ON INFORMATION AND SUFFICIENCY

By S. Kullback and R. A. Leibler

The George Washington University and Washington, D. C.

1. Introduction. This note generalizes to the abstract case Shannon's definition of information [15], [16]. Wiener's information (p. 75 of [18]) is essentially the same as Shannon's although their motivation was different (cf. footnote 1, p. 95 of [16]) and Shannon apparently has investigated the concept more completely. R. A. Fisher's definition of information (intrinsic accuracy) is well known (p. 709 of [6]). However, his concept is quite different from that of Shannon and Wiener, and hence ours, although the two are not unrelated as is shown in paragraph 2.

R. A. Fisher, in his original introduction of the *criterion of sufficiency*, required "that the statistic chosen should summarize the whole of the relevant information supplied by the sample," (p. 316 of [5]). Halmos and Savage in a recent paper, one of the main results of which is a generalization of the well known Fisher-Neyman theorem on sufficient statistics to the abstract case, conclude, "We think that confusion has from time to time been thrown on the subject by . . . , and (c) the assumption that a sufficient statistic contains all the information in only the technical sense of 'information' as measured by variance," (p. 241 of [8]). It is shown in this note that the information in a sample as defined herein, that is, in the Shannon-Wiener sense cannot be increased by any statistical operations and is invariant (not decreased) if and only if sufficient statistics are employed. For a similar property of Fisher's information see p. 717 of [6], Doob [19].

We are also concerned with the statistical problem of discrimination ([3], [17]), by considering a measure of the "distance" or "divergence" between statistical populations ([1], [2], [13]) in terms of our measure of information. For the statistician two populations differ more or less according as to how difficult it is to discriminate between them with the best test [14]. The particular measure of divergence we use has been considered by Jeffreys ([10], [11]) in another connection. He is primarily concerned with its use in providing an invariant density of *a priori* probability. A special case of this divergence is Mahalanobis' generalized distance [13].

We shall use the notation of Halmos and Savage [8] and that of [7].

2. Information. Assume given the probability spaces $(X, \mathbf{S}, \mu_i)$, $i = 1, 2$, such that $\mu_1 \equiv \mu_2$[1] (cf. p. 228 of [8]) and let λ be a probability measure such that $\lambda \equiv \{\mu_1, \mu_2\}$ (e.g., λ may be μ_1, or μ_2 or $\frac{1}{2}(\mu_1 + \mu_2)$, etc.). By the Radon-Nikodym theorem [7] there exist $f_i(x)$, $i = 1, 2$, unique up to sets of measure zero in λ,

[1] If $\mu_1(E) \neq 0$, $\mu_2(E) = 0$ or $\mu_1(E) = 0$, $\mu_2(E) \neq 0$ for $E \in \mathbf{S}$ then we can discriminate perfectly between the populations. The assumption $\mu_1 \equiv \mu_2$ that is, that μ_1 and μ_2 are absolutely continuous with respect to each other is made to avoid this situation.

Institute of Mathematical Statistics is collaborating with JSTOR to digitize, preserve, and extend access to *The Annals of Mathematical Statistics.* www.jstor.org

measurable λ with $0 < f_i(x) < \infty$ $[\lambda]$, $i = 1, 2$, such that

$$\mu_i(E) = \int_E f_i(x)\, d\lambda(x), \qquad i = 1, 2, \tag{2.1}$$

for all $E \in \mathbf{S}$. If H_i, $i = 1, 2$, is the hypothesis that x was selected from the population whose probability measure is μ_i, $i = 1, 2$ then we define

$$\log \frac{f_1(x)}{f_2(x)} \tag{2.2}$$

as *the information*[2] *in* x *for discrimination between* H_1 *and* H_2. The mean information for discrimination between H_1 and H_2 per observation from $E \in \mathbf{S}$ for μ_1 is given by (cf. pp. 18, 19 of [16]; p. 76 of [18])

$$\begin{aligned} I_{1:2}(E) &= \frac{1}{\mu_1(E)} \int_E d\mu_1(x) \log \frac{f_1(x)}{f_2(x)} = \frac{1}{\mu_1(E)} \int_E f_1(x) \log \frac{f_1(x)}{f_2(x)}\, d\lambda(x) \\ &\qquad \text{for } \mu_1(E) > 0, \\ &= 0 \qquad \text{for } \mu_1(E) = 0. \end{aligned} \tag{2.3}$$

