contents

Abstract

Statistical and, subsequently, econometric inferences have not undergone a cumulative, progressive process. We have seen instead the emergence of a number of different views, which have often been confused with each other in textbook literature on the subject. It therefore makes sense to approach the issue from a historical-scientific angle rather than a systematic one. We intend, using the extraordinarily complex development as a basis, to give a historical overview of the emergence of concepts that are of particular importance from the point of view of cliometrics. We shall start by describing the beginnings of modern probability theory, along with its connection with other statistical approaches.

T. Rahlf (✉)
German Research Foundation, Bonn, Germany
e-mail: thomas.rahlf@dfg.de

C. Diebolt, M. Haupert (eds.), *Handbook of Cliometrics*,
https://doi.org/10.1007/978-3-030-00181-0_16

The following overview covers the basic principles of the current concepts of inference developed by R. A. Fisher on one hand and by J. Neyman and E. S. Pearson on the other. Neo-Bayesian approaches have meanwhile been developed in parallel, although they were not taken into account during the initial founding phase of econometrics. A "classic" approach was instead adopted in this respect, albeit with an additional difficulty: the taking into account of time. Cliometrics initially followed a Bayesian approach, but this did not finally prevail. Following on from econometrics, a correspondingly classic, inference-based position was adopted. This chapter concludes with a reference to a fundamental critique of the classic position by Rudolf Kalman, which we also find very promising as an inference-related concept for cliometrics. We often quote authors directly, in an effort to portray developments more vividly.

Keywords
Probability · Inference · Bayesianism · Frequentism · System theory

Introduction

Statistical inference possesses an ambivalence that is present in virtually no other field of science. Current doctrine is built up consistently on one hand (an impression furthermore reinforced an interdisciplinary examination of the relevant literature) across all disciplinary boundaries and along the same strictly schematic lines. The impression given is that it is a logically structured, self-contained edifice possessing universal validity. While the significance of individual methods can differ from subject to subject, their inherent statistical inference-related principles (with particular reference to the method of testing hypotheses and, more generally, assessment of the "evidence" supplied by statistical data) appear to be universally valid.

This was the objection expressed by Gerd Gigerenzer et al. (1989, p. 105f) regarding contradictory and illogical "hybridization":

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

A look "inside" the statistics gives a variegated impression. Certain quotations from the literature on statistics serve to illustrate the controversies within this area of science. Examples include R. A. Fisher (1956, p. 9), who stated, "The theory of inverse probability is founded upon an error, and must be wholly rejected." Von Mises (1951, p. 188) admitted, with reference to Fisher's "likelihood approach": "The many fine words that Fisher and his followers use to justify the likelihood theory are incomprehensible to me. The main argument [. . .] has nothing to say to me." A. Birnbaum, who brought up the likelihood concept in a widely read contribution to the likelihood principle as a fundamental basis of statistical inference, rejected the confidence principle developed by J. Neyman and Pearson on the grounds of its opposition to the likelihood principle (Birnbaum 1962; Neyman and

Pearson 1928a, b, 1933). He went on to reject the likelihood principle a few years later, however, precisely because of its opposition to the confidence principle.[1] Stegmüller (1973, p. 2) refers to Neyman, who had claimed that the test methods developed by Fisher "[. . .] were, *in a mathematically-definable sense*, 'worse than useless' [. . .]."[2] B. de Finetti (1981), one of the main representatives of a subjectivist theory of probability, was convinced that Fisher "[. . .] showed his feel for the necessity of a conclusion in Bayesian form (with the illusion of being able to express them with an indefinable 'fiducial probability'), with a desire to present the problem in a way that was opposed to the Bayesian approach (like Neyman, essentially)." L. J. Savage (1954), another important defender of a subjectivist approach, who wanted to incorporate into his influential book, *The Foundations of Statistics*, the conventional statistical inference methods developed by him as part of an axiomatic system of a subjectivist doctrine, wrote in the book's second edition: "Freud alone could explain how the rash and unfulfilled promise (made early in the first edition, to show how frequentist ideas can be justified by means of personalistic probabilities) went unamended through so many revisions of the manuscript."[3] O. Kempthorne (1971) finally characterized the various concepts of inference in a way that caused J. W. Pratt (1971 commentary, p. 496) to summarize Kempthorne's theses as follows: "Fiducial and structural methods are nonsense. Jeffrey's Bayesian and subjective Bayesian methods are nonsense. Likelihood methods are nonsense. He doesn't say directly that orthodox methods are nonsense, but he says it implicitly by his remarks [. . .]. In short, he says all methods are nonsense, therefore use orthodox methods." This list could easily be extended, but the impressions given should suffice.

Probability and Inference in Statistics

We would like to start by giving a broad-brush description of how the central concepts have developed.[4] The following milestones mark the most important steps along the way:

Historical milestones in the field of statistical inference

1700–1730	The first systematic definitions of the terms "probability" and "chance" (G. W. Leibniz, J. Bernoulli) and the attempt to arrive at statistical inference (as a conclusion) from probability theory (J. Bernoulli)
1750–1775	Inversion of the probability concept in connection with the error function of Laplace
	Inversion of the probability concept in connection with Bayesian binomial distribution

(*continued*)

[1] Cf. Birnbaum (1962, 1968, 1977).

[2] Original author's italics.

[3] Quoted from DuMouchel (1992, S. 527). The first edition was published in 1954. Cf. Savage (1954).

[4] A detailed treatment of the topic of this chapter can be found at Rahlf (1998) and Gigerenzer/Swijtink/Porter/Daston/Beatty/Krüger (1989).

Around 1810	Synthesis of error function and probability by P. S. Laplace and C. F. Gauss
1820–1840	The further development of statistical inference concepts (law of errors, law of large numbers) and their incorporation into the "social physics" of A. Quetelet
1870–1885	The incorporation of Quetelet's concepts into biology by F. Galton and the conceptual foundation of correlation and regression
1840–1870	Philosophical investigations into the concept of probability (as a parallel development)
1880–1895	The systematization and formalizing of statistical inference concepts by F. Y. Edgeworth and K. Pearson
1895–1900	The application of these systematized concepts of statistical inference to social science data and development into multiple regression by G. U. Yule
	Attempts by K. Pearson and G. U. Yule to clarify the concepts of correlation, spurious correlation, and causality
Around 1900	The concept of the significance test is developed by K. Pearson
1929/1930	The criteria of "good" valuations postulated by R. A. Fisher, a quantitative assessment of the quality of these valuations using the fiducial principle also developed by Fisher
1933	"Classic" test theory and confidence inference according to J. Neyman and E. S. Pearson
1926–1954	Subjectivist Bayesian approaches, such as those of F. P. Ramsey, B. de Finetti, H. Jeffreys, or L. J. Savage
1955	Objectivist Bayesian approaches, such as those of H. Robbins
1949/1962	The likelihood principle developed by Fisher and expanded by G. Barnard and A. Birnbaum

The theory of probability was regarded as something of a "brainteaser" until the middle of the seventeenth century, in the sense of pure combinatorics. The chance of rolling a certain dice number, of a tossed coin falling on one face or the other, or of drawing a particular card from the pack could be indicated without any profound philosophical consideration of the nature of probability. The probability of, for example, tossing a coin ten times and having it come up "heads" four times and "tails" six was to be determined by a combination of purely mathematical considerations, as a coin-tossing "experiment" could be based on a fully specified theoretical model: The events are mutually independent, thus making their sum binomial, the parameter being $\pi = 0.5$.

The questions that arose in a socioeconomic context at this time were, however, only apparently of the same nature. Even variables such as overall gender ratio, life expectancy, infant mortality rates, the proportion of the population available for military service, etc., were considered legitimate in this respect. But how was one to assess the *reliability* of the results obtained?

It was decisive that studies like those carried out by J. Graunt (1662 [1939]) of the register of deaths in London or by E. Halley (1693) of births and deaths in Breslau (present-day Wroclaw) attracted the attention of mathematicians such as G. W. Leibniz, J. Bernoulli, or A. de Moivre, thereby obliging those concerned to

consider the problem of inference. The proportion of people possessing a certain characteristic was unknown, as long as the characteristic concerned could not be computed for a given (sub)population. A possible entitlement to apply the binomial model existed, but there was definitely no theory capable of postulating a value for the parameter that was to be verified. Furthermore, this value could only be determined on the basis of the data and a measure indicated – by means of an interval – for the accuracy of the "estimate." An *inversion* of probability was therefore necessary, although neither Bernoulli nor de Moivre was able to complete this step. We follow S. Stigler at this point while assuming that the conceptual difficulties could be overcome only via the detour of the error function, ultimately by T. Bayes and P. S. Laplace. This "Copernican Revolution" in the development of theoretical statistics[5] was connected with the intention of Bernoulli. It is somewhat curious that this concept is nowadays associated with Bayes rather than Laplace. Bayes had a groundbreaking idea that was nevertheless developed at the same time and, presumably independently, by Laplace. However, Laplace had also constructed a systematic theory of probability that went on to form the basis for a number of applications over many years. The key statisticians (Gauss, Galton, and Edgeworth) subsequently followed a mainly Bayesian line of argument. The most probable parameter value for Gauss, for example, was the maximum of the likelihood function, since it emanated, as it also did for Laplace, from the principle of insufficient reason and thus from an a priori uniform distribution. K. Pearson meanwhile followed a (mostly) sampling-based approach however, and G. U. Yule worked within the same framework, albeit without attributing much importance, in general, to the question of inference.[6]

K. Pearson and G. U. Yule

The works of K. Pearson were of great significance for the further development of statistical inference. Pearson's first independent contribution to the field of statistics, which formed the basis of his subsequent fame, consisted of a system of frequency distributions included in two extensive papers published in the *Philosophical Transactions of the Royal Society* under the title *Contributions to the mathematical theory of evolution* (1894, 1895), which led to him being elected a fellow of the society. The question regarding the form of frequency distributions had been a fundamental issue since the end of the eighteenth century. There was a prevailing general belief that individual phenomena, which were homogeneous in the sense of many individually insignificant influencing factors, had to follow a normal distribution. Not everyone regarded normal distribution as being universally valid however, and collections of data that accumulated over the years implied a series of "skewed" distributions. Pearson above all regarded this fact as a challenge, and he eventually developed a

[5] Stigler (1986, p. 122).

[6] See, for example, Yule (1895, 1896a, b) and Pearson (1898).

"family" of curves, each based on four parameters, by which data could be assigned to different types of curve using their first four moments.

Pearson supplied not only the formulae but also a wealth of practical examples (distribution of air pressure, heights of schoolchildren, and sizes of crustaceans; statistics on poverty and divorce rates; etc.) and showed that these variables could be reconciled to a large extent by using his system. He went even further than Quetelet in this respect. It was not only data with a normal distribution that followed a uniform distribution law, without a need to isolate groups or major factors, but also many others whose distribution was in fact skewed but no less legitimate in this respect. If this were the case, the search for causative factors, as introduced by Galton as part of biology, was invalid:

> The law of frequency is based on the assumption of perfect ignorance of causes, but we rarely *are* perfectly ignorant, and where we have any knowledge it ought of course to be taken into account.[7]

The further application of Pearson to areas that are not necessarily closely subject to a law of constant distribution has been criticized[8]:

> [. . .] I see that there are many cases of 'skew' variation: but all cases which he has given, of variation with an unmistakably skew frequency, are taken from phenomena which are changing with a rapidity much greater than that of any organs in crabs, or such creatures. Pauperism, divorces, and the like, have only been invented, in their present form, for a short time, and as he himself shows, the maximum frequency changes its position at least in ten years.[9]

But the most important counterargument was that the numerous forms that could be adapted using Pearson's frequency curves lacked a theoretical foundation, as they were purely empirical constructs. If a frequency distribution did not lend itself to being represented by a normal distribution, the concept of causation based on a large number of random causes could not be effective. It is precisely this last point that was however, according to Stigler (1986, p. 339), likewise not the intention of Pearson, who was seen to represent a philosophy of science that had been guided by Kantian nominalism. On this basis, Pearson regarded frequency curves only as mental constructs that summarize empirical evidence, without providing any statements on possible causes. Pearson nevertheless also searched in this respect for a formal criterion for assessing deviation in the empirical distributions of his frequency curves and finally found one in the form of his chi-squared (χ^2) test, which he made public in 1900.

