AD-A020 217

THE STATISTICAL ESTIMATION OF ENTROPY IN THE NON-PARAMETRIC CASE

Bernard Harris

Wisconsin University

Prepared for:

Army Research Office

December 1975

MRC Technical Summary Report #1605

THE STATISTICAL ESTIMATION OF ENTROPY IN THE NON-PARAMETRIC CASE

Bernard Harris

Mathematics Research Center
University of Wisconsin–Madison
610 Walnut Street
Madison, Wisconsin 53706

December 1975

Received October 8, 1975

Sponsored by

U. S. Army Research Office
P. O. Box 12211
Research Triangle Park
North Carolina 27709

UNIVERSITY OF WISCONSIN - MADISON
MATHEMATICS RESEARCH CENTER

THE STATISTICAL ESTIMATION OF ENTROPY IN THE NON-PARAMETRIC CASE

Bernard Harris

A

Technical Summary Report #1605
December 1975

ABSTRACT

Assume that N mutually independent observations have been taken from the population specified by

$$P\{X_i \epsilon M_j\} = p_j, \qquad i = 1, 2, \ldots, N, \quad j = 1, 2, \ldots$$

where X_i denotes the ith observation and M_j denotes the jth class. The classes are not assumed to have a natural ordering. Then the entropy is defined by

$$H = -\sum_j p_j \log p_j \ .$$

The natural estimator $\hat{H} = -\sum_j \hat{p}_j \log \hat{p}_j$ is shown to have certain deficiencies when the number of classes is large relative to the sample size or is infinite. A procedure based on quadrature methods is proposed as a means of circumventing these deficiencies.

AMS(MOS) Subject Classification: 62G05, 94A15

Key Words: Estimation of entropy

Sponsored by the United States Army under Contract No. DAAG29-75-C-0024.

THE STATISTICAL ESTIMATION OF ENTROPY IN THE NON-PARAMETRIC CASE

Bernard Harris

1. Introduction and Summary. Assume that a random sample of size N has been drawn from a "multinomial population" with an unknown and possibly countably infinite number of classes. That is, if X_i is the ith observation and M_j is the jth class, then

$$(1) \qquad P\{X_i \in M_j\} = p_j \geq 0, \quad j = 1, 2, \ldots, \quad i = 1, 2, \ldots, N \ ,$$

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

In such statistical populations, the entropy, defined by

$$(2) \qquad H = H(p_1, p_2, \ldots) = -\sum_{j=1}^{\infty} p_j \log p_j$$

is a natural parameter of interest. For technical reasons, natural logarithms will be employed throughout, rather than the more customary base 2 logarithms. This modification is equivalent to a change of scale and will have no essential effect on the subsequent discussion. We also assume throughout $H < \infty$. Some examples for which $H = \infty$ are given in Appendix 4.

Some concrete examples for which the entropy is a natural parameter are the frequencies of words in a language and the frequencies of species of plants or insects in a region. For such populations, the entropy may be regarded as a natural measure of heterogeneity. Many other measures of

Sponsored by the United States Army under Contract No. DAAG29-75-C-0024.

heterogeneity depend on the classes being numerically indexed, which is a stronger assumption than having a natural ordering.

We define the random variables Y_{ij}, $i = 1, 2, \ldots, N$; $j = 1, 2, \ldots$ by

$$(3) \qquad Y_{ij} = \begin{cases} 1 & \text{if} \quad X_i \in M_j \ , \\ 0 & \text{otherwise.} \end{cases}$$

Then

$$\sum_{ij} Y_{ij} = N$$

and

$$\sum_{i=1}^{N} Y_{ij} = Z_j$$

is the number of observations in the jth class.

The "natural" estimator of H, denoted by $\hat{H}$, where

$$(4) \qquad \hat{H} = -\sum_{j=1}^{\infty} \hat{p}_j \log \hat{p}_j$$

and

$$(5) \qquad \hat{p}_j = Z_j / N, \qquad j = 1, 2, \ldots \ ,$$

has been studied extensively for the case where the number of classes for which $p_j > 0$ is known and finite. We denote the number of such classes by s in this case and assume that these classes are indexed by $1, 2, \ldots, s$. Then, G. A. Miller and W. G. Madow [9] showed that the limiting $(N \to \infty)$

distribution of $\sqrt{N}\,(\hat{H} - H)$ is normally distributed with mean zero and variance $\sigma^2 = \sum_{j=1}^{s} p_j(\log p_j + H)^2$, provided that not all $p_j = 1/s$. They also showed that if $p_j = 1/s$, $j = 1, 2, \ldots, s$, then $2N(H - \hat{H})$ has a limiting chi-square distribution with $s - 1$ degrees of freedom. The Miller-Madow paper is summarized in R. D. Luce [7]. An asymptotic evaluation of $E(\hat{H} - H)$ is given in G. A. Miller [8]. The above results also appear as special cases of the more general problem of obtaining the limiting distribution of the amount of transmitted information, studied by Z. A. Lomnicki and S. K. Zaremba [6]. Subsequently G. P. Bašarin [1] also obtained the asymptotic mean and variance of $\hat{H}$ and determined the limiting normal distribution as above, however, he failed to note that if $p_j = 1/s$, $j = 1, 2, \ldots, s$, then $\sqrt{N}\,(H - \hat{H})$ does not have a proper limiting distribution. Note that in this case,

$$\sum_{j=1}^{s} p_j(\log p_j + H)^2 = 0 \quad .$$

The paper by G. P. Bašarin was subsequently generalized by A. M. Zubkov [10], who permitted $p_1, p_2, \ldots, p_s$ and s to depend on N in such a way that for some $\varepsilon > \delta > 0$, if

$$\frac{N^{1-\varepsilon}}{s}\left[\sum_{j=1}^{s} p_j \log^2 p_j - H^2\right] \to \infty$$

as $N \to \infty$ and $\max_{1 \le j \le s} (Np_j)^{-1} = O(s/N^{1-\delta})$, then $\sqrt{N}(\hat{H} - E\hat{H})/(\Sigma\, p_j \log^2 p_j - H^2)^{\frac{1}{2}}$ had a limiting standard normal distribution. He also showed that if s is fixed, then $2N(H - \hat{H})$ has a limiting chi-square distribution when

$\max_{1 \leq j \leq s} |p_j - s^{-1}| = o(N^{-\frac{1}{2}})$. In particular, note that in Zubkov's theorem, he considered $\hat{H} - E\hat{H}$ rather than $\hat{H} - H$ and required the additional condition that $s/\sqrt{N(\sum_j p_j \log^2 p_j - H^2)} \to 0$ as $N \to \infty$ in order to replace $E\hat{H}$ by H in the statement of his theorem. This last condition will be violated in many of the applications for which the present technique is intended. In Section 2 we will study the behavior of $\hat{H}$; here we observe that for the problem at hand, $\hat{H}$ has certain deficiencies. Roughly speaking, if too much of the probability is distributed over classes with "small p_j's", $\hat{H}$ will not be a satisfactory estimator. A method for circumventing some of these difficulties is given in Section 3. The alternatives presented here are arrived at through intuitive considerations and a detailed picture of their statistical behavior is not available at present. Some preliminary empirical investigations are presented to suggest the utility of the proposed techniques.

2. Properties of $\hat{H}$. Here we present a somewhat refined version of some of the Bašarin, Miller-Madow results. The refinement is needed to connect one known error in Bašarin's paper and to also revise his computation of the asymptotic variance of $\hat{H}$, which is inadequate when $p_1 = p_2 = \ldots = p_s = 1/s$ and $p_j = 0$, $j > s$.

Bašarin considered a multinomial population with a known finite number of classes, that is, we have $p_j > 0$, $j = 1, 2, \ldots, s$ and $p_j = 0$, $j > s$. For the present, we adopt this assumption. Then, expanding in a Taylor series, we can write

(6) $$\hat{H} = H - \sum_{j=1}^{s} (\hat{p}_j - p_j) \log p_j + \sum_{m=2}^{r} \frac{(-1)^{m-1}}{m(m-1)} \sum_{j=1}^{s} \frac{(\hat{p}_j - p_j)^m}{p_j^{m-1}} + R_{r+1} ,$$

where

(7) $$R_{r+1} = \frac{(-1)^r}{r(r+1)} \sum_{j=1}^{s} \frac{(\hat{p}_j - p_j)^{r+1}}{\xi_j^r}$$

and

(8) $$\xi_j = \lambda_j p_j + (1-\lambda_j)\, \hat{p}_j, \qquad 0 < \lambda_j < 1 \ .$$

