II: NEYMAN AND THE THEORY OF STATISTICAL INFERENCE
BY M. S. BARTLETT

1. *Background*

Jerzy Neyman began his statistical career before he came to England in 1925; but it was his professional encounters with British statisticians, especially the Pearsons (father and son) and R. A. Fisher, that were to be the catalyst stimulating his interest in statistical problems, and in particular in the theory of statistical inference. Among his many contributions to statistical inference, the most important resulted directly from his personal collaboration with E. S. Pearson (cf. also [**R3**], and indirectly from the powerful stimulus of R. A. Fisher, about whom Neyman acknowledged [**III14**, p. ix]:

> 'Even though, in a quarter of a century long dispute, I combated certain views of Fisher, there is not the slightest doubt that his many remarkable achievements had a profound influence on my own thinking and work'.

Neyman's introduction to probability and statistics started, however, in Russia and Poland.† His teacher in probability was Serge Bernstein, who had suggested to Neyman that he read Karl Pearson's *Grammar of science.* In 1915, while still a student at the University of Kharkov, he studied measure theory at the instigation of another of his university teachers, C. K. Russyan, reading in particular Lebesgue's *Leçons sur l'intégration.* After the revolution, Neyman moved in the summer of 1921 to Poland, where he acquired a post as statistician at the Agricultural Research Institute in Bydgoszcz. His initiation into the use of statistical methods in agricultural experimentation might be compared with Fisher's when appointed in 1919 to a statistical post at Rothamsted Experimental Station, though Neyman's acceptance of the Polish post was a much more reluctant one forced on him by his lack of a job.

One or two papers published during this period earned Neyman a doctorate in 1923; and in 1925 he took up a post-doctoral Fellowship granted by the Polish National Culture Fund for study abroad. The first of his two years abroad was spent at University College London, in Karl Pearson's department; and here he first met Karl Pearson's son Egon, who had not long joined his father's staff. His encounter with E. S. Pearson appears at first to have stimulated Pearson more than Neyman, for Pearson refers [**R6**, p. 4] to a conversation at the end of Neyman's stay in London about a 'general statistical problem' which Pearson had been puzzling over, with the suggestion of future collaboration. Neyman, on the other hand, reports about gravitating to Paris in his second Fellowship year abroad, attending Lebesgue's lectures, and added [**I160**, p. 151]:

> '... were it not for Egon Pearson, I would have probably drifted to my earlier passion for sets, measure and integration, and returned to Poland as a faithful member of the Warsaw school, and a steady contributor to *Fundamenta Mathematicae.*'

---

† See David Kendall's biographical account for further details of Neyman's early education.

2. *The Neyman–Pearson collaboration*

However, E. S. Pearson, who had been struggling for some years trying to perceive the principles underlying the statistical methodology he had grown up with, and who found it helpful to formulate his difficulties at some length in correspondence with W. S. Gossett, J. Neyman and others, wrote a letter to Neyman apparently towards the end of 1926, on the logic of statistical tests. Neyman has noted [**I160**, p. 151]:

> 'I am afraid I was not very helpful in my reply. Frankly, prior to this letter of Pearson, and also for a period thereafter, I did not notice the existence of the problems that bothered him. Later on, ... I understood the issues. The general question was how to formulate the problem of statistical tests so that it would have a mathematical meaning. The problem was that of delineating the contents of mathematical statistics as a proper discipline'.

To understand the scientific climate at the time, it should be recalled that the *concept* of probability, on which statistical theory depended, was still somewhat nebulous. R. A. Fisher had just published the second of his two influential papers on statistical estimation [**R7, R8**], these being based on the objective statistical notion of probability as a relative frequency in a 'large' population, as noted by Fisher in a preface to his 1925 paper.