[7] F. Galton in a letter to K. Pearson of 18 Nov 1893, quoted by Stigler (1986, p. 336). Original author's italics.

[8] Despite criticism, Pearson's frequency curves soon became part of the standard repertoire of statistics.

[9] W. F. R. Weldon in a letter to F. Galton of 27 Jan 1895, quoted by Stigler (1986, p. 337).

Pearson made another important contribution to modern statistics in the field of correlation. He considered two variables with a normal bivariate distribution, deduced the correlation coefficient and a posteriori distribution[10] (on the basis of empirical standard deviations), and systematized the findings obtained to date. The theoretical derivation was followed by a series of applied examples, which he took from Galton. He did not admit any major possibilities regarding the application to social phenomena:

> Personally I ought to say that there is, in my own opinion, considerable danger in allying the methods of exact science to problems in descriptive science, whether they be problems of heredity or of political economy; the grace and logical accuracy of the mathematical process are apt to fascinate the descriptive scientist that he seeks for sociological hypotheses which fit his mathematical reasoning and this without first ascertaining whether the basis of his hypotheses is as broad as that human life to which the theory is to be applied.[11]

This move was finally made by Pearson's student G. U. Yule in a series of studies of Poor Law legislation. One important question in this respect was the extent to which the proportion of poor people in a given district was connected with its structure of care provision. Yule (1895, 1896b) found a "significant" link, which he nevertheless described as "suggestive," as the distributions of both variables were clearly shown to be skewed. In a subsequent step, he established a "regression line" between the two variables by minimizing the distances between this straight line and the data concerned. He perceived that this approach was easy to extend to higher dimensions, thereby leading to the "normal" system of equations that had been introduced by Gauss several decades earlier in the field of astronomy. From here, it was merely a technical matter, no longer requiring any conceptual step, to extend the approach to more than two variables.

Irrespective of the different views held by K. Pearson and Yule in this context regarding the concepts of correlation and causality, the general question surrounding all these considerations was the following: Did inference refer to a *population* or to *laws*? This was clear in Pearson's case and also, subsequently, in that of Fisher. The aim of studying biological data was to investigate conformity to natural laws. The situation was more difficult when it came to the investigations of socioeconomic data carried out by Yule or studies, such as those of Gosset, of the correlations between

[10]K. Pearson explicitly rejected the concept of inverse probability, although E. S. Pearson was of the view that he implicitly followed this approach on at least one occasion. Cf. Pearson (1898). "The basic of the approach used here is a little obscure and there seems to be implicit in it the classical concept of inverse probability" (Pearson 1967, p. 347), quoted by Dale (1991, p. 379). Pearson expressed himself most extensively on this issue in his paper *The fundamental problem of practical statistics* (1920), which has provoked different interpretations up to the present day. While Fisher (1922, p. 311), for example, believed he recognized a proof of Bayes' theorem in it, Dale (1991, p. 388) considered this as a "totally inaccurate observation." For further interpretations, cf. ibid., pp. 377-391. According to Stigler (1986, p. 345), Pearson worked on multiple occasions "[. . .] (implicitly) in a Bayesian framework."

[11]Pearson (1898, p. 1f), quoted by Stigler (1986, p. 304).

cancer rates and apple consumption, which included at least one exploratory element.[12] The interpretation of a correlation coefficient could only be hypothetical according to Yule, as it was normally possible to give a variety of alternative explanations whose distinction could not be provided by statistics. This problem would prove to be fundamental for statistical inference-based interpretations in the field of social science.

R. A. Fisher

The further development of statistical methodology in the field of biology has been characterized, since at least the time of Karl Pearson and R. A. Fisher, by the possibility of its application to the natural sciences. Fisher (1955, 1956, 1959) attempted to solve, by means of his *Design of Experiments*, the problems of inference-based conclusions in biology caused by their dependence on the conditions that prevail when taking samples.

Fisher's concept of inference was initially characterized by its explicit rejection, directed against Pearson in particular (in 1922), of inverse probability. This view was mainly due, in his opinion, to the confusing of theoretical parameters and estimates:

> It is this last confusion, in the writer's opinion, more than any other which has led to the survival of the present day of the fundamental paradox of inverse probability, which like an impenetrable jungle arrests progress towards precision of statistical concepts.[13]

He nevertheless developed a certain understanding at the same time:

> The criticisms (. . .) have done something towards banishing the method, at least from the elementary text-books of Algebra; but though we may agree wholly (. . .) that inverse probability is a mistake (perhaps the only mistake to which the mathematical world has so deeply committed itself), there yet remains the feeling that such a mistake would not have captivated the minds of Laplace and Poisson if there had been nothing in it but error.[14]

Although Fisher's concept of probability was frequentist, he vehemently rejected a definition of probability as a limit value applying to relative frequency in an unlimited number of repeated attempts (i.e., the von Mises definition subscribed to

[12]Cf. ibid., p. 373.

[13]Fisher (1922 [1992], p. 13), similar also to Fisher (1959, p. 34). There is in the case of Fisher (1956, p. 9) a (more or less) clear rejection of the Bayesian approach. He emphasized that he was "personally convinced" that "the theory of inverse probability is founded upon an error, and must be wholly rejected."

[14]Fisher (1922 [1992], p. 13). Ambiguities such as these are characteristic of Fisher's work. According to Geisser (1992, p. 4), Fisher subscribed – until at least 1912 – to approaches based on Bayesian logic. He then (p. 26f) explicitly rejected the validity of Bayes' theorem. Cf. Barnard (1988) regarding this question.

by most frequentists)[15]: "For Fisher, a probability is the fraction of a set, having no distinguishable subsets, that satisfies a given condition [. . .]."[16]

Fisher postulated that statistical inference should refer to theoretical, and thus fixed, parameters of hypothetically infinite populations, thereby determining the direction of research in the field of theoretical statistics for the following 50 years.[17] Otherwise, his concept of a statistical or "scientific" inference could not prevail. He used the term "inductive logic," not at least in order to set himself apart from the approach of his intellectual rival J. Neyman, who spoke of "inductive behavior."[18] It was possible, in cases where there was an indisputable a priori distribution, to speak of the probability of events, which were to be described as fiducial probabilities.[19] Intervals that express the uncertainty of an estimate were always to be construed as fiducial intervals.

The problem of the "significance test" is closely connected to the problem of using intervals to indicate the accuracy of an estimate. What we now understand as the logic of the significance test became increasingly important during the first two decades of the twentieth century.[20] It can largely be traced back to Fisher and has remained in force alongside the concept of the hypothesis test developed by Neyman and Pearson (see below). For Fisher, the level of significance of a test is a *measure of evidence*, which should neither be defined a priori nor regarded as unalterable, nor established as a guiding principle:

> 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. For when the hypothesis is correct he will be mistaken in just 1% of these cases, and when it is incorrect he will never be mistaken in rejection. This inequality statement can therefore be made. 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

[15]See supporting evidence in Savage (1976, p. 461). In Fisher (1959, p. 32), he emphasized, for example, that no probability of individual events could be established with such a definition.

[16]Savage (1976, p. 461) with corresponding supporting evidence. Savage observes in this respect: "Such a notion is hard to formulate mathematically, and indeed Fisher's concept of probability remained very unclear, which must have contributed to his isolation from many other statistical theorists" (p. 462).

[17]Cf. Geisser (1992). Partly ambiguous terms such as "mean," "standard deviation," or "correlation coefficient" have remained in use to this day to indicate, in various contexts, either theoretical variables or estimators for these theoretical variables.

[18]Cf. Savage (1976, S. 462) with supporting evidence.

[19]Ibid., p. 466: "Nobody knows just what they mean [. . .]. In a word, Fisher hopes by means of some process – the fiducial argument – to arrive at the equivalent of posterior distributions in a Bayesian argument without the introduction of prior distributions [. . .]." We would like to join in with this criticism. As observed by Menges (1972, p. 275): "The fiducial concept considers the results of an observation as indisputable fact in this respect, and as the basis on which to build inference. *It can thus do justice, in principle, to the historical character of social phenomena*" (original author's italics), although this also applies to Bayesian logic in our opinion.

[20]Such as Pearson's chi-squared goodness-of-fit test of 1900, Student's *t*-test, developed in 1908 and formalized by Fisher, or the *F*-test applied to the analysis of variance by Fisher.

> circumstances, he rejects hypotheses; he rather gives his mind to each particular case in the light of his evidence and his ideas.[21]

This criticism was directed against a concept that had been propagated by J. Neyman and E. S. Pearson since the 1930s and which had quickly become the dominant view.

J. Neyman and E. S. Pearson

The works of J. Neyman and E. S. Pearson are likewise unanimously considered to be milestones in the history of theoretical statistics. While Fisher wished to allow, in relation to the testing of hypotheses, only the alternatives "rejection" and "no statement possible," Neyman and Pearson developed a closed test theory which introduced differentiated levels of rejection and acceptance, along with concepts such as the "power" of a test, Type I and Type II errors, and "uniformly most powerful test." Until the end of the nineteenth century, the testing of hypotheses was based on distributions of test statistics which were (1) best suited for use with large samples and (2) employed for intuitive reasons. The introduction of the *t*-distribution by W. S. Gosset (1908) and the contributions of R. A. Fisher, who differentiated the exact distributions of t, χ^2, F, and certain correlation coefficients in normal distributions, meant that at least problem (1) could be overcome. With this problem solved, the question then posed was that of a formally satisfactory test *theory*. E. S. Pearson stated in a review that the idea for this theory came to him via an observation made by Gosset:

> I had been trying to discover some principle beyond that of practical expediency which would justify the use of "Student's" ratio $z = (-m)/s$ in testing the hypothesis that the mean of the sample population was at m. Gosset's reply (to the letter in which Pearson [...] had raised the question) 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 hypothesis explains the observed events with a greater degree of probability.[22]

Gosset argued in this letter that not even a probability value as small as 0.0001 could lead per se to rejection of a hypothesis for a random sample. Only comparison with an *alternative* hypothesis, "which will explain the occurrence of the sample with a more reasonable probability, say 0.05 (such as that it belongs to a different

[21] Fisher (1959, p. 41f). Fisher's failure to include tables of *p*-values in his famous textbook *Statistical Methods for Research Workers* (rather than the tables of significance values that he *did* include) arose from the fact that K. Pearson held the copyright to the former. Cf. Watson (1983, p. 714).

[22] Pearson in a paper from 1939 quoted from Lehmann's comments (1992, p. 68) on Neyman/Pearson (1933) (our italics).

