The paper chosen for the 'Classics' Section in this issue is, 'Information and accuracy attainable in the estimation of statistical parameters', *Bulletin of the Calcutta Mathematical Society*, Vol.37, No.3, pp.81–91, 1945 and the author is C R Rao, or, in full, Calyampudi Radhakrishna Rao, writing from the 'Statistical Laboratory, Calcutta', which is now the Indian Statistical Institute, Kolkata.

To get a feel for the issues that Rao's paper addresses, let us look at two very simple examples. The fraction of all the voters who favour a given candidate in an election is clearly a very important quantity, and even before the election, or before the votes are counted, one can try and estimate this by an opinion poll of a small number of voters. Another example of estimation is taking the average height and weight of school children in different districts, to see if the mid-day meal scheme is effective. Clearly, one would not measure all the children in the state. In these two simple problems, it is very clear how to estimate the desired parameter, but it is still important to know how much one's estimate may be in error. In more complex, real world problems, there could be many possible alternative estimators (i.e., mathematical formulae involving the measured data), and the challenge is to pick the one where the average error is zero, and the mean squared error is the least.

This paper made a major contribution using the concept of 'information' which was introduced by Fisher [1] who is one of the pioneers of modern statistics. The paper went well beyond the earlier work by giving a rigorous lower limit to the attainable error for each estimator. This limit was independently published in 1946, one year after Rao's paper, in a book authored by the renowned statistician, Harald Cramer. It is known as the Cramer–Rao lower bound (the order of names is clearly alphabetical, not temporal!). The beauty of the result is that it is valid no matter what the law governing the randomness in the data may be – most of the earlier results were for a particular law. Immediately following this result is another which proved important later – how to improve an estimator by a certain averaging process. This was independently discovered by Blackwell in 1947, and now goes by the name of Rao–Blackwellisation [2]. All this by the end of Section 3 of the article!

The paper's contributions do not end here. For those familiar with some more advanced statistics, we can mention more important ideas introduced in the paper. Sections 4 and 5 generalize the earlier results to the case of many parameters. The lower bound on the variance of an ~~esti~~mator that C R Rao derived is applicable beyond unbiased (meaning average error is zero) estimators. Under certain regularity conditions on the probability distribution, this value is the variance of the asymptotic (large sample) distribution of maximum likelihood estimator and other optimal estimators. This is, in fact, useful for providing estimation errors for large samples. In sections 6 and 7, the paper introduces and analyses the notion of the distance between

probability distributions. This is actually important in the context of Bayesian inference [3, 4]. These sections discuss how the topological space of a parametric family of densities can be treated geometrically as a Riemannian manifold. In fact, the author freely uses notions like Christoffel symbols which are usually studied in general relativity! The volume element $p(\theta)d(\theta) = \sqrt{I(\theta)}\ d(\theta)$ (where θ is the parameter and $|I(\theta)|$ is the determinant of the Fisher information matrix) is then the uniform measure on this space. This is coordinate free and is invariant under reparametrization. It turns out that this measure is exactly the Jeffreys' prior distribution on the parameter, which can be considered an objective choice for a prior distribution.

The paper is remarkable in other ways. The author was just 25, and did not have a PhD degree! It is published in an Indian journal, something which most Indian scientists today would shy away from. There is no acknowledgement of any supervision – this reflects the independence of the author, and the vision of the founder of the group, P C Mahalanobis. The language and style are mature – there is no lack of confidence in this young research student! He did get his PhD soon after, working with Fisher in Cambridge between 1946 and 1948, and went on to a brilliant career at the ISI and, after retirement, at Pittsburgh and Penn State Universities in the US, rich with accomplishments and honours. He continues to be academically active even today [5]! This paper deserves to be read as a classic, both for its content and as marking a landmark event in Indian science.

Suggested Reading

[1] T Krishnan, Fisher's contributions to statistics, *Resonance*, Vol.2, No.9, pp.32–37, 1977.

[2] K B Athreya, M Delampady and T Krishnan, Markov chain Monte Carlo methods, Part IV, *Resonance*, Vol.8, No.12, pp.18–32, 2003.

[3] M Delampady and T Krishnan, Bayesian statistics, *Resonance*, Vol.7, No.4, pp.27–38, 2002.

[4] R Nityananda, The importance of being ignorant, *Resonance*, Vol.6, No.9, pp.8–18, 2001.

[5] B L S Prakasa Rao, C. R. Rao: A life in statistics, *Current Science*, Vol.107, pp.895–901, 2014.