From (6), for fixed j, $1 \leq j \leq s$, we have

(9) $$-\hat{p}_j \log \hat{p}_j + \hat{p}_j \log p_j - \sum_{m=2}^{r} \frac{(-1)^{m-1}}{m(m-1)} \frac{(\hat{p}_j - \hat{p}_j)^m}{p_j^{m-1}} = R_{r+1,\, j}$$

and

$$R_{r+1} = \sum_{j=1}^{s} R_{r+1,\, j} \ .$$

Then for any ε, $0 < \varepsilon < 1$ and $|\hat{p}_j - p_j| < (1-\varepsilon)p_j$, we can write

$$R_{r+1,\, j} = \sum_{m=r+1}^{\infty} \frac{(-1)^{m-1}}{m(m-1)} \frac{(\hat{p}_j - p_j)^m}{p_j^{m-1}}$$

and

$$|R_{r+1,j}| \le \sum_{m=r+1}^{\infty} |\hat{p}_j - p_j|^m / p_j^{m-1}$$

$$\le \frac{|\hat{p}_j - p_j|^{r+1}}{p_j^r} \Bigg/ \left(1 - \frac{|\hat{p}_j - p_j|}{p_j}\right)$$

$$\le \frac{|\hat{p}_j - p_j|^{r+1}}{\varepsilon p_j^r} \quad .$$

Now let $A_\varepsilon(p_j, \hat{p}_j) = \{\hat{p}_j \colon |\hat{p}_j - p_j| \ge (1-\varepsilon)p_j,\ \ 0 \le \hat{p}_j \le 1\}$. Then from (7) and (8)

$$R_{r+1,j} = R_{r+1,j}(\hat{p}) = \frac{(-1)^r}{r(r+1)} \frac{(\hat{p}_j - p_j)^{r+1}}{\xi_j^r}$$

and since $0 < p_j < 1$, $0 < \xi_j < 1$ and $R_{r+1,j} = 0$ if and only if $\hat{p}_j = p_j$. Thus on $A_\varepsilon(p_j, \hat{p}_j)$, $R_{r+1,j} \ne 0$. Consequently,

$$\lambda_j = \left\{ \left[\frac{(-1)^r}{r(r+1)} \frac{(\hat{p}_j - p_j)^{r+1}}{R_{r+1,j}} \right]^{\frac{1}{r}} - \hat{p}_j \right\} \Bigg/ (p_j - \hat{p}_j) \quad .$$

Now $A_\varepsilon(p_j, \hat{p}_j)$ is a compact set and $\lambda_j = \lambda_j(\hat{p}_j)$ is a continuous function of $\hat{p}_j$ on that set. Thus $\min\limits_{\hat{p}_j \epsilon A(\hat{p}_j)} \lambda_j(\hat{p}_j)$ is attained and is positive. Hence

$\min\limits_{1 \le j \le s} \min\limits_{\hat{p}_j \epsilon A(\hat{p}_j)} \lambda_j(\hat{p}_j) = \lambda_s^* > 0$. Further note that λ_s^* is independent of N .

Hence define

(10) $$\lambda^* = \min(\lambda_s^*, \varepsilon^{\frac{1}{r}}) \quad .$$

To understand the behavior of H and to motivate the subsequent decision, we proceed to obtain asymptctic estimates of the mean and variance of $\hat{H}$ by employing (6). To facilitate the evaluation of these expected values, a tabulation of some required auxiliary formulas and some comments concerning them are contained in Appendix 1 to this paper. In fact, we provide somewhat more formulas than are actually needed, since both the Bašarin paper [1] and the book by F. N. David and D. E. Barton [3, page 146] contain some misprints or errors, also these formulas have frequent applications in problems dealing with multinomial distributions and hopefully will prove to be useful in further studies in the direction on the present paper.

From (6) and (A.1.1-A.1.6), we have

$$(11) \qquad E\hat{H} = H + \sum_{m=2}^{r} \frac{(-1)^{m-1}}{m(m-1)} \sum_{j=1}^{s} \frac{1}{p_j^{m-1}} \; E\{(\hat{p}_j - p_j)^m\} + E R_{r+1} \; .$$

Then letting $\mu_m(j) = E\{(Z_j - Np_j)^m\}$ and noting that $E\{(\hat{p}_j - p_j)^m\} = \mu_m(j)/N^m$, we have

$$(12) \qquad E\hat{H} = H + \sum_{m=2}^{r} \frac{(-1)^{m-1}}{m(m-1)} \sum_{j=1}^{s} \frac{\mu_m(j)}{N^m \, p_j^{m-1}} + E R_{r+1} \; .$$

From (7), (8), (9) and (10), we have

$$|R_{r+1}| \le (r(r+1))^{-1} \sum_{j=1}^{s} \frac{|(\hat{p}_j - p_j)|^{r+1}}{\xi_j^r}$$

$$\le \frac{1}{r(r+1)} \sum_{j=1}^{s} \frac{|\hat{p}_j - p_j|^{r+1}}{\lambda^{*r} \, p_j^r} \; .$$

Consequently,

$$|E R_{r+1}| \le E|R_{r+1}| \le \frac{1}{r(r+1)} \sum_{j=1}^{s} \frac{E|\hat{p}_j - p_j|^{r+1}}{\lambda^{*^r} p_j^r} ,$$

and if r is an odd integer ≥ 1 ,

$$|E R_{r+1}| \le \frac{1}{r(r+1)N^{r+1}} \sum_{j=1}^{s} \mu_{r+1}(j) / \lambda^{*^r} p_j^r$$

Thus, from (A.1.16), for r an odd integer ≥ 1, we have

(13) $$|E R_{r+1}| = O(N^{-(r+1)/2}) .$$

Specifically, using (A.1.1)-(A.1.5) and (13), for $r = 5$, we get,

$$E\hat{H} = H - \frac{1}{2N} \sum_{j=1}^{s} \frac{(p_j - p_j^2)}{p_j} + \frac{1}{6N^2} \sum_{j=1}^{s} \frac{(p_j - 3p_j^2 + 2p_j^3)}{p_j^2}$$

$$- \frac{1}{4N^2} \sum_{j=1}^{s} \frac{(p_j^2 - 2p_j^3 + p_j^4)}{p_j^3} + O(N^{-3}) .$$

Thus

(14) $$E\hat{H} = H - \frac{s-1}{2N} + \frac{1}{12N^2} (1 - \sum_{j=1}^{s} \frac{1}{p_j}) + O(N^{-3}) .$$

Next ve evaluate the mean squared error of $\hat{H}$, that is, $E\{(\hat{H} - H)^2\}$.

From (6), we have

$$(15)\qquad (\hat{H} - H)^2 = \sum_{j=1}^{s} \sum_{k=1}^{s} (\hat{p}_j - p_j)(\hat{p}_k - p_k) \log p_j \log p_k$$

$$-2 \sum_{j=1}^{s} (\hat{p}_j - p_j) \log p_j \sum_{m=2}^{r} \frac{(-1)^{m-1}}{m(m-1)} \sum_{k=1}^{s} \frac{(\hat{p}_k - p_k)^m}{p_k^{m-1}}$$

$$+ \sum_{m=2}^{r} \sum_{\ell=2}^{r} \frac{(-1)^{m+\ell-2}}{m(m-1)\ell(\ell-1)} \sum_{j=1}^{s} \sum_{k=1}^{s} \frac{(\hat{p}_j - p_j)^m (\hat{p}_k - p_k)^\ell}{p_j^{m-1} \, p_\ell^{m-1}}$$

$$-2 R_{r+1} \sum_{j=1}^{s} (\hat{p}_j - p_j) \log p_j + 2 R_{r+1} \sum_{m=2}^{r} \frac{(-1)^{m-1}}{m(m-1)} \sum_{j=1}^{s} \frac{(\hat{p}_j - p_j)^m}{p_j^{m-1}}$$

$$+ R_{r+1}^2 \; .$$

We compute the expected value of (15), employing (A.1.1-A.1.13) and (A.1.20), obtaining, for $r = 3$,

$$(16)\qquad E \sum_{j=1}^{s} \sum_{k=1}^{s} (\hat{p}_j - p_j)(\hat{p}_k - p_k) \log p_j \log p_k$$

$$= \sum_{j=1}^{s} \frac{\log^2 p_j \mu_2(j)}{N^2} + \sum_{j,k} \frac{\log p_j \log p_k \, \mu_{11}(j,k)}{N^2} - \sum_{j=1}^{s} \frac{\log^2 p_j \mu_{11}(j,j)}{N^2}$$

$$= \frac{1}{N} \left(\sum_{j=1}^{s} p_j \log^2 p_j - H^2 \right) ,$$

(17) $$E\sum_{m=2}^{3}\sum_{\ell=2}^{3}\frac{(-1)^{m+\ell-2}}{m(m-1)\ell(\ell-1)}\sum_{j=1}^{s}\sum_{k=1}^{s}\frac{(\hat{p}_j-p_j)^m(\hat{p}_k-p_k)^\ell}{p_j^{m-1}\,p_k^{\ell-1}}$$

$$=\frac{1}{4N^4}\left\{\sum_{j,k}\frac{\mu_{22}(j,k)}{p_j\,p_k}+\sum_j\frac{\mu_4(j)}{p_j^2}-\sum_j\frac{\mu_{22}(j,j)}{p_j^2}\right\}+O(N^{-3})$$

$$=\frac{1}{4N^2}\left\{(3-2s+s^2)+3(s-2+\sum p_j^2)-(3\sum p_j^2-2+s)\right\}+O(N^{-3})$$

$$=\frac{1}{4N^2}(s^2-1)+O(N^{-3})\ ,$$

and

(18) $$-2E\sum_{m=2}^{3}\frac{(-1)^{m-1}}{m(m-1)}\sum_{j=1}^{s}\sum_{k=1}^{s}\frac{\log p_j(\hat{p}_j-p_j)(\hat{p}_k-p_k)^m}{p_k^{m-1}}$$

$$=\frac{1}{N^3}\left\{\sum_{j,k}\frac{\log p_j\mu_{21}(k,j)}{p_k}+\sum_j\frac{\log p_j\mu_3(j)}{p_j}-\sum_j\frac{\log p_j\mu_{21}(j,j)}{p_j}\right\}$$

$$-\frac{1}{3N^4}\left\{\sum_{j,k}\frac{\log p_j\mu_{31}(k,j)}{p_k^2}+\sum_j\frac{\log p_j\mu_4(j)}{p_j^2}-\sum_j\frac{\log p_j\mu_{31}(j,j)}{p_j^2}\right\}$$

$$=\frac{1}{N^2}(\sum_j\log p_j+sH)-\frac{1}{N^2}(sH+\sum_j\log p_j)+O(N^{-3})$$

$$=O(N^{-3})\ .$$

We now consider the three terms in (15) which contain R_4 as a factor. To consider the first of these terms, we write

