Reprint from Statistica Neerlandica vol. 11, 1957, nr 4.

Statistical Priesthood II

Sir Ronald on Scientific Inference *

door Prof. Dr D. van Dantzig

Samenvatting

Ter gelegenheid van het verschijnen van Sir Ronald Fisher's nieuwste boek [1]) *wordt een kritische beschouwing gegeven van zijn aannemelijkheidstheorie en de theorie van ,,fiducial inference''. Wat de laatste betreft wordt geconstateerd, dat deze theorie, in de vorm waarin zij door Fisher gegeven wordt, fouten bevat, hoewel een interpretatie mogelijk is, die in overeenstemming schijnt te zijn met Fisher's ideeën en die een wiskundig correcte behandeling mogelijk maakt. Daartoe is een duidelijk onderscheid, ook in notatie, nodig tussen stochastische grootheden en getallen. De hier gegeven definitie van stochastische grootheden, die Fisher's fiduciële verdelingen bezitten, kan wellicht een gemeenschappelijke basis vormen, waarop aanhangers en tegenstanders van Fisher's ideeën tot een beter wederzijds begrip kunnen komen. Een gedeelte van Fisher's methoden en resultaten kan eveneens gerechtvaardigd worden en ,,fiducial inference'' neemt dan het karakter aan van een eliminatie-methode voor onbekende parameters. Als zodanig heeft deze theorie ongetwijfeld verdiensten, maar het gebied van toepassing is nogal beperkt. Dezelfde resultaten kunnen echter ook bereikt worden langs andere weg, in het bijzonder met behulp van de theorie van betrouwbaarheidsgrenzen.*

Bij andere toepassingen echter, die niet gedekt worden door de theorie van Neyman en Pearson, in het bijzonder bij de toets van Behrens-Fisher, heeft de verwarring van stochastische grootheden en getallen tot onherstelbare fouten geleid. Ondanks alle pogingen, Fisher's dikwijls onduidelijke verklaringen in overeenstemming met zijn philosophische gedachtengang te interpreteren, kan geen rechtvaardiging voor deze toepassingen gevonden worden. Daar vroegere kritiek van andere schrijvers geldig blijft, kan er geen twijfel meer bestaan, dat deze toepassing fout is.

Summary

Partly as a critical review of Sir Ronald Fisher's latest book, partly as an essay, Fisher's theory of likelihood and fiducial inference is carefully considered. As to the latter, it is found that in the form presented it contains

---

*) Report SP 59 of the Statistical Department of Het Mathematisch Centrum, Amsterdam.

[1]) "Statistical Methods and Scientific Inference", Oliver and Boyd, Edinburough and London 1956, pp. 175, 16/—.

errors, but that an interpretation is possible which seems to be in agreement with Fisher's views and which admits of a mathematically correct treatment. This necessitates a clear distinction, also formally, between random variables and numbers. The definition given here of random variables possessing Fisher's fiducial distributions might perhaps provide a common platform on which adherents and opponents of Fisher's view could meet and maybe understand each other. Part of the method and its results can also be justified, and in this light fiducial inference appears as an elimination method for unknown parameters. As such it has indubitable merits, but a rather limited domain of applicability. In these cases the same results can also be obtained by other methods, in particular the theory of confidence domains. In other applications, however, not covered by the Neyman-Pearson theory, in particular in the Behrens-Fisher test, the confusion between random and constant quantities has caused irreparable errors. Notwithstanding a strenuous and prolonged effort to interpret Fisher's often obscure statements in agreement with his philosophical ideas, no justification for these applications could be found. Since the criticism by previous authors remains valid, or comes back in other places, there can no longer be any doubt that these applications are erroneous [1].

1. Introduction

Whichever level of significance be taken, the appearance of a new book by Sir Ronald Fisher can not fail to be a significant event.

Fisher's contributions to mathematical statistics are too numerous and too widely known to be enumerated here, and hardly anybody would deny that no one since Karl Pearson at least, has contributed so many, such important, and such fruitful methods to this science as he did.

A book by Fisher on Scientific Inference is all the more welcome as he has developed on this subject several concepts and methods which have given rise to a sharp controversy and almost to a split between some of the British statisticians and the greater part of the rest of the statistical world. This controversy is utterly regrettable, the more so as the difference in methods leads, in some cases at least, to differences in results, so that either Fisher and his friends use methods which are erroneous, or else a majority of statisticians remain, by lack of understanding, deprived of methods which could be very useful to them.

The controversy concerns in particular Fisher's so-called "fiducial methods" vs. Neyman and Pearson's theory of hypothesis testing,

---

[1]) The author wishes to express his thanks to Prof. Hemelrijk for his valuable help, a.o. in applying the axioms of choice and well-ordering to the paper.

and is carried on with a vehemence which luckily is exceptional in science and gives rise to the suspicion that non-scientific elements like personal dislikes, personal historical backgrounds and perhaps even political ideologies play a role in it.

