

Translation

Statistical methods and the subjective basis of scientific knowledge

G. Malécot

ANNALES DE L'UNIVERSITÉ DE LYON, Année 1947-X-pp. 43 à 74.

"Without a hypothesis, that is, without anticipation of the facts by the minds, there is no science." Claude BERNARD

(Translated from French and commented
by Professor Daniel Gianola; received April 6, 1999)

Preamble – When the Editor of Genetics, Selection, Evolution asked me to translate this paper by the late Professor Gustave MALÉCOT into French, I felt flattered and intimidated at the same time. The paper was extensive and highly technical, and written in an unusual manner for today's standards, as the phrases are long, windy and, sometimes, seemingly never ending. However, this was an assignment that I could not refuse, for reasons that should become clear subsequently.

I have attempted to preserve MALÉCOT's style as much as possible. Hence, I maintained his original punctuation, except for a few instances in which I was forced to introduce a comma here and there, so that the reader could catch some breath! In those instances in which I was unsure of the exact meaning of the phrase, or when I felt that some clarification was needed, I inserted footnotes. The original paper also contains footnotes by MALÉCOT; mine are indicated as "Translator's Note", following the usual practice; hence, there should be little room for confusion. There are a few typographical errors and inconsistencies in the original text, but given the length of the manuscript and that it was written many years before word processors had appeared, the paper is remarkably free of errors.

This is undoubtedly one of the most brilliant and clear statements in favor of the Bayesian position that I have encountered, specially considering that it was published in 1947! Here, MALÉCOT uses his eloquence and knowledge of science, mathematics, statistics and, more fundamentally, of logic, to articulate a criticism of the points of view advanced by FISHER and by NEYMAN in connection with statistical inference. He argues in a convincing (this is my subjective opinion!) manner that in the evaluation of hypotheses, speaking in

a broad sense, it is difficult to accept the principle of maximum likelihood and the theory of confidence intervals unless BAYES formula is brought into the picture. In particular, his discussion of the two types of errors that arise in the usual "accept/reject" paradigm of NEYMAN is one of the strongest parts of the paper. MALÉCOT argues effectively that it is impossible to calculate the total probability of error unless prior probabilities are brought into the treatment of the problem. This is probably one of the most lucid treatments that I have been able to find in the literature.

The English speaking audience will be surprised to find that the famous CRAMER-RAO lower bound for the variance of an unbiased estimator is credited to FRECHET, in a paper that this author published in 1943. C.R. RAO's paper had been printed in 1945! The reference given by MALÉCOT (FRECHET, 1934) is not accurate, this being probably due to a typographical error. If it can be verified that actually FRECHET (or perhaps DARMOIS) discovered this bound first, the entire statistical community should be alerted, such that history can be written correctly. In fact, some statistics books in France refer to the FRECHET-DARMOIS-CRAMER-RAO inequality, whereas texts in English mention the CRAMER-RAO lower bound or the "information inequality" ..

On a personal note, I view this paper as setting one of the pillars of the modern school of Bayesian quantitative genetics, which would now seem to have adherents. For example, when Jean-Louis FOULLEY and I started on our road towards Bayesianism in the early 1980s, this was (in part) a result of the influence of writings of the late Professor LEFORT, who, in turn, had been exposed to MALÉCOT's thinking. In genetics, MALÉCOT had given a general solution to the problem of the resemblance between relatives based on the concept of identity by descent (G. MALÉCOT, *Les mathématiques de l'hérédité* Masson et Cie., Paris, 1948). In this contemporary paper, we rediscover his statistical views, which point clearly in the Bayesian direction. With the advent of Markov chain Monte Carlo methods, many quantitative geneticists have now implemented Bayesian methods, although probably this is more a result of computational, rather than of logical, considerations. In this context, I offer a suggestion to geneticists that are interested in the principles underlying science and, more particularly, in the Bayesian position: read MALÉCOT.

Daniel Gianola, Department of Animal Sciences, Department of Biostatistics and Medical Informatics, Department of Dairy Science, University of Wisconsin-Madison, Wisconsin 53706, USA

1. BAYES FORMULA

The fundamental problem of acquiring scientific knowledge can be posed as follows. Given: a system of knowledge that has been acquired already (certainties or probabilities) and which we will denote as K ; a set of mutually exclusive and exhaustive assumptions θ_i , that is, such that one of these must be true (but without knowing which); and an experiment that has been conducted and that gives results E : what new knowledge about θ_i is brought about by E ?

A very general answer has been given in probabilistic terms by Bayes, in his famous theorem; let $P(\theta_i|K)$ be the probabilities of the θ_i based on K , or

prior probabilities of the hypotheses; $P(\theta_i|EK)$ be their posterior probabilities, evaluated taking into account the new observations E ; $P(E|\theta_iK)$ be the probability that the hypothesis θ_i , supposedly realized, gives the result E , a probability that we call the *likelihood of θ_i as a function of E* (within the system of knowledge K); the principles of total and composite probabilities give then:

$$P(\theta_i|EK) = \frac{P(E|\theta_iK) P(\theta_i|K)}{P(E|K)}$$

the denominator $P(E|K) = \sum_i P(E|\theta_iK) P(\theta_i|K)$ does not depend on i . One can say, then, that the *probabilities a posteriori (once E has been realized) of the different hypotheses are respectively proportional to the products of their probabilities a priori times their likelihoods as a function of E* (all this holding in the interior of system K). The proportionality constant can be arrived at immediately by writing that the sum of posterior probabilities is equal to 1. The preceding rule still holds in the case where one cannot specify all possible hypotheses θ_i or all the probabilities $P(E|\theta_iK)$ of their influence on E , but then the sum of posterior probabilities $P(\theta_i|EK)$ of all the hypotheses that one has been able to formulate their consequences would be lesser and not equal to 1.

We will show how BAYES formula provides logical rules for choosing one θ_i over all possible θ_i , or among those whose consequences can be formulated; further, it will be shown how the rules adopted in practice cannot have a logical justification outside of the light of this formula.

2. THE RULE OF THE MOST PROBABLE HYPOTHESIS

We shall begin a critical discussion of the methods proposed by FISHER's school by posing *the rule of the most probable value: choose the hypothesis θ_i having the largest posterior probability*, with the risk of error given by the sum of the probabilities of the hypotheses discarded (when one can formulate all such hypotheses)(the risk will be small only if this sum is small; it may be reasonable to group together several hypotheses having a total probability close to 1, without making a distinction between them; this we shall do in Section VII)

In order to apply this rule, it is necessary to determine the θ_i giving the maximum of $P(E|\theta_iK) P(\theta_i|K)$. It follows that the choice of θ_i depends not only on the likelihoods of the θ_i but also on their prior probabilities, often subjective and variable between individuals, even within individuals depending on the state of their knowledge or of their memory. However, it must be noted that the presence of the prior probability in the formula is in perfect agreement with the rule, admitted by most experimenters, of combining (weighted naturally) all observations that provide information about a certain hypothesis. Suppose that after the experiments E , another set of experiments E' is carried out: collecting all such experiments one has:

$$\begin{aligned} P(\theta_i|E'EK) &= \frac{P(E'E|\theta_iK) \times P(\theta_i|K)}{P(E'E|K)} \\ &= \frac{P(E'|\theta_iEK) \times P(E|\theta_iK) \times P(\theta_i|K)}{P(E'E|K)} \end{aligned}$$

and the rule leads to choosing the θ_i that maximizes the numerator; however, the first term represents the likelihood of θ_i as a function of E' within the system EK , and the product of the last two is proportional to the probability of θ_i within the system EK , that is:

$$P(\theta_i|EK) = \frac{P(E|\theta_i K) \times P(\theta_i|K)}{P(E|K)}$$

which is the probability a priori of θ_i before realization of E' ; it follows then that one would obtain the same result maximizing $P(E'|\theta_i EK) \times P(\theta_i|EK)$, that is, the product of the likelihood times the new prior probability.

The rule of the most likely value, as stated, takes into account all our knowledge, at each instant, about all hypotheses examined, and every new observation is used to update their probabilities by replacing the probabilities evaluated before such observation by posterior probabilities. The delicate point is what values should be assigned to the probabilities a priori before any experimentation providing information about the hypotheses takes place. LAPLACE and BAYES proposed to take the prior probabilities of all hypotheses as equal, which makes the posterior probabilities proportional to the likelihood, leading in this case to the *rule of maximum likelihood proposed by Mr. Fisher*¹, a rule that, unlike him, does not seem possible to me to adopt as a first principle, because of the risk of applying it to a given group of observations without considering the set of other observations providing information about the hypotheses considered. A striking example of this pitfall is the contradiction, noted by Mr. Jeffreys², between the principle of maximum likelihood and the underlying principle of "significance criteria". In this context, the objective is to determine if the observed results are in agreement with a hypothesis or with a simple law (the "null hypothesis" of Mr. Fisher), or if the hypothesis must be replaced by a more complicated one with the alternative law being more global, including the old and the new parameters. To be precise, if the old law depends on parameters $\alpha_1, \dots, \alpha_p$, the new one will depend in addition on $\alpha_{p+1}, \dots, \alpha_{p+q}$ and will reduce to the old one at given values of $\alpha_{p+1}, \dots, \alpha_{p+q}$ which can always be supposed to be equal to 0 (that is why the name "null hypothesis" is given to the assumption that the old law is valid). The maximum of $P(E|\alpha_1, \dots, \alpha_{p+q}, K)$ when all the α_i vary will be larger in general than its maximum when $\alpha_{p+1} = \dots = \alpha_{p+q} = 0$, hence, the rule of maximum likelihood will lead, almost always, to adopting the most complicated law. On the other hand, the usual criterion in this case is to investigate if there is not a great risk of error made by adopting the simplest law: to do this one can define a "deviation" between the observed results and those that would be expected, on average, from the simplest law, and then find the prior probability from such law of obtaining a deviation that is at least as large as the observed distance. It is convenient not to reject the simplest law unless this probability is very small. This is the principle of criteria based on "significant deviations".

