LXXV. *On Some Aspects of the Theory of Probability.* By DOROTHY WRINCH, *Lecturer at University College, London, and* HAROLD JEFFREYS, *M.A., D.Sc., Fellow of St. John's College, Cambridge* *.

I. *The Nature of Probability.*

THE theory of probability suffers at the present time from the existence of several different points of view, whose relations to one another have apparently never been adequately discussed. On the one hand some authorities follow de Morgan and Jevons in regarding probability as a concept comprehensible without any definition, and perhaps indefinable, satisfying certain definite laws the logical basis of which is not yet clear. On the other hand, attempts have been made to give definitions of probability in terms of frequency of occurrence; of these one is due to Laplace, who was largely followed by Boole, and another to Venn. Frequency of occurrence being a well-understood mathematical concept, such a definition would be important if it could be carried out; for then the undefined notion of probability would be expressed in terms of others that are better understood, and its laws, if true, would become demonstrable theorems in pure mathematics instead of postulates. Thus the subject would acquire the certainty of any other portion of pure mathematics and it would be unnecessary to investigate its foundations independently. It appears, however, as we hope to show in the first part of the present paper, that the definitions offered either implicity involve the very notion they are meant to avoid, or else make assumptions which are actually erroneous. We therefore consider it best to regard probability as a primitive notion not requiring definition.

Laplace defines probability † as the ratio of the number of favourable cases to that of all possible cases, and then goes on to say "but that supposes the various cases equally possible," so that to understand this definition it is necessary to examine what Laplace meant by *equally possible.* The expression is meaningless as it stands, for a proposition relative to a set of data is always either possible or impossible; there can be no degrees of possibility. He indicates later that if a coin is unsymmetrical the probability of throwing a head may be greater than that of throwing a tail, though the difference may be small; yet both are

* Communicated by the Authors.

† *Théorie analytique des probabilités*, troisième édition, p. 7 of introduction.

possible. In fact it seems that by *equally possible* he meant *equally probable*. Thus, as Poincaré has pointed out, it seems useless to attempt to make this definition satisfactory; it defines the probability of one proposition in terms of those of a set of others and not in terms of frequency alone, so that the notion Laplace set out to define reappears in the undefined concept of *equally possible*. The statement is, in fact, not a definition, but a simple and important principle of probability inference. Nor does it appear that there is any prospect of making any modification of it into a definition of probability; for there will always be the difficulty of deciding what are to be considered as unit alternatives. It is clear that even if it were possible to avoid introducing the notion of equally probable alternatives, some other way of distinguishing between sets of mutually exclusive and exhaustive alternatives would have to be found, and the immense variety of the circumstances to which it would have to apply seems to indicate that its scope must be at least as wide as that of truth; and it is very unlikely that a notion so general is capable of definition.

The view of Venn * is much more complex. He considers that the notion presupposes a series, the terms of which are indefinitely numerous and represent the cases of an attribute ϕ. From these one can pick out a smaller class, the members of which possess the further attribute ψ. If, then, we have chosen n members in all and m of them belong to the smaller class, the probability of ψ given ϕ is defined as the limit of m/n when n becomes indefinitely great. The form of this definition restricts the field of probability very seriously. In the first place it seems impossible to apply it to any case where the number of members of the first series is finite; one could attach no meaning to a statement that it is probable that the solar system was formed by the disruptive approach of a star larger than the sun, or that it is improbable that the stellar universe is symmetrical, for the indefinite repetition of entities of such large dimensions is utterly fantastic. Yet such cases as these are the very ones where the notion of probability is particularly valuable in science, and any definition that will not cover them is not satisfactory.

It may be urged, however, that this theory gives an adequate treatment of probability as applied to the class of cases with which it deals. Serious difficulties nevertheless present themselves. The existence of a probability on this theory requires that a limit shall exist to which a certain ratio tends in the long run; and one is led to ask what the