Rajaram Nityananda
Editor
rajaram.nityananda@gmail.com

Mohan Delampady
Indian Statistical Institute, Bengaluru
mohan.delampady@gmail.com

Information and the Accuracy Attainable in the Estimation of Statistical Parameters

C Radhakrishna Rao

(Communicated by Mr. R C Bose–Received August 23, 1945

Introduction

The earliest method of estimation of statistical parameters is the method of least squares due to Markoff. A set of observations whose expectations are linear functions of a number of unknown parameters being given, the problem which Markoff posed for solution is to find out a linear function of observations whose expectation is an assigned linear function of the unknown parameters and whose variance is a minimum. There is no assumption about the distribution of the observations except that each has a finite variance.

A significant advance in the theory of estimation is due to Fisher (1921) who introduced the concepts of *consistency, efficiency and sufficiency* of estimating functions and advocated the use of the maximum likelihood method. The principle accepts as the estimate of an unknown parameter θ, in a probability function $\phi(\theta)$ of an assigned type, that function $t(x_1, \ldots, x_n)$ of the sampled observations which makes the probability density a maximum. The validity of this principle arises from the fact that out of a large class of unbiased estimating functions following the normal distribution the function given by maximising the probability density has the least variance. Even when the distribution of t is not normal the property of minimum variance tends to hold as the size of the sample is increased.

Taking the analogue of Markoff's set up Aitken (1941) proceeded to find a function $t(x_1, \ldots, x_n)$ such that

$$\int t\phi(\theta)\pi dx_i = \theta$$

and

$$\int (t-\theta)^2\phi(\theta)\pi dx_i \text{ is minimum.}$$

Estimation by this method was possible only for a class of distribution functions which admit sufficient statistics. Some simple conditions under which the maximum likelihood provides an estimate accurately possessing the minimum

variance, even though the sample is finite and the distribution of the estimating function is not normal, have emerged.

The object of the paper is to derive certain inequality relations connecting the elements of the *Information Matrix* as defined by Fisher (1921) and the variances and covariances of the estimating functions. A class of distribution functions which admit estimation of parameters with the minimum possible variance has been discussed.

The concept of distance between populations of a given type has been developed starting from a quadratic differential metric defining the element of length.

Estimation by minimising variance

Let the probability density $\phi(x_1, \ldots, x_n; \theta)$ for a sample of n observations contain a parameter θ which is to be estimated by a function $t = f(x_1, \ldots, x_n)$ of the observations. This estimate may be considered to be the best, if with respect to any other function t', independent of θ, the probabilities satisfy the inequality

$$P(\theta - \lambda_1 < t < \theta + \lambda_2) \not< P(\theta - \lambda_1 < t' < \theta + \lambda_2) \tag{2.1}$$

for all positive λ_1 and λ_2 in an interval $(0, \lambda)$. The choice of the interval may be fixed by other considerations depending on the frequency and magnitude of the departure of t from θ. If we replace the condition (2.1) by a less stringent one that (2.1) should be satisfied for all λ we get as a necessary condition that

$$E(t - \theta)^2 \not> E(t' - \theta)^2, \tag{2.2}$$

where E stands for the mathematical expectation. We may further assume the property of unbiasedness of the estimating functions *viz.*, $E(t) = \theta$, in which case the function t has to be determined subject to the conditions $E(t) = \theta$ and $E(t - \theta)^2$ is minimum.

As no simple solution exists satisfying the postulate (2.1) the inevitable arbitrariness of these postulates of unbiasedness and minimum variance needs no emphasis. The only justification for selecting an estimate with minimum variance from a class of unbiased estimates is that a necessary condition for (2.1) with the further requirement that $E(t) = \theta$ is ensured. The condition of unbiasedness is particularly defective in that many biased estimates with smaller variances lose their claims as estimating functions when compared with unbiased estimates with greater variances. There are, however, numerous examples where a slightly biased estimate is preferred to an unbiased estimate with a greater variance. Untill a unified solution of the problem of estimation is set forth we have to subject

the estimating functions to a critical examination as to its bias, variance and the frequency of a given amount of departure of the estimating function from the parameter before utilising it.

