ON INFORMATION IN STATISTICS

By D. A. S. Fraser

University of Toronto

1. Summary and introduction. The three familiar definitions of statistical information, Fisher (1925), Shannon (1948), and Kullback (1959), are closely tied to asymptotic properties, hypothesis testing, and a general principle that information should be additive. A definition of information is proposed here in the framework of an important kind of statistical model. It has an interpretation for small samples, has several optimum properties, and in most cases is *not* additive with independent observations. The Fisher, Shannon and Kullback informations are aspects of this information.

A variety of definitions of information may be found in the statistical literature. Perhaps the oldest and most familiar is that due to Fisher (1925). For a real parameter and a density function satisfying Cramér-Rao regularity conditions it has the form

$$\begin{aligned} I_F(\theta) &= \int [(\partial/\partial\theta) \ln f(x \mid \theta)]^2 f(x \mid \theta)\, dx \\ &= \int -(\partial^2/\partial\theta^2) \ln f(x \mid \theta) f(x \mid \theta)\, dx; \end{aligned}$$

for a vector parameter $\boldsymbol{\theta}$ it becomes a matrix

$$\begin{aligned} \mathbf{I}_F(\boldsymbol{\theta}) &= \operatorname{cov}\{(\partial/\partial\boldsymbol{\theta}) \ln f(x \mid \boldsymbol{\theta}) \mid \boldsymbol{\theta}\} \\ &= \int -(\partial/\partial\boldsymbol{\theta})(\partial/\partial\boldsymbol{\theta}') \ln f(x \mid \boldsymbol{\theta}) f(x \mid \boldsymbol{\theta})\, dx \end{aligned}$$

where cov stands for covariance matrix and where the integrand being averaged in the second expression is the matrix of partial derivatives of the log-likelihood.

Shannon (1948) proposed a definition of information for communication theory. In its primary form it measures variation in a distribution; with a change of sign it measures concentration and is thereby more appropriate for statistics:

$$I_S(\theta) = \int \ln f(x \mid \theta) f(x \mid \theta)\, dx.$$

Kullback (1959) considers a definition of information for "discriminating in favor of $H_1(\theta_1)$ against $H_2(\theta_2)$":

$$I_K(\theta_1, \theta_2) = \int \ln [f(x \mid \theta_1)/f(x \mid \theta_2)] f(x \mid \theta_1)\, dx.$$

These information functions are additive with independent observation; in fact, additivity is taken as an essential property in most developments of information. The three information functions are defined for quite general statistical models, (mild regularity required for Fisher's definition). And the Fisher and the Kullback definitions are tied closely to large sample theory and to Bayes' theory respectively. The emphasis in this paper is not on information in a

Received 12 October 1964.

general model based on general principles. Rather it is on information in an important and somewhat special model—the location model and, more generally, the transformation-parameter model.

2. For a real variable and a real parameter. Consider the location model $f(x - \theta)$ involving a real variable x and a real parameter θ. As a measure of *information concerning the parameter value θ given the outcome x*, consider

$$I(\theta \mid x) = \ln f(x - \theta).$$

For interpretation, this can be viewed as the log-likelihood (old definition without indeterminate constant). Or, alternatively, for those that accept structural probability [Fraser, (1965)] this can be viewed as the logarithm of the structural density at the parameter value θ as determined by the outcome x. The information function is zero for unit structural density and measures logarithmically the structural density with respect to that reference value; a large positive value for $I(\theta \mid x)$ is strong information *for* that θ value and a large negative value is strong information *against* that value.

If θ^0 denotes a true value for the parameter, the value determining the distribution of x, then *the mean information attached to the value θ when the distribution is θ^0* is

$$\begin{aligned} I(\theta \mid \theta^0) &= \int_{-\infty}^{\infty} \ln f(x - \theta) f(x - \theta^0)\, dx \\ &= \int_{-\infty}^{\infty} \ln f(y - \delta) f(y)\, dy \end{aligned}$$

where $\delta = \theta - \theta^0$.

Consider some aspects or characteristics of this information function. First, the information at the true parameter value is

$$\begin{aligned} I(\theta^0 \mid \theta^0) &= \int \ln f(x - \theta^0) f(x - \theta^0)\, dx \\ &= I_S(\theta^0). \end{aligned}$$

Second, consider how much this exceeds the information attached to some other parameter value:

$$I(\theta^0 \mid \theta^0) - I(\theta \mid \theta^0) = \int \ln [f(x - \theta^0)/f(x - \theta)] f(x - \theta_0)\, dx = I_K(\theta^0, \theta).$$

This difference is always nonnegative [Kullback, (1959), p. 14]; accordingly $I(\theta \mid \theta^0)$ as a function of θ attains its maximum at the true value $\theta = \theta^0$.

