Asymptotic Bias in Information Estimates and the Exponential (Bell) Polynomials

Jonathan D. Victor
Department of Neurology and Neuroscience, Weill Medical College of Cornell University, New York City, New York 10021, U.S.A.

We present a new derivation of the asymptotic correction for bias in the estimate of information from a finite sample. The new derivation reveals a relationship between information estimates and a sequence of polynomials with combinatorial significance, the exponential (Bell) polynomials, and helps to provide an understanding of the form and behavior of the asymptotic correction for bias.

1 Introduction

In its most basic form, application of the tools of information theory to laboratory data relies on the estimation of the information in a process consisting of independent occurrences of K kinds of mutually exclusive events, each of which occurs with a probability q_j ($j = 1, \ldots, K$) (Rieke, Warland, de Ruyter van Steveninck, & Bialek, 1997). The quantity

$$H = -\frac{1}{\ln 2} \sum_{j=1}^{K} q_j \ln q_j \tag{1.1}$$

is the information (in bits) associated with a single observation. Typically, the probabilities q_j are not known and must be estimated from a finite set of observations. It is well known that the naive estimate for H, based on replacement of the exact probabilities q_j by their empirical probabilities observed from N observations, downwardly biases the estimate of equation 1.1. Essentially this is because equation 1.1 is a concave-downward function, so an average estimate for H derived from a range of estimates of the true probabilities q_j is less than the value of H given by equation 1.1 at the center of this range.

Several authors (Carlton, 1969; Miller, 1955) have derived asymptotic estimates for this bias in the limit of large N. The leading term in the asymptotic estimate of the bias depends on K, the number of kinds of events, and N, the number of observations but, remarkably, is independent of the probabilities q_j of the events. These calculations are readily extended to estimates of mutual information (Miller, 1955; Treves & Panzeri, 1995), since mutual information in a table is the sum of the information in the distributions of

the marginal probabilities, minus the information in the distribution of table entries. Because the number of rows (say, K_R) and columns (say, K_C) of a nontrivial table is fewer than the number of entries in the table ($K_R K_C$), the downward bias in simple estimates of information (see equation 1.1) translates into an upward bias in estimates of mutual information.

This article presents a new and concise derivation of the bias estimate. The new derivation clarifies the basis for the lack of dependence of the bias on the probabilities q_j, and reveals a relationship between information estimates and the exponential (Bell) polynomials (Bell, 1934b), a sequence of integer polynomials with a well-known combinatorial interpretation.

2 Results

We consider an estimate of information from a set of N events, each of which can independently have one of K possible mutually exclusive outcomes. The probability of the jth outcome is denoted q_j, and $\sum_{j=1}^{K} q_j = 1$.

We define

$$U(N, s) = \left\langle \sum_{j=1}^{K} p_j^s \right\rangle_N . \tag{2.1}$$

The quantities p_j are estimates of the probabilities q_j, which are considered to be definite but unknown. That is, $p_j = n_j/N$ is the empirical probability of outcome j, as estimated from a set of N observations in which this outcome occurred n_j times. $\langle\ \rangle_N$ denotes an average over all sets of N observations drawn from this universe. The expected value $\langle H \rangle_N$ for the estimate of the information from N observations is thus

$$\langle H \rangle_N = -\frac{1}{\ln 2} \frac{\partial}{\partial s} U(N, s)|_{s=1}. \tag{2.2}$$

