American J. of Mathematics and Sciences
Vol. 3, No -1 ,(January 2014)
Copyright © Mind Reader Publications
ISSN No: 2250-3102

APPROXIMATION OF A GIVEN PROBABILITY DISTRIBUTION THROUGH MAXIMUM ENTROPY PRINCIPLE

G.S.Buttar
Deptt. of Statistics
Punjabi University
Patiala (India)

P.K.Sharma
Deptt. of Mathematics
Hindu College
Amritsar (India)

Mukesh Sharma
Deptt. of Statistics
Punjabi University
Patiala (India)

ABSTRACT

This is worth mentioning that measures of information find tremendous applications in a variety of disciplines including Mathematics, Statistics and Operations Research. In the present communication, we have provided the applications of some specified measures of entropy and directed divergence to the field of statistics for evaluating approximately the probability of a given distribution by using the maximum entropy principle.

Key Words: Probability distribution, Entropy, Uncertainty, Directed divergence, Moments, Maximum entropy principle.

INTRODUCTION

After the introduction of the concept of entropy by Shannon [8], it was realized that entropy is a property of any stochastic system and the concept is now used widely in many fields. The tendency of the systems to become more disordered over time is described by the second law of thermodynamics, which states that the entropy of the system cannot spontaneously decrease. Today, information theory is still principally concerned with communication systems, but there are widespread applications in statistics, information processing and computing. A great deal of insight is obtained by considering entropy equivalent to uncertainty and a generalized theory of uncertainty has well been explained by Zadeh [9]. This uncertainty is called entropy, since this is the terminology that is well entrenched in the literature. Shannon [8] introduced the concept of entropy by associating uncertainty with every probability distribution $P = \left(p_1,\ p_2,\ ....,\ p_n\right)$ and found the following unique function that can measure it:

$$H(P) = -\sum_{i=1}^{n} p_i \text{ In } p_i \qquad (1.1)$$

The probabilistic measure of entropy (1.1) possesses a number of interesting properties. Immediately, after Shannon gave his measure, research workers in many fields saw the potential of the application of this expression and a large number of other information theoretic measures were derived. Renyi [7] defined entropy of order α as:

$$H_{\alpha}(P) = \frac{1}{1-\alpha} In\left(\sum_{i=1}^{n} p_i^{\alpha} \Big/ \sum_{i=1}^{n} p_i\right), \alpha \neq 1, \alpha > 0 \qquad (1.2)$$

which includes Shannon's [8] entropy as a limiting case as α→ 1. Zyczkowski [10] explored the relationships between the Shannon's [8] entropy and Renyi's [7] entropies of integer order. Some work related with the discontinuity of Shannon's measure has been done by Ho and Yeung [4].

Havrada and Charvat [3] introduced first non-additive entropy, given by:

$$H^{\alpha}(P) = \frac{\left[\sum_{i=1}^{n} p_i^{\alpha}\right] - 1}{2^{1-\alpha} - 1}, \alpha \neq 1, \alpha > 0 \qquad (1.3)$$

Kapur [5] introduced the following generalized measure of entropy:

$$H_{\beta}^{\alpha}(P) = \frac{1}{\alpha+\beta-2}\left[\sum_{n=0}^{\infty} p_n^{\alpha} + \sum_{n=0}^{\infty} p_n^{\beta} - 2\right], \alpha \geq 1, \beta \leq 1 or\, \alpha \leq 1, \beta \geq 1 \qquad (1.4)$$

Dehmer and Mowshowitz [2] described methods for measuring the entropy of graphs and to demonstrate the wide applicability of entropy measures. The authors have discussed the graph entropy measures which play an important role in a variety of problem areas, including biology, chemistry, and sociology, and moreover, developed relationships between selected entropy measures, illustrating differences quantitatively with concrete examples.

