203

THE LOGICAL INVERSION OF THE NOTION OF THE RANDOM VARIABLE

By R. A. FISHER
University of Cambridge

The mathematical concept of the random variable has been of value in giving formal precision to all statements of the theory of probability which depend on the notion of frequency. For the definition of such a random variable we require a probability P, such that $0 \leqslant P \leqslant 1$, defined as a function of the variable, in such a way that the definition may involve parameters,

$$\mathrm{P}\,(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{(x-\mu)/\sigma} e^{-\frac{1}{2}u^2}\,du \qquad \ldots \quad (1)$$

in which the distribution of the random variable x is defined, and is dependent on the two parameters μ and σ. P is then said to be the probability that the random variable shall be less than any assigned value x.

This definition brings into clear light a point, which has often been indicated by logical writers, but which is obscured by the forms of common speech, that in respect of any concrete object or event, the value of the probability will depend on the population to which, for purposes of discussion, that object or event is appropriately deemed to belong. The probability, or frequency, with which it rains at Manchester (tacitly including equally all periods of the day) is not necessarily the same as the frequency with which it rains at 9 a.m. in Lancashire (tacitly including equally all areas in the County) ; consequently an evaluation of probability, to be mathematically correct, must utilise all the information supplied by the data, and must explicitly exclude only those elements which in any particular problem are known to be irrelevant. Recognising this, however, we shall be free, in relation to any concrete event, to evaluate probabilities relevant to that event regarded as a member of any hypothetical population to which it belongs, and for which we possess the appropriate information.

In the observational sciences, the observations, which are particular and concrete events, seem to be capable of use, and are now freely used, as a basis for probability statements respecting parameters, the values of which, apart from the information given by such observations, are unknown. This type of reasoning arises from what have been called the tests of significance, and the probability statements arrived at are distinguished as statements of fiducial probability, to avoid confusion with the earlier usage of "inverse probability" on the basis of Bayes' postulate. The object has been to arrive at probability statements of practical use without assuming Bayes' postulate, not, of course, to find another method of calculating the probabilities a *posteriori* of Bayes, which, without that postulate, are obviously indeterminate.

In recent times one often-repeated exposition of the tests of significance, by J. Neyman, a writer not closely associated with the development of these tests, seems liable to lead mathematical readers astray, through laying down axiomatically, what is not agreed or generally true, that the level of significance must be equal to the frequency with which the hypothesis is rejected in repeated sampling of any fixed population allowed by hypothesis. This intrusive axiom, which is foreign to the reasoning on which the tests of significance were in fact based, seems to be a real bar to progress, for I am not aware that Neyman, or his followers, have succeeded in putting forward any tolerable solution even to the problem of the comparison of the means of two samples drawn from different normal populations ; a satisfactory solution of which was given by Behrens in 1929. The current tables, both those of Sukhatme and my own, are based on formulae equivalent to those of Behrens though somewhat more general.

The purpose of this note is therefore to discuss, without prejudice to the question of axiomatic foundations, the process of reasoning by which we may pass, without arbitrariness or ambiguity, from forms of statement in which observations are regarded as random variables, having distribution functions involving certain fixed but unknown parameters, to forms of statement in which the observations constitute fixed data, and frequency distributions are found for the unknown parameters regarded as random variables.

In tests of significance we define a class of events (samples) having a certain specified frequency (5 per cent, 1 per cent etc., according to the level of significance chosen) among a population A, of samples to which we regard the sample observed, to which the test of significance is to be applied, to belong. This population must resemble the sample observed in all relevant respects ; indeed the argument is not impeded if it be taken to resemble the sample in all observable respects.

Thus in testing the significance of the mean of a normal sample, we have as data the observed values $x_1, x_2, \ldots, x_n$, and are concerned to draw what inference may be possible about the mean of the hypothetical normal population, from which these observations form by hypothesis a random sample. It will in this case be sufficient to consider a population of samples characterised by the two statistics $\bar{x}$ and s, and by the ancillary statistic N, defining these by the relations

$$\mathrm{N}\bar{x} = \mathrm{S}(x)$$
$$(\mathrm{N}-1)\mathrm{s}^2 = \mathrm{S}(x-\bar{x})^2$$

The reason for which the limited specification is sufficient is that it has been shown that, for given values of N, $\bar{x}$ and s, the distribution of any functionally independent statistic T is independent of the population (supposed normal) from which the sample has been drawn.

The members of this population A of samples differ among themselves in the values, μ and σ , of the means and standard deviations of the parent populations from which they are drawn. In designing a test involving the mean only we require a test or pivotal quantity, involving μ only, the distribution of which is wholly independent of the parameters. This is supplied by $\mathrm{t} = (\bar{x}-\mu)\sqrt{\mathrm{N}}/\mathrm{s}$

The distribution of this quantity depends on N, for the number of degrees of freedom is $n = N-1$. This is a matter of no consequence, for the function of t and N,

$$\mathrm{P}(t) = \frac{\left(\frac{n-1}{2}\right)!}{\left(\frac{n-2}{2}\right)!}\ \frac{1}{\sqrt{\pi n}} \int_t^{\infty} \left(1+\frac{u^2}{n}\right)^{-(1+n)/2} du, \qquad .. \quad (2)$$

might equally well have been chosen as the pivotal quantity, and its distribution $0 \leqslant P \leqslant 1$, $df = dP$, is independent of N also. The essential step which we now take is to assert that, whatever may be the character of our sample, we shall derive the probability that μ shall fall in any range from the known probability that P and t shall fall in the ranges corresponding for these quantities. It is in this sense that we interpret the somewhat paradoxical statement, with which the problem at first confronts us, that a sample having *known* characteristics is a *random* sample of an unknown population. The properties of variable statistics derived from observations, defined as random variables with distribution functions given in terms of known parameters, are used to establish the connections, (1) and (2), which once established serve to give meaning to the situation which we encounter in practice, in which the statistics are observable, but the parameters unknown.

