A note on inverse probability. By Mr J. B. S. HALDANE, Trinity College.

[*Received* 19 November, *read* 7 December 1931.]

The problem of statistical investigation is the description of a population, or Kollektiv, of which a sample has been observed. At best we can only state the probability that certain parameters of this population lie within assigned limits, i.e. specify their probability density. It has been shown, e.g. by von Mises(1), that this is only possible if we know the probability distribution of the parameter before the sample is taken. Bayes' theorem is based on the assumption that all values of the parameter in the neighbourhood of that observed are equally probable *a priori*. It is the purpose of this paper to examine what more reasonable assumption may be made, and how it will affect the estimate based on the observed sample.

Let us first consider a population of which a proportion x possess the character X, and of which a sample of n members is observed, n being so large that one of the exponential approximations to the binomial probability distribution is valid. Let a be the number of individuals in the sample which possess the character X. Let $f(x) = e^{\phi(x)}$ be the *a priori* probability density of x, i.e. let the *a priori* probability that x lies between x and $x + \delta x$ be $f(x)\,\delta x$, where $f(x)$ is defined for all values of x between 0 and 1 inclusive. We shall assume that $f(x)$ is regular except for a finite number of singularities. If this is not the case we can always approximate to it by smoothing out $f(x)$ so that irregularities confined to intervals much smaller than $n^{-\frac{1}{2}}$ are eliminated, and this will have little effect on our final result. It is an important fact that in almost all scientific problems we have a rough idea of the nature of $f(x)$ derived from the past study of similar populations. Thus, if we are considering the proportion of females in the human population of any large area, $f(x)$ is quite small unless x lies between ·4 and ·6. It is sometimes more convenient to say that all values of a function $F(x)$ are equally probable *a priori*. Then, if we multiply by a suitable constant so that $F(1) - F(0) = 1$, we have $f(x) = F'(x)$, and $\phi'(x) = F''(x)/F'(x)$.

It is known that after the sample has been observed when $f(x) = 1$, three types of distribution of x are obtained. If a and $n - a$ are not small compared with $n^{\frac{1}{2}}$, the probability that x lies between x and $x + dx$ is given by

$$dp = \frac{1}{(2\pi)^{\frac{1}{2}}\sigma}\, e^{-\frac{1}{2\sigma^2}\left(x - \frac{a}{n}\right)^2} dx, \text{ where } \sigma^2 = \frac{a(n-a)}{n^3}.$$

https://doi.org/10.1017/S0305004100010495 Published online by Cambridge University Press

If a is small compared with $n^{\frac{1}{2}}$, but finite, then

$$dp = \frac{e^{-nx}x^a}{a!}\,dx,$$

and, if $a = 0$, $$dp = e^{-nx}dx.$$

The last two formulae admit of obvious alterations when $n - a$ is small or zero.

Now consider the first of these cases when $f(x)$ is not unity, but is regular in the neighbourhood of $x = a/n$. The probability that, for a given x, we shall obtain the observed sample, is

$${}_nC_a\,x^a(1-x)^{n-a},$$

and hence the probability density of x is given by

$$dp = \frac{x^a(1-x)^{n-a}f(x)\,dx}{\int_0^1 x^a(1-x)^{n-a}f(x)\,dx}.$$

This is a maximum when the numerator is a maximum, i.e. when $a\log x + (n-a)\log(1-x) + \phi(x)$ is a maximum, so that

$$\frac{a}{x} + \frac{n-a}{x-1} + \phi'(x) = 0;$$

hence
$$x = \frac{a}{n} + \frac{x(1-x)\phi'(x)}{n}. \qquad (1)$$

Let $\bar{x} = a/n$. Then if $\phi'(x)$ is small compared with n in the neighbourhood of $x = \bar{x}$, and does not vary rapidly with x, we have approximately, for the optimal value of x,

$$x = \bar{x} + \frac{\bar{x}(1-\bar{x})\phi'(\bar{x})}{n} = \bar{x} + \sigma^2\phi'(\bar{x}). \qquad (2)$$

Moreover the deviation due to the fact that $\phi'(x)$ does not vanish is small compared with the standard error $\sigma = \{\bar{x}(1-\bar{x})/n\}^{\frac{1}{2}}$, provided that $\phi'(\bar{x})$ is small compared with $n^{\frac{1}{2}}$.