Third, consider the curvature of the information function at the true value $\theta = \theta^0$. For this, assume the Cramér-Rao regularity conditions that allow θ-differentiations to be carried inside the integral. The slope of $I(\theta \mid \theta^0)$ is

$$\begin{aligned} (\partial/\partial\theta) I(\theta \mid \theta^0) &= \int (\partial/\partial\theta) \ln f(x - \theta) f(x - \theta^0)\, dx, \\ (\partial/\partial\theta^0) I(\theta \mid \theta^0) &= 0 \end{aligned}$$

where $\partial/\partial\theta^0$ is abbreviated notation for the operator defined by

$$(\partial/\partial\theta^0) f(\theta, \theta^0) = [(\partial/\partial\theta) f(\theta, \theta^0)]_{\theta = \theta^0}.$$

And the curvature at $\theta = \theta_0$ is

$$-[(\partial^2/\partial\theta^2)I(\theta \mid \theta_0)]_{\theta=\theta^0} = [\int -(\partial^2/\partial\theta^2) \ln f(x - \theta)f(x - \theta^0)\, dx]_{\theta=\theta_0}$$
$$= I_F(\theta^0).$$

Thus the information $I(\theta \mid \theta_0)$ as a function of θ has maximum value $I_S(\theta^0)$ at the true value, has curvature $I_F(\theta^0)$ at the true value, and has discrepancy to another θ value $I_K(\theta^0, \theta)$.

3. For a vector variable and a vector parameter. Consider the location-model $f(\mathbf{x} - \boldsymbol{\theta})$ involving a vector variable $\mathbf{x} = (x_1, \cdots, x_n)$ and a vector parameter $\boldsymbol{\theta} = (\theta_1, \cdots, \theta_n)$; this is an example of a transformation-parameter model. As a measurement of information concerning $\boldsymbol{\theta}$ given $\mathbf{x}$, take

$$I(\boldsymbol{\theta} \mid \mathbf{x}) = \ln f(\mathbf{x} - \boldsymbol{\theta}).$$

The mean information attached to the value θ when the distribution is θ^0 is

$$I(\boldsymbol{\theta} \mid \boldsymbol{\theta}_0) = \int \ln f(\mathbf{x} - \boldsymbol{\theta})f(\mathbf{x} - \boldsymbol{\theta}^0)\, d\mathbf{x}$$
$$= \int \ln f(\mathbf{y} - \boldsymbol{\delta})f(\mathbf{y})\, d\mathbf{y}$$

where $\boldsymbol{\delta} = (\theta_1 - \theta_1^0, \cdots, \theta_n - \theta_n^0)$.

In the pattern of the preceding section, the information $I(\boldsymbol{\theta} \mid \boldsymbol{\theta}^0)$ as a function of θ has maximum value $I_S(\boldsymbol{\theta}^0)$ at the true value θ_0, has discrepancy to another θ value $I_K(\boldsymbol{\theta}^0, \boldsymbol{\theta})$ and has curvature matrix $\mathbf{I}_F(\boldsymbol{\theta}^0)$ at the true value.

4. For a vector variable and a real parameter. Consider n independent variables $x_1, \cdots, x_n$ with location models $f_1(x_1 - \theta), \cdots, f_n(x_n - \theta)$. In particular this might be a sample from a location model; or it could be generalized to a variable $\mathbf{x} = (x_1, \cdots, x_n)$ from $f(\mathbf{x} - \theta\mathbf{1})$ where $\mathbf{1} = (1, \cdots, 1)$.

The orbit under translation can be described by the ancillary statistic $\mathbf{d} = (d_2, \cdots, d_n)$ where $d_i = x_i - x_1$ and the conditional sufficient statistic by x_1; the density for d is

$$h(\mathbf{d}) = \int_{-\infty}^{\infty} f_1(t)f_2(t + d_2) \cdots f_n(t + d_n)\, dt$$

and the conditional density for x_1 is

$$g(x_1 - \theta \mid \mathbf{d}) = f_1(x_1 - \theta)f_2(x_1 - \theta + d_2) \cdots f_n(x_1 - \theta + d_n)/h(d_2, \cdots, d_n).$$

As a measure of *information concerning θ given the outcome* $\mathbf{x} = (x_1, \cdots, x_n)$ consider $I(\theta \mid \mathbf{x}) = \ln g(x_1 - \theta \mid \mathbf{d})$. This can be viewed as the log-likelihood based on the conditional distribution; or as the logarithm of the structural density at the parameter value θ as determined by the outcome x.