The *mathematical* theory of probability was put a few years later on a firm measure theory basis in the fundamental monograph by A. N. Kolmogorov [**R14**], and Neyman's own earlier studies made it particularly natural for him to adopt such a basis. The logical approach of writers like Harold Jeffreys to the general problem of inductive inference [**R12**], in which probability was regarded as a 'degree of reasonable belief' obeying certain axioms, was rejected both by Fisher and by Neyman. Nevertheless, it is important to distinguish between a valid mathematical basis and an acceptable logical basis; and Neyman's rigorous mathematical formulations of the principles of statistical inference did not *in themselves* ensure scientific validity and acceptance. The summary below of his main contributions includes commentary on their logical content.

To return first, however, to the cooperation between Neyman and Pearson, it seems clear that Pearson raised the problems and Neyman developed their mathematical formulation and solution. Fisher's theory of estimation, in which Fisher had stressed the role of the *likelihood function*, and of further important concepts such as *statistical information* and *sufficiency*, represented a considerable advance; but Pearson felt that the logic of *statistical tests* still required elucidation. Fisher tended to discuss significance tests purely on the basis of what he termed the *null hypothesis*, and it was Neyman and Pearson who first formulated an approach explicitly introducing the class of 'alternative hypotheses'. Pearson himself has acknowledged (cf. [**R3**, p. 433]) that the idea of introducing the class of alternatives was suggested in a letter from W. S. Gosset to him dated 11 May 1926.

At first the results were still somewhat tentative. In a long paper [**I14**] in 1928, Neyman and Pearson introduced (at Pearson's suggestion) their generalization of the likelihood ratio criterion to cover *composite* hypotheses, in which the null hypothesis H_0, say, is not fully specified.

Let us suppose observations $x_1, x_2, \ldots, x_n$ define a sample point Σ in the sample space W, with probability density $p_H(x_1, x_2, \ldots, x_n)$ depending on the hypothesis H. If p_H is completely specified by H, the hypothesis H is called *simple*. If its functional

form is known, but involves c unspecified parameters $\theta_1, \ldots, \theta_c$, it is termed a composite hypothesis with c degrees of freedom. The set of admissible simple hypotheses is denoted by Ω. The null hypothesis H_0 will belong to Ω if simple; if composite, it defines a subset ω of Ω. Let A_Σ be the set of probabilities p_H determined by simple hypotheses in Ω, and let the upper bound of A_Σ be denoted by $p_\Omega(\Sigma)$. If H_0 is composite, denote by $A_\Sigma(\omega)$ the subset of A_Σ corresponding to the subset ω, and by p_ω the upper bound of $A_\Sigma(\omega)$. Then the generalized likelihood ratio associated with the composite hypothesis H_0 is defined by

$$\lambda_1 = p_\omega(\Sigma)/p_\Omega(\Sigma).$$

In the special case when H_0 is simple, λ_1 reduces to $\lambda = p_0(\Sigma)/p_\Omega(\Sigma)$, where p_0 is the probability density for H_0.

The authors remark [**I14**, p. 265]:

> 'The value of this criterion in the case of testing composite hypotheses will perhaps be questioned. It may be argued that it is impossible to estimate the probability of such a hypothesis without a knowledge of the relative *a priori* probabilities of the constituent simple hypotheses. But in general it is quite impossible even to attempt to express our *a priori* knowledge in exact terms. Can we then get no further? Certainly we are faced with a problem whose solution cannot take the same form as that which is possible when a simple hypothesis is tested, yet we are inclined to think this ratio of maximum chances or frequencies of occurrence provides perhaps the best numerical measure that can be found under the circumstances to guide our judgement.
>
> 'As for λ, so in the case of λ_1, a knowledge of the ratio alone is not, however, sufficient. It has been pointed out that the errors of judgement which must inevitably occur will be of two kinds (1) we sometimes reject the hypothesis when it is in fact true, and (2) we sometimes accept it when Σ has been drawn from some population not belonging to the subset ω. The form of the criterion λ_1 has been chosen to minimize the effect of (2), but it is necessary to know also the probability distribution of λ_1 in sampling from the members of ω in order to control the source of error (1). In the cases which we have so far come across this distribution has possessed the essential property of being the same or approximately the same whatever member of ω may have been sampled'.

