NOTE ON AN ARTICLE BY SIR RONALD FISHER

By JERZY NEYMAN

Department of Statistics, University of California, Berkeley, California

[Received January, 1956. Revised June, 1956.]

Summary

(1) FISHER's allegation that, contrary to some passages in the introduction and on the cover of the book by Wald, this book does not really deal with experimental design is unfounded. In actual fact, the book is permeated with problems of experimentation. (2) 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. (3) The conceptual fallacy of the notion of fiducial distribution rests upon the lack of recognition that valid probability statements about random variables usually cease to be valid if the random variables are replaced by their particular values. The notorious multitude of "paradoxes" of fiducial theory is a consequence of this oversight. (4) The idea of a "cost function for faulty judgments" appears to be due to Laplace, followed by Gauss.

1. *Introduction*

In a recent article (Fisher, 1955), Sir Ronald Fisher delivered an attack on a substantial part of the research workers in mathematical statistics. My name is mentioned more frequently than any other and is accompanied by the more expressive invectives. Of the scientific questions raised by Fisher many were sufficiently discussed before (Neyman and Pearson, 1933; Neyman, 1937; Neyman, 1952). In the present note only the following points will be considered: (i) Fisher's attack on the concept of errors of the second kind; (ii) Fisher's reference to my objections to fiducial probability; (iii) Fisher's reference to the origin of the concept of loss function and, before all, (iv) Fisher's attack on Abraham Wald.

2. *Attack on Abraham Wald*

On page 70 of the article, Sir Ronald writes:

> "But first I must exemplify ... by quoting some rather simple phrases from Wald's book on Decision Functions." (Wald, 1950.)
>
> " On the outside of the cover we read, 'Particularly noteworthy is the treatment of experimental design as a part of the general decision problem'.
>
> "On the inside, 'The design of experimentation is made a part of the general decision problems—a major advance beyond previous results ...'
>
> "These claims seem very much like an afterthought, of a kind which is sometimes suggested by a publisher; for, apart from these three quotations, the design of experiments is scarcely mentioned in the rest of the book. For example, the index does not contain the word 'replication', or 'control', or 'randomization'; ... Of authorities, the bibliography does not contain the names of Yates, of Finney, or of Davies; ... My own book is indeed mentioned, but no use seems to have been made of it. The obvious inference is that Wald was quite unaware of the nature and scope of the subject of experimental design, but had simply assumed that it *must* be included in that of acceptance procedures, to which his book is devoted."

First, the 167-page book of Wald is literally permeated with problems of experimentation. The words "experimentation", "experiments", and "experimenter", are found on the following pages: 2–14, 17–22, 26, 28, 29, 31, 59, 63–65, 67, 68, 86, 88, 89, 92–94, 103, 105, 107, 111, 115, 119, 120, 123, 124, 130, 131, 138, 147, 151, 152, 156, 157, 159, 160, 162, 164–166, i.e. on 58 pages

out of 167. This enumeration does not include pages where instead of "experiment" or "experimentation" a synonym is used, such as "taking of observations", etc. As to Sir Ronald's regret that his own book is merely mentioned by Wald, the reason is of course that the problems of experimental design treated by Wald, as well as his method and mathematical level, are entirely different from those of Fisher.

In Wald's case the problem of experimental design is understood to mean the search for optimum procedures to design not just one experiment but a sequence of them which, in principle, could be continued indefinitely. At each stage, the experimentation may be either discontinued and closed with a "terminal decision", or resumed, with proper adjustments dictated by what was found in the earlier stages. On the other hand, Fisher's own work is chiefly concerned with procedures relating to what in Wald's treatment are separate single stages in experimentation. On p. 19 Wald shows how, by imposing a number of restrictions including the restriction that the experimentation is to be terminated after the first stage, his general problem is reduced to Fisher's Latin Square design. The relation between the two schools of thought might be compared to that between tactics (Fisher) and strategy (Wald).

