A note on defining information in random censoring

T. Papaioannou
Department of Statistics and Actuarial Science
University of Piraeus
18534 Piraeus, Greece

Ch. Tsairidis
Department of Mathematics
University of Ioannina
45110 Ioannina, Greece

Summary

In this paper we show that the two ways of defining information in random censoring, the Hollander, Proschan and Sconing approach and the classical or likelihood approach, lead to the same measure of information, if censoring is non-informative.

Keywords: Fisher information; φ-Divergence; Non-informative censoring.

1. Introduction

Information measures can be classified as parametric or Fisher-type and nonparametric or divergence-type [cf. Ferentinos and Papaioannou (1981), Papaioannou (1985), and references therein]. The main representative of the first type is Fisher's measure which is defined for parametric families of distributions, while for the second type, Csiszar's φ-divergence or the Kullback-Leibler measure of information are the main representatives [cf. Csiszar (1963), Ali and Silvey (1966), and Kullback and Leibler (1951)].

In random censoring, where X is the variable of interest (e.g. failure times) and Y the censoring variable, both parametric and nonparametric measures of information can be defined in two ways. The first way is to condition first on Y and then average Y out. The second is to evaluate information based on the likelihood of a single observation, this observation being either the true value X or the value of the censoring Y, if X is censored.

The first approach is known as the Hollander, Proschan and Sconing way of defining measures of information for censored data. It is based primarily on entropy [cf. Hollander *et al.* (1987) and (1990), Baxter (1989), and Stute (1992)]. It provides a formal definition of measures of information and has some nice properties which Hollander *et al.* (1987) call the "acid test", and which are pertinent to censored data. The second is known as the classical or the likelihood approach [cf. Tsairidis *et al.* (1996)] and has formed the basis for developing results in statistical information theory. The two approaches do not lead, in general, to the same value for the measure.

The merits of these two approaches in the context of random censoring have been compared in Tsairidis *et al.* (1996). On the basis of various fundamental properties of the measures, the likelihood approach was preferred. In a given situation, however, and with respect to maximum likelihood estimation in particular, there still exists a dilemma as to which approach to use. Stute (1992), for example, studies the strong consistency of the maximum likelihood estimator in random censoring and for his measure of infomation follows the Hollander, Proschan and Sconing approach. Miller (1981), however, in his classic book uses the likelihood approach. The same is true for Abdushukurov and Kim (1987) who study Cramer-Rao and Bhattacharyya bounds for randomly censored observations. Thus, it is of interest to explore the cases where the two approaches lead to the same value, in view also of the computational effort involved for the evaluation of each measure due to the appearance of c.d.f's in their definitions. The purpose of this note is to clarify further the issues regarding the definition of information in random censoring, and primarily to show that, for both Fisher-type and divergence-type measures of information, the two ways of defining information in random censoring lead to the same measures of information, if censoring is non-informative.

Censoring is called non-informative if the likelihood function can be decomposed into the product of two components: one that involves only parameters for the distribution of the variable X of interest and the other only parameters for the distribution of variable Y or no parameters at all. Parameters in the latter component can be disregarded in making inferences about the distribution of failure times [cf. Kalbfleisch and Prentice (1980), p. 121, Lawless (1982), p. 42, and Wu and Carroll (1988)]. There are other aspects of non-informativeness such as conditional independence between censoring and failure times for a given set of covariates [cf. Slud *et al.* (1988)] or the constant-sum condition [cf. Kalbfleisch and Mackay (1979), and Williams and Lagakos (1977)]. Here we shall assume that Y has a fixed, possibly known, distribution or that it does not depend on unknown parameters. The mode is non-Bayesian.

The statistical and notational setup will be as follows: X will be the variable of interest, for example, the time taken for an event to occur, and Y the censoring variable, independent of X. In right random censoring we observe $Z = \min(X, Y)$ and $\delta = I_{(X \leq Y)}$. If $\delta = 1$ we observe $Z = X$, an uncensored observation. If $\delta = 0$ we observe $Z = Y$, a censored observation. F will be the c.d.f. of X and G the c.d.f. of Y, and $\overline{F}$ and $\overline{G}$ the corresponding survival functions. f and g will be the corresponding p.d.f's.