¹ Translator's Note: Fisher's name is in italics and not in capital letters in the original paper. I have left this and other minor inconsistencies unchanged.

² Translator's Note: References to Jeffreys made later in the paper appear in capital letters.

Hence, the simplest law benefits from a favorable prejudice, that is, of having a prior probability that is larger than that assigned to more complex laws. Why is it prejudged more favorably? Sometimes this is the result of our belief on the simplicity of the laws of nature, a belief that may stem from convenience (examples: the COPERNICUS system is more convenient than that of PTOLEMY to understand the observations and to make predictions; fitting of an ellipse to the trajectory of Mars by KEPLER without consideration of the law of gravitation), or from previous experience.

Consider the example of a fundamental type of experiment in agricultural biology: comparing the yields of two varieties of some crop, by planting varieties V and V' adjacent to each other at a number of points A_1, \dots, A_N of an experimental field, so as to take into account variability in light and soil conditions. If x_1, \dots, x_N and x'_1, \dots, x'_N are the yields of V and V' measured at the N points, two main attitudes are possible when facing the data: those inclined to believe that the difference between V and V' cannot affect yield will ask themselves if all x_i and x'_i can be reasonably viewed as observed values of two random variables X and X' following the same law; for this, they will adopt a significance test based on the difference between the means, and they will maintain their hypothesis if this difference is not too large. On the other hand, those whose experience leads them to believe that the difference in varieties should translate into a difference in yield will admit a priori that the random variables X and X' are different, introducing right away a larger number of parameters (for example, $\bar{X}, \sigma, \bar{X}', \sigma'$ if it is accepted that X and X' are Laplacian) and they will be concerned immediately with the estimation of these parameters, in particular $\bar{X} - \bar{X}'$, by the method of maximum likelihood for example (which in the case of laws of LAPLACE with the same standard deviation, gives as estimator of $\bar{X} - \bar{X}'$ the difference between arithmetic means of the x_i and x'_i); this method assumes implicitly that the prior probabilities of the values of $\bar{X} - \bar{X}'$ are all equal and infinitesimally small, which is quite different from the first hypothesis where a priori we view the value $\bar{X} - \bar{X}' = 0$ (corresponding to identity of the laws) as having a finite probability. These two different attitudes correspond to different states of information a priori, of prior probabilities; the statistical criteria are, *thus, not objective*, because there could not be a contradiction between the two: it is not possible that one leads to the conclusion that $\bar{X} - \bar{X}' = 0$ and the other to conclude that $\bar{X} - \bar{X}' \neq 0$. This discrepancies result from the fact that the criteria are subjective and correspond to different states of information or experience.

We shall now take an example from genetics. A problem of current interest is that of linkage between Mendelian factors. When crossing a heterozygote $AaBb$ with a double homozygote recessive, we observe in the children, if these are numerous, the genotypes $ABab, abab, Abab, aBab$ in numbers $\alpha, \beta, \gamma, \delta$ ($\alpha + \beta + \gamma + \delta = N$), leading to admit that, independently, each child can possess one of the 4 genotypes with probabilities $\frac{1-r}{2}, \frac{1-r}{2}, \frac{r}{2}, \frac{r}{2}$, with r being a "coefficient of linkage" having a value between 0 and 1. If all available knowledge were based on a certain number of crossing experiments in *Drosophila*, one would be led to state that all values of r inside of an interval are equally likely, and then take the maximum likelihood estimate as value of r ,

for each experiment. However, if one brings information from human genetics into the picture, this shows that r is almost always near to $\frac{1}{2}$, which would tend to give a privileged prior probability to $\frac{1}{2}$ when interpreting each measurement taken in human genetics. At any rate, more advanced experimentation on the behavior of chromosomes gives us a more precise basis for interpretation; if the two factors are "located" in different chromosomes, $r = \frac{1}{2}$, there is "independent segregation" of the two characters. There is "linkage" $r < \frac{1}{2}$: "coupling"; $r > \frac{1}{2}$: "repulsion" only when the two factors reside in the same chromosome, a fact which, in the absence of any information on the localization of the two factors considered, would have a prior probability of $\frac{1}{24}$ (because there are 24 pairs of chromosomes in humans).

In the light of this knowledge, one can start every study of linkage between new factors in humans by assigning $\frac{23}{24}$ and $\frac{1}{24}$ as values of the prior probabilities of $r = \frac{1}{2}$ and $r \neq \frac{1}{2}$; if one can view the values $r \neq \frac{1}{2}$ as equally likely, that is, take $\frac{1}{24}dr$ as the probability that $r \neq \frac{1}{2}$ lies between r and $r + dr$, then it is easy to form the posterior probabilities of $r = \frac{1}{2}$ and $r \neq \frac{1}{2}$; the likelihood of r (the probability that a given value r produces numbers $\alpha, \beta, \gamma, \delta$ in the four categories will be:

$$2^{-N} (1-r)^{\alpha+\beta} r^{\gamma+\delta}$$

which gives, letting E be the observation of $\alpha, \beta, \gamma, \delta$:

$$P\left(r = \frac{1}{2} \middle| EH\right) \propto 2^{-2N} \frac{23}{24}$$

$$P\left(r \neq \frac{1}{2} \middle| EH\right) \propto \int_0^1 2^{-N} (1-r)^{\alpha+\beta} r^{\gamma+\delta} \frac{1}{24} dr$$

Of these two, we will retain the hypothesis having the largest posterior probability; if this is hypothesis $r \neq \frac{1}{2}$, we would take as estimate of r , within all values $r \neq \frac{1}{2}$, the one maximizing the posterior probability, that is, the maximizer of the likelihood $2^{-N} (1-r)^{\alpha+\beta} r^{\gamma+\delta}$, which has as value $r = \frac{\gamma+\delta}{N}$.

I have deliberately presented the problem in a somewhat shocking manner, emphasizing that the prior probabilities are known. Nevertheless, it cannot be argued that the rule at which we arrive is not that in current use, or that at least it is in close numerical proximity³: reject the "null hypothesis" if this

³ Translator's Note: In the original, there is a delicate interplay of double negatives which is difficult to translate. The phrase is: "On ne peut néanmoins contester que la

gives a large discrepancy with the observations; subsequently, estimate the parameters by maximum likelihood. My objective has been to show on what type of assumptions one operates, *willingly or unwillingly*, when these rules are applied. Using prior probabilities, it is possible to see the logical meaning of the rules more clearly, and a possibly precarious state of the assumptions made a priori can be thought of as a warning against the tendency of attributing an absolute value to the conclusions (as done by Mr. MATHER who gives a certain number of rules as being objectively best, even if these are contradictory): we take note of the arbitrariness in the choice of the prior probabilities and in the manner of contrasting the hypotheses $r = \frac{1}{2}$ and $r \neq \frac{1}{2}$; and we also see how the conclusion about the value of r is subjective.

3. OPTIMUM ESTIMATION

We shall now examine another aspect of the question of the rule of maximum likelihood, which Mr. FISHER (7) thought could be justified independently of prior probabilities, with his rule of optimum estimation. Suppose the competing hypotheses are the values of a parameter θ , with each value giving to the observed results E a probability $\pi(E|\theta)$ before observation, which is a function of θ , its likelihood function; we will call an *estimator* of θ , extracted from observations E , any function H of the observations only giving information about the value of θ ; same as with the observations, this estimator is a random variable *before the data are observed*, its probability law depending on θ . (In the special case where, once the value H is given, the conditional probability law of E no longer depends on θ , it is unnecessary to give a complete description of E once H is known, because this would not give any supplementary information about θ , and we then say that H is an *exhaustive*⁴ estimator of θ .)

It is said that H is a *fair estimator*⁵ of θ if its mean value $M(H)$ ⁶ is always equal to the true value θ irrespective of what this is. It is said that H is *asymptotically fair*⁷ if $M(H) - \theta$ is infinitesimally small with $\frac{1}{N}$, N being the number of observations constituting E .

It is said that H is *correct*⁸ if it always converges in probability towards θ when N tends towards infinity. (For this, it suffices that H be asymptotically fair and that it has a fluctuation⁹ tending towards 0. Conversely, every fair estimator admitting a mean is asymptotically fair).

règle à laquelle nous arrivons ne soit, aux valeurs numériques des probabilités près, celle qui est d'un usage courant..."

⁴ Translator's Note: The English term is *sufficient*. Malécot's terminology is kept whenever it is felt that it has anecdotal value, or to reflect his style.

⁵ Translator's Note: Unbiased estimator.

⁶ Translator's Note: It is useful to remember hereinafter that M (expression) denotes the expected value of the expression. The M comes from "moyenne" = mean value.

⁷ Translator's Note: Asymptotically unbiased.

⁸ Translator's Note: Consistent.

⁹ Translator's Note: Fluctuation = Variance.

It is said that H is *asymptotically Gaussian* if the law of H tends towards one of the type LAPLACE-GAUSS when N increases indefinitely. In statistics, it is frequent to encounter estimators that are both correct and asymptotically Gaussian; we shall denote such estimators as C.A.G (see, DUGUE, 5). The precision of such an estimator is measured perfectly by $M \left[(H - \theta)^2 \right] = \zeta^2$, this becoming infinitesimally small with $\frac{1}{N}$; the precision will increase as ζ^2 decreases, hence $I = \frac{1}{\zeta^2}$, which will be termed the *quantity of information* extracted by the estimator, will be larger.

