Information theoretic approach in Parameter Estimation

Sandeep Kumar[1], Parmil Kumar[2], Mamta Khajuria[3], Ameena Rajput[4]

Department of Statistics, University of Jammu, Jammu, Jammu and Kashmir, 180006, INDIA.

Corresponding Addresses:

[1]sandeep06sharma@gmail.com , [2]parmil@yahoo.com ,[3]khajuriamamta@gmail.com , [4]ameenarajput@gmail.com

Research Article

Abstract: Let $f(x,\theta)$ be the probability density function of a random variable X, where functional form of pdf is known except for the parameterθ. This parameter θ can be a scalar or a vector. One of the most important tasks in statistical inference is of estimating θ on basis of a random sample $(x_1, x_2, \ldots, x_n)$ drawn from the population. The traditional methods of parameter estimation are methods of moments, least squares, minimum chi-square, maximum likelihood, minimum distance and recent one called method of probability weighted moment due to Greenwood *et al* [4]. Amongst all methods, Fisher [3] method of maximum likelihood is widely accepted and is considered as one of the best method for parameter estimation. Akaike [1] work paved the way for the information theoretic approach in parameter estimation. Lind and Solana [8] method is based on the principle of least information. Kapur [6] compared the Gauss' method of estimation with a method based on the principle of maximum entropy. In the present communication we have used Parameter estimation methods using entropy optimization principles and compare these with classical methods such as method of moments and method of m.l.e. The basic principle is that, subject to the information available we should choose θ in such a way that the entropy is as large as possible or the distribution as nearly uniform as possible. We have also derived some parameter estimation methods from entropy optimization principles, while their relation among methods of parameter estimation is also discussed. Further, the asymptotic behaviour of the estimator is also studied for exponential and geometric distribution.

1. Introduction

Let f(x, θ) be the probability density function of a random variable X, where functional form of probability density function is known except for the parameter θ. This parameter θ can be a scalar or a vector quantity. One of the most important task in statistical inference is of estimating the parameter θ on basis of a random sample (x_1, x_2,…,x_n) drawn from the population. The most commonly used traditional methods of parameter estimation are: methods of moments, least squares, minimum chi-square, maximum likelihood, minimum distance and recent one called method of probability weighted moment due to Greenwood *et al* [4]. Amongst all these methods, Fisher [3] method of maximum likelihood is widely accepted, often used and is considered as one of the best method for parameter estimation. But with the growth of information theoretic methods in Statistics, efforts were made by researchers in using the information theory in estimating the parameters and other problems.

Akaike [1] work paved the way for the information theoretic approach in parameter estimation. This paper gave the direction to researchers not only to estimate parameters but also of the model building. Further development took place for estimation when the information is not complete. Lind and Solana [8] method is based on the principle of least information. Kapur [6] compared the Gauss' method of estimation with a method based on the principle of maximum entropy. In this paper, we present a critical appraisal of parameter estimation methods using entropy optimization principles and compare these with classical methods such as method of moments and method of maximum likelihood. The basic principle is that, subject to the information available we should choose θ in such a way that the entropy is as large as possible or the distribution as nearly uniform as possible. In section 2, we discuss the problem of parameter estimation using maximum entropy principle. In section 3, we derive some parameter estimation methods from entropy optimization principles, while their relation among methods of parameter estimation is discussed in section 4. In section 5, we discuss method of parameter estimation using entropy optimization principle when population proportions are given and the asymptotic behaviour of the estimator is also studied for exponential and geometric distribution.

