THE ROLES OF MAXIMUM-ENTROPY AND MINIMUM-DISCRIMINATION INFORMATION PRINCIPLES IN STATISTICS*

By

J.N. KAPUR

Indian Institute of Technology, Kanpur

INTRODUCTION

I consider it a great honour to have been elected as sessional president of the Indian Society of Agricultural Statatics. I had the previlege of being in the first batch of students trained by the Institute of Agricultural Research Statistics (then Statistical Wing of the ICAR) about thirty-eight years ago. I also recall may close association with the Indian Society of Agricultural Statitics in its first ten years.

Though I have continued my interest in the teaching of statistics throughout the last four decades, both by direct teachning and through my book which has been used by over 200,000 students in India and abroad, my research interests have undergone a full cycle. I started with Statistics and then worked successively in Ballisiics, Fluid Dynamics, Operation Research, Biomathematics, Pattern Recognition and Information Theory.

My current interests are mainly statistical in nature. I am interested in stochastic birth-death-immigration-emigration processes, stochastic models in compartment analysis, statistical measures of entropy and divergence and applications of maximum-entropy principle to pattern recognition, time-series analysis, non-linear spectral estimation, estimation of missing values and non-parametric density estimation.

I would like to use the present occasion to make a strong plea for a greater role for principles of maximum entropy, minimum discrimination information, minimum inter dependence, minimax entropy etc. in the development of statistical theory.

*Technical Address at 38th Annual Conference of ISAS, 1984.

Statistics is concerned with inductive inference and in particular with drawing of inferences about populations from knowledge about samples. The principle of maximum entropy is also concerned with drawing of most unbiased inferences when only partial information is available about a probabilistic system.

In the present address, I shall discuss some of the applications of the principles of maximum entropy and minimum discrimination information in Statistics.

1. The Maximum-Entropy Principle

Suppose we know that a random variable can take only values $x_1, x_2, \ldots, x_N$, but we do not know the probability with which the variate values are taken. The only information we have about the probability distribution is that the sum of the probabilities must be unity, *i.e.*,

$$\sum_{i=1}^{N} p_i = p_1 + p_2 + \ldots + p_N = 1 \tag{1}$$

We have an infinity of probability distribution satisfying (1) and we have to have a principle to be able to choose, in some sense, the 'best' out of these.

Laplace, very early, gave his principle of insufficient reason, that since we have no reason to give a greater chance to one value than to another, let us choose

$$p_1 = p_2 = \ldots = p_N = \frac{1}{N} \tag{2}$$

This distribution may also be regarded as the 'most uniform' or 'most smooth' or 'most unbiased' or 'least committed' distribution we can assign. Any other distribution will be less uniform, will be more biased and will imply conscious and unconscious use of information which we do not possess and have no right to use. This distribution also maximizes Shannon's measure of uncertainty or entropy

$$s = -\sum_{i=1}^{N} p_i \, l_n \, p_i \tag{3}$$

subject to (1) being satisfied. Thus we may regard (2) as the distribution which maximizes the uncertainty subject to use being made of the given information (1). Now suppose some divine power also gives us the information that

$$\sum_{i=1}^{N} p_i \, g_r (x_i) = a_r, \qquad r=1, 2,..., m \tag{4}$$

i.e., it gives us the value of m population moments where $m<(n-1)$. We still have an infinity of choices of probability distributions and we have to make a choice. Again we would like to be as objective and as unbiased as possible. We should like to make use of all the information we have and scrupulously avoid making use of any information that we do not have. We should like to use the whole truth and use nothing else but the truth. According the principle of maximum-entropy, we choose the probability distribution which maximizes (3) subject to (1) and (4). Using Lagrange's method, this gives

$$p_i = \exp \left[-\lambda_o - \lambda_1 \, g_1 (x_i) - ... - \lambda_m \, g_m (x_i)\right], \tag{5}$$

were using (1) and (4)

$$\exp \lambda_o = \sum_{i=1}^{N} \exp \left[-\lambda_1 \, g_1 (x_i) - \lambda_2 \, g_2 (x_i) - ... - \lambda_m \, g_m (x_i)\right] \tag{6}$$

$$a_r \exp \lambda_o = \sum_{i=1}^{N} g_r (x_i) \exp \left[-\lambda_1 \, g_1 (x_i) - \lambda_2 \, g_2 (x_i) - ... - \lambda_m \, g_m (x_i)\right] \quad r=1, 2,..., m \tag{7}$$

From (6) we can determine λ_o as a function of $\lambda_1, \lambda_2, \cdots, \lambda_m$ and from (7) we can determine $\lambda_1, \lambda_2, \cdots, \lambda_m$ as functions of $a_1, a_2, \cdots, a_m$. Instead of (7) we can use

$$a_r = \frac{\sum_{i=1}^{N} g_r (x_i) \exp \left[-\lambda_1 \, g_1 (x_i) - ... - \lambda_m \, g_m (x_i)\right]}{\sum_{i=1}^{N} \exp \left[-\lambda_1 \, g_1 (x_i) - ... - \lambda_m \, g_m (x_i)\right]} \tag{8}$$

$$r=1, 2,..., m$$

Thus the maximum-entropy probability distribution is known if the functionss $g, (x)$ and the expected values a_1 are known for

$$r=1, 2, \ldots, m$$

2. MAXIMUM-LIKELIHOOD ESTIMATORS FOR a's

Suppose the divine power gives us only the function g's, and not the values of a's, so that we get a probability density function with unknown parameters $a_1, a_2, \ldots, a_m$.

We draw a random sample of size n in which x_1 may occur k_1 times, x_2 may occur k_2 times, ... and x_N may occur k_N times so

$$k_1+k_2+\ldots+k_N=n \tag{9}$$

Here, of course, some of the k's can be zero. To obtain estimates for a's, we use Fisher's method of maximum likelihood. The likelihood function is

$$L\equiv \exp\,[-n\lambda_o, -n\,\lambda_1\,\bar{g}_1-n\lambda_2\,\bar{g}_2-\,\ldots\,-n\,\lambda_m\,\bar{g}_m] \tag{10}$$

where

$$\bar{g}_r = \frac{\sum_{j=1}^{N} k_j\, g_r\,(x_j)}{\sum_{j=1}^{N} k_j} = \frac{\sum_{j=1}^{N} k_j\, g_r\,(x_j)}{n}, r=1,2,\ldots,m \tag{11}$$

are the sample means of the given functions $g_1(x), g_2(x), \ldots, g_m(x)$. Differentiating (1) logarithmically, we get

$$-\frac{1}{n}\frac{\partial}{\partial a_r}\,(lnL) = \sum_{j=1}^{m}\frac{\partial\lambda_o}{\partial\lambda_j}\frac{\partial\lambda_j}{\partial a_r}+\sum_{i=1}^{m}\frac{\partial\lambda_j}{\partial a_r}\;\bar{g}_j \tag{12}$$

From (6) and (7)

$$\exp\lambda_o\,\frac{\partial\lambda_o}{\partial\lambda_j}=\sum_{i=1}^{M}-g_j\,(x_i)$$

$$\exp\left[-\sum_{k=1}^{m}\lambda_k\, g_k\,(x_i)\right]=-a_j\,\exp\lambda_o \tag{13}$$

From (12) and (13)

$$-\frac{1}{n}\frac{\partial}{\partial a_r}(lnL)=\sum_{j=1}^{m}\frac{\partial\lambda_j}{\lambda a_r}(\bar{g}_j-a_j), \qquad r=1,2,\ldots,m \tag{14}$$

Differentiating again, we get

$$-\frac{1}{n}\frac{\partial^2}{\partial a_r\partial a_s}(lnL)=\sum_{j=1}^{m}\frac{\partial^2\lambda_j}{\partial a_r\partial a_s}(\bar{g}_j-a_j)-\frac{\partial\lambda_j}{\partial a_r} \tag{15}$$

If the determinant $|\partial\lambda_j/\partial a_r|$ is not zero, all the first order partial derivatives of lnL will vanish if

$$a_1=\bar{g}_1,\ a_2=\bar{g}_2,\ \ldots,\ a_m=\bar{g}_m \tag{16}$$

and when this condition is satisfied, the Hessian matrix of the second order partial derivatives of lnL is given by the matrix $n[\partial\lambda_j/\partial a_r]$.

Now,

$$\left[\frac{\partial\lambda_j}{\partial a_r}\right]\left[\frac{\partial a_r}{\partial\lambda_j}\right]=I_m \tag{17}$$

where I_m is the unit mxm matrix. Also, from (13)

$$\frac{\partial a_r}{\partial\lambda_j}=-\frac{\partial^2\lambda_o}{\partial\lambda_r\partial\lambda_j} \tag{18}$$

and

$$\exp\frac{\partial^2\lambda}{\partial\lambda_r\partial\lambda_j}+\exp\lambda_o\frac{\partial\lambda_o}{\partial\lambda_r}\frac{\partial\lambda_o}{\partial\lambda_j}=\sum_{i=1}^{N}g_j(x_i)\,g_r(x_i)$$

$$\times\exp\left[-\sum_{k=1}^{m}\lambda_k\,g_k(x_i)\right] \tag{19}$$

so that

$$\frac{\partial^2\lambda_o}{\partial\lambda_r\partial\lambda_j}+\frac{\partial\lambda_o}{\partial\lambda_r}\frac{\partial\lambda_o}{\partial\lambda_j}=E\left[g_j(x)\,g_r(x)\right] \tag{20}$$

or

$$\frac{\partial^2\lambda_o}{\partial\lambda_r\partial\lambda_j}=\left[g_j(x)\ g_r(x)\right]-E\left[g_j(x)\right]E\left[g_r(x)\right]$$

$$=\text{cov}\left[g_j(x)\ g_r(x)\right] \tag{21}$$

From (18) and (21)

$$\frac{\partial a_r}{\partial\lambda_j}=-\text{cov}\left[g_j(x)\ g_r(x)\right] \tag{22}$$

so that the matrix $[\partial g_r/\partial\lambda_j]$ is the negative of the variance-covariance matrix Z given by

$$Z=\begin{bmatrix} \text{var}(g_1) & \text{cov}(g_1,g_2) & \ldots\ldots & \text{cov}(g_1,g_m) \\ \text{cov}(g_2,g_1) & \text{var}(g_2) & \ldots\ldots & \text{cov}(g_2,g_m) \\ \ldots & \ldots & \ldots\ldots & \ldots\ldots \\ \text{cov}(g_m,g_1) & \text{cov}(g_m,g_2) & \ldots\ldots & \text{var}(g_m) \end{bmatrix} \tag{23}$$

This matrix is positive definite unless the constrains are linearly dependent, i.e., unless the set of functions

$$[1, g_1(x), g_2(x), \ldots, g_m(x)] \tag{24}$$

is a linearly dependent set. We assume that this is not the case, i.e., we deal with only linearly independent constraints. In this case the matrix Z is positive definite so that Z^{-1} is also positive definite and $-Z^{-1}$ is negative definite. Thus from (17), the matrix $[\partial\lambda_j/\partial a_r]$ is also negative definite, but this is the Hessian matrix of second order partial derivatives of lnL at the points where the first order partial derivatvies all vanish. Thus we establish that lnL is maximum when $a_1, a_2, \ldots, a_m$ are given by $\bar{g_1}, \bar{g_2}, \ldots, \bar{g_m}$.