On p. 3 of the foreword F i s h e r makes a few remarks about the personality and the work of K a r l P e a r s o n († 1936) which, even when one is not in a position to verify whether or not there is any objective justification for them, can hardly be qualified otherwise than as rather nasty [1]). They remind one of those artists who, whenever they paint a portrait, turn it subconsciously into a self-portrait. Anyhow, one wonders how this shooting at dead lions — although it must be admitted that some of the living ones are not treated much nicer —, can be reconciled with the ideal of sportsmanship which the world owes to the author's country.

The attitude of most statisticians not directly concerned with this controversy is: to use the N e y m a n - P e a r s o n method, the mathematical background of which they consider to be as sound as one could wish today, together with F i s h e r's older, non-fiducial, and many other methods; not to be much interested in personal conflicts, and, sometimes, to admit that they do not understand F i s h e r's ideas, a fact which not all of them believe to be their own fault. To some of them the Fisherians look less like a scientific school than like a priest-school, or a religious sect, the adepts of which adhere to some kind of esoteric wisdom, which remains incomprehensible except to the initiated. As, however, among those who accept F i s h e r's methods, some — though not all — of the most prominent British statisticians are found, who naturally have the best opportunity to understand him well, this fact necessarily must make us hesitate to dismiss this esoterism lightly.

For all these reasons a clarification of the situation is highly desirable, and it would be a rejoiceable event if F i s h e r in his new book succeeded in putting his ideas on a firm ground and in removing all doubts, so that his methods became acceptable to the whole statistical world.

Now, F i s h e r's book is fascinating, provoking the reader's thinking and deepening his understanding, and containing many remarks both lucid and sound on the foundations of statistics. In particular his criticism of the B a y e s-

---

[1]) "The terrible weakness of his mathematical and scientific work resulted from his incapability of self-criticism, and his unwillingness to admit the possibility of learning something from others. . . . His mathematics . . . were usually clumsy, and often misleading. In controversy, to which he was much addicted, he constantly showed himself to be without a sense of justice. In his dispute with Bateson . . . he was the bull to a skilful matador . . . Much as he would have disliked the future of statistical science, his activities have a real place in the history of a greater movement."

Laplace-method of probabilities is fuller and clearer than his first treatment of the subject in a paper of 1922, and, in my opinion, on the whole quite correct.

So is his stressing the logical difference between the procedures of testing a scientific hypothesis and of industrial acceptance-sampling, which induces him to reject the so-called statistical decision-methods, based on computation of cost, loss and risk outside the latter domain, although this is not a necessary consequence of the logical difference mentioned. Fascinating, though not quite convincing, also is his treatment (p. 131-136) of two independent normal unit variates, bound to a linear, circular, or general curvilinear relation between their means, as well as several other parts of the book [1]).

Nevertheless the aim as sketched above has not been reached. Although time and again Fisher's exposition is so suggestive that one is fascinated on a first impression, still, critical rereading shows many weaknesses, and closer inspection makes some of them unacceptable to an unprejudiced reader who takes the trouble of going independently through the underlying mathematics, sketched furtively or omitted by Fisher.

In the first place Fisher, notwithstanding his insistence upon the logical aspect of his method and his repeated reproofs of his opponents for faulty logic, does not give a single clear definition of any of his fundamental concepts. Nor does he give any systematic description of his methods, delimitating precisely the domain where they are applicable. That he does not prove or even formulate precisely any mathematical theorems is a matter of course for anyone knowing his work. The only help he renders in understanding the verbal description of general ideas is by exemplification. This queer kind of "logic", without systematics, exactitude or precision, makes the reader giddy; he never knows whether he has understood the author rightly; he floats in a hazy cloud upon a wavy verbalism, interspersed with a "digest" — though hard to digest — of mathematical treatment.

The fact, however, that we get no clear definition of fundamental concepts like "likelihood", "level of significance", "fiducial distribution", "fiducial inference", etc. is not necessarily fatal to them. We remember the fact that

---

[1]) A few minor remarks may be added here. The arithmetical triangle (p. 112) is not due to Fermat, but to Pascal. The Bessel function (p. 135) is I_0, not J_0. On p. 14 Jacob Bernoulli, and on p. 20-21 Von Kries and others should have been mentioned. The "Problem of the Nile" (p. 118) was not a new problem in 1936. It is a nice and witty form of a problem due to Neyman and Pearson (1933), viz. to determine "similar regions". A good deal of work has been done about the conditions under which it is solvable —Neyman 1937, 1941; Lehman and Scheffé, 1947, 1950, Hoel 1948, Ghosh 1948, a.o.).

in his 1922 and 1925 papers Fisher introduced in a similar way the concepts of "consistency", "efficiency", "sufficiency", "intrinsic accuracy", "amount of information", etc., and stated properties of them, accompanied by a faint smile of a mathematical proof. Nevertheless most of these concepts have later proved to be very useful and susceptible of precise definition. Also the properties stated there could be proved under not too strong conditions, and not always did his opponents, when introducing related methods or generalizations, do full justice to the original author. If lack of precision and of correct proof were a reason to disregard claims of authorship, almost no theorem could be ascribed to Laplace, Euler and many others either.