2. Maximum Entropy Principle in Parameter Estimation

In this section, we shall discuss the problem of parameter estimation using entropy optimization principle when along with the known form of density function, a random sample from the population is also given. Let us consider f(x, θ) as the given functional form of probability density estimation and we have to estimate the parameter θ for a given random sample x_1,x_2,..., x_n from the population. Fisher [3] suggested the method of maximum likelihood

i.e. θ should be chosen such that it maximizes the likelihood function

$$L(x,\theta) = \prod_{i=1}^{n} f(x_i, \theta) \quad (2.1)$$

$$\text{or } \log L(x, \theta) = \sum_{i=1}^{n} \log f(x_i, \theta) \quad (2.2)$$

Now a probability distribution can be formed such that

$$p_i = \frac{f(x_i,\theta)}{\sum_{i=1}^{n} f(x_i,\theta)}, \quad i = 1, 2,...,n \quad (2.3)$$

Where $f(x_i, \theta)$ is the value of pdf at $X = x_i$. For making p_i's as equal as possible, we choose parameter θ such that it maximizes Burg's[2] measure of entropy for this distribution. However, it may be noted that we can use any measure of uncertainty. Burg's entropy measure for probability distribution ($p_1, p_2, ..., p_n; p_i > 0; \sum_{i=1}^{n} p_i = 1$) is given by

$$H(P) = \sum_{i=1}^{n} \log p_i \quad (2.4)$$

Substituting (2.3) in (2.4), we have

$$H(P) = \sum_{i=1}^{n} \log \frac{f(x_i,\theta)}{\sum_{i=1}^{n} f(x_i,\theta)} \quad (2.5)$$

For maximizing (2.5) w.r.t. θ, we put the first derivative of (2.4) w.r.t.θ equal to zero and thus we get

$$\sum_{i=1}^{n} \left(\frac{1}{f(x_i,\theta)} \cdot \frac{\partial f(x_i,\theta)}{\partial \theta}\right) - n \frac{\sum_{i=1}^{n} \frac{\partial f(x_i,\theta)}{\partial \theta}}{\sum_{i=1}^{n} f(x_i,\theta)} = 0 \quad (2.6)$$

But Fisher's method of maximum likelihood requires to solve

$$\sum_{i=1}^{n} \left(\frac{1}{f(x_i,\theta)} \cdot \frac{\partial f(x_i,\theta)}{\partial \theta} \right) = 0$$

Since $\sum_{i=1}^{n} f(x_i,\theta)$ is not independent of θ, therefore (2.5) and (2.6) will give different estimates of θ.

It is worth mentioning here that $f(x_1,\theta), f(x_2,\theta),..., f(x_n,\theta)$ are not probabilities. Actually, these are the values of pdf at $(x_1, x_2, \ldots, x_n)$. Their sum is not necessarily unity or independent of θ as $x_1, x_2,..., x_n$ represents only a random sample and not all the values which the variate X can take.

3. Principles of Entropy Optimization, Maximum Likelihood and Minimum Chi-Square

In this section, we discuss the conventional estimation methods vis-a-vis entropy optimization principle.

Principle of Maximum Likelihood:

Let $(x_1, x_2, \ldots, x_n)$ be a random sample from a population with pdf $f(x,\theta)$. We choose or estimate parameter θ in terms of the sample values such that it maximizes likelihood function. But according to Maximum Entropy Principle, we choose the value of θ such that the uncertainty that remains after the sample values are known is as large as possible. Or, we can say that the entropy of the sample itself has to be a minimum. Thus, the sample entropy is given by

$$H_S = -\sum_{I=1}^{n} f(x_i,\theta) \log f(x_i,\theta) \quad (3.1)$$

$$= -\sum_{i=1}^{n} p_i \log p_i$$

$$H_S = -\frac{1}{n}[\log f(x_1,\theta) + \log f(x_2,\theta) + \cdots + \log f(x_n,\theta)]$$

$$= -\frac{1}{n}\left[\log L(x,\theta)\right] \quad (3.2)$$

where $L(x,\theta)$ is the maximum likelihood function given by (2.1).

Thus, we choose θ such that it minimizes the entropy of the sample or maximizes the likelihood function. It implies that maximum entropy principle leads to the principle of maximum likelihood.

