NOTE ON ESTIMATING INFORMATION[1]

By Colin R. Blyth

University of Illinois and *Stanford University*

1. Summary. This note is concerned with estimation of the Shannon-Wiener measure of information. Low bias estimates are obtained, and their bias and variance. These estimates are extended to the case where the number of possible values of the random variable is not known. The estimates are compared asymptotically with the maximum likelihood estimates. They are also compared with the minimax estimates (for squared error loss function) for a few special cases where these are easily found.

2. Introduction. Consider a random variable Y with finitely many distinct possible values:

$$P(Y = a_i) = p_i, \qquad i = 1, \cdots, k.$$

A *metric* measure of dispersion of Y measures how spread out the distribution of Y is, in terms of distance in the space of Y. If this space has no relevant distance function (e.g., $k = 3$, $a_1 =$ green, $a_2 =$ red, $a_3 =$ white), there is no relevant metric measure of dispersion. An *absolute* measure of dispersion of Y measures the degree to which the total probability of 1 is broken up into pieces in the distribution of Y. Such a measure is a function of $p_1, \cdots, p_k$ only; is free from dependence on the a_i's; is large when the probability is much broken up (e.g., $p_1, \cdots, p_k = 1/k, \cdots, 1/k$), small when it is not much broken up (e.g., $p_1, \cdots, p_k = .99, .01, 0, \cdots, 0$).

In handling both kinds of dispersion measures the following addition property plays the same important role: {Divide the values of Y into groups. Dispersion of Y = between group dispersion + expected within group dispersion.}. Knowing the distribution of Y gives information useful in predicting Y. Actually observing Y gives additional information—enough for perfect prediction. This additional information can be called information in Y or unpredictability of Y and can be measured by a measure of dispersion. In this language the addition property says that the information in observing Y equals the information in observing which group Y falls in plus the expected information in observing which member of that group.

For real valued Y the addition property identifies variance (except for a constant multiplier) among all metric dispersion measures of the form $Ef(|Y - EY|)$ with f continuous. This easily extends to weighted averages of the partial variances when Y has values in a Euclidean n-space. Similarly the addition property identifies information or entropy $H = -\sum p_i \log_2 p_i$ (except for a constant multiplier) among all absolute dispersion measures $f_k(p_1, \cdots, p_k)$ with

Received January 23, 1958; revised April 18, 1958.

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

f_k continuous and $f_k(1/k, \cdots, 1/k)$ an increasing function of k, as is proved in [1].

The addition property leads to very convenient mathematical simplifications. For this reason variance and information H are very widely used dispersion measures. But there seems to be little intuitive necessity for the addition property. Thus in the metric case it seems quite reasonable to use measures like $E|Y - EY|$ which lack the property- and loss functions other than squared error. Similarly in the absolute case it would seem quite reasonable to use measures like the natural chi-square measure $(k-1) - k\sum(p_i - 1/k)^2$ which lack the property. Essentially equivalent to this chi-square measure is the following linear function of it, which is the terms of order up to 2 in a Taylor series for H:

$$H_2 = 1 - \frac{C}{2}\left\{k - 2 + \sum_{i=1}^{k}(2p_i - 1)^2/2\right\}.$$

If $f(1, 0, \cdots, 0) = 0$ is desired we could make the necessary subtraction from the measure H_2.

This note is concerned with estimation of information or entropy of Y:

$$H = H(p_1, \cdots, p_k) = -C\sum_{i=1}^{k} p_i \log p_i,$$

where $C = \log_2 e = 1.442695$ and $p_i \log p_i$ is taken to be 0 whenever $p_i = 0$. Our estimate is to be based on independent repetitions $Y_1, \cdots, Y_n$ of the experiment Y. Then $X_1, \cdots, X_k$, where X_i is the number of Y's with the value a_i, is a sufficient statistic for $p_1, \cdots, p_k$ and has the following multinomial family of possible distributions:

$$P(X_1, \cdots, X_k = x_1, \cdots, x_k) = n!\prod_{i=1}^{k} p_i^{x_i}/x_i!,$$

$$x_i = 0, 1, \cdots, n, \qquad \sum_{i=1}^{k} x_i = n, \qquad 0 \leqq p_i \leqq 1, \qquad \sum_{i=1}^{k} p_i = 1. \tag{1}$$

We are now concerned, then, with the problem of what function $f_k(X_1, \cdots, X_k)$ to use as an estimate for H. The maximum likelihood estimate is considered by Miller and Madow [2]; it is good when n is large, but is likely to be poor for n small. One reasonable estimate would be the best unbiased one. Upon noticing that there is no unbiased estimate we will consider instead best estimates with low bias.