One of the basic concepts in the applications of information theory is that of entropy whereas another concept which is of basic importance is that of "distance" or of "directed divergence". In fact, of the two concepts, the concept of directed divergence is the more fundamental, since the concept of entropy can be derived from it and it is of great importance in all applications of mathematics to science and engineering. Naturally attempts were made to extend the concepts of distance for application to problems in other fields. Such a measure of distance usually denoted by $D(P:Q)$ which is defined as the discrepancy of the probability distribution P from another probability distribution Q has been developed by Kullback and Leibler [6]. In some sense, it measures the distance of P from Q and is given by

$$D(P:Q)=\sum_{i=1}^{n} p_i \ln \frac{p_i}{q_i} \tag{1.5}$$

Recently, Cai, Kulkarni and Verdu [1] remarked that Kullback-Leibler's [6] divergence is a fundamental information measure, special cases of which are mutual information and entropy, but the problem of divergence estimation of sources whose distributions are unknown has received relatively little attention.

Some parametric measures of directed divergence are:

$$D_\alpha(P:Q)=\frac{1}{\alpha-1}\ln\sum_{i=1}^{n} p_i^{\alpha} q_i^{1-\alpha}, \alpha\neq 1, \alpha>0 \tag{1.6}$$

which is Renyi's [7] probabilistic measure of directed divergence.

$$D^\alpha(P:Q)=\frac{1}{\alpha-1}\left[\sum_{i=1}^{n} p_i^{\alpha}\ q_i^{1-\alpha}\ -1\right], \alpha\neq 1, \alpha>0 \tag{1.7}$$

which is Havrada and Charvat's [3] probabilistic measure of divergence.

These measures of information including both, that is, measures of entropy and divergence find tremendous applications in a variety of disciplines. In the present communication, we have provided the applications of these measures to the field of probability theory by using the maximum entropy principle.

2. APPROXIMATING A PROBABILITY DISTRIBUTION VIA MAXIMUM ENTROPY PRINCIPLE

It is known fact that in many practical problems while dealing with the various disciplines of Operations Research and Statistics, we do not get simple expressions for the probability distributions. In all such cases, it becomes very difficult to apply these complicated expressions for further mathematical treatment in the manipulation of new results. Thus, it becomes the desirability to approximate these probability distributions. The approximating probability distributions should have some common properties with the given distribution. The simplest property is of having some common moments.

There may be an infinite number of distributions with the same first moment as the given distribution but we are interested with only that probability distribution which is most unbiased and from the theory of maximum entropy principle, we accept only that distribution which has maximum entropy. This fundamental principle will provide our first approximation to the given distribution. This result is based upon the postulate that most probability distributions are either maximum entropy distributions or very nearly so. To find a better approximation, we try to find that maximum entropy probability distribution which has two moments in common with the given probability distribution. As the number of moments goes on increasing, we get much better and better approximations and as a result of this we obtain the desired approximation of the given probability distribution. We illustrate the above mentioned principle by considering the following numerical example:

Numerical Example:

Let us consider the theoretical probability distribution P, given by

i :	0	1	2	3	4
p_i :	0.4	0.3	0.2	0.07	0.03

Our problem is to find MEPD with

(I) same mean;

(II) same mean and p_o ;

(III) same mean, p_o and p_1 ; and

(IV) same first two moments

Our purpose is also to find that maximum entropy probability distribution which is closest to the given probability distribution P.

To solve the above problem, we make use of maximum entropy principle by using Havrada and Charvat's [3] entropy of order 2 and our problem becomes:

(I) Maximize Havrada and Charvat's [3] entropy of order 2 given by

$$H^{\alpha}(P) = \frac{1}{\alpha(1-\alpha)}\left[\sum p_i^{\alpha} \; -1\right]; \alpha \neq 1, \alpha > 0$$

$$H(P) = \frac{1}{2}(1-\sum_{i=0}^{4} p_i^2) \tag{2.1}$$

subject to the following set of constraints

(i) $$\sum_{i=0}^{4} p_i \; = 1 \tag{2.2}$$

(ii) $$\sum_{i=0}^{4} ip_i \; = 1.03 \tag{2.3}$$

The corresponding Lagrangian is given by

$$L = \frac{1}{2}(1-\sum_{i=0}^{4} p_i^2) - \lambda\left\{\sum_{i=0}^{4} p_i \; -1\right\} - \mu\left\{\sum_{i=0}^{4} ip_i \; -1.03\right\}$$