Now let us consider $\phi(x,\theta)$ as the cumulative density function of the second distribution in case of minimum cross entropy principle. We shall choose θ such that for this value of θ the distribution function $f(x, \theta)$ is as close as possible to the distribution function determined by the random sample $x_1, x_2,..., x_n$.

Thus, Minimum Discrimination Information Statistic based on Kullback Leibler [7] measure is

$$D(\phi' : f) = \sum_{i=1}^{n} \phi'(x_i,\theta) \log \frac{\phi'(x_i,\theta)}{f(x_i,\theta)}$$

$$= \sum_{i=1}^{n} \phi'(x_i,\theta) \log \phi'(x_i,\theta) - \sum_{i=1}^{n} \log f(x_i,\theta) d\phi(x_i,\theta) \quad (3.3)$$

Equation (3.3.) attains minimum when its second part is maximum. It means, we choose θ which can maximize

$$\sum_{i=1}^{n} \log f(x_i,\theta) d\phi(x_i,\theta)$$

$$= \frac{1}{n}[\log f(x_1,\theta) + \log f(x_2,\theta) + \cdots + \log f(x_n,\theta)]$$

$$= \frac{1}{n}\left[\log L(x,\theta)\right] \quad (3.4)$$

Hence, we choose θ such as to maximize $L(x, \theta)$. Thus, both Maximum Entropy and Minimum Cross Entropy Principles lead to Maximum Likelihood Principle.

Principle of Minimum Chi-square:

Let us consider that there are n classes and $Np_1, Np_2, ..., Np_n$ be the expected frequencies on the basis of parameter θ in these classes, where N is total frequency. Further, we consider that

$Nq_{1,}Nq_2,.....,Nq_{n,}$ are the observed frequencies in these n classes. Then we choose θ so as to minimize divergence measure D (P:Q) or D (Q:P).

Let qi = pi + ε_i , where ε_i is very small

Then $\sum_{i=1}^{n} \varepsilon_i = 0$, since $\sum_{i=1}^{n} p_i = \sum_{i=1}^{n} q_i = 1$

We have, D (P: Q) = $\sum_{i=1}^{n} p_i \log \frac{p_i}{q_i}$ (3.5)

$\cong \frac{1}{2}\sum_{i=1}^{n} \frac{\varepsilon_i^2}{p_i} = \frac{1}{2}\sum_{i=1}^{n} \frac{(q_i - p_i)^2}{p_i}$ (3.6)

Next, similarly we have

$$D(Q:P) \cong \frac{1}{2}\sum_{i=1}^{n} \frac{(q_i - p_i)^2}{q_i} \quad (3.7)$$

It may be pointed here that (3.6) corresponds to modified chi-square while (3.7) is chi-square statistic. Thus, from (3.6) and (3.7) we can infer that θ is chosen to minimize either $\frac{1}{2}\sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i}$ or $\frac{1}{2}\sum_{i=1}^{n} \frac{(O_i - E_i)^2}{O_i}$, where O_i and E_i are observed and expected frequencies in the ith class and E_i's are function of θ.

Fisher's Measure of Information (FMI)

Let f(x, θ) = f and f (x, θ+Δθ) = g, be the two density functions, then divergence measure of f from g is given by

$$D(f:g) = \int_X f \log \frac{f}{g} dx$$

$$= -\int_X f(x,\theta) \log(1 + \frac{f(x,\theta+\Delta\theta) - f(x,\theta)}{f(x,\theta)}) dx$$

Since $\Delta\theta \to 0$,we have

$$D(f:g) = -\int_X f(x,\theta) \log(1 + \frac{\partial f(x,\theta)}{\partial\theta} \frac{\Delta\theta}{f(x,\theta)}) dx$$

$$= -\int_x f(x,\theta) \left[\frac{\partial f(x,\theta)}{\partial\theta} \frac{\Delta\theta}{f(x,\theta)} - \left(\frac{\partial f(x,\theta)}{\partial\theta} \frac{\Delta\theta}{f(x,\theta)} \right)^2 + \cdots \right] dx \quad = (3.8)$$