We form the generating function

$$\tilde{U}(z, s) = \sum_{N=0}^{\infty} \frac{N^s z^N}{N!} U(N, s). \tag{2.3}$$

To evaluate the expectation implicit in equations 2.1 and 2.3, we assign to each possible set of observations (say, n_j occurrences of each outcome j, with $N = \sum_{j=1}^{K} n_j$) its corresponding multinomial probability,

$$\frac{N!}{\prod_{j=1}^{K} n_j!} \prod_{j=1}^{K} (q_j)^{n_j}.$$

In this expression, the factorial terms count the number of ways of ordering n_j occurrences of each outcome j into a sequence of N observations, and

the terms involving q_j specify the probability of any one such sequence of observations. Inclusion of the multinomial probability in equation 2.1 converts it to

$$U(N,s) = \frac{N!}{N^s} \sum_{n_j} \left(\sum_{J=1}^{K} n_J^s \right) \left[\prod_{j=1}^{K} \frac{(q_j)^{n_j}}{n_j!} \right], \tag{2.4}$$

and equation 2.3 to

$$\tilde{U}(z,s) = \sum_{n_j=0}^{\infty} \left(\sum_{J=1}^{K} n_J^s \right) \left[\prod_{j=1}^{K} \frac{(q_j z)^{n_j}}{n_j!} \right], \tag{2.5}$$

where the outer sum is over all sets of nonnegative integers n_j, satisfying $\sum_{j=1}^{K} n_j = N$ in equation 2.4, but unrestricted in equation 2.5. Term-by-term consideration of the J sum in equation 2.5 and rearrangement lead to

$$\begin{aligned} \tilde{U}(z,s) &= \sum_{J=1}^{K} \left(\sum_{n_j=0}^{\infty} n_J^s \frac{(q_J z)^{n_J}}{n_J!} \right) \left[\prod_{j \neq J} \left(\sum_{n_j=0} \frac{(q_j z)^{n_j}}{n_j!} \right) \right] \\ &= e^z \sum_{J=1}^{K} b(q_J z, s), \end{aligned} \tag{2.6}$$

where we have defined

$$b(z,s) = e^{-z} \sum_{n=0}^{\infty} n^s \frac{z^n}{n!}, \tag{2.7}$$

and made use of $\sum_{J=1}^{K} q_J = 1$.

For nonnegative integer values of s, $b(z,s)$ are polynomials and have the generating function

$$\sum_{s=0}^{\infty} b(z,s) \frac{u^s}{s!} = e^{z(e^u - 1)}. \tag{2.8}$$

Thus, $b(z,s)$ are a simple example of the exponential (Bell) polynomials in z (the quantities η in Bell, 1934b), and $b(1,s)$ are the exponential (Bell) numbers (Bell, 1934a) (history reviewed in Rota, 1964) . For all (not necessarily integral) $s \geq 0$, $b(z,s)$ satisfies the recurrence relation

$$b(z, s+1) = z \left[b(z,s) + \frac{\partial}{\partial z} b(z,s) \right], \tag{2.9}$$

as can be verified from equation 2.7 or 2.8. Also from equation 2.7, $b(z, 0) = 1$, and for integer $s > 0$, the leading terms of $b(z, s)$ are

$$b(z,s) = z^s + \frac{s(s-1)}{2}z^{s-1} + \frac{s(s-1)(s-2)(3s-5)}{24}z^{s-2} + \frac{s(s-1)(s-2)^2(s-3)^2}{48}z^{s-3} + \cdots. \quad (2.10)$$

The Bell polynomials have a combinatorial interpretation (Bell 1934a, 1934b; Rota, 1964): the coefficient of z^t in $b(z, s)$ is the number of ways of placing s distinguishable objects into t indistinguishable containers. In particular, the coefficient of z^{s-1} is $s(s-1)/2$, the number of ways of choosing one pair of the s objects to share a container.