The final probability distribution of x is approximately given by

$$dp = \text{constant} \times e^{\phi(x) - \frac{1}{2}\left(\frac{x-\bar{x}}{\sigma}\right)^2}\,dx.$$

The distribution is thus Gaussian provided that $n^{-\frac{1}{2}}\phi'(x)$ is small. The correct value of the standard error is

$$\sigma[1 - \sigma^2\phi''(x)]^{-\frac{1}{2}}.$$

If $\phi'(x)$, though small compared with $n^{\frac{1}{2}}$ in the neighbourhood of $x = a/n$, has an infinity or becomes very large for some other

https://doi.org/10.1017/S0305004100010495 Published online by Cambridge University Press

value of x (other than 0 or 1), and if a/n is finite, then the probability distribution is approximately Gaussian in the neighbourhood of $x = a/n$, but has an infinity or a secondary maximum at the other point or points.

An illustration from genetics will make this point clear. The plant *Primula sinensis* possesses twelve pairs of chromosomes of approximately equal size. A pair of genes selected at random will lie in different chromosomes in $\frac{11}{12}$ of all cases, giving a proportion $x = 0{\cdot}5$ of "cross-overs." In $\frac{1}{12}$ of all cases they lie in the same chromosome, the values of the cross-over ratio x ranging from 0 to 0·5 without any very marked preference for any part of this range, except perhaps for a tendency to avoid values very close to 0·5. $f(x)$ is thus approximately $\frac{1}{6}$ for $0 \leqslant x < 0{\cdot}5$; it has a discontinuity at $x = 0{\cdot}5$, such that the probability of this value is $\frac{11}{12}$; while, for $0{\cdot}5 < x \leqslant 1$, $f(x) = 0$.

Now if a family of 400 seedlings from the cross between a doubly heterozygous plant and a double recessive contains 160 "cross-overs" we have two alternatives. The probability of getting such a family from a plant in which the genes lie in different chromosomes is $\frac{11}{12}\,{}_{400}C_{160}\,2^{-400}$, or $1{\cdot}185 \times 10^{-6}$. The probability of getting it from a plant in which they lie in the same chromosome is

$$\tfrac{1}{6}\,{}_{400}C_{160}\int_0^{\frac{1}{2}} x^{160}(1-x)^{240}\,dx.$$

Since this integral is very nearly equal to

$$\int_0^1 x^{160}(1-x)^{240}\,dx, \text{ or } \frac{160!\,240!}{401!},$$

this probability is approximately $\dfrac{1}{6 \times 401}$, or $4{\cdot}156 \times 10^{-4}$. Thus the probability that the family is derived from a plant where the genes lie in different chromosomes and $x = 0{\cdot}5$ is ·028. Otherwise the mean value of x is 0·4, with standard error ·0245. The overall mean value, or mathematical expectation, of x is ·4028, and the graph of the probability density $\dfrac{dp}{dx}$ is an approximately normal error curve centred at $x = 0{\cdot}4$ with standard deviation ·0245, together with an infinity at $x = 0{\cdot}5$. The error of the mean 0·4 is $\cdot 0006\phi'(x)$. This exceeds the standard error if $\phi'(x) > 40{\cdot}8$, i.e. if the *a priori* probability that $x = {\cdot}41$ is $e^{3{\cdot}16}$ or 2·26 times the *a priori* probability that $x = {\cdot}39$, which is almost certainly not the case.

We next consider the case when a, though not zero, is small compared with $n^{\frac{1}{2}}$. As before, equation (1) gives the optimum value of x. But this is not in general the mean value, since the distribution of probabilities is asymmetrical, being given by a Pearsonian

https://doi.org/10.1017/S0305004100010495 Published online by Cambridge University Press

curve of Type III when $f(x)$ is constant. Since $(1-x)^{n-a}$ approximates to e^{-nx}, we may write