Thus the problem of estimation of probability distribution is reduced to the following steps :

(*i*) Specify functions $g_1(x), g_2(x), \ldots, g_m(x)$;

(*ii*) Based on a random sample of size n, find $\bar{g_1}, \bar{g_2}, \ldots, \bar{g_m}$.

(*iii*) Find the probabilities by using

$$p_i=\frac{e^{-\lambda_1 g_1(x_i)-\lambda_2 g_2(x_i)-\ldots-\lambda_m g_m(x_i)}}{\sum_{i=1}^{N} e^{-\lambda_1 g_1(x_i)-\lambda_2 g_2(x_i)-\ldots-\lambda_m g_m(x_i)}} \tag{25}$$

(*iv*) Find $\lambda_1, \lambda_2, \ldots, \lambda_m$ in terms of $a_1, a, \ldots, a_m$ from (8).

(*v*) Replace $a_1, a_2, \ldots, a_m$ by $g_1, g_2, \ldots, g_m$.

II. MINIMUM CROSS ENTROPY (INACCURACY) AND MAXIMUM LIKELIHOOD

If we have reasons to believe, on the basis of institution and experience, that the probability distribution before the moments are prescribed, is given by $q_1, q_2, \ldots, q_N$ rather than by the uniform distribution, then we choose $p_1, p_2, \ldots, p_N$ in such a way that this distribution is as 'close' to $q_1, q_2, \ldots, q_N$ at possible and at the same time satisfies the given constraints. For this purpose, we minimize Kullback's information discrimination function

$$\sum_{i=1}^{N} p_i \ ln \ \frac{p_i}{q_i} \tag{26}$$

subject to the given constraints. The equations (5), (6), (7) and (10) are modified to

$$p_i=q_i \exp\left[-\lambda_0-\lambda_1 g_1(x_i)-\lambda_2 g_2(x_i)-\ldots-\lambda_m g_m(x_i)\right] \tag{27}$$

$$\exp \lambda_0=\sum_{i=1}^{N} g_i \exp\left[-\lambda_1 g_1(x_i)-\ldots-\lambda_m g_m(x_i)\right] \tag{28}$$

$$\exp \lambda_0 \, a_r=\sum_{i=1}^{N} q_i \, g_r(x_i) \exp\left[-\lambda_1 g_1(x_i)-\ldots-\lambda_m g_m(x_i)\right] \tag{29}$$

$$L=(q_1, q_2, \ldots, q_N) \exp\left(-N\lambda_0-N\lambda \bar{g}_1-\ldots-N\lambda_m \bar{g}_m\right] \tag{30}$$

The values (16) still give the maximum likelihood estimates for the parameters.

4. COMPARISON WITH FISHER'S THEORY OF ESTIMATION

Given a set of observations $x_1, x_2 ... x_N$; Fisher regards these as a random sample from a population and the aim of his theory is to get, from the sample, as much information about the population as possible. His three steps are :

(i) *Specification* : i.e., specify the density function of the population, say $f(x, a_1, a_2, ... a_m)$. This can be done on the basis of intuition and experience.

(ii) *Estimation*: Here the parameters $a_1, a, ... a_m$ have to be estimated as functions of the observed value $x_1, x_2 ..., x_N$. Fisher laid down the criteria of consistency, efficiency and sufficiency and gave the method of maximum likelihood which gives estimates that, in general, satisfy these criteria.

(iii) *Distribution :* Here the distributions of the estimates in random samples as obtained in order to determine how good the estimators are :

The critical difference between Fishher's Method of Estimation (FM) and the Maximum-Entropy Method (MEM) of estimation is in first step. Whereas Fisher's method proceeds by specifying the density function, MEM starts by specify certain moments corresponding to the functions $g_1, g_2, ..., g_m$.

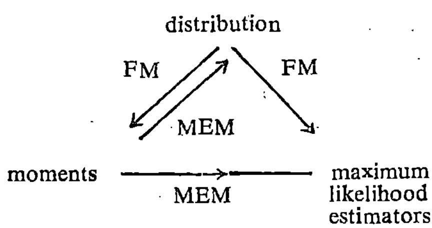
Since we can have a large number of density functions with the same moments, we use the MEP or MIP to get a unique most unbiased distribution with these moments. Thus, while FM specifies f directly, MEM specifies $g_1, g_2 ... g_m$ and then uses MEP to determine f.

In both methods, the population values of the parameters need not be given, but can be estimated in terms of sample values by using the method of maximum likelihood. The estimation is easier in the MEM since here the maximum likelihood estimators for $a_1, a_2 ... a_m$ are $g_1, g_2, ... g_m$ and can be obtained at once. In FM, for every density function, we shall have to obtain estimates for $a_1, a_2, ... a_m$ by solving equation $\partial/\partial a_r (lnL)=0$, $r=1, 2, ..., m$ de novo in every case.

There is no objective method for specifying either f or $g_1, g_2, \ldots, g_m$. It may even be argued in favour of *FM* that specifying the function f may be easier than specifying m functions $g_1, g_2, \ldots g_m$. That this is not so is seen by considering that f determines $g_1, g_2, \ldots, g_m$. uniquely, but $g_1, g_2, \ldots, g_m$ do not determine f uniquely without the use of the maximum entropy principle Thus, in some sense f contains more information than *g's* and its specification should require greater divine assistance than specifications of g's.

Actually specification of f implies the specification of all infinity of moments while specification of *g's* requires the knowledge of only a finite number of moments. In most cases $m=1$ or 2.

In many practical problems knowledge of f implies the knowledge of microscopic structure of a population, while knowledge of *g's* implies only a knowledge of some macroscopic observable quntities. The moments can be interpreted in terms of some measurable entities. Thus, in thermodynamics, these may stand for average energy or temperature or pressure; in social sciences these may stand for budget or number of jobs, or number of hours, etc. In fact, in these cases specifying moments is realistic while specifying f is much more difficult to interpret.



The above figure illustrates the relation between the two methods. In *FM* we go from the distribution to the moments and the maximum likelihood estimators. In *MEM* we go from the moments to distribution and to maximum likelihood estimators.

In almost all cases, the choice of *g's* is confined to the functions

$$x, x^2, x^n, \ln x, (\ln x)^2, \ln(1+x), \ln(1+x^2), \quad |x-m| \tag{31}$$

For specifying f, the choice is much larger. Even the catalogue of standard distributions is large and one has to be sufficiently familiar with all these distributions in order to be able to make an intelligent specification in *FM*.

5. DERIVATION OF STANDARD DISTRIBUTIONS BY USING MAXIMUM-ENTROPY PRINCIPLE

One way of preparing a catalogue of standard distributions is to find all the maximum-entropy distributions which can be obtained when expected values of one or two of the function given in (31) are prescribed. The *ME* distribution will also depend on the range of values permitted for x, e.g., on whether x takes on a finite and discrete set of values or x can take all values in a finite interval (a,b) or in a semi-infinite interval $(0, \infty)$ in the infinite interval $(-\infty, \infty)$.

The distribution will also depend on the a priori probability density function that may be specified.

Multi-variate distributions may be obtained either by :

(*i*) specifying covariances between pairs of variates, or by

(*ii*) specifying expected values like $E(x_1+x_2+\ldots+x_k)$, or by

(*iii*) specifying a relation among the variates, e.g., by specifying $x_1+x_2+\ldots+x_k=1$, or by

(*iv*) specifying an order relation among the variates, e.g., by specifying $x_1\leqslant x_2\leqslant x_3\leqslant\ldots\leqslant x_k$.

and then by applying the *MEP* or *MIP*.

For the discrete case if the a priori probability distribution is given by $q_1, q_2,\ldots,q_m$ and the constraints are given by

$$\sum_{i=1}^{N} p_i=1, \quad \sum_{i=1}^{N} p_i\, g_r(x_i)=a_r, \quad r=1,2,\ldots,m \tag{32}$$

then the *ME* or *MI* distribution is given by

$$p_i=q_i \exp\left[-\lambda_0-\lambda_1\, g_r(x_i)-\ldots-\lambda_m\, g_m(x_i)\right] \tag{33}$$

where

$$\exp \lambda_0 = \sum_{i=1}^{N} \exp [-\lambda_1 g_1 (x_i) - \ldots - \lambda_m g_m (x_i)] \qquad (34)$$

$$a_r \exp \lambda_1 = \sum_{i=1}^{N} g_r (x_i) \exp [-\lambda_1 g_1 (x_i) - \ldots - \lambda_m g_m (x_i)] \qquad (35)$$

$$r = 1, 2, \ldots, m$$

For the continuous case, if the a priori probability density function is $f_0(x)$, then

$$f(x) = f_o (x) \exp [-\lambda_0 - \lambda_1 g_1 (x) - \ldots - \lambda_m g_m (x)] \qquad (36)$$

$$\exp \lambda_0 = \int_a^b f_0 (x) \exp [-\lambda_1 g_1 (x) - \ldots - \lambda_m g_m (x)] \, dx \qquad (37)$$

$$a_r \exp \lambda_0 = \int_a^b f_0(x) \, g_r(x) \exp[-\lambda_1 g_1(x) - \ldots - \lambda_m g_m(x)] \, dx \qquad (38)$$

We now give some distributions obtained by using these results.