In what follows, we will restrict attention to the case where E consists of N *independent* observations x_1, \dots, x_N with their distribution functions being a priori:

$$F_1(x_1, \theta), \dots, F_N(x_N, \theta)$$

The probability of a set E of observations is:

$$\pi(E|\theta) = dF_1(x_1, \theta) \dots dF_N(x_N, \theta)$$

(Stieltjes multiple differential) with

$$\int \pi(E|\theta) = 1$$

with the integration covering the entire space \mathfrak{R}_N described by the x_1, \dots, x_N . It is then easy to show, with Mr. FRECHET (8), that the fluctuation ζ^2 of *any fair estimator* has a *fixed lower bound*. Let $H(x_1, \dots, x_N)$ be one such estimator. For any θ :

$$M(H - \theta) = \int (H - \theta)\pi(E|\theta) = 0$$

from where, taking derivatives of this identity with respect to θ :

$$-1 + \int (H - \theta) \frac{\delta \pi}{\delta \theta} = 0$$

leading to

$$M \left[(H - \theta) \frac{\delta \log \pi}{\delta \theta} \right] = 1.$$

Observing that

$$M \left[\frac{\delta \log \pi}{\delta \theta} \right] = \int \frac{\delta \pi}{\delta \theta} = 0$$

and letting

$$M \left[\left(\frac{\delta \log \pi}{\delta \theta} \right)^2 \right] = \sigma^2$$

it is seen that the square of the coefficient of correlation between $(H - \theta)$ and $\frac{\delta \log \pi}{\delta \theta}$ is

$$\frac{1}{\sigma^2 M [(H - \theta)^2]}$$

from where:

$$\zeta^2 = M [(H - \theta)^2] \geq \frac{1}{\sigma^2}.$$

^{10, 11}

The equality holds only if $(H - \theta) = \frac{\delta \log \pi}{\delta \theta} \times \text{constant}$ almost everywhere; it is easy to show that this cannot hold unless H is an exhaustive estimator, for, in making a change of variables in the space \mathfrak{R}_N , with the new variables being $H, \xi_1, \dots, \xi_{N-1}$, functions of x_1, \dots, x_N , the distribution function of H will be $G(H, \theta)$ and the joint distribution function of the ξ_i inside of the space $\mathfrak{R}_{N-1}(H)$ that they span will be $k(H, \xi_1, \dots, \xi_{N-1}, \theta)$ ¹²; then one has $\pi(E|\theta) = dG[dk]$ ¹³ with

$$\int_{\mathfrak{R}_{N-1}(H)} [dk] = 1$$

and

$$\int_{H=-\infty}^{H=+\infty} dG = 1;$$

further, because

$$M \left[\frac{\delta \log dG}{\delta \theta} \frac{\delta \log [dk]}{\delta \theta} \right] = \int_{-\infty}^{+\infty} \frac{\delta (dG)}{\delta \theta} \int_{\mathfrak{R}_{N-1}(H)} \frac{\delta [dk]}{\delta \theta} = 0$$

¹⁰ (1) Mr. Frechet has shown more generally that for an asymptotically fair estimator, for N sufficiently large, it is always true that

$$\zeta^2 \geq \frac{1 - \epsilon}{\sigma^2}$$

for an arbitrarily small ϵ .

¹¹ Translator's Note: This is a statement of the Cramer-Rao lower bound for the variance of an unbiased estimator. It is historically remarkable that FRECHET, to whom MALÉCOT attributes the result, seems to have published this in 1943 (1934 is given incorrectly in the References). The first appearance of the lower bound in the statistical literature is often credited to: Rao C.R., Information and accuracy attainable in the estimation of statistical parameters, Bull. Calcutta Math. Soc. 37 (1945) 81-91. According to C. R. Rao (personal communication) Cramer mentions this inequality in his book, published two years later. Neyman named it as Cramer-Rao inequality.

¹² Translator's Note: Although perhaps obvious, Malécot's notation hides somewhat that this is the conditional distribution of all ξ'_i s, given H .

¹³ The bracket denotes a multiple differential of the Stieltjes type, relative to variables ξ_i (Translator's Note: In the original paper, Malécot has ζ_i instead of ξ_i in the footnote, which is an obvious typographical error).

one has:

$$\sigma^2 = M \left[\left(\frac{\delta \log \pi}{\delta \theta} \right)^2 \right] = M \left[\left(\frac{\delta \log dG}{\delta \theta} \right)^2 \right] + M \left[\left(\frac{\delta \log [dk]}{\delta \theta} \right)^2 \right]$$

also, the formula:

$$0 = \int (H - \theta) \pi = \int_{-\infty}^{+\infty} (H - \theta) dG \int_{\mathfrak{R}_{N-1}(H)} [dk] = \int_{-\infty}^{+\infty} (H - \theta) dG$$

gives again, by taking derivatives with respect to θ :

$$\zeta^2 \geq \frac{1}{M \left[\left(\frac{\delta \log dG}{\delta \theta} \right)^2 \right]}$$

ζ^2 cannot be equal to $\frac{1}{\sigma^2}$ unless

$$M \left[\left(\frac{\delta \log [dk]}{\delta \theta} \right)^2 \right] = 0$$

that is if $[dk]$ and, therefore, also k is independent of θ nearly everywhere, that is, if H is an exhaustive estimator; the general form of laws admitting an exhaustive estimator has been given by Mr. DARMOIS (3) and Mr. FRECHET has verified (8) that the exhaustive estimator meets the condition $\zeta^2 = \frac{1}{\sigma^2}$

The condition $\zeta^2 = \frac{1}{\sigma^2}$ ¹⁴ cannot be met for finite N unless an exhaustive estimator exists. However, Mr. FISHER had shown earlier (7) that it would always exist, or at least that the condition would be met asymptotically when $N \rightarrow \infty$, when an estimator is obtained by producing as a function of E a value of θ which maximizes the likelihood function $\pi(E|\theta)$, that is, by applying *the rule of maximum likelihood*; this estimator H_o , being C.A.G. under fairly wide conditions, and its fluctuation $\zeta_o^2 \propto \frac{1}{\sigma^2}$ being asymptotically smaller or equal than that of any other such estimators, would be in the limit one of the most precise C.A.G. estimators and would merit the name of *optimum estimator*. Its amount of information will be

$$I_o = \frac{1}{\zeta_o^2} \propto M \left[\left(\frac{\delta \log \pi}{\delta \theta} \right)^2 \right]$$

¹⁴ Translator's Note: This is a typographical error since the ξ' 's were defined as random variables. The correct expression is $\zeta^2 = \frac{1}{\sigma^2}$.

For any other C.A.G. estimator obtained from the same observations E and with amount of information $I = \frac{1}{\zeta^2}$, the ratio $\frac{I}{I_o} = \frac{\zeta_o^2}{\zeta^2} = \frac{1}{\sigma^2 \zeta^2}$, which is smaller or equal to 1, will be called "efficiency" of the estimator; it gives the loss of precision accruing from using an estimator other than the optimum.

We shall now give a rigorous and general presentation of Mr. FISHER's theory, extending results of Mr. DOOB and of Mr. DUGUE (5).

Let $g(x_i, \theta)$ be a function of random variable x_i and of the unknown parameter θ , and suppose that the N random variables $g(x_i, \theta)$ have true means for each value of θ that are "equally convergent", that is, that the N probabilities

$$P [|g(x_i, \theta)| > t]$$

have an upper bound given by a function $p(t)$ independent of i which generates a finite integral $\int_0^{+\infty} t dp(t)$. If we suppose that

$$\sum_{i=1}^N \frac{M[g(x_i, \theta)]}{N}$$

tends towards a limit $\varphi(\theta)$ as $N \rightarrow \infty$, for every value of θ in an interval A...B, the extension of a result of Mr. KOLMOGOROFF (9)¹⁵ shows that the quantity

$$\Psi(\theta, N) = \sum_{i=1}^N \frac{M[g(x_i, \theta)]}{N}$$

deduced from N observations x_1, \dots, x_N , tends almost surely, when $N \rightarrow \infty$, towards $\varphi(\theta)$. If one supposes that the $g(x_i, \theta)$ are almost surely functions of θ with variation bounded by the same fixed number K ("equally bounded variation", the same holding for $\Psi(\theta, N)$), an extension of POLYA-CANTELLI's theorem shows that when $N \rightarrow \infty$, $\Psi(\theta, N)$ converges almost surely towards $\varphi(\theta)$ in the interval A...B¹⁶, which means that the probability that

$$\varphi(\theta) - \eta < \Psi(\theta, N) < \varphi(\theta) + \eta$$

tends towards 1 as $N_o \rightarrow \infty$, whatever the value of θ is and for $N > N_o$ (η being an arbitrary, fixed, number).

¹⁵ Translator's Note: The English spelling is KOLMOGOROV.

¹⁶ This holds even if there are discontinuities (of the first kind) by considering, instead of the value of θ , the limiting values at right and left (supposed to satisfy the same conditions):

$$\varphi(\theta + o), \varphi(\theta - o), \psi(\theta + o, N), \psi(\theta - o, N).$$

In what follows, it will be convenient to represent by $\varphi(\theta)$ the set of values comprised between $\varphi(\theta - o)$ and $\varphi(\theta + o)$, and by $\psi(\theta, N)$ the set of values comprised between $\varphi(\theta - o, N)$ and $\varphi(\theta + o, N)$.

Consider now a root θ_0 of $\varphi(\theta)$, suppose that it can be found and that it corresponds to a change of sign of $\varphi(\theta)$: more precisely, suppose that in every interval $\theta_1 \dots \theta_2$ surrounding θ_0 there is at least one value between θ_1 and θ_0 for which $\varphi(\theta)$ is negative, and that there is at least one value between θ_2 and θ_0 for which it is positive. If we let η be the smallest of the two corresponding $|\varphi(\theta)|$ it follows from the preceding that, for $N > N_o$, the probability that all the $\Psi(\theta, N)$ change from positive to negative inside the interval $\theta_1 \dots \theta_2$ and, therefore, the values cancel each other (in view of the statement in the preceding footnote, for the points in which there is discontinuity), tends towards 1 when $N \rightarrow \infty$. Because the interval $\theta_1 \dots \theta_2$ in the neighborhood of θ_0 can be taken to be arbitrarily small, this means that the equation $\Psi(\theta, N) = 0$ admits at least a root converging almost surely to θ_0 when $N \rightarrow \infty$.