It should be noted that $I_{1:2}(E)$ in (2.3) is well defined in that the integral in its definition always exists even though it may be $+\infty$, since the measures are finite measures.[3] It is shown in Lemma 3.2 that

$$I_{1:2}(E) \geqq \log \mu_1(E)/\mu_2(E) \quad \text{for} \quad \mu_1(E) > 0.$$

We shall denote by $I(1:2)$ the mean information for discrimination between H_1 and H_2 per observation from μ_1; i.e.,[4]

$$\begin{aligned} I(1:2) = I_{1:2}(X) &= \int d\mu_1(x) \log \frac{f_1(x)}{f_2(x)} \\ &= \int f_1(x) \log \frac{f_1(x)}{f_2(x)}\, d\lambda(x). \end{aligned} \tag{2.4}$$

---

[2] It follows from Bayes' Theorem [12] that

$$\log \frac{f_1(x)}{f_2(x)} = \log \frac{P(H_1 \mid x)}{P(H_2 \mid x)} - \log \frac{\alpha_1}{\alpha_2} \quad [\lambda]$$

where α_i, $i = 1, 2$, are the *a priori* probabilities and $P(H_i \mid x)$, $i = 1, 2$, the *a posteriori* probabilities of H_i, $i = 1, 2$, respectively.

[3] We are indebted to a referee for this remark as well as for the following example which shows that the assumptions at the beginning of this paragraph do not imply finiteness of information. Take $E = (0, 1)$, $\mu_1 =$ Lebesgue measure, $f_2(x)/f_1(x) = ke^{-1/x}$, where $k^{-1} = \int_0^1 e^{-1/t}\, dt$. It is easily verified that $I(1:2)$ is infinite (cf. also p. 137 [9]).

[4] We shall omit the region of integration when it is the entire space.

Set

$$
\begin{aligned}
J_{12}(E) &= I_{1:2}(E) + I_{2:1}(E) \\
&= \frac{1}{\mu_1(E)} \int_E d\mu_1(x) \log \frac{f_1(x)}{f_2(x)} + \frac{1}{\mu_2(E)} \int_E d\mu_2(x) \log \frac{f_2(x)}{f_1(x)} \\
&= \int_E \left(\frac{f_1(x)}{\mu_1(E)} - \frac{f_2(x)}{\mu_2(E)} \right) \log \frac{f_1(x)}{f_2(x)} \, d\lambda(x).
\end{aligned} \tag{2.5}
$$

We denote by $J(1, 2)$ the "divergence" between μ_1 and μ_2 (cf. p. 158 of [11]) so that[4]

$$
J(1, 2) = J_{12}(X) = \int (f_1(x) - f_2(x)) \log \frac{f_1(x)}{f_2(x)} \, d\lambda(x). \tag{2.6}
$$

Shannon ([15], [16]) defined information on a finite discrete space and we note that $I_{1:2}(E)$ defined in (2.3) is precisely the generalization of that information which is obtained when one replaces the finite space by $\mathsf{S} \cap E$, the measure of equidistribution by $\mu_2/\mu_1(E)$ and the measure whose information is being defined by $\mu_1/\mu_1(E)$. Just as Shannon observed that certain theorems were carried over to the Lebesgue case, we shall see here that they maybe formally carried over to the general case.[5]

For the parametric case in which $f_1(x) = f(x, \theta)$ and $f_2(x) = f(x, \theta + \Delta\theta)$, where θ and $\theta + \Delta\theta$ are neighboring points in the k-dimensional parameter space, with suitable assumptions on the density function (e.g., see p. 774 of [4]), to within second order terms it is found that

$$
I(\theta; \theta + \Delta\theta) = \tfrac{1}{2} \Sigma g_{\alpha\beta} \Delta\theta_\alpha \Delta\theta_\beta, \qquad \alpha, \beta = 1, \cdots, k, \tag{2.7}
$$

$$
J(\theta, \theta + \Delta\theta) = \Sigma g_{\alpha\beta} \Delta\theta_\alpha \Delta\theta_\beta, \qquad \alpha, \beta = 1, \cdots, k, \tag{2.8}
$$

where

$$
g_{\alpha\beta} = \int f \left(\frac{1}{f} \frac{\partial f}{\partial \theta_\alpha} \right) \left(\frac{1}{f} \frac{\partial f}{\partial \theta_\beta} \right) d\lambda \tag{2.9}
$$

are the elements of Fisher's information matrix (cf. par. 3.9 of [11]).