The mean information at θ when the distribution is θ^0 is

$$I(\theta \mid \theta^0, \mathbf{d}) = \int \ln g(x_1 - \theta \mid \mathbf{d})g(x_1 - \theta^0 \mid \mathbf{d})\, dx_1$$

given the ancillary $\mathbf{d}$, and is

$$I(\theta \mid \theta_0) = \int \ln g(x_1 - \theta \mid \mathbf{d})\Pi f_i(x_i - \theta)\Pi\, dx_i$$

marginally.

The curvature of the information at $\theta = \theta^0$ is given by $I_F(\theta_0)$ since

$$(\partial^2/\partial\theta^2) \ln g(x_1 - \theta \mid \mathbf{d}) = (\partial^2/\partial\theta^2) \ln \Pi f_i(x_i - \theta).$$

The discrepancy from θ^0 to another θ is given by $I(\theta^0 \mid \theta^0) - I(\theta \mid \theta^0) = I_K(\theta^0, \theta)$; and it follows that $I(\theta \mid \theta^0)$ is a maximum at $\theta = \theta^0$.

Before considering the information at the true θ^0 it is convenient to consider a more general model of the type discussed in the preceding section. Suppose that $(x_1, d_2, \cdots, d_n)$ has a density function $g(x_1 - \theta \mid d_2 - \tau_2, \cdots, d_r - \tau_r)$ $h(d_2 - \tau_2, \cdots, d_r - \tau_r)$, with marginal density for $(d_2, \cdots, d_n)$ given by $h(d_2 - \tau_2, \cdots, d_n - \tau_n)$. Then

$$\begin{aligned} I(\theta \mid \theta^0) &= \int [\ln f(\mathbf{x} - \theta\mathbf{1}) - \ln h(\mathbf{d})] f(\mathbf{x} - \theta^0\mathbf{1})\, d\mathbf{x} \\ &= [I^{\mathbf{x}}(\theta, \boldsymbol{\tau} \mid \theta^0, \boldsymbol{\tau}) - I^{\mathbf{d}}(\boldsymbol{\tau} \mid \boldsymbol{\tau})]_{\boldsymbol{\tau}=0} \end{aligned}$$

where $I^{\mathbf{d}}$ is the information function calculated for the marginal distribution of the d's.

The information at the true parameter value can be obtained from the preceding expression: $I(\theta^0 \mid \theta^0) = I_S{}^{\mathbf{x}}(\theta_0) - I_S{}^{\mathbf{d}}$. The information at the true parameter value is thus the Shannon information in the full distribution *less* the Shannon information in the distribution of the ancillary variable.

As an example consider a variable x that is uniformly distributed on the interval $\theta \pm \frac{1}{2}$. For a single observation

$$\begin{aligned} I(\theta \mid \theta^0) &= 0 && \text{if } \theta = \theta^0 \\ &= -\infty && \text{otherwise;} \end{aligned}$$

for $\theta \neq \theta_0$ the mean information *against* θ is infinite. For two observations

$$\begin{aligned} I(\theta \mid \theta^0) &= \int_0^1 - \ln (1 - R) 2(1 - R)\, d\mathrm{R} \\ &= \tfrac{1}{2} && \text{if } \theta = \theta^0 \\ &= -\infty && \text{otherwise.} \end{aligned}$$

The information with two observations is *more* than twice the information in a single observation. This is in accord with the intuitive picture for information in such sampling.

5. For the transformation-parameter model. For a generalization, let e be an error variable on a space X, let $G = \{g\}$ be a unitary group of transformations (if $x' = gx = hx$, then $g = h$), and consider the model $x = [\theta]e$ where $[\theta]$, a typical element in G can be indexed alternatively by $\theta = [\theta]\theta_0$ taking values in a parameter space Ω. Suppose that e has a density function $f(e)$ with respect to an invariant measure m on X: $m(gA) = m(A)$.

Let $[x]$ be the transformation that produces x from a reference point $a(x)$ on the orbit through x. Then the ancillary statistic can be expressed as $a(x) = [x]^{-1}x$, and the conditionally sufficient statistic as $[x]$.