* 'Logic of Chance,' pp. 162 *et seqq.*

evidence is for the existence of such a limit. Suppose, for instance, that the probability of ψ given ϕ is $\frac{1}{2}$. Then the numbers of both ψ's and not-ψ's are infinite, and selections of ϕ's may therefore be made so that the ratio of ψ's to all ϕ's will tend to any limit whatever between 0 and 1; it may even tend to no limit at all. If, for instance, every time a ψ occurs we write 1, and every time a not-ψ occurs we write 0, m/n will be the mean of the first n terms of the series thus obtained. If then they occur in such an order as to give the series

$$1011000011111111 \ldots \quad . \quad . \quad . \quad . \quad (1)$$

where the number of digits in any block of similar digits after the first is equal to the total number of digits that have occurred previously, let us consider the r-th block, starting at the $(2^{r-2}\pm 1)$th figure. If r is even, the number of 1's that have already occurred is

$$1+2+8+\ldots+2^{r-3}=\tfrac{1}{3}(2^{r-1}+1), \quad . \quad . \quad . \quad (2)$$

so that

$$m/n \text{ when } n=2^{r-2} \text{ is } \tfrac{1}{3}(2+2^{-(r-2)}). \quad . \quad . \quad . \quad (3)$$

The r-th block consists of 2^{r-2} zeros, and at the end of it m/n has fallen to $\frac{1}{3}(1+2^{-(r-1)})$. In the next block it rises again to $\frac{1}{3}(2+2^{-(r-1)})$. Thus, however great r may be, we can find values of n greater than 2^{r-2} such that m/n is greater than $\frac{2}{3}$, and others such that m/n is less than $\frac{1}{3}+\epsilon$, however small ϵ may be. Thus m/n tends to no limit whatever. The notion that all series picked from the class of entities with the property ϕ will give series of values of m/n tending to the same limit is therefore incorrect, unless some further criterion be introduced to exclude all those that do not behave in this way, whose number is infinite; and the task will not be an easy one, for any criterion based on the mode of occurrence of long runs of ψ's or not-ψ's is liable to be found invalid in instances occurring in practice.

The origin of the idea that such a limit must exist may be considered at this stage, as it involves a theoretical point that may be of importance in future developments of the subject. The belief was based on a well-known theorem of James Bernoulli, a proof of which, based on Stirling's approximation to $n!$ for large values of n, is given by Laplace *. This theorem answers the following question: If the prior probability of a ψ be r, however many ψ's and not-ψ's have

* *Loc. cit.* pp. 275 *et seqq.* A proof based on the same principle, but more elegant and easily applied, is given by Bromwich, Phil. Mag. August 1919, pp. 231–235.

been chosen already, what is the probability that when n ϕ's have been selected m/n will lie between $r-\epsilon$ and $r+\epsilon$? It is shown that the probability of any particular value of m is

$$\frac{n!}{m!(n-m)!} r^m (1-r)^{n-m}, \quad . \quad . \quad . \quad . \quad (4)$$

and when Laplace approximates according to the formula

$$n! = n^n e^{-n} \sqrt{(2\pi n)} \left(1 + \frac{1}{12n} + \frac{1}{288n^2} + \dots\right) \quad . \quad (5)$$

he shows that this is a maximum when m is rn, which is not necessarily an integer, and if m/n is equal to $r+\xi$, he gives a formula which reduces to

$$\left\{\frac{n}{2\pi r(1-r)}\right\}^{\frac{1}{2}} e^{-\frac{n\xi^2}{2r(1-r)}} \{1 + O(n\xi^3)\} d\xi. \quad . \quad (6)$$

The probability that ξ does not lie between $\pm\epsilon$ is then

$$2\left\{1 - \operatorname{Erf} \epsilon \left(\frac{n}{2r(1-r)}\right)^{\frac{1}{2}}\right\}$$
$$+ O\left\{n^{-\frac{1}{2}} e^{-n\xi^2/2r(1-r)} \left(\epsilon^2 - \frac{2r(1-r)}{n}\right)\right\}. \quad (7)$$

Now, no matter how small ϵ and η may be, it is always possible to choose n large enough to make this less than η; accordingly, by making n great enough we can make the probability that m/n differs from r by more than any quantity assigned beforehand as small as we please. This is Bernoulli's theorem.

This does not, however, give the probability that m/n will tend to a limit as n tends to infinity. For if ϵ be a small quantity fixed beforehand, the necessary and sufficient condition that m/n tend to a r as a limit is that a value of n_0 can always be found such that *for all values of n greater than* n_0, $m/n - r$ shall be less numerically than ϵ. Now if x be great enough to make e^{-x^2} small, we have the relation

$$1 - \operatorname{Erf} x = \frac{e^{-x^2}}{x\sqrt{\pi}} \left\{1 + O\left(\frac{1}{x^2}\right)\right\}, \quad . \quad . \quad . \quad (8)$$

and the probability that m/n does not lie between $\pm\epsilon$ is therefore, when n is great enough,

$$\frac{2}{\epsilon} e^{-n\epsilon^2/2r(1-r)} \sqrt{\left(\frac{2r(1-r)}{n\pi}\right)} \left\{1 + O\left(\frac{2r(1-r)}{n\epsilon^2}\right)\right\}. \quad . \quad (9)$$