It is possible to go further if one supposes that the quantities $\frac{\partial g(x_i, \theta)}{\partial \theta}$ and, hence, $\frac{\partial \Psi}{\partial \theta}$ are almost surely uniformly continuous with respect to θ , with equally bounded variation in $A \dots B$, and that these have "equally convergent true means". It follows easily that $\frac{\partial \Psi}{\partial \theta}(\theta, N)$ converges almost surely and uniformly towards a continuous function which is surely the derivative of $\varphi(\theta)$, that is, $\varphi'(\theta)$ and then that one can associate to every ε an interval $\theta_0 - \alpha$ and $\theta_0 + \alpha$ such that the probability that

$$\left| \frac{\partial \Psi}{\partial \theta} - \varphi'(\theta) \right| < \varepsilon$$

for all $N > N_o$ and for all θ between $\theta_0 - \alpha$ and $\theta_0 + \alpha$ tends towards 1 when $N_o \rightarrow \infty$.

Now, from the formula of finite increments, these inequalities imply, for $N > N_o$ and for all θ between $\theta_0 - \alpha$ and $\theta_0 + \alpha$:

$$\Psi(\theta, N) = \Psi(\theta_o, N) + (\theta - \theta_o)(D + \varepsilon_i) \text{ with } |\varepsilon_i| < \varepsilon$$

(where D is the fixed number $\varphi'(\theta_o)$); this shows that the equations $\Psi(\theta, N) = 0$ will have, for $N > N_o$ and within the interval $\theta_0 - \alpha$ and $\theta_0 + \alpha$, a single root, and that this root will be each time between

$$\theta_0 - \frac{\Psi(\theta_o, N)}{D + \varepsilon} \text{ and } \theta_0 - \frac{\Psi(\theta_o, N)}{D - \varepsilon}$$

provided that these quantities take values between $\theta_0 - \alpha$ and $\theta_0 + \alpha$: this will be attainable with probability tending to 1 when $N_o \rightarrow \infty$ because $\Psi(\theta_o, N)$ tends almost surely to $\varphi(\theta_o) = 0$. Hence, it is seen that the equation $\Psi(\theta_o, N) = 0$ admits only one root θ_N tending almost surely to θ_0 ; the probability that (for each value of $N > N_o$) this root is equal to

$$\theta_0 - \frac{\Psi(\theta_o, N)}{D + \varepsilon_i}$$

with $\varepsilon_i < \varepsilon$, tends towards 1 when $N_o \rightarrow \infty$ irrespective of the value of ε . θ_N is then a correct estimator of θ_0 ¹⁷.

Let us make now the following additional assumptions: the N random variables $g(x_i, \theta_0)$ constitute a *normal family* in the sense of Mr. P. LEVY (for this, it suffices to suppose, using the notation of Mr. P. LEVY, that $\int_0^\infty t^2 dp(t)$ is finite, which implies that the fluctuations σ_i^2 of the random variables $g(x_i, \theta_0)$ are a bounded set and that the fluctuation $\sigma^2 = \sum_i^N \sigma_i^2$ of

their sum $\sum_i^N g(x_i, \theta_0) = N\Psi(\theta_0, N)$ increases indefinitely with N . It is known (P. LEVY, 11) that then the type of law of this sum tends to a Gaussian one, and one can deduce easily (DUGUE, 5) that this law is the same as that of

$$\theta_N = \theta_0 - \frac{\Psi(\theta_0, N)}{D + \varepsilon_i}$$

θ_N is, thus, not only a correct estimator of θ_0 but C.A.G. as well. Because $\frac{N\Psi(\theta_0, N)}{\sigma}$ has a law that tends towards a standard Gaussian one, this being the same for $\frac{ND}{\sigma(\theta_N - \theta_0)}$, the fluctuation of the estimator θ_N is then:

$$\zeta^2 = M [(\theta_N - \theta_0)^2] \propto \frac{\sigma^2}{N^2 D^2} = \sum_i^N \frac{\sigma_i^2}{N^2 \varphi^2(\theta_0)}$$

Here we have a very general procedure for obtaining C.A.G. estimators. If, in particular, we take as $g(x_i, \theta)$ pertaining to the *ith* observation the function

$$\frac{\partial \log dF_i(x_i, \theta)}{\partial \theta}$$

which has a null mean value when θ is equal to the true value θ_0 , giving $\varphi(\theta_0) = 0$, then the equation $\Psi(\theta, N) = 0$ becomes the equation of *maximum likelihood*

$$\frac{1}{N} \frac{\partial \log \pi(E|\theta)}{\partial \theta}$$

If the conditions of continuity and convergence given previously are met, this equation leads to a C.A.G. estimator, θ_{N_o} , with a fluctuation involving:

$$\begin{aligned} \sigma_i^2 &= M \left[\frac{\partial \log dF_i(x_i, \theta)}{\partial \theta_0} \right]^2 = \int \frac{\partial \log dF_i}{\partial \theta_0} \frac{\partial dF_i}{\partial \theta_0} \\ &= \frac{\partial}{\partial \theta_0} \left[\int \frac{\partial \log dF_i}{\partial \theta_0} dF_i \right] - \int \frac{\partial^2 \log dF_i}{(\partial \theta_0)^2} dF_i \\ &= -M \left[\frac{\partial^2 \log dF_i}{(\partial \theta_0)^2} \right] \end{aligned}$$

¹⁷ Translator's Note: Recall that *correct* means *consistent*.

which shows that $\sum \sigma_i^2 = -N\varphi'(\theta_o)$, from where

$$\zeta^2 \propto -\frac{1}{N\varphi'(\theta_o)} = \frac{1}{\sum \sigma_i^2} = \frac{1}{\sigma^2}$$

hence, for a sufficiently large N , $\zeta^2 \leq \frac{(1+\varepsilon)}{\sigma^2}$, the maximum likelihood estimator is among the estimators having a minimum fluctuation. Henceforth, we will call this an *optimal estimator*.

Suppose in particular that two sets with N_1 and N_2 observations, respectively, have been collected, and that the observations within each set follow the same law, that is, there are laws dF_1 and dF_2 . The maximum likelihood equation for the entire collection of observations is:

$$(N_1 + N_2) \Psi(\theta, N_1 + N_2) = N_1 \Psi(\theta, N_1) + N_2 \Psi(\theta, N_2) = 0$$

and put

$$N_i \Psi(\theta, N_i) = \sum \frac{\partial \log dF_i}{\partial \theta_0}$$

This gives the solution:

$$\theta_{N_1+N_2} - \theta_0 = -\frac{\Psi(\theta_0, N_1 + N_2)}{[\varphi'(\theta_o) + \varepsilon_1]} = -\frac{N_1 \Psi_1 + N_2 \Psi_2}{(N_1 + N_2) [\varphi'(\theta_o) + \varepsilon_1]}$$

Now:

$$(N_1 + N_2) \varphi'(\theta_o) = N_1 \sigma_1^2 + N_2 \sigma_2^2$$

If we let θ_{N_1} and θ_{N_2} be the estimators obtained from each of the two sets separately, one has

$$\theta_{N_1} - \theta_0 = -\frac{\Psi_1}{[\varphi'(\theta_o) + \varepsilon'_1]} \propto \frac{\Psi_1}{\sigma_1^2}, \text{ etc., hence:}$$

$$\theta_{N_1+N_2} - \theta_0 \propto \frac{N_1 \sigma_1^2 (\theta_{N_1} - \theta_0) + N_2 \sigma_2^2 (\theta_{N_2} - \theta_0)}{N_1 \sigma_1^2 + N_2 \sigma_2^2}$$

The optimum estimator for the entire data set is, thus, the weighted average of the optimum estimators obtained from each of the individual sets, with the weights being $N_1 \sigma_1^2$ and $N_2 \sigma_2^2$, that is, the reciprocal of the fluctuations ζ_1^2 and ζ_2^2 ("quantities of information") of the two estimators. One finds the classical rule for combining observations deduced by Gauss from a principle identical to that of maximum likelihood.

This result highlights again that the rule of maximum likelihood is not valid if applied to only a part of the observations, as the only result worth keeping is that pertaining to the entire set of observations. The rule of maximum

likelihood is just a particular case of the rule of the most likely value¹⁸; that is the special case where any information about θ comes through the observations E , while knowledge K obtained previously does not contribute at all, so an uniform prior probability is assigned to θ . Furthermore, it must be observed, with Mr. JEFFREYS, that if one takes any *continuous* probability law for θ , $h(\theta) d\theta$, having *continuous* first and second derivatives, the effect of this law on the estimator obtained using the rule of the most likely value with N independent observations is negligible as $N \rightarrow \infty$. In fact, if we let E denote the set of such N observations, and let $\pi(E|\theta)$ be the corresponding likelihood function, the posterior probability of a value θ will be $\pi(E|\theta) h(\theta) d\theta$, so the most likely value will, thus, be the root of the equation

$$\frac{\partial \log \pi}{\partial \theta} + \frac{\partial \log h}{\partial \theta} = 0$$

from where, putting $\frac{\partial \log h}{\partial \theta} = l(\theta)$:

$$\Psi(\theta, N) + \frac{1}{N}l(\theta)$$

and, rearranging the calculations on page 54¹⁹ slightly, the estimator based on the most likely value is

$$\bar{\theta}_N = \theta_0 - \frac{\Psi(\theta_0, N) + \frac{1}{N}l(\theta_0)}{\varphi'(\theta_0) + \frac{1}{N}l'(\theta_0) + \varepsilon_1}$$

If $h(\theta_0) \neq 0$, $l(\theta_0)$ and $l'(\theta_0)$ are bounded, so when $N \rightarrow \infty$, $\bar{\theta}_N - \theta_0 \approx \theta_N - \theta_0$, with θ_N being the maximum likelihood estimator; the influence of the prior probability law becomes negligible. However, it must be emphasized that for large but finite N this influence is negligible only if $l(\theta_0)$ and $l'(\theta_0)$ are sufficiently small relative to N ; on the other hand, if $l'(\theta_0)$ is of the order of N , that is, if the curve representing $\log h$ and, hence, that representing $h(\theta)$ (elementary prior probability)²⁰ has a sharp peak, this is not so; it is patent, furthermore, that in this case, with the observations K made before E , having already given precise information about θ , then the maximum likelihood

¹⁸ Translator's Note: MALÉCOT refers to the mode of the posterior distribution.