Single parameter and the efficiency attainable

Let $\phi(x_1, \ldots, x_n)$ be the probability density of the observations $x_1, x_2, \ldots, x_n$, and $t(x_1, \ldots, x_n)$ be an unbiased estimate of θ. Then

$$\int \ldots \int t\phi\pi dx_i = \theta. \tag{3.1}$$

Differentiating with respect to θ under the integral sign, we get

$$\int \ldots \int t\frac{d\phi}{d\theta}\pi dx_i = 1 \tag{3.2}$$

if the integral exists, which shows that the covariance of t and $\frac{1}{\phi}\frac{d\phi}{d\theta}$ is unity. Since the square of the covariance of two variates is not greater than the product of the variances of the variates we get using V and C for variance and covariance

$$V(t)V\left(\frac{1}{\phi}\frac{d\phi}{d\theta}\right) \not< \left\{C\left(t, \frac{1}{\phi}\frac{d\phi}{d\theta}\right)\right\}^2 \tag{3.3}$$

which gives that

$$V(t) \not< 1/I$$

where

$$I = V\left(\frac{1}{\phi}\frac{d\phi}{d\theta}\right) = E\left\{-\frac{d^2 \log \phi}{d\theta^2}\right\} \tag{3.4}$$

is the intrinsic accuracy defined by Fisher (1921). This shows that *the variance of any unbiased estimate of θ is greater than or equal to the inverse of I which is defined independently of any method of estimation.* The assumption of the normality of the distribution function of the estimate is not necessary.

If instead of θ we are estimating $f(\theta)$, a function of θ, then

$$V(t) \not< \{f'(\theta)\}^2/I. \tag{3.5}$$

If there exists a sufficient statistic T for θ then the necessary and sufficient condition is that $\phi(x;\theta)$ the probability density of the sample observations satisfies the equality

$$\phi(x;\theta) = \Phi(T,\theta)\psi(x_1, \ldots, x_n), \tag{3.6}$$

where ψ does not involve θ and $\Phi(T,\theta)$ is the probability density of T. If t is an unbiased estimate of θ then

$$\theta = \int t\phi\pi dx_i = \int f(T)\Phi(T,\theta)dT \tag{3.7}$$

which shows that there exists a function $f(T)$ of T, independent of θ and is an unbiased estimate of θ. Also

$$\int (t-\theta)^2\phi\pi dx_i = \int [t-f(T)]^2\phi\pi dx_i + \int [f(T)-\theta]^2\Phi(T,\theta)dT$$
$$\geq \int [f(T)-\theta]^2\Phi(T,\theta)dT, \tag{3.8}$$

which shows that

$$V[f(T)] \ngeq V(t) \tag{3.9}$$

and hence we get the result that *if a sufficient statistic and an unbiased estimate exist for θ, then the best unbiased estimate of θ is an explicit function of the sufficient statistic.* It usually happens that instead of θ, a certain function of θ can be estimated by this method for a function of θ may admit an unbiased estimate.

It also follows that if T is a sufficient statistic for θ and $E(T) = f(\theta)$, then there exists no other statistic whose expectation is $f(\theta)$ with the property that its variance is smaller than that of T.

It has been shown by Koopman (1936) that under certain conditions, the distribution function $\phi(x,\theta)$ admitting a sufficient statistic can be expressed as

$$\phi(x,\theta) = \exp(\Theta_1 X_1 + \Theta_2 + X_2), \tag{3.10}$$

where X_1 and X_2 are functions of $x_1, x_2, \ldots, x_n$ only and Θ_1 and Θ_2 are functions of θ only. Making use of the relation

$$\int \exp(\Theta_1 X_1 + \Theta_2 + X_2)\pi dx_i = 1, \tag{3.11}$$

we get

$$E(X_1) = -\frac{d\Theta_2}{d\Theta_1} \text{ and } V(X_1) = -\frac{d^2\Theta_2}{d\Theta_1^2}. \tag{3.12}$$

If we choose $-\frac{d\Theta_2}{d\Theta_1}$ as the parameter to be estimated we get the minimum variance attainable is, by (3.5),

$$\left\{\frac{d}{d\theta}\frac{d\Theta_2}{d\Theta_1}\right\}^2 \Big/ \left\{\frac{d^2\Theta_2}{d\Theta_1^2}\frac{d\Theta_1}{d\theta}\right\} = -\frac{d^2\Theta_2}{d\Theta_1^2} = V(X_1). \tag{3.13}$$

Hence X_1 is the best unbiased estimate of $-\frac{d\Theta_2}{d\Theta_1}$. Thus for *the distributions of the type (3.10), there exists a function of the observations which has the maximum precision as an estimate of a function of* θ.

