International Journal of Pure and Applied Mathematics
Volume 84 No. 5 2013, 517-524
ISSN: 1311-8080 (printed version); ISSN: 1314-3395 (on-line version)
url: http://www.ijpam.eu
doi: http://dx.doi.org/10.12732/ijpam.v84i5.5

RELATION BETWEEN INFORMATION MEASURES AND CHI-SQUARE STATISTIC

Om Parkash[1], Mukesh[2] §
[1,2]Department of Mathematics
Guru Nanak Dev University
Amritsar, 143005, INDIA

Abstract: In the existing literature of Information theory and Statistics, there are many well known information theoretic measures, each with its own merits, limitations and areas of application. In the present communications, we have provided the applications of parametric information theoretic measures to one of the disciplines of Statistics. By using optimization principles, we have proved that the information contents both for entropy and divergence are distributed asymptotically as Chi-square statistic.

AMS Subject Classification: 94A
Key Words: parametric entropy, optimization principles, parametric cross entropy, Chi-square statistic

1. Introduction

In communication theory, the concept of entropy was introduced by Shannon [12] whereas an important and most widely used concept of divergence was introduced by Kullback and Leibler [5]. In the literature of information theory, both concepts are prevalent and find tremendous applications in a variety of disciplines. Shannon's [12] measure of entropy is probabilistic in nature and is

Received: October 17, 2012

§Correspondence author

given by

$$H(P) = -\sum_{i=1}^{n} p_i \log p_i. \tag{1}$$

Eguchi and Kato [3] has remarked that in statistical physics, Boltzmann-Shannon entropy provides good understanding for the equilibrium states of a number of phenomena. In statistics, the entropy corresponds to the maximum likelihood method, in which Kullback-Leibler divergence measure connects Boltzmann-Shannon entropy and the expected log-likelihood function. Rao [10] has provided an axiomatic setup for an entropy function as a measure of diversity and provided a general definition of cross entropy or directed divergence along with its use in solving a variety of stochastic and non-stochastic optimization problems. Kumar and Taneja [6] have studied a generalized cumulative residual information that parallels Shannon's [12] entropy.

Kullback-Leibler [5] have provided a measure of divergence given by

$$D(P;Q) = \sum_{i=1}^{n} p_i \log \frac{p_i}{q_i}. \tag{2}$$

In the literature of distance measures, usually only one single asymmetric (Alpha, Beta or Gamma) divergence is considered. Cichocki and Amari [2] have shown that there exist families of such divergence measures with the same consistent properties. Moreover, the authors have established links and correspondences among these divergence measures by applying suitable non-linear transformations. Their paper bridges these divergences and shows also their links to Tsallis and Renyi entropies. Botey and Kroese [1] have used generalized measures of cross entropy and described a framework for density estimation. The effectiveness of their approach is demonstrated through an application to some well-known density estimation test cases. Recently, Parkash and Mukesh [9] have provided the applications of divergence measure by developing an optimizational principle for minimizing risk in portfolio analysis. Some other characterizations and generalization of the measures of entropy and directed divergence along with their detailed properties have been provided by Renyi [11], Havrada and Charvat [4], Parkash and Mukesh [7,8].

In the present communications, we have developed the relations between parametric information measures (both entropy and divergence) and chi-square statistic by using the optimization principles applied to the parametric measures of entropy and cross-entropy introduced recently by Parkash and Mukesh [8].

2. Relation between Information Theoretic Measures and Chi-Square Statistic

Before developing these relations, we first of all construct the following two optimization principles applied to the parametric measures of entropy and cross entropy.