Although an unkind reader might say that one should not even *try* to make sense of Jabberwockian, we shall attempt to interpret the concepts Fisher has introduced in a way as generally acceptable as possible, admitting that we are groping in the mist for anything palpable, and never can be sure that our interpretation will coincide with that of the author himself.

The hypothesis we want to test, is therefore a paraphrase of a Polonian statement, viz.:

"Though this be logic, yet there is method in 't".

2. Likelihood

The gist of Fisher's ideas on likelihood seems to us to admit of the following formulation. [1])

1. There is a fundamental difference between the uncertainty implied in predictions of future events, occurring under a procedure implying some kind of randomness under given conditions, and the uncertainty implied in statements about such conditions ("hypotheses") which may have lead, under a procedure as above, to observed results.

2. It is neither necessary nor desirable, to treat both kinds of uncertainty by means of a single concept. For the former ones the concept of "probability" is adequate; for the latter a different, though related one, called "likelihood".

3. The concept of "probability" should not be considered as a subjective, but as an objective one: a probability statement may be correct or erroneous (cf. p. 33, l. 7-17) and is a mathematical idealization of the long run frequency with which an event occurs on repeated trials under constant conditions (p. 33 l.c., p. 45 line 3 f.b., p. 14 l. 7 f.b., p. 16 l. 3). It follows the well-known rules (or axioms). It refers to a set of events ("reference set", p. 110 l. 16 f.b., "well-defined aggregate", or "population", p. 33) and is applicable to a single one of these events only if it is impossible to determine in advance

[1]) The main text contains our wording of ideas we believe to be essentially those of Fisher. A few comments have been added between square brackets.

its belonging or not-belonging to any subset leading to long-run frequencies differing from those belonging to the whole set (condition of randomness) (p. 32-33, 51, 55, 110, a.o.).

[This materially coincides with Von Mises' definition; Fisher's "reference set" is related to Fréchet's "catégorie d'épreuves", Kolmogoroff's "probability field", and Neyman's "fundamental set", etc. Hence with respect to the concepts of probability there seems to be no difference with the current one, except with the so-called "subjective" view, advocated by De Finetti, Savage and others.]

4. In the Bayes-Laplace theory of the probability of causes it was assumed that every "cause" or hypothesis had a definite probability a priori. If $H_1, H_2, \ldots$ form a complete set of exclusive hypotheses, if $P\{H_i\}$ were the prior probability of H_i, if $P\{A|H_i\}$ is the conditional probability of some observable event A under condition H_i, then the existence of the conditional probability $P\{H_i|A\}$ of H_i under condition A satisfying Bayes' identity:

$$(1) \qquad P\{H_i|A\} = \frac{P\{H_i\} \cdot P\{A|H_i\}}{\Sigma_j P\{H_j\} \cdot P\{A|H_j\}}$$

would follow. It was interpreted as the probability of H_i if A was observed. There are cases where the assumption of the existence of definite prior probabilities $P\{H_i\}$ is justified, but this is not necessarily the case (p. 45, l. 14). Then the Bayes-Laplace theory loses its base. According to 1. the uncertainty about H_i if an observation of A is made, needs not necessarily be forced into the scheme of the probability concept. This can be done by defining the "likelihood" (denoted here by $\boldsymbol{L}$) of H_i, given A, as the probability of A, given H_i:

$$(2) \qquad \boldsymbol{L}\{H_i|A\} \overset{\text{def}}{=\!=} P\{A|H_i\}\ ^{1}).$$

It does not depend on any unknown or even non-existing prior probabilities.

[The transition from the probability to the likelihood terminology can be considered as a "grammatical transformation", like that from "John is older than Peter" to "Peter is younger than John", sometimes useful, never unavoidable. The "likelihood ratio" also is generally accepted, its importance was stressed especially by Neyman and Pearson.]

5. When testing a hypothesis it is possible — and in the case of scientific hypotheses even desirable —, not to restrict oneself to the two "decisions":

---

[1]) The symbol $\overset{\text{def}}{=\!=}$ denotes an equality defining the *left* hand member.

acceptance or rejection [1]), inasmuch as the acceptance or rejection of a scientific hypothesis never can nor should have any finality. Instead one can use the whole scale of likelihoods in order to express the "degree of reluctance" against acceptance of the hypothesis tested.

[The term "reluctance" seems not to be used here in a subjective, but in a normative sense. A likelihood of H_i, not being a probability, does not refer to an even fictitious sampling experiment on a population of hypotheses (or of populations; p. 59). The Neyman-Pearson theory can be summarized in the statement: "If a statistician chooses a level of significance α, and A as a critical event ("critical region"), then, if A is observed he should reject any H_i with

$$\boldsymbol{L}\{H_i|A\} \leqslant \alpha''. \tag{3}$$

Similarly the theory of likelihood could be summarized as follows: "If A satisfies certain conditions with respect to the set of hypotheses, making it fit for being regarded as a criterion, then, if A is observed, the statistician should prefer H_i to H_j if $\boldsymbol{L}\{H_i|A\} > \boldsymbol{L}\{H_j|A\}$". In this form, I do not see that the Neyman-Pearson school or other statisticians will have any fundamental objection against this extended use of the likelihood function. Fisher himself seems to hesitate between the normative interpretation ("should prefer"), the subjective one ("will prefer"), and the objectivistic one ("must prefer"). The preference itself, however, will, should or must be based on an objective situation, determined by the observation A and by the range of likelihood values, i.e. by the probabilities $P\{A|H_j\}$.]