Case of several parameters

Let $\theta_1, \theta_2, \ldots, \theta_q$ be q unknown parameters occurring in the probability density $\phi(x_1, \ldots, x_n; \theta_1, \theta_2, \ldots, \theta_q)$ and $t_1, t_2, \ldots, t_q$ be q functions independent of $\theta_1, \theta_2, \ldots, \theta_q$ such that

$$\int \ldots \int t_i \phi \pi dx_j = \theta_i. \tag{4.1}$$

Differentiating under the integral sign with respect to θ_i and θ_j, we get, if the following integrals exist,

$$\int \ldots \int t_i \frac{\partial \phi}{\partial \theta_i} \pi dx_k = 1, \tag{4.2}$$

and

$$\int \ldots \int t_i \frac{\partial \phi}{\partial \theta_j} \pi dx_k = 0. \tag{4.3}$$

Defining

$$E\left[-\frac{\partial^2 \log \phi}{\partial \theta_i \partial \theta_j}\right] = I_{ij}, \tag{4.4}$$

and

$$E[(t_i - \theta_i)(t_j - \theta_j)] = V_{ij}, \tag{4.5}$$

we get the result that the matrix of the determinant

$$\begin{pmatrix} V_{ii} & 0 & \ldots & 1 & \ldots & 0 \\ 0 & I_{11} & \ldots & I_{1i} & \ldots & I_{1q} \\ \ldots & \ldots & \ldots & \ldots & \ldots & \ldots \\ 1 & I_{i1} & \ldots & I_{ii} & \ldots & I_{iq} \\ \ldots & \ldots & \ldots & \ldots & \ldots & \ldots \\ 0 & I_{q1} & \ldots & I_{qi} & \ldots & I_{qq} \end{pmatrix} \tag{4.6}$$

being the dispersion matrix of the stochastic variates t_i and $\frac{1}{\phi}\frac{\partial \phi}{\partial \theta_j} (j = 1, 2, \ldots, q)$ is positive definite or semi-definite. If we assume that there is no linear relationship of the type

$$\sum \lambda_j \frac{1}{\phi} \frac{\partial \phi}{\partial \theta_j} = 0 \tag{4.7}$$

among the variables $\frac{1}{\phi}\frac{\partial\phi}{\partial\theta_j}(i=1,2,\ldots,q)$ then the matrix $\|I_{ij}\|$, which is known as the information matrix due to $\theta_1,\theta_2,\ldots,\theta_q$, is positive definite in which case there exists a matrix $\|I^{ij}\|$ inverse to $\|I_{ij}\|$. From (4.6) we derive that

$$V_{ii} - I^{ii} \geq 0 \tag{4.8}$$

which shows *that minimum variance attainable for the estimating function of* θ_i *when* $\theta_1,\theta_2,\ldots,\theta_q$ *are not known is* I^{ii}*, the element in the* i*-th row and the* i*-th column of the matrix* $\|I^{ij}\|$ *inverse to the information matrix* $\|I_{ij}\|$.

The equality is attained when

$$t_i - \theta_i = \sum \mu_j \frac{1}{\phi}\frac{\partial\phi}{\partial\theta_j}. \tag{4.9}$$

We can obtain a generalisation of (4.8) by considering the dispersion matrix of $t_1,t_2,\ldots,t_i$ and $\frac{1}{\phi}\frac{\partial\phi}{\partial\theta_r}(r=1,2,\ldots,q)$.

$$\begin{bmatrix} V_{11} & \ldots & V_{1i} & 1 & 0 & \ldots & 0 & \ldots & 0 \\ V_{21} & \ldots & V_{2i} & 0 & 1 & \ldots & 0 & \ldots & 0 \\ \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots \\ V_{i1} & \ldots & V_{ii} & 0 & 0 & \ldots & 1 & \ldots & 0 \\ 1 & \ldots & 0 & I_{11} & I_{12} & \ldots & I_{1i} & \ldots & I_{1q} \\ \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots \\ 0 & \ldots & 0 & I_{q1} & I_{q2} & \ldots & I_{qi} & \ldots & I_{qq} \end{bmatrix} \tag{4.10}$$

This being positive definite or semi-definite we get the result that the determinant

$$|V_{rs} - I^{rs}| \geq 0, \quad (r,s=1,2,\ldots,i) \tag{4.11}$$

for $i=1,2,\ldots,q$. The above inequality is evidently independent of the order of the elements so that, in particular, we get that the determinant

$$\begin{vmatrix} V_{ii} - I^{ii}, & V_{ij} - I^{ij} \\ V_{ji} - I^{ji}, & V_{jj} - I^{jj} \end{vmatrix} \geq 0, \tag{4.12}$$

which gives the result that if $V_{ii} = I^{ii}$, so that maximum precision is attainable for the estimation of θ_i, then $V_{ij} = I^{ij}$ for $(j=1,2,\ldots,q)$.