The probability that any value of m/n for $n > n_0$ lies outside these limits is therefore not greater than the sum of these expressions for all values of n from n_0+1 to infinity. We see that this is less than

$$\frac{2}{\epsilon}\sqrt{\frac{2r(1-r)}{n_0\pi}}\, e^{-n_0\epsilon^2/2r(1-r)}\left\{1+O\left(\frac{2r(1-r)}{n_0\epsilon^2}\right)\right\}$$
$$\{1+e^{-\epsilon^2/2r(1-r)}+e^{-2\epsilon^2/2r(1-r)}+\ldots\}. \quad (10)$$

The sum of the series is finite and independent of n_0; hence we see that n_0 can always be chosen so as to make the probability that, for all values of n greater than n_0, the value of m/n will lie between $r\pm\epsilon$, differ from unity by as small a quantity as we like.

The proposition required for the validity of Venn's theory is: n_0 can always be chosen so as to make the probability that, for all values of n greater than n_0, the value of m/n will lie between $r\pm\epsilon$, exactly equal to unity.

These two propositions bear a close resemblance to each other, but they are not equivalent. In fact, in consequence of the existence of modes of selection for which m/n does not tend to r as a limit, we know that the second proposition must be false. The first, on the other hand, has just been proved true; but it does not even establish a high probability for the proposition that m/n tends to r as a limit in any particular case. For it has been shown only that a certain result will have a very high probability when a single value of ϵ has been assigned; but there is no reason to infer from this that the probability is high that it will hold for all values of ϵ whatever, which would have to be true if m/n were to tend to a limit. The difficulty is somewhat similar to that in the theory of infinite series, with regard to series that "converge with infinite slowness."

II. *The Mathematical Theory of Probability.*

An essential assumption in order that analytical methods may be applicable to the theory of probability must now be stated, namely, that a correspondence can be established between positive real numbers and the propositions to which the fundamental notion of probability is applicable (relative in each case to the appropriate data) which shall have the following properties.

1. To each combination of proposition and data corresponds one and only one number.
2. If in one combination the proposition is more probable relative to the data than in another, the number corresponding to the first is greater than that corresponding to the second.
3. If two propositions referred to the same data are mutually exclusive, the number corresponding to the proposition that one of them is true is the sum of those corresponding to the two original propositions.
4. The greatest and least numbers correspond to those combinations and only those in which the data imply that the proposition is true or untrue respectively.

Several writers have defined probability as "quantity of belief," or somewhat better, "quantity of knowledge"; these are somewhat vague terms, and the transition from these expressions to the number series has usually been carried out without any explanation. Yet the use of numbers for the comparison of probabilities at all was perhaps the greatest advance ever made in the theory.

The above assumptions are independent; they are involved implicitly in every theory of probability yet introduced; and with their aid it is possible to make some progress with a logical theory. In the first place, we can show that the number corresponding to a proposition incompatible with the data is zero. For let a datum be that $x=1$; then on this datum the propositions $x=2$ and $x=3$ are both false, and each corresponds to the number a, where a is the least possible number of those involved in the correspondence. Further, $x=2$ is incompatible with $x=3$; hence by axiom 3 the probability that one of them is true is $2a$. But the proposition "$x=2$ or $x=3$" is incompatible with the datum "$x=1$," and therefore corresponds to a. Thus $2a$ is equal to a, and a is therefore zero. If, then, probabilities are to be represented by numbers, zero must be the least number involved; but adjustments could be made in our assumptions which would allow any other number to represent the minimum on the scale.

If we divide all numbers of the series by that corresponding to a proposition implied by the data, all the above axioms will apply equally to the numbers of the new series. We shall henceforth use the notation $P(p:q)$ to denote the number of this series corresponding to the proposition p on the data q; we have $P(q:q)=1$, and $P(\text{not-}q:q)=0$. $P(p:q)$ may be read "the probability of p given q."

Consider two propositions p and q which are not mutually exclusive, referred to data f. Then the following four propositions are mutually exclusive, namely

$$p.q\,;\ p.\sim q\,;\ \sim p.q\,;\ \sim p.\sim q\,^{*}.$$

Then by axiom 3

$$P(p:f)=P(p.q:f)+P(p.\sim q:f)$$
$$P(q:f)=P(p.q:f)+P(\sim p.q:f)$$
$$P(p\vee q:f)=P(p.q:f)+P(p.\sim q:f)+P(\sim p.q:f).$$

By addition we find

$$P(p:f)+P(q:f)=P(p\vee q:f)+P(p.q:f) \quad . \quad . \quad (1)$$

which is regarded by Jevons and de Morgan as axiomatic.