3. Fiducial distributions

In 1930 Fisher introduced "fiducial limits" for an unknown parameter (viz. the correlation coëfficient). At about the same time Jerzy Neyman developed in Poland for the same purpose his theory of "confidence limits", which he introduced in England in 1932 as a generalization of Fisher's result. In the discussion of Neyman's talk before the *Royal Statistical Society*, however, Fisher stated that the two concepts differed fundamentally. At first, in the examples treated, according to both theories the numerical results agreed. In some problems treated later, in particular

[1]) Nevertheless, Fisher *does* accept and reject hypotheses. If e.g. a so-called "test of normality" gives a non-significant result, Fisher would probably accept the hypothesis of normality; so he would accept absence of interaction on the basis of a non-significant test-result. Moreover he often does so on a fixed level of significance (or one among a few ones), notwithstanding his criticizing this method in Neyman-Pearson's work; he even has shown no objection against tables which can be used only on a 1% or 5% level.

the Behrens-Fisher test, this was no longer the case. (W. H. Behrens 1929; R. A. Fisher 1935). The correctness of the arguments underlying this test was contested by M. S. Bartlett (1937) and many other authors, in particular Neyman (1941), but Fisher argued that this was due only to their lack of understanding the logic of statistics and their trying to force his concepts into their own too narrow scheme of reasoning. At present, after some recent papers in the *Journal of the Royal Statistical Society*, the confusion is as great as ever and it is too much to hope that the mutual animosity raised during a quarter of a century could be mediated by an outsider. Nevertheless there might be some use in trying to erect a common platform on which the adherents of both the Fisher school and the Neyman and Pearson school could climb in order to understand each other better. For Fisher's complaint that his opponents do not understand him, is not entirely unjustified.

On the other hand, there is no doubt that Fisher's argument contains errors. If e.g. $\underline{x}$ is a N (μ, σ^2) variate (i.e. normally distributed with mean μ and variance σ^2) then everyone agrees that

$$\text{(4)} \qquad \mathrm{P}\{x < \mu\} = \tfrac{1}{2}.$$

If, in particular, $\mu = 0$, then $\mathrm{P}\{x < 0\} = \frac{1}{2}$. On the other hand, if μ is unknown, and an observation of x yields the value $x = 1{,}37$, say, then Fisher would derive from (4): $\mathrm{P}\{\mu > 1{,}37\} = \frac{1}{2}$. In a 1955 paper Fisher is very strong on this point and even reproaches Neyman's lack of understanding classical logic, for not admitting substitution of a special value of x into a general formula like (4), admitted to be generally true. This criticism, however, is just a blunder, clearly pointed out by Neyman (1956). In fact, a formula like $\sin^2 x + \cos^2 x = 1$ remains valid if x is replaced by an arbitrary number, but a formula like $\int_0^1 x\mathrm{d}x = \frac{1}{2}$ does not: substitution of $x = 2$, say, leads to $\int_0^1 2\mathrm{d}2 = \frac{1}{2}$, which is nonsense. In the first example x was a *free* variable, in the second one a *bound* one. This distinction, made in any elementary text book on symbolic logic, is disregarded by Fisher. Actually x in (4) is a random variable, and $\mathrm{P}\{x < \mu\}$ is not a function of x at all, but a so-called "functional". Hence (4) has the same nature as a statement like

"log x is a convex function"

which deals with the *function* log, not with its particular values, so that substitution of e.g. $x = 2$ yields the meaningless statement

"log 2 is a convex function".

In order to avoid this kind of confusion we are accustomed to *underline* random variables, or distinguish them from numbers by some other printing device (e.g. printing them in bold type). Thereby (4) should be written as

$$(4') \qquad P\{\underline{x} < \mu\} = \tfrac{1}{2},$$

whereas Fisher draws his conclusion from

$$(4'') \qquad P\{x < \underline{\mu}\} = \tfrac{1}{2}.$$

In both cases the non-underlined symbols represent numbers, by which they may be replaced, whereas the underlined symbols do not.

Hence it seems that fiducial inference is all wrong, because random variables and numbers are all mixed up (Cf. Neyman, 1941). This conclusion, however, is somewhat too hasty, and does not do full justice to Fisher.

Fisher's opponents have sometimes believed that statements like (4'') should be interpreted as relating to *conditional* probabilities, given $\underline{x} = x$, whereas $\underline{\mu}$ was thought to refer to an a priori distribution for μ in accordance with Bayes-Laplace. Although some of Fisher's remarks point in this direction, he has declined this interpretation, and rightly so.

