NOTES

PROBABILITIES OF HYPOTHESES AND INFORMATION-STATISTICS IN SAMPLING FROM EXPONENTIAL-CLASS POPULATIONS

By Morton Kupperman

The George Washington University

1. Summary. This paper is concerned with inequalities connecting probabilities of hypotheses using Bayes' theorem (a posteriori probabilities), a priori probabilities, and Kullback-Leibler information-statistics in sampling from populations belonging to the exponential class of populations. As a corollary, it is shown that if it is known that the a priori probabilities are all equal, the choice of the hypothesis with the minimum Kullback-Leibler information-statistic is the same as the choice of the hypothesis with the maximum a posteriori probability, and conversely.

2. Introduction. Suppose that an event E can occur only if one of the set of r exhaustive and incompatible (mutually exclusive) events $H_1, H_2, \cdots, H_r$ occurs. The a priori probabilities of these latter events (which we may call hypotheses) are denoted by $\alpha_1, \alpha_2, \cdots, \alpha_r$ respectively, where $\alpha_m > 0$ and $\sum_{m=1}^{r} \alpha_m = 1$. The conditional probabilities for E to occur, assuming the occurrence of H_m, are denoted by $p(E \mid H_m)$, $m = 1, 2, \cdots, r$. The a posteriori probabilities of H_m, given that E has occurred, are denoted by $p(H_m \mid E)$. Bayes' theorem (see, for example, Uspensky [16]) states that

$$p(H_m \mid E) = \frac{\alpha_m p(E \mid H_m)}{\sum_{j=1}^{r} \alpha_j p(E \mid H_j)}, \qquad \text{for } m = 1, 2, \cdots, r.$$

A discrete multivariate and multiparameter population will be said to belong to the exponential class of populations (cf. Blackwell and Girshick [1] and Girshick and Savage [5]) if its probability distribution can be represented by

$$p(\mathbf{x}, \boldsymbol{\theta}) = q(\boldsymbol{\theta})\, r(\mathbf{x}) \exp\left\{\sum_{i=1}^{h} s_i(\boldsymbol{\theta})\, t_i(\mathbf{x})\right\},$$

where $\mathbf{x}$ is the row vector $\mathbf{x} = (x_1, x_2, \cdots, x_k)$, $\boldsymbol{\theta}$ is the row vector $\boldsymbol{\theta} = (\theta_1, \theta_2, \cdots, \theta_h)$, $q(\boldsymbol{\theta})$ and $r(\mathbf{x})$ are nonnegative functions of $\boldsymbol{\theta}$ and $\mathbf{x}$ respectively, and the parameter space is assumed to be an open convex set in an h-dimensional Euclidean space. We have k variates and h parameters, with the number of products in the exponent of e being h. Examples of discrete populations of the exponential class are the binomial distribution with the single parameter p,

Received September 25, 1956; revised November 2, 1957.

the Poisson distribution, the geometric distribution, the multinomial distribution, and the multivariate Poisson distribution.

Now consider r populations of the exponential class, each of the same functional form but differing only in their parameters. Let the probabilities be given by $p(\mathbf{x}, \boldsymbol{\theta}_m) > 0$, where $\sum \mathbf{x} p(\mathbf{x}, \boldsymbol{\theta}_m) = 1$ for $m = 1, 2, \cdots, r$. Suppose that we have a single random sample of N independent observations from one of these r populations (we do not know which population) and we wish to decide, on the basis of the sample values, which of the r populations is the most likely source of the sample. We shall use the term "most likely" in the sense of "having the largest a posteriori probability" and we shall assume that the a priori probabilities α_m are already known.

Let E denote the random sample and let $\hat{\boldsymbol{\theta}} = (\hat{\theta}_1, \cdots, \hat{\theta}_h)$ denote the maximum-likelihood estimate of $\boldsymbol{\theta}$.