Hence $\frac{\partial L}{\partial p_i} = 0$ gives

$$p_i = -(\lambda + i\mu) \tag{2.4}$$

Applying (2.2), we get

$$\lambda + 2\mu = -0.2 \tag{2.5}$$

Applying (2.3), we get

$$\lambda + 3\mu = -0.103 \tag{2.6}$$

From (2.5) and (2.6), we have $\lambda = -0.394$ and $\mu = 0.097$

With these values of λ and μ , equation (2.4) gives the following set of probability distribution:

$p_0 = 0.3940$, $p_1 = 0.2970$, $p_2 = 0.2000$, $p_3 = 0.1030$, $p_4 = 0.0006$. Obviously, $\sum_{i=0}^{4} p_i = 0.9946 \cong 1$

Thus, the first MEPD P_1 is given by $P_1 = \{0.3940, 0.2970, 0.2000, 0.1030, 0.0006\}$

(II) In this case, our problem is to maximize Havrada and Charvat's (1967) entropy (2.1) under the set of constraints (2.2), (2.3) and $p_o = 0.4$.

Now $\sum_{i=0}^{4} p_i = 1$ gives that $p_o + \sum_{i=1}^{4} p_i = 1$

or $4\lambda + 10\mu = -0.6$ (2.7)

Applying (2.3), we get

$10\lambda + 30\mu = -1.03$ (2.8)

Equations (2.7) and (2.8) together give $\lambda = -0.385$ and $\mu = 0.094$

With these values of λ and μ, equation (2.4) gives the following set of probability distribution:

$p_0 = 0.4000$, $p_1 = 0.2910$, $p_2 = 0.1970$, $p_3 = 0.1030$, $p_4 = 0.0090$. Obviously, $\sum_{i=0}^{4} p_i = 1$

Thus, the second MEPD P_2 is given by $P_2 = \{0.4000, 0.2910, 0.1970, 0.1030, 0.0090\}$

(III) In this case, our problem is to maximize Havrada and Charvat's [3] entropy (2.1) under the set of constraints (2.2), (2.3), $p_o = 0.4$ and $p_1 = 0.3$.

$$p_o + p_1 + \sum_{i=2}^{4} p_i = 1$$

This gives $\sum_{i=2}^{4} p_i = 0.3$

Applying (2.4), we get

$\lambda + 3\mu = -0.1$ (2.9)

Also (2.3) gives

$9\lambda + 29\mu = -0.73$ (2.10)

Equations (2.9) and (2.10) together give $\lambda = -0.355$ and $\mu = 0.085$

With these values of λ and μ, equation (2.4) gives the following set of probability distribution:

$p_0 = 0.4000$, $p_1 = 0.3000$, $p_2 = 0.1850$, $p_3 = 0.1000$, $p_4 = 0.0150$. Obviously, $\sum_{i=0}^{4} p_i = 1$

Thus, the third MEPD P_3 is given by $P_3 = \{0.4000, 0.3000, 0.1850, 0.1000, 0.0150\}$

(IV) In this case, our problem is to maximize Havrada and Charvat's [3] entropy of order 2 subject to the set of constraints (2.2) and (2.3) along with the additional constraint given by

$$\sum_{i=0}^{4} i^2 p_i = 2.21 \qquad (2.11)$$

The corresponding Lagrangian is given by

$$L = \frac{1}{2}(1-\sum_{i=0}^{4} p_i^2) - \lambda\left\{\sum_{i=0}^{4} p_i - 1\right\} - \mu\left\{\sum_{i=0}^{4} ip_i - 1.03\right\} - w\left\{\sum_{i=0}^{4} i^2 p_i - 2.21\right\}$$

Hence $\frac{\partial L}{\partial p_i} = 0$ gives

$$p_i = -(\lambda + i\mu + i^2 w) \qquad (2.12)$$

Applying (2.2), equation (2.12) gives

$$5\lambda + 10\mu + 30w = -1 \qquad (2.13)$$

Applying (2.3), equation (2.12) gives

$$10\lambda + 30\mu + 100w = -1.03 \qquad (2.14)$$

Applying (2.11), equation (2.12) gives

$$30\lambda + 100\mu + 354w = -2.21 \qquad (2.15)$$

After solving equations (2.13), (2.14) and (2.15), we get $w = -0.0064$, $\mu = 0.1226$ and $\lambda = -0.4068$