6. Maximum-Entropy Discrete-Variate Probability Distributions

Range of Variate	*Sepecified Moments*	*Prior Distribution* q_i	*ME/MI Distribution* p_i	*Name*
1, 2, 3,...,n	—	—	$\frac{1}{n}$	uniform
1, 2, 3,...,n	mean m	uniform	ab^i	geometric
0,1, 2,3,...,n	mean	$\binom{n}{i}$	$\left[\begin{matrix} n \\ i \end{matrix}\right] p^i q^{n-i}$	binomial
1, 2, 3,...,n	mean m	improper uniform	ab^i	geometric
0, 1, 2, 3,...	mean m	$(i!)^{-1}$	$\frac{e^{-m} m^i}{i!}$	Poisson
1, 2, 3,...	mean m	i^{-1}	$-\frac{1}{ln(1-q)} \frac{q^i}{i}$	Log Series
1,2,3,...	mean m	i^{-d}	$\frac{\Sigma q^i / i^d}{\sum_{i=1}^{\infty} \frac{q^i}{i^d}}$	generalized geometric

7. Maximum-Entropy Continuous-Variate Probability Distributions

(a) Range $(-\infty, \infty)$

Specified Moments	*Distribution*
$E(x)$	Does not exist
$E(x^2)=\sigma^2$	$N(0, \sigma^2)$
$E(x-m)^2=\sigma^2$	$N(m, \sigma^2)$
$E(x)=m$, $E(x-m)^2=\sigma^2$	$N(m, \sigma^2)$
$E(x)=m$, $E(x^2)=\sigma_0^2$	$N(m, \sigma_0^2-m^2)$
$E(x-\bar{x})^2=\sigma^2$	$N(m, \sigma^2)$ (m arbitrary)
$E(x^r)=a_r$	Does not exist if k is odd
$r=1,2,\ldots,k$	$f(x)=\exp[-\lambda_0-\lambda_1 x-\ldots-\lambda_k x^k]$, if k is even
$E(\mid x \mid)=\sigma$	Laplace
$E(\mid x-m \mid)=\sigma$	Laplace with mean m
$E(x)=m$, $E\mid x-m \mid=\sigma$	Laplace with mean m
$E\ln(1+x^2)$	Generalized Cauchy

(b) Range $[0,\infty]$

$E(x)$	exponential
$E(x)$, $E(\ln x)$	gamma
$E(x)$, $E[\ln(1+x)]$	beta
$E(\ln x)$, $E(\ln x)^2$	log normal
$E\ln(1+x^2)$	utilateral generalized Cauchy
$E(x)$, $E(x^2)$	truncated normal if $\sigma^2<m^2$ exponential if $\sigma^2=m^2$ does not exist if $\sigma^2>m^2$

(*c*) **Range (0, 1)**

Specified Moments	*Distribution*
nil	uniform
mean	truncated exponential
$E(x)$, $E(x^2)$	truncated normal or truncated u or uniform depending on prescribed values
$E(\ln x)$, $E[\ln(1-x)]$	beta

8. MAXIMUM-ENTROPY MULTIVARIATE DISTRIBUTIONS

8.1 Discrete Variate Distributions

If the variates take integral values 0, 1, 2, 3,..., if the mean of each variate is prescribed; and the prior probability distribution is given by :

$$\frac{(r_1+r_2+\ldots r_n)\,!}{r_1\,!\;r_2\,!\ldots r_n\,!}$$

then the maximum-entropy probability density function is given by :

$$p(r_1,r_2,\ldots,r_n)= A\;\frac{(r_1+r_2+\ldots r_n)\,!}{r_1\,!\;r_2\,!\ldots r_n\,!}\,q_1^{r_1}\;q_2^{r_2}\;\ldots\;q_n^{r_n} \tag{39}$$

where

A is a normalizing constant, and

q's are to be determined in terms of the prescribed means. We get the following special cases :

(*i*) If $r_1+r_2+\ldots+r_n=N$, we get the multinomial distribution ;
(*ii*) If $r_1, r_2,\ldots, r_n$ take all non-negative integral values, we get :

$$p(r_1,r_2,\ldots,r_n)=(1-q_1-q_2-\ldots-q_n)\frac{r_1+r_2+\ldots+r_n)\,!}{r_1\,!\;r_2\,!\ldots r_n\,!}$$

$$\times\; q_1^{r_1}\;q_2^{r_2}\;\ldots\;q_n^{r_n}$$

$$r_i\geqslant 0;\; i=1,\ldots,m \quad q_i<1;\;\; q_1+q_2+\ldots+q_n<1 \tag{40}$$

This gives the multivariate geometric distribution ;

(*iii*) If $r_1, r_2, \ldots, r_n$ takes all non-negative integral values except that $r_1 = r_2 = \ldots = r_n = 1$ is not allowed, we get :

$$p(r_1, r_2, \ldots, r_n) = \frac{1 - q_1 - \ldots - q_n}{q_1 + q_2 + \ldots + q_n} \frac{(r_1 + r_2 + \ldots + r_n)!}{r_1!\, r_2! \ldots r_n!}$$

$$\times\, q_1^{r_1}\, q_2^{r_2} \ldots q_n^{r_n} \qquad (41)$$

(*iv*) If $r_1, r_2, \ldots r_n$ takes all positive integral values, we get :

$$p(r_1, r_2, \ldots, r_n) = \frac{1 - q_1 - \ldots - q_n}{q_1 + q_2 + \ldots + q_n} \frac{(r_1 + r_2 \ldots + r_n)!}{r_1!\, r_2! \ldots r_n!}$$

$$\times\, q_1^{r_1}\, q_2^{r_2} \ldots q_n^{n} \qquad (42)$$

If in addition to prescribing the arithmetic mean of each variate, we also prescribe $E\,[ln(r_1 + r_2 + \ldots + r_n)]$, then the maximum-entropy density function is :

$$p(r_1, r_2, \ldots, r_n) = \frac{1}{\varphi(q, d)} \frac{(r_1 + r_2 + \ldots + r_n)!}{r_1!\, r_2! \ldots r_n!}$$

$$\times\, q_1^{r_1}\, q_2^{r_2} \ldots q_n^{r_n}\, (r_1 + r_2 + \ldots + r_n)^d \qquad (43)$$

where

$q = q_1 + q_2 + \ldots q_n$, and

$$\phi(q, d) = \sum_{i=1}^{\infty} q^k\, k^d \; ; \qquad r_i \geqslant 0; \text{ all } r\text{'s not zero}$$

If $d = 0$, this gives the multivariate geometric distribution.

If $d = -1$, this gives the multivariate log series distribution :

$$p(r_1, r_2, \ldots, r_n) = \frac{1}{\ell n(1 - q_1 - q_2 - \ldots - q_n)}$$

$$\times \frac{(r_1 + r_2 + \ldots + r_n^{-1})!}{r_1!\, r_2! \ldots r_n!}\, q_1^{r_1}\, q_2^{r_2} \ldots q_n^{r_n} \qquad (44)$$

If we take other values of d, we get the family of multivariate generalized gemoetric distributions.

If we take the prior as :

$$\frac{(r_1+r_2\cdots+r_n+M-1)!}{r_1!\ r_2!\ \dots\ r_n!}$$

and prescribe the means only, then we get the multivariate negative binomial distribution :

$$p(r_1,\ r_2,\cdots,r_n)=\frac{Q^{-M}}{\tau(A)}\ \frac{\Gamma(M+r_1+_2+\dots+r_n)}{r_1!\ r_2!\ \dots\ r_n!}$$

$$\times\left[\frac{P_1}{Q}\right]^{r_1}\dots\ \left[\frac{P_n}{Q}\right]^{r_n} \tag{45}$$

Similarly, we can obtain the multivariate generalized negative binomial distribution :

$$p(r_1,r_2,\dots,r_n)=C\ \frac{\Gamma[M+\beta(r_1+r_2+\dots+r_n)]}{\Gamma[M+(\beta-1)(r_1+r_2+\dots+r_n)]}$$

$$\times\frac{q_1^{r_1}\ q_2^{r_2}\ \dots\ q_n^{r_n}}{r_1!\ r_2!\dots r_n!} \tag{46}$$

where

$$C=(1+z)^{-M}$$

where

$$\frac{z}{(1+z)\beta}=q=q_1+q_2+\dots q_n \tag{47}$$

If we take the prior as:

$$\frac{r_1+r_2+\dots+r_n}{r_1!\ r_2!\dots r_n!}$$

and the means are prescribed, we get the multivariate Poisson distribution :

$$p(r_1,r_2,\dots,r_n)=\frac{\bar{e}(q_1+q_2+\dots+q_n)}{q_1+q_2+\dots+q_n}\ \frac{r_1+r_2+\dots+r_n}{r_1!\ r_2!\dots r_n!}$$

$$\times q_1^{r1}\ q_2^{r2}\ \dots\ q_n^{rn} \tag{48}$$

8.2 Coatinuous-Variate Multivariate Distributions

(1) If the range of each variate is $(-\infty, \infty)$, and if the means, variances and covariances are all prescribed. the maximum-entropy distribution is the multivariate mormal distribution.

(2) If the range of each variate is $(0,\infty)$, and if $E(\ln x_i)$ $E(\ln x_i)^2$, and $\text{cov}\,(\ln x_i, \ln x_j)$ are all prescribed, the maximum-entropy distribution is the multivariate log normal distribution.

(3) If $E(\ln x_1)$, $E(\ln x_2)$, ..., $E(\ln x_{n-1})$, $E(\ln(1-x_1-x_2-\dots-x_{n-1}))$ are prescribed, and each $x_i \geqslant 0$, and $x_1+x_2+\dots+x_{n-1} \leqslant 1$, the maximum-entropy distribution is the Districhlet distribution.

(4) If $E(\ln x_1)$, $E(\ln x_2)$,..., $(E\ln x_{n-1})$, $E(\ln(1+x_1+\dots+x_{n-1}))$ are prescribed and all $x_i \geqslant 0$, the maximum entropy distribution is the multivariate beta distribution of the second kind,

(5) If in (4) $x_i = e^{-z_i}$, we get a generalized multivariale logistic distribution of which the ordinary logistic distribution is obtained as a particular case.

(6) If $E(\ln(1+x_1^2+x_2^2+\dots x_n^2))$ is prescribed, we get a generalized multivariate Cauchy distribution of which the ordinary multivariate Cauchy distribution is a special case.

(7) If the only information about the variates is that $x_i \geqslant 0$ and $x_1+x_2+\dots+x_n=1$, then the maximum entropy density for the ith variate is $(n-1)\,(1-x_i)^{n-2}$ and the joint density for two variates is $(n-1)\,(n-2)\,(1-x_i-x_j)^{n-3}$.