3. Inequalities. The information measure $I(1:2)$ was introduced by Kullback and Leibler [12] as a generalization to the abstract case of a definition of information independently introduced in 1948 by Shannon [15] and by Wiener [17]. (See also Kullback [9], [10], and [11] for uses in statistics of $I(1:2)$. $I(1:2)$ has recently been termed "Kullback-Leibler information number" (Chernoff [3]) and "K-L information number" (Bradt and Karlin [2]).)

We obtain for two discrete populations of the exponential class

$$I(1:2) = \sum_{\mathbf{x}} p(\mathbf{x}, \boldsymbol{\theta}_1) \log \frac{p(\mathbf{x}, \boldsymbol{\theta}_1)}{p(\mathbf{x}, \boldsymbol{\theta}_2)}$$

$$= \log \frac{q(\boldsymbol{\theta}_1)}{q(\boldsymbol{\theta}_2)} + \sum_{i=1}^{h} \left[\left\{ s_i(\boldsymbol{\theta}_1) - s_i(\boldsymbol{\theta}_2) \right\} \cdot E_1 \left\{ t_i(\mathbf{x}) \right\} \right],$$

where the probabilities for the first population are given by $p(\mathbf{x}, \boldsymbol{\theta}_1)$, the probabilities for the second population are given by $p(\mathbf{x}, \boldsymbol{\theta}_2)$, and E_1 denotes expected values with respect to the first population. The logarithms are natural logarithms.

We now define the Kullback-Leibler information-statistic for a random sample of N independent observations from the mth population as

$$\hat{I}_m = N \sum_{\mathbf{x}} p(\mathbf{x}, \hat{\boldsymbol{\theta}}) \log \frac{p(\mathbf{x}, \hat{\boldsymbol{\theta}})}{p(\mathbf{x}, \boldsymbol{\theta}_m)}$$

$$= N \log \frac{q(\hat{\boldsymbol{\theta}})}{q(\boldsymbol{\theta}_m)} + N \sum_{i=1}^{h} \left[\left\{ s_i(\hat{\boldsymbol{\theta}}) - s_i(\boldsymbol{\theta}_m) \right\} \cdot \left(E \left\{ t_i(\mathbf{x}) \right\} \right)_{\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}} \right].$$

In $I(1:2)$, which is a functional of the vectors $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$ only, $\boldsymbol{\theta}_1$ has been replaced by the maximum-likelihood estimate $\hat{\boldsymbol{\theta}}$ and $\boldsymbol{\theta}_2$ has been replaced by the set of parameters $\boldsymbol{\theta}_m$ of the hypothetical mth population. The sum has been multiplied by N since the information measure for N independent observations is N times the information measure for a single observation.

The Kullback-Leibler information-statistic for samples from discrete populations of the exponential class (as well as for samples from more general statistical populations, discrete or continuous, univariate or multivariate, uniparameter or multiparameter) has useful applications in mathematical statistics. If we set up a null hypothesis that the given sample of N independent observations was randomly drawn from the specified mth population, then it can be shown that $2\hat{I}_m$ as defined above is asymptotically distributed as chi-square with h degrees of freedom when the null hypothesis is true (Kupperman [13], [14]).

We shall now show that the following inequality relationships exist connecting the a posteriori probabilities, the a priori probabilities, and the Kullback-Leibler information-statistics:

Theorem. *For two discrete populations H_m and H_n of the exponential class we have $p(H_m \mid E) \geqq p(H_n \mid E)$ if and only if $\hat{I}_m \leqq \hat{I}_n + \log (\alpha_m/\alpha_n)$, with both relations being equalities or strict inequalities simultaneously.*