When μ_1 and μ_2 are multivariate normal populations with a common matrix of variances and covariances then

$$
J(1, 2) = \Sigma \delta_\alpha \delta_\beta \sigma^{\alpha\beta}, \qquad \alpha, \beta = 1, \cdots, k, \tag{2.10}
$$

where δ_α, $\alpha = 1, \cdots, k$, are the differences of the respective population means and $\sigma^{\alpha\beta}$, $\alpha, \beta = 1, \cdots, k$, are the elements of the inverse of the common matrix

[5] We are indebted to a referee for the comments with respect to Shannon's definition as well as for the comment that this should be of interest to anyone who has puzzled over Wiener's statement that his definition of "information" can be used to replace Fisher's definition in the technique of statistics (p. 76 of [18]).

of variances and covariances; i.e., $J(1, 2)$ in (2.10) is k times Mahalanobis' generalized distance [13].

3. Some properties of information.

LEMMA 3.1. *$I(1:2)$ is almost positive definite; i.e., $I(1:2) \geqq 0$ with equality if and only if $f_1(x) = f_2(x)$ [λ].*

PROOF.[6] Let $g(x) = f_1(x)/f_2(x)$. Then

$$I(1:2) = \int f_2(x)g(x) \log g(x)\, d\lambda(x) = \int g(x) \log g(x)\, d\mu_2(x). \tag{3.1}$$

If we write $\varphi(t) = t \log t$, then since $0 < g(x) < \infty$ [λ] and

$$\int g(x)\, d\mu_2(x) = \int f_1(x)\, d\lambda(x) = 1, \tag{3.2}$$

we may write

$$\varphi(g(x)) = \varphi(1) + [g(x) - 1]\varphi'(1) + \tfrac{1}{2}[g(x) - 1]^2\varphi''(h(x)) [\lambda], \tag{3.3}$$

where $h(x)$ lies between $g(x)$ and 1 so that $0 < h(x) < \infty$ [λ]. Therefore

$$\int \varphi(g(x))\, d\mu_2(x) = \frac{1}{2}\int [g(x) - 1]^2\varphi''(h(x))\, d\mu_2(x), \tag{3.4}$$

where $\varphi''(t) = \dfrac{1}{t} > 0$ for $t > 0$. It therefore follows from (3.4) that

$$\int g(x) \log g(x)\, d\mu_2(x) \geqq 0 \tag{3.5}$$

with equality if and only if $g(x) = 1$ [λ].

LEMMA 3.2.

$$I_{1:2}(E) \geqq \log \frac{\mu_1(E)}{\mu_2(E)} \qquad \text{for } \lambda(E) > 0.$$

PROOF. If $I_{1:2}(E) = \infty$, the result is trivial. For finite $I_{1:2}(E)$ apply Lemma 3.1 to

$$I_{1:2}(E) - \log \frac{\mu_1(E)}{\mu_2(E)} = \int_E \frac{d\mu_1(x)}{\mu_1(E)} \log \frac{f_1(x)/\mu_1(E)}{f_2(x)/\mu_2(E)}.$$

THEOREM 3.1. *$I(1:2)$ is additive for independent random events*[7]*; i.e.,*

$$I_{xy}(1:2) = I_x(1:2) + I_y(1:2).$$

---

[6] This is essentially the proof on p. 151 of [9].

[7] Shannon (p. 21 of [16]) and Wiener (p. 77 of [18]) prove similar results. This is clearly a fundamental property which information must possess, and is one of the *a priori* requirements set down by Shannon in arriving at his definition.