The second axiom yields as an obvious corollary the famous "principle of sufficient reason"; according to this, equal probabilities are assigned to propositions relative to data when the data give no reason for expecting any one rather than any other. In discussing the problems of probability † Poincaré, after disposing of the view of Laplace and Boole, seems inclined to consider this principle as the only possible basis of the theory. Substantially the same view is held by Jevons. There is, however, an objection to basing the whole theory on the principle of sufficient reason. For the only way of passing from the notion of "more probable" to the numerical estimate of probability in any particular case is to discover some set of mutually exclusive and exhaustive alternatives, from which we can pick out some by our judgment as more probable than others; the most probable on the data then receives the greatest numerical estimate. But if we restrict ourselves to cases where we can obtain a set of alternatives that shall be all equally probable, we are arbitrarily limiting the field to which the theory can be applied. We could, indeed, only deal with those cases where some proposition that is certain on the data can be expressed as the disjunction of a number of equally probable and mutually exclusive propositions; the probability of any proposition that can be expressed as, or is implied by, the disjunction of any sub-class of these could then be assessed by means of the principle of sufficient reason and axioms 3 and 4. Now there is no reason to believe that the notion of probability is applicable to no

* $\sim p$ denotes the proposition that p is false, and $p.q$ denotes the proposition that p and q are both true. Thus $\sim p.\sim q$ denotes the proposition that p and q are both false. The proposition that at least one of p and q is true is denoted by $p \vee q$, or the disjunction of p and q.

† *La Science et l'Hypothèse*, 1904, 213–245.

propositions other than those expressible in this way ; and it is habitually employed in scientific practice and everyday life in cases where it seems likely that such expression is impossible. Most of the problems of inverse probability, for instance, seem to introduce propositions not so decomposable. If then we wish to retain the customary applications of the theory (and this seems desirable at any cost), we must assume that axiom 2 is correct. The assumption that information can be obtained from the notion of "equally probable" alone without that of "more probable," which seems as intelligible *a priori* anyhow, demands that propositions can be decomposed in this way in all these cases, whether there is any warrant for assuming this possibility or not. Thus axiom 2 is preferable to the principle of sufficient reason as a primitive proposition, since it covers as much ground and involves fewer assumptions.

The use of the principle of sufficient reason in the cases where it is applicable leads to a proof of another proposition which is an axiom in Jevons's theory. Suppose we have a class of n propositions, of which we know that one and only one is true, and any one is as likely to be true as any other. Then if any m of them are selected, the probability that one of these m is true is m/n. Let q then denote the proposition that one of these m is true. Consider another class of the original propositions, and let p denote the probability that some member of this class is true. The probability that p and q are both true is then the probability that some member of the common part of the two sub-classes is true. Let the number of propositions in this common part be l. Then if h denote the data we have at the beginning, we have

$$\begin{aligned} P(p\,.\,q:h) &= l/n \\ &= \frac{l}{m}\,.\,\frac{m}{n} \\ &= P(p:q\,.\,h)\,.\,P(q:h). \end{aligned}$$

We see that all cases where the probabilities of propositions can be determined by decomposing a certain proposition into a finite number of equally probable alternatives can be treated in this way, so that the relation

$$P(p\,.\,q:h) = P(p:q\,.\,h)\,.\,P(q:h) \quad . \quad . \quad . \quad (2)$$

is always true when the principle of sufficient reason is applicable.

But is there any reason to suppose that this relation still holds when the principle is not applicable? Some further assumption is necessary before it can be proved in these cases, and various suggestions could be offered that would bridge the gap without making it necessary to suppose that the relation is known *a priori* in these cases. There seems, however, to be little or no ground for deciding between them, and the proposition may as well be assumed to hold in general without further discussion. From the propositions so far assumed or proved, with judgments of greater, equal, or less probability in particular cases, the mathematical theory of probability can be developed.