Proof. From $p(H_m \mid E) \geqq p(H_n \mid E)$ we obtain, using Bayes' theorem and simplifying,

$$[q(\boldsymbol{\theta}_m)]^{-N} \exp \left\{ - \sum_{j=1}^{N} \sum_{i=1}^{h} s_i(\boldsymbol{\theta}_m)\, t_i(\mathbf{X}_j) \right\} \leqq [q(\boldsymbol{\theta}_n)]^{-N} \exp \left\{ - \sum_{j=1}^{N} \sum_{i=1}^{h} s_i(\boldsymbol{\theta}_n)\, t_i\,(\mathbf{X}_j) \right\} \cdot \frac{\alpha_m}{\alpha_n},$$

where $\mathbf{X}_j$ is the value of the jth observation on $\mathbf{x}$, $j = 1, 2, \cdots , N$. Now it can be shown (Kupperman [14]) that for populations of this class we have identically

$$\left(E\left\{ t_i(\mathbf{x}) \right\} \right)_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}} = \frac{1}{N} \sum_{j=1}^{N} t_i\,(\mathbf{X}_j).$$

(The discrete populations of the exponential class now being considered belong to the class of distributions admitting sufficient estimates of the parameters $\boldsymbol{\theta}$; these are distributions of the Koopman-Pitman type.) Hence by multiplying both sides of the inequality by the positive quantity

$$[q(\hat{\boldsymbol{\theta}})]^{N} \exp \left\{ N \sum_{i=1}^{h} s_i(\hat{\boldsymbol{\theta}}) \cdot \left(E\left\{ t_i(\mathbf{x}) \right\} \right)_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}} \right\}$$

and taking logarithms, we obtain

$$\hat{I}_m \leqq \hat{I}_n + \log \frac{\alpha_m}{\alpha_n}.$$

Since the steps of the proof are all reversible, the theorem is proved. The following corollary is an immediate consequence of this theorem:

Corollary. *If the a priori probabilities are all equal, the choice of the hypothesis (or hypotheses) with the minimum Kullback-Leibler information-statistic is the same as the choice of the hypothesis (or hypotheses) with the maximum a posteriori probability, and conversely.*

4. Continuous exponential-class populations. Although the preceding two sections are concerned with discrete populations, it may be remarked that the theorem and corollary may easily be extended to samples from continuous populations of the exponential class (such as the univariate normal distribution, the chi-square distribution, and the multivariate normal distribution with k means and $k(k + 1)/2$ parameters in the variance-covariance matrix). We make use of Bayes' theorem for continuous distributions (Kolmogorov [8], p. 46), use the likelihood of the observed sample instead of the probability of the observed sample, and follow the same steps as in the proof in the discrete case. The statements concerning $(E\{t_i(\mathbf{x})\})_{\theta=\hat{\theta}}$ and the asymptotic distribution of $2\hat{I}_m$ remain valid for continuous as well as discrete distributions of the exponential class.

5. Application. The theorem and the corollary are applicable to problems in which the a priori probabilities can be expressed in exact numerical form and thus the application of Bayes' theorem is legitimate, as, for example, in Mendelian hypotheses (see David [4], Chapter VIII).

In connection with the theorem and corollary, it may be remarked that the statements hold true if common logarithms (or logarithms to any base) are used in place of natural logarithms. This point is of importance, for in practical work common logarithms are more frequently used. However, in connection with the approximation of the large-sample distribution of $2\hat{I}$ by a chi-square distribution, it is important that natural logarithms be used, or that if common logarithms have been used $2\hat{I}$ be multiplied by $\log_e 10$, or 2.30259 approximately.

In conclusion, it may be remarked that if we were to use the corollary and decide always to accept the hypothesis for which $\hat{I}$ is the minimum without regard to the a priori probabilities involved, then we are in effect tacitly assuming that the a priori probabilities are equal, which is Bayes' postulate (as distinguished from Bayes' theorem).

The connection between information theory and inverse probability has been noted by Good [7], who is also concerned with the terminology and notation of information theory, particularly as it is applicable to communication theory. Reference should also be made to Good [6] for an informative discussion on Bayes' theorem and inverse probability.