For Fisher-type measures of information we shall assume that the distribution of X belongs to a parametric family of distributions $\mathcal{M} = \{P_\theta, \theta \in \Theta\}$ with p.d.f. $f(x;\theta) = \frac{dP_\theta}{d\mu}$, relative to a dominating measure μ on the real space $\mathcal{X}$, satisfying some appropriate regularity conditions [cf. Papaioannou (1985)]. $\mathcal{X}$ is nonnegative and the parameter space is an open subset of the Euclidean space R^k. The distribution of Y does not depend on θ. We denote by $p((z,\delta);\theta)$ the p.d.f. of (Z,δ). The distribution of (Z,δ) is given by

$$p((z,\delta);\theta) = \left[f(z;\theta)\overline{G}(z)\right]^{\delta}\left[g(z)\overline{F}(z;\theta)\right]^{1-\delta}, \ \delta = 0,1. \tag{1.1}$$

For divergence-type measures of information we consider two probability measures P_1 and P_2 on the measurable space $(\mathcal{X},\mathcal{T})$, $\mathcal{T}$ a σ-algebra associated with $\mathcal{X}$. Let $f_i(x) = \frac{dP_i}{d\mu}$ and $F_i(x)$, $i = 1,2$, be the p.d.f and c.d.f of X, respectively, where μ is a dominating finite or σ-finite measure on $\mathcal{X}$. For the censoring variable Y we shall assume a p.d.f. $g(y)$ and a c.d.f. $G(y)$ as before, independent of θ. The distribution of Y need not belong to $\mathcal{M}$. For the φ-divergence or Csiszar's measure of information we shall also assume that φ is a convex function satisfying appropriate regularity conditions [cf. Csiszar (1963)]. Again the distribution of (Z,δ), for each (f_i,F_i) is given by (1.1) without θ.

We shall use the notation $I_{(X,Y)}$ and $I_{(Z,\delta)}$ to denote the measures of information defined by Hollander, Proschan and Sconing, and the likelihood approach, respectively.

The result of this note is helpful particularly when one considers the asymptotic distribution of maximum likelihood estimates and wonders whether to use the one or the other approach in defining Fisher's information for the asymptotic variance, each having certain computational advantages.

2. Main results

2.1 Fisher-type measures of information

In view of (1.1), Fisher's measure of information about θ contained in the censored data (Z,δ) and based on the likelihood of a single observation is given by

$$\begin{aligned} I^F_{(Z,\delta)}(\theta) &= \left\| \sum_{\delta=0,1} \int_{-\infty}^{\infty} p((z,\delta);\theta)\left(\tfrac{\partial}{\partial\theta_i}\log p((z,\delta);\theta)\tfrac{\partial}{\partial\theta_j}\log p((z,\delta);\theta)\right)dz \right\|_{k\times k} \\ &= \left\| \int_{-\infty}^{\infty} \tfrac{f_{(i)}(z;\theta)f_{(j)}(z;\theta)}{f(z;\theta)}\overline{G}(z)dz + \int_{-\infty}^{\infty} \tfrac{\overline{F}_{(i)}(z;\theta)\overline{F}_{(j)}(z;\theta)}{\overline{F}(z;\theta)}g(z)dz \right\|_{k\times k}, \end{aligned} \tag{2.1}$$

($\|\cdot\|_{k\times k}$ denotes a $k\times k$ matrix), where

$$f_{(i)}(z;\theta) = \tfrac{\partial f(z;\theta)}{\partial\theta_i}, \quad \overline{F}_{(i)}(z;\theta) = \tfrac{\partial\overline{F}(z;\theta)}{\partial\theta_i}, \quad i,j = 1,2,\ldots,k.$$

