School of Economics and Management
TECHNICAL UNIVERSITY OF LISBON

Department of Economics

Francisco Louçã

Should The Widest Cleft in Statistics - How and Why Fisher opposed Neyman and Pearson

WP 02/2008/DE/UECE

WORKING PAPERS

ISSN Nº 0874-4548

The Widest Cleft in Statistics

- How and Why Fisher opposed Neyman and Pearson

Francisco Louçã (UECE-ISEG, UTL, Lisbon)
flouc@iseg.utl.pt

Abstract

The paper investigates the "widest cleft", as Savage put it, between frequencists in the foundation of modern statistics: that opposing R.A. Fisher to Jerzy Neyman and Egon Pearson. Apart from deep personal confrontation through their lives, these scientists could not agree on methodology, on definitions, on concepts and on tools. Their premises and their conclusions widely differed and the two groups they inspired ferociously opposed in all arenas of scientific debate.

As the abyss widened, with rare exceptions economists remained innocent of this confrontation. The introduction of probability in economics occurred in fact after these ravaging battles began, even if they were not as public as they became in the 1950s. In any case, when Haavelmo, in the 1940s, suggested a reinterpretation of economics according to the probability concepts, he chose sides and inscribed his concepts in the Neyman-Pearson tradition. But the majority of the profession indifferently used tools developed by each of the opposed groups of statisticians, and many puzzled economists chose to ignore the debate.

Economics became, as a consequence, one of the experimental fields for "hybridization", a synthesis between Fisherian and Neyman-Pearsonian precepts, defined as a number of practical proceedings for statistical testing and inference that were developed notwithstanding the original authors, as an eventual convergence between what they considered to be radically irreconcilable.

"But the age of chivalry is gone.
That of sophisters, economists,
and calculators, has succeeded;
and the glory of Europe
is extinguished for ever."
Edmund Burke, 1790

1.
The cleft

Modernity was, since its inception, a hurricane devastating previous traditions, modes of thought, habits and creeds. Its sweep represented as well a loss of innocence for the new ages, and this is far clearer in science than in any other domain: looking for facts, for causes and for consequences, as an exhaustive quest for quantification and for certainty, science superseded all other forms of knowledge and imposed the social authority of numbers and experts moved by the superior ideal of control of nature. If a book evokes no numbers and experimental reasoning, "commit it then to the flames for it contains nothing but sophistry and illusion", urged Hume (1748: 165). The laments of the likes of Edmund Burke against the emergence of modernity and expressing the fear of enlightenment could not challenge this movement.

Science was therefore developed as measurement and quantification, as the perception of the numbers ruling the universe. The age of chivalry was indeed gone, and the epoch of calculators and economists succeeded: mechanics was the name for that knowledge. Mechanisms were discovered to rule everything, from the Newtonian cosmic laws to geometry, from clocks taming time to motors and other artefacts moving the world. All science could not be but mechanistic. Even Darwin's anti-deterministic *The Origin of the Species* was celebrated by Boltzmann as the introduction of "a mechanical view of Nature" (Gigerenzer et al., 1989: 136).

This view was certainly shared by major biologists. For Ronald Aylmer Fisher, one of the protagonists of this narrative, evolutionary theory was constructed in analogy with statistical mechanics, since evolution by natural selection "may be compared to the analytic treatment of the theory of gases, in which it is possible to make the most varied assumptions as to the nature of the individual molecules, and yet to develop the general

laws as to the behavior of gases" (Fisher, 1922: 324). Fisher's model of Mendelian populations was in fact metaphorised from the molecular models of statistical mechanics applied to gases, as he always made clear (Morrison, 2002: 64).

This analytical treatment of such diverse subjects as gases and genes was the function of statistics, the modern form of measurement and quantification, that corpus of techniques permitting the nomad analogy to travel from the theory of gases to all other domains and to impose mechanics as the paradigm.

Statistics, therefore, was the *motum* for the modern revolution in sciences. But statistics developed through time in two different directions. The first was measurement as such, with statistics posing as the representation of reality as data gathering. For classical mechanics, probability was a condition for measurement since human ignorance impinged errors in statistics, and the computation should either clean or at least miminize the error abusively superimposed on perfect determination: the error is part of the observer. The second direction was, on the contrary, to conceive of probability as the expression of the very structure of nature, as quantum mechanics or evolutionary biology suggested: the error is part of nature.

This was more difficult to generalise, in particular in social sciences, since it required a new conception of human agency and indeed "probability has not been a basic building block of scientific theorizing except in physics and in evolutionary biology. (…) Economists, for example, construct their theories as if they were mechanical systems; yet in applications these theories are given a stochastic representation for the purposes of empirical adequacy" (Kruger et al., 1987: 3), and consequently the exact statute of the stochastic process is not frequently established in economic models – a theme for the last section of this paper.

Modern frequencists, as the probability revolution unfolded under the lead of quantum mechanics, shared both the epistemic view of statistics as measurement and the ontologic view of the intrinsic stochastic nature of nature. In biology, they were all the better prepared to accept both the requirement for measurement and the concept of variation, since that was the core of Darwinism. As a consequence, in the 1930s evolutionary biology was one of the first sciences to be reconstructed on probabilistic foundations. Biologists had therefore the theory, the motivation, the laboratories, the

ability to measure and both the possibility and the will to make inferences. And the most distinguished of them had also an immense capacity to fight. Evolutionary biology became consequently the field of experimentation of new statistical methods and of confrontation of different emerging approaches.

The divergences opposing some of the founders of the modern statistics and evolutionary biology were epic, "one of the most bitter disputes in science" (Keuzenkamp and Magnus, 1995: 10). As Savage put it in a volume edited by Neyman, "the widest cleft between frequencists is that between R.A. Fisher and those who side closely with him on the one hand and those who more or less associate themselves with the school of Jerzy Neyman and Egon Pearson" (Savage, 1961: 577). The two groups opposed on almost everything, from the concept of scientific inference to the definition of statistics.

Yet, this was not necessarily the only possible destiny of their conversation. In fact, both Fisher and Neyman felt the vocation to supersede Karl Pearson's heritage in statistics: the former was a personal adversary of Pearson, and the second underrated his mathematical capacities and, although less outspoken given his cooperation with Karl's son, Egon, did never hide the feeling that a new start was required (Louçã, 2007). For a time, this provided motivation for convergence between both men.

When trying to move to Western Europe to enhance his academic opportunities and to avoid danger in those frightful 1930s, the Polish Jerzy Neyman corresponded with Fisher, only four years his elder, believing he was the man to help him. They were indeed quite close in the vision of statistics as the language for the new science to rule all other sciences. Gosset, a friend of both, emphasised their motivational vicinity when he wrote to Fisher in order to arrange for the visit of Neyman to Rothamstead: "He is fonder of algebra than correlation tables [meaning, against Karl Pearson] and is the only person except yourself I have heard talk about maximum likelyhood (sic) as if he enjoyed it" (quoted in Joan Fisher, 1978: 451).

A curious letter in 1932 presents some preliminary results by Neyman, claiming "these results [to be] in splendid disagreement with the last articles of Professor Pearson", certainly a conclusion meant to please Fisher. The suggested best tests "guarantee the minimum frequency of errors both in rejecting the true hypothesis and in accepting a

false one", an anticipation of the line of research Neyman was then following with Egon Pearson (Neyman to Fisher, 9 February 1932). Neyman insisted on the opportunity for working under Fisher's guidance: "I am often thinking that it would be very useful for me to work in contact with you. Unfortunately this requires considerable amount of money – without speaking of your consent – of course" and later asked again for a place.[1]

The opportunity to leave Poland only presented later, and it was Egon Pearson who managed to get a job for his co-author, who came in 1934 to University College, London, to find Fisher segregated in a different floor and department. There is evidence that, at first, they got along well enough.[2] Neyman presented a paper in 1934 phrasing his proposal of confidence intervals as a reformulation of Fisher's fiducial argument, and the latter was "ironically one of the few to comment favourably on Neyman's paper", although criticising him for the limited inferential value since not using all available information (Zabell, 1992: 374).