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."[23]

This idea was then jointly developed by Neyman and Pearson (1928a, b) in an extensive two-part paper, published in *Biometrika*, on the concept of the likelihood ratio test. While Pearson now saw in this the uniform method for which they had been seeking, Neyman was clearly still not satisfied:

> It seemed to him that the likelihood ratio principle itself was somewhat ad hoc and was lacking a fully logical basis. His search for a firmer foundation, which constitutes the third of the three steps, eventually led him to a new formulation: The most desirable test would be obtained by maximizing the power of the test, subject to the condition that under the hypothesis, the rejection probability has a preassigned value, the level of a test.[24]

The result was Neyman and Pearson (1933), which also includes the famous Neyman-Pearson lemma. This states that in the class of all tests with probability α, the criterion function of the likelihood ratio test dominates the criterion function of any other test (i.e., every other test has a greater probability of including a Type II error). Neyman and Pearson used a series of examples to demonstrate the application of this principle and thus laid the foundation for a widely recognized general test theory which today continues to be regarded as "classic," along with the "confidence interval" likewise formulated by Neyman (1937). The method based on Neyman-Pearson logic can be described, after Lehmann, in terms of four steps[25]:

1. Specification of a model using a parametric family of distributions which has produced the data
2. Specification of a hypothesis with regard to a parameter of interest, H_0: $\theta = \theta_0$, and one simple or one class of alternatives H_1, e.g., $\theta \leq \theta_0$
3. Specification of a level of significance α, indicating the maximum allowable probability of a Type I error
4. Selection of the optimum method for testing H_0 against H_1 by minimizing the ß-error[26]

Lehmann finally added a – quite fundamental – fifth item, but this is more of a prerequisite than a procedure:

1. All (four) steps must be completed "before any observations have been seen."

[23]Ibid.

[24]Lehmann (1992, p. 68). This highly important aspect of the Neyman-Pearson theory is often not taken into account. As Borovcnik (1992, p. 92) rightly points out, "[. . .] a frequency interpretation places too much emphasis on the α-error during testing, while the real trick with this method is to minimise the ß-error."

[25]According to Lehmann (1992, p. 69f).

[26]We do not intend to go into the corresponding techniques here but refer instead to textbook literature on the subject.

The approach postulated by Neyman and Pearson actually amounted only to a set of *guidelines*. The two authors expressed, as follows, the conviction that lay behind their theory:

> Without hoping to know whether each separate hypothesis is true or false, we may search for rules to govern our behaviour with regard to them, in following which we insure that, in the long run of experience, we shall not be too often wrong.[27]

Inference statements are therefore hypothetical-deductive and only possible *before* events occur. They therefore do not refer to specific hypotheses but to future *actions* in the long term. This approach was consequently extended by A. Wald (1950) to form a pure decision theory, with Neyman repeatedly emphasizing this behavior theory aspect in his later work.

There was however vehement criticism from no less a person than R. A. Fisher, who might have wished to recognize the Neyman-Pearson theory for situations where permanent decisions had to be taken but was in no way willing to accept statistical inference-based assessments in a *scientific* sense. A further argument concerned the claim of "repeated sampling *from the same population*." Fisher pointed out, following on from J. Venn, that a given sample could always have resulted from a variety of conceivable populations: "so [. . .] the phrase 'repeated sampling from the same population' does not enable us to determine which population is to be used to define the probability level, for no one of them has objective reality, all being products of the statistician's imagination."[28]

These approaches were met with further reservations: On one hand, models would mostly be chosen *in practice* on the basis of data while often examining not just one but several hypotheses using the same data. In many situations, the eventual reduction of inference to a yes/no decision was not appropriate.

It has furthermore been demonstrated that optimum (i.e., uniformly most powerful) tests exist only for limited situations or are so complex (when maximizing their minimum power) that their application presents considerable problems. It should however be emphasized that these reservations are the exception and that an overwhelming majority has, particularly in the field of applied statistics, unconditionally accepted the Neyman-Pearson approach, which has become something of a paradigm, even though today's statisticians continue to argue about where the precise differences between this approach and Fisher's test concept lie.[29]

If we compare this approach to that of Fisher, point 5 (see above) becomes particularly decisive. The method according to Neyman and Pearson is therefore strictly deductive, while Fisher's approach is (also, at least) inductive, with assessment taking place only after obtaining evidence based on data and above all without

[27]Neyman/Pearson (1933 [1992], p. 74). Kyburg (1985, p. 119) sums up their intention in the observation: "That says nothing about the case before us, but it may make us feel better."

[28]Fisher (1955, S. 71).

[29]Cf. Lehmann (1993), for example.

considering an alternative hypothesis. Neyman and Pearson surely did not intend to promote a universal and constant level of significance but rather only in this sense: Even if they allow for different levels in different situations, these must be determined *before* the experiment and/or *before* obtaining any knowledge of the data evidence. The second fundamental difference lies in the *direction* of the inference. Fisher's test concept – and, in this respect, K. Pearson's logically equivalent significance test concept – applies to a state that exists or which, strictly speaking, may already have passed. The inference of Neyman and Pearson, on the other hand, applies to the future: If we act in one way or another in the future on the basis of the test, how often are we then likely to commit an error? The current practice is in fact to apply a blending of both concepts.[30]

Statistical inference was now reduced to the creation of guidelines for conduct in the long term. No contentious epistemological issues were settled using the Neyman-Pearson theory; it related only to a clear statement. Its success can perhaps also be explained by the fact that other positions (K. Pearson, Fisher) lacked such clarity.

The dominant approach since then has in any case been a supposedly objective, frequency-theory, and inference-based position, although a modern, Bayesian, statistical inference has continued to develop in parallel. It is remarkable that modern, subjectivist probability theory was not established by social scientists, who regarded as problematic its individual prerequisites or implications with regard to long-term experimental inference, but – without exception – by mathematicians (Ramsey, de Finetti, Savage) or geophysicists (Jeffreys) who saw problems of logic in the predominant frequency theory-based approaches.

This development took place in three stages: the reestablishing, by F. P. Ramsey, B. de Finetti, H. Jeffreys, and L. J. Savage, of Bayesian probability theories; the expanding, by G. A. Barnard and especially A. Birnbaum, of various likelihood-based approaches to form a likelihood *principle*; and finally the combining of these two components to create a modern Bayesian inference, which has come to exist in numerous forms. The following section considers at first the development of subjectivist probability theories.

Bayesian Probability

These are based on the following three basic assumptions, according to Howson (1995, p. 2):

[30]Johnstone (1986, p. 6) aptly describes the prevailing approach: "In general, tests of significance in practice follow Neyman formally, but Fisher philosophically. Formally, there is mention of 'alternative' hypotheses, errors 'of the second kind', and the 'power' of the test, which are terms due to Neyman (and his colleague Pearson). But philosophically, the result in a test, e.g. the result that the level of significance P equals 0.049, or that P is less than or equal to 5%, is interpreted as a measure of evidence, which is the interpretation following Fisher, and denied repeatedly by Neyman."

1. A hypothesis *A* is, in extreme cases, certainly true or certainly false. Intermediate degrees of belief in *A* are permitted.
2. These degrees of belief can be expressed numerically.
3. If they are rational and measured against the closed unit interval, they satisfy the finite additivity axioms.

The subjectivist Bayesian concepts of F. P. Ramsey, B. de Finetti, H. Jeffreys, and L. J. Savage were developed successively but independently of each other. We will now deal with them briefly in chronological order.

The first "modern" subjectivist probability theory was established by F. P. Ramsey in papers written in 1926 and 1928 but published posthumously in 1931.[31] As we have seen, the epistemological conception of probability from Bernoulli to Laplace was subjective as well as in the case of Gauss, Galton, and Edgeworth: "Probability" was interpreted by C. Huygens in terms of betting odds, with chance defined as ignorance. The principle of insufficient reason implied an a priori uniform distribution, which was linked, via Bayes' theorem, to the evidence from data in form of an a posteriori probability for a given parameter value.

Ramsey argued along similar lines, albeit combined with a critique of the logical and frequency theory-based interpretation. His starting point was John Maynard Keynes' *Treatise on Probability* (1921). For Keynes, probability meant a logical relationship between two different sets of propositions that are interconnected via a "degree of belief":

> Let our premises consist of any set of propositions h and our conclusion consist of any set of propositions a, then if a knowledge of h justifies a rational degree of belief in a of degree A, we may say that there is a probability-relation of degree A between a and h.[32]

Keynes did not however require all degrees of belief to be numerically measurable or comparable, thereby avoiding major difficulties. Ramsey postulated instead that probabilities should be expressed as betting odds, which must be rational (i.e., consistent and coherent). Ramsey's observations were of a purely philosophical nature and did not constitute a concept of *inference*. This was supplied in a famous paper by Bruno de Finetti (1937). It was totally clear to Finetti that the basis of all probability was subjective in nature.[33] Bayes' theorem was of central importance in this respect: Subjective assessments/probabilities must be revised constantly in the light of Bayes' theorem on the basis of data and knowledge obtained. This meant that subjectivist probabilities converge to relative frequencies as evidence accumulates. De Finetti did not criticize classical statistics for false results but for its false foundations:

[31] Ramsey (1931a, b).

[32] Keynes (1921, S. 4), quoted by Kyburg/Smokler (1964, p. 9).

[33] Cf. de Finetti (1937).

> The overwhelming majority of modern statistics are in practice completely normal, but their foundations are false. Intuition has however prevented statisticians from making mistakes. My thesis is that the Bayesian method justifies what they have always done, and that they are developing new methods which are missing in the orthodox approach.[34]

Harold Jeffreys (1939) argued along similar lines. He combined a probability theory with a theory of induction. Jeffreys stressed (like de Finetti) that a fundamental problem of science lay in learning from experience:

> Knowledge obtained in this way is partly merely description of what we have already observed, but partly consists of making inferences from past experience to predict future experience. This part may be called generalization or induction. It is the most important part; events that are merely described and have no apparent relation to others may as well be forgotten, and in fact usually are.[35]

It therefore follows that probability is not a frequency but a "reasonable degree of belief, which satisfies certain rules of consistency and can in consequence of these rules be formally expressed by numbers."[36] If an explanation is given for an observed event, a researcher might determine that it is "probably true." It is thus implied that he has a high degree of confidence in a hypothesis, which is in turn (1) quantifiable and (2) based on experience and information.[37] A rule now states how the cognitive process should operate: This is none other than Bayes' theorem. In every probability to which we assign a hypothesis, that hypothesis is conditioned by the information available to us. If this changes (increasingly), the probability associated with the hypothesis must be revised accordingly. This approach is what constitutes the basis of learning from experience, which is formalized using Bayes' theorem: A posteriori probabilities result from the evaluation of a priori probability with the data evidence, using the likelihood function.

L. J. Savage was another important forerunner of modern Bayesian probability theory. Savage, who was influenced mainly by Milton Friedman and John von Neumann, formulated his concept of probability in the late 1940s/early 1950s, on the basis of a utility theory. The year 1954 saw the publication of his seminal work *The Foundations of Statistics*, in which he tried to arrange within a unified framework the (in his view) rather loosely connected set of techniques developed by R. A. Fisher and J. Neyman/E. S. Pearson, intended to be based on a theory of decision-making under uncertainty. However, an examination of the details showed that the venture was doomed to failure. H. E. Robbins (1955) took a different path. He postulated probabilities that were "objective" and a priori rather than epistemic. He started with the question as to whether one could apply the Bayesian approach even if the a priori probability of a parameter is unknown but nevertheless "exists." This

[34] De Finetti (1981, p. 657).

[35] Jeffreys (1939, p. 8).

[36] Ibid., p. 401.

[37] Although the hypothesis can still be false in terms of rule 4.

supposition of an objectively existing a priori probability is not shared by most Bayesians however nor is it, in a positive sense, required.

Bayesian Inference

We have, in the case of the Bayesian works cited above, placed the issue of probability in the foreground. But there is a second Bayesian inference: the likelihood element. Approaches to likelihood initially emerged independently of Bayesian concepts. The likelihood ideas created by Fisher were further developed mainly by G. A. Barnard.[38] These ideas were given a basic theoretical foundation by the pioneering work of A. Birnbaum, who developed them into a likelihood *principle* (LP).[39] By this time, the field of statistics was already being dominated by the Neyman-Pearson approach and its decision theory-based further development by A. Wald (1950).

The likelihood principle had radical consequences. It stated that all the evidence from data was contained in the likelihood function. This made the sample space irrelevant, *after* the data had been obtained. It means that measures of evidence referring to the space of all possible data (i.e., the probability or parameter space), such as p-values or the confidence level, are irrelevant to inference *after* a given piece of data has been created. This was otherwise a rejection of the frequentist position, without having to resort to Bayesian arguments.