3. Low bias estimation, k known. Since (1) is a complete family of distributions, the problem of unbiased estimation of any function $g(p_1, \cdots, p_k)$ is solved by Lehmann and Scheffe [3]. In fact, since $Ef(X_1, \cdots, X_k)$ is for every every function f a polynomial in $p_1, \cdots, p_k$ of degree at most n, no functions of $p_1, \cdots, p_k$ other than such polynomials possess unbiased estimates. And using the usual factorial notation $x^{(\nu)} = x(x-1)\cdots(x-\nu+1)$ we have

$$E\{X_1^{(\nu_1)} \cdots X_k^{(\nu_k)}\} = n^{(\nu_1+\cdots+\nu_k)}p_1^{\nu_1} \cdots p_k^{\nu_k},$$

which reduces to $0 \equiv 0$ whenever $\sum_{i=1}^{k} \nu_i > n$ but not otherwise. We therefore have

$$E \sum c(\nu_1, \cdots, \nu_k) X_1^{(\nu_1)} \cdots X_k^{(\nu_k)} / n^{(\nu_1+\cdots+\nu_k)} = \sum c(\nu_1, \cdots, \nu_k) p_1^{\nu_1} \cdots p_k^{\nu_k},$$

where summation is over any set of $(\nu_1, \cdots, \nu_k)$'s with $\sum_{i=1}^{k} \nu_i \leqq n$ for every member. It follows from the completeness that $\sum c(\nu_1, \cdots, \nu_k) X_1^{(\nu_1)} \cdots X_k^{(\nu_k)} / n^{(\nu_1+\cdots+\nu_k)}$ is the unique uniformly minimum variance (U.M.V.) unbiased estimate of $\sum c(\nu_1, \cdots, \nu_k) p_1^{\nu_1} \cdots p_k^{\nu_k}$ whenever $\sum_{i=1}^{k} \nu_i \leqq n$ for every term of the sum. This solves the problem of unbiased estimation of all functions $g(p_1, \cdots, p_k)$ because the U.M.V. unbiased estimate of every degree $\leqq n$ polynomial in $p_1, \cdots, p_k$ has been written down and no other functions of $p_1, \cdots, p_k$ possess unbiased estimates.

It is now clear that there is no unbiased estimate of H. If low bias is what we want the next best thing would be to use the U.M.V. unbiased estimate of the degree n polynomial which is in some sense (smallest maximum distance apart, for example) closest to H. A much more easily obtained polynomial which agrees quite closely with H is the terms of degree $\leqq n$ in the Taylor series expansion of H about the point $(\frac{1}{2}, \cdots, \frac{1}{2})$. We will consider use of the U.M.V. unbiased estimate of this polynomial as an estimate for H.

Writing $\gamma_i = p_i - \frac{1}{2}$ we have

$$\begin{aligned} p_i \log p_i &= (\tfrac{1}{2} + \gamma_i) \log (\tfrac{1}{2} + \gamma_i) \\ &= -\frac{1}{2C} + \left(1 - \frac{1}{C}\right)\gamma_i + \frac{1}{2}\left\{\frac{(2\gamma_i)^2}{1 \cdot 2} - \frac{(2\gamma_i)^3}{2 \cdot 3} + \frac{(2\gamma_i)^4}{3 \cdot 4} - \cdots\right\}. \end{aligned}$$