As far as I am aware, although the persons listed by Fisher as "authorities" have made valuable contributions applying, extending and perfecting Sir Ronald's ideas on the tactics of the design of experiments, none of them ever contributed anything that is conceptually different. Also, I do not remember seeing any contribution of these persons to the problems of strategy in experimentation as raised by Wald. For these reasons, it appears perfectly natural that these authors are not mentioned by Wald. On the other hand, there are a number of references to papers by David Blackwell, M. A. Girshick, E. L. Lehmann, C. M. Stein, and, particularly, J. Wolfowitz who, at the time when the book was written in the late 1940's, could perhaps be considered as authorities in the problems of experimental design raised by Wald. At the present time still another book, by Blackwell and Girshick (1954), should be mentioned.

The new school of thought symbolized by the above names grew and bypassed Sir Ronald, without his noticing the fact, and now he appears to be entirely unaware of the role of Wald in the history of statistics. Wald worked in statistics but a very brief time, from about 1937 to 1950, when he perished in an aeroplane accident. However, during this period he managed to advance the theory of statistics tremendously. Also, his results on the design of experiments, incorporated as they are in the general theory of statistical decision functions, appear to merit a separate term, perhaps "Wald's theory of experimentation". The reverberations of Wald's work are currently felt all over the world, including the British Isles. However, Wald's approach happens to be characterized by the phrase "statistical decisions" which he used as a substitute for the earlier "inductive behavior". And, as witnessed by the article of Fisher under discussion, the latter phrase, or the point of view behind it, is something that Sir Ronald disapproves of, forcefully and vociferously. Since he is faithfully seconded by some of his followers (particularly by Dr. Frank Yates) who are almost equally forceful, this creates a very curious atmosphere, exemplified by the following passage (Box and Anderson, 1955, p. 34).

> "The difficulties that face us in attempting to reply to Dr. Yates' final point are similar to those which must be experienced by neutral countries who try to keep out of the 'cold war'. We suspected that if we included statements by Fisher and Pearson in the same paragraph, even though these were about different things, this could lead to a discussion generating more heat than light. We therefore were at some pains to explain that the points of view we quoted were complementary rather than contradictory. Alas, we are now not only accused of saying that Pearson and Fisher agree but in the very next sentence are charged with 'gross misrepresentation' for saying (which we did not) that Fisher and Pearson do not agree about the importance of powerful tests. Against this double-edged weapon no defence is possible and we admit defeat."

3. "*Errors of the Second Kind*"

The contents of Fisher's section under the above title might be exemplified by the following passages:

> "The phrase 'errors of the second kind' although apparently a harmless piece of technical jargon, is useful as indicating the type of mental confusion in which it was coined . . .

> Such errors are therefore incalculable both in frequency and in magnitude merely from the specification of the null hypothesis, and would never have come into consideration in the theory only of tests of significance, had the logic of such tests not been confused with that of acceptance sampling."*

Errors of the second kind came under study in the joint work of E. S. Pearson and myself, because without them no purely probabilistic theory of tests is possible. This was first proved in 1929 (Neyman, 1930) and the proof, with some extensions, was later reproduced in a book (Neyman, 1952, pp. 43–47). The concept of errors of the second kind has led to considering probabilities of committing or of avoiding these errors (that is, to the power of tests). Since that time (Neyman and Pearson, 1933), these concepts have become the cornerstone of what some authors are kind enough to call the Neyman-Pearson theory of testing statistical hypotheses.†

The set of probabilities of errors of the second kind in using a given test or, equivalently, power, has three important applications. First, these concepts serve as a basis for the *deduction* of tests that are the most powerful either absolutely or compared with a specified class of tests. Second they serve as a means of comparison and evaluation of two or more suggested alternative tests. In particular, the power of nonparametric tests is frequently evaluated for this specific purpose. Third, 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.