¹⁹ Translator's Note: The reference is to the page of the *original* paper. MALÉCOT is pointing out towards the developments leading to:

$$\theta_0 - \frac{\Psi(\theta_0, N)}{D + \varepsilon_i}$$

in connection with maximum likelihood estimation.

²⁰ Translator's Note: The meaning of elementary, an adjective used often by French mathematicians, is unclear here. Presumably, MALÉCOT means density, an infinitesimally small element of a probability (in the continuous case).

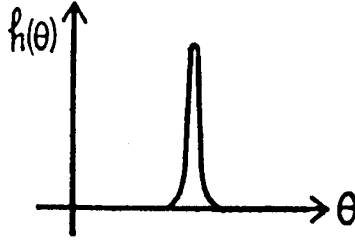


Figure 1.

estimator θ_N deduced from only E , is not the best; it is necessary to combine E with the previous observations by applying the rule of the probable value²¹, which gives the value $\bar{\theta}_N$.

Because the mean value of $\bar{\theta}_N$ is

$$\theta_0 - \frac{\frac{1}{N}l(\theta_0)}{\varphi'(\theta_0) + \frac{1}{N}l'(\theta_0) + \varepsilon_1}$$

with ε_1 being almost surely uniformly small with $\frac{1}{N}$, its fluctuation will be

$$\bar{\zeta}^2 \approx \frac{M[\Psi(\theta_0, N)]^2}{\left[\varphi'(\theta_0) + \frac{1}{N}l'(\theta_0)\right]^2} = \frac{\sum \sigma_i^2}{\left[-\sum \sigma_i^2 + l'(\theta_0)\right]^2} = \frac{\sigma^2}{[\sigma^2 - l'(\theta_0)]^2}$$

This can be larger or smaller than $\zeta^2 = \frac{1}{\sigma^2}$ (fluctuation of θ_N) depending on whether $l'(\theta_0)$ is > 0 or < 0 , that is, depending on whether the true value θ_0 lies in the neighborhood of a "valley" or of a "peak" of the curve representing the prior probability $h(\theta)$. In the case where $\bar{\zeta}^2 < \zeta^2$, there is no contradiction with the result given on page 50²², because this result establishes that ζ^2 is the minimum fluctuation for all estimators H such that $M(H) \equiv \theta$ for any θ ; it can be expected that when one does not have any prior knowledge about the true value θ_0 of θ the precision of the best estimator will be ζ^2 . On the other hand, if one knows that a value θ_0 is more probable than others, the condition $M(H) = \theta$ for any θ can be a nuisance²³ and give less precision than when would try to estimate in a region near the most probable value.

²¹ Translator's Note: The author probably means "the most probable value".

²² Translator's Note: This is the page of the original paper where the lower bound for the variance of an unbiased estimator is presented.

²³ Translator's Note: MALÉCOT employs the term "parasite". Although descriptive, such a term is not a part of the statistical lexicon in English.

4. THE PROBLEM OF INDUCTION

The decreasing importance of the prior probability as the number of observations increases describes certain aspects of the problem of induction in a remarkably clear manner. This problem consists essentially of extracting from the results observed a law summarizing them (and which also allows to forecast future results); this law is never dictated by the observed results, rather, it is a construction of the mind chosen for reasons of simplicity or convenience (naturally taking into account all previous experience); one can always suppose many laws; these play the role of the different hypotheses θ_i of our scheme; each of these, if formulated with sufficient precision, generates the observed results E with a known probability $P(E|\theta_i K)$, the likelihood of θ_i . The choice between the θ_i is dictated by the posterior probabilities $P(\theta_i|EK)$, depending both on the likelihoods, which are objective (because these depend only on the observations) and on the prior probabilities $P(\theta_i|K)$ which are more or less subjective; the evaluation of likelihoods is deductive (often in its more refined form, the mathematical deduction); however, the subjective part always enters in the evaluation of prior probabilities, illustrating wonderfully that every induction is subjective. It is true that when the number of observations increases, the subjective part decreases, as we saw previously. Further, the prior probabilities can be right away in more or less agreement with subsequent experience; when KEPLER viewed as very probable that an ellipse would fit his observations on Mars, he was in immediate agreement with all subsequent astronomical observations; on the other hand, the a priori belief that planets moved in circles around the earth led PTOLEMY and his predecessors to formulate laws which, by integrating all past observations, made difficult, because of their complexity, to predict subsequent observations. The scheme a priori was excessively subjective and had to be updated constantly in order to account for new observations. These examples show that as science progresses, that is, as new observations accumulate, its subjective part diminishes, although it would be an illusion to believe that it could be eliminated totally. In fact, experimental progress always allows us to choose, in the long run, between several hypotheses that have been *formulated completely* (by evaluating their likelihood deduced from all observations made), but we will always be incapable of formulating *precisely* (that is, making their consequences explicit) *all* possible hypotheses and, consequently, of calculating the likelihoods of *all* hypotheses. This is the reason why every law, every possible physical theory, will always become inadequate for explaining new facts: it has been chosen as the most likely of all the laws *among those that can be formulated*, but more advanced experimentation will make it appear less likely than *new laws that one would be led to formulate*; in this form, the system of PTOLEMY was replaced by that of KEPLER-NEWTON, and then by relativist mechanics. Each law is valuable for representing both the old field of observations and the new field motivating it; however, the law cannot pretend to represent the totality of future observations, because it is not more than a choice between a small number of laws that *our mind conceives and*, because of the weakness of our senses and of our mind, these laws are rough and incomplete blueprints of the rich complexity of natural phenomena. Of course, as experiences develop, the increasing finesse of our theories molds reality better but cannot pretend to grasp it completely. "There are more things

in heaven and earth than in all our philosophy". There is more complexity in the mechanisms of nature than we can think of and all the laws that we can construct, even if better than the preceding ones, are just an approximation to reality, an approximation that will become insufficient, eventually. OHM's law, although translating electrodynamic phenomena remarkably to our scale, becomes inadequate when an extension of our senses places us at the scale of the electrons, so it becomes just a statistical law. Is it not possible that even the laws of atomic physics behave eventually as statistical laws? A scientific law is never "true", that is, a definitive one, it is only more or less convenient for representing and anticipating phenomena viewed at a certain scale. When it is said that "a physical theory is justified by its consequences", this only has a relative meaning, that is, that among all theories formulated, this is the one having consequences that agree best with the observations. In induction, there are two very distinct parts: a deductive part that formulates the consequences of each hypothesis considered, and a part that is not amenable to deduction and which postulates hypotheses and assigns prior probabilities to these; there is where the genius of invention and the mind are manifested; then, the rest consists in choosing the most probable hypothesis after the consequences. The rule of the "most probable hypothesis" underlies every induction, translating precisely the logic of induction and, at the same time, highlighting its subjectivity. It does not seem possible to take the rule of maximum likelihood as a base of the logic of induction, as Mr. FISHER does, because this rule applied to different series of measurements will lead to contradictory consequences (and must be completed using significance tests, which are in contradiction with this rule!), while a logic must be a set of principles from which one can accept all consequences, this being certainly the case, as we have argued, for a logic based on BAYES formula.

5. "SUBJECTIVE" AND "OBJECTIVE" PROBABILITIES

If, with Mr. DE FINETTI (6), we view probability theory as a "logic of subjective judgements", how is it possible to have an agreement between statements derived from this logic and the objective reality? This is the objection made frequently to the formula of BAYES. The arbitrary form in which prior probabilities are evaluated confers a similar arbitrariness to the evaluation of posterior probabilities. Now, aren't there events whose probabilities have an objective meaning, as suggested by an agreement between observed frequencies and probabilities assigned by an a priori reasoning? We believe that the remarks made previously permit responding to this objection. Every evaluation of probabilities is a construct of the mind, and relative to a theoretical setting imagined by the mind to limit our ignorance, and based on the principle of indifference. For example, the statement that the value 6 in the toss of a die has a probability of $\frac{1}{6}$ is, at the same time, the result of ignorance about the movement of the die in the dice-box, and of the statement that there is no reason to believe that this movement favors a side over the others, hence all sides are equi-probable. This is relative to a certain theoretical scheme, to a certain hypothesis: a perfect die tossed fairly. Others may make a very different evaluation, by admitting a personal influence of the "lucky" player on the values

observed. At any rate, in the evaluation of probabilities, there will always be hypotheses a priori that, although more or less suggested by previous observations, will never dominate absolutely, will never be certain a priori, this being so because it is never possible to know the totality of circumstances giving rise to a phenomenon. (In passing, we dismiss the objection that it is not possible to speak about "probabilities of causes" because these would not be "random", one must be "true" and the others "false": if one admits determinism, the same is true of the effects; in fact, it is not the phenomena that are random, rather, it is the knowledge that we have about them; the probabilistic logic attempts to identify the limits of our ignorance). The role of experimentation is to confirm or question some of the assumptions made or, more generally, to update their probabilities; if one of these appears clearly as more probable than the others, it would be retained as the best, but it should be kept in mind that this superiority is temporary, and that the hypothesis could be demolished by subsequent experimentation. For example, consider games of chance, such as playing dice, to illustrate ideas. Experience has led us to abandoning the hypothesis, which perhaps may be natural for a primitive mind, that there is an influence of the player on the outcome, and to adopting the assumption that all sides of the die are equally likely, as the best explanation for the observed results. However, Weldon's experiments show, in turn, that this assumption is false, as the theoretical scheme of the perfect die does not hold in practice; there are always some sides that are favored: the probability of $\frac{1}{6}$ is then relative to a theoretical scheme deduced from reality by abstraction and simplification, and it will never be the limit of the observed frequencies.