In the case of the normal distribution

$$\phi(x;m,\sigma) = \text{ const. } \exp -\frac{1}{2}\{\Sigma(x_i-m)^2/\sigma^2\}, \tag{4.13}$$

we have

$$I_{mn} = n/\sigma^2, I_{m\sigma} = 0, I_{\sigma\sigma} = 2n/\sigma^2. \tag{4.14}$$

Since the mean of observations $(x_1+x_2+\ldots+x_n)/n$ is the best unbiased estimate of the parameter m and the maximum precision is attainable *viz.*, $V_{mm} = I^{mm}$, it follows that any unbiased estimate of the parameter σ is uncorrelated with the mean of observations for $V_{m\sigma} = I^{m\sigma} = 0$. *Thus in the case of the univariate normal distribution any function of the observations whose expectation is a function of σ and independent of m is uncorrelated with the mean of the observations.* This can be extended to the ease of multivariate normal populations where any unbiased estimates of the variances and covariances are uncorrelated with the means of the observations for the several variates.

If there exists no functional relationships among the estimating functions $t_1, t_2, \ldots, t_q$ then $\|V^{ij}\|$ the inverse of the matrix $\|V_{ij}\|$ exists in which case we get that the determinant

$$|V^{rs} - I_{rs}|, \quad (r, s = 1, 2, \ldots, i) \tag{4.15}$$

is greater than or equal to zero for $i = 1, 2, \ldots, q$, which is analogous to (4.11).

If a sufficient set of statistics $T_1, T_2, \ldots, T_q$ exist for $\theta_1, \theta_2, \ldots, \theta_q$ then we can show as in the case of a single parameter that the best estimating functions of the parameters or functions of parameters are explicit functions of the sufficient set of statistics.

Koopman (1936) has shown that under some conditions the distribution function $\phi(x_1, x_2, \ldots, x_n; \theta_1, \theta_2, \ldots, \theta_q)$ admitting a set of statistics $T_1, T_2, \ldots, T_q$ sufficient for $\theta_1, \theta_2, \ldots, \theta_q$ can be expressed in the form

$$\phi = \exp(\Theta_1 X_1 + \Theta_2 X_2 + \ldots + \Theta_q X_q + \Theta + X) \tag{4.16}$$

where X's are independent of θ's and Θ's are independent of x's. Making use of the relation

$$\int \phi dv = 1, \tag{4.17}$$

we get

$$\left.\begin{aligned} E(X_i) &= -\tfrac{\partial\Theta}{\partial\Theta_i}, \\ V(X_i) &= -\tfrac{\partial^2\Theta}{\partial\Theta_i^2}, \\ \operatorname{cov}(X_i, X_j) &= -\tfrac{\partial^2\Theta}{\partial\Theta_i\partial\Theta_j}. \end{aligned}\right\} \tag{4.18}$$

This being the maximum precision available we get that for this class of distribution laws there exist functions of observations which are the best possible

estimates of functions of parameters.

Loss of Information

If $t_1, t_2, \ldots, t_q$, the estimates of $\theta_1, \theta_2, \ldots, \theta_q$, have the joint distribution $\Phi(t_1, t_2, \ldots, t_q; \theta_1, \theta_2, \ldots, \theta_q)$ then the information matrix on $\theta_1, \theta_2, \ldots, \theta_q$ due to $t_1, t_2, \ldots, t_q$ is $\|F_{ij}\|$ where

$$F_{ij} = E\left\{-\frac{\partial^2 \log \Phi}{\partial\theta_i \partial\theta_j}\right\}. \tag{5.1}$$

The equality

$$I_{ij} = (I_{ij} - F_{ij}) + F_{ij} \tag{5.2}$$

effects a partition of the covariance between $\frac{1}{\phi}\frac{\partial\phi}{\partial\theta_i}$ and $\frac{1}{\phi}\frac{\partial\phi}{\partial\theta_j}$ as within and between the regions formed by the intersection of the surfaces for constant values of $t_1, t_2, \ldots, t_q$. Hence we get that the matrices

$$\|I_{ij} - F_{ij}\| \text{ and } \|F_{ij}\| \tag{5.3}$$

which may be defined as the dispersion matrices of the quantities $\frac{1}{\phi}\frac{\partial\phi}{\partial\theta_i}(i = 1, 2, \ldots, q)$ within and between the meshes formed by the surfaces of constant values of $t_1, t_2, \ldots, t_q$, is positive definite or semi-definite. This may be considered as a generalisation of Fisher's inequality $I_{ii} \geq F_{ii}$ in the case of a single parameter.