How deeply these concepts have penetrated modern statistical thinking might be illustrated by the fact that the same issue of the Society's Journal that carries the article by Fisher now under discussion contains a paper (Foster and Teichroew, 1955), specifically dealing with the power of certain tests, and another paper (Box and Anderson, 1955) in which there is a section on power.

As Sir Ronald remarks correctly, merely from the specification of the null hypothesis, the probabilities of errors of the second kind are certainly not calculable. However, the main point of the modern theory of testing hypotheses is that, for a problem to make sense, its datum must include not only a hypothesis to be tested, but in addition, the specification of a set Ω of alternative hypotheses that are also considered admissible. Also, ordinarily,‡ when a scientist designs an experiment he, consciously or subconsciously, has in mind at least a general outline of the set of admissible hypotheses, against which a particular hypothesis is to be tested. This general outline serves then as a lead towards a precise definition of the set Ω. With this kind of datum, the probabilities of errors of the second kind are certainly calculable and, in fact, a considerable number of tables of such probabilities, or of their equivalents, namely of power functions, are now available.

Another item worth mentioning in Sir Ronald's section on errors of the second kind is the passage: "These examples show how badly the word 'error' (of the second kind) is used in describing the situation. Moreover, it is a fallacy so well known as to be a *standard* example, to conclude from the test of significance that the null hypothesis is thereby established . . ." Although no names are mentioned, the context suggests that the fallacy in question is committed by the same people who are guilty of considering alternative hypotheses and the power of tests. Whereas Sir Ronald abstains from quoting any specific instance, it is easy to quote those in which he himself, and his followers, acted precipitately and advised others to do likewise, when a test failed to detect a significant effect.

A case in point is the design of the factorial experiment which, by itself, represents one of the most valuable inventions of Fisher. However, in the early days (Fisher and Wishart, 1930, pp. 17–23; Fisher, 1933, p. 6; Yates, 1935) factorial field experiments were recommended with a very small number of replicates, three, two, or even one. The procedure of analysis prescribed that one test for significance of interactions and then, if the interactions do not prove significant,

* The earlier objections of Sir Ronald were even more sweeping than the present: "The 'error', so called, of accepting the null hypothesis 'when it is false', is thus always ill defined both in magnitude and frequency".

† An excellent systematic exposition of this theory is available in the recent book by Schmetterer (1956).

‡ There are, of course, exceptions to this general rule. When such an exception occurs, I am inclined to think that the experimenter is at fault in not visualizing exactly what he intends to discover by means of his experiment. A case in point is Fisher's famous experiment (Fisher, 1937) concerned with the Lady Tasting Tea. Specifically, the first of the designs contemplated appears to be open to criticism. The details have been explained elsewhere (Neyman, 1950).

that one act as if it were known that the interactions are non-existent. Sir Ronald's attitude (Fisher and Wishart, 1930) is illustrated by the following: "The interactions of nitrogen and phosphate are not significant, and it therefore becomes clear how we are to present the results". This is seconded by Yates thus (Yates, 1935), "In general, if there is no evidence of interaction the mean responses to the two factors may be taken ('may be taken'—is this some sort of 'inductive reasoning' or a decision to take a certain action?) as the appropriate measures of the responses of these factors, which may be regarded as additive". On the other hand, both authors are frequently emphatic that interactions are of common occurrence: "Cases in which the interactions are certainly negligible are, in fact, rather rare". (Yates, *ibid.*, p. 211.)

The problem before us is, then, as follows. Given that interactions between treatments occur from time to time, given that, if the interactions are not found significant, then the treatments will be judged by their "average responses", and given a typical value of the error variance per plot to decide whether it is safe to lay down a trial with only three replications. The reader will realize that this is exactly a problem in which the numerical values of the probabilities of errors of the second kind are very important.

The solution of the problem was provided at the discussion of Dr. Yates's paper (Neyman, 1935). It was shown, that, if, in conditions comparable to Yates's experiments, factorial trials are performed with only three replications, then the frequency of entirely misleading results, caused by errors of the second kind, may be out of all proportion to the level of significance adopted. In these circumstances it was a pleasure to find that in his "The Design of Experiments" (Fisher, 1937) Sir Ronald contemplates factorial field trials with six replicates rather than with three.