It is instructive to compare the general form of the fiducial argument set out above with a special case of great simplicity, suitable for examining its logical cogency.

Let μ be the median of a distribution of which nothing is known save that its probability integral is continuous (e.g. it need not be differentiable). Let x_1 and x_2 be two observations of the variate ; then for any given value of μ it will be true that

(1) in one case out of 4 both x_1 and x_2 will exceed the median,

(2) in two cases out of 4, one value will exceed and the other be less than the median,

(3) in one case out of 4, both will be less than the median.

If a stands for the number of observations less than the median, then a will be a pivotal quantity involving both the unknown parameter and the observations, and having a sampling distribution independent of the parameter, i.e. a takes the values 0, 1, and 2 with probabilities $\frac{1}{4}$, $\frac{1}{2}$ and $\frac{1}{4}$ respectively.

Recognising this property we may argue from two given observations, now regarded as fixed parameters that the probability is $\frac{1}{4}$ that μ is less than both x_1 and x_2, that the probability is $\frac{1}{2}$ that μ lies between x_1 and x_2, and that the probability is $\frac{1}{4}$ that μ exceeds both x_1 and x_2. The argument thus leads to a frequency distribution of μ, now regarded as a random variate.*

The idea that probability statements about unknown parameters cannot be derived from *data* consisting of observations can only be upheld by those willing to reject this simple argument.

I do not think that any other interpretation of the known as a random variable of the unknown would be rational and consistent. It is, however, worth while to endeavour to see how, long after the tests had become established in statistical practice, a misunderstanding of their logic should have been possible.

Now, the derivation of Student's distribution involves the removal of a variable by integration, in a way which happens to be capable of such misinterpretation. The relation between a sample from a normal population and the population from which it is drawn is expressible in terms of two independent random variables,

* I have, of course, given elsewhere (Obituary "Student") the general fiducial distribution for this problem, namely

$$N\,!\; dp_1\, dp_2 \ldots dp_N, \qquad p_1 > p_2 > \ldots > p_N$$

where $p_1\, p_2 \ldots p_N$ are pivotal quantities defined as the fractions of the population sampled exceeded by each of N observations in descending order of magnitude.

$$(\bar{x}-\mu)\sqrt{\mathrm{N}}\,/\,\sigma = \xi, \qquad .. \quad (3)$$

where ξ is a random variable, such that

$$\mathrm{P}(\xi)=\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\xi} e^{-\frac{1}{2}x^2}\,dx \qquad .. \quad (4)$$

and

$$\tfrac{1}{2}\,ns^2/\sigma^2=u$$

where u is a random variable, with the Eulerian distribution

$$\mathrm{P}(u)=\frac{1}{\left(\frac{n-2}{2}\right)!}\int_{0}^{u} x^{(n-2)/2}\,e^{-x}\,dx.$$

Both random variables involve the unknown parameter σ, and to eliminate this we put

$$t=(\bar{x}-\mu)\sqrt{\mathrm{N}}/s=\xi\sqrt{n/2u}, \quad \text{so that} \quad \tfrac{1}{2}\xi^2=(t^2/n)u,$$

and a random variable independent of σ may be obtained by integrating, with respect to u the simultaneous frequency element

$$\frac{1}{\sqrt{2\pi}\left(\frac{n-2}{2}\right)!}\,e^{-u(1+t^2/n)}\,\sqrt{2u/n}\;u^{(n-2)/2}\,du\,dt$$

leading to the familiar distribution of t. It so happens in this case, though it is of no advantage to the test of significance, that, with the unknown σ, the known quantity s is also eliminated. The misunderstanding to which I refer is that which represents the elimination of s as the purpose of the integration, as though we needed to trace out the frequency distribution of t in successive samples from a population with σ given. This is wholly illogical since s and not σ is in fact used in making the test.

Such a mistake is of only academic importance in the simple case of Student's test, but becomes a real hindrance to understanding the more complex tests based upon it. When, for example, we have samples from two different populations, their two unknown variances are both eliminated, giving the simultaneous distribution of two independent values of t. To test whether the means could be equal, if μ stand for the common mean, we have

$$\bar{x}_1-\mu=s_1t_1, \quad \bar{x}_2-\mu=s_2t_2, \quad \text{whence } \bar{x}_1-\bar{x}_2=s_1t_1-s_2t_2,$$

and, knowing the distribution of t_1 and t_2 we are in a position to calculate what value of $\bar{x}_1-\bar{x}_2$ will be exceeded with any chosen frequency. The view that the purpose of the integration was to eliminate s_1 and s_2 would naturally require that the frequency distribution of $\bar{x}_1-\bar{x}_2$ would be determined for all sampling values of s_1 and s_2, i.e. as a normal distribution with variance $\sigma_1{}^2/\mathrm{N}_1+\sigma_2{}^2/\mathrm{N}_2$ so reintroducing the unknown σ_1 and σ_2. Reference to σ_1 and σ_2 puts the argument back to the stage for which Bayes method was originally developed.

The population of samples of which some standard fraction has a value of $\bar{x}_1-\bar{x}_2$ exceeding the level found for significance is not, however, a population of samples drawn in succession from some fixed population of unknown characteristics, but the population of samples (identical with the observed in the whole set of statistics N_1, N_2, s_1, s_2) determined by the fundamental random-variable relation by which sample and population are connected.

Paper received : 18 *April*, 1945.