AN ASYMPTOTIC LOWER BOUND FOR THE ENTROPY OF DISCRETE POPULATIONS WITH APPLICATION TO THE ESTIMATION OF ENTROPY FOR APPROXIMATELY UNIFORM POPULATIONS*)

E. B. COBB AND BERNARD HARRIS

(Received Jan. 6, 1965; revised July 30, 1966)

1. Introduction and summary

Assume that a random sample of size N has been drawn from a multinomial population with an unknown and perhaps countably infinite number of classes. That is, if X_j is the jth observation, and M_i the ith class, then

$$\mathrm{P}\{X_j \in M_i\}=p_i \geqq 0 \qquad i=1, 2, \cdots ; \; j=1, 2, \cdots, N$$

and $\sum_{i=1}^{\infty} p_i=1$. The classes are not assumed to have a natural ordering.

Let n_r be the number of classes which occur exactly r times in the sample. Then $\sum_{r=0}^{\infty} rn_r=N$.

Defining the entropy of the population by

$$(1) \qquad H(p_1, p_2, \cdots)=-\sum_{i=1}^{\infty} p_i \log p_i$$

we can show that for the cumulative distribution function $F^*(x)$ defined by

$$(2) \qquad F^*(x)=\sum_{Np_j \leqq x} Np_j e^{-Np_j} \Big/ \Big(\sum_{j=1}^{\infty} Np_j e^{-Np_j} \Big)$$

we have

$$(3) \qquad H(p_1, p_2, \cdots) \sim \frac{1}{N} \mathrm{E}(n_1) \int_{-\infty}^{\infty} e^x \log\left(\frac{N}{x}\right) dF^*(x) .$$

In addition, in Harris [1], it is shown that the moments of $F^*(x)$,

*) This research was sponsored by the Mathematics Research Center, U.S. Army, University of Wisconsin, under Contract No. DA-11-022-ORD-2059.

$\mu_1, \mu_2, \cdots$, are approximately given by

$$\mu_r \sim \frac{(r+1)!\,\mathrm{E}(n_{r+1})}{\mathrm{E}(n_1)}\,, \tag{4}$$

where

$$\mathrm{E}(n_r) \sim \frac{1}{r!} \sum_{j=1}^{\infty} (Np_j)^r e^{-Np_j}\,.$$

Let $\mathfrak{F}^{[a,b]}_{(\mu_1,\mu_2,\cdots,\mu_k)}$ be the set of cumulative distribution functions with $F(a-0)=0$, $F(b)=1$ and

$$\int_{-\infty}^{\infty} x^j\,dF(x)=\mu_j\,, \qquad j=1,2,\cdots,k\,.$$

It is clear that $F^*(x)$ must be an element of $\mathfrak{F}^{[0,N]}_{(\mu_1,\mu_2,\cdots,\mu_k)}$ for every k. Hence, for every k,

$$\min_{F\in\mathfrak{F}^{[0,N]}_{(\mu_1,\mu_2,\cdots,\mu_k)}} \frac{\mathrm{E}(n_1)}{N}\int_{-\infty}^{\infty} e^x \log\left(\frac{N}{x}\right) dF(x)$$

is an asymptotic (in N) lower bound for $H(p_1, p_2, \cdots)$. In general, the supremum will not exist.

In this paper, the minimum is explicitly computed for $k=2$, thus obtaining what may be regarded as an asymptotic "Tschebycheff type inequality" for the entropy of the population.

In particular, if the set of cumulative distribution functions $\mathfrak{F}^{[0,N]}_{(\mu_1,\mu_2,\cdots,\mu_k)}$ consists of only one element, then the solution to the minimizing problem must be $F^*(x)$. Similarly, if "most of the information provided by the moments" is concentrated in $\mu_1, \mu_2, \cdots, \mu_k$, then the solution to the minimizing problem should provide an approximation to $H(p_1, p_2, \cdots)$.

