STATISTICS IN FUNCTION SPACE

BY

D D. KOSAMBI, *Poona.*

[Received 23 August 1943]

The main purpose of this note is to develop statistical methods for discrimination between samples consisting of whole curves.

We take the observables as simple curves of type $y=f(x)$, the functions $f(x)$ being all single-valued, of bounded variation, continuous (though step-wise continuity—as for a sample of histograms—would cause no difficulties), defined on a finite closed interval which may be taken without loss of generality as $0 \leqslant x \leqslant 1$ by suitable choice of origin and scale. The methods developed for such curves apply directly to diagrams in polar co-ordinates $r=f(\theta)$, $0 \leqslant \theta \leqslant 2\pi$; by an obvious extension, to suitably restricted surfaces, (say crania) or multidimensional varieties. Peano's space-filling curves, Jordan curves of positive area, are naturally excluded.

The problem clearly resolves itself into four components: (1) to define a normal distribution in function-space, (2) to deduce useful consequences of such normality, assumed to hold for the population of curves, (3) to devise new methods of calculation where necessary, and (4) to examine the generality of the approach.

1. The probability P associated with a multivariate normal distribution is the definite integral, over the proper region, of

$$(2\pi)^{-k/2} e^{-\phi/2}\, dV, \tag{1.1}$$

where k is the number of variates in which ϕ is a positive definite quadratic form, and dV is the associated volume element. That is, the same transformation that reduces ϕ to a sum of squares makes $dV = dx_1 dx_2 \ldots dx_k$, the whole space being recognizable as an ordinary k-dimensional Euclidean manifold with $\phi = r^2$ as the square of the distance. A continuous function is determined completely by its values on a set of points everywhere dense on (0, 1), say all rational points; a function in general, therefore, has an infinity of co-ordinates. As approximation is possible by increasing k indefinitely, the first step would be to generalize distance and the quadratic form ϕ.

To this end, we assume the existence of a continuous symmetric kernel $K(s, t)$, positive definite or semi-definite. Then the distance between any two of our functions $f(x)$, $g(x)$ is given by

$$r(f, g) = \phi(f-g) = \iint K(s,t)\{f(s)-g(s)\}\{f(t)-g(t)\}\,ds\,dt. \tag{1.2}$$

The range for all otherwise undefined integrals in s and t is (0, 1) for each variable; here, the unit square. Restricting the population of functions to be such that $K(s, t)$ gives a definite ϕ therein according to the definition of (1.2), we shall have $r(f, g) = 0$ if and only if $f \equiv g$ with respect to the mechanism of observation; it follows, therefore, that r obeys all the basic postulates for distance including the triangular inequality. The normal distribution in function-space could be taken as defined by $c \exp(-\phi/2)\, dV$.

Unfortunately, not all the terms of this probability-density can be given a meaning that is useful in practice. As $(2\pi)^{-k/2} \rightarrow 0$ when $k \rightarrow \infty$, c can only be specified by the restriction that the total probability equals unity; also, the "volume element" dV may be given a direct mathematical meaning (**1**), but not one of much real use. To surmount this obstacle, we resort to a choice of independent variates that reduces ϕ to a diagonal form. This amounts to taking $K(s, t)$ in its canonical form (**2**, 117, 114),

$$K(s, t) = \Sigma\, \sigma_i^2 \phi_i(s) \phi_i(t). \tag{1.3}$$

The ϕ_i are the orthonormal characteristic (eigen-) functions of the kernel, σ_i^2 the corresponding characteristic values ($= 1/\lambda_i$ in the notation of **2**), all positive with $\Sigma\sigma_i^4$ convergent (**2**, 111). The orthogonal or independent co-ordinates for any function $f(t)$ are obviously the "Fourier" co-efficients $x_1, x_2, \ldots, x_r, \ldots$ with

$$x_r = \int f(t)\, \phi_r(t) dt, \quad f(t) = \sum_r x_r \phi_r(t). \tag{1.4}$$

As $K(s, t)$ is definite *for the population,* every function observed can be so represented. The series will converge uniformly and absolutely (**2**, 114) for every function that is the K-transform of any piecewise continuous function, a restriction which we place upon the population.