From equation 2.3,

$$\begin{aligned} -\frac{1}{\ln 2}\frac{\partial}{\partial s}\tilde{U}(z,s)|_{s=1} &= -\frac{1}{\ln 2}\sum_{N=0}^{\infty}\frac{N\ln N}{N!}z^N + \sum_{N=0}^{\infty}\frac{N}{N!}z^N\langle H\rangle_N \\ &= -\frac{e^z}{\ln 2}\frac{\partial}{\partial s}b(z,s)|_{s=1} + \sum_{N=0}^{\infty}\frac{z^N}{(N-1)!}\langle H\rangle_N, \end{aligned} \quad (2.11)$$

where we have used $U(N, 1) = 1$ (from equation 2.1) and equation 2.2 in the first step, and equation 2.7 in the second step. Combining this with equation 2.6 yields

$$\sum_{N=0}^{\infty}\frac{1}{(N-1)!}z^N\langle H\rangle_N = \frac{e^z}{\ln 2}\left[\frac{\partial}{\partial s}b(z,s)|_{s=1} - \frac{\partial}{\partial s}\sum_{J=1}^{K}b(q_J z, s)|_{s=1}\right]. \quad (2.12)$$

Equation 2.12 is exact. To derive an asymptotic estimate, we estimate the partial derivatives in equation 2.12 by assuming that formula 2.10, a finite series for integer values of s, is a useful approximation at noninteger values as well. That is, we use the approximation

$$\frac{\partial}{\partial s}b(z,s)|_{s=1} = z\ln z + \frac{1}{2} + \frac{1}{12z} + \frac{1}{12z^2}\cdots. \quad (2.13)$$

The behavior of this approximation is illustrated in Figure 1. Note that as terms beyond the constant term are added, the improvement in the approximation for large values of z is accompanied by a worsening for values of $z < 1$.

Inserting this approximation into equation 2.12 and identifying corresponding coefficients of z^N on its two sides leads directly to

$$H = \langle H\rangle_N + \frac{1}{\ln 2}\left[\frac{K-1}{2N} + \frac{1}{12N(N+1)}\left(\sum_{J=1}^{K}\frac{1}{q_J} - 1\right) + \ldots\right]. \quad (2.14)$$

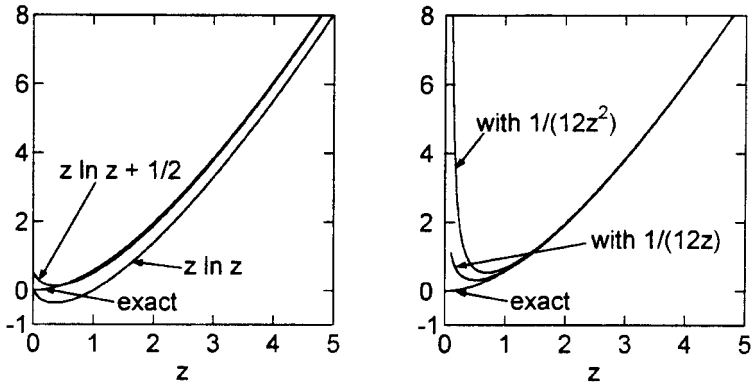

Figure 1: Behavior of an asymptotic estimate for $\frac{\partial}{\partial s} b(z, s)|_{s=1}$. (Left) Approximations provided by one and two terms of equation 2.13. (Right) Approximations provided by three and four terms of equation 2.13.

The first correction term corresponds to the previous results of several authors (Carlton, 1969; Miller, 1955; Treves & Panzeri, 1995). The fact that it is independent of the probabilities q_J reflects the constant value (1/2) of the first correction term in equation 2.13.

The mth correction term derived from our approach has a denominator $N(N+1)\ldots(N+m-1)$. This is different from the asymptotic series derived by Treves and Panzeri (1995), which is strictly in inverse powers of N. Nevertheless, the approaches agree. For example, the second correction term in equation 2.14, whose value depends on the probabilities q_J, differs from the results of Treves and Panzeri (1995), but the difference is third order (i.e., $O(N^{-3})$), and thus is subsumed in the third correction term.

