연세경제연구 제Ⅲ권, 제2호
Yonsei Economic Studies, Vol. Ⅲ, No. 2 (September 1996)

Fisher Information in Statistical Theory and Models

Sangun Park

We summarize some topics in the present statistical theory and discuss over the role of the Fisher information in those topics. We first study some measures of information and their roles in statistics. Then we look over the data reduction and the loss of information due to the reduction. We interpret some present statistical arguments on point and interval estimation in terms of the Fisher information. We further comment on some statistical models related with the Fisher information.

Ⅰ. Introduction

Suppose that X is a random variable whose probability density function (p.d.f.) is $f(x, \theta)$ where θ is a scalar parameter. There are lots of information measures defined by some statisticians. While the Fisher information measure is relevant with the parametric probability approach, while the entropy measure is relevant with the Bayesian approach. The Fisher information in X about θ, which is the most well-known information measure among statisticians, has been defined by the eminent statistician Sir R. A. Fisher as

$$I(\theta) = \int_{-\infty}^{\infty} \left(\frac{\partial}{\partial\theta} \log f(x, \theta) \right)^2 f(x, \theta) dx. \tag{1}$$

It has been discussed in Efron and Johnstone [7] that the transformation $(\partial/\partial\theta)$ log $f(x, \theta) \rightarrow (\partial/\partial\theta)$log $h(x, \theta)$ preserves the length in L^2 norm where $h(x, \theta)$ is the hazard function. For a multiparameter case, the marginal Fisher information about a smooth scalar function $g(\theta)$ has been used in the comparison of estimators in Severini and Wong [23]

Part-time lecturer, Department of Applied Statistics, Yonsei University, Seoul, 120-749, Korea.

and the optimal spacing choice in Park [20].

The entropy measure introduced by Shannon [25] is defined as

$$H = -\int_{-\infty}^{\infty} f(x)\log f(x)dx.$$

The Entropy is actually a measure of uncertainty in terms of concentration of probabilities, so we may take the negative of the entropy measure to use as a measure of information. It has been also noted in Park [19] that the transformation $-\log f(x) \rightarrow 1 - \log h(x)$ also preserves the length in L^1 norm. Shannon's measure has been extended to Kullback and Leibler information (Kullback and Leibler [11]), the mutual information as Bayesian information (Lindley, [13]) and the maximum entropy principle (Jayes, [9]). Their relations have been elegantly studied by Soofi [24].

The Fisher information has received most attention among present measures of information because of its deep relation with the statistical concepts, *sufficiency* and *efficiency*. If we have an independently and identically distributed (i.i.d.) sample of size n, the Fisher information in the sample is $nI(\theta)$. In the data reduction through a set of statistics, we will consider the Fisher information as a criterion of a good reduction. We propose an unbiased and most informative estimator as a point estimator which has the most Fisher information among unbiased estimators. We also give an interpretation of the conditional inference on ancillary statistics in terms of completing the Fisher information. Finally we comment on the role of the Fisher information in some topics like the comparison of the exact and the asymptotic information, the Fisher information in order statistics, robust estimation, sample estimate of the information and the information in some statistical models.

Ⅱ. Data Reduction and Some Basic Concepts

Let $X_1, \cdots, X_n$ be an independent sample of size n from a common probability density function $f(x, \theta)$ where the functional form of f is known and θ is unknown. Suppose that we intend to reduce the dimension of the original sample for some statistical purposes like estimation or data storage. In this case, we need a criterion about how to reduce the data. It seems that we better preserve the information about θ as much as possible after reduction. We simplify our argument by taking the single parameter case.

Let **T** be the functions of the original data, which are assumed to use instead of the original data in answering a variety of questions about the unknown population. Then we might question the efficiency of the chosen statistics. We wish to calculate how much information we lose in interpreting the data through the reduction of data. Rao [22] has considered this problem, but he said that no objective measurement of information is possible and we have hence the difficulty in the formulation of a suitable criteria. In the parametric approach, we can immediately consider the reciprocal of the variance a measure of information as did Fisher. Fisher later introduced the concept likelihood and a measure of information, (1), which is called now the Fisher information. Fisher information has its statistical meaning in that it is closely related with the important statistical concepts sufficiency and efficiency. In comparing two statistics, we usually consider the variance in comparing the degree of uncertainty. However, we can consider a measure of information in comparing the degree of certainty, thus we can consider the Fisher information as a measure of information, not the reciprocal of the variance.