The key to Fisher's fiducial inference is his view that a probability distribution does not (as Laplace thought) arise out of nothing or mere ignorance, but only out of observations. So, referring to the same simple example mentioned above, *before* an observation is made μ is just an unknown number, not having any (a priori) distribution at all. But *after* an observation of $\underline{x}$ is made, the fiducial argument uses it "to change the logical status of the parameter from one in which nothing is known of it, and no probability statement about it can be made, to the status of a random variable having a well-defined distribution" (p. 51).

Hence *before* the observation we have a random variable $\underline{x}$ and an unknown number μ; *after* it has been made we have a known number x and a new random variable $\underline{\mu}$, and fiducial argument concerns the latter. The confusion has arisen because Fisher uses the (fiducial) distribution of $\underline{\mu}$ as an a priori distribution of the parameter for *subsequent* observations.

A second objection, however, must be made. Fisher seems to think that a random variable (viz. $\underline{\mu}$) is determined by its probability distribution. This, however, is clearly untrue, for otherwise all N (0,1) variates would be identical. A random variable is defined only if the value it will take in any

observation is determined (although perhaps unknown). And Fisher nowhere defines the new random variable $\underline{\mu}$. We can remove this objection by defining $\underline{\mu}$ in the way in which it implicity is introduced by Fisher:

$$\underline{\mu} \overset{\text{def}}{=\!=} \mu - \underline{x} + x. \tag{5}$$

Here μ is the unknown value of the parameter, and x the known value the random variable $\underline{x}$ has taken in the observation. In the special observation we *have* made, we had $\underline{x} = x$, whence $\underline{\mu} = \mu$. Now $\underline{\mu}$ *has* been defined — and it is difficult to see how it could be defined differently — and we can determine its distribution. If $\Phi(x)$ denotes the N (0,1)-distribution function, we have

$$\text{(6)} \quad \mathrm{P}\{\underline{\mu} \leqslant c\} = \mathrm{P}\{\underline{x} - \mu \geqslant x - c\} = 1 - \Phi\left(\frac{x-c}{\sigma}\right) = \Phi\left(\frac{c-x}{\sigma}\right),$$

as $\underline{x} - \mu$ is N $(0,\sigma^2)$ distributed. Hence $\underline{\mu}$ is N (x, σ^2) distributed. Similarly if $\underline{m}$ is the mean of a random normal sample of size $n = \nu + 1$ and $\underline{s}^2$ the unbiased estimate of the variance of the mean [1]), then

$$\underline{t}_\nu \overset{\text{def}}{=\!=} \frac{\underline{m} - \mu}{\underline{s}} \tag{7}$$

is distributed according to Student's distribution with ν degrees of freedom. Fisher, not using underlinings, thereby determines the fiducial distribution of $\mu = m - st$. We must, however, *define* $\underline{\mu}$, viz. by

$$\underline{\mu} \overset{\text{def}}{=\!=} m - s\underline{t}_\nu \tag{8}$$

where m and s are the values $\underline{m}$ and $\underline{s}$ take in the sample observed, whereas $\underline{t}_\nu$ is defined by (7). As in the sample observed $\underline{m} = m$ and $\underline{s} = s$, it follows that then also $\underline{\mu} = \mu$. Generally, however, the relation between $\underline{\mu}$ and μ is rather remote: substitution of (7) into (8) gives

$$\underline{\mu} = m - s\,(\underline{m} - \mu)/\underline{s}. \tag{9}$$

If $S_\nu(t)$ is the distribution function of $\underline{t}_\nu$, i.e. $S_\nu(t) \overset{\text{def}}{=\!=} \mathrm{P}\{\underline{t}_\nu \leqslant t\}$ and if $t_\nu(\alpha)$ for any α with $0 \leqslant \alpha \leqslant 1$ is the value which $\underline{t}_\nu$ exceeds with probability α, we have by (8)

$$\text{(10)} \quad \mathrm{P}\{\underline{\mu} \leqslant m - st_\nu(\alpha)\} = P\{m - s\underline{t}_\nu \leqslant m - st_\nu(\alpha)\} = \mathrm{P}\{\underline{t}_\nu \geqslant t_\nu(\alpha)\} = \alpha$$

[1]) Many authors use the symbol s^2 as a denotation of $\frac{1}{n}\Sigma(x_i - m)^2$, many other authors use it for $\frac{1}{n-1}\Sigma(x_i - m)^2$; Fisher uses it to denote $\frac{1}{n(n-1)}\Sigma(x_i - m)^2$. Isn't it time that something be done about this notational confusion?

so that $m - st_\nu(\alpha)$ is a (onesided) fiducial limit for $\underline{\mu}$ on the level of significance α: we have

(11) $\qquad \underline{\mu} > m - st_\nu(\alpha)$ except for a probability α.

This corresponds with the Neyman-Pearson confidence limit, which, however, should be written as

(11′) $\qquad \mu > \underline{m} - \underline{s}t_\nu(\alpha)$ except for a probability α

which follows from (7).