If $I_{ii} = F_{ii}$, then it follows that $I_{ij} = F_{ij}$ for all j for otherwise the determinant

$$\begin{vmatrix} I_{ii} - F_{ii} & I_{ij} - F_{ij} \\ I_{ij} - F_{ij} & I_{jj} - F_{jj} \end{vmatrix} < 0. \tag{5.4}$$

If in the determinant

$$|I_{ij} - Fij|, \quad (i, j = 1, 2, \ldots, q), \tag{5.5}$$

the zero rows and columns are omitted, the resulting determinant will be positive and less than the determinant obtained by omitting the corresponding rows and columns in $|I_{ij}|$. If we represent the resulting determinants by dashes, we may define the loss of information in using the statistics $t_1, t_2, \ldots, t_q$ as

$$|I_{ij} - F_{ij}|'/|I_{ij}|'. \tag{5.6}$$

If Φ is the joint distribution of $t_1, t_2, \ldots, t_q$ the estimates of $\theta_l, \theta_2, \ldots, \theta_q$ with the dispersion matrix $\|V_{ij}\|$ then we have the relations analogous to (4.11) and (4.15)

connecting the elements of $\|V_{ij}\|$ and $\|F_{ij}\|$ defined above. Proceeding as before we get that the determinants

$$|V_{rs} - F^{rs}| \text{ and } |F_{rs} - V^{rs}|, (r, s = 1, 2, \ldots, i), \tag{5.7}$$

are greater than or equal to zero for all $i = 1, 2, \ldots, q$.

The population space

Let the distribution of a certain number of characters in a population be characterised by the probability differential

$$\phi(x, \theta_1, \ldots, \theta_q)dv. \tag{6.1}$$

The quantities $\theta_1, \theta_2, \ldots, \theta_q$ are called population parameters. Given the functional form in x's as in (6.1) which determines the type of the distribution function, we can generate different populations by varying $\theta_1, \theta_2, \ldots, \theta_q$. If these quantities are represented in a space of q dimensions, then a population may be identified by a point in this space which may be defined as the population space (P.S).

Let $\theta_1, \theta_2, \ldots, \theta_q$ and $\theta_1 + d\theta_1, \theta_2 + d\theta_2, \ldots, \theta_q + d\theta_q$ be two contiguous points in (P.S). At any assigned value of the characters of the populations corresponding to these contiguous points, the probability densities differ by

$$d\phi(\theta_1, \theta_2, \ldots, \theta_q) \tag{6.2}$$

retaining only first order differentials. It is a matter of importance to consider the relative discrepancy $d\phi/\phi$ rather than the actual discrepancy. The distribution of this quantity over the x's summarises the consequences of replacing $\theta_1, \theta_2, \ldots, \theta_q$ by $\theta_1 + d\theta_1, \ldots, \theta_q + d\theta_q$. The variance of this distribution or the expectation of the square of this relative discrepancy comes out as the positive definite quadrate differential form

$$ds^2 = \sum\sum g_{ij} d\theta_i d\theta_j, \tag{6.3}$$

where

$$g_{ij} = E\left(\frac{1}{\phi}\frac{\partial\phi}{\partial\theta_i}\right)\left(\frac{1}{\phi}\frac{\partial\phi}{\partial\theta_j}\right). \tag{6.4}$$

Since the quadratic form is invariant for transformations in (P.S) it follows that g_{ij} form the components of a covariant tensor of the second order and is also symmetric for $g_{ij} = g_{ji}$ by definition. This quadratic differential form with its *fundamental tensor* as the elements of the *Information matrix* may be used as a

suitable measure of divergence between two populations defined by two contiguous points. The properties of (P.S) may be studied with this as the *quadratic differential metric* defining the element of length. The space based on such a metric is called the Riemanian space and the geometry associated with this is the Riemanian geometry with its definitions of distances and angles.