If one uses the Hollander, Proschan and Sconing approach, Fisher's amount of information about θ contained in the censored experiment (X, Y) must be defined by

$$I^F_{(X,Y)}(\theta) = \left\| \int_{-\infty}^{\infty} g(y) \left[\int_{-\infty}^{y} \frac{f_{(i)}(x;\theta) f_{(j)}(x;\theta)}{f(x;\theta)} dx + \frac{F_{(i)}(y;\theta) F_{(j)}(y;\theta)}{\overline{F}(y;\theta)} \right] dy \right\|_{k\times k}, \tag{2.2}$$

[cf. Tsairidis *et al.* (1996) and (2001)]. One advantage of the Hollander, Proschan and Sconing definition is that it does not require knowledge in closed form of G (and $\overline{G}$) which in many cases is not available.

If censoring is informative, i.e., the distribution of Y depends on θ, Y provides information about θ, and this should be taken into account in defining measures of information. Thus, it is not appropriate to use formally the Hollander, Proschan and Sconing approach to define Fisher-type measures of information, as in (2.2), since we should also consider the score based on $g(y;\theta)$.

Theorem 2.1 In non-informative right random censoring we have

$$I^F_{(Z,\delta)}(\theta) = I^F_{(X,Y)}(\theta),\ \theta \in \Theta.$$

Proof. Only the case where θ is univariate will be treated. The k-variate case is straightforward. To prove the equivalence of (2.1) and (2.2), it is enough to show that

$$\int_{-\infty}^{\infty} \frac{(f_\theta(x;\theta))^2}{f(x;\theta)} \overline{G}(x) dx = \int_{-\infty}^{\infty} g(y) \left[\int_{-\infty}^{y} \frac{(f_\theta(x;\theta))^2}{f(x;\theta)} dx \right] dy,$$

where differentiation with respect to θ is denoted by $f_\theta(x;\theta) = \frac{\partial f(x;\theta)}{\partial \theta}$. Note also that $\overline{F}_\theta(\cdot;\theta) = -F_\theta(\cdot;\theta)$. We have

$$\begin{aligned}
\int_{-\infty}^{\infty} \tfrac{(f_\theta(x;\theta))^2}{f(x;\theta)} \overline{G}(x) dx &= \int_{-\infty}^{\infty} \tfrac{(f_\theta(x;\theta))^2}{f(x;\theta)} dx - \int_{-\infty}^{\infty} \tfrac{(f_\theta(x;\theta))^2}{f(x;\theta)} G(x) dx \\
&= I^F_X(\theta) - \int_{-\infty}^{\infty} \tfrac{(f_\theta(x;\theta))^2}{f(x;\theta)} \left[\int_{-\infty}^{x} g(y) dy \right] dx \\
&= I^F_X(\theta) - \int_{-\infty}^{\infty} g(y) \left[\int_{y}^{\infty} \tfrac{(f_\theta(x;\theta))^2}{f(x;\theta)} dx \right] dy \\
&= I^F_X(\theta) - \int_{-\infty}^{\infty} g(y) \left[I^F_X(\theta) - \int_{-\infty}^{y} \tfrac{(f_\theta(x;\theta))^2}{f(x;\theta)} dx \right] dy \\
&= \int_{-\infty}^{\infty} g(y) \left[\int_{-\infty}^{y} \tfrac{(f_\theta(x;\theta))^2}{f(x;\theta)} dx \right] dy.
\end{aligned}$$

Thus

$$I^F_{(Z,\delta)}(\theta) = I^F_{(X,Y)}(\theta).$$

Remark 2.1 A weakness of the Hollander, Proschan and Sconing approach is that it does not lead to a natural definition of the observed Fisher information [cf. Efron and Hinkley (1978)] which has certain advantages over the expected Fisher information. For measures derived from the Fisher information matrix which have been proposed as simpler measures of information, such as the trace, the determinant, the eigenvalues etc or even Vajda's measure of information, the result of Theorem 1 remains true [cf. Ferentinos and Papaioannou (1981)].