Let us suppose that a random variable takes values $x_1, x_2, ..., x_n$, but we do not know the probabilities $p_1, p_2, ..., p_n$ with which these values are taken. We also assume that the only information available with us is the natural constraint on probabilities, that is,

$$\sum_{i=1}^{n} p_i = 1. \tag{3}$$

We now have infinity of probability distributions satisfying the constraint (2.1) and we need to have a principle to be so chosen, which in some sense provides the best distribution. Now suppose we have some additional information about the probability distribution, given by

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

that is, it gives us the value of the m population moments where $m < (n-1)$. We still have 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 avoid making use of any information not given to us. For this purpose, we make use of a parametric measure of entropy (2.3) introduced by Parkash and Mukesh [8] and by applying the principle of maximum-entropy, we choose the probability distribution which maximizes this measure subject to the constraints (2.1) and (2.2).

$$S_\alpha = \frac{1}{\alpha} \sum_{i=1}^{n} (1 - p_i^{\alpha p_i}). \tag{5}$$

Consider the Lagrangian function given by

$$L = \frac{1}{\alpha} \sum_{i=1}^{n} (1 - p_i^{\alpha p_i}) - \lambda_0 (\sum_{i=1}^{n} p_i - 1) - \lambda_r (\sum_{i=1}^{n} p_i g_r(x_i) - a_r), r = 1, 2, \ldots, m. \tag{6}$$

Differentiating (2.4) with respect to p_i and setting equal to zero, we get

$$p_i^{\alpha p_i}(1 + \log p_i) = -[\lambda_0 + \lambda_1 g_1(x_i) + \lambda_2 g_2(x_i) + \cdots + \lambda_m g_m(x_i)]. \tag{7}$$

This expression will correspond to the maximum entropy probability distribution.

Now, if we suppose that the probability distribution before the moments are prescribed, is given by $q_1, q_2, ..., q_n$, then we shall choose $p_1, p_2, ..., p_n$ in such a way that this probability distribution is as close to $q_1, q_2, ..., q_n$ as possible and at the same time satisfies the given constraints. For this purpose, we again consider a parametric measure of cross-entropy (2.6) developed by Parkash and Mukesh [8] and minimize this measure of cross-entropy under the set of constraints (2.1) and (2.2).

$$I_\alpha = -\frac{1}{\alpha}\sum_{i=1}^{n} q_i[1 - (\frac{p_i}{q_i})^{\alpha\frac{p_i}{q_i}}]. \tag{8}$$

Proceeding as above, we get the minimum cross entropy probability distribution given by

$$(\frac{p_i}{q_i})^{\alpha\frac{p_i}{q_i}}[1 + \log(\frac{p_i}{q_i})] = \lambda_0 + \lambda_1 g_1(x_i) + \lambda_2 g_2(x_i) + \cdots + \lambda_m g_m(x_i). \tag{9}$$

Next, we use these optimization principles to establish the relation between information measures and chi-square statistic.

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

$$\Delta S_\alpha = (S_\alpha)_{max} - S_\alpha. \tag{10}$$

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.

Now

$$\begin{aligned}\Delta S_\alpha &= (S_\alpha)_{max} - S_\alpha \\ &= \frac{1}{\alpha}\sum_{i=1}^{n}(p_i^{\alpha p_i} - p_{i0}^{\alpha p_{i0}}).\end{aligned}$$

Taking limit as $\alpha \to 0$, we get

$$\lim_{\alpha\to 0}\Delta S_\alpha = \lim_{\alpha\to 0}\frac{1}{\alpha}\sum_{i=1}^{n}(p_i^{\alpha p_i} - p_{i0}^{\alpha p_{i0}})$$