(8) If, in addition, the means of the variates are also prescribed, the maximum entropy density for each variate is the sum of exponential functions.

(9) If $E[f(x)]$ is prescribed for each variate and, in addition, we are given that $x_1 \leqslant x_2 \leqslant x_3 \leqslant \dots \leqslant x_n$, we can find the distribution of ordered statistics. In the usual discussion,

all the unordered variates are supposed to be independently and identically distributed. In our case, they need not be identically distributed.

(10) In general, to get a multivariate distribution using the maximum entropy principle, we have to prescribe $E(x_1+x_2+\ldots+x_n)$ or prescribe some expected values of some functions of $x_1, x_2, \ldots, x_n$. In addition, we have to prescribe the expected values of functions of x_i separately.

The properties of most of the univariate and multivariate distributions obtained here are available in Johnson and Kotz [19–21], Consul and Jain [5], Consul and Shenton [6] Jain and Consul [16], Patil and Joshi [43] and Patil, Kapadia and Bowen [44].

9. Entropy-Concentration Theorem

Let $P_0=(p_{10}, p_{20}, \ldots, p_{n0})$ be the maximum-entropy probability distribution and let $P=(p_1, p_2, \ldots, p_n)$ be any other probability distribution consistent with the given constraints. Let S_{max} and S be their respective entropies and let

$$\triangle S=S_{max}-S \qquad \ldots(49)$$

Let C be the class of all probability distributions consistent with the constraints, then in this class, P_0 has a favoured status. It is most unbiased since it does not make use of any other information than what is given by the constraints. The distribution P can be obtained only by using some additional information, consciously or unconsciously. P_0 is also as near to the uniform distribution as possible since it minimizes the directed divergence between P and the uniform distribution $(1/n, 1/n, \ldots, 1/n)$.

The following questions naturally arise :

(1) Can we measure the degree of bias of P? Can we use $\triangle S$ as a measure of bias ? Which will be best; $\triangle S$, $\triangle S/S_{max}$, $\triangle S/\ln n$, and why ?

(2) What proportion of probability distributions in C have their entropies greater than $.95\ S_{max}$ or $>.9\ S_{max}$ or $>.5\ S_{max}$? Will this proportion depend on the nature of the constraints or on their numbers only ?

(3) If we consider n-dimensional space with coordinates $(p_1, p_2, \ldots, p_n)$, then the set of points corresponding to the class C from a closed convex set (why?). Can we consider $\triangle S$ or

$$\sum_{i=1}^{n} p_i \quad ln \ \frac{p_i}{p_{i0}} \qquad \ldots(50)$$

as the distance of any point in it from the point corresponding to the maximum-entropy distribution P_0 ? Can we say that P_1 is more biased than P_2 if $\triangle S_1 > \triangle S_2$?

(4) Can we find the additional constraint or additional information presumed which can make P a maximum-entropy distribution ? Can we at least find the measure of information contained in this constraint ?

Recently Jaynes [18] gave the following entropy-concentration theorem as a step towards answering these questions:

"In N random trials, $2N\triangle S$ is asymptotically distributed as chi-square with $k=n-m-1$ degrees of freedom."

Thus, we get :

$$P\left[\ S_{max} - \frac{\chi_k^2\,(0.5)}{2N} \leqslant S \leqslant S_{max}\right] = .95 \qquad \ldots(51)$$

$$P\left[\ S_{max} - \frac{\chi_k^2\,(.01)}{2N} \leqslant S \leqslant S_{max}\right] = .99 \qquad \ldots(52)$$

so that there is an 'entropy fiducial interval' of thength $\chi_k^2\ (P)/2N$ with 'confidence coefficient' $1-P$. The length of this entropy interval :

(1) decreases fast with N, in fact, it decreases inversely as N;

(2) increases with confidence level;

(3) increases with n;

(4) decreases wih m.

The probability distribution P in C for which the entropy S lies outside the 99% entropy interval is not likely to arise. In fact, a more correct statement would be that value of P strongly suggests the existence of an additional constraint on the system and urges us to search for it.

Thus, for an unbiased die, $S_{max}=ln6=1.792$, $k=5$, $M=1000$, χ^2_5 (.05)=11.07, χ^2_5 (.005)=16.75 so that 95% entropy interval is (1.786, 1.792) and 99.% entropy interval is (1.783, 1.792) so that if the entropy of the observed distribution is less than 1.783, we can rule out the possibility of the die being unbiased.

We can now introduce another constraint that the mean is prescribed. We throw the die a large number of times and observe the mean number of points. Suppose it is 4.5, It can be shown that in this case, S_{max} is 1.614, $k=4$, χ^2_4 (0.05)=9.49 and the 95% confidence entropy interval is (1.609, 1.614). If the entropy of the observed distribution is less than 1.609, it indicates the existence of another constraint or it may suggest that a constraint prescribing a moment other than the mean may be operative and we may look for it.

We may note that Jaynes' theorem is asymptotically valid, *i.e.* valid for large values of N only.

For smaller values of N, it may sometimes be possible to do complete enumeration. Thus for 20 throws of a coin, the $2^{20}=10^6$ possibilities are distributed as follows:

# of heads	0/20	1/19	2/18	3/17	4/16	5/15	6/14	7/13
# of states:	1	20	190	1140	4845	15504	38760	77520
# of heads:	8/12	9/11	10/10					
# of states:	125970	167960	184756					

Thus, the number of ways is maximum for 10 heads and 10 states, and this the most likely state to occur. In fact 9 and 11 heads have also a large number of ways associated with them and these states together account for more than 50% of the total number of ways.

Jaynes' proof is based on the concepts of n-dimensional space and is an adaptation of Pearson's proof of the chi-square distribution. An anatytical proof is given in the next section which shows that if $\triangle I = I - I_{min}$, then $2N \triangle I$ is also distributed as chi-square with k d.f. The proof can be easily adapted to measures of entropy other than Shannon's, provided these are concave functions.

10. MAXIMUM ENTROPY, MINIMUM INFORMATION, MAXIMUM LIKELIHOOD AND MINIMUM CHI-SQUARE

$$\triangle S = S_{max} - S = -\sum_{i=1}^{n} p_{i0} \ ln\ p_{i0} \sum_{i=1}^{n} p_i \ ln\ p_i$$

$$= \sum_{i=1}^{n} p_i \ ln \frac{p_i}{p_{i0}} + \sum_{i=1}^{n} (p_i - p_{i0}) ln\ p_{i0}$$

$$= \sum_{i=1}^{n} p_i \ ln \frac{p_i}{p_{i0}} + \sum_{i=1}^{n} (p_i - p_{i0})$$

$$[-\lambda_0 - \lambda_1 g_1 (x_i) - \ldots - \lambda_m g_m (x_i)]$$

$$= \sum_{i=1}^{n} p_i \ ln \frac{p_i}{p_{i0}} = -\sum_{i=1}^{n} p_i \ ln \frac{p_{i0}}{p_i}$$

$$= -\sum_{i=1}^{n} p_i \ ln \left[1 + \frac{p_{i0} - p_i}{p_i} \right]$$

$$= -\sum_{i=1}^{n} p_i \left[\frac{p_{i0} - p_i}{p_i} - \frac{(p_{i0} + p_i)^2}{2p_i^2} + \frac{(p_{i0} - p_i)^3}{3p_i^3} - \ldots \right]$$

$$= \frac{1}{2} \sum_{i=1}^{n} \frac{(p_{i0} - p_i)^2}{p_i} - \frac{1}{3} \sum_{i=1}^{n} \frac{(p_{i0} - p_i)^3}{p_i^2} + \ldots$$

$$= \frac{1}{2} \sum_{i=1}^{n} \frac{(p_{i0} - p_i)^2}{p_{i0}} \left[1 + \frac{p_{i0} - p_i}{p_i} \right]$$

$$- \frac{1}{3} \sum_{i=1}^{n} \frac{(p_{i0} - p_i)^3}{p_{i0}} \left[1 + \frac{p_{i0} - p_i}{p_i} \right]^2 + \ldots$$

$$= \frac{1}{2} \sum_{i=1}^{n} \frac{(p_{i0} - p_i)^2}{p_{i0}} + \frac{1}{6} \sum_{i=1}^{n} \frac{(p_{i0} - p_i)^3}{p_{i0}^2} - \ldots \quad (53)$$

Similarly,

$$\Delta I = I - I_{min} = \sum_{i=1}^{n} p_i \;ln\; \frac{p_i}{q_i} - \sum_{i=1}^{n} p_{i0} \;ln \frac{p_{i0}}{q_i}$$

$$= \sum_{i=1}^{n} p_i \;ln\; \frac{p_i}{p_{i0}} + \sum_{i=1}^{n} (p_i - p_{i0}) \;ln\; \frac{p_{i0}}{q_i}$$

$$= \sum_{i=1}^{n} p_i \;ln\; \frac{p_i}{p_{i0}} + \sum_{i=1}^{n} (p_i - p_{i0})$$

$$\times [-\lambda_0 - \lambda_1 \, g_1(x_i) - \ldots - \lambda_m \, g_m(x_i)]$$

$$= \sum_{i=1}^{n} p_i \;ln\; \frac{p_i}{p_{i0}}$$

$$= \frac{1}{2} \sum_{i=1}^{n} \frac{(p_{i0} - p_i)^2}{p_{i0}} + \frac{1}{6} \sum_{i=1}^{n} \frac{(p_{i0} - p_i)^3}{p_{i0}^2} - \quad \ldots (54)$$

As such up to a first approximation :

$$2N \, \Delta S = 2N \Delta I = \sum_{i=1}^{n} \frac{(Np_{i0} - Np_i)^2}{Np_{i0}} = x_1^2, \qquad (55)$$

since Np_i are the observed frequencies and Np_{i0} are the expected frequencies. Again, since there are $m+1$ constraints, the number of degrees of freedom is $n-m-1=k$. This gives the proof of Jayne's entropy concentration theorem that $2N \, \Delta S$ (or $2N \Delta I$ is distributed asymptotically as chi-square with k *d.f.*

The proof also gives an interesting interpretation for the chi-square which is now seen to represent twice the difference between the observed entropy and the maximum entropy. Many statisticians have lamented that in spite of its usefulness, chi-square does not represent anything meaningful. In fact, chi-square is intimately connected with entropy maximization. Akaike [1] considered this as a confirmation of his thesis that "some of the most siguificant successes in the history of statistics were obtained when the statistician was directly dealing with the entropy and its maximization.".