This suggests the following application. Replace the expected values in (4) by the observed values by defining

$$m_r = \frac{(r+1)!\,n_{r+1}}{n_1}\,,$$

and thus estimates of the moments of $F^*(x)$ are obtained. Then let $\mathfrak{F}^{[a,b]}_{(m_1,m_2,\cdots,m_k)}$ be the set of cumulative distribution functions with $F(a-0)=0$, $F(\mathrm{b})=1$, and

$$\int_{-\infty}^{\infty} x^j\,dF(x)=m_j\,, \qquad j=1,2,\cdots,k\,.$$

Since $p_1, p_2, \cdots$ are all assumed to be unknown, $F^*(x)$ is unknown and an asymptotic lower bound to (3) may be found by minimizing

$$\int_{-\infty}^{\infty} e^x \log\left(\frac{N}{x}\right) dF(x)$$

over the set $\mathfrak{F}^{[0,N]}_{(m_1,m_2,\cdots,m_k)}$. This process uses only the information contained in the first $k+1$ occupancy numbers $n_1, n_2, \cdots, n_{k+1}$, and is particularly useful, when the sample information concerning the parameters $p_1, p_2, \cdots$ is concentrated in the low order occupancy numbers. This occurs, for example, if as $N \to \infty$, $p_j \to 0$, $j=1,2,\cdots$, in such a way that $0 \leqq Np_j < \lambda$, where λ is approximately $k+1$.

The maximum likelihood estimator $\hat{H}$ is a good estimator of $H(p_1, p_2, \cdots)$ if there is an integer M, such that for N sufficiently large, $Np_i \to \infty$, $i=1,2,\cdots,M$ and in addition, for sufficiently small $\varepsilon>0$,

$$\sum_{i=M+1}^{\infty} p_i \log p_i < \varepsilon .$$

It will perform rather poorly when the p_i's are uniformly small. On these heuristic grounds the above procedure is suggested for this case. We will exhibit this for uniform populations in several examples given in this paper.

In subsequent papers, the problem of non-parametric estimation of entropy for arbitrary discrete populations will be discussed as well as the uses of entropy estimates in non-parametric testing of hypotheses.

2. The computation of the lower bound for entropy

In Harris [1], it was shown that for $r^2=O(N)$ as $N \to \infty$,

$$\text{(5)} \qquad \mathrm{E}(n_r) \sim \frac{1}{r!} \sum_{j=1}^{\infty} (Np_j)^r e^{-Np_j} ,$$

where the approximation is valid, in the sense that either both sides are negligible, or the ratio of the two sides approaches unity.

In particular,

$$\text{(6)} \qquad \mathrm{E}(n_1) \sim \sum_{j=1}^{\infty} Np_j e^{-Np_j} ;$$

hence

$$\frac{1}{N} \mathrm{E}(n_1) \int_{-\infty}^{\infty} e^x \log\left(\frac{N}{x}\right) dF^*(x) \sim \frac{1}{N} \sum_{j=1}^{\infty} e^{Np_j} \log\left(\frac{1}{p_j}\right) Np_j e^{-Np_j}$$
$$= H(p_1, p_2, \cdots) .$$

Let $h(x)=e^x \log(N/x)$. Then we wish to determine $F_0(x) \in \mathfrak{F}^{[0,N]}_{(m_1,m_2)}$ such that

$$\min_{F(x)\in\mathfrak{F}^{[0,N]}_{(m_1,m_2)}}\int_{-\infty}^{\infty}h(x)\,dF(x)=\int_{-\infty}^{\infty}h(x)\,dF_0(x)\ . \tag{7}$$

Since $h(0)$ does not exist, we consider instead $\mathfrak{F}^{[\varepsilon,N]}_{(m_1,m_2)}$, where $\varepsilon>0$, is arbitrary. Then $h(x)$ is bounded on $[\varepsilon, N]$ for every $\varepsilon>0$ and it is well-known [1] that $F_\varepsilon(x)$ defined by

$$\min_{F(x)\in\mathfrak{F}^{[\varepsilon,N]}_{(m_1,m_2)}}\int_{-\infty}^{\infty}h(x)\,dF(x)=\int_{-\infty}^{\infty}h(x)\,dF_\varepsilon(x)\ , \tag{8}$$

is obtainable as a discrete cumulative distribution function with at most three jumps, say at x_1, x_2, x_3, $\varepsilon\leqq x_1<x_2<x_3\leqq N$. Hence, there exists $\lambda_1, \lambda_2, \lambda_3\geqq 0$, $\sum_{i=1}^{3}\lambda_i=1$, with

$$\begin{cases}\lambda_1x_1+\lambda_2x_2+\lambda_3x_3=m_1\\ \lambda_1x_1^2+\lambda_2x_2^2+\lambda_3x_3^2=m_2\ ,\end{cases} \tag{9}$$

such that

$$F_\varepsilon(x)=\begin{cases}0\ , & x<x_1\\ \lambda_1\ , & x_1\leqq x<x_2\\ \lambda_1+\lambda_2\ , & x_2\leqq x<x_3\\ 1\ , & x\geqq x_3\end{cases} \tag{10}$$

whenever $m_2\geqq m_1^2$, a condition which we will assume throughout the remainder of this discussion. Without loss of generality, we may assume that $m_2>m_1^2$, since otherwise $F_\varepsilon(x)$ is a cumulative distribution function with exactly one jump, and (8) has a trivial solution.