The above example illustrates two points: the usefulness of calculated numerical probabilities of errors of the second kind and the fact that the "fallacy, so well known as to be a standard example" was committed by Sir Ronald himself when he advised the assessment of treatments by their "average responses" when the interactions are not found significant.

4. *Objections to the Notion of Fiducial Probability*

In his section under the title "Inductive Behavior" Sir Ronald writes: "A complementary doctrine of Neyman violating equally the principles of deductive logic is to accept a general symbolical statement such as

$$Pr\{(\bar{x} - ts) < \mu < (x + ts)\} = \alpha, \quad . \quad . \quad . \quad . \quad (1)$$

as rigorously demonstrated, and yet, when numerical values are available for the statistics $\bar{x}$ and s, so that on substitution of these and use of the 5 per cent. value of t, the statement would read

$$Pr\{92{\cdot}99 < \mu < 93{\cdot}01\} = 95 \text{ per cent.}, \quad . \quad . \quad . \quad . \quad (2)$$

to *deny* to this numerical statement any validity. This evidently is to deny the syllogistic process of making a substitution in the major premise of terms which the minor premise establishes as equivalent".

The discussion that follows refers specifically to the alleged violation of "principles of deductive logic". The last sentence quoted makes it clear that to Sir Ronald the relation between (1) and (2) is the same as that between

$$AB = BA, \quad . \quad . \quad . \quad . \quad . \quad . \quad (1')$$

a formula established for all *numbers*, and the result of substituting, say, $A = 2$ and $B = 3$,

$$2 \times 3 = 3 \times 2. \quad . \quad . \quad . \quad . \quad . \quad . \quad (2')$$

Naturally, if (1′) is established for *all numbers*, then it must be true irrespective of what *numbers* we substitute for the letters A and B. But what if somebody attempts to substitute for A and B *not numbers* but some other mathematical entities for which multiplication is defined, for example matrices? Then, and this is well within the domain of deductive logic, the result of substitution need not be true. This is precisely the case with the original formulae under discussion, (1) and (2).

With the appropriate meaning of the symbols used, Fisher's assertion that while accepting (1) I deny the validity of (2) is perfectly correct. Equality (1) is rigorously demonstrated on the

assumption that $\bar{x}$ and s are *random variables* with a strictly defined distribution depending on μ. On the other hand, the expression within the braces in (2) depends on "sure numbers", not on random variables. It is true that every sure number may be considered as a random variable with a "degenerate" distribution. However, the particular distribution of $\bar{x}$ and s the postulation of which leads to (1) is not degenerate. Thus (2) is certainly not a particular case of (1). The details of the situation are explained elsewhere (Neyman, 1937; 1952). When these explanations did not suffice to convince Sir Ronald of his mistake, I was tempted to give up. However, in a private conversation David Blackwell suggested that Fisher's misapprehension may be cleared up by the examination of several simple examples. They illustrate the general rule that *valid probability statements regarding relations involving random variables may cease and usually do cease to be valid if random variables are replaced by their observed particular values.** Thus, if X is a normal random variable with mean zero and an arbitrary variance greater than zero, then I expect Sir Ronald to accept the "general symbolical statement"

$$Pr\{X < 0\} = \tfrac{1}{2}$$

as rigorously demonstrated. Yet, presumably, when a numerical value $+ 1{\cdot}72$ is available for X, so that on substitution the statement would read

$$Pr\{+ 1{\cdot}72 < 0\} = \tfrac{1}{2},$$

Sir Ronald will join me in denying this statement any validity.