We classify a statistic according to how it contains the Fisher information, and define some basic statistical concepts. The information in the original sample can be divided into the information about θ and the ancillary information. The ancillary information is related with the further configuration of data, though not relevant with the unknown parameter. A statistic, a function of the original data, can be divided as follows.

i) **sufficient statistic** : a statistic which has the full information about θ whether it contains the ancillary information or not.

ii) **complete sufficient statistic** : a sufficient statistic which has no ancillary information.

iii) **ancillary statistic** : a statistic which has no information about θ or which has only ancillary information.

Any non-constant function of a complete sufficient statistic should not have ancillary information, where we usually take the concept of the first order ancillarity. The function may be restricted to a bounded function, then we concern a weaker concept, bounded completeness. The ancillary statistic is usually confined to be a function of minimal sufficient statistics (Cox and Hinkley [5]). This is quite a reasonable restriction, since the purpose of the ancillary statistic is to complete the loss of information. Sufficiency and conditionality principles, which are actually twin concepts, can be understood as reasonable arguments in view of the information oriented definitions in the above. When it is hard to find the ancillary statistic or sufficient statistic, we may consider the local ancillary statistic

(Cox [4]) or the local sufficient statistic (McCullagh [14]) through the score function.

Ⅱ. Unbiased and Most Informative Estimator

We call the process of finding an estimate estimation as conceived by Fisher. Many requirements for a good estimator have been discussed for reducing the class of statistics. Some approaches focused on first two moments by considering the unbiasedness and the minimum variance where the loss function is quadratic. The minimum variance is a good (maybe unanimously) criterion. However, the distribution of a statistic is determined by its all moments not just by first two moments. We here consider the maximum Fisher information criterion, i.e, we prefer one which has more Fisher information between two unbiased estimators. Since the Fisher information can be approximated by the reciprocal of the variance, a uniformly unbiased minimum variance estimator (UMVUE) can be understood as an approximation of the unbiased and most informative estimator. The unbiased and most informative estimator is defined to be one which has the most Fisher information among the class of the unbiased estimators. If the density function of a UMVUE belongs to the natural exponential family, the unbiased and most informative estimator is equal to the UMVUE, since the the variance of the UMVUE achieves the lower bound of the information inequality. Thus we have the natural question whether one unbiased estimator with smaller variance can have more Fisher information than the other unbiased estimator. The answer is yet known.

The minimum requirement for a good estimator is the consistency. Then we choose one which has the most Fisher information among the class of the consistent estimators. We call it a consistent and most informative estimator. When the calculation of the Fisher information is difficult, we may use the asymptotic information if the difference is not more than $o(n)$. If we consider the class of consistent and asymptotically normally distributed (CAN) estimators, the consistent and most informative estimator is equal to the estimator which has the least asymptotic variance.

We may give a restriction on the form of the statistic and find an unbiased and most informative estimator. For example, suppose that the class of statistics is restricted to the class of L-statistics which has received much attention because of their convenience and robustness. We now let T_n be a linear function of order statistics as $c_1X_{(1)}+ \cdots c_n+X_{(n)}$.

In choosing the coefficients, we put first the condition of the unbiasedness ($\sum_{i=1}^{n} c_i = 1$) or consistency. Then we can consider $\max_{c_1 \cdots c_n} I_{T_n}(\theta)$. Since the answer depends on some conditions of coefficient and functional form, we may take the class to be the class of L-statistics whose asymptotic distribution is a normal distribution. The conditions for the coefficients and the functional form have been studied in Stigler ([26], [27]). Then we can use the asymptotic information (the reciprocal of the asymptotic variance in this case) and find the suitable coefficients which give the maximum asymptotic information. Tukey [29] is the first one who studied the information in order statistics with a linear sensitivity measure. He actually used a moment based lower bound of $I_{T_n}(\theta)$ whether he intended to or not (see also Nagaraja [17]). We may approximate the Fisher information with the moment based lower bound or the asymptotic Fisher information. Then the best linear unbiased estimator or the asymptotically best linear estimator can be interpreted as the estimator which maximizes the approximated Fisher information.

There are some convenient general methods which produce a good estimator like maximum likelihood estimator, or minimum quantile distance estimator (LaRiccia and Wehrly [12]). Suppose that a real function $g(x, \theta)$ of x and θ is an estimating function. We may take it to be unbiased such that it has zero mean for all θ in the parameter space Θ Then the estimating function produces an estimator of θ by solving the estimating function, $g(x, \hat{\theta}) = 0$. We will define the most informative estimating function to be one which has the maximum value of $E\left(\frac{\partial}{\partial \theta} \; g(x, \theta) \right)^2 / V(g(x, \theta))$, which is the information in an estimating function defined by Bhapkar (1972).