Now, let there be N independent trials and at each trial let there be n possible results with probabilities $p_1, p_2, \ldots, p_n$ depending on the parameter μ. If $x_1, x_2, \ldots, x_n$ are the observed frequencies, the likelihood function is given by :

$$L(x_1, x_2, \ldots, x_n, \mu) = \frac{N!}{x_1!\, x_2! \ldots x_n!} p_1^{x_1} p_2^{x_2} \cdots p_n^{x_n} \quad (56)$$

$$\ln L = \ln \frac{N!}{x_1!\, x_2! \ldots x_n!} + \sum_{i=1}^{n} x_i \ \ln \frac{x_i}{N} + \sum_{i=1}^{n} x_i \ \ln \frac{\wedge p_i}{x_i}$$

$$= \ln C - \sum_{i=1}^{n} x_i \ \ln \frac{x_i}{Np} \quad (57)$$

where C is independent of p_i's and therefore of μ. Since :

$$\sum_{i=1}^{n} x_i = \sum_{i=1}^{n} Np_i = N, \quad (58)$$

by Shannon's inequality, the second term on the right$\geqslant 0$, and it will vanish iff $p_i = x_i/N$ so that $\ln C \geqslant \ln L$ so that C is the maximum value of L for variations in p_i's. Thus,

$$\ln L = \ln L_{max} + \sum_{i=1}^{n} x_i \ \ln\left[1 + \frac{Np_i - x_i}{x_i}\right]$$

$$= \ln L_{max} + \sum_{i=1}^{n} x_i \frac{Np_i - x_i}{x_i} - \frac{1}{2} \sum_{i=1}^{n} x_i^2 \left[\frac{Np_i - x_i}{x_i^2}\right]^2 + \ldots$$

$$(59)$$

or

$$\ln L = \ln L_{max} - \frac{1}{2} \sum_{i=1}^{n} \frac{(Np_i - x_i)^2}{x_i} + \frac{1}{3} \sum_{i=1}^{n} \frac{(Np_i - x_i)^3}{x_i^2} - \ldots$$

$$= \ln L_{max} - \frac{1}{2} \chi_i^2 + \ldots \quad (60)$$

where

$$\chi_3^2 = \sum_{i=1}^{n} \frac{(Np_i - x_i)^2}{x_i} \tag{61}$$

which differs only slightly from :

$$\chi_2^2 = \sum_{i=1}^{n} \frac{(Np_i - x_i)^2}{Np_i} \tag{62}$$

and

$$\chi_1^2 = \sum_{i=1}^{n} \frac{(Np_i - x_i)^2}{Np_{io}} \tag{63}$$

where p_{i0} is an estimate for p_i. By minimizing χ_1^2 or χ_2^2 or χ_3^2 with respect to μ, we can get minimum chi-square estimator for it.

This discussion connects chi-square with log likelihood function. Earlier we had related it with change in entropy so that we get :

$$2N \triangle S = 2N \triangle I = 2\triangle lnL = \chi_k^2 \tag{64}$$

or

$$2N(S_{max} - S) = 2N(I - I_{min}) = 2(\ell nL_{max} - lnL) = \chi_q^2 \tag{65}$$

This relation is true only asymptotically for large values of N. However, it gives a basic relationship between methods of maximum entropy, minimum entropy, maximum likelihood and minimum chi-square.

This gives an alternative method of defining entropy. Deviation from Maximum Entropy is the deviation from log maximum likelihood per trial. When observed frequencies are equal to expected frequencies, $L = L_{max}$, $S = S_{max}$, $I = I_{min}$.

Another important link between maximum-likehood, chi-square and Kullback's directed divergence is provided by the following result of Kuppermann [39].

Let $\underset{\sim}{x}_1, \underset{\sim}{x}_2, \ldots, \underset{\sim}{x}_N$ be a random sample from an exponential population with density function:

$$P(\underset{\sim}{x}, \underset{\sim}{\theta}) = q(\underset{\sim}{x})\, r(\underset{\sim}{\theta}) \exp\left[-\sum_{j=1}^{m} \lambda_j(\theta)\, g_j(\underset{\sim}{x})\right] \tag{66}$$

where $\underset{\sim}{x}$ is a k-dimensional and $\underset{\sim}{\theta}$ is an h-dimensional vector. Let $\hat{\theta}$ be the maximum-likelihood estimator for θ, then:

$$2N \sum_{i=1}^{N} p(x_i, \hat{\theta}) \ ln \ \frac{p(x_i, \hat{\theta})}{p(x_i, \theta)} \tag{67}$$

is distributed asymptotically as chi-square with k d,f.

According to the minimum information divergence principle, we usually are given $g(x)$ and we seek to find $f(x)$ minimizing :

$$I(f:g) = \int_{-\infty}^{\infty} f(x) \ ln \ \frac{f(x)}{g(x)} \ dx = \int_{-\infty}^{\infty} f(x) \ ln f(x) dx$$

$$- \int_{-\infty}^{\infty} f(x) \ ln g(x) \ dx \tag{68}$$

and satisfying certain constrains. Alternatively, we may be given $f(x)$ and we may seek to find $g(x)$ so that we have to maximize:

$$- \int_{-\infty}^{\infty} [ln \ g(x)] f(x) \ dx = = - \int_{-\infty}^{\infty} ln \ g(x) \ dF(x) \tag{69}$$

Now let $x_1, x_2, \ldots, x_n$ be a radom sample and let $F(x), -\infty < x < \infty$ correspond to the sample distribution defined by:

$$F(x) = \text{fraction of } x_1, x_2, \ldots, x_n \leq x \tag{70}$$

so that if $x_1, x_2, \ldots, x_n$ are in increasing order, we have

$$F(x) = 0 \quad \text{when } x < x_1, F(x) = \frac{1}{n}, x_1 \leq x < x_2, \ldots,$$

$$f(x) = 1 \text{ when } x \geq x_n \tag{71}$$

and (69) becomes :

$$- \frac{1}{n} \sum_{i=1}^{n} ln \ g(x_i) \tag{72}$$

Now let the density function $g(x)$ be indexed by a parameter θ so that $g(x,\theta)$ is a known function with an unknown parameter θ so that we have to choose θ so as to minimize:

$$- \frac{1}{n} \sum_{i=1}^{n} ln \ g(x_i, \ \theta) - \frac{1}{n} \log L(x_1, x_2, \ldots, x_n, \theta) \tag{73}$$

where L is the likelihood function so that minimizing the divergence information of $g(x, \theta)$ from $f(x)$ (where $f(x)$ corresponds to the sample distribution) is equivalent to maximizing the likelihood function. The function defined in (69), i.e.,

$$H(g:f) = -\int_{-\infty}^{\infty} \ln g(x, \theta) f(x)\, dx \tag{74}$$

is called the cross-entropy of g and f and we have minimized it to choose θ. For the discrete case, we get the expression:

$$-\sum_{i=1}^{n} p_i \ln q_i \tag{75}$$

which is called the inaccuracy [41]. In fact, we have

$$\sum_{i=1}^{n} p_i \ln \frac{p_i}{q_i} = -\sum_{i=1}^{n} p_i \ln q_i - \left[-\sum_{i=1}^{n} p_i \ln p_i\right] \tag{76}$$

$$\int_{-\infty}^{\infty} f(x) \ln \frac{f(x)}{g(x)}\, dx = -\int_{-\infty}^{\infty} f(x) \ln g(x)\, dx - \left[-\int_{-\infty}^{\infty} f(x) \ln f(x)\, dx\right] \tag{77}$$

so that

$$I(f:g) = H(g:f) - H(f:f) \tag{78}$$

or

$$\text{Information Divergence} = \text{Cross-entropy} - \text{Entropy} \tag{79}$$

This is an identity if $f=g$. If $g(x)=1$, it shows that minimizing information divergence in this case would give same results as maximizing entropy.

Our discussion shows that maximum likelihood principle can be regarded as a special case of the minimum information principle.

12. COMPARISON WITH METHOD OF MOMENTS

The Maximum Entropy Principle Method of Estimation has some similarities with the Method of Moments used by Karl Person; but which was strongly criticized by Fisher. Fisher assumed that he had the correct model $f(x, \theta_1, \theta_2, \dots, \theta_n)$ and his object was to estimate the parameters $\theta_1, \theta_2, \dots, \theta_n$. He gave the method of maximum likelihood and the criteria of consistency, efficiency and sufficiency and showed the superiority of his procedure over that of method of moments. This superiority was based on the assumption that the correct f was known [8, 42].

Pearson did not have one model, but a family of models in terms of his family of curves. He chose a member with four parameters and compared the first four moments of the observations with four moments of the distribution to get the estimates for the four parameters. Later he carried out a goodness of fit. If the fit was not good, he proceeded with another family of his family.

Pearson used $E(x^1)$, $E(x^2)$, $E(x^3)$, $E(x^4)$. In MEM. we also consider moments, but we also consider $E(lnx)^2$, $E(lnx)^2$, $Eln(1\text{-}x)$, etc.,. Pearson's method could lead to the family of maximal-entropy distributions :

$$f(x)=\exp[-\lambda_0-\lambda_1 x-\lambda_2 x^2-\lambda_3 x^3-\lambda_4 x^4] \tag{80}$$

This leaves out a large number of other distributions occuring in practice.

The main difference between the MEM and the MM is that the former has a sound theoretical principle to back it.

The MEP gives us which moments we should calculate from the data in order to estimate the parameters. Thus, for the beta distribution, we should calculate geometrical means of x and $1\text{-}x$. For the gamma distribution, the moments to be calculated are the arithmetic and geometric means of the observations.

13. GAUSS' PRINCIPLE OF ESTIMATION

Let $f(x, a_1, a_2, \dots, a_m)$ be the density function and let

$$E[g_r(x)]=a_r, \qquad r=1,2,\dots,m \tag{81}$$

then Gauss' principle considers those density functions for which the maximum likelihood estimators for a_r are :

$$\hat{a}_r = \frac{1}{N}[g_r(x_1)+g_r(x_2)+\ldots+g_r(x_n)] \tag{82}$$

Gauss considered the particular case of normal distribution only. It is obvious that Gauss' Principle of Estimation and Maximum Entropy Principle are equivalent. It can be shown that Gauss' principle leads to the exponential family of distributions and vice versa [3,35].