Similarly, if Y is a binomial variable, say the number of "heads" obtained in $n = 2$ independent throws of a fair coin, then the "general symbolical statement"

$$Pr\{Y = 1\} = \tfrac{1}{2} \quad . \quad . \quad . \quad . \quad . \quad . \quad (3)$$

is rigorously demonstrated. However, if upon observing two heads in two throws we substitute in (3) $Y = 2$, we obtain the nonsensical result

$$Pr\{2 = 1\} = \tfrac{1}{2},$$

etc.

Sir Ronald's mistaken opinion that (1) implies (2) is at the base of what he calls "fiducial distributions". The conceptual fallacy of this notion was pointed out quite early (Neyman, 1934). However, it took a problem of some complexity to bring out the numerical consequences of the mistake. This was first done by Bartlett (1936, 1939). Bartlett's revelation that the frequencies in repeated sampling (from the same or from different populations conforming with the hypothesis tested) need not agree with Fisher's solution of the Fisher-Behrens problem brought about an avalanche of rebuttals by Fisher and by Yates. Because of Fisher's remarkable talent in polemics, the analysis of his statements would take an excessive amount of space. On the other hand, some of Yates's (Yates, 1939) assertions are frank, simple and quite revealing:

> "To avoid this objection we must frankly recognize that we have here introduced a new concept into our methods of inductive inference, which cannot be deduced by the rules of logic from already accepted methods . . . The approach by inverse probability enables fiducial statements about μ to be derived from the classical theory of probability, without the introduction of any new principle, but only at the cost of postulating a particular *a priori* distribution of σ. In the fiducial approach such *a priori* postulation is regarded inadmissible, but in order to discard it a new principle, that of utilizing the fiducial distribution of σ must be introduced . . . If the principle is not accepted, then it appears that we must either assume an *a priori* distribution of σ, or deny that there is any possibility of making fiducial statements about μ."

But just a few lines before, Yates writes that the assumption of a fiducial distribution of σ "is open to objection on the ground that in subsequent samples σ may in fact be distributed in any manner . . ." Thus, already in the late 1930's the whole basis of fiducial argument in general

*There are trivial exceptions to this rule. Thus, for example, the equality $P\{(X = 0)(X = 1)\} = 0$ remains true whether X is a random variable or an arbitrary number.

and of the Behrens-Fisher test in particular appeared to be all tied up in knots. The present situation is certainly no better (Breny, 1955). In the course of time more problems came under consideration and the words "Paradox of fiducial theory" became a familiar feature in the statistical literature. Indeed, an article with these words in the title (Mauldon, 1955) immediately follows the paper by Sir Ronald Fisher now discussed. Several other "paradoxes" were mentioned at the Symposium on Interval Estimation recently held before the Research Section. A particular problem was treated independently by Mr. Fieller (1954) and by Miss Creasy (1954), the latter working under the guidance of Dr. Finney. In both cases efforts were made to follow exactly the rules of fiducial argument, but the solutions arrived at were different. Sir Ronald himself took part in the discussion and also produced a solution. Unfortunately, at least one of the main authors could not understand Sir Ronald and frankly admitted that this was the case. Another authority, whose familiarity with fiducial theory goes back to 1930, recounted that after much thought and discussion with three other scholars, he arrived at a tenative conclusion that both solutions, one by Miss Creasy and the other by Mr. Fieller, may be right but that they refer to different parameters.

Taking into account that the problem under discussion was relatively trivial, that of the ratio of means of two normal populations, it must be clear that no such diversity of opinion among scholars could have arisen if the subject of their discussion was clear. Actually, several participants in the discussion called for a clear-cut operational definition of fiducial distribution. However, both Miss Creasy and Mr. Fieller, while insisting that their own deductions were valid, seemed to think that no definition is necessary. Mr. Fieller's final comment was: "I personally have no difficulty in deciding that ... Miss Creasy's limits are not what I call 'fiducial limits'; I merely note that they are not quite the same as mine".

Thus, given the conditions of a problem and given a function, there appears to be no operational means for determining whether this function is a fiducial distribution or not. All one can do is to apply to an authority. But then the authority will only decide whether the function is *what he calls* a fiducial distribution. However, what if, as in the present case, several authorities are applied to and if they disagree?