In their 1933 *Phil. Trans.* paper [**I33**], the authors attempted a more formal and systematic development; and though the nature of the problem prevented a complete solution in the framework of sampling theory they were able to obtain an imposing set of general results. Their own summary includes the following [p. 185 of **III15**]:

> '1. A new basis has been introduced for choosing among criteria suitable for testing any given statistical hypothesis, H_0, with regard to an alternative H_1. If θ_1 and θ_2 are two such possible criteria and if in using them there is the same chance, ε, of rejecting H_0 when it is in fact true, we should choose that one of the two which assures the minimum chance of accepting H_0 when the true hypothesis is H_1.
>
> 2. Starting from this point of view, since the choice of a criterion is equivalent to the choice of a critical region in multiple space, it was possible to introduce the conception of the best critical region with regard to the alternative hypothesis H_1. This is the region, the use of which, for a fixed

value of ε, assures the minimum chance of accepting H_0 when the true hypothesis is H_1. The criterion, based on the best critical region, may be referred to as the most efficient criterion with regard to the alternative H_1.

3. It has been shown that the choice of the most efficient criterion, or of the best critical region, is equivalent to the solution of a problem in the Calculus of Variations. We give the solution of this problem for the case of testing a simple hypothesis.

To solve the problem in the case where the hypothesis tested is composite, the solution of a further problem is required; this consists in determining what has been called a region similar to the sample space with regard to a parameter.

We have been able to solve this auxiliary problem only under certain limiting conditions; at present, therefore, these conditions also restrict the generality of the solution given to the problem of the best critical region for testing composite hypotheses'.

The quantity ε above was termed the *size* of the test. The Neyman–Pearson *Fundamental Lemma* established that the chance of rejecting H_0 when a single possible alternative hypothesis H_1 was true was maximized in the case of simple alternatives when the critical region was based on λ. In a subsequent paper [**I34**] this chance was called the *power* of the test, and the test based on λ the most powerful test.

Further elaboration, including the notions of *uniformly most powerful tests* [**I34**], 'unbiased' tests [**I51**], in which the power is never less than the size ε, etc., resulted in a considerable theoretical superstructure which overall made quite an impact. The unequivocal mathematical basis for the theory rendered it in particular attractive to many university teaching departments of statistics, and, in due course, to writers of textbooks. Nevertheless Fisher, while apparently a referee of the 1933 paper†, became one of its strongest critics. This may partly have been because he considered the development of the Neyman–Pearson theory of testing statistical hypotheses *ab initio* tended to obscure the fundamental role of his own concepts of *likelihood* and *sufficiency*, even though the importance of the former had been vindicated by the Neyman–Pearson Fundamental Lemma, and the latter, if appropriately extended, was related to their idea of similar regions. However, a more basic difference of viewpoint was to emerge, associated with the notion of 'repeated sampling'.

This is a fundamental problem, which will arise again later, as it impinges strongly on Neyman's subsequent work and general attitude. For the moment this problem will be discussed, not in relation to Fisher's criticisms, as Fisher was far from consistent or objective in these, but to the critique by I. Hacking [**R11**]. Hacking notes that uncritical appeal to the concepts of size and power alone (the foundation of the Neyman–Pearson theory) is inadequate, and could lead to absurdity—for example, if an event with zero probability on H_0, but non-zero on H_1, has been included in the acceptance region for H_0, but has actually occurred. Hacking distinguishes between the value of tests *before* and *after* a trial has been carried out. The Neyman–Pearson theory is very valuable for 'before-trial' assessments, but not necessarily so relevant after the results of the trial are known (unless it has been decided to ignore the further details of the data, *given* that the sample point has fallen in the acceptance region). While the above counter-example

† See, however, Constance Reid's remarks in her biography *Neyman—From life* (Springer-Verlag, 1982, p. 104).

could be judged inadmissible, as its acceptance region corresponds to a test in conflict with the Fundamental Lemma, more subtle examples of *conditional* inference, in which the sampling distribution envisaged is conditional on what Fisher termed *ancillary* information, cannot readily be discussed within the Neyman–Pearson framework.