Another point in connexion with the theory may be briefly mentioned. All that is strictly necessary in order that the notions of probability may be capable of logical treatment is that combinations of propositions and data can be arranged in a series so that whenever a combination A is not more probable than another, B, B shall not precede A in the series. With suitable assumptions regarding the position in the series of a combination, such as the disjunction of two contradictory propositions referred to the same data, a theory could be constructed. There is no reason save convenience why the number series should be the one employed for this purpose. So long as we confine ourselves to those cases where a proposition certain on the data can be decomposed into a finite number of equally probable alternatives, and the proposition whose probability is to be estimated is expressible as or equivalent to the disjunction of a class of these, the number series is obviously adequate; in fact the series of all rational proper fractions in ascending order of magnitude would be adequate. This latter series is, however, at once found to be insufficient when we attempt to deal with cases where the number of equally probable alternatives required to cover the case considered is infinite. This difficulty was thought to be removed by using the series of all the real numbers less than unity instead of that of the rational numbers. But the question that arises now is, whether the series of all the real numbers is itself adequate for the purpose, and the answer seems to be in the negative, for there are evidently cases where the use of infinitesimals is necessary to a complete theory, and the discovery of others, necessitating the introduction of infinitesimals of different orders, is practically certain. For instance, suppose we are given that x is a whole number, and that all whole numbers are equally probable values of x. What is the probability of any particular value of x, say 1053? Clearly it is not finitely different from 0; for if it were λ say, we could find a whole number M whose reciprocal would be less than λ; but $1/M$ is the probability of M being 1053 when there are only M possible alternatives, and the probability cannot be increased by increasing the number of alternatives. Hence the probability that x is 1053 is less than any finite number, contrary to what was assumed. It is nevertheless different from zero, for then there would be no means of distinguishing between the probability of this, which is a perfectly possible proposition on the data, and that of a proposition known to be impossible on the data. Hence this probability is less than any finite number, and yet is different from zero; in other words, it is an infinitesimal, in the original sense of the term*. Again, we can see that the probability of a particular real number chosen at random being rational is infinitesimal; so is the probability that a function is analytic, given that all functions are equally probable. Now a complete theory of probability must cover all these cases; but so long as we are confined to the series of the real numbers that is impossible; for if this has C members, the number of possible functions whose values are real numbers is C^C, which is greater; hence problems arising in connexion with the probability of functions demand the use of a series for comparison whose members are more numerous than the real numbers. Such series are known; and perhaps one suitable for the purpose may be constructed which will include among its members the real numbers themselves.

III. *On Probability Inference.*

The characteristic feature of the type of inference with which classical logic is primarily concerned is that given the premises it is possible to establish the conclusions with absolute certainty from them. In many cases, however, such a result is unobtainable when it is nevertheless possible to

* M. E. Borel remarks (*Leçons sur la Théorie des Fonctions*, 1914, p. 184) that "there is a true discontinuity between an infinitely small probability, *i. e.*, a variable probability tending towards zero, and a probability equal to zero. However small be the probability of the favourable case, this is possible; whereas it is impossible if the probability be zero..... The same is not true of continuous probabilities; the probability that a number taken at random may be rational is 0; this must not be considered as equivalent to impossibility." This use of zero to denote the probability of both an impossible alternative and a possible alternative with no finite probability seems likely to lead to confusion. The introduction of the conception of a limit does not help matters, for in making a single trial the probability of success is quite definite, and involves no notion of a limit.

show that the conclusion has a certain probability relative to the premises; an inference of this kind may be called a "probability inference." The establishment by this means of a high probability in favour of the conclusion relative to the premises is often as useful as the inference that does not involve the notion of probability. The course thus indicated is always followed in empirical generalization, for in such a generalization it is never possible to establish the certainty of the conclusion from the data. The principles employed in such inference are therefore of extreme importance; but as yet they are not well understood.

Detailed treatment is most applicable to the type of probability inference known as sampling induction, and numerous discussions of this have been given, but even here various errors seem to have survived. The problems capable of solution by this method are analogous to the following. Suppose that a bag contains m balls, an unknown number of which are white. Of these $p+q$ have been drawn and not replaced; p of them have been white and q not white. What is the probability that the number of white balls in the bag is n?