Let us now turn to the linking of a priori probabilities and likelihood inference to Bayesian inference. The Bayesian breakthrough eventually succeeded, in practical terms, with a paper by W. Edwards, H. Lindman, and L. J. Savage (1963), which finally made the corresponding approaches available to a wider public.[40]

Edwards, Lindman, and Savage dealt with the main reservations affecting the Bayesian approach, such as how scientific objectivity could be possible if different scientists held different a priori views, thereby creating different a priori probabilities (and probability distributions).[41] They did not bring in the argument proposed by Laplace and Edgeworth[42] (whereby an increase in the range of data causes the influence of a priori distribution to diminish progressively, before eventually disappearing altogether) but opted rather for the question as to whether an a priori distribution can be assumed to be uniform or whether the exact form of the a priori distribution is of no great importance to a posteriori distribution. They showed that "it suffices that your actual prior density change gently in the region favored by the

[38]Barnard (1947, 1949). For historical development, see Berger/Wolpert (1988, p. 22ff).

[39]Birnbaum (1962). Cf. also Bjornstad (1992) on the following. A "standard" work on the subject is that of Berger/Wolpert (1988).

[40]Edwards/Lindman/Savage (1963 [1992]). Our intention from here on is to deal only with certain ideas without going into technical detail.

[41]Ibid., pp. 534–540.

[42]For example, Laplace (1812) and Edgeworth (1884).

data and not itself too strongly favor some other region."[43] These vague indications were then given a mathematical form, thereby showing that such an approach is indeed justified under somewhat weak assumptions.[44]

The authors did however acknowledge, on the other hand, that there are also situations where the exact characteristics of a priori distribution are decisive.[45]

The following includes a section on "Bayesian hypothesis testing." If an alternative to the prevailing classical statistics was to be provided (and this was their claim), this would also have to include such a central aspect as the testing of scientific hypotheses.[46] They started by clarifying the terms "odds" and "likelihood ratios." Using the example of checking to see if a dice is "fair," the application of likelihood ratios in a Bayesian sense was then compared to the classic approach of Neyman/Pearson (see above). They paid particular attention to clarifying the problem whereby classical statistics favored a consideration of Type I and Type II errors on the basis of this test variable:

> The interesting point is made that a Bayesian hypothesis test can add extensive support to the null hypothesis whenever the likelihood ratio is large. The classical test can only reject hypotheses, and it is not clear just what sort of evidence classical statistics would regard as a strong confirmation of a null hypothesis.[47]

We would like to avoid going into the – mostly highly technical – details in this respect. Solutions have meanwhile been found for numerous individual problems and fundamental questions, such as the Bayesian interpretation of frequency theory-based points of view, purely empirical Bayesian approaches, or even a theory of Bayesian data analysis.

One important issue in this context is the assessment of significance tests and confidence intervals.[48] The use of significance tests in their frequency theory-based sense enjoys wide support from a number of Bayesians for use as a heuristic tool, while others reject this approach. If a priori information is lacking, the confidence intervals of classical statistics and the Bayesian probability intervals may be almost numerically identical. They should, however, be interpreted in totally different ways.[49] In the classic, frequentist interpretation, a confidence interval of 95% means that, with the indicated (identical) sample ranges n for $m \to \infty$ (where m is the number of samples), 95% of intervals cover the true, unknown, fixed parameter

[43]Ibid., p. 541. This is referred to as "stable estimation."

[44]DuMouchel (1992, p. 521) points out that this approach is closely related to the "reference priors" subsequently proposed by other Bayesians for use in situations where little a priori information is available, which are also acceptable to classical statisticians.

[45]Edwards/Lindman/Savage (1963 [1992], p. 546).

[46]Bayesian literature does not adopt a uniform position regarding the need for a test theory.

[47]DuMouchel (1992, p. 523). Cf. example no. 3 in appendix A3 and also example no. 2 in appendix A4.

[48]General reference is made to Hodges (1990) in this respect.

[49]The following according to Iversen (1984, p. 31).

and 5% do not. We do not know however (and can only hope) whether the specific interval concerned covers the parameter or not. A Bayesian analysis assumes, in contrast, that the unknown parameter has an (usually subjective) a priori distribution. There is still uncertainty after the data have been obtained, but less so than in the previous case. This uncertainty is still expressed in probabilities but with a wholly different interpretation: The parameter θ lies, with a probability of 95%, between the two values c_u and c_o. Such an interpretation is not possible in terms of classical statistical inference,[50] although misleading interpretations of this Bayesian epistemology can still be found to this day in classic literature on the subject.

The alternative definition of the concept of probability is fundamental, regardless of individual formulations. In order to highlight better the contrast with the classic approach, we should first turn to the classic concept of probability and its weaknesses.

W. Stegmüller counts eight objections, put forward in literature on the subject, to the frequency theory arising from von Mises' definition,[51] regarding at least the last of them as "deadly": He confuses practical certainty with logical necessity.[52] A particular weakness of this concept of probability was seen to lie in its rejection of individual probabilities. According to von Mises' definition, it was impossible, for example, to indicate the probability of a certain throw of a particular dice at a particular location.

K. R. Popper (1990), for example, one of the most vehement opponents of subjectivism, used this problem to develop his own concept of probability (mainly related to the problems of physics) which evolved over the years into a so-called propensity theory.

No agreement has been reached up to the present day (nor is such a clarification likely to be achieved in the near future) about the final definition of probability, as, for example, C. Howson established:

> It would be foolhardy to predict that philosophical probability has entered a final stable phase; surveys of the field tend to have useful lifetimes of a decade or so, at most two. It would also probably be incorrect to pretend that there is likely in the near future to be any settled consensus as to which interpretations of probability make viable and useful theories, and which are dead ends.[53]

Bayesian concepts of inference are however not limited to a subjective element that formalizes a priori probability but link it, by means of Bayes' theorem, with the "evidence of the data," which is in turn formalized in the likelihood function. The likelihood function already played an important role for Bernoulli, Laplace, and Gauss. Its importance as a central element of statistical inference was emphasized by

[50]Ibid: "This is the way many users of confidence intervals want to interpret a confidence interval, but in classical statistical inference such an interpretation is not possible."

[51]See above, p. 86f.

[52]Cf. Stegmüller (1973, p. 32ff, particularly p. 37).

[53]Howson (1995, p. 27).

A. Birnbaum in particular, who introduced the concept of the likelihood principle in this context.[54] The main difference between the likelihood principle and the frequency principle can be formulated as a question: Is it possible to obtain evidence about a parameter on the basis of a specific piece of data (i.e., a "sample")? Adherents of the frequency concept (particularly J. Neyman) emphasize that we can only assess the performance of a procedure if it is carried out repeatedly and measured on the basis of long-term averages.

However, if it is not possible to conduct experiments, and conclusions can only be drawn using existing, repeatable data that have not been scrutinized (e.g., as is the case in cliometrics), the relevance of such a concept must be seriously questioned. If repeatability is purely hypothetical, it should also be explicitly defined as a (subjective) conviction and not as an objective possibility. We therefore find it more reasonable, for such situations, to define probability as a degree of belief, which is then assigned to a parameter value. The evaluation and revision of this conviction with the evidence of existing, non-hypothetical data obtained by applying the likelihood function are also logically consistent in our opinion, especially as it does not depend on asymptotic generalizations. We would like to subscribe to the opinion of D. Lindley in this respect:

> The present position in statistical inference is historically interesting. The bulk of practitioners use well-established methods like least squares, analysis of variance, maximum likelihood and significance tests: all broadly within the Fisherian school and chosen for their proven usefulness rather than their logical coherence. If asked about their rigorous justification most of these people would refer to ideas of the NPW [Neyman-Pearson-Wald, T. R.] type; least-square estimates are best, linear unbiased; F-tests have high power and maximum likelihood values are asymptotically optimal. Yet these justifications are far from satisfactory: the only logically coherent system is the Bayesian one which disagrees with the NPW notions, largely because of their violation of the likelihood principle.[55]

Inference in Econometrics

Let us now turn to inference in econometrics. Two phases can be distinguished in economic statistics and econometrics: an initial phase, in which the description and exploration of economic series or processes predominated, and a second phase of inference and modeling.

The first phase can be characterized by its adoption of correlation concepts developed by Galton (1888) and Pearson. There was, however, a crucial difference: A body of theory did in fact exist in economics, but it was neither uniform nor

[54] See above, p. 99.

[55] Lindley (1991, p. 493).

sufficiently established to make it accessible for direct empirical application.[56] An explorative character was therefore dominant from the beginning in this respect. Phenomena such as "trade cycles" were not physical variables that only had to be measured, nor were they biological variables with distribution that could be determined with arbitrary precision and influencing factors that could be analyzed by experiment. On the contrary, the data were (1) passive in nature and not immediately suitable for reproducing, they had to be (2) precisely defined, and they were not (3) subject to universally stable distribution.

The use of the correlation calculation was theoretically based in the case of Galton. As the observed data came, for example, from a bivariate normal distribution, their relationship to each other could be expressed in a coefficient. But this theoretical reasoning was already abandoned by Yule upon its first application in the context of social science.[57] The functional relationships were considered linear for computational processing reasons, while the parameters were determined, on the same grounds, by means of the method of least squares. Yule's authority (he was one of the leading statisticians of his day) justified the application of biometric techniques, even though the theoretical justification for this approach was doubtful.

Two aspects are of particular significance in this context: Firstly, no in-depth statistical knowledge was needed in order to recognize that the structure of socio-economic phenomena was different to the structure used to determine the growth of plants or relationships between organism body sizes. Secondly, this was made all the more clear as attention turned to the analysis of data that represented *time series*.

The Time Dimension

The analysis of economic events in terms of their processuality did not find, in economic theory, any concrete statements regarding duration, form, or relationships of trade cycles to each other. The pioneers of empirical studies thus went their own ways, with H. L. Moore and W. S. Jevons seeking replacement in the field of astronomy. Not only astronomical phenomena, such as the periodically varying number of sunspots or the strictly periodic path of Venus (an 8-year cycle between the Sun and Earth), were used to provide explanations; the mechanics of astronomy, in the form of periodogram analysis, were also employed. A method such as this had the advantage of being able to make "hidden" periodicities visible. However, the initial euphoria created by the use of the periodogram analysis soon gave way to the sobering realization that the application lacked an important prerequisite: the stability of the object being examined. Trade cycles were not like the planets, with their constant movements of a duration that could be computed with fixed margins of

[56]Economic theories, from L. Walras to A. Marshall, started out from states of equilibrium, which were adapted, independently of historical context, by the same perpetual motives of human action. The economic laws contained in these theories were timeless.

[57]See above, p. 76 f.

error, but were instead phenomena whose length and intensity varied both with time and the intensity of their disturbance factors.

And even this was not enough, as economic data generally tended to be subject to trends. Their long-term development was therefore not distributed on the basis of stable averages. In these cases, there were no timeless states of equilibrium from which (at the most) transient deviations were possible. There was instead an irreversible development.

The solution to this problem did not however lie in using this irreversibility as an opportunity to adopt a fundamentally different view. Instead, two alternatives were taken up: one postulated, even for this long-term development, either a functional, measurement-error-conditioned context in the form of a polynomial or some other trend function (if the long-term curve had a reasonably smooth appearance). The method of least squares was used to determine this trend. This had already developed a life of its own, and its progress was barely stoppable. Either that or one could decide completely against a long-term development model and exclude it by observing the deviations from a moving average. In both cases however, the goal was not a comprehensive analysis of (historical) development, but rather an "exclusion" of whatever could not be incorporated into the scheme of identical timeless structures.[58]

It is to this extent obvious that a component-based concept dominated further research. Mutually independent explanatory factors therefore determined the long-, medium-, and short-term curves by which the trend component was found to be just as disruptive as its short-term "residual" counterpart. It was difficult in this context to respond to the question of correlation. The study of trends and cycles on one hand and of correlations on the other was not a separate epistemic interest but an interrelated factor. According to the statisticians, the trend first had to be excluded to allow the examination of correlations, while the goal of correlation analysis was to examine the conformity of medium-term (i.e., cyclic) curves.