Yet, the following year their relation broke and the pretext was the discussion on Neyman's paper on agricultural experiments. The clash was violent, as the flourished rhetorical bombardments witness and, as Fisher mounted his attack on Neyman, Egon Person came to the rescue of his co-author, accusing Fisher of nothing less than ignorance, after Neyman having been accused of the same sin.[3]

But by that time the "widest cleft" was already deep enough and, when in 1938 Neyman moved to Berkeley and was in position to otherwise influence the development of his

[1] Neyman to Fisher, 28 October 1932 and 13 June 1933. When he approached Fisher for a job, Neyman had already developed and published with Egon Pearson the first building-blocks of his own theory of estimation. It is obvious he did not consider that to be in contradiction with Fisher's ideas. Fisher highlighted this connection as he accepted, even when the quarrel was installed, that "For several years it is certain that Neyman thought, in speaking of 'confidence-intervals, that he was only systematising and developing a new exposition appropriate to the fiducial argument I had put forward and pressed upon his notice" (Fisher to M.G. Kendall, 3 November 1943).

[2] Neyman's biographers state that "Initially Fisher was gracious and Neyman admiring, but they soon came into conflict" (Lehmann and Reid, 1982: 161). Joan Fisher declares that Neyman, after a short period of peace, "sniped" at her father since the 1935 debate (Fisher, 1978: 262-3).

[3] Fisher on Neyman: "I suggest that before criticizing previous work it is always wise to give enough study to the subject to understand its purpose". Pearson on Fisher, speaking in defence of Neyman: "Dr. Pearson said while he knew that there was a widespread belief in Professor Fisher's infallibility, he must, in the first place, beg leave to question the wisdom of accusing a fellow worker of incompetence without, at the same time, showing that he had succeeded in mastering the argument" (minutes of the discussion, in Neyman, 1935: 202).

science, the offences of the civil war in statistics were numerous.[4] Neyman himself later addressed this topic in a paper on the "Silver Jubilee" of his quarrel with Fisher, although presenting an unfairly biased account of the confrontation.[5]

It is interpretative and therefore highly subjective and impossible to establish the complete set of personal reasons for the opposition between these scientists. Some clues are given in the correspondence by Fisher to some colleagues, but no uncontroversial reason can be identified as the main cause for this *malaise*. Fisher, who "sometimes became an irascible protagonist" (Bennett, 1990: vii), accused Neyman of using his own work without reference,[6] but still this would imply their theories to be related, a claim both vehemently denied. In any case, the confrontation grew exponentially, to the point of reference to "mad Neymanians in California",[7] to people "being 'bawled down' by Neymanians", who were supposed to "intimidate Americans successfully enough, especially refugees anxious to get posts in American Universities",[8] since "Neyman is, judging by my own experience, a malicious mischief-maker. Probably by now this is sufficiently realised in California. I would not suggest to anyone to engage in scientific controversy with him, for I think that scientific discussion is only profitable when good faith can be assumed in the common aim of getting at the truth",[9] all this from a man who candidly confessed "averse to controversy in print or in letters".[10]

Neyman retorted in the same measure, simply accusing his opponent of incompetence: "the theory of fiducial inference is simply non-existent in the same way as, for example, a theory of numbers defined by mutually contradictory definitions" (Neyman, 1941: 149).

---

[4] It is to be noted that Egon Pearson ceased the intense collaboration he had with Neyman since the latter moved to the US.

[5] Neyman argued that Fisher intensely persecuted his views through his entire career, but this description is wrong. This could only be true as far as the animosity between both men led to a campaign of private letters mixing disdain and argument; but Fisher only twice criticised his opponents in public until the 1955 book although from then he did not stop (Zabell, 1992: 376). It is certainly true to say that, in this story, neither side spared the other.

[6] Fisher often suggested that Neyman and Pearson had taken from him their theory of inference and confidence intervals "and proceeded to expound it to the world with the minimum reference to its origin" (Fisher to M.G. Kendall, 9 but also 2 November 1942, just as he had wrote to W.A. Shewhart, 2 February 1940).

[7] Fisher to D.D.A. Sprott, 13 January 1962.

[8] Fisher to D.A.S. Fraser, 11 January 1962.

[9] Fisher to H. Gray, 2 July 1951.

[10] Fisher to G.S. James, 16 August 1955.

In any case, as powerful as he was in theory making and in practice and teaching, Fisher was confronted with the vigorous success of the Neyman-Pearson method. One of his friends, G.A. Barnard, who was the vice-president the Royal Statistical Society under Fisher's presidency, noted that his antagonists were fashionable enough:

> If I may say so, it has for a long time struck me that Neyman and Pearson's ideas have caught on widely, because they are based on an explicit theory of probability (the neo-classical theory), and they are therefore more easily put in 'clear' mathematical language than your own ideas. (…) As far as I know (and I should be very grateful for correction), your own ideas on probability, as distinct from statistical testing, have not been set out explicitly in a complete form. (G.A. Barnard to Fisher, 14 October 1945)

Much later, Barnard still complained in the same sense: "almost everyone agrees with your criticisms of Neyman and Jeffreys, and with at least a part of your general theory. But nearly all become highly convoluted because (I think) their mathematical training has followed the now fashionable trend towards axiomatics to such an extent as to make them sometimes incapable of following a semantic argument".[11] As Barnard noticed, the explicit support of the Neyman-Pearson technology on a theory of probability and its axiomatics propelled it to large audiences.[12] In fact, this gained Neyman a large support that Fisher was wrong to undervalue. As Seidenfeld put it, "by the end of World War II the overwhelming majority of practising statisticians took it [Neyman-Pearson's theory] as the received position" (Seidenfeld, 1979: 30), and others state that "at the formal level Neyman and Pearson seem to have won the battle with Fisher" (Keuzenkamp and Magnus, 1995: 13). But Fisher had weapons to oppose his adversaries and did not hesitate to fight back.

The temptation to explain this confrontation through the idiosyncrasies of the contenders frequently emerged. Indeed, in particular the personal characteristics of Fisher generated a legend, and it was not without reason. His daughter and biographer, Joan Fisher Box, wrote that "He grew up without developing a sensitivity to the ordinary humanity of his fellows. He was unaware of the effects of his own behavior, and often expressed his love ineptly". Furthermore, "He was at once exceedingly self-

[11] Barnard to Fisher, 27 February 1958.
[12] In contrast, Fisher either neglected the "continental" work on axiomatics (Bartlett, 1965: 407) or distasted it (Zabell, 1992: 387).

centered and utterly self-forgetful, charming and impossible. And his friends learned to accept his inconsiderate and irritable behavior because they perceived the greatness of his character", in spite of the fact that "He would 'fly off the handle' for not apparent reason and was fiercely intolerant of petty inconveniences. His anger was unpredictable" (Joan Fisher, 1978: 10, 12, 170).

Karl Pearson, one of his opponents, accepted he was an "apt controversialist" (Pearson, quoted in Inman, 1994: 6), and Neyman emphasised "Fisher's remarkable talent in polemics" (Neyman, 1956: 292), although the younger Bartlett – who was another fighter in these early skirmishes – was less measured, noticing he had been "downright rude" and "did not argue fairly" against Neyman and Pearson, to conclude that "The trouble with great men, especially those with temperaments of comparable stature, is that they are liable to excite either allegiance or rebellion" (Bartlett, 1965: 397, 408). For Savage, Fisher "sometimes published insults that only a saint could entirely forgive" and added "I am surely not alone in having suspected that some of Fisher's major views were adopted simply to avoid agreeing with his opponents" (Savage, 1976: 446). In the same mood, Mackenzie registered his "extreme egocentricity and violent temper" (Mackenzie, 1981: 184). In any case, Fisher certainly impressed by animosity and imposing attitude, namely his "English spoken in a Shakespearean style and delivered in the manner of a Spanish grandee" (Hoyle, 1999).

However relevant these personal characteristics may have been, the nature of the debate about the foundations of statistics invites for a balanced assessment of the very arguments. These are the theme for next section.

2.
Ravaging battles between R.A. Fisher and Jerzy Neyman-Egon Pearson

R.A. Fisher's best selling books[13] and acclaimed papers were the driving force for a major dislocation of the themes of statistics: instead of the correlation established between variables describing large populations in a causeless world, in the mood of Karl Pearson, under Fisher rigorous protocols of experimentation, randomisation and modelling were defined, allowing for sampling and inference. Among Fisher's major contributions, one can count the analysis of variance and the definition of the concept of maximum likelihood.[14] As a consequence, statistics gained both a solid mathematical basis and an experimental reference imposing its authority.