For exponential family members, the calculations of maximum likelihood estimates is relatively easy. For others it is relatively complicated. In fact, in the earlier stages the MM was sometimes preferred because it led to easier calculations. With the advent of computers, this advantage of the MM was lost. However, the representation of distributions in Monte Carlo studies, when only moments are known, borrows from the ideas of Karl Pearson and is strengthened by the Maximum Entropy Principle.

14. CONTINGENCY TABLES

For an m×n contingency table, in which all elements and totals are divided by the grand total, let S_1, S_2 and S_{12} denote the entropies of the marginal totals distributions and of the complete table. Then it is easily shown that :

$$-S_{12}+S_1+S_2=\sum_{j=1}^{n}\sum_{i=1}^{m} p_{ij}\ ln\ \frac{p_{ij}}{p_{i.}\ p_{.j}} \tag{83}$$

which $\geqslant 0$ by Shannon's inequality, and vanishes only when $p_{ij}=p_{i.}\ p_{.j}$, i.e., when the two attributes are independent. Thus, $S_1+S_2-S_{12}$ is a measure of dependence in the table. Substituting in (33)

$$p_{ij}=p_{j.}\ p_{.j}+e_{ij}, \tag{84}$$

we get :

$$S_1+S_2-S_{12}=-\sum_{j=1}^{n}\sum_{i=1}^{m} p_{ij}\ \ln\left[1-\frac{e_{ij}}{p_{ij}}\right]$$

$$=\frac{1}{2}\sum_{j=1}^{n}\sum_{i=1}^{m}\left[\frac{p_{ij}-p_{i.}\ p_{.j}}{p_i\ p_{.j}}\right]^2 \tag{85}$$

so that up to a first approximation 2 $(S_1+S_2-S_{12})$ is distributed as chi-square with appropriate degrees of freedom, and chi-square test appears as a test of 'closeness' of the entropy of the table to the entropy calculated on the hypothesis of independence of attribuies.

For a $d_1\times d_2\times\ldots\times d_k$ contingency table, we find similarly that $2(S_1+S_2+\ldots+S_k-S_{12\ldots,k})$ is distributed as chi-square. Here $S_1,S_2, \ldots, S_k$ are entropies of the marginal totals and $S_{12\ldots,k}$ is the entropy of the table as such.

We can similarly calculate entropies for other hypotheses of independence, e.g., of no second order interactions, of no third order interactions or of conditional independence of two attributes, knowing the third and then express the difference in entropies in terms of chi-squares [10, 12].

For transportation problems [23, 53], let x_{ij} denote the proportion of persons going from i^{th} origin to the j^{th} destination, then maximizing :

$$-\sum_{j=1}^{n}\sum_{i=1}^{m} x_{ij}\ ln\ x_{ij}$$

subject to :

$$\sum_{j=1}^{n} x_{ij}=x_i=O_i,\ \sum_{i=1}^{m} x_{ij}=x_{.j}=D_j,$$

$$\sum_{j=1}^{n}\sum_{i=1}^{m} x_{ij}\, c_{ij}=C \tag{86}$$

we get :

$$x_{ij}=A_i\ O_i\ B_j\ D_j\ e^{-\nu c_{ij}} \tag{87}$$

and the maximum entropy is given by :

$$S_{max}=-\sum_{i=1}^{m} x_{i.}\ ln\ A_i-\sum_{i=1}^{m} x_{i.}\ ln\ x_{i.}$$

$$-\sum_{j=1}^{n} x_{.j}\ ln\ x_{.j}-\sum_{j=1}^{n} x_{.j}\ ln\ B_j+\nu C \tag{89}$$

The quantity $2\left[S_{max}+\sum_{j=1}^{n} \sum_{i=1}^{m} x_{ij} \ln x_{ij} \right]$ is distributed as chi-square with $(m-1)(n-1)-1$ degrees of freedom.

Deeper results are obtained if we do not regard x_{ij}'s as fixed numbers, but rather as random variables satisfying constraints (86). We can then use the principle of Maximum Entropy to obtain the maximum entropy distributions of x_{ij}'s, both individually and jointly [36]

15. Some Historical Perspectives

When Shannon [47] discovered in 1948 his measure of uncertainty or information given by $-\sum_{i=1}^{n} p_i \ln p_i$, he first thought of calling it 'information', but he felt that this word was already overworked, so he consulted the great mathematician Von Neumann about the name for this measure. His response was direct, "You call it 'entropy', and for two reasons : (1) the function is already in use in thermodynamics under that name; (2) and more importantly, most people do not know what entropy really is, and if you use the word 'entropy' in an argument, you will win every time !" [51].

In retrospect, the advice appears to have been unsound on both counts. Shannon had discovered a measure for uncertainty associated with a probability distribution and the only thing common between his measure and thermodynamic entropy was that they had a common mathematical expression. Even in 1948, the expressions for entropy for Bose-Einstein and Fermi-Disc distributions were different from $-\sum_{i=1}^{n} p_i \ln p_i$. Later it was established that the thermodynamic entropy could be obtained from information-theoretic entropy through the principle of maximization of entropy.

However, the word entropy has been so well-entrenched in thermodynamics that even after twenty-five years, many persons consider the maximum-entropy principle as a principle of thermodynamics. The misunderstanding has been partly caused by the fact that the maximum entropy principle was first stated in 1957 by E.T. Jaynes [17] in the context of statistical mechanics. Also, in this way the maximization of entropy came to be associated with the second law of thermodynamics which states that the entropy always increases.

The unfortunate nomenclature may have been partly responsible for this principle not obtaining its rightful place in statitsical theory Jaynes [17] did say that he would consider entropy as equivalent to uncertainty. If he had gone a step further and had called his principle *The Principle of Maximum Uncertainty*, statisticians might have looked at it more closely because uncertainty is certainly the subject matter of statistics. Even if had called it *The Principle of Minimum Bias*, statisticians would have been interested because many statistical investigations are motivated by the consideration of minimizing bias.

Kulkback and Leibler [38] in 1951 gave the measure I (1:2) for discrimination between hypotheses H_1 and H_2. The MDI principle was not stated here. It was not stated by Kullback even in his book [37] published in 1959 where he stated : "Information theory is relevant to statistical inference and should be of basic interest to statisticians. Information theory provides a unification of known results, and leads to natural generalization of known results. The subject of this book is the study of logarithmic measures of information and their application to the testing of statistical hypotheses.'

Kullback concentrated exclusively on the testing of hypotheses and this became the main application of information theory in statistics. Kullback did not refer to Jaynes' work. We have of course, to make a distinction between application of information theory and applications of the maximum entropy principle and the MDI principle. The motivation for the MDI principle came much later jointly from Jaynes' maximum entropy principle and the Kullback-Leibler discrimination information number.

In statistics, Fisher [8] had defined information prior to Shannon. He considered the object of statistical inference to be to get as much information about the population as possible, ideally the whole of the information contained in the sample, but then proceeded to give a technical definition of information. The main goal of statistical inference was not worked out in detail in terms of this definition of information, though in the theory of optimal designs, the maximization of the determinant of the information matrix is considered.

If $f(x,\theta)$ and $f(x,\theta+\Delta\theta)$ are two density functions, where θ is a vector, then it is easy to show that :

$$I(\theta_i, \theta+\Delta\theta)=\frac{1}{2}\sum \sum g_{\alpha\beta}\, \Delta\theta_\alpha\, \Delta\theta_\beta \tag{90}$$

Where

$$g_{\alpha\beta}=\int f\left(\frac{1}{f}\frac{\partial f}{\partial \theta_\alpha}\right)\left(\frac{1}{f}\frac{\partial f}{\partial \theta_\beta}\right) dx \tag{91}$$

are the elements of Fisher's information matrix.

For the case of a single parameter, this shows that the greater the value of Fisher's information, the greater is the information for discriminating between $f(x,\theta)$ and $f(x,\theta+\Delta\theta)$ and so the density function $f(x,\theta)$ can be clearly determined.

Although Fisher's and Shannon's concepts of information are related, the prior introduction of information by Fisher may have inhibited statisticians from exploiting fully the powerful and general concept given by Shannon, with the important exception of Kullback who exploited it fully for generating known results about testing of hyphotheses.

Jaynes' work showed that statistical mechanics was more of a statistical theory than physical theory and this could have led to statistical mechanics being considered as a branch of mathematical statistics. On the other hand, the use of the word 'entropy' almost led to the feeling that the use of the entropy concept, on a large scale in statistics, may make mathematical statistics a branch of statistical mechanics !

Whenever the principle of maximum entropy is used in economics, geography, urban structure studies, marketing, etc., a feeling is unfortunately created that some arguments by analogy with physics or thermodynamics are being used, while essentially one is using probabilistic or statistical arguments.

Earlier we said that both the arguments of Von Neumann for recommending the use of the word 'entropy' were unfortunate. The first argument created a lot of confusion and misunderstanding. The second agrument was right; the use of the word entropy created a lot of mystery and awe and always enabled one to win an argument. However, ti delayed nhe penetration of statistics by this principle and the power of this principle could not be exploited in statistics. By using the word, one may create the feeling that one is using the laws of physics, while one may be using only the laws of uncertaintly and of statistics. Winning an argument is not as important as winning scientific truths.

16. The Role of Maximum Entropy Principle in Statistics

The Maximum Entropy Principle has been used in the discussion of the following problems in statistic:

(1) Characterisation of Probability Distributions;

(2) Estimation of Probability Distributions;

(3) Analysis of Categorical Data;

(4) Testing of Hypotheses;

(5) Time Series Analysiy;

We discuss these in turn below. In the present paper, we have discuss (1) and (2) in detail and (5) partially. In part II we shall discuss (3), (4) and (6) morefully.

16.1 Characterization of Probability Distributions

Reza [46] and Goldman [11] obtained uniform, exponential, gamme and normal distributions as maximum entropy distributions. Tribus [49] derived these and also beta and truncated normal distributions, but stated that Cauchy and Weibull distributions could not be dedtbed from the maximum entropy principle. Kagan, Linnin and Rao [22] characterized these as well as the Laplace distribution as maximum entropy distributions through the MEP. Lisman and Van Zuylen [40] gave maximum entropy characterization of geometric, chi-square, Cauchy and Weibull distributions, as well. Gokhale [11] also gave maximum entropy characterisation of some distributions. Dewson and Wragg [7, 54] discussed the maximum entropy distributions when the first two moments are prescribed over the semi-infinite interval $[0, \infty]$

In a series of papers, Kapur [24-27, 30] has systematically and comprehensively discussed the characterization of maximmm entropy distributions including the discrete distributions such as binomial, Poisson, geometric, generalized geometric, discrete normal, log series, negative binomial, generalized Poisson and Lagrangian distributions and the continuous-variate distributions over the intervals $(-\infty, \infty)$, $(0, \infty)$ and [a, b] when some of the moments $E(x)$, $E(x^2)$, $E(lnx)$, $E(ln\ (1-x))$, $E(ln\ (1+x))$, $E(ln\ (1+x^2)$, $E(lnx)^2$ *are* prescribed.