Ⅳ. Recovery of the Loss of Information

In comparing two confidence intervals for an unknown parameter, the first principle comes from the behaviour of two intervals at the true parameter, i.e., we prefer a shorter interval. This is quite easy principle itself, but we have much controversies over it. The second one comes from the behaviour of two intervals at the alternative, i.e., we prefer an interval which covers the alternative with lower probability. We may take the alternative to be fixed or locally around the true parameter. This is quite related with the theory of testing hypothesis, so may be preferred. In determining the critical region, we may consider unbiased critical region or likelihood based on critical region.

When we construct a confidence interval based on a suitably chosen statistic, we have some loss of information if the statistic is not a sufficient statistic. The density form of the non sufficient statistic does not give the full information. Thus we need to seek a density form which gives the full information. This can be done by considering the conditional distribution. Then what should be the conditioning statistic? The conditioning statistic needs to be one which has the full information about θ jointy with the chosen statistic but does have no information about θ itself. Thus the answer is the ancillary statistic. A similar rule also holds for the definition of the sufficiency. This argument gives the conditional inference on ancillary statistic in terms of completing the Fisher information, while Fisher [8] has first given this idea in terms of transforming the data configuration.

The ancillary statistic alone has no information about θ, but it provides some information about θ, jointly with a statistic which is not a sufficient statistic. We note that two ancillary statistics jointly may have some information about θ. Thus the recovery of lost information can be done by conditioning the ancillary statistic. The lost information can be all recovered when we consider the ancillary statistic. For the location and scale family, the choice of a statistic does not matter as far as it is location and scale equivariant. However, the ancillary statistic may not be unique in some cases; McCullagh [15] consider the Cauchy model which is closed under Mobius transformation and studied the difference in the coverage probabilities due to the nonunique ancillary statistics.

V. Asymptotic Information

Usually it is quite difficult to get the exact distribution of a statistic, so we use its asymptotic distribution. Then the natural question is to compare the exact distribution (moment) and the asymptotic distribution (moment) of the statistic. In the distribution case, the error bound has been studied for the normalized mean of the i.i.d. sample, which is known as Berry-Essen rate. In the variance case, the asymptotic variance is less than or equal to the normalized limit of the exact variance. The conditions for the equality has been studied. The most interesting statistic is a maximum likelihood estimator, but we do not have any formal proof about the equality of both variances, which still remains an open problem in statistics, but Professor R. R. Bahadur personally commented to me that the problem is believed to be true though he does not have any formal proof, but not recommendable to statisticians of any age. Here we have the similar question whether the asymptotic information is equal to the normalized limit of the exact Fisher information. The answer will clarify the the efficiency of using the asymptotic distribution instead of the exact distribution.

Ⅵ. Some Statistical Models

Random censored model : In a random censorship model where the censoring time is random, the partial likelihood part has most Fisher information as claimed by Cox [3] and studied by Efron [6] and Oakes [18]. We can further consider the marginal likelihood of only the censoring indicator by ignoring the survival time of uncensored observations. Then we may have an advantage of pertaining robustness instead of losing some information. We may not lose much information for the usual survival analysis models. Cox's proportional hazard model actually with the assumption of the parent distribution in Lehman alternative, can be solved by the partial likelihood.

Regression model : One approach in robust regression model is to use regression quantile suggested Koenker and Bassett [10]. The regression quantile has a similar interpretation to the sample quantile. Thus the trimming appear in this topic and the trimming portion should be considered in terms of the information.

Type 2 censored model : We have some situations where only a part of order statistics is available. The study about the distribution of the information among order statistics has been started by the eminent statistician J. W. Tukey who used the linear sensitivity as a measure of information. Mehrotra et al. [16] first considered the Fisher information and expressed the Fisher information in order statistics as a linear combination of moments of order statistics for some distributions. Nagaraja [17] studied the relation between the Fisher information and the linear sensitivity of order statistics (see also Park [20]). However, while the recipe for the Fisher information in a set of order statistics is simple, but it has been taken to be a messy problem (Arnold et al. [1]), since it contains an iterated integral to evaluate. Some inherent relations (recurrence relations) in the Fisher information have been derived by Park ([20], [21]), which enables us to get the exact Fisher information in consecutive order statistics for any parametric distribution. The result has its wide applications where a part of order statistics needs to be considered like in life testing and survival analysis, nonparametrics, robust inference on location and scale and regression parameters etc.