(19) $$R_4 \sum_{j=1}^{s} (\hat{p}_j - p_j) \log p_j = \sum_{j=1}^{s} (\hat{p}_j - p_j) \log p_j \left(\sum_{k=1}^{s} \frac{(-1)^3}{12} \frac{(\hat{p}_k - p_k)^4}{p_k^3} + \sum_{k=1}^{s} \frac{(-1)^4}{20} \frac{(\hat{p}_k - p_k)^5}{p_k^4} + R_6 \right) .$$

The expected value can easily be estimat^d using (A. 1. 5), (A. 1. 6), (11), (A. 1. 7), (A. 1. 2), the Cauchy-Schwarz inequality, (A. 1. 16) and (A. 1. 2C). We obtain

(20) $$E R_4 \sum_{j=1}^{s} (\hat{p}_j - p_j) \log p_j = O(N^{-3}) .$$

The extensive computation indicated in (19) appeared to be essential, since a direct application of the Cauchy-Schwarz inequality yields an estimate of $O(N^{-5/2})$.

Similarly, from (11), (A. i. 20), and (A. 1. 16), it follows readily that

(21) $$R_4 \sum_{m=2}^{3} \frac{(-1)^{m-1}}{m(m-1)} \sum_{j=1}^{s} \frac{(\hat{p}_j - p_j)^m}{p_j^{m-1}} + R_4^2 = O(N^{-3}) .$$

Combining (16)-(21), we obtain

(22) $$E(\hat{H} - H)^2 = \frac{1}{N} \left(\sum_{j=1}^{s} p_j \log^2 p_j - H^2 \right) + \frac{1}{4N^2} (s^2 - 1) + O(N^{-3}) .$$

From (14) and (22), we obtain

(23) $$\sigma_{\hat{H}}^2 = E(\hat{H} - H)^2 - (E\hat{H} - H)^2 = \frac{1}{N} \left(\sum p_j \log^2 p_j - H^2 \right) + \frac{s-1}{2N^2} + O(N^{-3}) .$$

The preceding discussion enables us to observe a variety of shortcomings, when one employs $\hat{H}$ as an estimator in the more general situation described in section one.

First, from (14), we see that the bias of $\hat{H}$ depends on s, the number of classes. If s is known, the bias can be largely removed by replacing $\hat{H}$ by $\hat{H} + \frac{s-1}{2N}$; however, we have assumed that s is unknown. Secondly, it should be noted that the bias increases with s . Thus if we permit s to grow, or if s is unknown, the bias may be large. In particular, we are interested in the case where s may be of the same magnitude as N . In this case, we would have to regard $s = s(N)$ and $p_i = p_i(N)$. However, from (22) or (14), it is apparent that $\hat{H} - H$ will not generally tent to zero in probability. Intuitively, it too much of the total probability is concentrated on cells that are too small, then H will not be a satisfactory estimator.

In the examination of the properties of $\hat{H}$, we found it desirable to extend Bas̆arin's computations to terms of $O(N^{-2})$. This is desirable whenever $p_i = 1/s$, $i = 1, 2, \ldots, s$. In that case,

$$G(p_1, p_2, \ldots, p_3) = \sum_{j=1}^{s} p_j \log^2 p_i - H^2$$

$$= s \frac{1}{s} \log^2 s = \log^2 s = 0$$

and a useful asymptotic estimate of σ_H^2 or $E(\hat{H} - H)^2$ is not obtained.

In summary, if s is known, or known to be bounded (independent of N) or if the total probability of "small classes" is known to be small, then $\hat{H}$ will have satisfactory properties. In Appendix 2, the maximum of $G(p_1, p_2, \ldots, p_s)$ is obtained. This can be utilized in determining the sample size necessary to obtain a specified mean squared error when s is known and $\hat{H}$ is used as the estimator of H.

3. Quadrature methods of estimating H. Let

$$(24) \qquad R(p_1, p_2, \ldots) = \sum_{j=1}^{\infty} Np_j e^{-Np_j} .$$

We define the distribution function

$$(25) \qquad F(x) = \sum_{Np_j \le x} Np_j e^{-Np_j} / R(p_1, p_2, \ldots) .$$

Then, it follows that

$$(26) \qquad \frac{R(p_1, p_2, \ldots)}{N} \int_0^N e^x \log(\frac{N}{x})\, dF(x) = \frac{1}{N} \sum_{j=1}^{\infty} e^{Np_j} \log(\frac{N}{Np_j})\, Np_j e^{-Np_j}$$

$$= -\sum_{j=1}^{\infty} p_j \log p_j = H .$$

Thus, it is clear that if we knew $p_1, p_2, \ldots,$ we would therefore know $F(x)$ and consequently know H. The procedure is to use the data to obtain an estimate of $F(x)$ and thus to obtain an estimate of H, which we denote by $\check{H}$.

Specifically, we propose to write (26) in the form

$$H = \int_0^N g(x)\, dF(x) \quad , \tag{27}$$

and to estimate H by

$$H = \sum_{i=1}^{d} g(x_i) w_i \, , \tag{28}$$

the points x_i and the weights w_i are to be determined from the data. We now proceed to the construction of quadrature formulas of the form of (28).

Let n_r be the number of cells occuring r times in the sample. Trivially, we have

$$\sum_{j=1}^{N} r\, n_r = N \quad . \tag{29}$$

From Appendix 3, we have that

$$E n_r \sim \sum_{j=1}^{\infty} \frac{(Np_j)^r}{r!} e^{-Np_j} , \qquad r = 1, 2, \ldots, k \quad , \tag{30}$$

where k does not depend on N. The reader should refer to the appendix for details concerning the sense in which the symbol "$\sim$" is used here. The moments of $F(x)$, denoted by μ_r are given by

$$(31) \qquad \int_0^N x^r \, dF(x) = \sum_j (Np_j)^{r+1} e^{-Np_j} / R(p_1, p_2, \dots)$$

$$\sim (r+1)!\, E(n_{r+1})/E(n_1) \ .$$

The observed values of n_r may be regarded as estimates of $E(n_r)$ whenever $n_1 \neq 0$. In this case, we can regard

$$(32) \qquad m_r = (r+1)!\, n_{r+1}/n_1, \qquad r = 1, 2, \dots, k$$

as estimates of the first k moments.

We proceed as follows. If $n_1 = 0$, estimate H by (4). If $n_1 \neq 0$, select k and determine $m_1, m_2, \dots, m_k$. Using these as estimates of the moments, we seek to determine a distribution function whose first k moments are $m_1, m_2, \dots, m_k$. Unfortunately, it may happen that the "sample moments" $m_1, m_2, \dots, m_k$ are inconsistent. That is, since these are estimates of the moments of (25) and subject to sampling fluctuations, it is possible that there is no distribution function on $[0, N]$ with $m_1, m_2, \dots, m_k$ as its first k moments. Consequently, we compare $m_1, m_2, \dots m_\ell$, $\ell \leq k$ with the consistency conditions, which may be found in B. Harris [4]; the simplest of these conditions is $m_2 \geq m_1^2$. If m_ℓ, $1 < \ell \leq k$, is the last moment estimate which satisfies these conditions, we employ $m_1, m_2, \dots, m_\ell$ in determining $\hat{F}_\ell(x)$, the estimator of $F(x)$ used in determining $\breve{H}$.