The distance between two populations

If two populations are represented by two points A and B in (P.S) then we can find the distance between A and B by integrating along a geodesic using the element of length

$$ds^2 = \sum\sum g_{ij} d\theta_i d\theta_j. \tag{7.1}$$

If the equations to the geodesic are

$$\theta_i = f_i(t), \tag{7.2}$$

where t is a parameter, then the functions f_i are derivable from the set of differential equations

$$\sum_1^q {}_j g_{jk} \frac{d^2\theta_j}{dt^2} \sum_1^q {}_{jl}[jl,k] \frac{d\theta_j}{dt}\frac{d\theta_l}{dt} = 0, \tag{7.3}$$

where $[jl,k]$ is the Christoffel symbol defined by

$$[jl,k] = \frac{1}{2}\left[\frac{\partial g_{jk}}{\partial\theta_l} + \frac{\partial g_{lk}}{\partial\theta_j} + \frac{\partial g_{jl}}{\partial\theta_k}\right]. \tag{7.4}$$

The estimation of distance, however, presents some difficulty. If the two samples from two populations are large then the best estimate of distance can be found by substituting the maximum likelihood estimates of the parameters in the above expression for distance. In the case of small samples we can get the fiducial limits only in a limited number of cases.

We apply the metric (7.1) to find the distance between two normal populations defined by (m_1, σ_1) and (m_2, σ_2) the distribution being of the type

$$\phi(x, m, \sigma) = \frac{1}{\sqrt{(2\pi\sigma^2)}} \exp -\frac{1}{2}\frac{(x-m)^2}{\sigma^2}. \tag{7.5}$$

The quantities g_{ij} defined above have the values

$$g_{11} = 1/\sigma^2, \quad g_{12} = 0, \quad g_{22} = 2/\sigma^2, \tag{7.6}$$

so that the element of length is obtained from

$$ds^2 = \frac{(dm)^2}{\sigma^2} + \frac{2}{\sigma^2}(d\sigma)^2. \tag{7.7}$$

If $m_1 \neq m_2$ and $\sigma_1 \neq \sigma_2$ then the distance comes out as

$$D_{AB} = \sqrt{2}\log\frac{\tan\theta_1/2}{\tan\theta_2/2} \tag{7.8}$$

where

$$\theta_i = \sin^{-1}\sigma_i/\beta \text{ and } \beta^2 = \sigma_1^2 + [(m_1-m_2)^2 + 2(\sigma_2^2-\sigma_1^2)]^2/8(m_1-m_2)^2. \tag{7.9}$$

If $m_1 = m_2$ and $\sigma_1 \neq \sigma_2$

$$D_{AB} = \sqrt{2}\log(\sigma_2/\sigma_1). \tag{7.10}$$

If $m_1 \neq m_2$ and $\sigma_1 = \sigma_2$

$$D_{AB} = \frac{m_1 - m_2}{\sigma}. \tag{7.11}$$

Distance in tests of significance and classification

The necessity for the introduction of a suitable measure of distance between two populations arises when the position of a population with respect to an assigned sets of characteristics of a given population or with respect to a number of populations has to be studied. The first problem leads to tests of significance and the second to the problem of classification. Thus if the assigned values of parameters which define some characteristics in a population are $\bar{\theta}_1, \bar{\theta}_2, \ldots, \bar{\theta}_q$ represented by the point O, and the true values are $\theta_1, \theta_2, \ldots, \theta_q$ represented by the point A, then we can define the divergence from the assigned sets of parameters by D_{A0}, the distance defined before in the (P.S). The testing of the hypothesis

$$\bar{\theta}_i = \theta_i, \quad (i = 1, 2, \ldots, q), \tag{8.1}$$

may be made equivalent to the test for the significance of the estimated distance D_{A0} on the large sample assumption. If $D_{A0} = \psi(\theta_1, \ldots, \theta_q; \bar{\theta}_1, \ldots, \bar{\theta}_q)$ likelihood estimates of $\theta_1, \theta_2, \ldots, \theta_q$ are $\hat{\theta}_1, \hat{\theta}_2, \ldots, \hat{\theta}_q$, then the estimate of D_{A0} is given by

$$\hat{D}_{A0} = \psi(\hat{\theta}_1, \ldots, \hat{\theta}_q; \bar{\theta}_1, \ldots, \bar{\theta}_q). \tag{8.2}$$

The covariance between the maximum likelihood estimates being given by the elements of the information matrix, we can calculate the large sample approximation to the variance of the estimate of D_{A0} by the following formula

$$V(\hat{D}_{A0}) = \sum\sum \frac{\partial\psi}{\partial\theta_i}\frac{\partial\psi}{\partial\theta_j}\,\mathrm{cov}(\hat{\theta}_i, \hat{\theta}_j) \tag{8.3}$$

We can substitute the maximum likelihood estimates of $\theta_1, \theta_2, \ldots, \theta_q$ in the expression for variance. The statistic

$$w = \frac{\hat{D}_{AO}}{[V(\hat{D}_{AO}]^{\frac{1}{2}}}, \tag{8.4}$$

can be used as a normal variate with zero mean and unit variance to test the hypothesis (8.1).