Forecasting records : Records can be interpreted as a sequence of smallest or largest order statistics. (see Tryfos [28]). Thus the information tend in the se‹quence will follow the assumed distribution.

◈ *REFERENCES* ◈

1. Arnold, B. C., Balakrishnan, N. and Nagaraja, H. N., *A first Course in Order Statistics*, 1992, New York : Wiley.
2. Bhapkar, V. P., "On a Measure of Efficiency in an Estimating Equation," *Sankhya*, A, 1972, pp. 467~472.
3. Cox, D. R., "Regression Models and Life-tables," *Journal of Royal Statistical Society*, B. 34, 1972, pp. 187~202.
4. ________ , "Local Ancillarity," *Biometrika*, 67, 1980, pp. 279~286.
5. ________ , and Hinkley, D. V., *Theoretical Statistics*, London: Chapman and Hall, 1974.
6. Efron, B., "The Efficiency of Cox's Likelihood Function for Censored Data," *Journal of American Statistical Association*, 72, 1977, pp. 557~565.
7. _______, and Johnstone, I., "Fisher Information in Terms of the Hazard Rate," *Annals of Statistics*, 72, 1990, pp. 557~565.
8. Fisher, R. A., "Two New Properties of Mathematical Likelihood," *Proceedings of Royal Statistical Society*, A. 144, 1934, pp. 285~307.
9. Jayes, E. T., "Information Theory and Statistical Mechanics," *Physical Review*, 106, 1957, pp. 620~630.
10. Koenker, R. and Bassett, G. Jr., "Regression Quantiles," *Econometrica*, 46, 1978, pp. 33~50.
11. Kullback, S. and Leibler, R. A., "On Information and Sufficiency," *Annals of Mathematical Statistics*, 22, 1951, pp. 79~86.
12. LaRiccia, V. N. and Wehrly, T. E., "Asymptotic Properties of a Family of Minimum Quantile Estimators," *Journal of American Statistical Association*, 80, 1985, pp. 742~747

13. Lindley, D. V., "On a Measure of the Information Provided by an Experiment," *Annals of Mathematical Statistics*, 27, 1956, pp. 986~1005.

14. McCullagh, P., "Local Sufficiency," *Biometrika*, 71, 1984, pp. 233~244.

15. __________, "Conditional Inference and Cauchy Model," *Biometrika*, 79, 1992, pp. 247~259.

16. Mehrotra, K. G., Johnson, R. A. and Bhattacharyya, G. K., "Exact Fisher Information for Censored Samples and the Extended Hazard Rate Functions," *Communications in Statistics*, A. 15, 1979, pp. 1493~1510.

17. Nagaraja, H. N., "Tukey's Linear Sensitivity and Order Statistics," *Annals of Institute of Statistics and Mathematics*, 46, 1994, pp. 757~768.

18. Oakes, D., "The Asymptotic Information in Censored Survival Model," *Biometrika*, 64, 1977, pp. 441~448.

19. Park, S., "Sample Entropy of Order Statistics," *Unpublished manuscript*, 1994.

20. _____, "Selection of Order Statistics through the Fisher Information," *Unpublished manuscript*, 1995.

21. _____, "Fisher Information in Order Statistics," *Journal of American Statistical Association*, 91, 1996, pp. 385~390.

22. Rao, C. R., "Criteria of Estimation in Large Samples," *Sankhya*, A. 25, 1963, pp. 189~206.

23. Severini, T. A. and Wong, W. H., "Profile Likelihood and Conditional Parametric Models," *Annals of Statistics*, 20, 1992, pp. 1768~1802.

24. Soofi, E. S., "Capturing the Intangible Concept of Information," *Journal of American Statistical Association*, 89, 1994, pp. 1243~1254.

25. Shannon, C. E., "A Mathematical Theory of Communication," *Bell Technical Journal*, 1948, pp. 279~423.

26. Stigler, S. M., "Linear Functions of Order Statistics," *Annals of Mathematical Statistics*, 40, 1969, pp. 770~788.

27. __________, "Linear Functions of Order Statistics with Smooth Weight Functions," *Annals of Statistics*, 2, 1974, pp. 676~693.

28. Tryfos, P., "Forecasting Records," *Journal of American Statistical Association*, 80, 1985, pp. 46~50.

29. Tukey, J. W., "Which Part of the Sample Contains the Information?" *Proceedings of National Academic Science*, 53, 1964, pp. 127~134.