From (31) and from Appendix 3, we can easily see that it is the "small probabilities" that contribute to $E n_r$, $r = 1, 2, \dots, k$ and thus an estimator of F(x) constructed in this manner will use mainly the information contained in the "small p_j's". For the "large p_j's", the estimation of p_j by $\hat{p}_j$ is satisfactory. To estimate the part of the data that should be assigned to "large p_j's", the following procedure is followed. Once $\hat{F}_\ell(x)$ is determined, we compute

$$(33) \qquad \mu_r(\hat{\Gamma}_\ell) = \int_0^N x^r \, d\hat{F}_\ell(x), \qquad r = \ell+1, \ \ell+2, \ \ldots$$

from which, we obtain, using (31),

$$(34) \qquad \hat{n}_{r+1} = \mu_r(\hat{F}_\ell)\, n_1/(r+1)!, \qquad r = \ell+1, \ \ell+2, \ \ldots .$$

From these estimates, we define

$$(35) \qquad w_{r+1} = \begin{cases} n_{r+1} - \hat{n}_{r+1} & \text{if} \quad n_{r+1} - \hat{n}_{r+1} > 0 , \\ 0 & \text{otherwise.} \end{cases}$$

w_{r+1} provides an estimate of the contribution to the occupancy numbers accounted for by the "large cells", that is, not included in $\hat{F}_\ell(x)$. A further modification is necessitated in the case of Gauss quadrature formula, which will be discussed subsequently.

Thus combining the heuristic arguments given above, we obtain

$$(36) \qquad \check{H}_\ell = \frac{n_1}{N} \int_0^N e^x \log\left(\frac{N}{x}\right) d\hat{F}_\ell(x) - \sum_{k \geq \ell} \frac{w_{k+1}}{N} \log\left(\frac{w_{k+1}}{N}\right) ,$$

which is easily seen to have the form (28).

To amplify and illustrate the above principles, we proceed by using the Gaussian quadrature formulas, which are the simplest to employ.

Then we have for $\hat{F}_{\ell}(x)$, $\ell = 1, 2, 3$, the following:

$$\hat{F}_1(x) = \begin{cases} 0 & x < m_1 \ , \\ 1 & m_1 \leq x \ ; \end{cases} \tag{37}$$

$$\hat{F}_2(x) = \begin{cases} 0 & x < \dfrac{Nm_1 - m_2}{N - m_1} \ , \\ \dfrac{(N-m_1)^2}{(N-m_1)^2 + (m_2 - m_1^2)} & \dfrac{Nm_1 - m_2}{N - m_1} \leq x < N \ , \\ 1 & x \geq N \ ; \end{cases} \tag{38}$$

$$\hat{F}_3(x) = \begin{cases} 0 & x < \dfrac{m_1 z + m_2 - m_1^2}{z} \\ \dfrac{z^2}{z^2 + m_2 - m_1^2} & \dfrac{m_1 z + m_2 - m_1^2}{z} \leq x < m_1 - z \\ 1 & m_1 - z \leq x \ , \end{cases} \tag{39}$$

where

$$z = \frac{-M_3 - \sqrt{M_3^2 + 4(m_2 - m_1^2)^6}}{2(m_2 - m_1^2)} \tag{40}$$

and

$$M_3 = m_3 - 3m_1(m_2 - m_1^2) - m_1^3 \ . \tag{41}$$

The Gauss quadrature formulas listed here have the attribute that for ℓ an even integer, positive probability is placed at N . Thus as $N \to \infty$, this provides an asymptotic lower bound for H (see E. B. Cobb and B. Harris [2]). Simultaneously, the use of (34) provides overestimates for $\hat{n}_{r+1}$. In odd values of ℓ, the use of $\hat{F}_\ell$ minimizes the higher moments, suggesting that this will account for the information contained in the "small p_j's" in a reasonable way. Accordingly, in the examples that follow, we have used the minimum values of the moments in (34), feeling that this will be appropriate. Thus (34) and $\hat{F}_\ell(x)$ for odd value of ℓ are to be regarded as providing the estimates we seek. We report the results for even values of ℓ as well in the numerical examples that follow for purposes of comparison. The apparent negative bias is to be noted in each example.

We now turn to some numerical examples to clarify the preceding discussion and to provide numerical comparisons for purposes of justifying the proposed technique and the heuristic arguments which suggest it.

4. Numerical examples. The examples which follow are intended to provide comparisons between $\hat{H}$ and $\check{H}_\ell$. We present these in substantial detail with extensive discussion so that the ideas and computational procedures are clear. Some are artificial in the sense that expected values are employed instead of "random data". This has the following purpose - if the techniques described here perform poorly when the data is "perfect", then it should do even worse when random fluctuations are imposed.

Example 1. $p_j = \frac{1}{4}$, $j = 1, 2, 3, 4$, $z_1 = 30$, $z_2 = 27$, $z_3 = 21$, $z_1 = 22$, $N = 100$, $H = \log 4 = 1.38629$, $\hat{H} = 1.37556$.

From (14), we have that $E\hat{H} \sim 1.37117$, and from (25), $\sigma^2_{\hat{H}} = .00015$ and $E(\hat{H} - H)^2 = .000375$. Note that if s is assumed known, we can improve $\hat{H}$ by correcting for the bias, obtaining $\hat{H} + \frac{s-1}{2N} = 1.39056$.

Example 2. $p_j = 10^{-3}$, $j = 1, 2, \ldots, 10^3$, $N = 100$. In such a population, $\hat{H}$ should not perform too well, since the cell probabilities are all very small compared to N. Here $H = 6.90776$. Thus type of population is very favorable to the quadrature method, since $F(x)$ is a degenerate distribution with probability one at $Np_j = .1$ and is therefore completely determined by μ_1 (that is, $\mu_2 = \mu_1^2$, $\mu_3 = \mu_1^3, \ldots$). The data is $n_1 = 85$, $n_2 = 6$, $n_3 = 1$. Thus, $m_1 = .14118$, $m_2 = .07059$. Further note that $\hat{H} = 4.48903$, also we always have $\hat{H} \leq \log 100 = 4.60517$.

For $k = 1$, we have $w_3 = .71765$. Thus $\breve{H}_1 = 6.49982$. For $k = 2$, we have $\breve{H}_2 = 6.42456$. We are not able to proceed to $k = 3$, since $n_4 = 0$ insures that the consistency conditions for m_1, m_2, m_3 to be a valid moment sequence on $[0, N]$ are not satisfied.

The estimates $\breve{H}_1$ and $\breve{H}_2$ are lower than H. However, this is precisely as it should be, since $En_1 \sim 90$, $En_2 \sim 4.5$, $En_3 \sim .15$ and thus, as a consequence of sampling fluctuations, the data looks as if it came from a distribution which does not have equal probabilities for all cells.

Example 3. $p_i = 10^{-3}$, $i = 1, 2, \ldots, 10^3$, $N = 1000$, $n_1 = 373$, $n_2 = 199$, $n_3 = 62$, $n_4 = 8$, $n_5 = 1$, $n_6 = 1$, $H = 6.90776$.

For this data $\hat{H} = 6.36438$, $m_1 = 1.06702$, $m_2 = .99732$, $m_3 = .51475$, $m_4 = .32172$, $m_5 = 1.93029$.

To compute $\breve{H}_\ell$, we first set $k = 1$, obtaining $w_3 = 0$, $w_4 = 0$, $w_5 = 0$, $w_6 = .28345$. Thus, we get

$$\breve{H}_1 = 7.42779 \; .$$

Since $m_2 < m_1^2$, the process terminates here. Here, the overestimate is precisely what one would expect from the data, since $E n_1 \sim 368$. The observed value of n_1 suggests a larger number of cells than are actually at hand.

Example 4. This example is identical with Example 3 except that $n_1 = 341$, $n_2 = 179$, $n_3 = 70$, $n_4 = 17$, $n_5 = 2$, $n_6 = 1$, $n_7 = 1$. Then $\hat{H} = 6.29417$, $m_1 = 1.04985$, $m_2 = 1.23167$, $m_3 = 1.19648$, $m_4 = .70381$, $m_5 = 2.11144$, $m_6 = 14.78006$.

For $k = 1$, we have $w_3 = 7.35875$, $w_4 = .55897$, $w_5 = 0$, $w_6 = .39596$, $w_7 = .90941$ and hence $\breve{H}_1 = 6.86725$.

For $k = 2$, $w_i = 0$, $i = 4, 5$, $w_6 = .05808$, $w_7 = .84214$, $\breve{H}_2 = 6.71320$.

We are unable to proceed to H_3, since the sequence m_1, m_2, m_3 is not a realizable moment sequence.

We now choose an example for which $F(x)$ is again a one-point distribution, but since $N p_j = 2$, the n_j's will be non-zero for larger value of j.

Example 5. $p_i = 2/1000$, $i = 1, 2, \ldots, 500$, $N = 1000$, $n_1 = 139$, $n_2 = 146$, $n_3 = 78$, $n_4 = 42$, $n_5 = 21$, $n_6 = 5$, $n_7 = 2$, $n_8 = 1$, $n_{10} = 1$. Then $m_1 = 2.10072$,

$m_2 = 3.36691$, $m_3 = 7.25180$, $m_4 = 18.12950$. Here $m_2 < m_1^2$, so that we compute $\breve{H}_1$. $H = 6.21461$ and $\hat{H} = 5.9257$. We obtain $\breve{H}_1 = 7.06270$.

Example 6. Let $p_1 = p_2 = p_3 = p_4 = 1/8$, $p_i = 1/2\ 10^{-3}$, $i = 5, 6, \ldots, 1004$, $N = 200$. The data obtained is $n_1 = 86$, $n_2 = 4$, $n_5 = 1$, $n_{23} = 1$, $n_{24} = 2$, $n_{30} = 1$. For this population $H = 4.84017$. From the data, we have $\hat{H} = 3.59686$ and $\breve{H}_1 = 4.75552$.