The 1933 paper [**I34**] on the testing of statistical hypotheses in relation to probabilities *a priori* is of interest not only for its introduction of the concept of *power*, but also for its recognition of problems where the economic consequences of decisions may be sufficiently relevant to be introduced explicitly into the analysis, thus being to some extent a forerunner of 'statistical decision theory'.

3. *The theory of confidence intervals*

A few years after the development of the Neyman–Pearson theory of statistical tests, Neyman published another paper [**I56**] in *Phil. Trans.* on interval estimation. This set out systematically his own theory of 'confidence intervals', and had arisen from attempts to describe the accuracy of estimates of parameters independently of the formal Bayes solution involving the assignment of prior probabilities. The solution may be illustrated in terms of a simple and familiar example. Suppose a sample of n independent observations $x_1, x_2, \ldots, x_n$ from a normal population with true (but unknown) mean μ has sample mean m with variance $\sigma_m^2 = \sigma^2/n$, where σ^2 is the (known) variance per observation. Then we know, say, the probability 0.95 that

$$M - \mu < 1.96\sigma_m$$

(or alternatively the probability 0.90 that $|M-\mu| < 1.96\sigma_m$) from the sampling distribution of the *random* quantity M, of which m is the realized value. If therefore we assert that $\mu > M - 1.96\sigma_m$ (or alternatively that $|\mu - M| < 1.96\sigma_m$), we shall be correct *in repeated sampling* on 95% of occasions (in the alternative 'two-tailed' version, on 90% of occasions).

The misunderstandings that such statements generated were remarkable, but were partly due to ambiguous discussion in this context by Fisher, who had already in 1930 published a note [**R9**] in which he first proposed what appeared to be the above form of statement under the name of *fiducial probability*. Unfortunately, while Fisher had always been careful to distinguish between a population parameter such as μ, and a sample estimate such as m, he did *not* distinguish in his notation between the observed value m on a particular occasion, and the random quantity M, of which m is a realization. At the time his discussion in terms of repeated sampling seemed, however, clear enough. For example, in the 1930 note when discussing the case of an unknown correlation coefficient ρ, Fisher said [p. 535]:

> '... if we take a number of samples ... from the same or different populations and for each calculate the fiducial 5 per cent value of ρ, then in 5 per cent of cases the true value of ρ will be less than the value we have found'.

Both Neyman and Pearson have referred to other early discussion on the problem of interval estimation. In 1927 Pearson was already engaged in correspondence (I am indebted to Professor G. A. Barnard for showing me a copy of this correspondence) with W. S. Gosset on intervals defined by relevant significance tests, and Neyman [**I160**, p. 154] refers to a question by Pytkowski, a student of his

in Warsaw, about 1930, on the precision of an estimated regression coefficient, which stimulated his own theory of confidence intervals. However, Fisher's 1930 note appears to have priority in print, and Neyman many years later has acknowledged [**I160**, p. 155]:

> 'in the first paper in which I presented the theory of confidence intervals, published in 1934 [**I35**], I recognized Fisher's priority for the idea that interval estimation is possible without any reference to Bayes's theorem and with the solution being independent from probabilities *a priori*'.

It was not until later that controversy over the interpretation of Fisher's theory was accentuated by my own criticism (on the assumption that a valid frequency interpretation was intended) of some of Fisher's extensions of his fiducial theory to more than one parameter [**R1**]. This controversy led to eventual general acceptance that Fisher's fiducial theory and Neyman's theory of confidence intervals must be distinguished. Thus in a critical review in 1941 [**I69**] Neyman concluded:

> 'the only thing that the present author ventures to profess is that the theory of fiducial probability is distinct from that of confidence intervals'.