Hence

$$\begin{aligned} H &= -C \sum_{i=1}^{k} p_i \log p_i \\ &= 1 - \frac{C}{2}\left\{\sum_{i=1}^{k} (2\gamma_i) + \sum_{i=1}^{k} \frac{(2\gamma_i)^2}{1 \cdot 2} - \sum_{i=1}^{k} \frac{(2\gamma_i)^3}{2 \cdot 3} + \cdots\right\} \\ &= 1 - \frac{C}{2}\left\{(2 - k) + \sum_{\alpha=2}^{\infty} \sum_{i=1}^{k} (-2\gamma_i)^\alpha / \alpha^{(2)}\right\}. \end{aligned}$$

Here $|2\gamma_i| \leqq 1$ so all series converge absolutely and can be rearranged. Also, $\sum_{i=1}^{k} (2\gamma_i) = 2 - k$. For any integer $r \leqq n$ we now write

$$\begin{aligned} H_r &= 1 - \frac{C}{2}\left\{(2 - k) + \sum_{\alpha=2}^{r} \sum_{i=1}^{k} (-2\gamma_i)^\alpha / \alpha^{(2)}\right\} \\ &= 1 - \frac{C}{2}\left\{(2 - k) + \sum_{\alpha=2}^{r} \sum_{i=1}^{k} \frac{1}{\alpha^{(2)}} \sum_{\nu=0}^{\alpha} \binom{\alpha}{\nu} (-2p_i)^\nu\right\}. \end{aligned}$$

The U.M.V. unbiased estimate of H_r is

$$Z_r = 1 - \frac{C}{2}\left\{(2 - k) + \sum_{\alpha=2}^{r} \sum_{i=1}^{k} \sum_{\nu=0}^{\alpha} \frac{1}{\alpha^{(2)}} \binom{\alpha}{\nu} (-2)^\nu \frac{X_i^{(\nu)}}{n^{(\nu)}}\right\}.$$

The bias of Z_r as an estimate for H is

$$B_r = EZ_r - H = H_r - H$$
$$= \frac{C}{2} \sum_{\alpha=r+1}^{\infty} \sum_{i=1}^{k} (-2\gamma_i)^\alpha / \alpha^{(2)}.$$

Now u^s is a convex function of u for s an even integer, and $-u^s$ is convex on $u \leqq 0$ for s an odd integer. From this it is easily shown that if $\sum_{i=1}^{k} u_i = 2 - k$ and $|u_i| \leqq 1$, then

$$(k - 1 + (-1)^s)\left(1 - \frac{2}{k}\right)^s \leqq (-1)^s \sum_{i=1}^{k} u_i^s \leqq k - 1 + (-1)^s.$$

These lower, upper bounds are achieved by the choices $(u_1, \cdots, u_k) = (2/k - 1, \cdots, 2/k - 1)$ and $(1, -1, \cdots, -1)$ respectively. Applying this to the series for B_r gives

$$\frac{C}{2} \sum_{\alpha=r+1}^{\infty} \frac{(k - 1 + (-1)^\alpha)(1 - 2/k)^\alpha}{\alpha^{(2)}} \leqq B_r \leqq \frac{C}{2}\left\{\frac{k - 1 - (-1)^r}{r} - 2 \sum_{\alpha=r+1}^{\infty} \frac{(-1)^\alpha}{\alpha}\right\}.$$

This lower bound is achieved when the p_i's are all $1/k$, and the upper bound is achieved when some $p_i = 1$. If $k > 2$ this lower bound is positive and we will use the estimate

$$Z_r' = Z_r - \frac{C}{2} \sum_{\alpha=r+1}^{\infty} \frac{(k - 1 + (-1)^\alpha)(1 - 2/k)^\alpha}{\alpha^{(2)}}$$

instead of Z_r for H because Z_r' has the same variance as Z_r and uniformly smaller bias. For a fixed set $\gamma_1, \cdots, \gamma_k$ we have

$$B_r \leqq \frac{Ck}{2r} (\max_i |2\gamma_i|)^{r+1}$$

and the corresponding result for the bias of the improved estimate Z_r'. To compute the variance of Z_r we now use the fact

$$X_i^{(\nu_1)} = \sum_{t=0}^{\nu_1} \frac{\nu_1^{(t)} \nu_2^{(t)}}{t!} (X_i - \nu_2)^{(\nu_1 - t)},$$

which gives

$$X_i^{(\nu_1)} X_i^{(\nu_2)} = \sum_{t=0}^{\min(\nu_1, \nu_2)} \frac{\nu_1^{(t)} \nu_2^{(t)}}{t!} X_i^{(\nu_1 + \nu_2 - t)}.$$