We now define normality in the function-space to mean normal distribution in each of the x_i. Without loss of generality, the population mean for the function-space and hence for each x_i may be taken as zero. The variances will be σ_i^2. That is, our ϕ has to be taken as generalizing not the k-dimensional population quadratic form but the one that enters into the characteristic function of the distribution, the Fourier transform. In Euclidean space, both become equal to r^2 when ϕ is expressed as a sum of squares. Our choice for function-space is dictated by the implication of (1.4) that Σx_i^2 converges whence $x_i \rightarrow 0$ as $i \rightarrow \infty$, which can be made to hold *in probability* for random x_i if and only if $\sigma_i^2 \rightarrow 0$. If ϕ were to be taken as the population (probability density) quadratic form, or variances would become $1/\sigma_i^2$; the two kernels must be reciprocals, as is seen by application of the Fourier transform to the k-variate distribution. Since we deal here with kernels of the first kind, only one can be properly defined, in general, the other "existing" only in symbolic form as is seen by the fact that except in the degenerate case, the series of squares of characteristic values and the series of squares of their reciprocals cannot both converge simultaneously. To sum up, we may formulate a

DEFINITION: *Normal distri ution in function-space will be taken to mean normal distribution for each variate x_i of an (independent) infinite sequence $x_1, x_2, \ldots$ with all population means zero and variances $\sigma_1^2, \sigma_2^2, \ldots$ These x_i are the "Fourier" coefficients of a random function of the space with respect to the orthonormal characteristic functions $\phi_1(t), \phi_2(t), \ldots$ with characteristic values $\sigma_1^2, \sigma_2^2, \ldots$ belonging to a continuous, symmetric, positive, definite (in the manifold of admissible functions) kernel $K(s, t)$ defined over the unit square $0 \leqslant s \leqslant 1, 0 \leqslant t \leqslant 1$. The kernel $K(s, t)$ thus completely determines the distribution.*

2. This definition has the initial advantage of covering all finite-dimensional cases, represented by degenerate kernels where all but a finite number of the variances σ_i^2 vanish. Conversely, it allows approximation by degenerate kernels and application of methods developed for *k*-variate distributions. These can be used together to prove, for example, that

The sum of two normally distributed function-variates is also normally distributed with mean the sum of the population means and kernel the sum of the two given kernels.

This follows directly from the definition since $\exp(-\phi/2)$ is not the probability density but the characteristic function of the distribution. The characteristic function of the sum of two variates is the product of the two characteristic functions. In particular, the mean of a sample of *n* functions chosen at random from the same normal population will have the population mean and kernel K/n; one degree of freedom will be lost in measuring from the sample mean, and so on. If the same set of orthonormal functions covers both kernels, then we may add corresponding variances as usual; if not, we can at least state that the sum-variances do not decrease (**2**, 113 with an obvious correction).

What seems to me to be the most important consequence of our definition rests upon a basic theorem of Kolmogoroff (**3**). Using the letter P to indicate the probability and E the expectation of the events bracketed, this may be stated as follows.

Given a random sequence $u_1, u_2, \ldots, u_r, \ldots$, *the probability of the convergence of the series* Σu_r *is unity if there exists some random sequence* $v_1, v_2, \ldots, v_r, \ldots$ *such that thethree series* $\Sigma P(u_r \neq v_r)$, $\Sigma E(v_r)$, $\Sigma E[(v_r - E(v_r))^2]$ *all converge. If no such sequence exists, the probability for the convergence of* Σu_r *is zero.*

In our case we take $u_r = v_r = x_r\phi_r(t)$, so that the first two series converge by hypothesis. The third is

$$\Sigma E(x_r^2\phi_r^2) = \Sigma\phi_n^2(t)\, E(x_n^2) = \Sigma\sigma_n^2\phi_n^2(t) = K(t, t) \text{ (by } \mathbf{2}, 110).$$

The structure of Kolmogoroff's proof shows that in our case convergence and uniform convergence go together. We conclude, therefore, that

A random sequence of the variates $x_1, x_2, \ldots, x_r, \ldots$ *of our definition represents with unit probability a function of the normally distributed function-population.*

This replaces the Riesz-Fischer theorem, proving a 1-1 correspondence between random functions and random sequences of coefficients, in the sense of unit probability. The Riesz-Fischer theorem would require, for unit probability, the convergence of $\Sigma\sigma_r^2$, and give only convergence in the mean for $\Sigma x_r\phi_r(t)$.