Let μ be the left measure on the group G: $\mu(gH) = \mu(H)$. Let ν be the right

measure on the group: $\nu(Hg) = \nu(H) = \mu(H^{-1})$. And let Δ be the modular function: $d\mu(g) = \Delta(g)\, d\nu(g)$.

The probability element for x is $f([\theta]^{-1}x)\ dm(x)$. This leads to the conditional distribution of $[x]$ given $a(x)$:

$$\begin{aligned} k(a(x))f([\theta]^{-1}x)\ d\mu([x]) &= k(a(x))f([\theta]^{-1}x)\Delta([x])\ d\nu([x]) \\ &= r([x] \mid a, \theta)\ d\nu([x]); \end{aligned}$$

and to the structural distribution for $[\theta]$ given x:

$$\begin{aligned} k(a(x))f([\theta]^{-1}x)\Delta([\theta]^{-1}[x])\ d\mu([\theta]) &= k(a(x))f([\theta]^{-1}x)\Delta([x])\ d\nu([\theta]) \\ &= r([x] \mid a, \theta)\ d\nu([\theta]). \end{aligned}$$

As a measure of information concerning θ given the outcome x consider

$$I(\theta \mid x) = \ln r([x] \mid a, \theta).$$

This can be viewed as the log-likelihood based on the conditional distribution; or as the logarithm of the structural density for $[\theta]$ determined by the outcome x. The mean information at θ when the distribution is θ^0 is

$$I(\theta \mid \theta^0, a) = \int I(\theta \mid x) r([x] \mid a, \theta^0)\ d\nu([x])$$

given the ancillary a, and is

$$I(\theta \mid \theta^0) = \int \ln r([x] \mid a, \theta) f([\theta^0]^{-1}x)\ dm(x)$$

marginally.

First, consider the information at the true parameter value θ^0. For this let n be the measure for the ancillary statistic determined by $dm(x) = dn(a)\ d\mu([x])$, and let $s(a) = k^{-1}(a)$ be the density function for a with respect to n:

$$f([\theta]^{-1}x)\ dm(x) = r([x] \mid a, \theta) s(a)\ d\nu([x])\ dn(a).$$

Then

$$\begin{aligned} I(\theta^0 \mid \theta^0) &= \int [\ln \{f([\theta^0]^{-1}x)\Delta([x])\} - \ln \{s(a)\}] f([\theta^0]^{-1}x)\ dm(x) \\ &= \int \ln \{f([\theta^0]^{-1}x)\Delta([x])\} f([\theta^0]^{-1}x)\Delta([x]) \cdot \Delta^{-1}([x])\ dm(x) \\ &\quad - \int \ln \{s(a)\} s(a)\ dn(a) \\ &= I_s{}^x(\theta^0) - I_s{}^a; \end{aligned}$$

in this expression the Shannon information for x is calculated using a density with respect to the adjusted measure $\Delta^{-1}([x])\ dm(x)$, which corresponds to right invariant measure on the group. Thus the information at the true value θ^0 is the Shannon information in the full distribution less the Shannon information in the ancillary distribution.

The discrepancy in the information from θ^0 to θ is again given by

$$I(\theta^0 \mid \theta^0) - I(\theta \mid \theta^0) = I_K(\theta^0, \theta)$$

and is nonnegative. The information function thus attains its maximum at the true value θ_0.

Suppose now that the group G is continuous and has Euclidean coordinates. Then the matrix of partial derivatives satisfies

$$(\partial/\partial\boldsymbol{\theta})(\partial/\partial\boldsymbol{\theta}')\ln(r[x] \mid a, \boldsymbol{\theta}) = (\partial/\partial\boldsymbol{\theta})(\partial/\partial\boldsymbol{\theta}')\ln f([\boldsymbol{\theta}]^{-1}x).$$

If in addition the distribution satisfies the Cramér-Rao regularity conditions then

$$[-(\partial/\partial\boldsymbol{\theta})(\partial/\partial\boldsymbol{\theta}')I(\boldsymbol{\theta} \mid \boldsymbol{\theta}^0)]_{\boldsymbol{\theta}=\boldsymbol{\theta}^0} = \mathbf{I}_F(\boldsymbol{\theta}^0).$$

6. Three optimality properties. Consider a variable x having a transformation model as described in the preceding section. And, in addition consider a reduction $y = w(x)$ and suppose that the new variable y admits the definition of information in the preceding section. In this section, the effects on the information function of a statistical reduction will be examined.