Further routine calculations now give

$$E\left(\frac{X_i^{(\nu_1)}}{n^{(\nu_1)}} - p_i^{\nu_1}\right)\left(\frac{X_i^{(\nu_2)}}{n^{(\nu_2)}} - p_i^{\nu_2}\right) = \sum_{t=1}^{\min(\nu_1, \nu_2)} \frac{\nu_1^{(t)} \nu_2^{(t)}}{t!} \frac{p_i^{\nu_1 + \nu_2 - t}(1 - p_i)^t}{n^{(t)}},$$

$$E\left(\frac{X_i^{(\nu_1)}}{n^{(\nu_1)}} - p_i^{\nu_1}\right)\left(\frac{X_j^{(\nu_2)}}{n^{(\nu_2)}} - p_j^{\nu_2}\right) = \left(\frac{n^{(\nu_1 + \nu_2)}}{n^{(\nu_1)} n^{(\nu_2)}} - 1\right) p_i^{(\nu_1)} p_j^{(\nu_2)}, \qquad i \neq j.$$

From these, the variance of Z_r is seen to be

$$\begin{aligned}\operatorname{var} Z_r &= E(Z_r - H_r)^2 \\ &= \frac{C^2}{4} E\left\{\sum_{\alpha=2}^{r}\sum_{\nu=0}^{\alpha}\binom{\alpha}{\nu}\frac{(-2)^\nu}{\alpha^{(2)}}\sum_{i=1}^{k}\left(\frac{X_i^{(\nu)}}{n^{(\nu)}} - p_i^\nu\right)\right\}^2 \\ &= \frac{C^2}{4}\sum_{\alpha_1=2}^{r}\sum_{\alpha_2=2}^{r}\sum_{\nu_1=0}^{\alpha_1}\sum_{\nu_2=0}^{\alpha_2}\binom{\alpha_1}{\nu_1}\binom{\alpha_2}{\nu_2}\frac{(-2)^{\nu_1+\nu_2}}{\alpha_1^{(2)}\alpha_2^{(2)}}\left\{\sum_{i_1\neq i_2=1}^{k} p_{i_1}^{\nu_1} p_{i_2}^{\nu_2}\left[\frac{n^{(\nu_1+\nu_2)}}{n^{(\nu_1)}n^{(\nu_2)}} - 1\right]\right. \\ &\qquad \left. + \sum_{i=1}^{k}\sum_{t=1}^{\min(\nu_1,\nu_2)}\frac{\nu_1^{(t)}\nu_2^{(t)}}{t!}\frac{p_i^{\nu_1+\nu_2-t}(1-p_i)^t}{n^{(t)}}\right\}.\end{aligned}$$

Grouping together terms of like order in n, the asymptotic variance as $n \to \infty$ of the sequence $\{Z_n\}$ of estimates is seen to be

$$\operatorname{var}(Z_n) \sim \frac{C^2}{n}\sum_{i=1}^{k} p_i(\log p_i - \sum_{j=1}^{k} p_j \log p_j)^2$$

provided the non-zero p_i's are not all equal; and

$$\operatorname{var}(Z_n) \sim \frac{C^2}{n}\cdot\frac{k^* - 1}{2(n-1)}$$

if the non-zero p_i's are all equal, where k^* is the number of non-zero p_i's. And for bias of this sequence of estimates we have asymptotically as $n \to \infty$

$$\frac{kC}{2n(n+1)}\left(1 - \frac{2}{k}\right)^{n+2} \leqq B_n \leqq \frac{kC}{2n}\{\max_i |2\gamma_i|\}^{n+1}.$$

Comparison with asymptotic results obtained by Miller and Madow for the maximum likelihood estimates $\{H_n'\}$ and the estimates $\{H_n''\} = \{H_n' + C(k-1)/2n\}$ shows the following: No asymptotic differences in variance. Asymptotic differences in expected square error only in the special cases where this has order $1/n^2$ or smaller. Bias is asymptotically much smaller for $\{Z_n\}$ than for $\{H_n'\}$ or $\{H_n''\}$ except for the special case when some $p_i = 1$.