It is assumed that the balls are indistinguishable before being drawn, so that at any stage any individual ball is as likely to be drawn as any other. Let $f(n)$ be the prior probability of any particular number of white balls. If n were the true number of white balls in the bag the probability that p white balls and q others would be picked in $p+q$ trials would be $\frac{{}^{n}\mathrm{C}_{p}\,{}^{m-n}\mathrm{C}_{q}}{{}^{m}\mathrm{C}_{p+q}}$. It follows that the prior probability of a particular pair of values of p and q for a given n is $f(n)\frac{{}^{n}\mathrm{C}_{p}\,{}^{m-n}\mathrm{C}_{q}}{{}^{m}\mathrm{C}_{p+q}}$. Hence, by the law of inverse probability, which follows easily from the proposition

$$\mathrm{P}(p\,.\,q:h)=\mathrm{P}(p:q\,.\,h)\,.\,\mathrm{P}(q:h),$$

the probabilities on the data of particular numbers of white balls are in the ratio of the probabilities of the actual values of p and q for these numbers of white balls; thus we find that the probability that any particular value of n is the true number of white balls in the bag, given the composition of the sample, is

$$f(n)\,{}^{n}\mathrm{C}_{p}\,{}^{m-n}\mathrm{C}_{q}\div\Sigma_{n}f(n)\,{}^{n}\mathrm{C}_{p}\,{}^{m-n}\mathrm{C}_{q}, \quad . \quad . \quad (1)$$

where the summation is to be extended to all possible values of n.

This gives the solution of the problem in the most general case, but in most cases more concise information, even though it may be only approximate, is desirable. The case exclusively considered in the discussions hitherto given is the very simple one where $f(n)$ is the same for all values of n. The ground for this evaluation may be either complete ignorance of the relative number of white balls among the balls in the world, or knowledge that white and other balls have occurred equally frequently in all the ratios possible in this problem. On this hypothesis it can be shown that the probability of a white ball at the $(p+q+1)$th drawing is $(p+1)/(p+q+2)$; and if q is 0, the probability that all the balls in the bag are white is $(p+1)/(m+1)$.

It is however very rarely, if ever, possible to assume, on the data available before the sample is taken, that $f(n)$ is independent of n, and cases where it has other values are much more interesting. For instance, we know that there is a strong tendency for similar individuals to be associated, so that the greatest and least values of n are more probable on the initial data than the intermediate ones. Or suppose we are considering balls of another colour, say green. It would be absurd to suggest that a bag is as likely to contain green balls alone as to contain no green balls, for we know that in fact green balls are not nearly so common as balls of all other colours together. On the other hand in these cases it is not usually possible to decompose the propositions, whose probabilities we wish to assess in order to find $f(n)$, into equally probable alternatives, so that the principle of sufficient reason cannot be applied; thus though we may be confident that $f(n)$ lies within certain limits, we cannot say that it has any particular value. It will, however, be shown that unless the form of this function is something very remarkable the probabilities to be assigned to particular values of n are practically independent of the prior probabilities, depending almost wholly on the composition of the sample taken, provided this is large enough. To show how this comes about we need an approximation to ${}^nC_p\,{}^{m-n}C_q$ when p and q are fairly large. This is best obtained by a method analogous to that adopted by Dr. Bromwich *. If r and s are both large and r is large compared with s, formula (2) of Dr. Bromwich's paper yields the approximation

$$\log\{(r+s)!\} = (r+s+\tfrac{1}{2})\log r - r + \tfrac{1}{2}\log 2\pi + \tfrac{1}{2}\frac{s^2}{r}$$

$$+ \text{ terms of order } \frac{s^3}{r^2} \text{ \&c.} \quad (2)$$

* *Loc. cit.*

Put $$n_0 = \frac{pm}{p+q}\,;\; n = n_0 + mx. \quad . \quad . \quad . \quad . \quad (3)$$

Then

$$\log({}^{n}C_{p}\,{}^{m-n}C_{q}) = \log\{(n_0+mx)!\} + \log\{(m-n_0-mx)!\}$$
$$-\log p! - \log q!$$
$$-\log\{(n_0-p+mx)!\} - \log\{(m-n_0-q-mx)!\}, \quad (4)$$

which gives on substituting the above approximation, provided mx is not so great as to invalidate it, an expression that simplifies to

$$(p+q)\log\frac{m}{p+q} + (m-p-q+1)\log\frac{m}{m-p-q} - \tfrac{1}{2}\log(pq)$$
$$-\log 2\pi - \tfrac{1}{2}x^2\frac{(p+q)^3 m}{pq(m-p-q)}. \quad . \quad (5)$$

Hence the function considered is a maximum for $n=n_0$, and its values for other values of n are distributed about this according to the Gauss law.