It is doubtful whether the chaos and confusion now reigning in the field of fiducial argument were ever equalled in any other doctrine. The source of this confusion is the lack of realization that equation (1) does not imply (2).

5. *Origin of the Concept of "Cost Function for Faulty Judgments"*

The main objection of Fisher to the work of "Neyman, Pearson, Wald and Bartlett" is that it is allegedly based on an attitude identifying ordinary research with industrial acceptance sampling, an attitude marked by the concept of "cost function for faulty judgments", appropriate to Russian and American mentalities but one that Sir Ronald finds objectionable.

It happens that the first scholars to think of "cost function for faulty judgments" appear to be Laplace and then Gauss. Since this circumstance has been overlooked by many contemporary authors, including E. S. Pearson and myself (Neyman and Pearson, 1933, p. 502) it seems appropriate to quote the relevant passage from Gauss (1887, pp. 5–6).

> "Die Bestimmung einer Grösse durch eine einem grösseren oder kleineren Fehler unterworfene Beobachtung wird nicht unpassend mit einem Glücksspiel verglichen, in welchem man nur verlieren, aber nicht gewinnen kann, wobei also jeder zu befürchtende Fehler einem Verluste entspricht. Das Risiko eines solchen Spieles wird nach dem wahrscheinlichen Verlust geschätzt, d. h. nach der Summe der Produkte der einzelnen möglichen Verluste in die zugehörigen Wahrscheinlichkeiten. Welchem Verluste man aber jeden einzelnen Beobachtungsfehler gleichsetzen soll, ist keineswegs an sich klar; hängt doch vielmehr diese Bestimmung zum Theil von unserem Ermessen ab. Den Verlust dem Fehler selbst gleichzusetzen, ist offenbar nicht erlaubt; würden nämlich positive Fehler wie Verluste behandelt, so müssten negative als Gewinne gelten. Die Grösse des Verlustes muss vielmehr durch eine solche Funktion des Fehlers ausgedrückt werden, die ihrer Natur nach immer positiv ist. Bei der unendlichen Mannigfaltigkeit derartiger Funktionen scheint die einfachste, welche diese Eigenschaft besitzt, vor den übrigen den Vorzug zu verdienen, und diese ist unstreitig das Quadrat. Somit ergiebt sich das oben aufgestellte Princip.

"*Laplace* hat die Sache zwar auf eine ähnliche Weise betrachtet, er hat aber den immer positiv genommenen Fehler selbst als Maass des Verlustes gewählt. Wenn wir jedoch nicht irren, so ist diese Festsetzung sicherlich nicht weniger willkürlich, als die unsrige . . . "

These considerations were later resumed on British soil by Edgeworth. Edgeworth's relevant paper (Edgeworth, 1908) should be well known to Sir Ronald. In fact, in this paper Edgeworth anticipated a number of Fisher's results concerned with maximum likelihood estimates and this circumstance was called to Fisher's attention by Sir Arthur Bowley twenty years ago.