2.2 Divergence-type measures of information

Csiszar's φ-divergence between f_1 and f_2 based on the randomly censored experiment at hand and the likelihood, is the φ-divergence between $p_1(z,\delta)$ and $p_2(z,\delta)$ which are generated from f_1 and f_2 via (1.1). Then it is defined as

$$\begin{aligned} I^C_{(Z,\delta)}(f_1,f_2) &= \sum_{\delta=0,1}\int_0^\infty p_2(z,\delta)\varphi\left(\tfrac{p_1(z,\delta)}{p_2(z,\delta)}\right)dz \\ &= \int_0^\infty f_2(z)\overline{G}(z)\varphi\left(\tfrac{f_1(z)}{f_2(z)}\right)dz + \int_0^\infty g(z)\overline{F}_2(z)\varphi\left(\tfrac{\overline{F}_1(z)}{\overline{F}_2(z)}\right)dz. \end{aligned}$$

But if one uses the Hollander, Proschan and Sconing approach, Csiszar's φ-divergence between f_1 and f_2 based on the censored experiment (X,Y) must be defined as follows [cf. Tsairidis *et al.* (1996) and (2001)]

$$I^C_{(X,Y)}(f_1,f_2) = \int_{-\infty}^{\infty} g(y)\left[\int_{-\infty}^{y} f_2(x)\varphi\left(\frac{f_1(x)}{f_2(x)}\right)dx + \overline{F}_2(y)\varphi\left(\frac{\overline{F}_1(y)}{\overline{F}_2(y)}\right)\right]dy. \tag{2.3}$$

Again the evaluation of $I^C_{(X,Y)}(f_1,f_2)$ does not require knowledge of $\overline{G}$.

Note however that in the informative case, Csiszar's measure of information may not be formally defined via (2.3), since we could also have two p.d.f's for Y. For this divergence-type measure of information we then have:

Theorem 2.2 In non-informative right random censoring we have

$$I^C_{(Z,\delta)}(f_1,f_2) = I^C_{(X,Y)}(f_1,f_2).$$

Proof. It is enough to show that

$$\int_{-\infty}^{\infty} f_2(x)\overline{G}(x)\varphi\left(\frac{f_1(x)}{f_2(x)}\right)dx = \int_{-\infty}^{\infty} g(y)\left[\int_{-\infty}^{y} f_2(x)\varphi\left(\frac{f_1(x)}{f_2(x)}\right)dx\right]dy.$$

We have

$$
\begin{aligned}
\int_{-\infty}^{\infty} f_2(x)\overline{G}(x)\varphi\left(\frac{f_1(x)}{f_2(x)}\right)dx &= \int_{-\infty}^{\infty} f_2(x)\varphi\left(\frac{f_1(x)}{f_2(x)}\right)dx - \int_{-\infty}^{\infty} f_2(x)G(x)\varphi\left(\frac{f_1(x)}{f_2(x)}\right)dx \\
&= I_X^C(f_1,f_2) - \int_{-\infty}^{\infty}\left[\int_{-\infty}^{x} g(y)dy\right] f_2(x)\varphi\left(\frac{f_1(x)}{f_2(x)}\right)dx \\
&= I_X^C(f_1,f_2) - \int_{-\infty}^{\infty} g(y)\left[\int_{y}^{\infty} f_2(x)\varphi\left(\frac{f_1(x)}{f_2(x)}\right)dx\right]dy \\
&= I_X^C(f_1,f_2) - \int_{-\infty}^{\infty} g(y) I_X^C(f_1,f_2)dy \\
&\qquad + \int_{-\infty}^{\infty} g(y)\left[\int_{-\infty}^{y} f_2(x)\varphi\left(\frac{f_1(x)}{f_2(x)}\right)dx\right]dy \\
&= \int_{-\infty}^{\infty} g(y)\left[\int_{-\infty}^{y} f_2(x)\varphi\left(\frac{f_1(x)}{f_2(x)}\right)dx\right]dy.
\end{aligned}
$$