Let $y_i = f(t_i) = \Sigma x_r\phi_r(t_i)$ be the ordinate at a fixed point t_i as abscissa. From $E(x_ix_j) = 0$, $E(x_i^2) = \sigma_i^2$, we obtain

$$E(y_i^2) = \Sigma\sigma_r^2\phi_r^2(t_i) = K(t_i, t_i)$$
$$E(y_iy_j) = \Sigma\sigma_r^2\phi_r(t_i)\phi_r(t_j) = K(t_i, t_j). \qquad (2.1)$$

The matrix $\| E(y_iy_j) \| = \| K(t_i, t_j) \|$ of covariances may be regarded as the symbolic product of $\| \sigma_r\phi_r(t_i) \|$ with the transposed matrix. In case the kernel is degenerate and there are more ordinates $y_1, y_2, \ldots, y_m$ than characteristic functions $\phi_1, \phi_2, \ldots, \phi_k$, it is obvious that the determinant $|E(y_i y_j)|$ will vanish. Conversely, if $|K(t_i, t_j)|$ vanishes identically, the kernel is necessarily degenerate as its Fredholm expansion [2,122] breaks down into the ratio of two polynomials. As the y_i are convergent (in the sense of unit probability) linear combinations of normally distributed variates, we have proved that

For any fixed abscissa t, in a normally distributed population of functions, the ordinate is normally distributed with variance $K(t, t)$. The covariance between values of the functions at two points s and t is $K(s, t)$. The distribution of ordinates at k fixed points is multivariate normal, and is a proper distribution except when the kernel $K(s, t)$ is itself degenerate with less than k characteristic values.

In this, of course, the nodal points of the entire set of functions, points where $K(t, t) = 0$ in particular, must be avoided when selecting the k points for measuring abscissae. An example would be $\phi_r = \sqrt{2} \sin \pi r t$ and the end points of the interval, $t_1 = 0$, $t_2 = 1$.

3. The usefulness of the preceding section is manifest, as the population kernel $K(s, t)$ would remain unknown in practice even when the hypothesis of normality is granted. Our theorems enable us to proceed by the methods of the ordinary multivariate normal distribution, measuring ordinates at suitably chosen abscissae. The meteorologist would be justified in working with temperatures taken at noon and midnight, but not necessarily with his maximum and minimum temperatures, which are measured at varying times of the day. The anthropologist's characters and indices would be less justified, than, say, measurements from the ear orifice to the profile at fixed angles from the line joining the orifice to the base of the nose. Coefficients on the harmonic analyzer, regression coefficients in properly chosen orthogonal functions (whether they belong to the kernel or not) are also to be regarded as co-ordinates in multivariate normal distribution, provided the fitting when done by values at fixed points is done with the same fixed abscissae for each curve.

Given a sample of n curves, $y = f_1(x), f_2(x), \ldots, f_n(x)$, the best estimate of the population mean $\mu(x)$ and the population kernel $K(s,t)$ are given respectively by

$$m(t) = \frac{1}{n}\sum f_i(t); \quad k(s,t) = \frac{1}{n-1}\sum [f_i(t)-m(t)]\,[f_i(s)-m(s)] \tag{3.1}$$

as follows obviously from the foregoing. Large sample theory means calculation of these sample-functions and therewith the characteristic-functions and values [which will approximate those of the population]. Hotelling's T^2, Fisher's discriminating-function and such methods for discrimination would also apply without restriction provided the number of points for taking ordinates did not exceed the number of functions in the samples.