This process of transformation was intensely disturbed when the "widest cleft" was declared. There was a "negative impact upon science from the disagreement" and it "had a significantly deleterious effect upon the practice of statistics in science, essentially because it has led to widespread confusion" (Berger, 2003: 2, 4).

This widespread confusion was to be expected since, as their investigations proceeded, Fisher and Neyman-Pearson generated successive domains of confrontation spreading over all the major topics of statistics. One of the protagonists, Egon Pearson, appreciated these shocks as harmful blows to statistics: "The titanic battles which have from time to time been waged across the statistical field were perhaps enlivening to the onlookers, but they were very real and I think harmfully moving to the participants" (Pearson, 1968: 456).

Let me state some of the differences of practical application, as both Fisher and Neyman-Pearson proposed their concepts (Table 1), before appreciating their rationale.

---

[13] The first book, "Statistical Methods for Research Workers" was initially ignored by the peers but acclaimed by the researchers: "it did not receive a single good review" (Joan Fisher, 1978: 130) and yet it became a bestseller.

[14] Not to mention another of his contributions, a relevant one for a new born science: the definition of operational concepts, which are now trivial, such as "parameters", to "specify the parent population", and statistic, "calculated from the observed sample" (Fisher to W.E. Deming, 25 September 1934).

Table 1: Fisher and Neyman compared: P-value (Fisher) versus fixed-error probabilities (Neyman-Pearson) testing

Significance testing	Hypothesis testing
Concept: 1. the information is an extraction from an infinite hypothetical population 2. the statistics is a property of the sample Procedure: 1. state the null, H_0 3. specify the test statistic T and its distribution under H_0 being true 4. apply to data and calculate T 5. determine the p-value corresponding to T under the assumption of H_0 being true 6. reject H_0 if p is small	Concept: 1. the test is conceived of as a frequency obtained from repeated sampling 2. size and power are properties of the test Procedure: 1. state the null H_0, and the alternative H_A 3. specify T and its distribution under Ho 4. specify the significance level α and define the rejection region R under H_0 5. apply to data and determine T 6. reject H_0 and accept H_A if T is in R

The application of these methodologies can deliver very different conclusions in some cases. Berger proposes the following example: suppose $\{X_1 \ldots X_N\}$ are *iid* from $N(\mu, \sigma)$ with known variance, n=10, and we test H_0:μ=0 versus H_A:μ#0. If z=2,3 (or z=2.9), then Fisher would report p=0,021 (or p=0,0037) and Neyman, prescribing an error Type I of 0,05, would report α=0,05 in either case (Berger, 2003: 1). Other authors suggest otherwise that the methods are generally equivalent: "For it turned out that both methods were mathematically equivalent. The tail area is, for the usual choices of alternative hypothesis, nothing more than the projections of a rejection region on the real line, and can be justified on the grounds of avoiding an error of the first kind. Furthermore, it turns out that many Fisherian choices of a test statistics are equivalent to a choice of an alternative hypothesis" (Gigerenzer et al., 1989: 99-100). In any case, either confusion is the result of these alternative presentations of testing, or divergence in conclusions challenges the rigour and objectivity of statistics, both implications being highly undesirable.

In spite of the impacts of the cleft on the general perception of the ability of statistical science, both sides proved to be unrepentant. Indeed, Fisher always rejected the core concepts of the Neyman-Pearson approach: the notion of error of the second kind, the idea of repeated samples from the same population and, moreover, the behavioristic approach to induction – acceptance of the alternative hypothesis being obtained through

rejection of the null. Alternatively, Neyman argued that Fisher had no logical basis for the choice of his test statistics and that it did not obey the frequencist criteria.

These divergences are recapitulated below in three sections. The first deals with the theory of induction and its foundations, while the second evaluates the views on refutation and confirmation and the third discusses the concepts of models and experiments. In each case, the cleft is reassessed and the views of both camps are compared.

The final result of the cleft is yet to be established. Some scholars argue that "At the formal level Neyman and Pearson seen to have won the battle with Fisher", in spite of the fact that econometric practice seems to follow Fisher rather than Neyman and Pearson: "But how often is the question that an econometrician has to answer a decision problem in the context of repeated sampling? Fisher's interpretation of a small P-value (which follows the tradition of Laplace to K. Pearson), that either something very unlikely has happened or the null is false, may be more useful in econometric practice" (Keuzenkamp and Magnus, 1995: 13, 18).

2.1. The first difference: the theory of induction

Probability, as the heart of statistical theory, is since its origins predicated upon a theory of induction. The success of R.A. Fisher's methods was precisely based on his contribution to the definition of statistical induction as a model of scientific inference itself and in that sense he stood almost alone among his contemporaries. Fisher was philosophically committed to the Laplacean ambition of a statistical redescription of a fully deterministic world, and he singled out as one of the few capable of providing the arguments for this ambitious scientism. The history of Fisher's adherence to the modern views on statistical inference is consequently very relevant for this narrative.

In the scholar year of 1912-3, having been granted a studentship in physics, Fisher dedicated his graduate research in Cambridge to study quantum mechanics with the physicist James Jeans and the theory of errors with F.J.M. Stratton. As Joan Fisher summarizes, as a consequence he accepted the divulgation of Heisenberg's principle of uncertainty in the 1930s as a natural implication of physics, unlike other scientists (Fisher, 1978: 33, 290).

From that time and through his whole life as he was exposed to biological experimentation and statistical inference, his heroes were the improbable pair of Darwin and Boltzmann (Depew and Weber, 1996: 246). From Boltzmann, Fisher concluded that probabilities were an objective feature of the world, unrelated to subjective states of mind; consequently, statistics should be able to decipher the natural laws and not just to point out the omnipresence of stochasticity (Depew and Weber 1996: 254, 272). Boltzmann and quantum mechanics, provided the methods, as Fisher unreluctantly emphasised: "the whole investigation may be compared to the analytical treatment of the Theory of Gases, in which it is possible to make the most varied assumptions as to the accidental circumstances, and even the essential nature of individual molecules, and yet to develop the general laws as to the behavior of gases" (Fisher, 1922: 321-2).

It is obvious Fisher was conceiving this analytical treatment not as forces in action, but as the physics of energy, namely fitness being the equivalent to entropy:

> It will be noticed that the Fundamental Theorem (…) bears some remarkable resemblances to the Second Law of Thermodynamics. Both are *properties of populations*, or aggregates, true irrespective of the nature of the units which compose them; both are statistical laws; each requires the constant increase of a measurable quantity, in the one case the *entropy* of a physical system and in the other the *fitness* (…) of a biological population. (Fisher, 1930: 36-7, my italics)

Fitness, like entropy, was consequently measured as part of a dynamic maximization process tending to a most probable distribution, its equilibrium (Depew and Weber, 1996: 262). As quantum physics suggests, these measures are supposed to be qualities of the populations: in that sense, Fisher was the biologist whose propositions were akin to the Boltzmannian approach. In that he differed from J.B.S. Haldane, who looked for the modification in one single gene, to be amplified in the population, whereas for Fisher only the dynamics of arrays of genes in the population mattered.

The concept of population therefore became a cornerstone of the introduction of probability in biology and, as a consequence, in other sciences as well.

2.1.1. The concept of population

The probabilistic revolution entered in biology as the Mendelian reconfiguration of Darwinism, and Fisher was instrumental to that synthesis. Indeed, after a generation of heated confrontation between biometricians and Mendelians, a convergence was obtained, and that was the result of proof via statistics. Evidence shows that, at least for Fisher, the synthesis in biology incorporated the benefits from an education in quantum physics, namely in the definition of probability itself. For Fisher, as he stated in an early statement, biological populations were conceived as an analogy with the populations of atoms in gas theory, following the approach of the molecular models of statistical mechanics in order to define variation at the level of Mendelian characters and not at the level of the phenotypes (Fisher and Stock, 1915).