Implementation of this bias correction is not completely straightforward. In the laboratory, one has access only to estimates of the event probabilities q_J, the quantities we have denoted p_J. Equation 2.14 states that the naive estimate of information $\langle H \rangle_N$, obtained by replacing the q_J in equation 1.1 by p_J, is too low. The first correction term in equation 2.14 requires knowing the number of possible kinds of events, K, but not their probabilities. But even K may not be known in advance. One can be sure that an event is possible only if one has observed it, but one does not know that additional kinds of events are impossible, merely that they have not been observed in a sample of size N. Panzeri and Treves (1996) suggest a sophisticated approach for modifying the count of the observed number of kinds of events. Here we initially consider the simple strategy of setting K equal to the number of kinds of events that were actually observed within N trials.

The left panels of Figure 2 demonstrate the effects of this strategy for an experiment in which there are two kinds of events. The initial correction

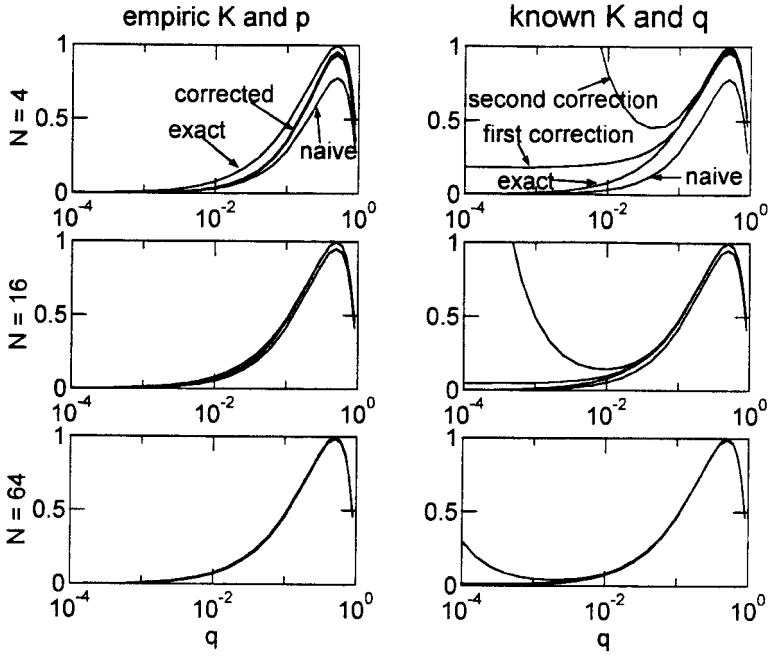

Figure 2: Comparison of strategies for the adjustment of empirical estimates of information in a simulation with two kinds of events (K) and 4, 16, or 64 trials (N). Exact value (see equation 1.1), naive estimate $\langle H \rangle_N$, and corrections of the naive estimate for bias. Abscissa: Probability q of the first kind of event ($q_1 = q, q_2 = 1 - q$). Ordinate: Information H. (Left panels) Number of kinds of events K and their probabilities estimated from the data. The two curves for the corrected estimates are virtually indistinguishable. (Right panels) Number of kinds of events K and their probabilities known in advance. On the right, the two-term correction leads to higher estimates of information (as labeled in the upper panel). All functions are symmetric about $q = 0.5$, but are plotted for q in $[0.0001, 0.9]$ on a logarithmic scale to emphasize the behavior for extreme values of q.

closes most of the gap between the naive estimate $\langle H \rangle_N$ and the true value H, provided that the number of trials is sufficiently large so that each kind of event has a reasonable chance of occurring ($Nq \gg 1$). When the number of trials is so small that one of the events will probably not be observed ($Nq \ll 1$), the correction is ineffective—as would be expected, since there is no empirical evidence for more than one kind of event. The second-order correction is very small in both regimes (nearly superimposed on the initial correction).