It can be shown that $\lambda_i\geqq 0$, $i=1, 2, 3$, if and only if

$$(-1)^{i+j-1}(x_ix_j-m_1(x_i+x_j)+m_2)\geqq 0\ , \qquad 1\leqq i<j\leqq 3\ . \tag{11}$$

In addition, from Harris [1], there exist real numbers $\alpha_0, \alpha_1, \alpha_2$ such that x_1, x_2 and x_3 are roots of

$$g(x)=\sum_{i=0}^{2}\alpha_ix^i-h(x)=0\ , \tag{12}$$

and

$$\sum_{i=0}^{2}\alpha_ix^i-h(x)\leqq 0\ , \qquad \varepsilon\leqq x\leqq N\ . \tag{13}$$

From (11) and (12), we also have that for $\varepsilon<x_i<N$, $i=1, 2, 3$;

$$g'(x_i)=\alpha_1+2\alpha_2x_i-h'(x_i)=0\ . \tag{14}$$

To solve (9), (12), (13) and (14), observe that there exist numbers

$\delta_1, \delta_2, \delta_3$, $0<\delta_1<\delta_2<\delta_3<N$, such that

$$h'(x)\begin{cases} <0\,, & 0<x<\delta_1\,, \\ >0\,, & \delta_1<x<\delta_3\,, \\ <0\,, & \delta_3<x\leqq N\,, \end{cases}$$

and

$$h''(x)\begin{cases} >0\,, & 0<x<\delta_2 \\ <0\,, & \delta_2<x\leqq N \end{cases}$$

with

$$\begin{aligned} &\delta_1\to 0 &&, \quad N\to\infty \\ &\delta_2=(N-2)+O\left(\frac{1}{N}\right) &&, \quad N\to\infty \\ &\delta_3=(N-1)+O\left(\frac{1}{N}\right) &&, \quad N\to\infty \end{aligned}$$

and $h''(x)$ is strictly decreasing on $(0,\delta_1)$ and (δ_2, N). We now establish the following

LEMMA. *If* $\varepsilon<x_1<x_2<N$ $(0<\varepsilon<\delta_1)$, *the following conditions cannot be satisfied simultaneously*

$$\sum_{i=0}^{2}\alpha_i x^i \leqq h(x)\,, \qquad \varepsilon\leqq x\leqq N \tag{15}$$

$$\sum_{i=0}^{2}\alpha_i x_j^i = h(x_j)\,, \qquad j=1, 2\,. \tag{16}$$

PROOF. Assume (15) and (16) hold. Let $p(x)=\sum_{i=0}^{2}\alpha_i x^i$. Then

$$h'(x_j)=p'(x_j)\,, \qquad j=1, 2\,. \tag{17}$$

Let $I_1=(\varepsilon, \delta_1]$, $I_2=(\delta_1, \delta_2]$, $I_3=(\delta_2, N)$. Assume $\alpha_2>0$. Then if $x_2\in I_3$, since $p(x)$ is strictly convex and $h(x)$ is strictly concave in I_3, by (16) and (17), we have $p(x_0)>h(x_0)$ for some $x_0\in I_3$, contradicting (15). If $x_2\in I_2$, then $p'(x_2)>0$, hence $p(N)>p(x_2)>0=h(N)$, contradicting (15). If $x_2\in I_1$, then $\varepsilon<x_1<x_2\leqq\delta_1$, and by (16) and Rolle's Theorem, there exist ξ_1, ξ_2, $x_1<\xi_1<\xi_2<x_2$ such that $g''(\xi_j)=0$, $j=1, 2$. This, however, implies that $h''(\xi_j)=2\alpha_2$, $j=1, 2$, contradicting the monotonicity of $h''(x)$. If $\alpha_2<0$, the argument is similar. The case $\alpha_2=0$ is trivial. We now obtain $F_0(x)$.