Comparing (11) with (11′) the underlinings have changed place; because of our definition (8) this is not done erroneously, but (11) has been *proved* as well as (11′). The variate $\underline{t}_\nu = (\underline{m} - \mu)/\underline{s} = (m - \underline{\mu})/s$ has served as a so-called "pivotal" [1]) variate (p. 117). The distribution function of $\underline{\mu}$ follows immediately from (10):

$$(12) \qquad P\{\underline{\mu} \leqslant c\} = S_\nu\left(\frac{m-c}{s}\right).$$

Fisher uses such a distribution as an a priori distribution for a *second* sample from the same population in order to obtain predictions about the latter. The procedure is described in general terms as follows (p. 126).

If $f(\theta)$ is the fiducial probability density of an unknown parameter and if $P\{A|\theta\}$ is the probability of an event A given θ, then

$$(13) \qquad P\{A\} = \int P\{A|\theta\} f(\theta)\, d\theta.$$

In the examples given it is made clear, that a frequency interpretation for $P\{A\}$ is valid if *only one* prediction for a second sample is made, i.e. if a sequence of pairs of samples is considered. The procedure is, for these examples, equivalent with an elimination procedure of the unknown parameter, which can also be performed without the intervention of a fiducial distribution for θ. If the second sample is taken to be of "infinite" size, the "predictions" reduce to fiducial statements about the unknown parameters.

We have by now found that the following statements, corresponding roughly with ideas expressed on different occasions by Fisher can be justified if our interpretation is accepted.

I. A fiducial distribution is an ordinary probability distribution in the sense of the classical probability concept (p. 51), albeit not of the parameter, but of a new random variable replacing it.

II. It is not based on the assumption that an a priori distribution for the unknown parameter exists.

[1]) Our notation also clarifies this expression. The answer to the question "What pivots?" is: the character of randomness, represented here by the underlining.

III. In the examples mentioned above the fiducial distributions actually possess the form stated by Fisher.

IV. Fiducial distributions can be used to obtain predictions from one sample about *one* second sample in a sequence of pairs of samples, *not* about a sequence of samples all taken from the same population as the original one.

If fiducial inference were presented, more modestly, as an elimination method for unknown parameters, leading, in cases where this is possible, to probability statements which do not depend on the parameters, it might easily find general appreciation as an interesting method, and nobody would dream of being nasty towards other statisticians using different methods to the same purpose.

Moreover I might consider the following statements also as acceptable, without going now into an argument about them.

V. There is no coercive ground for using a fixed level of significance (p. 42). A statistician may use the whole range ("zoning") of probabilities, or, equivalently, likelihood values.

VI. The fiducial argument is restricted by rather severe conditions. As such Fisher mentions (p. 50-51): existence of an "exhaustive" statistic; absense of discontinuous observations; absense of information a priori [1]).

VII. In many applications of statistics to scientific problems neither the dichotomy: rejection or acceptance of a hypothesis [2]), nor the computation of expectations of economic risks is particularly appropriate.

4. The Behrens-Fisher problem

We now consider the Behrens—Fisher problem.

Let m_1, m_2, $n_1 s_1^2$, $n_2 s_2^2$ be the means and variances of two independent samples of sizes n_1, n_2 taken from a $N(\mu_1, \sigma_1^2)$ and a $N(\mu_2, \sigma_2^2)$ population respectively. It is demanded to test the hypothesis $\delta = 0$ where $\delta \overset{\text{def}}{=} \mu_2 - \mu_1$, by a test which does not depend on σ_1, σ_2. Fisher's own procedure is as follows. The ratio's

[1]) In the case of a Bernoullian distribution the a priori knowledge $0 \leqslant p \leqslant 1$ exists. Similarly in Fisher's first example (p. 52-55) $\theta > 0$. If a priori knowledge of this kind is allowed, the conditions mentioned are *not* sufficient; the range of the parameter may e.g. be smaller than that of the statistic. A simple example showing this feature is given by

$$F(x|\theta) = 1 - e^{-x^2 - \theta x} (x \geqslant 0, \ \theta \geqslant 0).$$

where there is a positive probability that the maximum likelihood estimate $\hat{\theta}$ of θ is zero, so that the fiducial distribution of θ in Fisher's sense is discontinous, although the original distribution is not.

[2]) But this dichotomy is handled far less dogmatically in the Neyman-Pearson school than Fisher seems to believe. And his insinuation: "had the authors . . . had any real familiarity with work in the natural sciences . . ." expresses insufficient knowledge of achievements as well as bad temper and bad taste.

(14) $$t_{\nu_1} = (m_1 - \mu_1)/s_1, \quad t_{\nu_2} = (m_2 - \mu_2)/s_2$$

are independent and have Student t-distributions with ν_1 and ν_2 degrees of freedom respectively. Hence, inserting for s_1 and s_2 the observational values, $T = s_2 t_{\nu_2} - s_1 t_{\nu_1}$ [1]) has a computable distribution, and we have

(15) $$\delta = d - T \text{ with } d \overset{\text{def}}{=\!=} m_2 - m_1,$$

so that $\delta - d$, hence also δ, has a known distribution. If $T(p) = T(s_1, s_2; p)$ is the value of T exceeded with probability p, then δ differs significantly from zero on the level of significance 2α if $|d| \geqslant T(\alpha)$.