The following examples are artificial in the sense that instead of random data, the expected values of the n_r are employed for "small" values of r.

Example 7. We are given 2000 cells, 1000 of which have $p = 1/4000$ and the balance of which have $p_i = 3/4000$. Two thousand observations are taken. We will examine the behavior of $\hat{H}$ and $\breve{H}_k$ as if the n_i's were exactly equal to En_i. Such examples serve to illustrate the motivation for the quadrature method. In this example $En_1 = 548.9751$, $En_2 = 157.1906$, $En_3 = 35.2414$, $En_4 = 6.3543$, $En_5 = .9404$, $En_6 = .1171$, $En_7 = .0125$, $En_8 = .0011$, $H = 7.47009$, and $\hat{H} = 6.52939$. Thus, even with the use En_i in r_i, $\hat{H}$ has a sizeable negative bias. On the other hand, $\breve{H}_i = 7.41776$, $\breve{H}_2 = 7.28016$, and $\breve{H}_3 = H$. This last occurs since

$$F(x) = \begin{cases} 0 & x < .25 \\ .35466 & .25 \leq x < .75 \\ 1 & x \geq .75 \end{cases}$$

That is, $F(x)$ is a two-point distribution and it can easily be seen that every two-point distribution is uniquely characterized by three moments. Thus, if $n_i = En_i$, it follows that $\hat{F}_3(x) = F(x)$ and $\check{H}_3 = H$.

Example 8. This example is extremely artifical, but serves nevertheless to illustrate one of the possible boundary situations which clarify the differences between $\hat{H}$ and $\check{H}$. Assume that we are sampling from a probability distribution that is absolutely continuous with respect to Lebesgue measure on the real line. Every real number is considered to be a separate class. Then $n_1 = N$ with probability one. Here, one should define $H = \infty$, $\hat{H} = \log N$ and $\check{H} = \infty$.

Example 9. The Zipf Distribution. A common mathematical model for describing linguistic as well as other data is the Zipf distribution given by

$$(42) \qquad p_j = (\zeta(s))^{-1} j^{-s} \qquad s > 1, \quad j = 1, 2, \ldots ,$$

where $\zeta(s)$ denotes the Riemann zeta function. This distribution is suited for a test of the quadrature estimates proposed in this paper, since for "small values" of ζ, there are both classes with large probabilities and a substantial concentration of the total probability in small cells. For a specific numerical illustration, we will take $s = 3/2$.

We evaluate H and $E(n_j)$ for the Zipf distributions by means of the Euler-MacLaurin formula, which is particularly suited for this case.

First we have

$$H = -\sum_{j=1}^{\infty} p_j \log p_j$$

$$= \sum (j^s \zeta(s))^{-1} (s \log j + \log \zeta(s))$$

(43) $$= \log \zeta(s) + s(\zeta(s))^{-1} \sum_{j=1}^{\infty} \log j/j^s \ .$$

We employ the Euler-Maclaurin formula to evaluate $\sum_{j=1}^{\infty} \log j/j^s$. To accomplish this we write

$$\sum_{j=1}^{\infty} \log j/j^s = \sum_{j=1}^{M-1} \log j/j^s + \sum_{j=M}^{\infty} \log j/j^s \ .$$

Then

(44) $$\sum_{j=1}^{\infty} \log j/j^s = \sum_{j=1}^{M-1} \log j/j^s + M^{-s+1} \left(\frac{\log M}{s-1} + \frac{1}{(s-1)^2} \right) + \frac{\log M}{2M^s} - \sum_{\nu=1}^{m-1} \frac{B_{2\nu}}{(2\nu)!} \frac{d^{2\nu-1}}{dM^{2\nu-1}} \left(\frac{\log M}{M^s} \right) + R_m(M) \ ,$$

where B_j are the Bernoulli numbers.

We can similarly estimate $E(n_r)$, $r = 1, 2, \ldots$. That is,

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

$$= \frac{1}{r!} \sum_{j=1}^{\infty} \left(N/\zeta(s) j^s \right)^r e^{-N/\zeta(s) j^s} \ .$$

Here the Euler-Maclaurin formula is also applicable and we obtain, using only the initial term of the expansion

$$(45) \qquad E(n_r) \sim \frac{N^{1/s}\,\Gamma(r-s^{-1})}{(\zeta(s))^{1/s}\, r!\, s} \; .$$

We now apply this to a specific numerical illustration setting $s = 3/2$, $N = 1000$.

Thus, we have, for $i = 1, 2, \ldots, 10$

$E(n_1) \sim 94.15584$

$E(n_2) \sim 15.69264$

$E(n_3) \sim 6.97451$

$E(n_4) \sim 4.06846$

$E(n_5) \sim 2.71231$

$E(n_6) \sim 1.95889$

$E(n_7) \sim 1.49249$

$E(n_8) \sim 1.18155$

$E(n_9) \sim .96275$

$E(n_{10}) \sim .80229$

To determine H, we note that using (44) with $M = 4$ and $m = 3$ we get

$$\sum_{j=1}^{\infty} \log j/j^{3/2} = \sum_{j=1}^{3} \log j/j^{3/2} + 4^{-1/2}\left(\frac{\log 4}{1/2} + \frac{1}{(1/4)}\right)$$

$$+ \frac{\log 4}{2\cdot 4^{3/2}} - \sum_{\nu=1}^{1} \frac{B_{2\nu}}{(2\nu)!}\,\frac{d^{2\nu-1}}{dM^{2\nu-1}}\left(\frac{\log M}{M^{3/2}}\right)\Bigg|_{M=4} + R_2(M) \; .$$

$$= .45649 \div 3.38629 + .08664 + .00281 + R_2(M) \ ,$$

where $|R_2(M)| \leq 2.5 \times 10^{-6}$. Thus it follows that we can write

$$H = \log \zeta(3/2) + (3/2\,\zeta(3/2))\Big(\sum_{j=1}^{\infty} \log j/j^{3/2}\Big)$$

$$\sim 3.21811 \ .$$

We make the assumption that for each j, $j = 1, 2, \ldots, n_j = E(n_j)$. This enables us to compare $\hat{H}$ and $\check{H}_k$ when the data are perfect, that is, the data is completely devoid of sampling errors. Given this artificial assumption, we have $\hat{H} = 2.82871$, $\check{H}_1 = 3.00146$, $\check{H}_2 = 2.84048$, $\check{H}_3 = 3.03918$.

Specifically H, $\check{H}_1$, $\check{H}_2$, and $\check{H}_3$ were computed here as follows in order to obtain a reasonable comparison of their behavior. For $p_1, p_2, \ldots, p_5$ the assumption that $\hat{p}_j = p_j$ was made. This was employed, since when p is large, both techniques gives virtually the same result. The remaining p's were distributed according to their contributions to $E(n_r)$, as in the preceding examples. For detailed information about the Zipf distribution and extensive references to articles about the distribution and its applications, see N. L. Johnson and S. Kotz [5, pp. 240-247].

5. <u>Concluding Remarks</u>. The estimator $\check{H}$ described in the preceding sections is to be regarded as a first attempt to produce an estimator which can circumvent the deficiencies of the natural estimator $\hat{H}$. The procedure is by no means completely analyzed and it is hoped that this work will stimulate further investigations into its behavior.

The following remarks are therefore relevant. We have chosen the Gauss quadrature formulas, because they are among the simplest. If we fail to analyze completely the problem for Gauss quadracure formulas, then we are unlikely to be successful in more complicated situations. The selection of k produces problems, since for greater values of k, we utilize more information from the sample; however, the higher moments are less reliable statistically. Thus, a way of balancing these two properties is needed. Second, if $m_2 < m_1^2$, we have set $m_2 = m_1^2$. However, we could also have increased m_1 to $\sqrt{m_2}$, or chosen any alternative in between. Here again, further investigation is needed. The same remarks apply to the determination of w_k (35). The procedure that we have used provides a sequence of quadrature formulas which give better estimates as we increase k, when the data are perfect, that is, $En_i = n_i$, $i = 1, 2, \ldots$. This is an ad hoc procedure and has not taken adequate account of sampling fluctuations.

There are two sources of errors in the quadrature methods. Quadrature formulas of the Gauss type integrate polynomials exactly, but $e^x \log \frac{N}{x}$ is not a polynomial. Secondly, we are aggregating the "small p_j's" and treating them as if they possessed relatively few values, whereas they are in general distributed over a region. This is a form of smoothing, whose properties are not completely understood at this time.

Further work in this direction is being continued by myself and my students and we hope to be able to report further results in this direction in the near future.

Appendix 1

Some formulas and relationships for multinomial distributions

In the evaluation of (10) and (22), the central moments of the multinomial distribution have been used. As a convenience to the reader, they have been tabulated here, along with some identities and inequalities which have been used to obtain order estimates.