$$dp = \frac{e^{-nx}x^a f(x)\,dx}{\int_0^1 e^{-nx}x^a f(x)\,dx},$$

so that the mean is

$$\frac{\int_0^1 e^{-nx}x^{a+1} f(x)\,dx}{\int_0^1 e^{-nx}x^a f(x)\,dx}.$$

If $f(x)$ is constant, the modal value of x is thus a/n, and the mean value is $(a+1)/n$. So although in n trials we have obtained a X's, we should expect to obtain $a+1$ in the next n trials. The absurdity of this result shows that it is illegitimate to assume the constancy of $f(x)$. Now when a/n is very small it is often reasonable to assume that, in the neighbourhood considered, all values of $\log x$ are equally probable *a priori*. Thus Blackett(2) found that 8 out of 415,000 α-ray tracks in air showed anomalous forks due to the penetration of nitrogen nuclei, giving $x = 1{\cdot}9 \times 10^{-5}$. In such a case one would naturally state the *a priori* probability in logarithmic terms, e.g. one might surmise a value of $\log x$ between -4 and -5. Now, if $F(x) = c\log x$, $f(x) = c/x$, and $\phi'(x) = -x^{-1}$. Hence the modal or optimal value of x is $(a+1)/(n+1)$, and the mean value is a. This is in accordance with expectation, and it appears reasonable, when a sample of n contains a members with the character x, and a is small compared with $n^{\frac{1}{2}}$, to take

$$dp = \frac{e^{-nx}x^{a-1}}{(a-1)!}\,dx \qquad (3)$$

as an expression for the probability distribution, instead of the usual formula given above. The probability of obtaining more than a given number in a subsequent sample can still be estimated by means of Pearson's(3) incomplete gamma function, but the value of a (Pearson's p) must be diminished by unity.

In particular if $a = 1$, $dp = e^{-nx}\,dx$. Thus although the probability that another X should be obtained in the next sample is $1/n$, as it should be, the probability that $x < 1/n$ is $(e-1)/e$, which exceeds $\frac{1}{2}$.

We now come to the case where $a = 0$. Clearly equation (3) breaks down. Unless we have reasons to the contrary, we can no longer assume that $x = 0$ is an infinitely improbable solution, as we should do even if we took a value of $f(x)$ having a logarithmic infinity at $x = 0$ (and no higher type of infinity is permissible if $f(x)$ is to be an analytical function, since $\int_0^1 f(x)\,dx = 1$). In the

https://doi.org/10.1017/S0305004100010495 Published online by Cambridge University Press

case of many classes both of logical and physical objects we can point to a finite, though undeterminable, probability that $x = 0$. Thus there are good, but inconclusive, reasons for the beliefs that there is no even number greater than 2 that cannot be expressed as the sum of 2 primes, and that no hydrogen atoms have an atomic weight between 1·9 and 2·1. In addition, we know of very large samples of each containing no members possessing these properties. Let us suppose then, that k is the *a priori* probability that $x = 0$, and that the *a priori* probability that it has a positive value is expressed by $f(x)$, where $\underset{\epsilon \to 0}{\text{Lt}} \int_{\epsilon}^{1} f(x)\,dx = 1 - k$.

Then the probability that a population where $x = 0$ should give a sample of n with $a = 0$ is unity. The same probability for a finite value of x is $(1-x)^n f(x)$, or approximately $e^{-nx} f(x)$. Hence the probability, after observing the sample, that $x = 0$ is

$$\frac{k}{k + \int_0^1 (1-x)^n f(x)\,dx}.$$

If $f(x)$ is constant this is $\frac{(n+1)k}{nk+1}$, and if, as seems plausible, $f(x)$ increases as x tends to zero, the order of magnitude of the definite integral is still $1/n$. This is so even if $f(x)$ has a logarithmic infinity at $x = 0$. Hence as n tends to infinity the probability that $x = 0$ tends to unity, however small be the value of k. The mean value of x, i.e. the probability that the next individual observed will have the character X, is

$$\frac{\int_0^1 x(1-x)^n f(x)\,dx}{k + \int_0^1 (1-x)^n f(x)\,dx}.$$

For constant $f(x)$ this becomes $\frac{1}{(n+2)(nk+1)}$, and in general is of the order n^{-2} for large values of n. Now the usual theory, according to which $k = 0$, gives $1/(n+2)$ for this probability. However, so long as k is not zero, the probability, for sufficiently large values of n, is clearly of the order n^{-2}, rather than n^{-1}. This seems to be a more reasonable estimate of the validity of an induction than that generally given.