In the form given the argument is subject to similar objections as mentioned before. Introducing the underlinings, we have

(16) $$\underline{t}_{\nu_1} = (\underline{m}_1 - \mu_1)/\underline{s}_1, \quad \underline{t}_{\nu_2} = (\underline{m}_2 - \mu_2)/\underline{s}_2$$

for the Student variables. Then either we have to put

(17) $$\underline{T} = \underline{s}_2 \underline{t}_{\nu_2} - \underline{s}_1 \underline{t}_{\nu_1},$$

or

(17′) $$\underline{T} = s_2 \underline{t}_{\nu_2} - s_1 \underline{t}_{\nu_1}.$$

In the first case, however, $\underline{T} = \underline{m}_2 - \underline{m}_1 - \mu_2 + \mu_1$ does not have the required distribution, but a $N\left(0, \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}\right)$ distribution. In the second case $\underline{T}$ *does* have the required distribution, but then (15) is no longer valid. For

$$d - \underline{T} = d - s_2 \underline{t}_{\nu_2} + s_1 \underline{t}_{\nu_1} = m_2 - s_2 (\underline{m}_2 - \mu_2)/\underline{s}_2 - m_1 + s_1 (\underline{m}_1 - \mu_1)/\underline{s}_1,$$

which is not equal to $\delta = \mu_2 - \mu_1$. [2])

We can, however, salvage the argument to a certain extent by introducing, as before, the "fiducial variates"

(18) $$\underline{\mu}_2 \overset{\text{def}}{=\!=} m_2 - s_2 \underline{t}_{\nu_2} \; ; \; \underline{\mu}_1 \overset{\text{def}}{=\!=} m_1 - s_1 \underline{t}_{\nu_1} \; ; \; \underline{\delta} \overset{\text{def}}{=\!=} \underline{\mu}_2 - \underline{\mu}_1.$$

From these follows immediately

(19) $$\underline{\delta} = d - \underline{T}$$

with

(17′) $$\underline{T} \overset{\text{def}}{=\!=} s_2 \underline{t}_{\nu_2} - s_1 \underline{t}_{\nu_1}$$

having the required distribution. It follows in particular that

---

[1]) Actually Fisher uses $D = T/\sqrt{s_1^2 + s_2^2}$, but this is of no influence on the argument.

[2]) If we omit the underlinings, i.e. muddle up random variables and numbers, everything cancels except $\mu_2 - \mu_1$.

$$(20)\quad P\{d - T(\alpha) < \underline{\delta} < d + T(\alpha)\} = P\{|\underline{T}| < T(\alpha)\} = 1 - 2\alpha,$$

as required, the distribution of $\underline{T}$ being symmetrical.

This does not, however, justify the use of (9), (17′) and (20) for testing the hypothesis $\delta = 0$. For, (20) is true for any pair of values s_1 and s_2, but only if $\underline{t}_{\nu_1}$ and $\underline{t}_{\nu_2}$ are random variables with Student distributions independent of these values s_1 and s_2. The statement (20) therefore holds if T is computed from (17′) after substitution of two values t_{ν_1} and t_{ν_2} chosen at random from such distributions, and if δ is computed from (19) with this value for T. If, however, t_{ν_1} and t_{ν_2} are computed from (14) and the *same* values s_1 and s_2 are inserted in (17′), then the abovementioned condition of independence does not hold and (20) is not true [1]).

Now, Fisher says explicitly (p. 59) that "probabilities obtained by a fiducial argument are objectively verifiable in exactly the same sense as are propabilities assigned in games of chance". Thus, *just as a statistician wants to test a scientist's hypothesis, the scientist may want to test the statistician's test.* With respect to the Behrens-Fisher test, he may do so by feeding the statistician with samples taken at random from pairs of normal populations having parameters he may choose *at his will.* Therefore he may choose pairs of populations with always $\mu_1 = \mu_2$, $\sigma_1/\sigma_2 = \varrho =$ constant [2]), and $\alpha = 0.05$. For the onesided case he then has to verify, for every pair of samples, the inequality

$$(21)\qquad m_2 - m_1 \geqslant T(s_1, s_2;\ 0.05)$$

and then Neyman's argument (1941; 1952, p. 242-245), which in itself is unassailable, becomes applicable. It shows that the limiting frequency quotient of the number of cases wherein this inequality is found to be true is equal to

$$P\{\underline{m}_2 - \underline{m}_1 \geqslant T(\underline{s}_1, \underline{s}_2;\ 0.05) \mid \mu_1 = \mu_2;\ \sigma_1 = \varrho\sigma_2\},$$

which depends on ϱ and is not constantly equal to 0.05.