PROOF.

$$I_{xy}(1:2) = \int f_1(x, y) \log \frac{f_1(x, y)}{f_2(x, y)} \, d\lambda(x, y)$$

$$= \iint f_1^{(1)}(x) f_1^{(2)}(y) \log \frac{f_1^{(1)}(x) f_1^{(2)}(y)}{f_2^{(1)}(x) f_2^{(2)}(y)} \, d\lambda_1(x) \, d\lambda_2(y) \tag{3.6}$$

$$= \int f_1^{(1)}(x) \log \frac{f_1^{(1)}(x)}{f_2^{(1)}(x)} \, d\lambda_1(x) + \int f_1^{(2)}(y) \log \frac{f_1^{(2)}(y)}{f_2^{(2)}(y)} \, d\lambda_2(y)$$

$$= I_x(1:2) + I_y(1:2).$$

4. Transformations and invariance of I(1:2). Consider the measurable transformation T of the probability spaces $(X, \mathbf{S}, \mu_i)$ *onto* the probability spaces $(Y, \mathbf{T}, \nu_i)$ and suppose for $G \in \mathbf{T}$, $\nu_i(G) = \mu_i(T^{-1}G)$, $i = 1, 2$. Then $\nu_1 \equiv \nu_2 \equiv \gamma$, where $\gamma = \lambda T^{-1}$. We define

$$I'_{1:2}(G) = \frac{1}{\nu_1(G)} \int_G d\nu_1(y) \log \frac{g_1(y)}{g_2(y)} = \frac{1}{\nu_1(G)} \int_G g_1(y) \log \frac{g_1(y)}{g_2(y)} \, d\gamma(y), \tag{4.1}$$

$$J'_{12}(G) = \int_G \left(\frac{d\nu_1(y)}{\nu_1(G)} - \frac{d\nu_2(y)}{\nu_2(G)} \right) \log \frac{g_1(y)}{g_2(y)}, \tag{4.2}$$

where $g_i(y)$ is defined by

$$\nu_1(G) = \int_G g_i(y) \, d\gamma(y), \qquad i = 1, 2, \tag{4.3}$$

for all $G \in \mathbf{T}$.

THEOREM 4.1. *$I(1:2) \geqq I'(1:2)$, with equality if and only if T is a sufficient statistic.*

PROOF. If $I(1:2) = \infty$ the result is trivial. By Lemma 3 of Halmos and Savage [8]

$$I'(1:2) = \int d\mu_1(x) \log \frac{g_1 T(x)}{g_2 T(x)}. \tag{4.4}$$

Then

$$I(1:2) - I'(1:2) = \int d\mu_1(x) \left[\log \frac{f_1(x)}{f_2(x)} - \log \frac{g_1 T(x)}{g_2 T(x)} \right]$$

$$= \int f_1(x) \log \frac{f_1(x) g_2 T(x)}{f_2(x) g_1 T(x)} \, d\lambda(x). \tag{4.5}$$

If we set $g(x) = \dfrac{f_1(x)g_2T(x)}{f_2(x)g_1T(x)}$, then

$$I(1:2) - I'(1:2) = \int \frac{f_2(x)g_1T(x)}{g_2T(x)}\, g(x) \log g(x)\, d\lambda(x) \tag{4.6}$$
$$= \int g(x) \log g(x)\, d\mu_{12}(x),$$

where $\mu_{12}(E) = \int_E \dfrac{f_2(x)g_1T(x)}{g_2T(x)}\, d\lambda(x)$ for all $E \in \mathbf{S}$.

Since

$$\int g(x)\, d\mu_{12}(x) = \int \frac{f_1(x)g_2T(x)}{f_2(x)g_1T(x)} \frac{f_2(x)g_1T(x)}{g_2T(x)}\, d\lambda(x) = 1,$$

the method of Lemma 3.1 leads to the conclusion that $I(1:2) - I'(1:2) \geqq 0$ with equality if and only if

$$\frac{f_1(x)}{f_2(x)} = \frac{g_1T(x)}{g_2T(x)}\ [\lambda]. \tag{4.7}$$

But (4.7) implies that

$$\frac{f_1(x)}{f_2(x)}\ (\varepsilon)T^{-1}(\mathbf{T})\ [\lambda], \tag{4.8}$$

which is by Corollary 2 of Halmos and Savage [8] necessary and sufficient that the statistic T be sufficient for a homogeneous set of measures on $\mathbf{S}$. If T is sufficient then by the same proof[8] as Theorem 1 of Halmos and Savage [8] $f_1(x)$ and $f_2(x)$ are $(\varepsilon)T^{-1}(\mathbf{T})[\lambda]$. Then by Lemma 2 of Halmos and Savage [8] and the definition of g_1 and g_2, $f_1(x) = g_1T(x)$ $[\lambda]$, $f_2(x) = g_2T(x)$ $[\lambda]$ and the result in (4.7) follows.