So

$$I_{(Z,\delta)}^C(f_1,f_2) = I_{(X,Y)}^C(f_1,f_2).$$

3. Conclusion

The results of this note, applicable in non-informative random censoring, are helpful when one considers the asymptotic distribution of maximum likelihood estimators for the parameters of the model and wonders whether to use the one or the other approach in defining Fisher's information for the asymptotic variance. Each approach has certain computational advantages, but both lead to the same value. If the c.d.f. G of the censoring variable is not available in closed form, a case not rare in practice, it may be easier to use (2.2) to compute Fisher's information than (2.1), which requires knowledge of both c.d.f's F and G. The same is true for φ-divergence.

The reader may note that in informative random censoring, i.e., when the distribution of the censoring variable Y depends on the same parameter θ of the model or another parameter η, the equivalence shown above is not true [cf. Tsairidis *et al.* (1996)].

References

Abdushukurov, A. A. and Kim, L. V. (1987). Lower Cramer-Rao and Bhattacharyya bounds for randomly censored observations. *J. Sov. Math*, **38**, 2171-2185.

Ali, S. M. and Silvey, S. D. (1966). A general class of coefficients of divergence of one distribution from another. *Journal of the Royal Statistical Society* B, **28**, 131-142.

Baxter, L. A. (1989). A note on information and censored absolutely continuous random variables. *Statistics and Decisions*, **7**, 193-198.

Csiszar, I. (1963). Eine Informationstheoretische Ungleichung und ihre Anwendung auf den Beweis der Ergodizitat von Markoffschen Ketten. *A Mayar Tudomanyos Academia Matematikai Kutato Intezetenek Kozlemenyei*, **8**, 85-108.

Efron, B. and Hinkley, D. V. (1978). Assessing the accuracy of the maximum likelihood estimator: Observed versus expected Fisher information. *Biometrika*, **65**, 457-487.

Ferentinos, K. and Papaioannou, T. (1981). New parametric measures of information. *Information and Control*, **51**, 193-208.

Hollander, M., Proschan, F. and Sconing, J. (1987). Measuring information in right-censored models. *Naval Research Logistics*, **34**, 669-681.

Hollander, M., Proschan, F. and Sconing, J. (1990). Information, censoring and dependence. *Institute of Mathematical Statistics, Lecture Notes - Monograph Series, Topics in Statistical Dependence*, **16**, 257-268.

Kalbfleisch, J. D. and Mackay, R. J. (1979). On constant-sum models for censored survival data. *Biometrika*, **66**, 87-90.

Kalbfleisch, J. D. and Prentice, R. L. (1980). *The Statistical Analysis of Failure Time Data.* Wiley, New York.

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

Lawless, J. F. (1982). *Statistical Models and Methods for Lifetime Data.* Wiley, New York.

Miller, R. (1981). *Survival Analysis.* Wiley, New York.

Papaioannou, T. (1985). Measures of information. *Encyclopedia of Statistical Sciences, Kotz and Johnson, Eds.* **5**, Wiley, New York, 391-397.

Slud, E. V., Byar, D. P. and Schatzkin, A. (1988). Dependent compenting risks and the latent-failure model. *Biometrics*, **44**, 1203-1205.

Stute, W. (1992). Strong consistency of the MLE under random censoring. *Metrika*, **39**, 257-267.

Tsairidis, Ch., Ferentinos, K. and Papaioannou, T. (1996). Information and randon censoring. *Information Sciences*, **92**, 159-174.

Tsairidis, Ch., Zografos, K., Ferentinos, K. and Papaioannou, T. (2001). Information in quantal response data and random censoring. *Annals of the Institute of Statistical Mathematics* (to appear), 1-24.

Williams, J. S. and Lagakos, S. W. (1977). Models for censored survival analysis: constant-sum and variable-sum models. *Biometrika*, **64**, 215-224.

Wu, M. C. and Carroll, R. J. (1988). Estimation and comparison of changes in the presence of informative right censoring by modeling the censoring process. *Biometrics*, **44**, 175-188.