THEOREM 1. *There exists a unique cumulative distribution function* $F_0(x)\in\mathfrak{F}^{[0,N]}_{(m_1,m_2)}$ *such that*

$$\int_{-\infty}^{\infty} h(x)\,dF_0(x) = \min_{F(x)\in\mathfrak{F}^{[0,N]}_{(m_1,m_2)}} \int_{-\infty}^{\infty} h(x)\,dF(x)$$

given by

$$F_0(x) = \begin{cases} 0 , & x < \dfrac{Nm_1 - m_2}{N - m_1} \\ \dfrac{(N-m_1)^2}{(N-m_1)^2 + (m_2 - m_1^2)} , & \dfrac{Nm_1 - m_2}{N-m_1} \leqq x < N \\ 1 , & x \geqq N . \end{cases} \tag{18}$$

PROOF. By the above lemma, we have $x_1 = \varepsilon$, $\varepsilon < x_2 < N$, $x_3 = N$. From (11), we have

$$\frac{Nm_1 - m_2}{N - m_1} \leqq x_2 \leqq \frac{m_2 - m_1\varepsilon}{m_1 - \varepsilon} . \tag{19}$$

Thus, by (9), we have

$$\lambda_1(x_2, \varepsilon) = \frac{Nx_2 - m_1(N + x_2) + m_2}{(x_2 - \varepsilon)(N - \varepsilon)}$$

$$\lambda_2(x_2, \varepsilon) = \frac{-(\varepsilon N - m_1(N + \varepsilon) + m_2)}{(x_2 - \varepsilon)(N - x_2)} ,$$

and

$$\lim_{\varepsilon\to 0} \lambda_1(x_2, \varepsilon) = \frac{Nx_2 - m_1(N + x_2) + m_2}{x_2 N} ,$$

$$\lim_{\varepsilon\to 0} \lambda_2(x_2, \varepsilon) = \frac{Nm_1 - m_2}{x_2(N - x_2)} .$$

This gives a parametric family of cumulative distribution functions $F_{0,x_2}(x)$. Since $\lim_{x\to 0+} h(x) = \infty$, we must have $\lambda_1(x_2, \varepsilon) = O\left(\dfrac{1}{h(\varepsilon)}\right)$, $\varepsilon \to 0$, since otherwise $F_\varepsilon(x)$ would not satisfy (8). Hence $\lim_{\varepsilon\to 0} \lambda_1(x_2, \varepsilon) = 0$ and $x_2 \to \dfrac{Nm_1 - m_2}{N - m_1}$ as $\varepsilon \to 0$. Since $\lambda_1(x_2, \varepsilon)h(\varepsilon) \geqq 0$ for every $\varepsilon > 0$, it follows that $\lambda_1(x_2, \varepsilon) = o\left(\dfrac{1}{h(\varepsilon)}\right)$ as $\varepsilon \to 0$, establishing the theorem.

Finally we have:

THEOREM 2. *The required lower bound for the entropy is*

$$\frac{n_1}{N} \int_{-\infty}^{\infty} h(x)\,dF_0(x)$$

$$= \frac{n_1}{N} \frac{(N-m_1)^2}{(N-m_1)^2 + (m_2 - m_1^2)} \exp\left(\frac{Nm_1 - m_2}{N - m_1}\right) \log \frac{N(N - m_1)}{Nm_1 - m_2} .$$

Remark. Krein [2] has studied minimization problems similar to (8). However, Krein's methods require that 1, x, x^2, $h(x)$ form a Tschebycheffian system of functions on $[\varepsilon, N]$. A necessary condition for the above (see Pólya and Szegö [3]) is that the Wronskians

$$W(x)=\begin{vmatrix} 1 & x & x^2 & h(x) \\ 0 & 1 & 2x & h'(x) \\ 0 & 0 & 2 & h''(x) \\ 0 & 0 & 0 & h'''(x) \end{vmatrix}, \qquad \varepsilon \leqq x \leqq N,$$

be non-negative (non-positive) on $[\varepsilon, N]$. This condition is clearly not satisfied in this case and Krein's methods are therefore inapplicable.