COROLLARY 4.1. *$I(1:2) = I'(1:2)$ if T is non-singular.*

PROOF. If T is non-singular, $T^{-1}(\mathbf{T})$ is $\mathbf{S}$ and therefore $f_i(x)(\varepsilon)T^{-1}(\mathbf{T})$, $i = 1, 2$. The result then follows from Theorem 4.1.

THEOREM 4.2.[9] *$I_{1:2}(T^{-1}G) = I'_{1:2}(G)$ for all $G \in \mathbf{T}$ if and only if*

$$I(1:2) = I'(1:2).$$

PROOF.

$$I'_{1:2}(G) = \int_G \frac{d\nu_1(y)}{\nu_1(G)} \log \frac{g_1(y)}{g_2(y)} = \int \chi_G(y)\, \frac{d\nu_1(y)}{\nu_1(G)} \log \frac{g_1(y)}{g_2(y)}$$
$$= \int \chi_{T^{-1}G}(x)\, \frac{d\mu_1(x)}{\mu_1(T^{-1}G)} \log \frac{g_1T(x)}{g_2T(x)} \tag{4.9}$$
$$= \int_{T^{-1}G} \frac{d\mu_1(x)}{\mu_1(T^{-1}G)} \log \frac{g_1T(x)}{g_2T(x)}.$$

Application of the method of Theorem 4.1 completes the proof.

[8] Note that the λ in Theorem 1 of [8] is different from the λ here. However, as remarked by a referee, the same proof will suffice.

[9] We are indebted to a referee for calling this to our attention.

5. Properties of J(1, 2). For each of the results in paragraphs 3 and 4 there can be stated an identical one for $J(1,2)$. This follows from its definition in (2.5) and (2.6). Also it should be noted that $J(1,2)$ is symmetric with respect to μ_1 and μ_2 and independent of the *a priori* probabilities. Jeffreys (par. 3.9 of [11]) mentioned the symmetry, positive definiteness and additivity, and invariance for non-singular transformations.

6. Application. Two indications of simple application of these concepts may be useful.

(1). Consider the problem of testing an hypothesis presented by Lehmann (p. 2 of [20]). Let the subscript 1 refer to Lehmann's hypothesis H, the subscript 2 refer to any of the alternatives, $F = \{-2, 2\}$, $G = \{0\}$, then

$$(6.1)\qquad \begin{aligned} I_{1:2}(F) &= \frac{1}{\alpha}\left(\frac{\alpha}{2}\log\frac{\alpha}{2pc} + \frac{\alpha}{2}\log\frac{\alpha}{2c(1-p)}\right), \\ I_{1:2}(G) &= \frac{1}{\alpha}\cdot\alpha\log\frac{1-\alpha}{1-c}. \end{aligned}$$

It may be readily verified that $I_{1:2}(G) < I_{1:2}(F)$ and therefore G i.e. $\{0\}$ should be used as the critical region.

(2). Suppose it is necessary to decide whether a sample of n observations has been drawn from the multinomial population $\{p_1, p_2, \cdots, p_k\}$ or $\left\{\frac{1}{k}, \frac{1}{k}, \cdots, \frac{1}{k}\right\}$. Because of certain limitations the test must be made under the following conditions:

a) Sequential analysis cannot be used.

b) The observations must be grouped into two mutually exclusive categories.