We denote $E\{(Z_1 - Np_1)^{j_1}(Z_2 - Np_2)^{j_2} \ldots (Z_s - Np_s)^{j_s}\}$ by $\mu_{j_1 j_2 \ldots j_s}$ for every $1 \leq j_i < \infty$, $1 \leq s < \infty$

(A.1.1) $$\mu_1 = 0$$

(A.1.2) $$\mu_2 = N(p_1 - p_1^2)$$

(A.1.3) $$\mu_3 = N(p_1 - 3p_1^2 + 2p_1^3)$$

(A.1.4) $$\mu_4 = 3N^2(p_1^2 - 2p_i^3 + p_1^4) + N(p_1 - 7p_1^2 + 12p^3 - 6p^4)$$

(A.1.5) $$\mu_5 = 10N^2(p_1^2 - 4p_1^3 + 5p_1^4 - 2p_1^5) + N(p_1 - 15p_1^2 + 50p_1^3 - 60p_1^4 + 24p_1^5)$$

(A.1.6)
$$\begin{aligned}\mu_6 = {} & N^3(15p_1^3 - 45p_1^4 + 45p_1^5 - 15p_1^6) + \\ & + N^2(25p_1^2 - 180p_1^3 + 415p_1^4 - 390p_1^5 + 130p_1^6) \\ & + N(p_1 - 31p_1^2 + 180p_1^3 - 390p_1^4 + 360p_1^5 - 120p_1^6)\end{aligned}$$

(A.1.7) $$\mu_{11} = -Np_1p_2$$

(A.1.8) $$\mu_{21} = N(2p_1^2p_2 - p_1p_2)$$

(A.1.9) $$\mu_{111} = 2Np_1p_2p_3$$

(A.1.10) $$\mu_{31} = 3N^2(p_1^3p_2 - p_1^2p_2) + N(-6p_1^3p_2 - p_1p_2 + 6p_1^2p_2)$$

(A.1.11) $$\begin{aligned}\mu_{22} &= N^2(3p_1^2p_2^2 - p_1^2p_2 - p_2^2p_1 + p_1p_2)\\ &+ N(-6p_1^2p_2^2 + 2p_1^2p_2 + 2p_1p_2^2 - p_1p_2)\end{aligned}$$

(A.1.12) $$\mu_{211} = N^2(3p_1^2p_2p_3 - p_1p_2p_3) + N(-6p_1^2p_2p_3 + 2p_1p_2p_3)$$

(A.1.13) $$\mu_{1111} = 3N^2p_1p_2p_3p_4 - 6Np_1p_2p_3p_4 \,.$$

These formulas may be obtained by completely elementary methods. Further, a number of these are given in G. P. Bašarin [1], although μ_3 is incorrectly stated there. Similarly, all of the above with $\Sigma j_i \leq 4$ may be found in F. N. David and D. E. Barton [3, p. 146], although μ_{22} is incorrectly given there.

From (A.1.1)-(A.1.6), we note that

(A.1.14) $$\mu_{22} = O(N^r), \;\; \mu_{2r-1} = O(N^{r-1}), \;\; r = 1, 2, 3.$$

From the well-known recursion $\mu_0 = 1$, (A.1.15) $$\mu_{r+1} = (p_1 - p_1^2)(Nr\mu_{r-1} + \frac{d}{dp_1}\mu_r), \quad r = 1, 2, \ldots$$

it follows that

(A.1.16) $$\mu_{2r} = O(N^r), \quad \mu_{2r-1} = O(N^{r-1}), \quad r = 1, 2, \ldots .$$

We get order estimates for the product moments, that is, those indexed by more than one subscript by use of repeated applications of the Cauchy-Schwarz inequality. Specifically, let $q \geq 1$ be an integer. Then for arbitrary random variables W_1, W_2, ..., W_{2^q} such that the moments given below all exist, we obtain

(A.1.17) $$(E(W_1 W_2 \ldots W_{2^q}))^{2^q} \leq \prod_{i=1}^{2^q} EW_i^{2^q}$$

When $q = 1$, this is the customary form of the Cauchy-Schwarz inequality. To apply (A.1.17) to the situation at hand, we define $W_i = (Z_i - Np_i)^{j_i}$ and if in $\mu_{j_1 j_2 \ldots j_s}$, $2^{q-1} < s < 2^q$, $q > 1$, then then we define $W_i = 1$, $i = s+1, s+2, \ldots, 2^q$. We write (A.1.17) in the form

(A.1.18) $$|E(W_1 W_2 \ldots W_{2^q})| \leq \left\{\prod_{i=1}^{2^q} EW_i^{2^q}\right\}^{\frac{1}{2^q}} .$$

Thus, for $2^{q-1} < s \leq 2^q$, $q \geq 1$

(A.1.19) $$|E\{(Z_1-Np_1)^{j_1}(Z_2-Np_2)^{j_2}\ldots(Z_s-Np_s)^{j_s}\}| \leq \left(\prod_{i=1}^{s} E\left(Z_i-Np_i\right)^{2^q j_i}\right)^{\frac{1}{2^q}} .$$

From (A.1.14), $E(Z_i-Np_i)^{2^q j_i} = O(N^{2^{q-1} j_i})$, $i = 1, 2, \ldots, s$ and hence

$$\prod_{i=1}^{s} E(Z_i-Np_i)^{2^q j_i} = O(N^{2^{q-1}\sum_{i=1}^{s} j_i}) .$$

Thus

(A.1.20) $$\mu_{j_1 j_2 \ldots j_s} = O(N^{[\sum_{i=1}^{s} j_i/2]}) ;$$

the integer part is a consequence of the fact that N can only appear in integer powers.

Appendix 2

The Behavior of the Asymptotic Variance and Mean Squared Error of $\hat{H}$. In this appendix we show that

$$(\text{A}2.1)\quad \max_{p_1, p_2, \ldots, p_s} G(p_1, p_2, \ldots, p_s) = \max_{p_1, p_2, \ldots, p_s} \left(\sum_{j=1}^{s} p_j \log^2 p_j - H^2\right)$$

$$= 4p^*(1-p^*)/(1-2p^*)^2 ,$$

where p^* is the largest solution in p of

$$e^2 = \left(\frac{1-p}{(s-1)p}\right)^{1-2p} .$$

This can be used to specify the sample size N^* necessary to obtain estimates of H of a given precision when using $\hat{H}$ and therefore is of importance when s is bounded; or more precisely, when s/N^* is sufficiently small. The minimum of $G(p_1, p_2, \ldots) = 0$ and is trivially attained when

$$\begin{cases} p_j = 0 & i \in I , \quad I \subset \{1, 2, \ldots, s\} , \\ p_j = |I^c|^{-1}, & i \in I^c, \quad I^c \neq \emptyset . \end{cases}$$

This is easily verified as follows, since then

$$\sum_{j=1}^{s} p_j \log^2 p_j = \sum_{j \in I^c} p_j \log^2 p_j$$

$$= |I^c| \, |I^c|^{-1} \log^2 |I^c| .$$

Further, in this case

$$H = -\sum_{j=1}^{s} p_j \log p_j = |I^c| \, |I^c|^{-1} \log |I^c| \quad ,$$

thus verifying the assertion.

To determine the maximum, we first note that

$$\text{(A.2.2)} \quad G^*(s) = \max_{p_1, \ldots, p_s} G(p_1, \ldots, p_s) \geq \max_{p_1, \ldots, p_{s-1}} G(p_1, \ldots, p_{s-1}, 0)$$

$$= \max_{p_1, \ldots, p_{s-1}} G(p_1, \ldots, p_{s-1}) = G^*(s-1) \quad .$$

Now let $p_s = 1 - \sum_{j=1}^{s-1} p_j$ and note that for $j = 1, 2, \ldots, s-1$

$$\text{(A.2.3)} \quad \frac{\partial G(p_1, \ldots, p_s)}{\partial p_j} = (\log p_j - \log p_s)(\log p_j + \log p_s + 2 + 2H) \quad .$$

Setting

$$\text{(A.2.4)} \quad \frac{\partial G(p_1, \ldots, p_s)}{\partial p_j} = 0, \qquad j = 1, 2, \ldots, s-1 \quad ,$$

we note that in each equation we must have either $\log p_j - \log p_s = 0$ or $\log p_j + \log p_s + 2 + 2H = 0$. Clearly if $\log p_j = \log p_s$ for $j = 1, 2, \ldots, s-1$, we have $p_j = 1/s$, $j = 1, 2, \ldots, s$ and $G(p_1, p_2, \ldots, p_s) = 0$, a minimum. Hence there must be at least one j with $\log p_j + \log p_s + 2 + 2H = 0$. Since for any solution of (A.2.4), any permutation of the indices $1, 2, \ldots, s$ is also a solution, with no loss of generality, we can set

$$\log p_j + \log p_s + 2 + 2H = 0, \qquad j = 1, 2, \ldots, t; \; 1 \leq t \leq s-1 \;;$$

then we have

$$\text{(A. 2. 5)} \qquad \begin{cases} p_j \, p_s = e^{-2-2H}, & j = 1, 2, \ldots, t \;, \\ p_j = p_s \,, & j = t+1, \; t+2, \; \ldots, \; s-1 \;, \end{cases}$$

the set of indices j for which $p_j = p_s$ possibly being empty. From (A. 2. 5), we have

$$\text{(A. 2. 6)} \qquad \sum_{j=1}^{t} p_j \, p_s = p_s(1 - (s-t)p_s) = te^{-2-2H} \;.$$