But in many cases the complete curves are recorded automatically with less trouble and more accuracy than for a finite number of observations on the same material. In that case, we could, if the proper machines were available, calculate the sample-functions in (3.1) and thereafter the "Student" ratio $t(x)$ or Fisher's $z(x)$ from two given samples for every point of the abscissa $0 \leqslant x \leqslant 1$. Corresponding to these or to any other statistics we shall get a local probability $p(x)$ as a function over the unit interval. In methods of discrimination, one may choose a single point, say the point where $p(x)$ takes on its maximum value in the closed unit interval; the corresponding value of x gives the abscissa where the maximum discrimination has been achieved. This, in a way, is the determination of the best [in the obviously restricted sense] linear combination of the unknown co-ordinates x_r, the "Fourier" coefficients with respect to the unknown population orthogonal functions. But, again, one is tempted to ask whether something more could not be done, whether one could not calculate or measure a single probability for the whole interval or for any given sub-interval, instead of a point probability. What is required is not $p(x)$ but a $P(\alpha, \beta)$ for any given $0 \leqslant \alpha < \beta \leqslant 1$, on the basis of the two samples and any given statistic. The most that can be done here is to show that such questions need not be meaningless.

Suppose the kernel to have a single non-negative characteristic function $\phi(t) \geqslant 0$ and characteristic value σ^2. We ask for the probability that a sample-function $f(t) = x\phi(t)$ lies between the two limits $a(t)$ and $b(t)$ throughout an interval (α, β) in $(0, 1)$. Then this probability is

$$\frac{1}{\sigma\sqrt{2\pi}}\int_{x_1}^{x_2} e^{\frac{-x^2}{2\sigma^2}}\,dx, \tag{3.2}$$

if x_2 is the greatest value of x satisfying $x\phi(t) \leqslant b(t)$ and x_1 the least for $x\phi(t) \geqslant a(t)$ in (α, β), with $x_2 > x_1$; the probability is zero otherwise. A similar approach is possible for $\phi(t)$ with changes of sign or more than one characteristic function. The general question, for examples of the type chosen for illustration, depends upon the correspondence that can be set up between two different types of function-*lattices*, not merely function-spaces, with measure and maps upon the unit hypercube in infinitely many dimensions (torus space).

The calculating machines, under the circumstances that now limit my activity, cannot go beyond the stage of design. The fundamental ideas will be made clearly by the two schematic figures appended here in the hope of doing service to some more fortunately situated experimenter. Figure 1 shows how $\Sigma f_i(x)$ may be drawn by means of templates and a wire passing over a system of alternately fixed and moving pulleys, suggested by Kelvin's tidal machine. The formulae of (3.1), in particular the most important ones for $m(t)$ and $k(t, t)$, depend upon the operations of addition, summation of the square, subtraction, and division by a chosen factor. For the pulley machine, subtraction is possible by reversing the direction

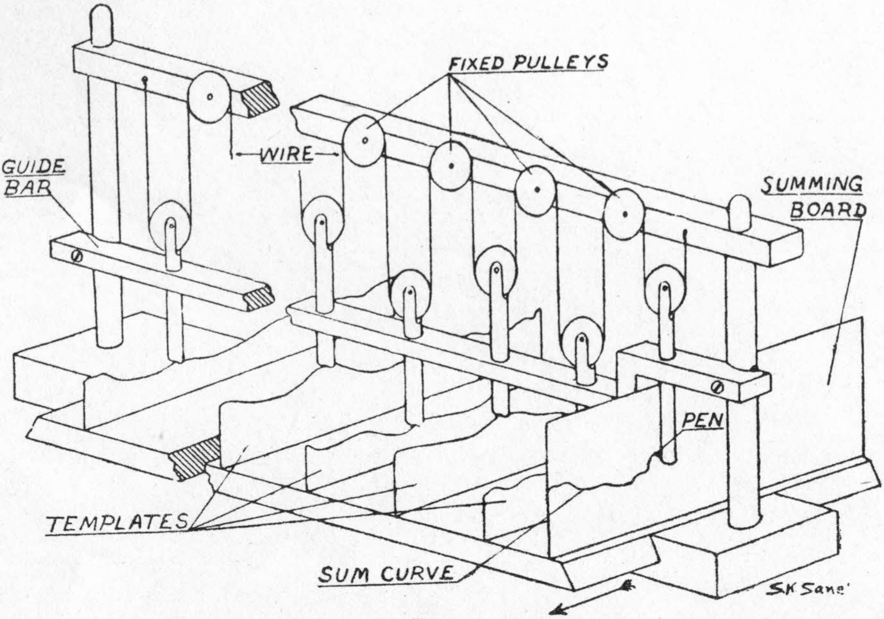

Fig. 1.

