Book Review

Information Theory and Statistics. By SOLOMON KULLBACK. John Wiley, New York, N. Y., 1959. xvii + 395 pp., $12.50.

Information theory has gradually achieved respectability in the eyes of mathematicians. This is borne out by several works (*1–10*) which are mainly concerned with generalizations and abstract formulations of the basic concepts. An important entity in information theory is the quantity "information" (or entropy). Generalizations and abstract formulations of this quantity and a study of its various properties have been the exclusive subject of many works (*1–3*, *5*, *6*, *10*). Information theory as a branch of mathematical theory of probability and statistics naturally finds applications to any probabilistic or statistical system of observations, e.g., the field of communication theory (which formulates a communication system as a stochastic or random process) (*11*), statistical thermodynamics (*7*), the theory of estimation and hypotheses testing in mathematical statistics, etc. It is this last application which forms an important part of the book under review.

This book can be roughly divided into two parts—part one (Chapters 1–5), which may be considered as the information theory part and part two (Chapters 6–13), which consists of applications of the material developed in part one to the problems of estimation and hypotheses testing.

In the first part the author introduces and defines a logarithmic measure of information "I" as follows. Let μ_1 and μ_2 be two probability measures defined on a measurable space (X,S). Let H_1 and H_2 be the two hypotheses corresponding to μ_1 and μ_2. Then the mean information for discrimination in favour of H_1 against H_2 per observation from μ_1 is defined as

$$I(1{:}2,X) = I(1{:}2) = \int f_1(x) \log[f_1(x)/f_2(x)]\, d\lambda(x)$$

where f_1, f_2 are the Radon-Nikodym derivatives (also generalized probability densities) i.e., $f_1 = d\mu_1/d\lambda$, $f_2 = d\mu_2/d\lambda$, and λ is a probability measure such that $\lambda \equiv \mu_1$, $\lambda \equiv \mu_2$. A corresponding definition holds for $I(2{:}1)$. The measure of information "I" has been termed as "directed divergence" and the quantity $J(1,2) = I(1{:}2) + I(2{:}1)$ as "divergence." The author points out that if X and Y are two random variables and H_1 is the hypothesis that X and Y depend with a probability density $f(x,y)$, and H_2 is the hypothesis that X and Y are independent with respective probability densities $g(x)$ and $h(y)$ then

$$I(1{:}2) = \int\int f(x,y) \log[f(x,y)/g(x)h(y)]\, dxdy = R(X,Y),$$

which is Shannon's rate of information transmission. In Chapter 2 some properties of "I" have been studied: the important ones are additivity, convexity, and invariance. In section 5 of this chapter a theorem (attributed to Sakaguchi) relating the channel capacity C to $J(1,2)$ has been stated. This theorem may be of interest to the communication engineer. Section 6 deals with the relationship of "I" with Fisher's information measure and sufficiency.† Many exercises at the end of this chapter state the corresponding properties for Shannon's information rate "H." A small point to note here is that there is no general relationship corresponding to $H(Y) \geqslant H(Y \mid X)$ ((c), exercise 8:30, p. 34) for the measure "I." In Chapter 3 "the minimum discrimination information" is defined as $I(*:2) = \inf I(1:2)$ where the inf is taken over all $f_1(x)$ subject to the condition that

$$\int T(x) f_1(x) \, d\lambda(x) = \theta$$

where θ is a constant, and $Y = T(x)$ is a measurable statistic. Then $I(1:2) \geqq I(*:2)$, with equality if and only if $f_1(x) = f^*(x) = f_2(x) \exp[T(x)/M_2(\tau)]$ $[\lambda]$ where $M_2(\tau) = \int f_2(x) \exp[\tau T(x)] \, d\lambda(x)$ (Theorem 2.1, p. 38). This theorem is considered as a generalization of the well-known Fréchet-Cramér-Rao inequality. In Chapter 4 some limiting properties concerning "I" have been considered and their relation to some results of Chernoff ("Large-sample theory: parametric case," *Ann. Math. Stat.* **27,** 1–22 (1956)) has also been pointed out. The estimators of the information measure and the general asymptotic distribution theory of these estimators have been considered in Chapter 5. When $f_2(x)$ corresponds to the generalized density of n independent observations $I(*:2)$ is estimated by using the observed value of $T(x)$ in a sample of n independent observations as an estimate of θ, $\hat{\theta}(x)$, and a related estimate of τ, $\hat{\tau}(x) = \tau(\hat{\theta}(x))$, such that

$$T(x) = \hat{\theta}(x) = \left[\frac{d}{d\tau} \log M_2(\tau)\right]_{\tau=\hat{\tau}(x)=\tau(\hat{\theta}(x))} .$$