On the other hand, one must however not overlook the fact that it was in this formulation phase that the issue of historical change in economic structures became highly problematic. If there was a long-term trend "component," why should the mutual links between economic variables not then also be made subject to long-term changes? The attempts by Hall (1925), Kuznets (1928a, b), Ezekiel (1928), or Frisch (1931) to extend existing concepts to include time-dependent models, or at least to point out the inadequacy of conventional formalizations, were therefore the obvious thing to do.

We can only speculate as to why this path was not pursued further. One possible explanation might be that the technical difficulties with regard to modeling were too great. However, as these papers were in any case barely implicated in statistical inference, another explanation seems plausible to us: the surprisingly great similarity between an economic index on the trade cycle and a series of computed random

[58]One of the few exceptions, who assigned independent significance to the trend, was S. Kuznets. See Kuznets (1930a, b) in particular.

variables, contained above all in a paper by Slutzky (1937) and presented shortly afterward to the English-speaking world by Kuznets (1929). Did this similarity mean that even trade cycles depended solely on random variables?

Research by Yule and Slutzky went on to form the conceptual basis of the modern theory of stochastic processes. Although both of them described different types of models – autoregressive processes in the case of Yule (1927) and so-called "moving average" processes in the case of Slutzky (1937) – their structures nevertheless had crucial factors in common. They regarded a time series as a realization of a stochastic differential equation. While Yule started with a trigonometric function that could be represented as a differential equation (albeit one in which the error term had a completely different effect to that of the functional form), Slutzky constructed various – at first glance rather arbitrary – sums of random variables. A deeper justification for the chosen type of model (e.g., regarding why a certain number of random variables was provided with different weightings and added up once or several times) was of less importance in this respect than the alarming fact that random variables could create cyclical phenomena.

It is highly surprising that there were apparently, in the case of this conceptualization, bigger problems regarding the acceptance of the idea of a random, yet legitimate, process than there were for cross-sectional regression analysis. Time series were therefore regarded either deterministically, in terms of their essential components (the component model), or as purely coincidental, with cycles which then had no significance. The key point was overlooked: It was not the random variables that were responsible for (pseudo-)cyclical character but the mechanism, i.e., the model.

This inner logic of these models remained hidden to Kuznets, just as it subsequently did to G. Tintner, J. Schumpeter, and John Maynard Keynes.[59] It is therefore not surprising that scientists with less of a mathematical background were no longer willing or able to follow the conceptual idea associated with such models.

R. Frisch (1933) on the other hand, an econometrician with physical background, had clearly recognized the inner logic of these models and had even included a corresponding economic justification of it in his famous article on propagation. In a dynamic model of an economy, certain parameter values not affected by disturbance factors could give rise to damped oscillations. The action of "shocks" could, on the other hand, produce the irregular cycles first referred to by Yule.[60]

[59]Even Tinbergen came to recognize that he "did not understand the role of the shocks as well as Frisch did" (Tinbergen in Magnus/Morgan (1987, p. 125)).

[60]The separation between the role of the mechanism and that of the shock was of great importance for the development of econometrics, even though Tinbergen regarded it critically in retrospect: "[. . .] I think that what interested economics most was not the shocks but the mechanism generating endogenous cycles, and it might very well be that we have overestimated the role of the mechanism. Maybe the shocks were really much more important. This problem was never solved, because the War came along and after the War we were not interested in business cycles anymore" (Tinbergen in Magnus/Morgan (1987, p. 125)).

With the reception of these models into economics, the ways divide. Kuznets' (1934) *Time Series* contribution to the *Encyclopedia of the Social Sciences* described only the component model, without any stochastic implications. No mention was made of models with variable parameters or of the fundamental significance of the models of Yule and Slutzky.[61] Papers by Schumpeter, and also by Burns and Mitchell (1946), took a similar line. Schumpeter did in fact write the opening article of the first issue of *Econometrica*, which was published in 1933, but played no further part in the development of econometrics.

It was of crucial importance to further development that the scientific orientation of econometrics was largely determined by individuals with an educational background in physics, such as Jan Tinbergen, Ragnar Frisch, Tjalling Koopmans, Charles Roos, or Harold T. Davis.[62] These thinkers possessed a different picture of economics to that of "traditional" empirical researchers. They brought a mechanistic, rigorously mathematical model of thinking to empirical research. One example of this development is an account by Koopmans of his career:

> Why did I leave physics at the end of 1933? In the depth of the worldwide economic depression, I felt that the physical sciences were far ahead of the social and economic sciences. What had held me back was the completely different, most verbal, and to me almost indigestible style of writing in the social sciences. Then I learned from a friend that there was a field called mathematical economics, and that Jan Tinbergen, a former student of Paul Ehrenfest, had left physics to devote himself to economics. Tinbergen received me cordially and guided me into the field in his own inimitable way. I moved to Amsterdam, which had a faculty of economics. The transition was not easy. I found that I benefited more from sitting in and listening to discussions of problems of economic policy than from reading the tomes. Also, because of my reading block, I chose problems that, by their nature, or because of the mathematical tools required, have similarity to physics.[63]

It was possible to have in this environment (1) modeling of the economic world in the form of differential equations and (2) a rigid stochastic process. It nevertheless appears strange, at first glance, that Koopmans should develop his approach using the theory of R. A. Fisher and did not see, as Frisch had, measurement errors, in physical analogy, as a justification for a stochastic approach but started out instead, in a biological analogy, from hypothetically infinite populations from which, with constant probabilities, the existing data would have stemmed. The basic stochastic concept was probably not of so much importance in this instance but rather the facts

[61] Cf. Kuznets (1934).

[62] See Epstein (1987, p. 75 note 39), Mirowski (1989, p. 234), and above all Boumans (1993). Even the statistician G. U. Yule, who was particularly involved in research in the field of time series analysis and its potential applications in economics, began his academic career in the study of electrical waves.

[63] Quoted from Mirowski (1991, p. 152). Frisch and Koopmans applied matrix calculus, which was being widely disseminated in physics in the mid-1920s, in the context of multiple regression analysis, to the field of econometrics, thereby making it more difficult for economists to comprehend the texts concerned. Cf. Mirowski (1989, p. 231).

that Fisher had developed a comprehensive statistical estimation theory and that he was regarded as a leading statistician.

Univariate time series analysis turned into a sideshow issue in this context, with thinking in terms of "complete" models coming to dominate instead.[64] These models did not however fully or consistently match, from the beginning, the theoretical economic models, although their consideration was the initial objective of econometrics. Tinbergen had already found himself forced into a series of compromises, as the existing economic theories of his day had not been specified to an extent that permitted direct empirical testing.

The uninhibited, iterative approach of Tinbergen infringed the rules of the stochastic concept of statistics that had just been adopted by Frisch and Koopmans. Some criticism of Keynes or Friedman was to this extent justified. The chosen way was nevertheless followed further and given a certain manifesto-like air by T. Haavelmo, a student of R. Frisch.

"Clarification": Trygve Haavelmo

Haavelmo's line of argument, which set the trend for further development, called – like Koopmans' – for a rigorously stochastic approach. Unlike Koopmans however, Haavelmo did not rely on Fisher's theory but on those of Neyman and Pearson. If we examine the foundations of this theory, its application to (macro)economic developments inevitably appears problematic.

We have seen that acceptance of the Neyman-Pearson approach brings with it a concept directed at rules of conduct. Even the application of Fisher's notion of hypothetically infinite populations, from which random samples are drawn, may appear strange. However, this is even more problematic for the Neyman-Pearson concept of "repeated sampling from the same population." When applying such a notion to macroeconomic time series, the question to ask is the following: "[. . .]how often is the question that an econometrician has to answer a decision problem in the context of repeated sampling?"[65]

Why did Haavelmo use precisely this approach as a basis?[66] One possible explanation could be that the rivalry of the early 1940s between the approaches of Fisher and Neyman/Pearson resulted in the latter emerging as the victor, thereby already representing a "paradigm" in the Kuhnian sense. There is also a personal reason: Haavelmo himself reported that he had for various months enjoyed the privilege of studying under the "world's famous statistician" J. Neyman. This may

[64]Research nevertheless still continued to take place in the "old" tradition, as econometrics began to develop. See, for example, Hotelling (1934), Schultz (1934), Greenstein (1935), and Regan (1936). Even the method of moving averages was still being recommended by Sasuly (1936) in this context.

[65]Keuzenkamp/Magnus (1995, p. 18).

[66]Heckman (1992, p. 881) also poses the question in this context, in criticism addressed to Morgan (1990): "Why was the Neyman-Pearson theory adopted as the paradigm of statistical inference in econometrics, and why were rival theories by Ronald Fisher and Harold Jeffreys less successful?".

have shown him, as someone who was then "young and naïve," "ways [. . .] to approach the problem of econometric methodology that were more promising than those that had previously resulted in so much difficulty and disappointment."[67]

Haavelmo certainly saw the problems that lay in a simple application of the Neyman-Pearson concept and therefore argued from an instrumentalist stance. His writings repeatedly contain remarks such as "it has been found fruitful" and similar. In addition, large parts of his explanations are based solely on "hopes":

> [. . .] we might hope to find elements of invariance in economic life, upon which to establish permanent laws [. . .]. Our hope in economic theory and research is that it may be possible to establish constant and relatively simple relations [. . .]. Our hope for simple laws in economics rests upon the assumption that we may proceed as if such natural limitations of the number of relevant factors exist.[68]

Is it justified, with a stance such as this, in starting out from objective inference? Even if we rule out the problematic underpinnings, there is a series of questions that the Neyman-Pearson approach fails to answer. As Heckman correctly notes, Haavelmo did not, for example, take into account the important aspect of model structure and selection:

> These claims have never been rigorously established, even for analyses conducted on large samples. There is no 'correct' way to pick an empirical model and the problems of induction, inference, and model selection are very much open. [. . .] The Neyman-Pearson theory espoused by Haavelmo and the Cowles group takes a narrow view of science. By its rules, hypotheses are constructed in advance of knowledge of the data and the role of empirical work is to test the hypotheses. This rigid separation of model construction and model verification was a cornerstone of classical statistics circa 1944. Even then, influential scholars, primarily Bayesians such as Harold Jeffreys quarreled with this view of empirical science. Since that time, the monopoly of classical statistics has broken.[69]

Haavelmo's application of the Neyman-Pearson paradigm nevertheless formed the basis in econometric research for several decades. Even Koopmans stopped citing Fisher and defended Haavelmo's approach with respect to R. Vining. The physical world view was thus cemented into place. Koopmans drew comparisons between the "complete" systems of structural equations and the explanatory power of Newton's theory of gravitation, while J. Marshak (1950), Chairman of the Cowles Commission, went so far as to regard the issue explicitly as "social engineering." But does this not ominously remind us of the "social physics" – vehemently rejected in its day – of Quetelet?

[67] Haavelmo (1994, p. 75).

[68] Haavelmo (1944, pp. 13, 22f, 24).

[69] Heckman (1992, p. 882). He gives reasons for Morgan's overestimation of Haavelmo's approach – rightly in our opinion – with the view, which can be traced back to the influence of Hendry, that these problems are generally solvable in the context of the Neyman-Pearson approach. This overestimation is also picked up by Malinvaud (1991, p. 635) and Zellner (1992, p. 220).