$$
\begin{aligned}
&= \lim_{\alpha \to 0} \sum_{i=1}^{n} (p_i p_i^{\alpha p_i} \log p_i - p_{i0} p_{i0}^{\alpha p_{i0}} \log p_{i0}) \\
&= \lim_{\alpha \to 0} \sum_{i=1}^{n} [p_i p_i^{\alpha p_i} \log p_i - p_{i0} p_{i0}^{\alpha p_{i0}} \log p_{i0} + p_{i0}^{\alpha p_{i0}} (p_i - p_{i0})] \\
&= \lim_{\alpha \to 0} \sum_{i=1}^{n} [p_i p_i^{\alpha p_i} \log p_i + p_{i0}^{\alpha p_{i0}} p_i - p_{i0} p_{i0}^{\alpha p_{i0}} (1 + \log p_{i0}) \\
&\quad + p_i \log p_{i0} (p_{i0}^{\alpha p_{i0}} - 1)] \\
&= \lim_{\alpha \to 0} \sum_{i=1}^{n} [p_i p_i^{\alpha p_i} \log p_i + (p_i - p_{i0}) p_{i0}^{\alpha p_{i0}} (1 + \log p_{i0}) - p_i \log p_{i0}] \\
&= \lim_{\alpha \to 0} \sum_{i=1}^{n} [p_i p_i^{\alpha p_i} \log p_i - (p_i - p_{i0})(\lambda_0 + \lambda_1 g_1(x_i) + \cdots + \lambda_m g_m(x_i)) \\
&\quad - p_i \log p_{i0}] \\
&= - \sum_{i=1}^{n} p_i \log (1 + \frac{p_{i0} - p_i}{p_i}) \\
&= - \sum_{i=1}^{n} p_i [\frac{p_{i0} - p_i}{p_i} - \frac{(p_{i0} - p_i)^2}{2 p_i^2} + \frac{(p_{i0} - p_i)^3}{3 p_i^3} - \cdots] \\
&= \frac{1}{2} \sum_{i=1}^{n} \frac{(p_{i0} - p_i)^2}{p_{i0}} + \frac{1}{6} \sum_{i=1}^{n} \frac{(p_{i0} - p_i)^3}{p_{i0}^2} + \cdots .
\end{aligned}
$$

Thus up to first approximation, we have

$$
\lim_{\alpha \to 0} 2N \Delta S_\alpha = \sum_{i=1}^{n} \frac{(N p_{i0} - N p_i)^2}{N p_{i0}} = \chi^2. \tag{11}
$$

Now Np_i are the observed frequencies and Np_{i0} are the expected frequencies. Again, since there are $m+1$ constraints, the degree of freedom is $n-m-1$. Thus, equation (2.9) gives that $\lim_{\alpha \to 0} 2N \Delta S_\alpha$ is distributed asymptotically as Chi-square with $n-m-1$ degrees of freedom.
Again, let

$$
\begin{aligned}
\Delta I_\alpha &= I_\alpha - (I_\alpha)_{min} \\
&= \frac{1}{\alpha} \sum_{i=1}^{n} q_i [(\frac{p_i}{q_i})^{\alpha \frac{p_i}{q_i}} - (\frac{p_{i0}}{q_i})^{\alpha \frac{p_{i0}}{q_i}}].
\end{aligned}
$$