Since $\int_X$ f (x, θ) dx = 1, therefore

$$\int_X \frac{\partial f}{\partial\theta} dx = 0 \text{ and } \int_X \frac{\partial^2 f}{\partial\theta^2} dx = 0 \quad (3.9)$$

(3.8) and (3.9) together gives

$$\text{D (f: g)} = \frac{1}{2}(\Delta\theta)^2 \int_X \frac{1}{f(x,\theta)} \left(\frac{\partial f(x,\theta)}{\partial\theta} \right)^2 dx + \ldots\ldots\ldots \quad (3.10)$$

$\int_X \frac{1}{f(x,\theta)} \left(\frac{\partial f(x,\theta)}{\partial\theta} \right)^2$ in (3.10) is called Fisher's information measure. It can be noted Fisher's Measure of Information measures the power of discrimination or divergence between two density functions $f(x,\theta)$ and $f(x,\theta+\Delta\theta)$ .Thus, greater the value of FMI, greater is the power of discrimination or it can be said that it gives us more information about θ.

Fisher's Measure of Information is different in many aspects from Shannon's measure of information and Kullback-Leibler's measure of divergence. Shannon's measure of information gives us information about the probability density functions while FMI gives information about the estimators of population parameters. When interval is finite FMI measures the directed divergence of $f(x,\theta)$ from $f(x,\theta+\Delta\theta)$,while Shannon's measure gives the directed divergence of f (x, θ) from uniform density function.

Fisher's Measure of Information gives directed divergence of f (x, θ) from density function depending on both f and q, while Shannon's measure gives the directed divergence of f(x, θ) from a density function which is independent of both f and q.

The Kullback-Leibler measure of directed divergence can discriminate between any two density functions f (x, θ) and g(x, θ) while FMI discriminate between f (x, θ) and $f(x,\theta+\Delta\theta)$ only. Thus, these measures have different purposes, while deciding the relative merits of information measures difficulty arises when the problems of discriminate are viewed in isolation. In generalized model, these measures are considered in relation with the probability distribution and their moment.

4. Equivalence of classical and information theoretic methods of parameter estimation

In this section, we have studied the relations between traditional and information theoretic methods of parameter estimation and observe that in most of cases these are equivalent.

Entropy optimization Principle and Laplace's principle of insufficient reasoning

If the constraints are absent in Jaynes', Maximum Entropy Principle (MEP), and then maximization of uncertainty gives the uniform distribution. Thus, the Laplace principle is a special case of MEP. However, Hadgiswas [5] has shown that the MEP and the MDI principles can be deduced from the principle of insufficient reasoning and thus, MEP and MDI can be regarded as the special case of Laplace's principle, while Laplace's principle can be regarded as a particular case of MDI principle when there are no constraints and the prior distribution is uniform.

Minimum discrimination Information and Maximum Likelihood principle

In section 4.3, a correspondence between the MDI and Fisher's maximum likelihood principle has been

established. Suppose we are given g(x) then we find f(x) which minimizes

$$D(f:g)=\int_X f(x)\log\frac{f(x)}{g(x)}dx$$

$$=\int_X f(x)\log f(x)\,dx-\int_X f(x)\log g(x)\,dx \qquad (4.1)$$

and satisfies the given constraints or we may be given f(x) and have to find g(x) so that we have to maximize

$$\int_X (\log g(x))\, f(x)\,dx=\int_X \log g(x)\,dF(x), \qquad (4.2)$$

where F(x) is the cumulative distribution function of X. In section 3 we have shown that maximization of (4.2) correspond to maximization of the likelihood function. Thus, Maximum Likelihood Principle can be regarded as a special case of MDI principle.

Entropy Optimization principle and Guiasu's principle of Minimum Interdependence (PMI)

If the probability distributions of the individual random variables are included in the set of constraints, as the marginal probability distributions of the joint probability distribution, the PMI is equivalent to the MEP. PMI is also a particular case of Kullback's MDI principle if a priori joint probability density function is the independent product density of n individual variables.