Alternatives

There have been increasing attempts, ever since the 1970s, to seek out alternative ways. C. Sims[70] proposed vector autoregressive time series models as a counter to traditional systems based on simultaneous equations. These models initially provided nothing more than a description of the delayed correlation structure present in existing time series. One could, in principle, regard vector autoregressive models as the ideal form for cliometrics. They are, however, associated with the same problem as univariate ARIMA models,[71] in that the "right" model must first be found on the basis of the data, which infringes in turn the assumptions of classical inference. It is moreover not possible, given the high degree of complexity of these models, to use the tools developed by Box and Jenkins for use in univariate time series analysis. Sims therefore proposed restricting the high number of parameters that result from such models, thereby ultimately advocating a Bayesian approach.

Bayesian approaches, which marked the beginnings of structural equation models in the econometrics of the 1960s, were still subject, in technical terms, to greater difficulties than classical statistical inference. These technical difficulties should not, however, obscure the fact that the Bayesian standpoint is considered by its representatives to be, from a conceptual point of view, a single approach:

> That there is a unified and operational approach to problems of inference in econometrics and other areas of science is a fundamental point that should be appreciated. Whether we analyze, for example, time series, regression, or 'simultaneous equation' models, the approach and principles will be the same. This stands in contrast to other approaches to inference that involve special techniques and principles for different problems.[72]

E. Leamer developed the most consistent Bayesian econometric methodology.[73] The main criticism of Leamer appears to us to be the part concerning modeling problems. Leamer rightly pointed out that the classical theory, in which the model is regarded as a given, required an almost "Orwellian" approach to econometrics:

> In such a fanciful world, personal uncertainties and public disagreements concerning how to interpret data would be completely resolved in advance. New data sets would not be distributed to humans at all, but instead would be delivered with elaborate security measures to a centralized warehouse where preprogrammed computers would pore over the numbers and pass the conclusions to the public. Once analyzed, the data would be entirely destroyed, to prevent the urge to try something else from becoming an unwanted reality.[74]

[70] See, for example, Sims (1980).

[71] They were also subject to the same statistical limitations, such as stationarity and linearity.

[72] Zellner (1971, p. 11).

[73] See references in Rahlf (1998).

[74] Leamer (1994, p. ix).

The nonexperimental nature of econometrics prohibits such a notion. Data relating to such factors as the development of a country's gross national product are available only once but are evaluated repeatedly. If there is uncertainty regarding the model and – with respect to the selection of relevant variables – (1) the data are not neutral and (2) the personal conviction of the scientist plays a role (e.g., selection of the determinants of criminality by conservative or liberal researchers, selection of the determinants of inflation by monetarists or Keynesians), then a Bayesian point of view is, in our opinion, the only one that can be justified. The indication of the effect of different assumptions and selected variables, or "sensitivity analysis," appears to offer a promising approach in this respect, although its future reliability would have to be underpinned by a larger number of applications.

D. Hendry (2001) has developed a third methodology. Hendry, unlike Leamer, is convinced that a model structure based on the intensive analysis of a data set can be justified by the methods of classical inference. One revealing example of his approach comes from the reanalysis of a selected model based on comprehensive research carried out by M. Friedman and A. Schwartz of monetary trends in Britain and the United States, although the individual steps of the modeling process involved remain partly obscure. The possibility of validation, using classical "testing" based on the theory of Neyman and Pearson, has therefore been questioned in literature on the subject.[75]

Milton Friedman, by his own account, put no trust in formal statistical criteria. He had already rightly pointed out, in his criticism of Tinbergen's consideration of economic theories, that the traditional testing of significance or of hypotheses becomes less meaningful when it is applied after the analysis of the same data. His own t-tests are therefore also more likely to be understood as pragmatic.

If we consider these methodologies and approaches as a whole, a natural science world view dominated econometric research. Most approaches are based above all on constant, time-invariant parameters. Although the consideration of parameter constancy is part of Hendry's testing batteries, seldom alternatives – other than dummy variables – are modeled. Friedman and Schwartz (1991) do in fact point out the significance of analyzing historically uniform periods, but they subject these periods, in turn, to rigid constraints. Complexity is as a rule reduced to a parameter matrix that reflects the time-invariant structure, regardless of whether it concerns short- or long-term relationships.

Inference for Cliometrics

The question regarding the importance of empirical research to economic history and economics was again picked up in 1949, by A. P. Usher. Usher offered numerous philosophical, psychological, and scientific approaches that should justify a modern

[75]Keuzenkamp (1995, p. 243) therefore uses, for Hendry's approach, the more apposite term "diagnostic *checks*" rather than "diagnostic *tests*."

take on empiricism while highlighting its relevance to economic history. However, his references regarding approaches to philosophical probability theory stand in isolation.[76] Seen as a whole, the discipline of economic history in the first half of the twentieth century was, even in the United States, geared more toward a qualitative approach, with a tendency to reject the quantitative.[77]

The Bayesian Origins of Cliometric Inference

What is the current position regarding the concept of inference in cliometrics? If we define cliometrics generally by (1) the application of explicitly theory-driven, neoclassically oriented economic history research along with (2) the intensive use of mass data and of formal methods for verifying the theories based on those data, the question immediately arises of what difference there is, if any, with respect to the intrinsic concept of econometrics. The headword entry for "cliometrics" in the *New Palgrave* defines the approach as an "amalgam of methods,[78] born of the marriage contracted between historical problems and advanced statistical analysis, with economic theory as bridesmaid and the computer as best man,"[79] while the *American Heritage Dictionary* lists it as "the study of history using economic models and advanced mathematical methods of data processing and analysis."[80]

If the predominant characteristic is therefore the use of certain methods,[81] it is consequently surprising that the criticism that cliometrics has attracted on the part of "traditional" economic history has not established methodological problems, in the strict sense of the term, as a subject for discussion. Discussions were centered on the question of whether the application of theoretical models and their verification was, if at all, the cognitive goal of economic history with respect to a specific time and place and whether historical data fulfilled the conditions for applying elaborate statistical methods. The methods themselves played no further role however.

One can say that cliometrics has followed, in terms of methodology, the "paradigm" of econometrics, thereby taking into account the problems described by this field.[82] If we start by assuming, as E. Heckscher (1939) did, that the purpose of

[76]Cf. Usher (1949, p. 148 and p. 155, note 29).

[77]Fogel (1995, S. 49): "The leading history journals, even in economic history, initially refused to accept articles with complex tables and even after such articles began to be accepted, equations were absolutely forbidden."

[78]Floud (1991, p. 452.

[79]Fogel/Elton (1983, S. 2), quoted by Floud (1991, p. 452.

[80]See above, p. 5.

[81]See also Fogel (1995, p. 52) on this subject: "By the early 1980s *cliometric methods* were so firmly established in certain fields of history that no scholar in these fields could afford to neglect them" (our italics).

[82]This is supported not least by the fact that cliometrics did emerge as an independent school of thought because an application for admission by a group of the founding fathers of cliometrics had been rejected by the *Econometric Society*. Cf. Hughes (1965).

economic history is not fundamentally different to that of economics (or econometrics), it becomes plain that the econometric tools available to it, which were already well developed and firmly established by the early 1960s, were accepted uncritically, because econometrics gave, at that time, the most complete impression of its entire history of development.

It is therefore surprising, against the background of this development, that the papers by A. Conrad and J. Meyer (1957, 1958), which are commonly regarded as the "starting pistol" of cliometrics, should go off in a completely different direction. The two economists presented, at a 1957 conference held jointly by the Economic History Association and the National Bureau of Economic Research, a paper entitled *The Economics of Slavery in the Antebellum South*, in which they expounded the thesis – based on statistical methods, data compiled from secondary literature, and a theoretical economic model – that the purchase of a slave in the period before the Civil War represented a profitable investment for a slave owner from the Southern United States. Their work, which was published the following year, raised a storm of protest – and not just because of their "econometric" approach.[83] Our intention here is not to follow this discussion however[84] but rather to examine their methodology. This was also pointed out by the authors in 1957, in a programmatic article on the relationship between economic theory, statistical inference, and economic history, which followed a surprisingly Bayesian line of argument.[85]

Conrad and Meyer set out here to emphasize the significance of a concept of causal orders, which should underpin every historical narrative. The denial of the possibility of causal explanations in history, which was put forward by a number of philosophers, is based mainly on the view that historical events are unique, complex, and unquantifiable.[86] They rightly pointed out that econometric modeling predetermines a causal order, which is only valid for the variables contained in the model[87]: "Causal order is an operational term, which does not require the involvement of any invisible forces or internal needs."[88] The claim that causal explanations are connected to the basic repeatability of an experiment, although historical events are unique, is likewise incorrect. Firstly, experiments are also essentially first-time events and, secondly, a science such as astronomy would no longer be capable of making causal statements, as it would then be dealing with non-repetitive phenomena.[89] This is where Bayesian reasoning came into play, as it is not based on the

[83]Conrad/Meyer (1958). The authors were at the time *assistant* professors of economics at Harvard. The expressions "starting gun" and "watershed" are therefore justified, since econometric methods were for the first time being applied to historical phenomena without any reference to the present.

[84]Cf. Conrad/Meyer (1964).

[85]Conrad/Meyer (1957).

[86]Cf. Conrad/Meyer (1957, p. 527).

[87]They refer here to an example given by Simon (1957) regarding the differing possible influences of the variables of weather, wheat harvest yield, and wheat price.

[88]Conrad/Meyer (1957, p. 147.

[89]They seek support, in this context, in the line of argument of H. Jeffreys.

repeatability of the events but is concerned rather with a subjective grasp of statements of probability:

> Explicitly, the formal tests attach an actual numerical probability to the correctness of the hypothesis in the light of the observed results. This introduces the question of relative plausibility into the empirical procedure and consequently helps the investigator to scale the degree of belief, an intrinsically ordinal concept at the very least, that should be placed in the hypothesis. There are, in sum, substantial advantages as well as disadvantages to the introduction of more formal procedures in the evaluation of historical hypotheses. The question therefore arises: Is there a satisfactory compromise that embodies maximum advantage with minimum disadvantage? Ideally, the best procedure would appear to be one in which the formal tests were adapted or altered to take account of a maximum of a priori information. This leads, admittedly, to an essentially Bayesian approach to statistical inference.[90]

They did indeed see it as problematic that the Bayesian approach was sinking into a "morass of subjectivism," in the immediate absence of a priori notions and probabilities. They were, however, confident that this could form a basis for creating guidelines and simplifying the communication of scientific results.

The discussion following the presentation of the paper, in which the economists present expressed their opposition to the application to historical data of econometric models and statistical tests, included (in the same manner as later papers on cliometrics) little evidence of this key difference with respect to the prevailing econometric approach.[91] Subsequent development tended rather to follow the path marked out by econometrics, albeit without influencing econometrics itself.

The cliometric (r)evolution, whose development had been swiftly gathering pace since the 1960s, then took over the methods of econometrics, along with its associated concepts. A way such as this was, for logical reasons, just as faintly compelling as the adoption by econometrics of "classic" statistical inference. The papers by Conrad and Meyer, which marked the beginnings of cliometrics, followed a Bayesian argument, although this was subsequently taken into account neither by cliometrics itself nor its critics. The course of econometrics was actually set by physicists in their role as "social engineers," harking back implicitly (or even explicitly) to Newton. We would like to conclude our overview by citing some criticism that is very revealing for statistical inference in the field of cliometrics: the critique of the mathematician Rudolf Kalman.

[90] Conrad/Meyer (1957, p. 544). Specific examples can be found in Conrad/Meyer (1964).

[91] Bayesian approaches were not to find fertile ground in the field of econometrics until several years later. It must however be emphasized that the line of argument maintained by Conrad and Meyer contained various terms and concepts (they speak of objective tests and significant differences, before returning to probabilities of hypotheses and a "morass of subjectivism") that cannot always be clearly differentiated from each other.