One might expect that the estimate of information could be improved if there were a better way to estimate either K or the event probabilities q_J. For illustrative purposes, the right panels of Figure 2 consider the extreme situation: that both are known exactly (but only the empiric values p_J are used to estimate information). The first correction is not as helpful. In the regime in which it is significant ($Nq \ll 1$), it amounts to an overcorrection, because the "correct" value $K = 2$ is always used, even though two kinds of events were not typically observed. Surprisingly, the second-order correction is even worse, resulting in large overestimates, because of the terms involving reciprocals of small probabilities q_J. In the first strategy considered, since the q_J are taken from empirical estimates, $1/q_J$ is limited by N, thus bounding the second-order correction. But there is no such limit here, since even smaller values of q_J, based on a priori knowledge, may occur. A hybrid strategy (using a priori knowledge of K, but not of the event probabilities q_J) results in performance that is worse than either of the two considered above (not shown).

This pattern of behavior was also seen in numerical experiments involving several kinds of events and a wide range of event probabililties. In sum, the most straightforward application of equation 2.14, making use of only what was observed, appears to be both conservative and effective. It fails under appropriate circumstances: when observation is so limited that possible modes of behavior have not been observed. Under these circumstances, higher-order corrections are not helpful.

3 Discussion

We derived an exact expression, equation 2.12, for the expected information estimate from an N-trial data set in terms of the partial derivative of the Bell polynomials with respect to their order, $\frac{\partial}{\partial s} b(z, s)|_{s=1}$. The asymptotic expansion for this derivative, equation 2.13, corresponds in a term-by-term fashion to an asymptotic expansion, equation 2.14, of the expected information estimate. The leading term in the expansion of $\frac{\partial}{\partial s} b(z, s)|_{s=1}$, namely $z \ln z$, corresponds to the (unbiased) estimate of information from an unlimited sample. The second term in the expansion of $\frac{\partial}{\partial s} b(z, s)|_{s=1}$, namely, $\frac{1}{2}$, corresponds to the initial correction due to finite sample size (the estimate of the bias), as derived by previous authors (Carlton, 1969; Miller, 1955; Treves & Panzeri, 1995). Since this term is a constant, the initial term in the asymptotic form for the bias depends on only the number of terms in the sum on the right-hand side of equation 2.12: the number of possible kinds of events K, and not on their probabilities q_J.

This analysis helps to explain why the high-order correction terms of equation 2.14 are not useful in practice. The higher-order correction terms reflect the successive approximations to the Taylor expansion of $\frac{\partial}{\partial s} b(z, s)|_{s=1}$ for small z. As is seen in Figure 1, the asymptotic series, equation 2.13, converges rapidly for large z, but diverges in the neighborhood of $z = 0$.

Thus, in the regime in which the higher-order correction terms might matter (low N and estimates of q below $1/N$), they worsen the estimate of bias (see the upper two right panels of Figure 2). When N is large, the second-order corrections do indeed improve the estimate of bias for some values of q, but the size of the correction is minuscule (see the lower panels of Figure 2).

Acknowledgments

This work was supported in part by EY9314. I thank Jeff Tsai and Danny Reich for comments on the manuscript.

References

Bell, E. T. (1934). Exponential numbers. *Amer. Math. Monthly*, *41*, 411–419.

Bell, E. T. (1934). Exponential polynomials. *Ann. Math.*, *35*, 258–277.

Carlton, A. G. (1969). On the bias of information estimates. *Psychological Bulletin*, *71*, 108–109.

Miller, G. A. (1955). Note on the bias on information estimates. *Information Theory in Psychology; Problems and Methods II-B*, 95–100.

Panzeri, S., & Treves, A. (1996). Analytical estimates of limited sampling biases in different information measures. *Network*, *7*, 87–107.

Rieke, F., Warland, D., de Ruyter van Steveninck, R. R., & Bialek, W. (1997). *Spikes: Exploring the neural code*. Cambridge, MA: MIT Press.

Rota, G-C. (1964). The number of partitions of a set. *Amer. Math. Monthly*, *71*, 498–504.

Treves, A., & Panzeri, S. (1995). The upward bias in measures of information derived from limited data samples. *Neural Computation*, *7*, 399–407.

Received September 27, 1999; accepted January 21, 2000.