of the wire or rather substituting a moving for a fixed pulley; or by using as template the conjugate curve to the one to be subtracted. Reduction of scale, i.e. division by a given number, will have to be done by a pantagraph, or some such device. Both of these introduce errors, and there is the additional difficulty of getting material for templates that will be stiff enough to stand up under the weights, and smooth enough to allow all the templates to be pulled through on their rack without sticking. For sum-squares, and sums of products, the arrangement has to be extended to the measurement of torque and moments, the simple pulley-machine being inadequate.

The second instrument is suggested by the high fidelity with which sound is recorded on and reproduced by cinematography. Here, the area under the curve is cut out of a standard sheet of paper, and scanned by means of movement past a narrow fixed slit. The light that falls on the slit is of uniform intensity throughout, in the first instance. The film is coupled with the template, and

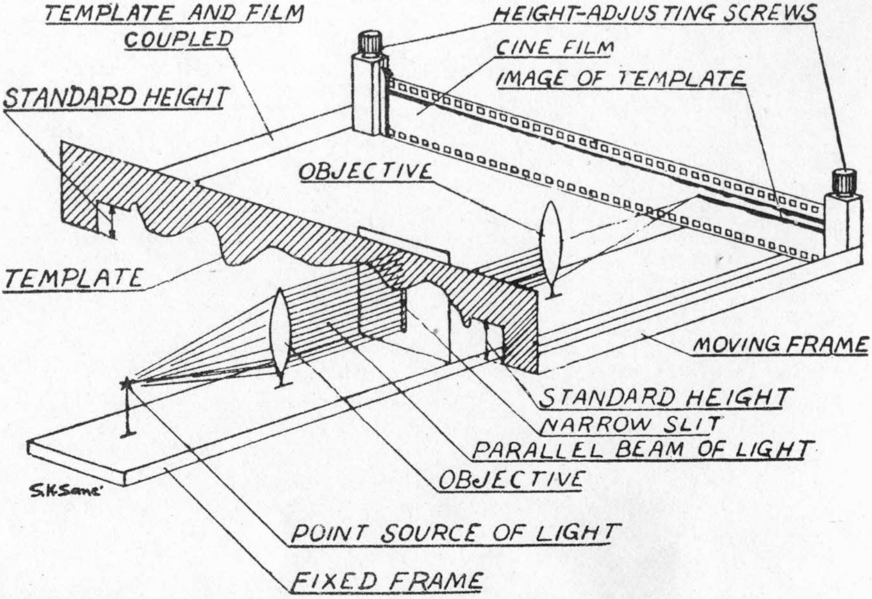

FIG. 2.

both are drawn through with uniform speed. The lens reduces the curve in height, but not in length, and by means of a vertically movable rack, many such curves, say at least a dozen, may easily be recorded on a single film. At the end of each curve-template, a standard height is cut out of the template material.

The film is developed and printed as usual, and focussed back in all its width through another slit on to a photo-electric cell. The current recorded will be proportional to $\Sigma f_i(x)$ at each point, standardization being achieved by means of the fixed heights cut out of each template at the end of the curve. The factor $1/n$ or $1/(n-1)$ can also be set thereby, adjusting the primary current, or putting the proper shunt across the current-recording device.

Sums of squares are easily obtained by varying the intensity as well. This is best done by means of a photo-electric cell coupled with the moving template rack. This cell would regulate the current supplied to the light, so that we should have the product of the height of the curve by the intensity as $f_i^2(x)$. The difficulty here, of course, is in the law of darkening, and very much more careful adjustments will have to be made. The same method allows function-covariances to be calculated, coupling one set of templates to the photo-electiic cell and the other to the slit-rack.