Fundamental Criticism: Rudolf Kalman

Rudolf Kalman took on a study, in the early 1980s, of the problem of model structure and inference in the field of econometrics and expressed fundamental criticism, in this context, from a system theory point of view.[92] In his opinion, econometrics mainly went along the following two paths:

1. Economic laws and relationships have been formulated as dynamic equations in terms of Newton's laws.
2. The coefficients of these equations have been determined quantitatively by the extraction, from real data, of statistically relevant information.

He establishes, with this development in mind, that the progress in knowledge subsequently achieved is, even in comparison to the 250 years that have elapsed since Newton, disappointingly low. He expounds the thesis (which requires, in his opinion, no discussion in terms of "hard" science):

> [. . .] that economics is not at all like physics and therefore that it is not accessible by a methodology that was successful for physics. Far from being governed by absolute, universal, and immutable laws, economic knowledge, unlike physical science, is strongly system (context) dependent; when economic insights are taken out of temporal, political, social, or geographical context, they become trivial statements with little information content. [. . .] Since economic 'laws' do not process the attributes of physical laws, writing down equations, in the style of physics, to translate economic statements into mathematics is not a productive enterprise. [. . .] System theory provides a simple but hard suggestion: Do not write equations expressing assumed relationships; deduce your equations from real data. [. . .] To put it differently, there will never be a Newton in economics; the path to be followed must be different.[93]

His opinion on the second step, the statistical determination of unique parameters, is even more negative. This only makes sense in his view if there are concrete, explicitly measurable parameters, as is the case, for example, with resistors in Ohm's law:

> Economists have often dreamed of imitating the simple situation characterized by Ohm's law just by hoping for the best, for example, by assuming that such a law (the Phillips curve) exists between inflation and unemployment. But unemployment and inflation, in any quantitative sense, are fuzzy and politically biased attempts to replace complex situations by (meaningless) numbers; consequently any hope that two such concepts can be tied to one another by a single coefficient is barbarously uninformed wishful thinking.[94]

[92] We rely mainly on Kalman (1982a, b) in this respect. We are therefore not concerned with the *application* of the so-called Kalman filter to econometrics.

[93] Kalman (1982a, p. 19f). Original author's italics.

[94] Ibid., p. 20.

Unique relationships such as these exist in astronomy, for example, where their parameters have a direct significance that is independent of any system, such as the determining of the position of an object as a function of moment and angle. It is not surprising, against this background, that Kalman is especially critical of Haavelmo's approach: "The aspiration of Haavelmo to give a solid foundation to econometrics by dogmatic application of probability theory has not been fulfilled (in the writer's opinion), no doubt because probability theory has nothing to say about the underlying system-theoretic problems."[95] He calls instead for a rigorous application of system theory. System theory does not set out from a directly measurable relationship between input and output: "instead of determining a single parameter, such as a resistance, system theory is concerned with the much more general question of determining a system."[96] Parameters contained in systems have, according to Kalman, a completely different significance to that hitherto assumed by econometricians; they are therefore to be defined only *locally*. It is by no means self-evident, for Kalman, that the cognitive goal of statistical analysis should lie in the obtaining of constant figures, such as with the application of maximum likelihood estimates or the method of least squares: "[. . .]common sense should tell us, that such a miracle is possible only if additional assumptions (*deus ex machina*) are imposed on the data which somehow succeed in neutralizing the intrinsic uncertainty."[97] The method of least squares is thus so popular in these terms because it delivers a clear ("unique") response. However, the assumptions associated with such an approach cannot normally be justified.

The common approach of using data that show variance to determine a specific value that reveals maximum likelihood or minimizes deviations (thereby making it preferable to all others) is, for him, "fundamentally wrong and extremely harmful to scientific progress."[98] Such an approach implies the following suppositions (or "prejudices"):

1. The data have been generated using a probabilistic mechanism.
2. This probabilistic mechanism is very simple; it is constant in terms of time, and a distribution function explains everything.
3. There is a "true" value, which can be regarded as the "particularly striking feature" of the hypothetical distribution function, such as the expected value, median, or modal value.
4. A single figure constitutes the response of a deductive process based on self-evident postulates.

[95]This sentence, which was supposed to appear in Kalman (1982c), was deleted at editorial request and included instead in Kalman (1982b, p. 194).

[96]Kalman (1982a, p. 23). Linearity and finiteness might be reasonable assumptions for such a system.

[97]Kalman (1982b, p. 162). Original author's italics.

[98]Ibid., p. 171.

The assumption of exact conformity to natural laws in probabilistic phenomena analogous to Newtonian physics, which is what such an approach is supposed to aspire to, has nevertheless long since proved to be an illusion. Apart from "mathematical artifacts," such as the law of large numbers, there have not been any universal laws of random phenomena – even in physics – but rather ones that depend on the very system that surrounds them.[99] A view such as this has profound implications:

> The implications of this situation for econometric strategy are devastating. Since the problem is to identify a system and since systems cannot be described in general by globally definable parameters, the whole idea of a parameter loses its (uncritically assumed) significance. [. . .] The Jugendtraum of econometrics, determining economically meaningful parameters from real data via dynamical equations supplied from economic theory, turns out to have been a delusion.[100]

This criticism of Kalman has not as yet – as far as we can see – had any impact on econometrics. Even if one does not wish to follow the path to its ultimate consequences, the fundamental reasonableness of applying physics-based approaches to economic developments should still be examined. It is surely a highly promising basis for inference statements in the field of cliometrics.

References

Barnard G (1947) The meaning of a significance level. Biometrika 34:179–182
Barnard G (1949) Statistical inference (with discussion). J R Stat Soc B 11:115–149
Barnard G (1988) R. A. Fisher – a true Bayesian? Int Stat Rev 55:183–189
Berger J, Wolpert R (1988) The likelihood principle. , vol 6, 2nd edn, Lecture notes – monograph series. Institute of Mathematical Statistics, Hayward
Birnbaum A (1962) On the foundations of statistical inference (with discussion). J Am Stat Assoc 57:269–306
Birnbaum A (1968) Likelihood. In: Sills D (ed) International encyclopedia of the social sciences. Macmillan, New York, pp 299–301
Birnbaum A (1977) The Neyman-Pearson theory as decision theory and as inference theory: with a criticism of the Lindley-Savage argument for Bayesian theory. Synthese 36:19–49
Bjornstad J (1992) Introduction to Birnbaum (1962) on the foundations of statistical inference. In: Kotz S, Johnson N (eds) Breakthroughs in statistics. Bd. I. Foundations and basic theory. Springer, New York, pp 461–477
Borovcnik M (1992) Stochastik im Wechselspiel von Intuitionen und Mathematik. Spektrum Akademischer Verlag, Mannheim
Boumans M (1993) Paul Ehrenfest and Jan Tinbergen: a case of limited physics transfer. In: de Marchi N (ed) Non-natural social sciences: reflecting on the enterprise of 'More heat than light',

[99]Cf. ibid., p. 172.

[100]Kalman (1982a, pp. 26, 27). He describes the calculation of a constant parameter (e.g., in the context of the Phillips curve) as a "conceptual absurdity" (ibid.). Kalman consequently also rejects any causal interpretation. Cf. Kalman (1982b, p. 177), for example.

vol 25, Supplement to history of political economy. Duke University Press, Durham/London, pp 131–156
Burns AF, Mitchell WC (1946) Measuring business cycles. National Bureau of Economic Research, New York
Conrad A, Meyer J (1957) Economic theory, statistical inference, and economic history. J Econ Hist 17:524–544
Conrad A, Meyer J (1958) The economics of slavery in the antebellum south. J Polit Econ 66:95–130
Conrad A, Meyer J (eds) (1964) The economics of slavery. Studies in econometric history. Aldine, Chicago
Dale A (1991) A history of inverse probability. From Thomas Bayes to Karl Pearson, vol 16, Studies in the history of mathematics and physical sciences. Springer, New York
de Finetti B (1937) La prévision: Ses lois logiques, ses sources subjectives. Ann V Institut Henri Poincaré 1:1–68
de Finetti B (1981) Wahrscheinlichkeitstheorie. Einführende Synthese mit kritischem Anhang. Oldenbourg, Wien/München
DuMouchel W (1992) Introduction to Edwards, Lindman, Savage (1963) Bayesian statistical inference for psychological research. In: Kotz S, Johnson N (eds) Breakthroughs in statistics. Bd. 1. Foundations and basic theory. Springer, New York, pp 519–530
Edgeworth FY (1884) The philosophy of chance. Mind 9(34):223–235
Edwards W, Lindman H, Savage L (1963) Bayesian statistical inference for psychological research. Psychol Rev 70:193–242 [Reprinted in: Kotz S, Johnson N (1992) (eds) Breakthroughs in statistics. Bd. 1. Foundations and basic theory, New York]
Epstein RJ (1987) A history of econometrics. North Holland, Amsterdam
Ezekiel M (1928) Statistical analysis and the law' of price. Q J Econ 42:199–227
Fisher R (1922 [1992]) On the mathematical foundations of theoretical statistics. Philos Trans R Soc Lond A 222:309–368 [Reprinted in: Kotz S, Johnson N (1992) (eds) Breakthroughs in statistics. Bd. 1. Foundations and basic theory, New York]
Fisher R (1955) Statistical methods and scientific induction. J R Stat Soc B 17:69–78
Fisher R (1956) Statistische Methoden für die Wissenschaft, 12 Aufl. Oliver and Boyd, Edinburg
Fisher R (1959) Statistical methods and scientific inference, 2nd edn. Oliver and Boyd, London
Floud R (1991) Cliometrics. In: Eatwell J, Milgate M, Newman P (eds) The new Palgrave. A dictionary of economics. Bd. 1, 2nd edn. Macmillan, London/New York/Tokyo, pp 452–454
Fogel R (1995) History with numbers: the American experience. In: Etemad B, Batou J, David T (eds) Pour une histoire économique et sociale internationale. Ed. Passé Présent. Genf, Genéve, pp 47–56
Fogel R, Elton G (1983) Which road to the past? Two views of history. Yale University Press, New Haven/London
Friedman M, Schwartz A (1991) Alternatives approaches to analyzing economic data. Am Econ Rev 81(1):39–49
Frisch R (1931) A method of decomposing an empirical series into its cyclical and progressive components. J Am Stat Assoc (Suppl) 26:73–78
Frisch R (1933) Propagation problems and impulse problems in dynamic economics. In: Essays in honour of Gustav Cassel. Allen & Unwin, London
Galton F (1888) Co-relations and their measurement. Proc R Soc Lond Ser 45:135–145
Geisser S (1992) Introduction to Fisher (1922) on the mathematical foundations of theoretical statistics. In: Kotz S, Johnson N (eds) Breakthroughs in statistics. Bd. 1. Foundations and basic theory. Springer, New York, pp 1–10
Gigerenzer G, Swijtink T, Porter T, Daston L, Beatty J, Krüger L (1989) The Empire of chance: how probability changed science and everyday life. Cambridge University Press, Cambridge/New York
Gosset WS (1908) The probable error of a mean. Biometrika 6:1–25