For small n, numerical checking shows the following: Z_n has a smaller bias than H_n' or H_n'' over most of the range of $(p_1, \cdots, p_k)$. Comparing expected squared error as a whole, H_n' is quite poor and there is little to choose between Z_n and H_n'': sometimes one seems better, sometimes the other. In these comparisons Z_n and H_n'' are modified by substituting $C \log k$ for any value exceeding $C \log k$, since this uniformly reduces expected squared error. For example in Table 1 for $k = 2$, every value exceeding $C \log 2 = 1$ would be replaced by 1. This table gives $Z_n(x_1)$ in the upper part of each column and $H_n''(x_1)$ in the lower part, for $n = 2, 3, \cdots, 7$ and all possible x_1. Values not tabled are obtained from $Z_n(n - x_1) = Z_n(x_1)$ and $H_n''(n - x_1) = H_n''(x_1)$.

When $k = 2$, minimax estimates (squared error loss function can be found for small n by the usual method of guessing a least favorable a priori distribution λ for $p_1 = p$ and finding the corresponding Bayes estimate Z_λ. If the risk function

TABLE 1

Estimates of H for $k=2$ with $Z_n(x_1)$ in upper part of each column, $H''_n(x_1)$ in lower part

x_1 \ n	2	3	4	5	6	7
0	.27865	.27865	.15843	.15843	.11034	.11034
1	1.72135 1.36067	1.24045	1.12022	.92787	.84772	.74238
2	.36067	1.15875	1.12022 1.18034	1.12022	1.00801	.96222
3		.24045	.99162	1.11524	1.08516 1.12022	1.09961
4			.18034	.86619	1.03852	1.08828
5				.14427	.77024	.96617
6					.12022	.69472
7						.10305

TABLE 2

Comparison of low bias Z'_n and minimax Z^* estimates of H for $k=2$

n	$Z^*(0)$, $Z^*(1)$	$\sup_p R_{Z^*}(p)$	$\sup_p R_{Z'_n}(p)$
1	1/2, 1/2	.2500	1
2	$\sqrt{2}-1$, 1	.1716	.2602
3	.33673, .94400	.1134	.1444

$R_{Z_\lambda}(p)$ assumes its maximum value with λ-probability 1 then this λ is indeed least favorable and Z_λ in minimax. Here for λ we take unspecified probabilities at $n+1$ unspecified points one of which is 0, with the restriction that λ be symmetric about $p=\frac{1}{2}$. The points and probabilities are then determined so that $R_{Z_\lambda}(p)$ will have equal maxima at these points. We will compare the risk functions $R_{Z'_n}(p)$ and $R_{Z^*}(p)$ [$Z'_n = Z_n$ except $Z'_n = 1$ when $Z'_n > 1$; Z^* is minimax]. One point of interest is the degree to which $\sup_p R_{Z'_n}(p)$ exceeds $\sup_p R_{Z^*}(p)$. This comparison, given in Table 2 for $n = 1, 2, 3$, is of particular interest for small values of n where Z'_n would be expected to show up most poorly. Actually Z'_n does quite well even for $n = 2, 3$: Z'_n seems to deviate from unbiasedness in the direction of being like the minimax estimate.

Similarly, for $k = 2$, we can compare the minimax and U.M.V. unbiased estimates of the chi square dispersion measure H_2. Equivalently we can compare the minimax estimate T^* and the U.M.V. unbiased estimate T of the binomial variance pq, of which H_2 is just a linear function. This comparison is given for $n = 1, 2, \cdots, 5$ in Table 3. Note that T is very poor compared to T^* for small n, compares more favorably as n increases. For example the ratio $\sup_p R_T(p)/\sup_p R_{T^*}(p)$ is 5.83 when $n = 2$, decreases to 3.37 by $n = 5$. The comparison indicates that for small n the minimax estimate for binomial variance is decidedly