The above theory may readily be generalized when x is not the proportion of a certain type, but any statistic of the population, e.g. its mean, median, or skewness. Here also we commonly have an idea of the nature of $f(x)$, from past work on similar populations. The formal solution is a modification of Fisher's(4) method of maximum likelihood. Let $e^{L(x)}$ be the probability that the observed

https://doi.org/10.1017/S0305004100010495 Published online by Cambridge University Press

sample should be drawn from a population in which the parameter considered has the value x. In this case Fisher showed (subject to the tacit assumption that all values of x in the neighbourhood of the optimal value are equally probable *a priori*) that the probability density of x is given by

$$dp = \frac{1}{(2\pi)^{\frac{1}{2}}\sigma} e^{-\frac{(x-\bar{x})^2}{2\sigma^2}} dx,$$

where $\bar{x}$, the optimal value, is the root of

$$L'(x) = 0, \text{ and } \sigma^{-2} = -L''(\bar{x}).$$

Fisher defines the likelihood of x as a quantity proportional to $e^{L(x)}$. This is a convenience of statement, but the introduction of the *a priori* probability density $f(x)$ allows the deduction of Fisher's results without introducing concepts other than those found in the theory of direct probability. The method of maximum likelihood in its complete form is only applicable where σ is somewhat smaller than the difference between $\bar{x}$ and the upper or lower limits which it can possibly attain, and where the graph of $L(x)$ can be adequately fitted by a parabola in the neighbourhood of $\bar{x}$.

In general when $f(x)$ is not constant, the probability that the observed sample should be drawn from a population in which the parameter x lies between x and $x+\delta x$ is $e^{L(x)+\phi(x)}\delta x$. Hence the optimal value of x is given by

$$L'(x) + \phi'(x) = 0.$$

Now provided that the approximation $L'(x) = (x-\bar{x})L''(\bar{x})$ is valid, and that $\phi'(x)$ does not vary rapidly with x in the neighbourhood of $\bar{x}$, we have approximately for the optimum

$$x = \bar{x} - \frac{\phi'(\bar{x})}{L''(\bar{x})} = \bar{x} + \sigma^2\phi'(\bar{x}). \qquad (4)$$

Since $L(x)$ and its derivatives vary with n, the number of individuals in the sample, the approximation $x = \bar{x}$ can be made as accurate as we desire by increasing n, so long as $\phi'(\bar{x})$ is finite. It is, however, also necessary, if the assigned standard error σ is to have any significance, that it should be large compared with $\sigma^2\phi'(\bar{x})$. Hence it is a sufficient condition for the validity of Fisher's theory that $[\phi'(x)]^2$ should be small compared with $-L''(\bar{x})$ in the neighbourhood of $x = \bar{x}$. This is entirely in accord with the dictates of common sense. For example, if we found a mean height of $6\frac{1}{2}$ feet in a sample of a human population we should regard such a value as very improbable *a priori*, i.e. we should consider $\phi(x)$ to be small. Further we should regard the *a priori* probability as rapidly diminishing with the height, i.e. we should assign a large negative value to $\phi'(x)$. To be convinced of the validity of our

https://doi.org/10.1017/S0305004100010495 Published online by Cambridge University Press

observation we should therefore demand an unusually large sample, i.e. an unusually large value of $-L''(\bar{x})$. So far, however, no attempt has been made to apply any quantitative criterion to aberrant statistics of this type. It would seem, however, that where the number of comparable populations already observed is large enough to yield a frequency curve (as is the case for example with human cranial indices), there might be a certain gain in accuracy were equation (4) applied to aberrant populations. The correction to the standard error is of the same form as that found in the simple case of sampling.

It will be seen then that the basic assumption by which the formulae of the theory of inverse probability can be deduced from those of direct probability is that a certain function should not only be regular, but make no approaches to irregularity such as are implied by large values of its derived function. This assumption may be regarded as a particular case of the postulate of the uniformity of nature. But it is definitely inapplicable in special cases, such as that where none, or very few, members of a class possess a particular character, and must everywhere be applied with caution. Such caution is much more probable if the precise nature of the assumption is realized.

REFERENCES.

(1) R. von Mises, *Wahrscheinlichkeitsrechnung* (1931).

(2) P. M. S. Blackett, *Proc. Roy. Soc.* 107 A, p. 349 (1925).

(3) K. Pearson, *Tables of the incomplete* Γ*-function* (1922).

(4) R. A. Fisher, *Phil. Trans. Roy. Soc.* 222 A, p. 309 (1922).

https://doi.org/10.1017/S0305004100010495 Published online by Cambridge University Press