In this sense, Fisher imposed the notion of data as being drawn from an infinite and therefore imagined population as the condition for statistical inference from an experiment:

> The object is accomplished by constructing a hypothetical infinite population, of which the actual data are regarded as constituting a random sample. The law of distribution of this hypothetical population is specified by relatively few parameters, which are sufficient to describe it exhaustively in respect of all qualities under discussion. (Fisher, 1922: 311)

A simple example argues for this notion of an infinite population:

> The idea of an infinite hypothetical population is, I believe, implicit in all statements involving mathematical probability. If, in a Mendelian experiment, we say that the probability is one half that a mouse born in a certain mating shall be white, we must conceive of our mouse as one of an infinite population of mice which might have been produced by that mating. The population must be infinite for in sampling from a finite population the fact of one mouse being white would affect the probability of others being white, and this is not in the hypothesis we wish to consider (…). Being infinite the population is clearly hypothetical (…). (Fisher, 1950: 699)

Both the concept of population and the definition of a model to interpret it fuelled the main early confrontations between Fisher and the elder Pearson. Karl Pearson strongly rejected Mendelism, since his radical positivism denied any theoretical entity such as a gene and favoured instead an acausal and purely descriptive strategy for biology. Yet, the core of his divergence with Fisher was the definition and model of a population: "The dispute between Pearson and Fisher centred on how populations should be characterised with respect to the individuals they comprise" (Morrison, 2002: 62). Indeed, Fisher's path-breaking contribution to the synthesis between Mendelism and Darwinism was precisely the notion of demonstration of natural selection operating in Mendelian populations, and his analysis of variance was conceived as a determination of the distinctive genetic and environmental influences.

Karl Pearson strongly opposed to the concept of an infinite population, something his positivism could not accept in any case; instead, for Fisher, the specification of an infinite and hypothetical population was the condition to infer the parameters of the model.[15] Consequently, statistics could not live without the concept of the infinite population. It could not come as a surprise if Fisher bombarded Neyman and the younger Pearson with the same argument he was so proud to differentiate him from Karl Pearson since the successful beginning of his career.

As the story unfolded, it became clear that it was due to these essential philosophical anxieties, namely coherence with the Boltzmannian concepts of probability, that Fisher opposed Neyman and Pearson about their concept of repeated samples, which was clearly in contradistinction with his own notion of information as an extraction from a population.

Although the background in theoretical physics and statistics of this debate was not invoked by the contenders, it is not without importance. In fact, physics was at the time of the formation of Fisher's and Neyman's approaches evolving from statistical mechanics (probability being applied to aggregates of elements and systems, measured as relative frequencies) to quantum mechanics (also extended to the probability of behaviour of a single element). Furthermore, Maxwell had introduced the ergodic

[15] The definition of probability was postulated in relation to the infinite hypothetical population: "the probability that an event *e* has property *F* is p_1 iff *e* is *known* to belong to a reference class of 'possible events' with proportion p_1 having property *F*, and such that *no subset* of this reference class having different proportions for *F*-ness *is known*, i.e. no *recognizable* subset is known. Under these conditions, *e* is said to be a random sample (or random member) from a hypothetical reference class with respect to the property *F*" (Seidenfeld, 1979: 73).

hypothesis, following Boltzmann, considering that a system in equilibrium assumes all states compatible with the conservation of energy, and promoting as a consequence an "ensemble approach" studying the population of possible systems distributed over all states compatible with the observed magnitudes, as Gibbs summarised it. The concept of population of possible events referred therefore to this epistemic interpretation of probability as the chance of finding the one event under scrutiny, an interpretation put forward by Fisher. Working in this frontier of the theory of errors inspired by Boltzmann and the "ensemble approach", Fisher felt secured by the coherence of this approach he was championing in statistical biology.

2.1.2. Laplace between Fisher and Neyman

A peculiar aspect of the "widest cleft" was the vindication of the authority of the fathers of statistics. Both groups of contenders frequently referred to the work of ancestors in order to reinforce their points, so that Laplace and Gosset were frequently quoted in their arguments and counter-arguments. Egon Pearson used a letter by "student", *alias* Gosset, to argue the pertinacity of the tests of hypotheses, but it was about Laplace that the dispute turned bitter.

Indeed, both sides had strong points. For Fisher, and he was right on that, Neyman and Pearson had abandoned the Laplacean pretence of a theory of induction, and therefore of a general method for scientific inference. In fact, Neyman rejected Fisher's programme of disproving the null as the basis for inductive inference, and instead argued for an inductive behaviour using the test of hypothesis as a decision rule (Kruger et al., 1987: 18; Gigerenzer et al, 1989: 207; Keuzenkamp and Magnus, 1995: 10). Consequently, Fisher "believed Neyman had mistakenly reinterpreted his tests of significance in terms of acceptance procedures, an ideological point of view that valued expediency over truth" (Gigerenzer et al, ibid.: 104).

The direct probability statement produced by Fisher was challenged by Neyman and Pearson, who preferred a model of decision, an acceptance procedure that was part of the behaviouristic approach they preached for statistics. In that sense, they abandoned Laplace.

Nevertheless, Neyman and Pearson considered that this was the price for fidelity, since a pure frequencist approach ought to sacrifice the philosophical pretence to the statistical convenience. The repeated sampling approach, according to them a condition to interpret the tests of hypotheses as probability statements, was considered to be the rationale for statistical decision: "Without consideration of hypotheses alternative to the one under test and without the study of probabilities of errors of the two kinds, no purely probabilistic theory of tests is possible" (Neyman, 1956: 288). In fact, Neyman – who was obviously the leading theoricist of this cooperation with Pearson – accused Fisher of subjectivism given his supposition of a hypothetical population, and furthermore suspected Fisher of abandoning a strict frequencist attitude for the sake of his design of scientific inference, and in that he was right. And so he could turn the accusation and vindicate Laplace: since Fisher denounced Neyman, Pearson, Wald and Bartlett "allegedly based on an attitude identifying ordinary research with industrial acceptance sampling", i.e. on the basis of the repeated sampling interpretation, Neyman could present a quotation from Laplace sustaining his views on that topic (ibid.: 293).[16]

Apart from this dispute on the borrowed credibility of a respected founder of statistics, the reference to Laplace highlights other difficulties shared by both Fisher and Neyman.

The first is related to the nature of the frequencist view. Indeed, although both the behaviouristic and the inductive model of inference were relative to an imaginary framework – since Fisher's population was as imaginary as Neyman's repeated sampling procedure[17] – Fisher could claim his method led to a direct probability statement, since his test was a property of the sample, whereas the operational concepts of size and power defined by Neyman and Pearson were conditional on the test itself.[18]

---

[16] Hacking supports Neyman's claim on the precedence of Laplace in the theory of hypothesis testing, given his emphasis on the importance of a rival hypothesis, and argues that Fisher "was wrong in this respect" (Hacking, 1964: 16).

[17] Fisher denounced the "fallacious approach to tests of significance introduced by Neyman and Pearson, and usually expressed in such phrases as, 'the frequency found in repeated samples from the same population'. (…) I have only realised recently (…) to what extent Neyman and Pearson's approach, including this question of 'repeated samples from the same population', is due to their thinking of a test of significance as though it were a kind of acceptance procedure in which the repeated samples have an objective reality, and are not, as they are with tests of significance, constructs of the statistician's imagination" (Fisher to H.F. Smith, 27 August 1954).

[18] Gigerenzer and his co-authors discuss the frequencist interpretation to the claim of repeated sampling and the contradiction with Fisher's views: "Second, [for Neyman and Pearson] the frequencies of the errors of the first and second kind are calculated on the basis of repeated sampling of the distributions in the original mathematical specification of the problem, and the probabilities have therefore a direct frequency interpretation (…). Recall Fisher's belief that, in scientific applications, the population of the appropriate statistical model for the analysis of experimental data cannot in any realistic sense be sampled

From that Fisher concluded that "in the end Neyman denies the possibility of making any probability statement about the natural facts behind the observations".[19]

The second difficulty is related to the very concept of probability. It is obvious Fisher was unease with the limiting frequency interpretation of probability, since it supports no proposition on a specific fact, for instance it allows for no statement of probability of a particular throw of dice (Seidenfeld, 1979: 71). Fisher argued that probability statements are only a peculiar class of strictly mathematical specifications, not to be generalised to all cases of inference from observations:

> From my point of view the important point is that the original concept of probability is not adequate to specify the nature of uncertainty inherent in many forms of inference from observations. From this point of view it is almost unfortunate that a group of cases has been found in which inductive inference may properly be expressed in terms of probability, using the fiducial mode of argument; for this has tempted some mathematicians, and will, I fear, tempt more, to imagine that this type of argument is more widely applicable than is really the case, and to avoid enlarging their imagination sufficiently to grasp the cases where no probability statement is adequate. This is, in my view, a decisive reason against enlarging the meaning of the theory of probability so as to cover all types of inductive inference, since the word 'probability' must be tied closely to one quite defined mathematical concept." (Fisher to A.C. Aitken, 23 January 1936)

This view he maintained through all his life work:

> I have reiterated repeatedly that the concept of mathematical probability is inadequate to express the nature and extent of our uncertainty in the face of certain types of observational material, while in all cases the concept of mathematical likelihood will supply very helpful guidance, if we are prepared to

---

repeatedly, and has 'no objective reality, being exclusively the product of the statistician's imagination'" (Gigerenzer et al, 1989: 103). Moreover, "In place of what Neyman and Pearson saw as Fisher's quasi-Bayesian view that the exact level of significance somehow measures the discordance of data with the null hypothesis, their interpretation of statistical inference was a purely behavioristic one that refrained from any epistemic interpretation. The concept of size and power apply to a test, whereas Fisher's significance level is a property of the sample" (ibid.).