TABLE 3

Comparison of minimax T^* and U.M.V. unbiased T estimates for binomial variance pq

n	$T^*(0), T^*(1), T^*(2)$	$\sup_p R_{T^*}(p)$	$T(0), T(1), T(2)$	$\sup_p R_T(p)$
1	.125, .125, —	.015625	—	—
2	.10355, .25, .10355	.010724	0, 1/2, 0	.062500
3	.08333, .25, .25	.006944	0, 1/3, 1/3	.027778
4	.07158, .20228, .24584	.005124	0, 1/4, 1/3	.018750
5	.06508, .17797, .22841	.004235	0, 1/5, 3/10	.014286

preferable to the U.M.V. unbiased estimate. The estimate T^* is found in the same way as Z^* except that for $n = 3, 4, 5$ T^* has constant risk and can be more easily found by showing that the only constant risk estimate can be Bayes.

4. Low bias estimation, k unknown. When k is unknown we shall consider the estimates obtained by acting as though k were equal to the observed number of different Y values and using the estimates of the preceding section. Now we have

$$Z_r = 1 - \frac{C}{2}\{W_1 + \cdots + W_k\},$$

where

$$W_i = \left(2\frac{X_i}{n} - 1\right) + \sum_{\alpha=2}^{r}\sum_{\nu=0}^{\alpha}\binom{\alpha}{\nu}\frac{(-2)^\nu}{\alpha^{(2)}}\frac{X_i^{(\nu)}}{n^{(\nu)}}.$$

The modification of Z_r just suggested for use in the case k unknown is

$$Z_r^* = 1 - \frac{C}{2}\{W_1^* + \cdots + W_k^*\},$$

where

$$W_i^* = W_i \qquad \text{if } X_i \neq 0,$$
$$= 0 \qquad \text{if } X_i = 0.$$

Since $W_i = -1/r$ when $X_i = 0$ we have

$$P(W_i^* = W_i) = 1 - (1 - p_i)^n,$$
$$P(W_i^* = W_i + 1/r) = (1 - p_i)^n.$$

Hence

$$EW_i^* = EW_i + \frac{1}{r}(1 - p_i)^n,$$

$$EZ_r^* = 1 - \frac{C}{2}\left\{EW_1 + \cdots + EW_k + \frac{1}{r}\sum_{i=1}^{k}(1 - p_i)^n\right\}$$
$$= EZ_r - \frac{C}{2r}\sum_{i=1}^{k}(1 - p_i)^n = H_r - \frac{C}{2r}\sum_{i=1}^{k}(1 - p_i)^n.$$

The bias of Z_r^* as an estimate for H is

$$B_r^* = EZ_r - H = H_r - H - \frac{C}{2r}\sum_{i=1}^{k}(1-p_i)^n$$

$$= B_r - \frac{C}{2r}\sum_{i=1}^{k}(1-p_i)^n.$$

The variance of Z_r^* is found as follows:

$$\text{var } Z_r^* = \frac{C^2}{4}\left\{\sum_{i=1}^{k}\text{var } W_i^* + \sum_{i\neq j=1}^{k}\text{cov } W_i^* W_j^*\right\},$$

$$\text{var } W_i^* = E(W_i^* - EW_i^*)^2$$

$$= [1-(1-p_i)^n]E\left\{(W_i - EW_i) - \frac{1}{r}(1-p_i)^n\right\}^2$$

$$+ (1-p_i)^n E\left\{(W_i - EW_i) + \frac{1}{r} - \frac{1}{r}(1-p_i)^n\right\}^2$$

