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.