For fixed t, let H^* be any H in the solution set of (A. 2. 6). Then, $p_s = p_s(H^*)$ has at most two solutions, say $p_{s1}(H^*)$ and $p_{s2}(H^*)$. Thus, from (A. 2. 5), we have, for every H^* and every $p_{si}(H^*)$, $i = 1, 2$,

$$\text{(A. 2. 7)} \qquad \begin{cases} p_k(H^*) = p_j(H^*), & 1 \leq j, k \leq t \;, \\ p_k(H^*) \neq p_{si}(H^*), & 1 \leq k \leq t \;. \end{cases}$$

Thus every solution to (A. 2. 5) has the form

$$\text{(A. 2. 8)} \qquad \begin{cases} p_j = \dfrac{1-(s-t)p_s}{t}, & j = 1, 2, \ldots, t \;; \\ p_j = p_s & j = t+1, \; \ldots, \; s-1 \;. \end{cases}$$

This yields

$$H = -(s-t)\, p_s \log p_s - (1 - (s-t)p_s) \log\left(\frac{1-(s-t)p_s}{t}\right) .$$

Substituting this into (A.2.6), we obtain

$$\text{(A.2.9)} \qquad e^2 = \left(\frac{1-(s-t)p_s}{tp_s}\right)^{1-2(s-t)p_s} , \qquad 0 \leq p_s \leq (s-t)^{-1} .$$

Now the logarithm of the right hand side of (A.2.9) is a convex function of p_s which assumes the value $+\infty$ at $p_s = 0$ and $p_s = (s-t)^{-1}$ and the values 0 at $p_s = 1/s$ and $1/2(s-t)$. Thus there are exactly two solutions of (A.2.9), p_{s1} and p_{s2} with $0 < p_{s1} < 1/s \leq \frac{1}{2(s-t)} < p_{s2} < 1/(s-t)$. As a consequence of the preceding discussion, we have that

$$\text{(A.2.10)} \quad G^*(s) = \max_{1 \leq t \leq s-1} \max_{i=1,2} G\left(\frac{1-(s-t)p_{si}}{t}, \ldots, \frac{1-(s-t)p_{si}}{t}, p_{si}, \ldots, p_{si}\right)$$

$$= \max_{1 \leq t \leq s-1} \max_{i=1,2} G_1(t, p_{si}) .$$

Further, note that if (t, p_{si}) is a solution of (A.2.5), then $(s-t, \frac{1-(s-tp_{si}}{t})$ is also a solution and

$$G_1(t, p_{si}) = G_1\left(s-t, \frac{1-(s-t)p_{si}}{t}\right) .$$

Thus, we can reduce (A.2.10) to

$$G^*(s) = \max_{s/2 \leq t \leq s-1} \max_{i=1,2} G_1(t, p_{si}) .$$

Hence one can determine the maximum by evaluating $G_1(t, p_{si})$ for s-1 choices of (t, p_{si}) . However, an exact computation is possible. Hence we proceed further.

Note that by using (A.2.1), we have, for $i = 1,2$;

$$G(t, p_{si}) = t\left(\frac{1-(s-t)p_{si}}{t}\right) \log^2\left(\frac{1-(s-t)p_{si}}{t}\right) + (s-t)p_{si} \log^2 p_{si}$$

$$-\left(t\left(\frac{1-(s-t)p_{si}}{t}\right)\log\left(\frac{1-(s-t)p_{si}}{t}\right) + (s-t)p_{si} \log p_{si}\right)^2$$

$$= (1-(s-t)p_{si})(s-t)\, p_{si}(\log(1-(s-t)p_{si}) - \log(tp_{si}))^2 \quad .$$

Then, since p_{si} is a solution of equation (A.2.9), we have

$$\log(1-(s-t)p_{si}) - \log(tp_{si}) = 2/(1-2(s-t)p_{si}) \quad ,$$

hence

$$\text{(A.2.11)} \qquad G_1(t, p_{si}) = \frac{4(s-t)\, p_{si}(1-(s-t)p_{si})}{(1-2(s-t)p_{si})^2}, \quad i = 1,2 \quad .$$

Consequently, we define

$$\text{(A.2.12)} \qquad G_2(\rho) = \frac{4\rho(1-\rho)}{(1-2\rho)^2}, \qquad 0 < \rho < 1 \quad ,$$

where $\rho_i = (s-t)p_{si}$. Thus $G_2(\rho_i) = G_1(t, p_{si})$, $i = 1,2$. Clearly $G_2(\rho)$ is symmetric about $\rho = 1/2$. Further, $G_1(\rho)$ is increasing for $0 < \rho < 1/2$ and decreasing for $1/2 < \rho < 1$. Consequently

$$\max_{i=1,2} G_1(t, p_{si}) = G_1(t, p_s^*) = G_2(\rho^*) \quad ,$$

where $\rho^* = \rho_2$ if $\rho_2 - 1/2 \leq 1/2 - \rho_1$ and ρ_1 otherwise and $p_s^* = \rho^*/(s-t)$. We now show that $\rho^* = \rho_2$.

We transform (A.2.9) similarly, obtaining

$$\text{(A.2.13)} \qquad [\frac{(s-t)(1-\rho)}{t\rho}]^{1-2\rho} = e^2, \quad 0 < \rho < 1 .$$

If $t = s/2$, then the left hand side of (A.2.13) is symmetric about $\rho = .5$ and consequently $\rho_2 - 1/2 = 1/2 - \rho_1$ in that case. Further, since $t \geq s/2$, for $\rho < 1/2$,

$$\text{(A.2.14)} \qquad (\frac{1-\rho}{\rho})^{1-2\rho} \geq (\frac{(s-t)(1-\rho)}{t\rho})^{1-2\rho}$$

and for $\rho > 1/2$

$$\text{(A.2.15)} \qquad (\frac{1-\rho}{\rho})^{1-2\rho} \leq (\frac{(s-t)(1-\rho)}{t\rho})^{1-2\rho} .$$

Thus in general, $\rho_2 - 1/2 \leq 1/2 - \rho_1$ and $\rho^* = \rho_2$. Hence $p_s^* = p_{s2}$. Consequently, in (A.2.12) and (A.2.13), we can restrict attention to the region $\rho \geq 1/2$. Thus, we have shown that

$$G^*(s) = \max_{s/2 \leq t \leq s-1} G(t, p_{s2}) .$$

Further, note that (A.2.13) depends on t and s only through $(s-t)/t$. Now let s and $\rho \geq 1/2$ be fixed and consider t_1, t_2 with $s/2 \leq t_1 < t_2 < s-1$. Then

$$\frac{s-t_1}{t_1} > \frac{s-t_2}{t_2}$$

and hence

$$\left(\frac{(s-t_1)(1-\rho)}{t_1 \rho}\right)^{1-2\rho} \leq \left(\frac{(s-t_2)(1-\rho)}{t_2 \rho}\right)^{1-2\rho} .$$

Thus the root $p_{s2}(t_2)$ of (A. 2. 9) is smaller than the root $p_{s2}(t_1)$ of (A. 2. 9) and we conclude

(A. 2. 16)
$$G^*(s) = G\left(\frac{1-p^*_{s2}}{s-1}, \ldots, \frac{1-p^*_{s2}}{s-1}, p^*_{s2}\right)$$
$$= \frac{4p^*_{s2}(1-p^*_{s2})}{(1-2\,p^*_{s2})^2} ,$$

where $p^*_{s2} = p_{s2}(s-1)$.

The above argument can also be employed to demonstrate the monotonicity of $G^*(s)$ as a function of s, however this follows immediately from (A. 2. 2).

Appendix 3

In this appendix we justify some of the approximations used in the quadrature method. In the subsequent discussion τ and λ will be given real numbers with $0 < \tau < 1/2 < \lambda < 1$ and $\tau + \lambda < 1$. We now establish the following computational lemmas.

Lemma A.3.1. Let $N \to \infty$. Then for any integer $r \geq 0$, and any $c > 0$, and $r \leq cN^{\tau}$,

$$\binom{N}{r} = \frac{N^r}{r!}\,(1 + O(N^{2\tau - 1})) \,. \tag{A.3.1}$$

The proof of this lemma is trivial and therefore omitted.

Lemma A.3.2. For $r \leq cN^{\tau}$ and $p \leq cN^{-\lambda}$, as $N \to \infty$,

$$(1-p)^{N-r} = e^{-Np}(1 + O(N^{1-2\lambda})) \,. \tag{A.3.2}$$

The proof of this lemma is trivial and therefore omitted.

Lemma A.3.3. For $r \leq cN^{\tau}$ and $p > cN^{-\lambda}$, for every $\epsilon > 0$ there is an N_{ε} sufficiently large so that for $N \geq N_{\varepsilon}$

$$\binom{N}{r} p^r (1-p)^{N-r} \leq e^{-N^{1-\lambda-\epsilon}} \tag{A.3.3}$$

and

$$\frac{(Np)^r}{r!}\, e^{-Np} \leq e^{-N^{1-\lambda-\epsilon}} \,. \tag{A.3.4}$$

The proof of this lemma is trivial and therefore omitted.