References

BARTLETT, M. S. (1936), "The information available in small samples", *Proc. Camb. Philos. Soc.*, **32**, 560–566.
—— (1939), "Complete simultaneous fiducial distributions", *Ann. Math. Stat.*, **10**, 129–138.
BLACKWELL, DAVID & GIRSHIK, M. A. (1954), *Theory of Games and Statistical Decisions.* New York: Wiley.
BOWLEY, A. L. (1935), "Discussion of Professor Fisher's paper", *J. R. Statist. Soc.*, **98**, 55–57.
BOX, G. E. P. & ANDERSEN, S. L. (1955), "Permutation theory in the derivation of robust criteria and the study of departures from assumption", *J. R. Statist. Soc.* B, **17**, 1–34.
BRENY, H. (1955), "L'ètat actuel du problème de Behrens-Fisher", *Trabajos de Estadistica*, **6**, 111–131.
CREASY, MONICA A. (1954), "Limits for the ratio of means", *J. R. Statist. Soc.* B, **17**, 186–194 and 221–222.
EDGEWORTH, F. Y. (1908), "On the probable error of frequency constants", *J. R. Statist. Soc.*, **71** (1908) 381–397, 499–512, 651–678; **72** (1909), 81-90.
FIELLER, E. C. (1954), "Some problems of interval estimation", *J. R. Statist. Soc.* B, **17**, 175–185 and 219–221.
FISHER, R. A. (1933), "The contributions of Rothamsted to the development of the science of statistics", Rothamsted Report of 1933, 1–8.
—— (1935), "The fiducial argument in statistical inference", *Ann. Eugen. Lond.*, **6**, 391–398.
—— (1937), *The Design of Experiments.* Edinburgh and London: Oliver and Boyd.
—— (1955), "Statistical methods and scientific induction", *J. R. Statist. Soc.* B, **17**, 69–78.
FISHER, R. A. & WISHART, J. (1930), "The arrangement of field experiments and the statistical reduction of the results", *Imp. Bureau of Soil Science, Tech. Comm.* No. 10, 1–24.
FOSTER, F. G. & TEICHROEW, D. (1955), "A sampling experiment on the powers of the records tests for trend in a time series", *J. R. Statist. Soc.* B, **17**, 115–120.
GAUSS, C. F. (1887), *Abhandlungen zur Methode der kleinsten Quadrate von Carl Friedrich Gauss.* Berlin, (Translation from Latin by A. Borsch and P. Simon.)
MAULDON, J. G. (1955),"Pivotal quantities for Wishart's and related distributions, and a paradox in fiducial theory", *J. R. Statist. Soc.* B, **17**, 79–90.
NEYMAN, J. (1930), "Mèthodes nouvelles de vérification des hypothéses", *Comptes Rendus du Premier Congrès des Mathématiciens des Pays Slaves*, Warszawa, 1929.
—— (1934), "On the two different aspects of the representative method: the method of stratified sampling and the method of purposive selection", *J. R. Statist. Soc.*, **97**, 558–625.
—— (1935), "Contribution to the discussion of the paper by F. Yates", *Suppl. J. R. Statist. Soc.*, **2**, 235–241.
—— (1937), "Outline of a theory of statistical estimation based on the classical theory of probability", *Philos. Trans. Roy. Soc., London*, Ser. A, **236**, 333–380.
—— (1942), "Basic ideas and some recent results of the theory of testing statistical hypotheses", *J. R Statist. Soc.*, **105**, 292–327.
—— (1950), *First Course in Probability and Statistics.* New York: Henry Holt.
—— (1952), *Lectures and Conferences on Mathematical Statistics and Probability.* Washington, D.C.: Publication of the Graduate School, U.S. Department of Agriculture.
NEYMAN, J. & PEARSON, E. S. (1933*a*), "On the problem of the most efficient tests of statistical hypotheses" *Philos. Trans. Roy. Soc., London*, Ser. A, **231**, 289–337.
—— (1933*b*), "The testing of statistical hypotheses in relation to probabilities *a priori*", *Proc. Camb. Philos Soc.*, **29**, 492–510.
PEARSON, E. S. (1929), "The distribution of frequency constants in small samples from non-normal symmetrical and skew populations", *Biometrika*, **21**, 259–286.
SCHMETTERER, L. (1956), "*Einführung in die mathematische Statistik*". Wien: Springer.
WALD, ABRAHAM (1950), *Statistical Decision Functions.* New York: Wiley.
YATES, F. (1935), "Complex experiments", *Suppl. J. R. Statist. Soc.*, **2**, 181–247.
—— (1939), "An apparent inconsistency arising from tests of significance based on fiducial distributions of unknown parameters", *Proc. Camb. Philos. Soc.*, **35**, 579–591.