5. Estimation of parameter when interval proportions are given

In this section, we discuss the problem of parameter estimation in case proportions in different intervals are given.

Let us consider a random variate X over the interval [a,b] and let the random sample be arranged in order as

$$a=x_0<x_1<x_2<\ldots<x_i<x_{i+1}<\ldots<x_n<x_{n+1}=b \qquad (5.1)$$

So that the interval [a, b] is divided into (n +1) subintervals and $Q_0, Q_1, \ldots, Q_n$ are the given proportions of the population in these (n + 1) subintervals.

Let us define a probability function over subinterval (x_i, x_{i+1}) as

$$P_i=\int_{x_i}^{x_{i+1}} f(x,\theta)\,dx, \qquad i=0,1,2,..,n \qquad (5.2)$$

where θ is the population parameter.

Thus, $(P_0, P_1, ..., P_n)$ gives us a probability distribution depending on θ. Now, we have to choose parameter θ such that $P_0, P_1,..., P_n$ are as close as possible to given $Q_0, Q_1,..., Q_n$. This can be achieved by minimizing the measure of cross entropy or directed divergence. We can make use of any measure of cross entropy that gives rise to a convex function of θ. But here, we minimize the Kullback Leibler measure of cross entropy,

$$D(Q:P)=\sum_{i=0}^{n}Q_i\log\frac{Q_i}{P_i}=\sum_{i=0}^{n}Q_i\log Q_i-\sum_{i=0}^{n}Q_i\log P_i \qquad (5.3)$$

Minimization of (5.3) is same as maximization of $\sum_{i=0}^{n}Q_i\log P_i$. So, we have to maximize

$$\sum_{i=0}^{n}Q_i\log P_i=\int_{x_i}^{x_{i+1}} Q_i\log\int_{x_i}^{x_{i+1}} f(x_i,\theta)\,dx \qquad (5.4)$$

This principle have wide applications in estimating parameters when interval propertions are given to us, e.g. proportions of students in different intervals of marks obtained, proportion of failed equipments in different intervals of time etc.

Let us consider the case when $f(x_i,\theta)$, functional form of distribution is exponentially distributed with unknown parameter θ. Then, (4.5.4) reduces to maximize

$$\phi=\sum_{i=0}^{n}Q_i\log\int_{x_i}^{x_{i+1}}\theta e^{-x\theta}dx=\sum_{i=0}^{n}Q_i\log(e^{-x_{i+1}\theta}+e^{-x_i\theta}) \qquad (5.5)$$

The above principle is illustrated in the following example having randomly generated population data. We have simulated the results for different sizes of the random samples.

Example: Let us consider a randomly generated population of size 50 (from exponential distributed with mean = 20) with interval proportions as

Intervals:	0-10	10-20	20-30	30-40	40-50	>75
Frequency:	19	13	4	4	7	3
Q_i = Proportion:	0.38	0.26	0.08	0.08	0.14	0.06

Here

$x_0=0$, $x_1=10$, $x_2=20$, $x_3=30$, $x_4=40$, $x_5=60$, $x_6=\infty$, we choose θ which maximizes (4.5.5) i.e.

$$\phi(\theta)=\sum_{i=0}^{n}Q_i\log(e^{-x_{i+1}\theta}+e^{-x_i\theta})$$

$$=0.38\log(1-e^{-10\theta})+0.26\log(e^{-10\theta}-e^{-20\theta})+0.08(e^{-20\theta}-e^{-30\theta})+0.08(e^{-30\theta}-e^{-40\theta})+0.14(e^{-40\theta}-e^{-50\theta})+0.06\,e^{-50\theta}$$

$$=-15.2\theta+0.94\log(1-e^{-10\theta}) \qquad (5.6)$$

To maximize (5.6), differentiate it w.r.t. θ and put the resultant form equal to zero, we get

$$\phi'(\theta)=-15.2+\frac{0.94\times10e^{-10\theta}}{1-e^{-10\theta}}=0$$

$$\hat{\theta}=\frac{1}{10}\log\frac{24.6}{15.2}$$

$$mean=\frac{1}{\hat{\theta}}\cong 20.77$$

The estimated value of the parameter is quite close to the population parameter value i.e. we have small bias.