Lemma A. 3. 4. If n_r is the number of classes occurring r times then

$$E(n_r) = \sum_{j=1}^{\infty} \binom{N}{r} p_j^r (1-p_j)^{N-r}$$

Proof. Let $Z_j = 1$ if the jth class occurs r times in the N observations and $Z_j = 0$ otherwise. Then $n_r = \sum_{j=1}^{\infty} Z_j$ and $En_r = E \sum_{j=1}^{\infty} Z_j = \sum_{j=1}^{\infty} EZ_j = \sum_{j=1}^{\infty} P\{Z_j = 1\} = \sum_{j=1}^{\infty} \binom{N}{r} p_j^r (1-p_j)^{N-r}$. Combining all of the above, we have the following theorem.

Theorem A. 3. 1. Let τ and λ be given real numbers with $0 < \tau < \frac{1}{2} < \lambda < 1$ and $\tau + \lambda < 1$. Then given a random sample $(X_1, X_2, \ldots, X_N)$ of N observations from the population with $P(X_i \epsilon M_j) = p_j$, $j = 1, 2, \ldots, \infty$, and if n_r is the number of cells such that exactly r X_i's ϵ M_j, then

$$E(n_r) = \sum_{j=1}^{\infty} \binom{N}{r} p_j^r (1-p_j)^{N-r}$$

and for $r \leq N^{\tau}$ and for every $\epsilon > 0$ there is an N_{ε} such that for $N \geq N_{\varepsilon}$

$$E(n_r) = \sum_{p_j \leq N^{-\lambda}} \frac{(Np_j)^r}{r\,!} e^{-Np_j}(1 + O(N^{1-2\lambda})) + O(e^{-N^{1-\lambda-\epsilon}}) \;.$$

Proof. There are at most N^{λ} cells with $p_j > N^{-\lambda}$; hence from Lemma A. 3. 3 and A. 3. 4,

$$\sum_{j:\, p_j > N^{-\lambda}} \binom{N}{r} p_j^r (1-p_j)^{N-r} \leq N^{\lambda} e^{-N^{1-\lambda-\epsilon}} \leq e^{-N^{1-\lambda-\epsilon_1}} \;.$$

The first term is direct from Lemmas A. 3. 1 and A. 3. 2.

We now obtain:

Theorem A. 3. 2. For τ and λ such that $0 < \tau < \frac{1}{2} < \lambda < 1$ and $\tau + \lambda < 1$, we have

$$E(n_r) = (1 + O(N^{\tau - \lambda})) \sum_{p_j \leq N^{-\lambda}} \frac{(Np_j)^r}{r!} e^{-NP_j} + O(e^{-N^{1-\lambda-\epsilon}}) .$$

Proof. We utilize the easily established fact that if a_i, b_i are positive numbers, $i = 1, 2, \ldots$, and $\frac{a_0}{b_0} \geq \frac{a_i}{b_i}$, $i = 1, 2, \ldots$ then

$$\frac{a_0}{b_0} \geq \frac{\Sigma a_i}{\Sigma b_i} .$$

Then for $p_j \leq N^{-\lambda}$, $r \leq N^{\tau}$, we have

$$\frac{\binom{N}{r} p_j^r (1-p_j)^{N-r}}{\frac{(Np_j)^r}{r\,!} e^{-Np_j}} = \frac{N(N-1)\ldots(N-r+1)}{N^r} e^{rp_j} \leq e^{N^{\tau-\lambda}} = 1 + O(N^{\tau-\lambda})$$

Thus

$$\frac{\sum_{p_j \leq N^{-\lambda}} \binom{N}{r} p_j^r (1-p_j)^{N-r}}{\sum_{p_j \leq N^{-\lambda}} \frac{(Np_j)^r}{r\,!} e^{-Np_j}} = 1 + O(N^{\tau - \lambda}) .$$

The conclusion follows from Theorem A. 3. 1.

Appendix 4

In this appendix we provide two examples of populations for which $H = \infty$.

Example 1. Since $\sum_{j=1}^{\infty} \frac{1}{(j+1)\log^2(j+1)} = c < \infty$. Let $p_j = 1/c(j+1)\log^2(j+1)$, $j = 1, 2, \ldots$. Then $\log p_j = -\log c$ $-\log(j+1) - 2\log\log(j+1)$. Then

$$-\sum p_j \log p_j = \sum_j \frac{\log c + \log(j+1) + 2\log\log(j+1)}{c(j+1)\log^2(j+1)} \geq -\sum_j \frac{1}{c(j+1)\log(j+1)} = \infty .$$

Example 2. Let m_k be the smallest non-negative integer such that $m_k \geq \frac{2^k}{k} - k$, $k \geq 1$. Let $M_0 = 0$, $M_k = \sum_{j=1}^{k} 2^{m_j}$. Define $p_i = 2^{-k-m_k}$ for $M_{k-1} < i \leq M_k$, i an integer. Thus $0 < p_i < 1$, $i \geq 1$,

$$\sum_{i=1}^{\infty} p_i = \sum_{k=1}^{\infty} \sum_{i=M_{k-1}+1}^{M_k} \frac{1}{2^{k+m_k}} = \sum_{k=1}^{\infty} \frac{2^{m_k}}{2^{k+m_k}} = \sum_{k=1}^{\infty} \frac{1}{2^k} = 1.$$

Using logarithms base 2, we have

$$\sum_{i=1} -p_i \log_2 p_i = \sum_{k=1}^{\infty} \sum_{i=M_{k-1}+1}^{M_k} -p_i \log p_i = \sum_{k=1} \frac{2^{m_k}(k+m_k)}{2^{k+m_k}} =$$

$$= \sum_{k=1}^{\infty} \frac{k+m_k}{2^k} \geq \sum_{k=1}^{\infty} \frac{k+(\frac{2^k}{k} - k)}{2^k} = \sum_{k=1}^{\infty} \frac{1}{k} = \infty.$$

REFERENCES

[1] G. F. Bašarin, On a statistical estimate for the entropy of a sequence of independent random variables, Teor. Verojatnost. i Primenen., 4 (1959), 361-364.

[2] E. B. Cobb and B. Harris, An asymptotic lower bound for the entropy discrete populations with application to the estimation of entropy for approximately uniform populations. Ann. Inst. Statist. Math., 18 (1966), 289-297.

[3] F. N. David and D. E. Barton, Combinatorial Chance, Griffin and Company, London, 1962.

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

[5] N. L. Johnson and S. Kotz, Discrete Distributions, Houghton Mifflin Company, Boston, Massachusetts, U.S.A., 1969.

[6] Z. A. Lomnicki and S. K. Zaremba, The asymptotic distributions of estimates of the amount of transmitted information, Information and Control, 2 (1959), 260-284.

[7] R. D. Luce, The theory of selective information and some of its behavioral applications, in Developments in Mathematical Psychology, R. D. Luce, Editor, The Free Press, Glencoe, Illinois, U.S.A. 1955.

[8] G. A. Miller, Note on the bias of information estimates, in Information Theory in Psychology, H. Quastler, Editor, The Free Press, Glencoe, Illinois, U.S.A., 1955.

[9] G. A. Miller and W.G. Madow, On the maximum likelihood estimate of the Shannon-Weaver measure of information, Air Force Cambridge Research Center Technical Report 54-75, 1954.

[10] A. M. Zubkov, Limit distributions for a statistical estimate of the entropy, Teor. Verojatnost i Primenen., 18 (1973), 643-650.

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE

READ INSTRUCTIONS BEFORE COMPLETING FORM

1. REPORT NUMBER
#1605

2. GOVT ACCESSION NO.

3. RECIPIENT'S CATALOG NUMBER

4. TITLE (and Subtitle)
THE STATISTICAL ESTIMATION OF ENTROPY IN THE NON-PARAMETRIC CASE

5. TYPE OF REPORT & PERIOD COVERED
Summary Report - no specific reporting period

6. PERFORMING ORG. REPORT NUMBER

7. AUTHOR(s)
Bernard Harris

8. CONTRACT OR GRANT NUMBER(s)
DAAG29-75-C-0024

9. PERFORMING ORGANIZATION NAME AND ADDRESS
Mathematics Research Center, University of Wisconsin
610 Walnut Street
Madison, Wisconsin 53706

10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS

11. CONTROLLING OFFICE NAME AND ADDRESS
U. S. Army Research Office
P.O. Box 12211
Research Triangle Park, North Carolina 27709

12. REPORT DATE
December 1975

13. NUMBER OF PAGES
43

14. MONITORING AGENCY NAME & ADDRESS(if different from Controlling Office)

15. SECURITY CLASS. (of this report)
UNCLASSIFIED

15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)

Approved for public release; distribution unlimited.

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

18. SUPPLEMENTARY NOTES

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

Estimation of entropy

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

A procedure based on quadrature methods as suggested as a means of overcoming some deficiencies of the natural estimator of entropy.

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)