So the scientist may ask: "what am I to think about a test, stating that a

[1]) Resuming: if we want to get rid of the ratios $s_2/\underline{s}_2$ and $s_1/\underline{s}_1$ in the expression for $d - \underline{T}$ (not bothering as yet about the remaining $m_2 - \underline{m}_2 - m_1 + \underline{m}_1$), we may: *a*) replace in (16) $\underline{s}_1$ by s_1, $\underline{s}_2$ by s_2, but then $\underline{t}_{\nu_1}$, $\underline{t}_{\nu_2}$ do not have Student's distributions but normal distributions; *b*) use (16) with (17) instead of (17′); then $\underline{t}_{\nu_1}$, $\underline{t}_{\nu_2}$ do have Student's distributions, but $\underline{T}$ has a normal distribution instead of a Behrens-Fisher one; *c*) use (16) with (17′), but impose the condition $s_1/\underline{s}_1 = s_2/\underline{s}_2$. Then $\underline{t}_{\nu_1}$, $\underline{t}_{\nu_2}$ do have Student's distributions, and $\underline{T}$ is a linear combination of them with constant coefficients, but they are *no longer independent.* Hence the convolution theorem upon which the Behrens-Fisher distribution is based does not apply, and $\underline{T}$ again has not the required distribution.

[2]) Without the statistician's knowing, of course.

definite inequality has a fiducial probability of 5%, *irrespective of the value of* ϱ, which I, using $\varrho = 0.01$ (with $n_1 = 12$, $n_2 = 6$; cf. Neyman, 1952, p. 245) find in the long run to hold in 3.4% of all cases, whereas my colleague, doing the same thing with $\varrho = 10$ finds it to hold in 6.6% of all cases?"

5. Conclusions

Summarizing this lengthy discussion we may say that a considerable effort has been made to ferret out all that might be correct in Fisher's theory and to bring "fiducial inference" into a form that might be expected to be acceptable to the majority of statisticians and that seems to incorporate all philosophical and logical principles put forward by Fisher. This effort, I hope, may have hit not too far from the mark and at least have brought some clarification.

The result, however, is rather disappointing. We have found that fiducial inference could be justified in just those cases where it agreed with the inference based on elimination of parameters or on confidence limits. In these cases the confusion between random and constant quantities is reparable. In the critical case of the Behrens-Fisher test, however, where no such agreement exists, this confusion is irreparable and enters so strongly into the theory of the test, that it can not be justified, whereas the criticism by previous authors remains completely unshaken. So it seems that even from Fisher's own standpoint the Behrens-Fisher test is not tenable.

If the "hypothesis" mentioned on p. 189 were tested according to Fisher's own principles, it may be said to have acquired a considerable likelihood. And if it were done according to Neyman and Pearson's methods it could certainly not be "rejected" on any reasonable level of significance. Fisher's logic doubtlessly contains a good deal of method, and besides, profound thinking and good sense, even if errors have been made in some of its applications, so that some of its results can not be justified, whereas, in those cases where its application is correct, the same results can be obtained by other methods also, which imply a considerably smaller danger of making errors.

Recently I have compared some modern statisticians to ancient priests. If it is allowed to pursue the metaphor a little further, and no offence is taken where none is meant, I might say that Sir Ronald doubtlessly would be the High Priest. One can well imagine, how hard it must have been, for some ancient tribes, not to be charmed by the passion of their

High Priest's dancing around the altar, enchanted by the sincere sonority of his singing praise, in unison with his priest choir, of his goddess Fiducia, spellbound by the enticing clouds of sweet exciting incense enveloping her, and awe-struck by the curse by which he damns his demons.

I am afraid that we are perhaps somewhat too sober-minded to participate in such a dance, whilst being eager for any real wisdom and genuine science which might be hidden behind the rites.

Returning, finally, to Alice: *has* Fisher slain the Jabberwock? It seems that he *did* use the vorpal blade. But I am not so sure that he succeeded in shunning the frumious Bandersnatch. "You see, she didn't like to confess, even to herself, that she couldn't make it out at all. Somehow it seems to fill my head with ideas . . . only I don't exactly know what they are. However, *somebody* killed *something* : that's clear, at any rate."

Literature

M. S. Bartlett (1936), The information available in small samples, Proc. Cambridge Phil. Soc. **32,** 560-566.

M. S. Bartlett (1939), Complete simultaneous fiducial distributions, Ann. of Math. Stat. **10,** 129-138.

Sir Ronald Fisher (1955), Statistical methods and scientific induction, Journ. Royal Stat. Soc. B **17,** 69-78.

Sir Ronald Fisher (1956), On a test of significance in Pearson's Biometrica table (10.11), id. **18,** 56-60.

Lewis Carroll (1872), Through the looking glass and what Alice found there.

J. Neyman (1941), Fiducial argument and the theory of confidence intervals. Biometrica, **32,** 128-150. Reprinted in: Lectures and conferences on mathematical statistics and probability, 2nd ed. 1952, pp. 229-254.

J. Neyman (1956), Note on an article by Sir Ronald Fisher, Journ. Royal Stat. Soc. B **18,** 288-294.

J. Neyman and E. S. Pearson (1933), On the problem of the most efficient tests of statistical hypotheses, Phil. Trans. **231,** 289-337.

W. Shakespeare, Hamlet, II. 2.