Nevertheless, I was always rather surprised to find no explicit reference to Fisher's 1930 note in the 'review of the solution of the problem of estimation advanced hereto' in Neyman's 1937 *Phil. Trans.* paper [**I56,** and p. 258 in **III14**].

Fisher's equivocal stance was not entirely without substance. In a Bayesian approach the inference on an unknown mean μ would be conditional on the observed m, though at the cost of introducing 'probabilities *a priori*'. Hacking [**R11**, p. 159], when discussing Neyman's theory of confidence intervals, remarks that

> '... it may be helpful to examine its relation to the fiducial argument. This is especially necessary, since Fisher's original expression of the fiducial argument is couched in terms more suitable to Neyman's theory than his own'.

What Hacking in effect emphasizes, in terms of our specific example, is that an inference in terms of the observed mean m, as distinct from the random M, implies a 'principle of irrelevance.' Indeed Jeffreys claimed [**R13**]:

> 'In fact the fiducial argument when completed, and the inverse probability argument, are simply different ways of saying the same thing'.

Hacking, however, did not go as far as this, pointing out that while the irrelevance principle might *imply* a 'prior distribution', this did not mean the whole-hearted acceptance of the machinery of prior distributions involved in the Bayesian approach. Neyman defended his theory of confidence intervals purely in terms of 'long-run behaviour'; but, while this is a consistent attitude, it is difficult (as noted for the Neyman–Pearson theory of tests) to uphold a *particular* inference in such terms if the inference is obviously false on that particular occasion. To put it bluntly, an engine driver perceiving a large obstruction on the railway line ahead naturally applies his brakes, and does not argue that if he ignores such obstructions he will be correct on most occasions. The difficulty here is that a statistical inference as such cannot be expected to cope efficiently with a unique, or even a very rare class of, event, but becomes merely part of a wider type of inductive inference which Neyman might reasonably have claimed was no longer classifiable within the discipline of *statistical* inferences.

4. *Other papers on statistical methodology and inference*

In addition to the main work on hypothesis testing and confidence intervals, Neyman published some other important papers during his period in England as a member of staff in E. S. Pearson's new Department of Statistics. These included a paper in 1934 [**I35**] on sampling theory presumably largely prepared while Neyman was still in the Biometric Laboratory, Warsaw, which is given as his address; this paper has already been cited in §3 for its acknowledgment of Fisher's theory of fiducial probability. Perhaps because of this, Fisher's attitude to the paper was reasonably amicable, in sharp contrast to the controversial exchanges following Neyman's second paper [**I43**] to the Royal Statistical Society in 1935, in cooperation with K. Iwaszkiewicz and S. Kołodziejczyk, on statistical problems in agricultural experimentation. In this paper Neyman had the temerity to question the theory underlying the statistical analysis of randomized blocks and Latin Squares, pointing out in particular that if the treatments reacted differently to differences in soil fertility, then the standard test of significance of treatment effects could be biased in the case of Latin Squares. Fisher's defence was to emphasize his own null hypothesis that the treatments had *no* effect on the yields; and, like the more permanent controversy that was to develop over the rival merits of fiducial and confidence intervals, there was some right on both sides. As I myself noted at the time [**I43**, p. 169]:

> 'it would ... have been simpler if Dr. Neyman had stated in words exactly why it was that the Latin Squares appeared to be giving a biased estimate of error; he had assumed that the treatments were having an effect on the yields, and if one made the "null hypothesis", that was not necessary for a test of significance. It was only fair to Dr. Neyman to mention that although on the null hypothesis the z test would be unbiased for testing the effects of the treatments, one would still get the possibility of bias in a t test for testing the difference between only two treatments, since usually one would be using the whole of the error for that particular comparison'.