6. Acknowledgment. The author wishes to thank Professor S. Kullback and the referees for suggesting the generalization incorporated in the results of this paper, which were derived originally for the special case of multinomial sampling.

REFERENCES

[1] D. Blackwell and M. A. Girshick, *Theory of Games and Statistical Decisions*, John Wiley & Sons, Inc., New York, 1954.

[2] R. N. Bradt and S. Karlin, "On the design and comparison of certain dichotomous experiments," *Ann. Math. Stat.*, Vol. 27 (1956), pp. 390–409.

[3] H. Chernoff, "Large-sample theory: Parametric case," *Ann. Math. Stat.*, Vol. 27 (1956), pp. 1–22.

[4] F. N. David, *Probability Theory for Statistical Methods*, Cambridge University Press, Cambridge, 1949.

[5] M. A. Girshick and L. J. Savage, "Bayes and minimax estimates for quadratic loss

functions," *Proceedings of the Second Berkeley Symposium on Mathematical Statistics and Probability*, University of California Press, Berkeley and Los Angeles, 1951, pp. 53–73.

[6] I. J. Good, *Probability and the Weighing of Evidence*, Charles Griffin & Company Limited, London, 1950.

[7] I. J. Good, "Some terminology and notation in information theory," *Monograph No. 155R*, The Institution of Electrical Engineers, London, 1955.

[8] A. Kolmogoroff, *Grundbegriffe der Wahrscheinlichkeitsrechnung*, Berlin, 1933.

[9] S. Kullback, "An application of information theory to multivariate analysis," *Ann. Math. Stat.*, Vol. 23 (1952), pp. 88–102.

[10] S. Kullback, "Certain inequalities in information theory and the Cramér-Rao inequality," *Ann. Math. Stat.*, Vol. 25 (1954), pp. 745–751.

[11] S. Kullback, "An application of information theory to multivariate analysis, II," *Ann. Math. Stat.*, Vol. 27 (1956), pp. 122–146.

[12] S. Kullback and R. A. Leibler, "On information and sufficiency," *Ann. Math. Stat.*, Vol. 22 (1951), pp. 79–86.

[13] M. Kupperman, "Further applications of information theory to multivariate analysis and statistical inference, (preliminary report)," (abstract), *Ann. Math. Stat.*, Vol. 27 (1956), p. 1184.

[14] M. Kupperman, "Further applications of information theory to multivariate analysis and statistical inference," Ph.D. dissertation on file at the library of The George Washington University, February 1957.

[15] C. E. Shannon, "A mathematical theory of communication," *Bell System Tech. J.*, Vol. 27 (1948), pp. 379–423 and 623–656.

[16] J. V. Uspensky, *Introduction to Mathematical Probability*, McGraw-Hill Book Company, Inc., New York, 1937.

[17] N. Wiener, *Cybernetics*, John Wiley & Sons, Inc., New York, 1948.

ON THE DISTRIBUTION OF 2 × 2 RANDOM NORMAL DETERMINANTS[1]

By W. L. Nicholson[2]

Princeton University

1. Summary. The c.d.f. of a 2 × 2 random determinant with mutually independent normally distributed entries is derived as an infinite series. Error functions that bound the tail of this series facilitate numerical calculation. Conditions are imposed on four variable quadratic forms for this distribution to apply. A normal approximation to the distribution is suggested.

2. Introduction. Let X_1, X_2, X_3 and X_4 be mutually independent random variables, each normally distributed, with means μ_1, μ_2, μ_3 and μ_4, and common variance σ^2. Let D be the random determinant,

$$D = \begin{vmatrix} X_1 X_2 \\ X_3 X_4 \end{vmatrix} = X_1 X_4 - X_2 X_3.$$

Received June 18, 1956; revised December 9, 1957.

[1] Work sponsored by the Office of Naval Research.

[2] Now employed by the General Electric Company, Richland, Washington.