What makes the theoretical scheme appealing is its convenience: with everything kept simple, it summarizes with sufficient precision the main aspects of an experiment, and it can be expressed through formulae that are simple and, at the same time, that allow making forecasts having a good precision. As it has been stated by Mr. DARMOIS (2): "making a probability calculation in a specific case, requires seeing clearly all that it is necessary to know, such that the study follows closely the *essential circumstances* of the phenomenon considered". Thus, the evaluation of a probability always results from a theoretical scheme permitting to assess, with more or less precision, the equal or unequal probability; it is completely legitimate, as stated by Mr. BOREL, to evaluate the probability of an isolated event provided that a scheme can be conceived where this probability is related to other known ones (for example in a lottery scheme)²⁴. However, the probabilities thus calculated will not be in reasonable agreement with the observed frequencies unless the theoretical scheme is in sufficient agreement with the real mechanism, for example, the equi-probable cases corresponding with the equally frequent cases, and this will happen when the scheme has been established after considering a sufficiently large number of experiments. It is in this situation that an "agreement between

²⁴ Translator's Note: It is unclear what MALÉCOT means here. In the original paper, he stated: "*Ainsi l'évaluation d'une probabilité résulte toujours d'un schéma théorique permettant d'évaluer, avec plus ou moins de précision, l'égalité ou l'inégalité de probabilité ; il est tout à fait légitime, comme le remarque M. BOREL, d'évaluer la probabilité d'un événement isolé dès qu'on peut concevoir un schéma ramenant cette probabilité à d'autres connues (par exemple, un schéma du tirage au sort)*".

individual opinions" (DE FINETTI) or an "agreement between equally well informed minds" will be obtained, a condition that Mr. BOREL confers to an "objective probability" (which, furthermore, is not a sufficient condition because errors of judgment or of expertise can be committed unanimously).

On the other hand, if the scheme is established from a weak knowledge about facts, the probabilities that can be deduced have the risk of not bearing any relationship with reality. This is what makes Mr. DE FINETTI to write: "if one does not want to take subjective factors into account explicitly, the question should be abandoned, by stating that it is not sensible". This is scarcely a reason—the opposite, rather— for rejecting the formula of BAYES, since there is a need for adopting a position (DE FINETTI, (6), p. 26)²⁵. The question brings into perspective the subjectivity of this view, as it was done in the linkage example. Also, the criticism of the formula made by Mr. NEYMAN (15) is somewhat surprising. Mr. NEYMAN takes as an example a set of individuals I , all dominant for a Mendelian factor²⁶; it is wished to use those having the homozygote genotype AA , and to discard the hybrid types (Aa); to do this, each I is crossed with an aa , and the k descendants from this cross are observed; if aa types are observed within these, then I is discarded, naturally; on the other hand, I is kept if the k descendants are of the dominant type. However, in so doing, some of the individuals I kept will be of the undesirable type Aa ; the problem is the evaluation of the risk of such an error. Because an I of the AA type produces only dominant descendants, and an I of the type Aa gives k descendants that are all dominant with probability $\left(\frac{1}{2}\right)^k$, the posterior probability of keeping an I which will be Aa , using BAYES formula, and letting p_0 be the prior probability that I is Aa will be:

$$p_1 = \frac{p_0/2^k}{1 - p_0 + p_0/2^k} \quad (2)$$

It is clear that if p_0 is "objective", that is, if it reflects an observable frequency, then p_1 provides a forecast of the frequency of errors. If, for example, it is known that the I individuals come from crossing heterozygotes, one would take $p_0 = \frac{2}{3}$, representing the frequency of heterozygotes in a large number of individuals I examined. Then:

$$p_1 = \frac{1}{1 + 2^{k-1}}$$

would sensibly represent the proportion of individuals that, although kept, possess the Aa type, that is, the proportion of errors. However, if the origin of I and, hence p_0 , is unknown, the equation evidently loses part of its specific meaning. Should one, then, with Mr. NEYMAN, declare it useless?²⁷. It is clear

²⁵ Translator's Note: I have translated "*adopter une ligne de conduite*" as "*for adopting a position*".

²⁶ Translator's Note: Although perhaps obvious from the context, MALÉCOT means that the set I includes individuals with at least a copy of the allele A .

²⁷ Translator's Note: The author refers to BAYES formula here.

at the onset that no other formula, in the absence of additional experiments, can give us the proportion of errors, because from the equation, this is linked to p_0 , and this is unknown. Any estimation of error needs a judgement, explicit or not, about the value of p_0 , and in the formula of BAYES this judgement must be made explicit. The formula shows, for example, that if $k = 6$, the statement that there is at least 1 error in 65 is equivalent to stating that p_0 is $\leq \frac{1}{2}$, which may or may not be viewed as reasonable depending on the information available about how the individuals I were obtained. None of the two statements has a stronger foundation than the other, and any reasoning attempting to give more credibility to the preceding one would be erroneous. BAYES formula, establishing an exact correspondence between the "prior" and the "posterior" probabilities shows clearly that a judgement based on the latter one is equivalent to a judgement on the former ones, and that this is unavoidable, except in some special cases to be discussed in Section 7. Further, this formula has value for the interpretation of subsequent experiments: if these involve a genetic analysis of the individuals I kept, from which it follows that the frequency of errors can be evaluated, this leads to an "objective" value of p_1 , that is, of the composition of the initial population, information which may be precious for other experiments.

6. NEYMAN'S POINT OF VIEW

After having shown that the statistical ideas advanced by Mr. Fisher's school of thought cannot be justified logically without introducing the "rule of the most probable value" deduced from BAYES formula, we will consider now the methods with which Mr. NEYMAN has thought it is possible to bypass this formula while providing "objective" criteria, expressible in terms of frequencies. The problem, as posed by Mr. NEYMAN, is to decide if a hypothesis H_0 is to be "rejected" or "accepted" according to whether the point E having as coordinates the N observed values x_1, \dots, x_N , is found inside of a certain "critical region" w or inside of a complementary region \bar{w} of the N -dimensional space \mathfrak{R}_N ("observations space") (classical examples: significance of the difference between a theoretical mean and an observed mean, by comparing their difference with their standard error; assessment of goodness of fit with the χ^2 method). This decision can produce an error in two different manners: if H_0 is rejected when it holds true, one makes a *type-1 error* (the only one that is classically taken into account in the two preceding examples). If one accepts H_0 when it is false, a *type-2 error* results. The idea of Mr. NEYMAN is evaluating the probabilities of these two errors separately and "objectively", that is, to predict their frequencies (by deduction and not by induction, as emphasized by Mr. NEYMAN).

Consider the case where the hypothesis to be examined concerns the value of a parameter θ intervening in the probability law $f(x, \theta)$ taken for each observation x . Because the function f is supposed to be known, one can calculate, as a function of θ , the probability that the point x_1, \dots, x_N falls in the critical region w . This probability, $P(E \subset w | \theta) = \beta(\theta, w)$ is called "power function" of the criterion based on w . If the hypothesis H_0 to be examined attributes a value θ_0 to the parameter, the probability of a type-1

error calculated under hypothesis H_0 will be $\beta(\theta_0, w)$, and that of a type-2 error, calculated supposing that the true value is θ_1 will be $\beta(\theta_1, \bar{w}) = 1 - \beta(\theta_1, w)$. Mr. NEYMAN proposes first to reduce the probability of errors of the first type to a fixed, sufficiently small value, α , defining a family of "equivalent critical regions" w in terms of the formula $\beta(\theta_0, w) = \alpha$: then, attempt to choose one of these regions such that the type-2 error is as small as possible, and this for any θ_1 in a certain domain; hence, this defines a criterion that is "uniformly most powerful" in this domain (but this criterion exists only for very specific laws f and, provided that the domain is restricted sufficiently. This is the reason why the domain is often restricted to the neighborhood of θ_0).

Our first criticism is as follows: why would one want first to minimize the type-1 error? Mr. NEYMAN points out to a case where the consequences of a type-1 error would be much more important than those of a type-2 error: for a pharmacological product which, by accident, can contain a toxic substance, and which has been assayed previously on some animals, it is essential not to discard the hypothesis H_0 : "the product is dangerous", because it is accurate; however, the consequences are not serious if this hypothesis is kept, even if it is false; the problem is, then, essentially, one of reducing the type-1 error. However, this is a very particular situation. In general, the cases where one will be concerned about the type-1 error are those where a priori there are strong reasons to believe that H_0 is accurate: in fact, reducing the type-1 error leads, most of the times, to an increase of the type-2 error in the neighborhood. If one can vary θ in a continuous manner and if $\beta(\theta, w)$ is a continuous function of θ , the two errors become evident in the curve representing the function, because the corresponding probabilities are, respectively, the ordinate at abscissa θ_0 (where θ_0 is the value under scrutiny) and the complement to 1 of the ordinate with abscissa θ_1 ($\theta_1 =$ true value); even if the region w is chosen such that one has a uniformly most powerful criterion, in those rare cases where it exists, it is still true that a reduction of α will cause in general a reduction of the neighboring coordinates, that is, an increase of the type-2 error, provided the true value θ_1 is not too far from the value θ_0 under scrutiny. For example, in the estimation of linkage, it is frequent to reject the hypothesis $r = \frac{1}{2}$ if the estimate of r obtained from the experiments it is away from $\frac{1}{2}$ by more than λ times its standard error. The larger λ is, the smaller the risk of rejecting the hypothesis $r = \frac{1}{2}$ if this holds; however, there will be some risk of discarding the hypothesis that r has a value other than $\frac{1}{2}$ but near $\frac{1}{2}$ when this hypothesis is true. In general, the weight to be assigned to the two types of error, that is, the choice of α , depends inevitably on assumptions made a priori about the probabilities of H_0 and of the other hypotheses. The method of Mr. NEYMAN cannot pretend to give an "objective" judgement about H_0 ; its appeal resides in making the distinction between the two distinct classes of error, but it is incapable, in the absence of any consideration a priori, of assigning appropriate weights to the two; now, the more clear manner of incorporating a priori considerations is to introduce prior probabilities; if these are subjective, so be it.