[19] Fisher to Finney, 24 March 1955.

> give up our irrational urge to express ourselves only in terms of mathematical probability. (Fisher to D.J. Finney, 14 September 1954)

This profound scepticism perplexed many of his contemporaries. Jeffreys, a leading Bayesian, questioned Fisher on the generalisation of probability statements,[20] and E.B. Wilson was tutored on the relevance of Keynes's concept of "the degree of rational belief" for those cases "from which no statement of probability can be derived".[21] The reason for this attitude was not mistrust in statistics but, on the contrary, a commitment to the role statistics should play as a rigorous instrument for scientific inference. This was precisely the reason for imposing the experimental nature of the probability statements:

> In the first place you seem to express surprise at my belief that 'statements of probability can be verified by observational frequencies'. It would have been more explicit had I said 'statements of probability in the Natural Sciences' or 'statements of probability referring to the real world' (…). In the second place you do not seem to grasp the central characteristic of the concept of mathematical probability, namely that it enables a statement of uncertainty to be made with rigorous exactitude. This requires a specification not only of what is known, or can be validly asserted on the data, but also of what is unknown, in order that the probability statement should be distinguishable from a statement of certainty. (Fisher to J. Tukey, 18 July 1955)

Of course, these observational frequencies obtained from investigation in natural sciences did not apply, according to Fisher, to the tests of hypotheses as suggested by

[20] When Jeffreys questioned him, since "I cannot follow your objection to the generalization to all probabilities of the laws of probability obtained for samples" (H. Jeffreys to Fisher, 24 February 1934), Fisher retorted: "I do not object to the generalization to all probabilities of the laws appropriate to the games of chance, but I do think, and indeed claim to have shown, that there are also logical situations in which a rigorous statement of the nature of uncertainty in our uncertain inferences is expressible not in terms of probability, but in terms of likelihood, a quantity which does not obey to these laws" (Fisher to Jeffreys, 26 February 1934).

[21] Fisher, who generally underrated Keynes's work, namely his book on probability theory, argued that "Keynes's excellent phrase, as I have always thought it, 'the degree of rational belief', is not really well applied to the concept of mathematical probability, in its classical sense implying the possibility of verification by observations of frequency, but can be seen to be more appropriate to the more primitive type of inference represented by those tests of significance from which no statement of probability can be derived" (Fisher to E.B. Wilson, 8 August 1956).

Neyman and Pearson: "The absence of reference to experience seems to me a serious flaw in their work".[22]

2.1.3. Fisher's criticism of Neyman's and Pearson's theory of hypotheses testing

Fisher conceived statistical research as the production of a direct probability statement for the limited class of cases for which that would be possible. The statement refers to the determination of the p-value, the test statistics being compared to its known distribution under H_o being true, and attributing to the researcher a decision on rejection or acceptance based upon experience (Fisher, 1922). Although Fisher loudly prevented the misinterpretation of *p*-values as error probabilities – as the obtained value of *p* is a function of a single data set –, Neyman criticised this approach as a violation of the frequencist principle and, at least at first, he conceived his own effort as an attempt to correct and extend this method to the choice of an efficient test based on a precise axiomatic and allowing for an objective protocol of acceptance or rejection.[23] But it soon became obvious that both Fisher and Neyman understood the divergence separating their investigations and theoretical foundations. Jeffreys, on the other hand, sharply rejected the use of p-values, arguing that a true hypothesis could eventually be rejected for not predicting values not having occurred (Jeffreys, 1961).

On the other hand, for Fisher, the Neyman-Pearson technology was unsound essentially given that the level of significance could not be defined as a relative frequency of repeated sampling from the same population: as his biographer stated, "the sampling property of confidence intervals (…) says nothing at all about the probability of θ given the result of the one particular sample of data actually obtained" (Joan Fisher, 1978: 451). Three reasons contributed to this denial. First, there is no opportunity for repeated sampling, because the values of the ancillary statistics cannot be fixed from one sample to the other simply given the fact that the population – or the reference set – changes from sample to sample. Second, the evidence against the null does not necessarily

[22] Fisher to C.I. Bliss, 6 October 1938.

[23] As previously argued, Neyman and Pearson at first saw their contribution as a refinement of Fisher's, although he took the opposite view, rejecting what he thought to be just an acceptance procedure similar to the control quality tests, leading to a distortion of the logic of inference: "I can now understand, much better than before, the early work of Neyman, or Neyman and Pearson, in the light of what you said the other afternoon, for it now seems clear to me, as it did not before, that Neyman, thinking all the time of acceptance procedures, was under the misapprehension that my own work on estimation had only the same end in view" (Fisher to Barnard, 9 February 1954).

match the relative frequency with which that p-value is attained (Johnstone, 1987: 492).[24] Third, Fisher pointed out a logical difficulty: if the true hypothesis is unspecifiable, it is immaterial to define the error of the second kind relative to the alternative hypothesis. Consequently, the sampling approach fails:

> There is a great deal in the approach chosen by Neyman and Pearson that I disagree with, but so far it seems to have led to nothing more than the conclusion that the tests of significance which I and those who agree with me had previously put forward were the best possible for their purposes (…). It is however, in my opinion, a pity that these writers have introduced the concept of 'errors of the second kind', i.e. of accepting an hypothesis when it is false, seeing that until the true hypothesis is specified, such errors are undefined both in magnitude and in frequency. (Fisher to Deming, 19 September 1935)

For Fisher, the null hypothesis simply cannot be considered true only because it failed to be contradicted, and the notion of the error of the second kind is incomputable whenever it depends on an unknown alternative.

Furthermore, Fisher strongly favoured applied research against mathematical abstractions and, namely, he rejected statistics to be part of deductive reasoning. In this he differed from Neyman: "Fisher was a research scientist using mathematical skills, Neyman a mathematician applying mathematical concepts to experimentation" (Joan Fisher, 1978: 265).

Other authors developed this discussion, suggesting that the logical mode of reasoning implied in Neyman's and Pearson's approach leads to the fallacy of denying the consequent, since obtaining atypical data under a certain hypothesis is no proof of the falsity of that hypothesis. Otherwise, the significance test under the null is not enough evidence for the hypothesis.

[24] The following example is given by Spielman: "The size of a test is the maximum relative frequency of erroneous rejections of the hypothesis tested among *all members of the reference class*, whereas the reliability of a negative diagnosis is determined by the relative frequency of erroneous rejections among *those members of the reference class for which a rejection will occur*. Those values can be radically different. It would be as stupid to regard one minus the size of a test as indicating the reliability of the test for rejection of the hypothesis tested as it would be for a male American of age fifty who has had a major coronary attack to regard the proportion of all fifty-year-old American males who will suffer a fatal coronary prior to their fifty-first birthday as indicating his chance of having a fatal coronary prior to his fifty-first birthday" (Spielman, 1973: 209).

The notions of size and power became the focus of much attention. Gill suggests they are contradictory: "Note that α and β are probabilities conditional on two mutually exclusive events: α is conditional on the null hypothesis being true, and β is conditional on the null hypothesis being false" (Gill, 1999: 651), whereas Spielman recapitulates an example by Neyman to argue that these concepts may be misleading and concurring to wrong decisions, and are therefore irrelevant as decision tools (Spielman, 1973: 202, 215). Consequently, these authors support Fisher's rejection of the Neyman-Pearson approach as unsuited for a theory of inference (Seidenfeld, 1979: 47).