The multi-variate normal distribution had been obtained quite early as a ME distribution. Kapur (28, 29, 34) has also characterized as ME distributions more multi-variate distributions including the following : long normal, Dirichlet, inverted Dirchlet, generalized Cauchy, generalized gamma, generalized logistics, negative binomal generalized negative binomial and Lagrangian distributions. Kapur [31-33] has also obtained generalized distributions of order statistics by using MEP. He has also obtained multi-variate distributions of random variates when the only information available about them is that they are $\geqslant 0$ and their sum is unity. He has also obtained the distributions when additionally the means of the variates are known. Kapur [36] has also obtained the distribution of cell entries in contingency tables by regarding them as random variates.

Usually in statistics text book, one obtains every distribution using a different set of assumptions. Karl Pearson's was one major attempt to get a family of distributions by obtaining density functions as solutions of a differential equations with four parameters. Many other ways of characterizing probability distributions are given in Kagan, Linnik and Rao [22]. However, maximum-entropy characterization is the most comprehensive and the simplest.

Almost all the uni-variate and multi-variate distributions used in statistics can be obtained by prescribing some very simple moments and even a good undergraduate student should be able to obtain these in a systematic and unified manner by using the maximum-entropy principle.

It is interesting to observe that though some of the probability density function expressions (specially the multi-variate ones) look very complicated, their description in terms of the characterizing moments is always very simple. It is also interesting to note that statisticians have used only those distributions which can be obtained from the MEP by prescribing very simple moments. Consciously or unconsciously, the principles of maximum-entropy and simplicity appear to have been the guiding principles.

The problem of finding distributions characterized by minimum Fisherian information has also been considered, e.g. it has been shown that out of all the distribution with a location parameter and known finite voriance, the normal distribution has the minimum Fisherian information and out of all the distributions with a scale parameter and with known first and second order moments, the gamma distribution has minimum Fisherian information [22]. Random sample from these disributions give minimum information about the location and scale parameters respectiveiy.

However, it will be more interesting to characterize distribution as maximum Fisherian information distributions since we will be interested in knowing the distribution. random samples from which give maximum information about location, scale and other parameters of the population. In optimal design theory [55], suitable functions of Fisher's information matrix are maximized to get optimal designs.

Finding minimum Fisherian Information Distribution is like finding minimum entropy distributions because these will give, in some sense, the most biased or the most predictable distributions in light of the available information. However, such distributions are also very interesting These are usually discrte, not unique, and the least likely to arise, but ahese can be useful in pattern recognition [55, 56],

For obtaining discrete distributions, usually the MDI principle in more useful because the choice of a suitable prior is necessary. For the continuous distribution, the MEP is usually quite sufficient. Hobson and Cheng [15] strongly pleaded for greater use of the MDI principle and claimed superiority for it. Tribus and Rosetti [50] on the other hand, strongly defended the MEP. In practice, they are based on the same principle of minimum bias or maximum uncertainty and we can use either one which is convenient in a problem.

16.2 Estimation of Probability Distribution

Given a random sample $x_1, x_2, \cdots, x_n$ from a population with density function $f(x, \theta)$, the principle of MDI shows that the value of θ which minimizes I $[f: g]$, where g is the sample distribution, is the one which maximizes the likelihood function. In this sense the Maximum Likelihood Principle is a special case of the MDI principle. We may even consider this as a 'proof' of the Maximum Likelihood Principle.

If the form of f is known, we find the moments for which f is the maximum entropy distribution. Then we find the sample values of these moments and use these as estimates for the population parameters.

If the form of f is not known, but the characterizing moments are given, we can use the MEP to find the form of f.

Jaynes [16] established his entropy concentration theorem viz. that $2N(S_{max}-S)$ is asymptotically distributed as chi-square with n-m-l degrees of freedom where N is the size of the sample, m is the number of moment constraints, and n is the number of classes. This enables us to know how close a given distribution with given moments is to the maximum entropy distributions with the same values for moments. This show show the chi-square test is a test of the closeness of entropy to the maximum entropy.

Theil [48] has recently given another version of the minimum information diveragence principle. He chooses g to minimize $\int f(x) ln(f(x)/g(x))dx$. When some partial information is available about both f and g, e.g., if the form of $g(x)$ is given and the moments for $f(x)$ are known, and those moments those which characterize $g(x)$ as a maximum entropy distribution, then he seowed that $g(x)$ has the same moments as $f(x)$. Parzen [45] has further discussed the implications of this result.

The MEP and MDIP are clasely related to the minimum chi-square estimation principle and the method of moments.

16.3 Analysis of Categorical Data

The generation of hypothesis for multi-dimensional contingency tables by using the maximum-entropy principle, has been discussed

For the special case $\mu_x=0$, auto-correlation function is the same as the auto-covariance function.

We now define the spectral density $S_x(f)$ by :

$$S_x(f)=\Delta t \sum_{m=-\infty}^{\infty} R_x(m) \exp(-i\,2\pi fm\,\Delta t) \tag{97}$$

so that

$$R_x(m)=\int_{-\frac{1}{2}\Delta t}^{\frac{1}{2}\Delta t} S_x(f) \exp(i\,2\pi fm\,\Delta t)df \tag{98}$$

$S_x(f)$ and $R_x(m)$ form a Fourier transform pair.

In practice, we have only a finite number, say $2M+1$ values, of the auto-correlation function of a weakly stationary time series $\{x_n\}$ of zero mean. If we know $R_x(m)$ for all m, we could find $S_x(f)$. Now our problem is to find a spectral density $S_x(f)$ which corresponds to the most random or most unpredictable or most unbiased of time series where the auto-correlation function is consistent with the set of known values. This requires the principle of maximum entropy. The *MEP* gives an estimate which is asymptotically normal and is asymptotically unbiased.

The basic idea of the method is to extrapolate the auto-correlation function of the given time series by maximizing the entropy of the process. The method is well-suited to the spectral analysis of relatively short data records and as such the resolution of the method is usually superior to that obtained by using the conventional linear methods.

The maximum entropy method for use in spectral analysis was developed by Burg [2] in his Ph. D. thesis, almost independently of the work of Jaynes. The method is described in a monograph by Haykin [13]. A book, edited by Childers [4] contains a dozen papers on *MEM* published in the period 1967-1978. It also contains an extensive bibliography. The Proceedings of the First ASSP workshop on Spectral Estimations held at McMaster University of August 17/18, 1981 [14] contains seven papers on MEM including the paper by Jaynes and Parzen referred to earlier.

by Good [12]. More recently, Gokhale and Kullback [10] have given a comprehensive discussion of the analysis of contingency tables by the use of the MDI principle. Kapur [36] has discussed the estimation of probability distributions of cell entries when these are regarded as random variables.

16.4 Testing of Hypotheses

The entire book by Kullback [34] is devoted to testing hypotheses, but it involves Kullback-Leibler measures. It does not make use of the MEP or MDI[illegible]

16.5 Time Series Analysis

One of the most powerful applications of the maximum-entropy principle is to non-linear spectral analysis of time series data. The statistical discrete time series :

$$\{X_n\}=\{x_1, x_2, \ldots, x_N\} \tag{92}$$

represents a particular realization of a stochastic process. This will be a weak stationary process of order two if the statistical moments of the process upto order two depend on time differences only. The mean of the process is :

$$\mu_x(n)=E(x_n) \tag{93}$$

The auto-correlation function of the process for lag m and time origin n is given by :

$$R_x(m, n)=E(x_{n+m}x_n^*) \tag{94}$$

where x_n^* denotes the complex conjugate x_n. The corresponding auto-covariance function of the process is defined by :

$$C_x(m,n)=E\{(x_{n+m}-\mu_x(n+m)\ (x_n^*-\mu_n^*(n))\} \tag{95}$$

In the case of weakly stationary process of order two, the mean $\mu_x(n)$ and the auto-correlation function $R_x\,(m, n)$ are both independent of the time origin n so that :

$$\left.\begin{array}{l}\mu_x(n)=\mu_x=\text{const},\ R_x\,(m, n)=R_x(m),\\ C_x(m, n)=C_x(m)\end{array}\right\} \quad (96$$

16.6 *Estimation of Missing Data*

We conclude by giving a very simple example of the application of the MEP. Suppose we are given a set of observations x_1, x_2, ..., x_n and we know one observation is missing. What is the most unbiased value for this? Let x be this value and let T be the total of the known observations, then maximizing:

$$-\sum_{i=1}^{n} \frac{x_i}{T+x} \ln \frac{x_i}{T+x} - \frac{x}{T+x} \ln \frac{x}{T+x} \qquad ...(99)$$

we get,

$$x = \left[x_1^{x_1}, x_2^{x_2}, ..., x_n^{x_n}\right]^{1/T}$$

Similarly, if two values, x and y are missing, then x and y are determined from:

$$x = \left[x_1^{x_1}, x_2^{x_2}, ..., x_n^{x_n}, y^y\right]^{\frac{1}{y+T}} \qquad ...(101)$$

$$y = \left[x_1^{x_1}, x_2^{x_2}, ..., x_n^{x_n}, x^x\right]^{\frac{1}{x+T}} \qquad ...(102)$$

If $x_1 = x_2 = ... x_n = \mu$, we get, of course, $x = \mu$, $y = \mu$. We shall give more examples of this type in Part II.

REFERENCES

[1] Akaike, H. (1977) "The entropy Maximization Principle", in *Applications of statistics*, edited by P.R. Krishniah, pp. 27-42, North Holland.

[2] Burg, J.P. (1973) *Entropy Spectral Analysis*, Ph. D. Dissertation, Stanford University, University Microfilm 77-25, p. 499.

[3] Campbell, L.L. (1970) "Equivalence of Gauss' Principle and Minimum Discrimination Estimation of Probabilities", *Ann. Math. Stat,, 41*, pp. 1011-1015.

[4] Childers, D.G. (ed.) (1978) *Modern Spectrum Analysis*, IEEE Press, New York, NY.