The law of resistance in electric circuits in parallel shows obvious means of calculating harmonic means. For the two-dimensional kernels $K(s, t)$, the best methods would seem to be those derived from television. Such instruments are now being devised by others for work in a single dimension. If successful, the need for cutting templates would be obviated, with a gain in accuracy.

4. From the purely theoretical point of view, we have ignored many other possibilities. Some of these were mentioned in a former exploratory approach [**1**]. I give an example to show that theoretical generality is certainly possible, in defining the normal distribution, but that the usual facilities such as the central limit theorem, the chi-square and other tests used in practice, in short the whole mechanism of everyday statistics is invalidated.

The population is defined by functions

$$\phi_r(t) = \sqrt{2} \sin \pi r t,$$

$$f(t) = \Sigma 3^{-n/2} a_n \phi_n(t), \text{ where } a_n = 0 \text{ or } 2. \qquad (4\ 1)$$

The kernel $K(s. t)$ is given by

$$K(s, t) = [\Sigma 3^{-i/2} \phi_i(s)]\ [\Sigma 3^{-j/2} \phi_j(t)], \qquad (4{\cdot}2)$$

being thus degenerate of the first order. It follows that $r(f, g)$ for two functions defined by sequences a_n, b_n is given by $r = |\Sigma(b_n - a_n)/3^n|$. If the sequence a_n is regarded as defining a point of Cantor's ternary set [Discontinuum], expressed by the same sequence of zeros and twos in a ternary expansion, it is seen that there is a 1-1 correspondence between points of the set and our population of admissible functions; moreover, the distance between two functions is now the distance between the two corresponding points of the line segment. Now Hausdorff (4) has shown that a measure can be defined over the Cantor set or over any similar set obtained by deletion of a central interval. If each of the surviving pieces is, at each stage, a fraction p of the original, the dimension r of the set is given by $2p^r = 1$, whence the Cantor set is of dimension log 2/log 3; a trifling extension of Hausdorff's argument will show that when the deletion is not symmetric, the surviving pieces being of fractions p and q at each deletion, the dimension r is given by $p^r + q^r = 1$. What concerns us here is the existence of the outer measure, by means of which we may define our integral of $c \exp(-\phi/2) dV$, where ϕ is the quadratic form defined by means of the kernel $K(s, t)$ of (4.2), and dV is the Hausdorff measure on the Cantor set, extended from the line segment (0, 1) to the entire line $-\infty, +\infty$ by simple translation, along with the coefficients and the functions of the space. This shows the possibility of generalizing the normal distribution beyond the needs of the statistician. Let it be noted that in choosing samples of functions from such a space, the gaps might pass unnoticed, because the set of points is perfect though nowhere dense, so that arbitrarily close to every function there could be other functions of the population. In this connection, we might also note the fundamental role of measure and distance, as contrasted to mere one-to-one correspondence. Every point on the line segment (0, 1) may be expressed by means of the two digits 0, 1 in the binary decimal scale; replacing the 1 of the binary by the 2 of the ternary set, we get a one-to-one correspondence, excepting for the points which, in the Cantor discontinuum, have an infinite sequence of 2's. As these doubly represented points are all rational in the binary scale, their totality forms a set of measure zero. But on the continuous line segment (0, 1), extended by translation, our normal statistics can be defined as usual.

This is of particular interest in considering such cases as the Kollektiv concept of von Mises, where we usually start by setting up a 1-1 correspondence between the throws of a coin and the binary expansion on (0, 1), which determines the measure *a priori* without further justification.

It gives me great pleasure to thank Mr. S. K. Sane for his careful execution of the two figures.

REFERENCES.

1. D.D. KOSAMBI : *Current Science*, 11 (1942), 271-4.

2. R. COURANT AND D. HILBERT : *Methoden d. Mathematischen Physik*, Vol. I, 2nd edition (Berlin), 1931.

3. A. KOLMOGOROFF : *Math. Annalen*, 99 (1928), 309-19.

4. F. HAUSDORFF : *Math. Annalen* : 79 (1919), 157-79.