3. The estimation of the entropy of uniform populations

Let

$$p_j=\begin{cases} \dfrac{1}{M} & j=1, 2, \cdots, M \\ 0 & \text{otherwise}. \end{cases}$$

Then,

$$F^*(x)=\begin{cases} 0 & x<N/M \\ 1 & x\geqq N/M \end{cases}$$

where

$$N\to\infty, \quad M\to\infty \qquad \text{so that} \quad N/M\to\lambda>0 .$$

In addition,

$$\mathrm{E}(n_r)\sim\frac{M}{r!}\lambda^r e^{-\lambda} \qquad r=1, 2, \cdots$$

and

$$\mu_r=\lambda^r \qquad r=1, 2, \cdots .$$

In this case,

$$\frac{1}{N}\,\mathrm{E}(n_1)\int_{-\infty}^{\infty} h(x)\,dF^*(x)=e^{-\lambda}h(\lambda)=\log M$$

as required.

In addition, the class $\mathfrak{F}^{[0,N]}_{(\mu_1,\mu_2)}$ contains only $F^*(x)$, so that the solution of (7) provides an estimate of $H(p_1, p_2, \cdots)$ rather than a lower bound.

, In the replacement of μ_1, μ_2 by the sample quantities m_1, m_2, it may happen that $m_2 < m_1^2$. This, of course suggests that $F^*(x)$ is degenerate, and in such cases, we take $m_2 = m_1^2$.

By way of contrast, the maximum likelihood estimate $\hat{H}$ is poor under the limiting process employed here, since

$$\mathrm{E}(\hat{H}) = \sum_{i=1}^{N} \mathrm{E}(n_i) \frac{i}{N} \log\left(\frac{i}{N}\right)$$

and for $M=1000$, $N=100$, we have $\mathrm{E}(n_1)=90.48$, $\mathrm{E}(n_2)=4.52$, $\mathrm{E}(n_3)=.15$ obtaining

$$\mathrm{E}(\hat{H}) = 4.271$$

and $H = \log M = 6.908$.

The examples given below are intended to exhibit the application of this procedure in estimating entropy.

Example 1. Three random samples were chosen with $M=1000$, $N=100$. The data are summarized below.

	Sample # 1	Sample # 2	Sample # 3
n_1	85	94	92
n_2	6	3	4
n_3	1	—	—
m_1	.1412	.0638	.0870
m_2	.0706	—	—
$\frac{n_1}{N}\int h(x)\,dF_0(x)$	6.4246	7.3714	7.0726
$H(p_1, \cdots, p_M)$	6.9076	6.9076	6.9076
$\hat{H}$	4.4890	4.5636	4.5497

Example 2. Three random samples were chosen with $N=1000$, $M=1000$. The data are summarized below.

	Sample # 1	Sample # 2	Sample # 3
n_1	373	341	377
n_2	199	179	169
n_3	62	70	60
n_4	8	17	25
n_5	1	2	1

n_6	1	1	0
n_7	0	1	0
m_1	1.067	1.050	.897
m_2	.997	1.232	.955
$\frac{n_1}{N}\int h(x)\,dF_0(x)$	7.419	6.683	6.486
$H(p_1, \cdots, p_M)$	6.908	6.908	6.908
$\hat{H}$	6.364	6.294	6.329

In samples #2 and #3 of example 1 and in sample #1 of example 2, $m_2 < m_1^2$ in those three instances.

These examples suggest strongly the superiority of $\frac{n_1}{n}\int h(x)\,dF_0(x)$ as an estimator of H in comparison with the maximum likelihood estimator $\hat{H}$. In fact, the trivial observation that $\hat{H}$ can never exceed $\log N$ can easily be made.

Mathematics Research Center, U.S. Army, University of Wisconsin[*]

References

[1] B. Harris, "Determining bounds on integrals with applications to cataloging problems," *Ann. Math. Statist.*, 30 (1959), 521-548.

[2] M. G. Krein, "The ideas of P. L. Čebyšev and A. A. Markov in the theory of limiting values of integrals and their further development," (In Russian), *Uspehi Mat. Nauk* (N. S.), 6 (1951), No. 4 (44), 3-120; in English: *Amer. Math. Soc. Translations Ser.* 2, 12 (1959), 1-121.

[3] G. Pólya and G. Szegö, *Aufgaben und Lehrsätze aus der Analysis.* Bd. II, Springer, Berlin, 1925.

[*] E. B. Cobb is now at the University of Kansas.