The given sample is said to be a fair one if $p/(p+q)$ is equal to the ratio of the true number of white balls to the whole number of balls in the bag. The deviation from fairness is therefore represented by x. Substituting the approximation (5) in the formula (1), we find by summation that the probability that x lies between $\pm\epsilon$ is

$$\frac{\Sigma' f(n)\exp -\frac{1}{2}\frac{(p+q)^3 mx^2}{pq(m-p-q)}}{\Sigma f(n)\exp -\frac{1}{2}\frac{(p+q)^3 mx^2}{pq(m-p-q)}}, \quad . \quad . \quad . \quad (6)$$

where in the denominator the summation covers all values of n and in the numerator all values between $m\left(\frac{p}{p+q}-\epsilon\right)$ and $m\left(\frac{p}{p+q}+\epsilon\right)$. Now the coefficient of x^2 in the exponent is always numerically greater than $2(p+q)$. If then $2^{\frac{1}{2}}(p+q)^{\frac{1}{2}}\epsilon$ is greater than h, the exponential is less than e^{-h^2}, which is very small even when h is not remarkably small. Outside of the range the exponential factor is even smaller, and unless $f(n)$ is so great that its greatness can counteract the smallness of the exponential factor, the contribution to the denominator from the values of x not between $\pm\epsilon$ is small. Thus the numerator and denominator are nearly equal and the probability that x lies between $\pm\epsilon$ is nearly 1.

We therefore have the theorem: if a selection of $p+q$ members from a class of m α's consists of p β's and q not-β's, and $f(n)$ the prior probability of there being n β's and $m-n$ not-β's in the class α is such that when x is numerically greater than ϵ, $f(n)$ is never so great that $f(n)\exp-2(p+q)(x^2-\epsilon^2)$ is comparable with $f\{mp/(p+q)\}$, then the probability that n/m lies within ϵ of $p/(p+q)$ differs from unity by a quantity of order not greater than $\exp-2(p+q)\epsilon^2$. Thus, unless the distribution of prior probability among various values of n is very remarkable, its precise form does not produce much effect on the probability that the true value lies within a certain range determined wholly by the constitution of the sample itself.

It is worthy of note that the range within which it is probable that n/m must lie is of length proportional to $(p+q)^{-\frac{1}{2}}$; it does not depend on m to any great extent, but if p and q are very different the range may be much shorter than this. This leads to the result that there is a strong presumption that a large sample is approximately a fair one even if it is small compared with the whole of the class; and that the range within which the fractional composition is as likely as not to lie is much the same however great the whole number of individuals may be. The fact that the error likely to be committed in sampling is, except in extreme cases, limited by the size of the sample itself, may be of some importance in electoral and economic questions. It is also easy to infer from the results obtained that the probability of drawing a β at the next trial is not likely to be far from $p/(p+q)$, agreeing with sufficient accuracy with the result of the ordinary theory.

In a recent paper* Mr. C. D. Broad has given a suggestive discussion of the problem of inductive inference, in which he adopts the ordinary theory, according to which when q is 0 the probability that all the members are β's is $(p+1)/(m+1)$. This is not necessarily true, for the reasons given above, but this does not affect Mr. Broad's main point, which is that in all ordinary cases the number of observed instances is so small compared with the total number of instances that it is impossible to arrive by this means at any noteworthy probability for a general law. General laws are, however, of various kinds. The type to which Mr. Broad devotes most attention is the statement that "all crows are black," based on the fact that all observed crows have been black. Now a crow is an object defined by the conjunction of a number of properties, which may or may not include blackness. In the

* 'Mind,' October 1918, pp. 389–404.

former case the inference becomes tautologous, and we are concerned only with the latter. But it has been shown above that if n be the number of black crows in the world, n/m is not likely to deviate from $p/(p+q)$ by more than a quantity of the order of $(p+q)^{-\frac{1}{2}}$; and in this case, as p is great and q is zero, we are justified in inferring that the number of crows that are not black is a small fraction of the whole, which is all that is inferred in practice; for the possibility in exceptional cases of sport, albinos, and so on is well known. The other type of general law is one that is held to be true in *every* instance of the entities to which it is held to apply. Such a law cannot be derived by means of probability inference, for it deals only with certainties. Here Mr. Broad's argument is valid, and no such law can derive a reasonable probability from experience alone; some further datum is required. One way of arriving at such laws may be suggested here. Suppose we have an *a priori* belief that either every x has the property ϕ or every x has the property ψ. If then a single x, say c, is found to satisfy ϕ but not ψ, we can infer deductively the universal proposition that all x's satisfy ϕ. Such cases are fairly frequent: if for instance we consider that either Einstein's or Silberstein's form of the principle of general relativity is true, a single fact contradictory to one would amount to a proof of the other in every case.