Let $A(y)$ be the ancillary statistic for y in reference point form, $y = [y]A(y)$. Towards an ancillary for x consider the solutions of the equation $w(x) = A(y)$. The relation $y = [y]A(y)$ shows that on any orbit for x there is exactly one solution; designate it by $a(x)$. Then $x = [x]a(x)$ where $[y] = [w(x)] = [x]$. In addition it follows that A is a reduction on a: $A = v(a)$. Thus the reduction $x \to w(x)$ is equivalent to $([x], a) \to ([x], v(a))$.

THEOREM 1. *Under a statistical reduction $y = w(x)$ that maintains a transformation model*

$$I^x(\theta^0 \mid \theta^0) \geqq I^y(\theta^0 \mid \theta^0),$$

with strict inequality unless the distribution of the transformation variable $[x]$ is determined by the ancillary statistic for y.

PROOF. Consider the conditional information given the ancillary A:

$$I^x(\theta^0 \mid \theta^0, A) = E\{I^x(\theta^0 \mid \theta^0, a)\}$$

where the expectation is taken with respect to the distribution of a given A. Then

$$\begin{aligned} I^x(\theta^0 \mid \theta^0, A) &= E\{\int \ln r([x] \mid a, \theta^0) r([x] \mid a, \theta^0)\, dv([x])\} \\ &= \int E\{\ln r([x] \mid a, \theta^0) r([x] \mid a, \theta^0)\}\, dv[x] \\ &\geqq \int \ln R([x] \mid A, \theta^0) R([x] \mid A, \theta^0)\, dv[x] \\ &= I^y(\theta^0 \mid \theta^0, A); \end{aligned}$$

the inequality uses the convex function $t \ln t$ and the expression $R([x] \mid A, \theta^0) = E\{r[x] \mid a, \theta^0\}$ is the average with respect to a of the conditional density of $[x]$ given a, and is therefore the conditional density of $[x]$ given A. With the straightforward checking of the equality case, this completes the proof.

THEOREM 2. *Under a statistical reduction $y = w(x)$ that maintains a transformation model*

$$I^x(\theta^0 \mid \theta^0) - I^x(\theta \mid \theta^0) \geqq I^y(\theta^0 \mid \theta^0) - I^y(\theta \mid \theta^0).$$

PROOF. This is a specialization of Corollary 4.3, Chapter 2, in Kullback (1959).

THEOREM 3. *Under a statistical reduction* $y = w(x)$ *that maintains a transformation model, the curvature matrix* $I_F(\theta^0)$ *of the information function at the true value* θ^0 *satisfies*

$$\mathbf{I}_F^x(\theta^0) \succ \mathbf{I}_F^y(\theta^0)$$

where $\succ$ *denotes the natural partial ordering on positive semi-definite matrices.* ($A \succ B$ *if and only if* $A - B$ *is positive semi-definite.*)

PROOF. Consider *any* statistical reduction $y = w(x)$. Let $g(y \mid \theta)$ be the density function for y with respect to a measure derived from the measure m for the variable x; then

$$f(x \mid \theta) = g(y \mid \theta)h(x \mid y, \theta)$$

where $h(x \mid y, \theta)$ is the conditional density for $x \mid y$ with respect to the conditional measure. The logarithmic gradient vectors at θ^0 satisfy

$$\partial \ln f(x \mid \theta)/\partial\theta^0 = \partial \ln g(y \mid \theta)/\partial\theta^0 + \partial \ln h(x \mid y, \theta)/\partial\theta^0.$$

The second term on the right side has a conditional expectation given y at θ^0 that is equal to zero; as a result, the cross covariance matrix between the two vectors on the right side is equal to zero. The covariance matrices then satisfy

$$\mathbf{I}^x(\theta^0) = \mathbf{I}^y(\theta^0) + \mathbf{I}^{x \mid y}(\theta^0)$$

and therefore $I^x(\theta^0) \succ I^y(\theta^0)$.

REFERENCES

FISHER, R. A. (1925). Theory of statistical estimation. *Proc. Cambridge Philos. Soc.* **22** 700–725.

FRASER, D. A. S. (1965). Structural probability and a generalization. Submitted to *Biometrika*.

KULLBACK, S. (1959). *Information Theory and Statistics*. Wiley, New York.

SHANNON, C. E. (1948). A mathematical theory of communication. *Bell System Tech. J.* **27** 379–423 and 623–656.