Let us go further: this method not only does not permit to evaluate the global frequency of errors in the absence of knowledge of prior probabilities,

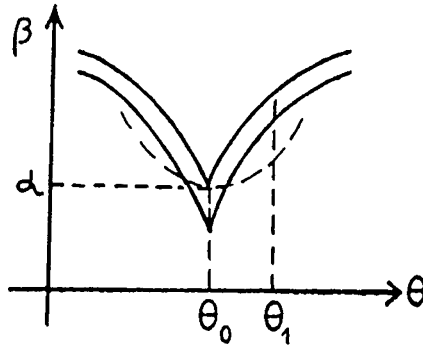


Figure 2.

as acknowledged by Mr. NEYMAN, but it does not allow evaluation of the frequency of errors of each type and, contrary to what seems to be stated by Mr. NEYMAN, it does not furnish any observable frequency. In fact, $\beta(\theta_0, w)$ just measures the frequency of errors of the first type that would take place if H_0 were always true; $1 - \beta(\theta_1, w)$ measures the frequency that the errors of the second type would have provided the hypothesis $\theta = \theta_1$ were always true; now, in practice, we do not have any certainty about these hypotheses, this being precisely the reason why we wish to arrive at a probabilistic judgement about these; hence, we are incapable of predicting to what extent the real frequencies of these errors correspond to the preceding probabilities unless, naturally, one knows for the different values of θ the "objective" prior probabilities, that is, expressible in terms of frequencies.

Let K be the prior probability that the hypothesis $\theta = \theta_0$ holds and $(1 - K) dg(\theta_1)$ (STIELTJES' differential) be the prior probability that $\theta = \theta_1 \neq \theta_0$ ($\int_L dg(\theta_1) = 1$, with L denoting the domain of variation of θ_1 , excluding θ_0); the posterior probabilities, when it is known that the observations have given a result falling in w , are respectively proportional to:

$$K\beta(\theta_0, w) \text{ and } (1 - K)\beta(\theta_1, w) dg(\theta_1)$$

giving as posterior probabilities of the errors of the first and second types:

$$P_1 = \frac{K\beta(\theta_0, w)}{K\beta(\theta_0, w) + \int (1 - K)\beta(\theta_1, w) dg(\theta_1)}$$

(probability that H_0 is true given that the observations fall in w , leading to rejection of H_0).

$$P_2 = \frac{\int (1 - K)\beta(\theta_1|\bar{w})^{28} dg(\theta_1)}{K\beta(\theta_0|\bar{w}) + \int (1 - K)\beta(\theta_1|\bar{w}) dg(\theta_1)}$$

(probability that H_0 is false given that the observations fall in to \bar{w} , leading to acceptance of H_0).

²⁸ Translator's Note: Without warning, MALÉCOT changes the notation $\beta(\theta, w)$ to $\beta(\theta|w)$ hereinafter.

The posterior probability of any error is:

$$P = K\beta(\theta_0, w) + \int (1 - K)\beta(\theta_1|\bar{w}) dg(\theta_1)$$

It is seen that the prior probabilities (K and $g(\theta)$) intervene in an essential manner in the expected frequencies of the two errors and in the weights to be assigned to these. The coefficients by which $\beta(\theta_1|w)$ ²⁹ and $\beta(\theta_1|\bar{w})$ must be weighted are the prior probabilities K and $(1 - K) dg(\theta_1)$; the choice of the size of α , for which Mr. NEYMAN does not offer any guidance, is implicitly equivalent to an assumption about the prior probability K of θ_0 ; by considering only the type-1 error and minimizing α (as in the usual case of evaluating the significance of deviations, or in the χ^2 test) this is equivalent to supposing that K is close to 1 so that $(1 - K) \int \beta(\theta_1|\bar{w}) dg(\theta_1)$ in P is negligible relative to $K\alpha$ (although the value of the integral, ranging between $1 - \alpha$ and 0 in the usual case where $\beta(\theta|w)$ is minimum for θ_0 , can be of the order of $1 - \alpha$ for certain laws of the prior probability $dg(\theta_1)$)

7. THE "CONFIDENCE INTERVALS"

The problem has been addressed in a different form by several authors, and by Mr. NEYMAN in another report (13). We shall modify the presentation of his theory by introducing prior probabilities. Let $dg(\theta)$ be the prior probability of an unknown parameter intervening in the probability law of the random variable under study (this parameter can vary within an interval $a \dots b$ which we shall denote as L), and let E_i ($i = 1, 2, \dots, n$) be the different possible outcomes (these being mutually exclusive) of the set of possible experiments involving this random variable. For each possible E_i we introduce a corresponding "estimating set" (supposed to be measurable) Θ_i contained in L , and we shall agree that if E_i is observed, the true value of θ will be regarded as belonging to the corresponding Θ_i . If Θ_i is an interval, we shall refer to it as a "confidence interval" associated to E_i .

(The situation in Section 6 was one where the E_i were distributed only into two categories, w and \bar{w} , and where the corresponding estimating sets are $\theta \neq \theta_0$ and $\theta = \theta_0$, thus non-overlapping; what it is different now is that the estimating sets Θ_i corresponding to the different values of i can overlap).

Let again $\pi(E_i|\theta)$ denote the probability of observing E_i when the parameter has value θ ; the total probability of observing E_i is

$$P_i = \int_{-\infty}^{+\infty} \pi(E_i|\theta) dg(\theta);$$

BAYES formula gives as posterior probability that θ is not in Θ_i (i.e., that it belongs to the complementary set $L - \Theta_i$), given that E_i has been observed:

$$Q_i = \frac{\int_{L-\Theta_i} \pi(E_i|\theta) dg(\theta)}{P_i};$$

²⁹ Translator's Note: MALÉCOT probably means $\beta(\theta_0|w)$.

consequently, the total prior probability that the rule " θ is in Θ_i when E_i has been observed" leads to a false statement is:

$$\gamma = \sum_i P_i Q_i = \sum_i \int_{L-\Theta_i} \pi(E_i|\theta) dg(\theta) \tag{3}$$

The interesting aspect of this formula is that, by choosing the Θ_i conveniently, is it possible to arrange it such that γ is always smaller than a fixed limit, *irrespective of the prior probability law $g(\theta)$ of the parameter*; suppose that when θ varies in the interior of $L - \Theta_i$, $\pi(E_i|\theta) \leq \delta$, with δ being a limit *independent of i* , which can be reduced arbitrarily by reducing the $L - \Theta_i$; the formula of the mean then gives that

$$\gamma \leq \delta \left[\sum_i \int_{L-\Theta_i} dg(\theta) \right]$$

and the sum inside the brackets cannot increase when the sets $L - \Theta_i$ are reduced and, hence, in particular, when δ is reduced; hence, this can be made arbitrarily small, which proves the statement. Therefore, one can always *choose the Θ_i* such that, *without knowing anything about $g(\theta)$* , it is assured that the probability that the rule adopted leads to an error that is smaller than a *fixed number ε* , hence, on average, one will make mistakes in a proportion of experiments that is smaller than ε . Thus, one can speak of an "objective" probability of error and "independent of the prior probabilities"; however, it should be pointed out that limiting "objectively" the probability of error has a penalty in terms of reduced precision of a statement concerning θ ; first, by use of the rule stated, we arrive only at the statement " θ is in a given set" and not: " θ has a specific value"; then, if the objective of the experiment is to judge a specific value of θ deduced from a theory, or to obtain a numerical value permitting subsequent evaluations, this value can be examined only in the light of certain prior probabilities, as we established in Section VI. Besides, even if one is satisfied with giving an indeterminate answer within a certain set, it must be noted that the sets Θ_i corresponding to the different results E_i could have considerable overlap, and in some cases there could be a part common to all Θ_i ; hence, the method will often be unable to choose, after the experiment, *one set* from a collection of overlapping sets, but will just allow to keep after the experiment a certain number of sets from this group without being able to choose among these (perhaps even some of these sets will *never* be rejected, irrespective of the results!). Nevertheless, these remarks should not make loose sight of the attribute of the method, which is to provide *an upper limit for the probability of error that is completely independent of the prior probabilities*, a limit which will be usable only in the case where we do not know absolutely anything about the latter.

The result is extended easily, by modifying the notation slightly, to the case where all the possible results form a measurable continuum \mathfrak{S} in a space \mathfrak{R} . If one lets $\pi(E|\theta) dE$ be the probability that when the parameter has value θ a result belonging to an element with volume dE is observed around a point E , and $\Theta(E)$ be the estimating set (supposed to be measurable) associated with

E , and if one adopts the rule "state that when one observes E , then θ is in $\Theta(E)$ ", the prior probability that this statement will be false is:

$$\gamma = \int_{\mathfrak{S}} dE \int_{L-\Theta(E)} \pi(E|\theta) dg(\theta) \quad (4)$$

To be more specific, let us adopt the presentation of Mr. NEYMAN, and put in brackets a generalization of his statements. Let E be the experimental point (set of N observations x_1, x_2, \dots, x_N) describing a continuum \mathfrak{S} in a space \mathfrak{R}_N ; to each value θ_0 of the parameter we associate an "acceptance set" $A(\theta_0)$ "of size equal [or larger than] to α ", which by definition is a measurable set (function of θ_0) of points in \mathfrak{R}_N chosen such that the probability that E belongs to this set, calculated under the hypothesis $\theta = \theta_0$, is equal [or larger than] to α . Further, associate to each experimental point E the set $\Theta(E)$ of values of θ_0 for which $A(\theta_0)$ contains E ; this set $\Theta(E)$ will be called "estimating set of θ , with a confidence coefficient equal [or larger than] to α ". If, for each E observed, we agree to state that the true value of θ is in the interior of the corresponding $\Theta(E)$, it is easy to show that the total prior probability that this rule leads to an error is independent of the prior probability of θ and is equal [or smaller than] to $1 - \alpha$. In fact, this probability γ is given by the above formula, that is, by a multiple integral over the domain:

$$\begin{array}{ll} \theta \in L - \Theta(E) & \text{or equivalently } E \in \mathfrak{S} - A(\theta) \\ E \in \mathfrak{S} & \text{or equivalently } \theta \in L \end{array}$$

(because there is a logical equivalence between the two propositions: " E is not a part of $A(\theta)$ " and " θ_0 is not a part of $\Theta(E)$ ") enabling us to write also:

$$\int_L dg(\theta) \int_{L-\Theta(E)} \pi(E|\theta) dE$$

However the integral to the right is, for any θ , by definition of $A(\theta)$, smaller or equal to $1 - \alpha$, the same holding for γ , thus completing the proof.