If it is assumed that $p_1 \geqq p_2 \geqq \cdots \geqq p_k$, then the most effective grouping is such that

$$(6.2)\qquad J' = \left(\sum_{i=1}^{r} p_i - \frac{r}{k}\right)\log\frac{\sum_{i=1}^{r} p_i}{r/k} + \left(\sum_{i=r+1}^{k} p_i - \frac{k-r}{k}\right)\log\frac{\sum_{i=r+1}^{k} p_i}{(k-r)/k}$$

is a maximum. The efficiency of the grouped test is measured by

$$(6.3)\qquad J'/J,$$

where

$$(6.4)\qquad J = \sum_{i=1}^{k}\left(p_i - \frac{1}{k}\right)\log\frac{p_i}{1/k}$$

in the sense that n observations of the grouped test will provide as much information as N observations of the ungrouped test where

$$(6.5)\qquad nJ' = NJ.$$

For example if $p_1 = .5$, $p_2 = .3$, $p_3 = .1$, $p_4 = .1$, then using logarithms to base 10, J' for $r = 1, 2, 3, 4$, becomes respectively

$$(6.6)\qquad .1193, \quad .0903, \quad .0716, \quad 0.0,$$

and in this case J is 0.1986. The most effective grouping is therefore (p_1), $(p_2 + p_3 + p_4)$ and the grouped case is $\frac{.1193}{.1986} = .6007$ times as efficient as the ungrouped test; i.e., there is a loss of 40% because of the grouping.

REFERENCES

[1] A. BHATTACHARYYA, "On a measure of divergence between two statistical populations defined by their probability distributions," *Bull. Calcutta Math. Soc.*, Vol. 35 (1943), pp. 99–109.

[2] A. BHATTACHARYYA, "On a measure of divergence between two multinomial populations," *Sankhyā*, Vol. 7 (1946), pp. 401–406.

[3] G. W. BROWN, "Basic principles for construction and application of discriminators," *Jour. Clinical Psych.*, Vol. 6 (1950), pp. 58–61.

[4] J. L. DOOB, "Probability and statistics," *Trans. Am. Math. Soc.*, Vol. 36 (1934), pp. 759–775.

[5] R. A. FISHER, "On the mathematical foundations of theoretical statistics," *Philos. Trans. Roy. Soc. London, Ser. A*, Vol. 222 (1921), pp. 309–368.

[6] R. A. FISHER, "Theory of statistical estimation," *Proc. Cambridge Philos. Soc.*, Vol. 22 (1925), pp. 700–725.

[7] P. R. HALMOS, *Measure Theory*, D. Van Nostrand, 1950.

[8] P. R. HALMOS AND L. J. SAVAGE, "Application of the Radon-Nikodym theorem to the theory of sufficient statistics," *Annals of Math. Stat.*, Vol. 20 (1949), pp. 225–241.

[9] G. H. HARDY, J. E. LITTLEWOOD, AND G. PÓLYA, *Inequalities*, Cambridge University Press, 1934.

[10] H. JEFFREYS, "An invariant form for the prior probability in estimation problems," *Proc. Roy. Soc. London, Ser. A.*, Vol. 186 (1946), pp. 453–461.

[11] H. JEFFREYS, *Theory of Probability*, 2nd ed., Oxford, 1948.

[12] A. KOLMOGOROFF, *Grundbegriffe der Wahrscheinlichkeitsrechnung*, Ergebnisse der Mathematik und ihrer Grenzgebiete, Julius Springer, Berlin, 1933.

[13] P. C. MAHALANOBIS, "On the generalized distance in statistics," *Proc. Nat. Inst. of Sciences of India*, Vol. 2 (1) (1936).

[14] E. MOURIER, "Étude du choix entre deux lois de probabilité," *C. R. Acad. Sci. Paris*, Vol. 223 (1946), pp. 712–714.

[15] C. E. SHANNON, "A mathematical theory of communication," *Bell System Technical Journal*, Vol. 27 (1948), pp. 379–423; pp. 623–656.

[16] C. E. SHANNON AND W. WEAVER, *The Mathematical Theory of Communication*, University of Illinois Press, Urbana, 1949.

[17] B. L. WELCH, "Note on discriminant functions," *Biometrika*, Vol. 31 (1939), pp. 218–219.

[18] N. WIENER, *Cybernetics*, John Wiley and Sons, 1948.

[19] J. L. DOOB, "Statistical estimation," *Trans. Am. Math. Soc.*, Vol. 39 (1936), pp. 410–421.

[20] E. L. LEHMANN, "Some principles of the theory of testing hypotheses," *Annals of Math. Stat.*, Vol. 21 (1950), pp. 1–26.