With these values of w, λ and μ equation (2.12) gives the following set of probability distribution:

$p_0 = 0.4068$, $p_1 = 0.2906$, $p_2 = 0.1872$, $p_3 = 0.0966$, $p_4 = 0.0188$. Obviously, $\sum_{i=0}^{4} p_i = 1$

Thus, the fourth MEPD P_4 is given by $P_4 = \{0.4068, 0.2906, 0.1872, 0.0966, 0.0188\}$

Our next aim is to find that maximum entropy probability distribution which is closest to the given probability distribution P.

For this purpose, we use Havrada and Charvat's [3] directed divergence of order 2 to determine the approximity of the distributions to P. We know that Havrada and Charvat's [3] directed divergence is given by

$$D^{\alpha}(P:Q) = \frac{1}{\alpha(\alpha-1)}\left[\sum_{i=0}^{4} p_i^{\alpha} q_i^{1-\alpha} - 1\right], \alpha \neq 1, \alpha > 0,$$

Thus for $\alpha = 2$, we have

$$D(P:Q) = \sum_{i=0}^{4} \frac{p_i^{\,2}}{q_i} - 1 \qquad (2.16)$$

Using (2.16), we get the following results:

$D(P_1:P) = 0.017439$, $D(P_2:P) = 0.015286$, $D(P_3:P) = 0.010741$, $D(P_4:P) = 0.007759$

Hence, we observe that MEPD P_4 is closest to the given probability distribution P.

Important observations. We also make out the following observations:

(a) Since, P_2 is based upon information about mean and p_0 whereas P_1 is based upon information about mean only, we must expect that

$$D(P_2 : P) < D(P_1 : P)$$

In our case, it is found to be true.

(b) Since, P_3 is based upon information about mean, p_0 and p_1 ,we must expect that

$$D(P_3 : P) < D(P_2 : P) < D(P_1 : P)$$

In our case, it is found to be true.

(c) Since, P_4 is based upon information about the first two moments, whereas P_1 is based upon information about mean only, we must expect that

$$D(P_4 : P) < D(P_1 : P)$$

In our case, it is found to be true.

(d) Since, P_2 is based upon information about mean and p_0 ,whereas P_4 is based upon mean and second moment and

$$D(P_4 : P) < D(P_2 : P)$$

Thus, we conclude that P_0 gives less information than the second moment.

REFERENCES

[1] Cai, H., Kulkarni, S. and Verdu, S. (2006). Universal divergence estimation for finite-alphabet sources. *IEEE Transactions on Information Theory* **52**: 3456-3475.

[2] Dehmer, M. and Mowshowitz, A. (2011). A history of graph entropy measures. *Inform. Sci.* **181**(1): 57-78.

[3] Havrada, J.H. and Charvat, F. (1967). Quantification methods of classification process:Concept of structural α-entropy. *Kybernetika* **3**: 30-35.

[4] Ho, S.W. and Yeung, R. R. (2009). On the discontinuity of the Shannon information measure. IEEE Trans. Inform. Theory 55(12): 5362-5374.

[5] Kapur, J.N. (1986). Four families of measures of entropy. *Indian Journal of Pure and Applied Mathematics* **17**: 429-449.

[6] Kullback, S. and Leibler, R. A. (1951). On information and sufficiency. *Annals of Mathematical Statistics* **22**: 79-86.

[7] Renyi, A. (1961). On measures of entropy and information. *Proceedings 4th Berkeley Symposium on Mathematical Statistics and Probability* **1**: 547-561.

[8] Shannon, C. E. (1948). A mathematical theory of communication. *Bell System Technical Journal* **27**: 379-423, 623-659.

[9] Zadeh, L. A. (2004). Precisiated natural language (PNL). *Al Magazine* **25**: 74-91.

[10] Zyczkowski, K. (2003). Renyi extrapolation of Shannon entropy. *Open Systems and Information Dynamics* **10**: 297-310.