Graunt J (1662 [1939]) Natural and political observations made upon the bills of mortality. Edited with an introduction by Willcox WF John Hopkins University Press, Baltimore
Greenstein B (1935) Periodogram analysis with special application to business failure in the U.S. 1867–1932. Econometrica 3:170–198
Haavelmo T (1944) The probability approach in econometrics. Econometrica 12(Suppl):1–115
Haavelmo T (1994) Ökonometrie und Wohlfahrtsstaat. Nobel-Lesung vom 7. Dezember 1989. In: Grüske K-D (ed) Die Nobelpreisträger der ökonomischen Wissenschaft. Bd. 3. 1989–1993. Wirtschaft und Finanzen, Düsseldorf, pp 71–80
Hall L (1925) A moving secular trend and moving integration. J Am Stat Assoc 20:13–24
Halley E (1693) An estimate of the degrees of mortality of mankind drawn from curious tables of the births and funerals at the city of Breslau; with an attempt to ascertain the price of annuities upon lives. Philos Trans R Soc 17:596–610. Electronic reprint: http://www.pierre-marteau.com/editions/1693-mortality.html
Heckman J (1992) Haavelmo and the birth of modern econometrics: a review of the history of econometric ideas by Mary Morgan. J Econ Lit 30:876–886
Heckscher E (1939) Quantitative measurement in economic history. Q J Econ 53:167–193
Hendry DF (2001) Econometrics: alchemy or science? 2nd edn. Oxford University Press, Oxford
Hodges J (1990) Can/may Bayesians do pure tests of significance? In: Geisser S, Hodges J, Press S, Zellner A (eds) Bayesian and likelihood methods in statistics and econometrics. Essays in honor of George A. Barnard, vol 7, Studies in Bayesian econometrics and statistics. North Holland Publishing, New York, pp 75–90
Hotelling H (1934) Analysis and correlation of time series. Econometrica 2:211
Howson C (1995) Theories of probability. Br J Philos Sci 46:1–32
Hughes J (1965) A note in defense of Clio. Explor Entrep Hist 3:154
Iversen G (1984) Bayesian statistical inference. Sage, Newbury Park
Jeffreys H (1939) Theory of probability. The Clarendon Press, London/New York
Johnstone D (1986) Tests of significance in theory and practice (with discussion). Statistician 35:491–504
Kalman R (1982a) Dynamic econometric models: a system-theoretic critique. In: Szegö G (ed) New quantitative techniques for economic analysis. Academic, New York, pp 19–28
Kalman R (1982b) Identification from real data. In: Hazewinkel M, Rinnooy Kan A (eds) Current developments in the interface: economics, econometrics and mathematics. Reidel, Dordrecht, pp 161–196
Kalman R (1982c) Identifiability and problems of model selection in econometrics. In: Hildenbrand W (ed) Advances in econometrics. Cambridge University Press, Cambridge
Kempthorne O (1971) Comment on 'Applications of statistical inference to physics'. In: Godambe V, Sprott D (eds) Foundations of statistical inference. Holt, Rinehart and Winston of Canada, Toronto, pp 286–287
Keuzenkamp H (1995) The econometrics of the Holy Grail – a review of Econometrics: alchemy or science? Essays in econometric methodology. J Econ Surv 9:233–248
Keuzenkamp H, Magnus J (1995) On tests and significance in econometrics. J Econ 67:5–24
Keynes J (1921) A treatise on probability. Macmillan, London
Koopmans T (1941) The logic of econometric business-cycle research. J Polit Econ 49:157–181
Kuznets S (1928a) On moving correlation of time sequences. J Am Stat Assoc 23:121–136
Kuznets S (1928b) On the analysis of time series. J Am Stat Assoc 23:398–410
Kuznets S (1929) Random events and cyclical oscillations. J Am Stat Assoc 24:258–275
Kuznets S (1930a) Secular movements in production and prices. Houghton Mifflin, Boston/New York
Kuznets S (1930b) Wesen und Bedeutung des Trends. Zur Theorie der säkularen Bewegung, Veröffentlichungen der Frankfurter Gesellschaft für Konjunkturforschung. Schroeder, Bonn
Kuznets S (1934) Time series. In: Seligman E, Johnson A (eds) Encyclopedia of the social sciences. Bd. 13. Macmillan, New York, pp 629–636

Kyburg H (1985) Logic of statistical reasoning. In: Kotz S, Johnson N (eds) Encyclopedia of statistical sciences. Bd. 5. Wiley, New York, pp 117–122
Kyburg H, Smokler H (eds) (1964) Studies in subjective probability. Wiley, New York
Laplace P-S (1812) Théorie analytique des probabilités. Courcier, Paris. https://archive.org/details/thorieanalytiqu01laplgoog
Leamer EE (1994) Introduction. In: Leamer EE (ed) Sturdy econometrics. Elgar, Aldershot, pp ix–xvi
Lehmann E (1992) Introduction to Neyman and Pearson (1933) on the problem of the most efficient tests of statistical hypotheses. In: Kotz S, Johnson N (eds) Breakthroughs in statistics. Bd. 1. Foundations and basic theory. Springer, New York, pp 67–72
Lehmann EL (1993) The Fisher, Neyman-Pearson theories of testing hypotheses: one theory or two. J Am Stat Assoc 88:1242–1249
Lindley D (1991) Statistical inference. In: Eatwell J, Milgate M, Newman P (eds) The new Palgrave. A dictionary of economics, vol 4, 2 Aufl. Macmillan, London/New York/Tokyo, pp 490–493
Malinvaud E (1991) Review of Morgan, Morgan M (1990) the history of econometric ideas. Econ J 101:634–636
Magnus J, Morgan M (1987) The ET interview: Professor J. Tinbergen. Econ Theory 3:117–142
Marshak J (1950) Statistical inference in economics. In: Koopmans T (ed) Statistical inference in dynamic economic models. Wiley, New York
Menges G (1972) Grundriß der Statistik. 1. Theorie, 2nd edn. Westdeutscher Verlag, Opladen
Mirowski P (1989) The probabilistic counter revolution, or how stochastic concepts came to neoclassical economic theory. Oxf Econ Pap 41:217–235
Mirowski P (1991) The when, the how and the why of mathematical expression in the history of economic analysis. J Econ Perspect 5:145–157
Morgan M (1990) The history of econometric ideas. Cambridge University Press, Cambridge
Neyman J (1937) Outline of a theory of statistical estimation based on the classical theory of probability. Philos Trans R Soc Lond Ser A Math Phys Sci 236(767):333–380
Neyman J, Pearson ES (1928a) On the use and interpretation of certain test criteria for purposes of statistical inference. Part I. Biometrika 20A:175–240
Neyman J, Pearson ES (1928b) On the use and interpretation of certain test criteria for purposes of statistical inference. Part II. Biometrika 20A:263–294
Neyman J, Pearson ES (1933) On the problem of the most efficient tests of statistical hypotheses. Philos Trans R Soc Lond Ser A, containing papers of a mathematical or physical character 231:289–337 [Reprinted in: Kotz S, Johnson N (1992) (eds) Breakthroughs in statistics. Bd. 1. Foundations and basic theory, New York]
Pearson K (1894) Contributions to the mathematical theory of evolution. Philos Trans R Soc Lond 85:71–110
Pearson K (1895) Contributions to the mathematical theory of evolution. II. Skew variation in homogeneous material. Philos Trans R Soc Lond 186:343–414
Pearson K (1898) Mathematical contributions to the theory of evolution: on the law of ancestral heredity. Proc R Soc Lond 62:386–412
Pearson K (1920) The fundamental problem of practical statistics. Biometrika 13(1):1–16
Pearson ES (1967) Some reflections on continuity in the development of mathematical statistics, 1885–1920. Biometrica 52:3–18
Popper K (1990) A world of propensities. Thoemmes, Bristol
Pratt J (1971) Comment on: 'probability, statistics and knowledge business' by O. Kempthorne. In: Godambe V, Sprott D (eds) Foundations of statistical inference. Holt, Rinehart and Winston, Toronto
Rahlf T (1998) Deskription und Inferenz Methodologische Konzepte in der Statistik und Ökonometrie, vol 9, Historical social research supplement. Zentrum für Historische Sozialforschung, Köln

Ramsey F (1931a) Truth and probability (1926). In: Braithwaite R (ed) The foundations of mathematics and other logical essays by Frank Plumpton Ramsey. International Library of Psychology, Philosophy and Scientific Method, London [Reprinted in Kyburg, Smokler (1964)]
Ramsey F (1931b) Further considerations (1928). In: Braithwaite R (ed) The foundations of mathematics and other logical essays by Frank Plumpton Ramsey. International Library of Psychology, Philosophy and Scientific Method, London, pp 199–211
Regan F (1936) The admissibility of time series. Econometrica 4:189
Robbins H (1955) An empirical Bayes approach to statistics. In: Neyman J (ed) Proceedings of the 3rd Berkeley symposium on mathematical and statistical probability, University of California. Statistical Laboratory: University of California Press, vol 1, pp 157–163 [Reprinted in Kotz/Johnson (1992)]
Sasuly M (1936) A method of smoothing economic time series by moving averages. Econometrica 4:206
Savage L (1954) The foundations of statistics. Wiley, New York
Savage L (1976) On rereading R. A. Fisher (with discussion). Ann Stat 4:441–500
Schultz H (1934) Discussion of the question 'Is the theory of harmonic oscillations useful in the study of business cycles?'. Econometrica 2:189
Sims C (1980) Macroeconomics and reality. Econometrica 48:1–48
Simon H (1957) Models of man. Wiley, New York
Slutzky E (1937) The summation of random causes as the source of cyclic processes. Econometrica 5:105–146 [originally published in Russian 1927]
Stegmüller W (1973) Personelle und Statistische Wahrscheinlichkeit. Erster Halbband: Personelle Wahrscheinlichkeit und Rationale Entscheidung. Zweiter Halbband. Statistisches Schließen, Statistische Begründung, Statistische Analyse. Probleme und Resultate der Wissenschaftstheorie und Analytischen Philosophie IV. Springer, Berlin/Heidelberg/New York
Stigler S (1986) The history of statistics: the measurement of uncertainty before 1900. Belknap Press of Harvard University Press, Cambridge, MA
Usher A (1949) The significance of modern empiricism for history and economics. J Econ Hist 9:131–155
von Mises R (1951) Wahrscheinlichkeit, Statistik und Wahrheit. Springer, Wien
Wald A (1950) Statistical decision functions. Wiley, New York
Watson G (1983) Hypothesis testing. In: Kotz S, Johnson N (eds) Encyclopedia of statistical sciences. Bd. 3. Wiley, New York, pp 712–722
Yule GY (1895) On the correlation of total pauperism with proportion of out-relief, I: all ages. Econ J 5:603–611
Yule GU (1896a) Notes on the history of pauperism in England and Wales from 1850, treated by the method of frequency-curves; with an introduction on the method. J R Stat Soc 59(2):318–357
Yule GY (1896b) On the correlation of total pauperism with proportion of out-relief, II: males over sixty-five. Econ J 6:613–623
Yule GY (1927) On a method of investigating the periodicities of disturbed series, with special reference to Wolfer's sunspot numbers. Philos Trans R Soc A 226(1927):267–298
Zellner A (1971) An introduction to Bayesian statistics in econometrics. Wiley, New York
Zellner A (1992) Review of Morgan, Morgan M (1990) the history of econometric ideas. J Polit Econ 100:218–222

Recommended Reading

The best starting point is still Gigerenzer et al. (1989). See Cited Literature. Other helpful overviews are:

Cohen IB (2005) The triumph of numbers: how counting shaped modern life. W. W. Norton, New York

Kotz S, Johnson NL (eds) (1992) Breakthroughs in statistics, 1. Foundations and basic theory. 2. Methodology and distribution, Springer series in statistics. Springer, New York

Lenhard J (2006) Models and statistical inference: the controversy between Fisher and Neyman-Pearson. Br J Philos Sci 57:69–91

Salsburg D (2001) The lady tasting tea: how statistics revolutionized science in the twentieth century. Freeman, New York

Sprenger J (2014) Bayesianism vs frequentism in statistical inference. In: Hájek A, Hitchcock C (eds) Handbook of the philosophy of probability. Oxford University Press, Oxford

Sprenger J, Hartmann S (2001) Mathematics and statistics in the social sciences. In: Jarvie IC, Bonilla JZ (eds) The SAGE handbook of the philosophy of social sciences. Sage, London, pp 594–612

Stigler SM (1999) Statistics on the table: the history of statistical concepts and methods. Harvard University Press, Cambridge, MA