If the hypothesis is that two populations have the same set of parameters then the statisic

$$w = \frac{\hat{D}_{AB}}{[V(\hat{D}_{AB}]^{\frac{1}{2}}}, \tag{8.5}$$

where $\hat{D}_{AB}$ is the estimate of the distance between two populations defined by two points A and B in (P.S) can be used as (8.4). This expression for variance has to be calculated by the usual large sample assumption.

If the sample is small the appropriate test will be to find out a suitable region in the sample space which affords the greatest average power over the surface in the (P.S) defined by constant values of distances. The appropriate methods for this purpose are under consideration and will be dealt with in a future communication.

This estimated distance can also be used in the problem of classification, it usually becomes necessary to know whether a certain population is closer to one of a number of given populations when it is known that populations are all different from one another. In this case the distances among the populations taken two by two settle the question. We take that population whose distance from a given population is significantly the least as the one closest to the given population.

This general concept of distance between two statistical populations (as different from test of significance) was first developed by Prof. P. C. Mahalanobis. The generalised distance defined by him (Mahalanobis, 1936) has become a powerful tool in biological and anthropological research. A perfectly general measure of divergence has been developed by Bhattacharya (1942) who defines the distance

between population as the angular distance between two points representing the population on a unit sphere. If $\pi_1, \pi_2, \ldots, \pi_k$ are the proportions in a population consisting of k classes then the population can be represented by a point with coordinates $\sqrt{\pi_1}, \sqrt{\pi_2}, \ldots, \sqrt{\pi_k}$ on a unit sphere in a space of k dimensions. If two populations have the proportions $\pi_1, \pi_2, \ldots, \pi_k$ and $\pi'_1, \pi'_2, \ldots, \pi'_k$ the points representing them have the co-ordinates $\sqrt{\pi_1}, \sqrt{\pi_2}, \ldots, \sqrt{\pi_k}$ and $\sqrt{\pi'_1}, \sqrt{\pi'_2}, \ldots, \sqrt{\pi'_k}$. The distance between them is given by

$$\cos^{-1}\{\sqrt{(\pi_1\pi'_1)} + \sqrt{(\pi_2\pi'_2)} + \ldots + \sqrt{(\pi_k\pi'_k)}\}. \tag{8.6}$$

If the population are continuous with probability densities $\phi(x)$ and $\psi(x)$ the distance is given by

$$\cos^{-1} \int \sqrt{\{\phi(x)\psi(x)\}}dx. \tag{8.7}$$

The representation of a population as a point on a unit sphere as given by Bhattacharya (1942) throws the quadratic differential metric (7.1) in an interesting light. By changing $\theta_1, \theta_2, \ldots \theta_q$ the parameters occurring in the probability density, the points representing the corresponding populations describes a surface on the unit sphere. It is easy to verify that the element of length ds connecting two points corresponding to $\theta_1, \theta_2, \ldots \theta_q$ and $\theta_1 + d\theta_1, \ldots, \theta_q + d\theta_q$ on this is given by

$$ds^2 = \sum (d\phi)^2/\phi = \sum\sum g_{ij}d\theta_i d\theta_j, \tag{8.8}$$

where g_{ij} are the same as the elements of the quadratic differential metric defined in (7.1).

Further aspects of the problems of distance will be dealt with in an extensive paper to be published shortly.

Statistical Laboratory, Calcutta

References

Aitken, A. C. (1941), On the Estimation of Statistical Parameters. *Proc. Roy. Soc. Edin.*, 61, 56--62.

Bhattacharya, A. (1942), On Discrimination and Divergence. *Proc. Sc. Cong.*

Fisher, R. A. (1921), On the Mathematical Foundation of Theoretical Statistics. *Phil. Trans. Roy. Soc. A*, 222, 309-368.

Koopman, B. O. (1936), On distribution admitting Sufficient Statistics. Trans. Am. Math. Soc., 39, 399--409.

Mahalanobis, P. C. (1936), On the Generalised Distance on Statistics. *Proc. Nal. Inst. Sc. Ind.*,2, 49--55.