Nevertheless, when he had presented his own alternative to Bayesian inverse probability, Fisher defined a confidence interval approach as part of the concept of fiducial probability.[25] In the early 1930s, most statisticians regarded fiducial probability and Neyman's confidence intervals as synonymous (Zabell, 1992: 371-3). Bartlett, a supporter of the Neymanian approach, argued that fiducial intervals are a particular case of confidence intervals, under a frequencist interpretation (Bartlett, 1965: 395). It is certainly obvious that afterwards, for polemic reasons as well as for deep convictions on the nature of statistics, Fisher rejected both the confidence intervals approach and its theoretical assumptions on the nature of frequencism, since evidence from the data should not be confused with a limit of actual frequencies but should be thought of as a frequency from the imagined population as well as a measure of rational belief. In this sense, the difference opposing Fisher and Neyman stood as philosophical questions on the essence of statistical inference. Whereas for Neyman probability was defined as the "idealization of long-run frequency in a long sequence of repetitions under constant conditions" (Lehmann, 1993: 1245), for Fisher the concept was ambiguous. Furthermore, for practical purposes, Fisher generally opposed the choice of fixed levels of significance for a decision procedure, therefore transformed into acceptance rules, a notion Pearson attributed to Neyman (Lehmann, 1993: 1244).

These differences opposing both camps confronted the inductive behaviour, solely based on deduction and theory of probability, against inductive inference, respectively championed by Neyman and by Fisher.

[25] Fiducial probability, as an alternative to the Bayesian concept of inverse inference based on prior probabilities, was Fisher's "most controversial proposal", and it proved to be "invalid" (Seidenfeld, 1979: 3, 105, 219).

2.2. The second difference: statistics as a tool for refutation

The value of statistical propositions has always been a mystery for a number of prestigious scientists. That was certainly the case in the period of emergence of statistics as the science of measurement and inference. One curious instance of that perplexity is the refrain of the song the director of Rothamstead wrote for the Christmas party of 1928:

> "Statistics, you see, is a wondrous cult
> For a non-mathematical mind,
> Which wants but the final, or end result –
> As to how its attained is quite blind"

The paradox in this blind process of a wondrous cult, the song went, was the importance of the error, a bizarre event only someone versed in statistics would be able to tame: as another verse assured, "Fisher can always allow for it" (Joan Fisher, 1978: 138-9). What to make of statements proved by error was in any case a matter of dispute among statisticians.

Indeed, if the first difference between Fisher and Neyman-Pearson concerns the logic of inference, the second is complementary if not derived from that and concerns the very nature of the statistical propositions. What they discussed was not at all trivial, since it provided an anticipation of the crucial confrontations in the twentieth-century epistemology on the nature of inference.

When it came to rationalize the practice of statistics, the Hypothetico-Deductive model was offered as a norm: it establishes a set of hypotheses and deduces a prediction out of them in order to be tested against the facts, asking then for confirmation (propounded by the Vienna Circle and Carnap) or, as it became the accepted view, for refutation (as suggested by Karl Popper). A disciple of Popper, Imre Lakatos, praised Neyman's and Pearson's test of hypothesis as "rest[ing] completely on methodological falsificationism" (Lakatos, 1978: 25), although the fact is that they developed their methodology independently and previously to Popper. Others emphasise, more rigorously, that the Fisherian test of significance is closer to Popperian falsificationism, since the null hypothesis can only be refuted (Gigerenzer et al 1989: 96), although also

in this case the concept was established well before Popper published his 1935 *magnum opus*.[26]

In fact, Fisher used the principle of refutation as an argument against his opponents, although he was at times ambiguous on the value of confirmation. What in any case it obvious is that Neyman fully endorsed a logic of confirmation, which was embedded in the choice of the most powerful test and in the concept of test of alternatives since "the numerical values of probabilities of errors of the second kind are most useful for deciding whether or not the failure of a test to reject a given hypothesis could be interpreted as any sort of 'confirmation' of this hypothesis" (Neyman, 1956: 290). Therefore, Neyman's approach is alien to Popperianism, which would question the mechanics of acceptance of whatever alternative hypothesis as a consequence of the rejection of the null.

Fisher rejected this approach: typically, he argued that "every experiment may be said to exist only to give the facts a chance of disproving the null hypothesis" (Fisher, 1935: 16). This radical claim was reiterated by Fisher on several occasions. One of such was a debate in the pages of *Nature* with the aged Karl Pearson, which was ignited by a letter by an applied scientist, John Buchanan-Wollaston, on the irrelevance of tests. Buchanan-Wollaston argued that one cannot test simultaneously "untruth on one hypothesis and the truth of the reverse hypothesis" (Inman, 1994: 3), and Fisher used the opportunity to tackle the Neyman-Pearson approach:

> It would, therefore, add greatly to the clarity with which the tests of significance are regarded if it were generally understood that tests of significance, when used accurately, are capable of rejecting or invalidating hypotheses, in so far as these are contradicted by the data; but that they are never capable of establishing them as certainly true. In fact the 'errors of the second kind' are committed only by those who misunderstood the nature and application of tests of significance. (quoted in Inman, 1994: 5)

Karl Pearson was also totally at odds with his son Egon, since his positivism radically prevented the acceptance of any claim of truth related to the hypothesis. In that he was

[26] Another peculiarity of this dispute on the attribution of Popperian credentials to any of the contenders is that Popper himself considered Darwinism to be a metaphysical theory and therefore non-scientific, which would exclude Fisher from considerations.

followed by Fisher: not rejecting the null does not establish its truth. But Karl Pearson was more radical than Fisher, since he considered that even rejection was restricted to the strange case of zero probability: "There is only one case in which a hypothesis can be definitely rejected, namely when its probability is zero" (quoted in Inman, 1994: 7). Therefore, except in this case, anything goes.

In this polemics with Neyman and Egon Pearson via Karl Pearson, R.A. Fisher attacked another assumption of the strategy of tests of hypotheses, its sampling interpretation, the condition for an acceptance procedure in a decision framework. Instead, Fisher proposed a nonsampling interpretation of the level of significance: the determination of a level of 0.05 does not imply that the investigator will be deceived once in twenty occasions.

The different interpretations of the meaning of a significance level opposed Fisher (interpreting the result as that of an individual test rather than of sequences of tests) and Neyman-Pearson (establishing the size and power of the test as derived from a series of similar tests). Furthermore, since the population is imaginary, the provisional conclusion is referred to the test and not to reality itself. In that, he distanced from Popper, who asked for refutation grounded on experimental evidence for a realistic conclusion. For Fisher, the whole procedure was only a matter of scientific attitude:

> What is particularly troublesome is that Neyman, in importing from Eastern Europe his misconceptions as to the nature of scientific research, should have chosen so ubiquitous a scientific tool as the test of significance as the subject on which to fasten ideas relevant only to the acceptance procedure. A typical test of significance is based on a probability statement derived from the hypothesis to be tested, and therefore existing only in the hypothetical world created by this hypothesis. Typically it leads to no probability statement in the real world, but to a change in the investigator's attitude towards the hypothesis under consideration, for which if we choose to use the word 'rejection' we must remember that the rejection is only provisional, and that our hypothetical calculations have shown that there would be a finite probability of our obtaining the observed level of significance even were the hypothesis true." (Fisher to N. Keyfitz, 21 November 1955)

Contradictorily and in certain occasions, Fisher indulged in the acceptance of a provisional confirmation: "if the observations are such that with reasonable probability they might have arisen on the hypothesis under test, this hypothesis, though not proved, has at least so far been confirmed, and, pending further and more stringent observations, may be accepted" (Fisher, 1951: 36-7).

Refuting or provisionally accepting the hypothesis, Fisher's approach established that "a significance test does not permit one to assign any specific degree of probability to the hypothesis" and that only in specific cases can the uncertainty of the hypothesis be expressed in probabilistic terms (Gigerenzer et al 1989: 93). In this framework, the test of significance tests the null for plausibility, measuring the deviations of observations, even if it is an unsettled question to know "if a significance level is a meaningful measure for discrepancy" (ibid.: 94). This is why Fisher adhered to the Keynesian notion of the "degree of rational belief" as a condition for inference in those cases in which no statement of frequencist probability could be derived: finally and in general, it would be up to the researcher, considering his knowledge and wisdom, to formulate his own intuition on the statistical evidence from the experiment, and no technology could substitute the experience of the experimenter.