Taking limit as $\alpha \to 0$, we get

$$\begin{aligned}
\lim_{\alpha\to 0}\Delta I_\alpha &= \lim_{\alpha\to 0}\frac{1}{\alpha}\sum_{i=1}^{n} q_i[(\frac{p_i}{q_i})^{\alpha\frac{p_i}{q_i}} - (\frac{p_{i0}}{q_i})^{\alpha\frac{p_{i0}}{q_i}}] \\
&= \lim_{\alpha\to 0}\sum_{i=1}^{n}[p_i(\frac{p_i}{q_i})^{\alpha\frac{p_i}{q_i}}\log\frac{p_i}{q_i} - p_{i0}(\frac{p_{i0}}{q_i})^{\alpha\frac{p_{i0}}{q_i}}\log\frac{p_{i0}}{q_i}] \\
&= \lim_{\alpha\to 0}\sum_{i=1}^{n}[p_i(\frac{p_i}{q_i})^{\alpha\frac{p_i}{q_i}}\log\frac{p_i}{q_i} - p_{i0}(\frac{p_{i0}}{q_i})^{\alpha\frac{p_{i0}}{q_i}}\log\frac{p_{i0}}{q_i} + (\frac{p_{i0}}{q_i})^{\alpha\frac{p_{i0}}{q_i}}(p_i - p_{i0})] \\
&= \lim_{\alpha\to 0}\sum_{i=1}^{n}[p_i(\frac{p_i}{q_i})^{\alpha\frac{p_i}{q_i}}\log\frac{p_i}{q_i} - p_{i0}(\frac{p_{i0}}{q_i})^{\alpha\frac{p_{i0}}{q_i}}(1+\log\frac{p_{i0}}{q_i}) + (\frac{p_{i0}}{q_i})^{\alpha\frac{p_{i0}}{q_i}}p_i \\
&\quad + p_i\log\frac{p_{i0}}{q_i}((\frac{p_{i0}}{q_i})^{\alpha\frac{p_{i0}}{q_i}} - 1)] \\
&= \lim_{\alpha\to 0}\sum_{i=1}^{n}[p_i(\frac{p_i}{q_i})^{\alpha\frac{p_i}{q_i}}\log\frac{p_i}{q_i} + (\frac{p_{i0}}{q_i})^{\alpha\frac{p_{i0}}{q_i}}(1+\log\frac{p_{i0}}{q_i})(p_i - p_{i0}) - p_i\log\frac{p_{i0}}{q_i}] \\
&= \lim_{\alpha\to 0}\sum_{i=1}^{n}[p_i(\frac{p_i}{q_i})^{\alpha\frac{p_i}{q_i}}\log\frac{p_i}{q_i} + (p_i - p_{i0})(\lambda_0 + \lambda_1 g_1(x_i) + \cdots + \lambda_m g_m(x_i)) \\
&\quad - p_i\log\frac{p_{i0}}{q_i}] \\
&= \sum_{i=1}^{n}(p_i\log\frac{p_i}{q_i} - p_i\log\frac{p_{i0}}{q_i}) \\
&= -\sum_{i=1}^{n} p_i\log\frac{p_{i0}}{p_i} \\
&= \frac{1}{2}\sum_{i=1}^{n}\frac{(p_{i0}-p_i)^2}{p_{i0}} + \frac{1}{6}\sum_{i=1}^{n}\frac{(p_{i0}-p_i)^3}{p_{i0}^2} + \cdots .
\end{aligned}$$

Thus up to first approximation, we have

$$\lim_{\alpha\to 0} 2N\Delta I_\alpha = \sum_{i=1}^{n}\frac{(Np_{i0} - Np_i)^2}{Np_{i0}} = \chi^2. \tag{12}$$

Thus, equation (2.10) gives that $\lim_{\alpha\to 0} 2N\Delta I_\alpha$ is distributed asymptotically as Chi-square with $n - m - 1$ degrees of freedom.

Acknowledgments

The authors are thankful to the University Grants Commission, New Delhi, for providing financial assistance for the preparation of the manuscript.

References

[1] Z.I. Botey, D.P. Kroese, Generalized cross entropy method, with applications to probability density estimation, *Methodol. Comput. Appl. Probab.*, **13** (2011), 1-27.

[2] A. Cichocki, S. Amari, Families of Alpha-Beta- and Gamma-divergences: Flexible and robust measures of similarities, *Entropy*, **12** (2010), 1532-1568.

[3] S. Eguchi, S. Kato, Entropy and divergence associated with power function and the statistical application, *Entropy*, **12** (2010), 262-274.

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

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

[6] V. Kumar, H.C. Taneja, Some characterization results on generalized cumulative residual entropy measure, *Statist. Probab. Lett.*, **81** (2011), 1072-1077.

[7] O. Parkash, Mukesh, Two new symmetric divergence measures and information inequalities, *International Journal of Mathematics and Applications*, **4** (2011), 165-179.

[8] O. Parkash, Mukesh, New generalized parametric measures of entropy and cross entropy, *American Journal of Mathematics and Sciences*, **1** (2012), 91-96.

[9] O. Parkash, Mukesh, Development of optimizational principle for minimizing risk in portfolio analysis, *Global and Stochastic Analysis: An International Journal* (2012).

[10] C.R. Rao, Entropy and cross entropy: characterizations and applications, *The legacy of Alladi Ramakrishnan in the mathematical sciences, Springer, New York* (2010), 359-367.

[11] A. Renyi, On measures of entropy and information, Proc. 4th Ber. Symp. Math. Stat. and Prob., **1** (1961), 547-561.

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