[5] Consul, P.C. & G.C. Jain (1972) "A Generalization of Poisson Distributions". *Technometrics, 15*, pp. 791-799.

[6] Consul, P.C. & L.R. Shenton (1972), "The Use of Lagrange's Expansion for Generating Discrete Generalized Probability Distributions", *SIAM J. Appl. Math.*, *23*, pp. 239-248.

[7] Dawson, D.C. & A. Wragg (1973) "Maximum-Entropy Distribution Having Prescribed First and Second order Moments", IEEE Trans., IT-19, pp. 689-693.

[8] Fisher, R.A. (1922), "On the Mathematical Foundations of Theoretical Statistics" *Phil. Trans.*, *222A*, pp. 309-368.

[9] Gokhale, D.V. (1975) "Maximum Entropy Characterization of Some Distrbutions", in *Statistical Distributions in Scientific Work*, Vol. 3. edited by Patel, Golz & Ord., M.A. Ridel, Boston, pp. 299-309.

[10] Gokhale. D V. & S. Kullback (1978) *The Information on Contingency Tables*, Marcel Dekker, NY.

[11] Goldman, S. (1953) *Information Theory*, Prentice-Hall, NY.

[12] Good, J.J. (1965) "Maximum Entropy for Hypotheses Formulation Especially for Multi-Dimensional Contingency Tables", *Ann. Math. Stat*,, *34*, pp. 911-934.

[13] Haykin, S. (ed.) (1979) *Non-Linear Methods in Spectral Analysis*, Springer-Verlag, NY.

[14] Haykin, S. & J. Codzow (eds.) (1981) *Proceedings of the First ASSP Workshop on Spectral Estimation*, McMaster University.

[15] Hobson, A. & B.K. Cheng (1973) "A Comparison of the Shannon and Kullback Information Measures;', *J. Stat. Phy.*, 7, pp. 306-319.

[16] Jain, G.C. & P.C. Consul (1971) "A Generalization of the Negative Bionomial Distribution", *SIAM J, Appl. Math.*, *21*, pp. 501-513

[17] Jaynes, E.T. (1957) "Information Theory and Statistical Mechanics", *Phy. Review*, *106*, pp. 620-630; *108* pp. 171-179.

[18] ——(1981) "What is the Problem?", in Reference 14, pp. 1. 1. 1-1. 1. 10.

[19] Johnson, M.L. & S. Kotz (1969), *Discrete Distributions*, Houghton-Miffin.

[20] ——(1970, 1971) *Continuous Distributions*, I-11, Houghton-Miffin.

[21] ——(1972) *Multivariate Distributions*, Wiley NY.

[22] Kagan, A.M., J.V. Linnik, & C.R. Rao (1973) *Characterization Problems in Mathematical Statistics*, Wiley, NY.

[23] Kapur, J.N. (1980a) *Entropy Maximization Models in Regional and Urban Planning*, Manitoba University Research Report. (Also Appendix 11)

[24] (1980b) *Characterization of Probability Distributions Through Entropy Maximization Principle*, Manitoba University Research Report.

[25] (1980c) *On the Entropy Characterization of Normal and Laplace Distributions*, Manitoba University Research Report.

[26]. (1980d) *Entropy Maximizing Probability Distribution for a Continuous Random Variate*, Manitoba University Research Report. (Also Appendix 26)

[27]. (1980e) *Maximum Entropy Probability Distributions for a Continuous Random Variate over a Finite Interval,* Manitoba University Research Report. (Also Appendix 14)

[28]. (1980f) *Maximum Entropy Continuous Multivariate Distributions*, Manitoba University Research Report.

[29]. (1980g) *Maximum Entropy Discrete Multivariate Distributions*, Manitoba University Research Report.

[30]. Kapur, J.N. & S. Bhatt (1980) *Derivation of Some Purchase Incidence Models from Entropy Maximization Principle*, Manitoba University Research Report. (Also Appendix 28)

[31]. Kapur, J.N. (1981a) *Some Multivariate Distributions for Ordered Random Variates*, Waterloo University Research Report.

[32[. (1981b) *Entropy Maximization Distribution for Ordered Statistics*, Waterloo University Research Report.

[33]. (1981c) *Entropy Maximizing and Stochastic Processes*, Waterloo University Research Report.

[34]. (1981d) *Maximum Entropy Formalism for Some Univariate and Multivariate Lagrangian Distribution*, Waterloo University Research Report. (Also Appendix 19).

[35]. (1981e) *On Equivalence of Gauss' and Minimum Discrimination Principles*, Waterloo University Research Report.

[36]. (1981g) *Entropy Maximization Distributions and Contingency Tables*, Waterloo University Research Report. (Also Appendix 15)

[37]. Kullback, S. (1959) *Informatisn Theory and Statistics*, Wiley, NY.

[38]. Kullback, S. & R.A. Leibler (1951) "On Information and Sufficiency", *Ann. Math. Stat.*, 22, pp. 79-86.

[39]. Kuppermann, M. (1957) *Further Applications of Informations Theory to Multivariate Analysis and Statistical Inference*, Ph. D. Dissertation, George Washington University.

[40]. Linsman, J.H.R. & M.C.A. Van Zuylen (1972) "Note on the Generation of the Most Probable Frequency Distribution", *Statistica Netherlander*, *25*, pp. 19-23.

[41]. Mathi, A.M. & P.N. Rathie (1975) *Basic Concepts of Information Theory and Statistics*, Wiley, NY.

PUBLICATIONS OF INFORMATION THEORY AND MAXIMUM-ENTROPY MODELS

BY

J. N. KAPUR

1. Generalised Entropy of order α and type β. *The Mathematics Seminar* **4**, 78-94, 1967.

2. On the postulates of entropy theory, *The Mathematics Seminar*, **4**, 95-102, 1967.

3. Some inequalities for generalised entropies, *Progress of Mathematics* **2**, 181-188, 1968.

4. On some Applications of dynamic programming to information theory, Proceedings Indian Academy of Sciences, **68A**, 1-11, 1968.

5. On information of order α and type β, Proceedings Indian Academy of Sciences **68A**, 65-75, 1968.

6. Reny'is entropy for generalised discrete and continuous probability distribution (with P.N. Chabra) *Def. Sci. Journ* **19**, 77-92.

7. Some properties of generalised entropies, *Indian Journal of Mathematics*, **9**, 427-442, 1967.

8. Some properties of entropy of order α and type β, Proceedings Ind. Acad. Sciences **69A**, 201-211, 1969.

9. Measures of uncertainty, mathematical programming and physics, *Journal of Indian Society of Agricultural Statistics* **24**, 47-66, 1972.

10. Generalised entropies for a continuous random variate, *The Mathematics Student* **42**, 353-360, 1974.

11. Entropy Maximization Models for Regional and Urban Planning *Int. Journ. Math. Edu. Sci. Tech.* **13**(6), (698-714, 1982.

12. Bayesian Entropy and its applications to Entropy Maximisation Models. *Selecta Statistica Canadiana* (To appear).

13 On Maximum-Entropy Complexity Measures, *Int. Journ. of General System* **9**, 95-102, 1983.

14. Maximum Entropy Probability Distributions for a continuous random variate over a finite interval, *J. Math. Phy Sciences* **16**(1), 97-103, 1982.

15. Maximum-Entropy Distributions for Contingency Tables, Acta Ciencia Indica (To appear).

16. On the estimation of the Entropy Parameter, *Acta Ciencia Indica* (To appear).

17. The Maximum-Entropy Principle and its Applications to Science and Engineering. Proc. Nat. Symposiom on Mathematical Modelling, *MRI* Allahabad, 75-98, 1984.

18. On Lee's Markovian Entropy-Maximization Model for Population Distributions, Environment and Planning, **15**, 1449-1455, 1983.

19 Maximum-Entropy Formalism for some univariate and multivariate Lagrangian distributions. *AMU Journal of Statistics* **2**, 1-16, 1982.

20. On bias and variance of estimates for the entropy parameter, *Jour. Math. Phys. Sciences* **16**(4), 329-346, 1982.

21. A comparative assessment of various measures of entropy. Journal of Information and Optimization Sciences, **4**(3), 207-232, 1983.

22. Non-additive measures of entropy and distributions of statistical mechanics *IJPAM*, 14(11), 1372-1387, 1983.

23. On Maximum-Entropy Estimation of Missing Values *Nat. Acad Sci. Letters*, **6** (2), 59-65, 1983.

24. On the relationship between some probability distribution measures of entropy, measures of directed divergence and distributions of statistical mechanics. *Indian Journal Pure Applied Mathematics*, **14**(2) 1435-1443, 1983.

25. Twenty-five years of Maximum-Entropy Principle. *Joutnal of Mathematical and Physical Sciences* **17**(2) 103-153, 1983.

26. Entropy-Maximizing Probability Distributions for Continuous random variates. *Journal Indian Soc. Ag. Stat.*, **35**(3), 91-103, 1983.

27. The generalised entropy model for brand switching (with C.R. Bector and U. Kumar) *Naval Research Logistics Quarterly*, **31**, 183-198, 1983.

28. Derivation of some purchase incidence models from entropy maximization principle (with S.K. Bhatt) Proc. Eleventh Annual Conference of Administrative Sciences Association of Canadian Vol. **4**, Part **3**, 19-35, 1983.

29. Generalisation of some functional equations in information theory, *Journ. Math. Phys. Sci.* **17**(4), 331-339, 1983.

30. Maximum-Entropy Principle and Search Theory. Journal of *National Academy of Mathematics*, **1**(2) 99-104, 1983.

31. Maximum-Entropy Principle and Flexible Manufacturing Systems *Defence Science Journal* (To appear).

32. A comparative assessment of measures of directed divergence *Advances in Management Sciences* (To appear).

33. Derivation of logistic law of population growth from maximum entropy principle. National Academy Science Letters **6** (12)

34. Generalisation of Logistic and Time delay models through maximum-entropy principle, *Mathematics Forum* (To appear).

35. Maximum-Entropy Models in Science and Engineering Presidential Address, Physical Sciences Section, National Academy of Sciences, Madurai.

36. On generalised entropies and divergence of order α and type β Journal of Organisational Behaviour (with C.R. Bector and B.L. Bhatia), *Journal of Organisational Behaviour* (To appear).

37. Voting Behaviour Through Entropy Approach (with C.R. Bector and Uma Kumar) Journal of Information and Optimization Sciences. (to appear).

38. *Maximum-Entropy Models in Science and Engineering*, South Asia Publishers, New Delhi (To appear).