2.3. The third difference: the concept of model and experiment

The notion of a statistical model was introduced by Fisher in 1922, with the paper "On the Mathematical Foundations of Theoretical Statistics". It refined the concept of variables and parameters and established the distinction between sample and population, defining the hypothetical infinite population. The subsequent work on statistics built on these definitions.

Neyman and Pearson initially shared this definition of the model, although suggesting a confirmationist interpretation of the results of the test:

> A model is a set of invented assumptions regarding invented entities such that, if one treats these invented entities as representations of appropriate elements of the phenomena studied, the consequences of the hypotheses constituting the model are expected to agree with observations. If, in all relevant trials, the

> degree of conformity appears to us satisfactory, then we consider the model an adequate model. (Neyman, 1957: 8)

What differenced Neyman and Fisher on the function of the model was the eventual incorporation of a decision framework as part of the definition of the test itself. For Neyman, the description of the model was reinterpreted in a behaviouristic framework and hypothesis testing was therefore part of a reiterated process, similar to quality control sampling, in order to establish the optimal procedure. In that sense, for Neyman the logic of inference emancipated from the notion of the hypothetical population (Lenhard, 2006).

Fisher rejected this concept of a behaviouristic approach to modelling and considered it was a consequence of alienation from the practical needs of applied research and explained the suport gained by Neyman: "I do not of course say, or, I hope, seem to imply, that in all mathematical departments the Neyman fog has settled in, but that it has settled in only in those departments which are insulated from practical research in the Natural Sciences" (Fisher to Barnard, 8 May 1961). Instead, he favoured a common sense approach to decision in testing:

> I am a little sorry that you have been worrying yourself at all with that unnecessary portentous approach to tests of significance represented by the Neyman and Pearson critical regions, etc. In fact, I and my pupils through the world would never think of using them. If I am asked to give an explicit reason for this I should say they approach the problem entirely from the wrong end, i.e. not from the point of view of a research worker, with a basis of well grounded knowledge on which a very fluctuating population of conjectures and incoherent observations is continually under examination. What he needs is a confident answer to the question 'Ought I to take notice of that?'. This question can, of course, and for refinement of thought should, be framed as 'Is this particular hypothesis overthrown, and if so at what level of significance, by this particular body of observations?'. It can be put in this form unequivocally only because *the genuine experimenter already has the answers to all the questions that the followers of Neyman and Pearson attempt, I think vainly, to answer by merely mathematical considerations*. (Fisher to W.E. Hick, 8 October 1951, my italics)

2.4. Hybridization

The final result of the widest cleft is to be established. What is certainly obvious is that the quarrel is essentially ignored in contemporary statistical theorising and that the practical workers indistinctively apply tools derived from one or the other of the contending camps, for instance the notion of the power of the test and that of significance. As Keuzenkamp and Magnus refer in relation to econometrics:

> Testing hypotheses belongs to the basic pastimes of econometricians. It is a compulsory topic in any course in introductory statistics and econometrics. In such a course, students are made familiar with notions like Type I and Type II errors, significance levels, and power. This is firmly in the tradition of statistical testing along the lines proposed by Jerzy Neyman and Egon Pearson. However, econometric practice seems closer to the approach of Sir R.A. Fisher, although he is rarely mentioned (apart from references to the F-test). (Keuzenkamp and Magnus, 1995: 6)

This is the result of what Gigerenzer and his co-authors named *hybridization*, the synthesis established in spite of the cleft and reconciling the points of view the original authors considered irreconcilable (Gigerenzer et al., 1989: 106 f.). Hybridization is a strategy of maximizing the benefits from both approaches, considering Neyman-Pearson's more suited for evaluation of the costs of alternative options, but less suited for scientific inference. Consequently, hybridization suggests a forest of interpretative deviations: "The researcher must specify the level of significance before conducting the experiment (following Neyman and Pearson rather than Fisher); he must not draw conclusions from a non-significant result (following Fisher's writings, but not Neyman-Pearson); and so on. Neyman's behavioristic interpretation did not become part of the hybrid, and Type I and Type II errors are given an epistemic interpretation. This has led to an enormous confusion about the meaning of a significance level" (ibid.: 107).

Confusion and personal bias, since, furthermore, it is accepted that all statistical alternatives require doubtful choices tainted with subjectivism: Fisher arbitrarily chooses the model and the test, Neyman and Pearson the class of hypotheses and the rejection region and the Bayesians the a priori probability (ibid.: 101, 105).

As a consequence, hybridization evolved to several attempts to correct subjectivism through the establishment of rigid protocols in testing. Berger suggested a conditional error probabilities approach, following Fisher rather than Neyman (Berger, 2003), and Lenhard suggested John Tukey's exploratory data analysis as a convenient synthesis (Lenhard, 2006). Lehmann, a disciple of Neyman, also battled for a unified approach, suggesting avoiding the omission of power (in Fisher's strategy) and the omission of conditioning (in Neyman-Pearson's; Lehmann, 1993: 1247). None of these was imposed as the authoritative synthesis in statistics and, as a consequence, hybridization remains an open menu for statisticians.

3. Conclusion

Modern statistics went through many great advancements: the extension of probability from games of chance to the concept of measurement and the further extension to the concept of nature itself. As a consequence, by the end of the nineteenth century statistics was reshaped by biology, engaged as it was in modelling variation and not so much in reducing error (Gigerenzer et al, 1989: 68). Karl Pearson and Ronald A. Fisher emerged as the constructors of modern statistics as they were both experimentally and theoretically involved in the explanation of variation, first under biometrics and then under the statistical exploration in the Darwin-Mendel synthesis.

The empirical foundation of biology favoured the adoption of a frequencist approach, but it was inside this camp that the wildest cleft developed opposing Fisherian tests of significance and concept of likelihood to the Neyman-Pearson theory of tests of hypotheses. The cleft distinguished between concepts of probability, concepts of sample and population, concepts of inference and as consequence the concepts of the tests themselves. Both contenders clearly defined their theories in the field they proceeded from: for Fisher, postulating an infinite hypothetical population was trivial, since the extraction of a small sample of biological entities was not supposed to change the nature of the population, whereas for the mathematically abstract entities defined by Neyman the probability could be conceived of as the frequency of successive sample extractions from the same population.

For economists, instead, none of the concepts apply trivially and that was the reason for so much resistance and miscomprehension. For in any realistic economic context, repeated extractions are not independent given time-dependency and, furthermore, there is not a stable population through time given structural change. Haavelmo addressed this problem suggesting a meta-historical view of a supra-population of "histories", and read this as the incorporation of the repeated sampling approach in order to sustain the applicability of Neyman-Pearson's theory. In this sense, he hybridized the available but contradictory theories in statistics.

The consequence was that economic variables were themselves to be reconsidered as stochastic processes, if this narrative were coherent. But the price for this is high enough and it was not unperceived by the contemporary economists, since this requires the abandonment of strict determinism: economics could not any more be written as the exploration in mechanical determination, in finding laws of behaviour and of action and consequence, but should instead consist in the description of laws of distribution and the determination of probabilities derived from them. It does not come as a surprise that this was accepted in words but not practiced in deeds.