$$= E(W_i - EW_i)^2 + \frac{1}{r^2}\{(1-p_i)^n - (1-p_i)^{2n}\}$$

$$= \text{var } W_i + \frac{1}{r^2}\{(1-p_i)^n - (1-p_i)^{2n}\}.$$

Furthermore, we have for $i \neq j$

$$P(W_i^*, W_j^* = W_i, W_j) = 1 - (1-p_i)^n - (1-p_j)^n + (1-p_i-p_j)^n,$$

$$P(W_i^*, W_j^* = W_i + 1/r, W_j) = (1-p_i)^n - (1-p_i-p_j)^n,$$

$$P(W_i^*, W_j^* = W_i, W_j + 1/r) = (1-p_j)^n - (1-p_i-p_j)^n,$$

$$P(W_i^*, W_j^* = W_i + 1/r, W_j + 1/r) = (1-p_i-p_j)^n.$$

So the covariance of W_i^*, W_j^*, $i \neq j$, is

$$\text{cov } W_i^*, W_j^* = E(W_i^* - EW_i^*)(W_j^* - EW_j^*)$$

$$= [1-(1-p_i)^n - (1-p_j)^n - (1-p_i-p_j)^n]$$

$$\cdot E\left\{(W_i - EW_i) - \frac{(1-p_i)^n}{r}\right\}\left\{(W_j - EW_j) - \frac{(1-p_j)^n}{r}\right\}$$

$$+ [(1-p_i)^n - (1-p_i-p_j)^n]$$

$$\cdot E\left\{(W_i - EW_i) + \frac{1-(1-p_i)^n}{r}\right\}\left\{(W_j - EW_j) - \frac{(1-p_j)^n}{r}\right\}$$

$$+ [(1-p_j)^n - (1-p_i-p_j)^n]$$

$$\cdot E\left\{(W_i - EW_i) - \frac{(1-p_i)^n}{r}\right\}\left\{(W_j - EW_j) + \frac{1-(1-p_j)^n}{r}\right\}$$

$$+ (1 - p_i - p_j)^n E\left\{(W_i - EW_i) + \frac{1 - (1 - p_i)^n}{r}\right\} \cdot \left\{(W_j - EW_j) + \frac{1 - (1 - p_j)^n}{r}\right\}$$

$$= \operatorname{cov} W_i, W_j + \frac{1}{r^2}\{(1 - p_i - p_j)^n - (1 - p_i)^n(1 - p_j)^n\}.$$

Hence

$$\operatorname{var} Z_r^* = \operatorname{var} Z_r + \frac{C^2}{4r^2}\left\{\sum_{i=1}^{k}[(1 - p_i)^n - (1 - p_i)^{2n}] + \sum_{i \neq j=1}^{k}[(1 - p_i - p_j)^n - (1 - p_i)^n(1 - p_j)^n]\right\}$$

$$= \operatorname{var} Z_r + \frac{C^2}{4r^2}\left\{\sum_{i=1}^{k}(1 - p_i)^n - \left[\sum_{i=1}^{k}(1 - p_i)^n\right]^2 + \sum_{i \neq j=1}^{k}(1 - p_i - p_j)^n\right\}.$$

When only one value of Y is observed, which happens with probability $\sum_{i=1}^{k} p_i^n$, the value of Z_r^* is $(C/2)\{(-1)^r/r - 2\sum_{\alpha=r}^{\infty}(-1)^\alpha/\alpha\}$. Since $u_m > 0$, $u_m \to 0$, $u_m > u_{m+1}$ and $u_{m+1} - u_m > u_{m+2} - u_{m+1}$ together imply convergence of $\sum_{m=1}^{\infty}(-1)^{m-1}u_m$ to a value $> u_1/2$, this value of Z_r^* has the sign of $(-1)^{r-1}$. In the case r even, this negative value should be replaced by 0; bias and variance of the resulting modification of Z_r^* are easily found. This point does not arise in Section 3 because if $k = 1$ is known, $H = 0$ is known and estimation is not needed.

A similar discussion can be given for the estimates $H_r''^* = H_r' + C(k^* - 1)/2r$. The maximum likelihood estimates H_r' do not require knowledge of k so can be used unchanged in the case k unknown.

REFERENCES

[1] C. E. Shannon and W. Weaver, *The Mathematical Theory of Communication*, University of Illinois Press, 1949.

[2] G. A. Miller and W. G. Madow, "On the maximum likelihood estimate of the Shannon-Wiener measure of information," Air Force Cambridge Research Center, 1954.

[3] E. L. Lehmann and H. Scheffé, "Completeness, similar regions and unbiased estimation I," *Sankhya*, Vol. 10 (1950), pp. 305–340.

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