Further, we can study the asymptotic behaviour of the estimator.

Table 1: Exponential Distribution (Mean=20)

Sample size	MLE	Estimates obtained by MEP	Bias
30	21	20.23	-0.23
200	21.55	21.414	-1.414
1000	21.8	20.18	-0.18
10000	21.492	19.84	0.16

Table 2: Geometric Distribution with p=0.2

Sample size	MLE	Estimates obtained by MEP	Bias
30	0.1668	0.1955	0.0045
200	0.1765	0.1933	0.0067
1000	0.1677	0.2177	-0.0177
10000	0.1690	0.2207	-0.0207

Fig.1 and Fig.2 shows the graph between the sample size and the estimates obtained by MLE, MEP and bias for geometric and exponential distribution respectively.

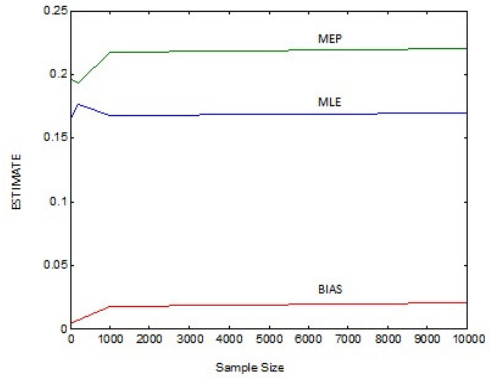

Figure 1

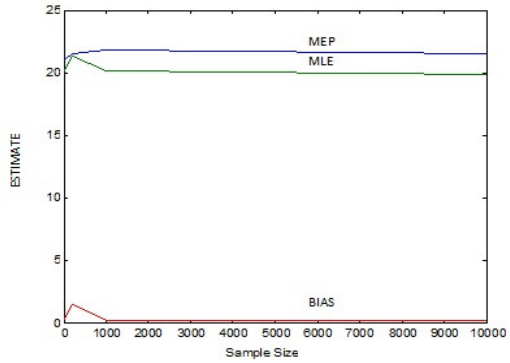

Figure 2

References

1. H. Akaike, Information-theoretical considerations on estimation problems. *Information and Control,* 19(3) (1971), 181-194.
2. J.P. Burg, The relationship between maximum entropy spectra and maximum likelihood spectra. In D.G.Childers, Editor, Modern *Spectral Analysis*. (1972). 130-131.
3. R.A. Fisher, On the mathematical foundations of theoretical Statistics. *Phil. Trans. Roy. Soc*. 222(A), (1921), 309-368.
4. A. J. Greenwood, N. C. Matalas, J.R.Wallis, Probability weighted moments; Definition and relation to parameters of several distributions expressible in reversible form. *Water Resources Researh*.15 (5), (1979), 1049-1055.
5. N.Hadgiwas, The maximum entropy principle as a consequence of the principle of Laplace . *J. Stat. Phy*.26, (1981), 807-815
6. J.N. Kapur, Maximum Entropy Models in Science in Engineering. *Wiley Eastern*, New Delhi. 1989.
7. S.Kullback, R.A. Leibler, On Information and Sufficiency. *Ann. Math. Stat*. 22, (1951),79-86
8. N.C. Lind, Solana, Cross Entropy Estimation of Random Variables with Fractile constraints. Paper no. 11, (1988), Institute for Risk Research, University of Waterloo, Canada.
9. C.E. Shannon, A Mathematical Theory of Communication. *Bell System Tech*. *J*.27, (1948), 379-423.