Let this estimate be $\hat{I}(*:2)$. $\hat{I}(*:2)$ is a measure of the "directed divergence" between the sample and $f_2(x)$. The larger the value of $\hat{I}(*:2)$ the worse is the resemblance between the sample and the population with generalized density $f_2(x)$. $\hat{I}(*:H_2) - \hat{I}(*:H_1) \geqslant c$, where c is a constant, is considered as a critical region for testing a null hypothesis H_2 against an alternative hypothesis H_1. When H_1 and H_2 are simple, and $T(x) = \log[f_1(x)/f_2(x)]$, this yields the most powerful critical region as given by the Neyman-Pearson lemma. An application of $\hat{I}(*:H)$ to the problem of classification has also been pointed out.

The second part of the book consists of applications. The main purpose of the author is to attempt a unification of a heterogeneous body of material connected with various statistical procedures which is scattered throughout the literature. The unification is attained by "a consistent application of the concepts and properties" developed in the earlier part of the book. In Chapters 6–13 many well-known

† Fisher's information measure, the measure "I," and their relationship with the problems of estimation and hypotheses testing have also been considered by D. D. Joshi in "L'information en statistique mathématique et dans la théorie des communications"—Publications de Institut de statistique de l'Université de Paris, vol. VIII, Fascicule 2, (1959), pp. 83–159. Incidentally this paper also contains some results pertaining to binary codes (cf. *Information and Control* (**1,** 289–295 (1958)) and a proof of Shannon's fundamental theorem for the case of an abstract alphabet.

problems in mathematical statistics have been considered and information theory has been applied to obtain already known results and present them in a unified manner. The author considers the analysis of one or more samples from a multinomial population in Chapter 6 and the analysis for Poisson population in Chapter 7. Chapter 8 presents an analysis of contingency tables. Chapter 9 is concerned with multinomial normal populations for the tests of statistical hypotheses and is in the same spirit as Chapters 6–8. The problem of testing the general linear hypotheses (for normal populations) is dealt with in Chapter 10, and in Chapters 11 and 12 the corresponding problems for multivariate linear hypotheses and hypotheses other than linear have been considered respectively. Chapter 13 deals with linear discriminant functions and some issues for further investigation have been raised. The author's claim of a unified treatment is especially borne out by the material in Chapters 8, 11, and 12.

In all chapters numerous examples have been worked out and at the end of each chapter several exercises have been set. These examples and exercises often bring out small points which have not been treated in the body of the text. They also enable one to see how the material can possibly be applied. The bibliography is very exhaustive.

For a statistician both the parts of this book are important. In fact the second part, in a way, justifies the first. For a communication engineer the first part will be interesting and useful, and it will naturally tempt him to extend and apply this material to communication theory (which may include applications to more general stochastic processes, including sequential analysis). I feel that such applications are possible and the work of Pérez (*9*) may be considered as an example.

Aravind K. Joshi
University of Pennsylvania
Philadelphia, Pennsylvania

References

1. Chover, J. On normalized entropy and the extensions of a positive-definite function. (unpublished work).
2. Gelfand, I. M., Kolmogorov, A. N., and Yaglom, A. M. On the general definition of amount of information. *Doklady Akad. Nauk SSSR* **111,** 745–748 (1956).
3. Halmos, P. R. "Entropy in Ergodic Theory." Math. lecture notes, University of Chicago, 1959.
4. Khinchin, A. I. "Mathematical Foundations of Information Theory." Dover, New York, 1957.
5. Kolmogorov, A. N. A new metric invariant of transitive dynamical systems and automorphisms of a Lesbesgue space. *Doklady Akad. Nauk SSSR* **119,** 861–864, (1958); On the entropy per unit time as a metric invariant of automorphisms. *Doklady Akad. Nauk SSSR,* **124,** 754–755 (1959).
6. Lloyd, S. P. On common entropy. (unpublished work).
7. Mandelbrot, B. An outline of a purely phenomenological theory of statistical thermodynamics. *IRE Trans. on Information Theory* **2,** 190–203 (1956).

8. McMillan, B. The basic theorems of information theory. *Ann. Math. Stat.* **24,** 196–219 (1953).

9. Pérez, A. Notions généralisées d'incertitude, d'entropie et d'information du point de vue de la théorie de martingales. Sur le théorie de information dans de le cas d'un alphabet abstrait. *Trans. of the first Prague Conf. on Information Theory, Statistical Decision Functions, and Random Processes,* pp. 183–208, 209–243. Publishing House of the Czechoslovak Academy of Sciences, Prague, 1957.

10. Schützenberger, M. P., On measures of information used in statistics. *In* "Information Theory—Third London Symposium," pp. 18–23. Academic Press, New York, 1956.

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