In later years Neyman continued to publish on theoretical aspects of statistical inference, but his most important contributions in this area had been completed. One paper [**I76**] published in 1949, but dealing with work started several years earlier, developed the theory of what Neyman called 'best asymptotically normal estimates', which were a generalization of Fisher's maximum likelihood estimates having the same asymptotic properties but intended to be more convenient to use. In 1959 [**I106**; cf. also **90**] he developed the theory of what he termed C(α) tests, which were asymptotically optimal for testing a null hypothesis $\theta_1 = 0$, say, against the alternative class $\theta_1 \neq 0$, when other parameters $\theta_2, \ldots, \theta_k$ were unknown.

Suppose a sample of n independent observations $x_1, \ldots, x_n$ has density function

$$\prod_{i=1}^{n} f(x_i \mid \theta_1, \ldots, \theta_k)$$

with log likelihood function $L = \sum_{i=1}^{n} \log f(x_i)$. Neyman considers the statistic

$$T = (\partial L/\partial \theta_1) - \sum_{j=2}^{k} \beta_j (\partial L/\partial \theta_j)$$

evaluated at $\theta_1 = 0$, $\theta_j = \hat{\theta}_j$ $(j > 1)$, where $\hat{\theta}_j$ are maximum likelihood estimates (more generally, estimates with what Neyman termed 'local root-n consistency'), and β_j are the regression coefficients of $\partial L/\partial\theta_1$ on the $\partial L/\partial\theta_j$, as estimated from the second-order derivatives. Neyman's procedures were, however, subsequently examined by P. A. P. Moran [**R17**], who showed that they were asymptotically equivalent to the use of the likelihood ratio test and to tests using maximum likelihood estimators, and noted that they had also previously been discussed by myself [**R2**]. Moran also noted the generalization to the case where the null hypothesis involved more than one parameter e.g. $\theta_1 = \theta_2 = \ldots = \theta_r = 0$ $(1 < r < k)$.

Neyman's wide-ranging contributions to applied fields are referred to elsewhere in this memoir; but within the framework of the general theory of statistical inference it might be added that his avowedly behaviouristic approach to inductive problems [e.g. **I132**] did not inhibit his emphasis on the need for specific stochastic models applicable to the particular problem under investigation. As his interest in applications developed, model-building became an even more important aspect of his researches. At the end of his own review of his illustrious and influential research career he notes [**I160**, p. 162]:

> 'particularly in the more recent decades, the delight I experience in trying to fathom the chance mechanisms of phenomena in the empirical world'.

III. Neyman as astronomer
by Thornton L. Page

I'd like to contribute as much as possible to the Royal Society Memoir on Jerry Neyman. He was a true friend, as well as my statistical mentor. As you may remember, he put me up in his house on Amherst Avenue; then, after I suffered a serious auto accident, arranged for me to spend a quarter doing research and editing several books at Berkeley. I have strong memories of his humour and devotion to work. One evening he joined us for supper in our Berkeley apartment, and I served herring snacks on crackers with 6 o'clock drinks. Jerry said, 'Oh, I like dead fish', and herring snacks have been 'dead fish' to us ever since. When we were working on the statistics of double galaxies, he would often terminate supper with 'It's time to get to *work*', which he pronounced to rhyme with York.

Jerry was intensely Polish. I first got to know him after the I.A.U. Congress in Moscow when we (together with 16 other western astronomers) toured southern Russia, visiting several Soviet observatories. At one point we were taken to a Soviet resort on the Black Sea for swimming and boating. I was swimming while Jerry and several others drank coffee in a small seaside restaurant. When I joined them, in my bathing suit, the Soviet police moved in and arrested me. It seems that there is a strict Soviet law against entering a public restaurant in a bathing suit. Jerry persuaded them that I was just a dumb American, but told me that I must change my clothes then and there under police escort. I was embarrassed as I did this in full view of several lady astronomers, but it satisfied the police.

On another occasion, Neyman showed his understanding of Russians and their language. At each observatory we were given a ceremonial luncheon highlighted by toasts in wine by the hosts 'to our guests from overseas', and by one of us 'to our hospitable hosts'. We took turns making the responses, and in Burakan, Soviet