Before leaving the important question of induction, we propose to consider it in relation to the Venn view. If Venn's definition of probability be adopted the existence of a numerical estimate of probability depends on the possibility (at least imagined) of indefinite repetition of the data, the truth or falsehood of the proposition whose probability relative to the data is to be estimated being recorded at each repetition. The probability is then the limit of the ratio of the number of favourable cases to the number of all cases. Now on this basis it is never possible, by what has been said already, to prove that in any given case such a limit will exist. All the axioms of the "undefined concept" theory are therefore indemonstrable, and must be assumed *a priori* in the same way. Even in the simple case of picking indistinguishable balls out of a bag the probability of picking any particular individual cannot be assessed without some hypothesis about the limit of the results obtained by making an indefinite number of selections, each ball being replaced after being drawn. In the problem of sampling induction we can therefore by making enough assumptions of this character, which there seems little or no reason to believe, obtain a proof of a proposition superficially the same as the chief theorem of this section; but let us consider what this result means on the Venn view. It would mean that if we had a large number of classes, each of m members, and from each we had picked out $p+q$ members, of which p were β's and q not-β's, then when the number of such classes is indefinitely increased the fraction of them in which the actual number of β's does not lie within certain limits would tend to zero as a limit. Thus the already hopeless task of proceeding to the limit of an infinite number of observations becomes in this case the still more complex one of repeating similar classes indefinitely.

Such indefinite repetition of *infinite* classes is called by Venn the construction of "cross-series" and forms an essential part of his theory of inference. It is necessary, for instance, in giving a meaning to the proposition connecting the probabilities of a proposition referred to different data $P(p.q:h)=P(p:q.h).P(q:h)$. For an infinite series is needed to give an account of $P(p:q.h)$, which is the limit derived from the frequency of the truth of p among entities for which q and h are true. Such entities, however, are only a part of those for which h holds. Thus to establish a meaning for the number $P(p.q:h)$ we must consider *all* entities satisfying h, whether they satisfy q or not. Thus further series must be constructed which will show how often q is actually true, and this requires, according to Venn, an infinite number of series of entities all satisfying h, so that we can examine in one direction to find the frequency of p given q and h and in the other to find that of q given h. Thus the difficulty of obtaining enough terms, acute in the simple case, is here intensified; further, there is no more reason to believe in the existence of limits in this case than there was in the other. The difficulties are merely complicated and not removed by the use of cross-series.

There is no evidence that Venn ever attempted to meet these difficulties. Indeed, we may conclude from some passages of his work that they had never suggested themselves. The following passage, for example, occurs in the third edition of his 'Logic of Chance,' page 208. "The opinion according to which certain inductive formulæ are regarded as composing a portion of probability cannot, I think, be maintained. It would be more correct to say . . . . that induction is quite distinct from probability, yet co-operates in almost all its inferences. By induction we determine for example whether and how far we can safely generalise the proposition that 4 men in 10 live to be 56:

supposing such a proposition to be safely generalised we hand it over to Probability to say what sort of inferences can be deduced from it."

The "undefined concept" view of probability can be developed so as to yield a theory of induction adequate for scientific purposes. There are difficulties in the way of obtaining such a theory from the frequency view, and we conclude that the balance is in favour of the "undefined concept" view.

Summary.

It is shown that the attempt to give a definition of probability in terms of frequency is unsuccessful. Laplace's definition, apparently in these terms, really involves implicitly the concept of probability and is therefore circular in character. Venn's definition in terms of the limit of a series is unsatisfactory because there is no reason to believe that his series do in fact usually tend to a limit; it is shown that there are many cases where they do not; and as his process is incapable of being carried out, the existence of such a limit can in any case only be known *a priori* if at all, so that his method offers no advantage over that of regarding probability as an entity known to exist independently of definition, intelligible without such definition, and perhaps indefinable.

A set of axioms on which a mathematical theory of probability can be based is then given, which seems to offer certain advantages over the current ones. In particular it is capable of covering cases where the principle of sufficient reason cannot be applied to assess probability. It is also shown that a complete theory of probability must allow for the use of infinitesimals.

A discussion of sampling induction is given, in which it is shown that when the sample is large enough the prior probabilities of different constitutions of the whole do not usually affect appreciably the probabilities inferred after the samples have been taken. Also the range within which the fractional constitution is as likely as not to lie includes the fractional constitution of the sample, and its extent is inversely proportional to the number of the sample itself, whatever be the number of the whole.