References

Arrow, K. (1978), "Jacob Marschak", *Challenge*, March-April, 69-71

Bartlett, M.S. (1965), "R.A. Fisher and the Last Fifty Years of Statistical Methodology", *Journal of the American Statistical Association*, 60(310): 395-409

Berger, J. (2003), "Could Fisher, Jeffreys and Neyman Have Agreed on Testing?", *Statistical Science,* 18(1): 1-32

Bjerkholt, O. (2007), "Writing 'The Probability Approach' with Nowhere to Go: Haavelmo in the United States, 1939-1944", *Econometric Theory*, 23: 775-837

Davidson, D. and Marschak, J. (1959), "Experimental Tests of a Stochastic Decision Theory", in Churchman, C.W., Ratoosh, P. (eds.), *Measurement: Definitions and Theories*, New York: Wiley. 233-69, also Cowles Foundation Paper 137

Depew, D. and Weber, B. (1996), *Darwinism Evolving – Systems Dynamics and the Genealogy of Natural Selection*, Cambridge, Mass.: Bradford

Ezekiel, M. (1928), "Statistical Analyses and the 'Laws' of Prices", *Quarterly Journal of Economics*, 42: 199-227

Fisher, J.B. (1978), *R.A. Fisher – The Life of a Scientist*, New York: Wiley

Fisher, R.A. (1922), "On the Dominance Ratio", *Proceedings of the Royal Society of Edinburgh*, 42: 321-41
- (1922), On the Mathematical foundations of Theoretical Statistics, *Philosophical Transactions of the Royal Society of London*, B, 17, 69-78
- (1923), "Mr Keynes's Treatise on Probability", *Eugenics Review*, 14: 46-50
- 1925? (1973), *Statistical Methods for Research Workers*, 14th ed, New York: Hafner Press
- (1930), *The Genetical Theory of Natural Selection*, Oxford: Clarendon
- (1935), *The Design of Experiments*, Oliver and Boyd
- (1950), "Introduction to 'Theory of Statistical Estimation'", in *Contributions to Mathematical Statistics*, New York: Wiley
- (1951), "Statistics" in Heath, A. (ed.), *Scientific Thought in the Twentieth Century*, 31-55, Watts
- (1951), "Review of the Second Edition of 'Hereditary Genius', by Galton", *Eugenics Review*, 43: 37
- (1983), *Natural Selection, Heredity, and Eugenics*, edited by JH Bennett, Oxford: Clarendon Press
- (1990), *Statistical Inference and Analysis – Selected Correspondence of R.A. Fisher*, ed. by JH Bennett, Oxford: Clarendon

Fisher and Stock, C.S. (1915), "Cuenot on Pre-Adaptation. A Criticism", *Eugenics Review*, 7: 46-61

Frisch, R. (1934), *Confluence Analysis by Means of Complete Regression Systems*, Oslo: University Institute of Economics

Gigerenzer, G., Swijtink, Z., Porter, T., Daston, L., Beatty, J., Kruger, L. (1989), *The Empire of Chance – How Probability Changed Science and Everyday Life*, Cambridge: Cambridge University Press

Gill, J. (1999), "The Insignificance of Null Hypothesis Significance Testing", *Political Research Quarterly*, 52 (3), 647-74

Gillham, N. (2001), "Sir Francis Galton and the Birth of Eugenics", *Annual Review of Genetics*, 35: 83-101

Haavelmo, T. (1941 or 1942?), "A Note on the Variate Difference Method", *Econometrica*, 9: 74-9

Hacking, I. (1964), "On the Foundations of Statistics", *British Journal for the Philosophy of Science*, 15 (57): 1-26

Hoyle, F. (1999), *Mathematics of Evolution*, Memphis: Acorn Enterprises

Huberty, C. (2004), "Historical Origins of Statistical Testing Practices: The Treatment of Fisher versus Neyman-Pearson Views in Textbooks", *Journal of Experimental Education*, 61(4): 317-33

Hume, D. (1748), *Enquiries Concerning Human Understanding and Concerning the Principles of Morals*, Oxford: Clarendon, reprint

Inman, H.K. (1994), Pearson and RA Fisher on Statistical Tests: a 1935 Exchange from Nature, *American Statistician*, 48 (1): 2-11

Koch, L. (2004), "The Meaning of Eugenics: Reflection son the Government of Genetic Knowledge in the Past and the Present", *Science in Context*, 17(3): 315-31

Koopmans, T. (1937), *Linear Regression Analysis of Economic Time Series*, Haarlem: Netherlands Economic Institute
- (1975)

Kruger, L., Gigerenzer, G. and Morgan, M. (eds., 1987), *Ideas in the Sciences, The Probabilistic Revolution*, vol. 2, Cambridge: Bradford

Lakatos, I. (1978), *The Methodology of Scientific Research Programmes*, Philosophical Volume Papers 1, Cambridge: Cambridge University Press

Lehmann. E.L. (1993), "The Fisher, Neyman-Pearson Theories of Testing Hypotheses: One Theory or Two?", *Journal of the American Statistical Association*, 88($24): 1242-9

Lehmann, E.L., Reid, C. (1982), "In Memoriam, Jerzy Neyman 1894-1981", *American Statistician*, 36 (3): 161-2

Lenhard, J. (2006), "Models and Statistical Inference: The Controversy between Fisher and Neyman-Pearson", *British Journal for the Philosophy of Science*, 57(1): 69-91

Levene, H. (1974), "Harold Hotelling, 1895-1973", *American Statistician*, 28 (2): 71-3

Louçã, F. (2007), *The Years of High Econometrics – A Short History of the Generation that Reinvented Economics*, London: Routledge
- (2007a), *The Flags of Our Fathers – Did Eugenics Create Statistics?*, working paper

Mackenzie, D. (1981), *Statistics in Britain 1865-1930, The Social Construction of Scientific Knowledge*, Edinburgh: Edinburgh University Press

Magnello, M.E. (1999), "The Non-Correlation of Biometrics and Eugenics: Rival Forms of Laboratory Work in Karl Pearson's Career at University College London", *History of Science*, 37: 79-106 and 123-150

Marschak, J. (1946), "Neumann's and Morgenstern's New Approach to Static Economics", *Journal of Political Economy*, 54 (2): 97-115
- (1947), "Statistical Inference from Non-experimental Observation: An Economic Example", *Proceedings of the International Statistical Conferences*, vol III, 289-96, also Cowles Foundation Paper 32
- (1954), "Probability in the Social Sciences", in Lazarsfeld, P. (ed.), *Mathematical Thinking in the Social Sciences*, Free Press, 166-219, also Cowles Foundation Paper 82

Ménard, C. (1987), "Why was there no Probabilistic Revolution in Economic Thought?", in Kruger et al, 139-46

Morgan, M. (1987), "Statistics without Probability and Haavelmo's Revolution in Econometrics, in Kruger et al, 171-97

Morrison, M. (2002), Modelling Populations: Pearson and Fisher on Mendelism and Biometry, *British Journal for the Philosophy of Science*, 53: 39-68

Neyman, J. (1941), "Fiducial Arguments and the Theory of Confidence Intervals", *Biometrika*, 32: 128-150
- (1956), Note on an Article by Sir Ronald Fisher, *Journal of the Royal Statistical Society*, Series B (Methodological), 18 (2): 288-94
- (1957), "'Inductive Behavior' as a Basic Concept of Philosophy of Science", *Review of the International Institute of Statistics*, 25: 7-22

Pearson, K. (1894), *Socialism and Natural Selection*, reprinted in *The Chances of Death and other Studies on Evolution*
- (1897), *The Chances of Death and Other Studies in Evolution*, London: Edward Arnold
- (1908), *Memories of My Life*, London: Methuen
- (1914-30), *The Life, Letters and Labours of Francis Galton*, Cambridge: Cambridge University Press

Pearson, E. (1968), "Some Early Correspondence between W.S. Gosset, R.A. Fisher and Karl Pearson, with Notes and Comments", *Biometrika*, 55(3): 445-57

Porter, T. (2002), "Statistical Utopianism in an Age of Aristocratic Efficiency", *Osiris*, 17: 210-27

Reid, C. (1998), *Neyman*, New York: Springer

Savage, L. (1961), "The Foundations of Statistics Reconsidered", in Neyman, J. (ed), *Proceedings of the Fourth Symposium on Mathematical Statistics and Probability*, Berkeley: University of California, vol. 1, pp 575-86
- (1976), "On Rereading R.A. Fisher (with discussion)", *Annals of Statistics*, 4: 441-500

Schaffer, G. (2005), "'Like a Baby with a Box of Matches': British Scientists and the Concept of 'Race' in the Inter-war Period", *British Journal of the History of Science*, 38(3): 307-24

Schultz, G. (1928), *Statistical Laws of Demand and Supply with Special Application to Sugar*, Chicago: Chicago University Press

Seidenfeld, T. (1979), *Philosophical Problems of Statistical Inference – Learning from R.A. Fisher*, Dordrecht: Reidel

Spielman, S. (1973), A Refutation of the Neyman-Pearson Theory of Testing, *British Journal for the Philosophy of Science*, 24(3): 201-22

Wald, A. (1939), "Contributions to the Theory of Statistical Estimation and Testing Hypotheses", *Annals of Mathematical Statistics*, 10: 299-326

Yule, U. (1936), "Karl Pearson, 1857-1936", *Obituary Notices of Fellows of the Royal Society of London*, 2: 73-104

Zabell, S.L. (1992), "R.A. Fisher and Fiducial Argument", *Statistical Science*, 7(3): 369-87