This proof puts in evidence, better than that of Mr. NEYMAN, the class of trials on which the probabilities are defined: is the set of all possible trials from all possible values of θ distributed according to an unknown law $dg(\theta)$. Mr. NEYMAN uses well the logical equivalence between the 2 propositions noted above, but he does not emphasize that this does not imply the equality of their probabilities unless these are defined over the same class of trials. For example, this would not give the probability of error in the set of cases where we observe a *given* E_i event (selection of results), because, from the formula on p. 68³⁰ giving Q_i , it would be necessary to know the prior probability of this event. If there is any conceptual confusion concerning the probability $1 - \alpha$ attached to a confidence interval, it is because there is an incomplete definition of the corresponding category of trials. It seems to us that one must see there a posterior probability of error calculated over the set of all possible trials,

³⁰ Translator's Note: Page 68 of the *original* paper.

and independently of the prior probability of θ , thus "objective". What Mr. FISHER cautiously calls "fiducial probability" is a true probability, as rightly observed by Mr. NEYMAN.

There is a well known application of this theory, this being the rule of "STUDENT". If the x_i are N observed, independent, values with mean \bar{x} of the same random variable following the law of LAPLACE-GAUSS with unknown expectation θ , we can take as estimating set with a confidence coefficient α the "confidence interval"

$$\bar{x} - ts \dots \bar{x} + ts, \quad \text{with } s = \sqrt{\frac{\sum (x_i - \bar{x})^2}{N - 1}}$$

with t linked to α through the formula:

$$\alpha = \frac{\int_0^t \left[1 + \frac{t^2}{N - 1} \right]^{-\frac{N}{2}} dt}{\int_0^\infty \dots}$$

The statement that θ belongs to such interval would give a frequency of errors equal to $1 - \alpha$, over a long series of experiments of the same type, and where there is no selection of results.

The theory of confidence intervals can be combined with that of estimation. Often, for a parameter θ with unknown true value θ_0 , one possesses an estimator E deduced from a large number N of observations, that it is correct³¹, asymptotically Gaussian, and with a known standard error, which is a function of θ_0 , that is, $\sigma(\theta_0)$. The interval $\theta_0 - \lambda\sigma(\theta_0) \dots \theta_0 + \lambda\sigma(\theta_0)$ is for E an acceptance set of size α connected to the "critical coefficient" λ by the formula

$$\alpha = \frac{2}{\sqrt{2\pi}} \int_0^\lambda e^{-\frac{x^2}{2}} dx;$$

if, within the interval where θ_0 can vary, $\sigma(\theta_0)$ admits an upper limit σ , it is seen that the interval $E - \lambda\sigma \dots E + \lambda\sigma$, entirely determined by the observations, will be a confidence interval for θ , with a confidence coefficient larger or equal than α .

In particular, if E is the maximum likelihood estimator, hence one of those minimizing $\sigma(\theta_0)$, and if the margin of uncertainty about θ_0 is small enough such that $\sigma(E)$ is not too different from $\sigma(\theta_0)$, *the interval $E - \lambda\sigma(E) \dots E + \lambda\sigma(E)$ will give a confidence interval of size α for θ* , and it will be, among all confidence intervals of size α derived from different C.A.G. estimators, the smallest one. This is why the rule indicated has practical value, by giving a maximum reduction of the uncertainty about θ while maintaining an "objective" probability of error (besides, as suggested already in Section 2, this rule has the effect of grouping the value with maximum likelihood, very

³¹ Translator's Note: This means *consistent*, as seen earlier.

unlikely by itself, with the neighboring values; however, we have now replaced the consideration of posterior probabilities of different values, which depend on the prior probabilities, by that of the total probability of error, which does not depend on these).

Nevertheless, it must be pointed out that possessing certain information about the prior probability of θ is necessary and sufficient to reduce even more the interval without increasing the probability of error $1 - \alpha$. In particular, one could not logically take a specific value of the interval without making assumptions, explicitly or not, about the prior probabilities. If, for example, an interval containing an integer value of θ has been obtained, adopting this value of θ rather than the estimate E will often depend on theoretical considerations a priori (for example, if θ is the linkage coefficient r defined already on page 47³², or if it is an atomic weight).

To finish, let us give an example of a confidence interval based on a small number of observations. Suppose, with Mr. FRECHET, that from an urn with a completely unknown composition, a *single* ball (suppose it is white) is drawn. What can we say then about the probability of drawing balls of the same color? If p is the (unknown) value of this probability and $f = 0$ or 1 is the frequency of white balls that can be observed in a single draw, an acceptance set of size $\geq \alpha$ would be defined by:

$$\begin{aligned} f = 0 & \text{ when } p < 1 - \alpha; \quad f = 1 \text{ when } p > \alpha; \\ f = 0 \text{ or } 1 & \text{ when } 1 - \alpha \leq p \leq \alpha. \end{aligned}$$

The confidence intervals for p with coefficient $\geq \alpha$ can be deduced to be:

$$\begin{aligned} p &\geq 1 - \alpha \text{ when } f = 1 \text{ (outcome is a white ball)} \\ p &\leq \alpha \text{ when } f = 0 \text{ (outcome is a ball other than white)} \end{aligned}$$

which implies, to clarify the ideas, that if one repeats the experiment in a large number of urns having an arbitrary composition, and that if one states each time that the prior probability of the observed result, no matter what this is, would be $\geq \frac{1}{100}$, one would be wrong in at most 1 of every 100 such trials. On the other hand, it is impossible to bound the probability that, in the case that one observes a white ball (selection of results), one makes a mistake by stating that the probability of whites is $\geq \frac{1}{100}$ (it is evident that all urns could contain less than $\frac{1}{100}$ of whites). The criterion does not allow us to choose one among several hypotheses, *these being stated before the experiment and mutually exclusive*, for example, between the hypotheses $p > \alpha$, $\alpha \geq p \geq 1 - \alpha$, $p < 1 - \alpha$; it only enables us, after each experiment, to *reject a single one* among these 3; it does not permit us, ever, to reject the second one because this is the common part of the 2 confidence intervals. This illustrates the remarks made on page 69³³.

³² Translator's Note: The page number of the *original* paper.

³³ Translator's Note: The page number of the *original* manuscript.

In summary, we see that the theory of confidence intervals allows us to make "objective" judgements free from a frequency of error that is known or bounded, but only in the following form: after the experiment, discard certain intervals where the bounds depend on the results of the experiment; however, this does not permit us to choose a given value, or often to choose between one or several values fixed a priori, so it becomes indispensable (unless one refuses to make this choice) to invoke a scheme of prior probabilities formulated in a more or less clear manner. This is necessary if one wishes to take into account previous experiments, unless their benefits are dispensed with willingly, as pointed out by STUDENT in the title of one of his tables (JEFFREYS, (10), p. 310).

8. INDETERMINACY OF A SET OF HYPOTHESES

In the preceding development, it was supposed that the probability law is known perfectly once θ is fixed and, hence, that all the consequences of all such possible hypotheses can be stated. In practice, as we observed in Section 4, this is not so: the hypotheses that one can state, and their consequences do not cover in an exhaustive manner the field of all possible hypotheses, so the sum of their probabilities, a priori or a posteriori, give a number < 1 ; the rules that we have given lead one to making a choice between the hypotheses stated, but do not prejudge at all about the probabilities of those that have not been formulated yet, and these may be appreciable, because the history of scientific theories is the history of the abandonment of old hypotheses and of the keeping of the newly formulated ones. For example, when a law $f(x, \theta)$ derived from theoretical considerations is fitted to data, it would be better to avoid suggesting, in agreement with Mr. MATHER, that all that can be extracted from the observations can be summarized in a confidence interval about θ , and it should be always kept in mind that $f(x, \theta)$ may be inexact! Certainly, in general, we will be incapable of formulating precisely all alternatives to the validity of $f(x, \theta)$, but it would be prudent to reserve a non-null prior probability for these alternatives, which will avoid a situation where $f(x, \theta)$ receives a brutal refutation in the case that, subsequently, the alternatives become more plausible and their posterior probabilities increase, at the expense of that of the former! As it has been said by CLAUDE BERNARD, we should not forget that the scientist must sacrifice as many theories as needed, "like the general that has had many horses killed but that still advances".

REFERENCES

- (1) E. BOREL, traité de Calcul des Probabilités, tome IV, fasc. III, 1939: valeur pratique et philosophie des probabilités.
- (2) G. DARMOIS, ouvrage ci-dessus, note VI.
- (3) G. DARMOIS, méthodes d'estimation, Actualité Hermann, No. 356, 1936.—, résumés exhaustifs, 23^e session de l'Institut International de Statistique, Athènes, 1936.—, Comptes Rendus, 200, 1935, p. 1265.
- (4) J. L. DOOB, Transact. of Amer. Math. Soc. 1934, p. 759 et 1936, p. 410.
- (5) D. DUGUE, Comptes-Rendus, 202, 1936, p. 193 et 1733.—Journal de l'École Polytechnique, 1937.

- (6) B. de FINETTI, colloque de Genève, Actualités Hermann, No. 766, 1939
- (7) R. A. FISHER, Philos. Transact., A 222, 1922, p. 309.—Journal of the Royal Statistical Society, 98, 1935, p. 39.—Statistical Methods for Research Workers, London, 1934.
- (8) M. FRECHET, Revue de L'Inst. Int. de Statistique, 1934, p. 182³⁴.
- (9) Traité, E. BOREL, t. 1, fasc. III, 1937.
- (10) H. JEFFREYS, Theory of probability, Oxford, 1939.
- (11) P. LEVY, L'addition des variables aléatoires, Paris, 1937.
- (12) K. MATHER, Statistical Analysis in Biology, London, 1943.
- (13) J. NEYMAN, colloque de Genève, Actualités Hermann, No. 739, 1938.
- (14) Biometrika 32, 1941, p. 128.
- (15) L'application du Calcul des Probabilités, Institut International de Coopération intellectuelle, 1945.

³⁴ Translator's Note: There is an error, probably typographical: the paper was published in 1943.