the largest solution of a certain equation. In addition, the proportion of letter i in the maximum scoring subsequence was given explicitly. These results were extended by Karlin and Altschul (1990), who consider more general scoring schemes. They require the expected score of two letters to be negative and do not allow insertions or deletions. They describe generalizations of Arratia, Morris and Waterman (1988) and give a Chen–Stein style formula to assess statistical significance:

$$P\left(M_{n,m} > \frac{\ln(nm)}{\lambda^*} + x\right) \le Ce^{-\lambda^* x}.$$

It is straightforward to simulate random sequences and to obtain estimates of the 95th percentile of the score distribution. Why then is the Chen–Stein analysis valuable to molecular biology? An important part of the answer lies in the number of sequence comparisons that are made. When a newly determined sequence is compared to the full GenBank database, the sequence is compared to about 49,000 sequences of average length 1,000 nucleotides. These sequences are of differing compositions and could each require a simulation. Since comparison of sequences is the rate limiting step in database searches, we need to have a rapid, accurate way to assess statistical significance. The Chen–Stein method provides that. In addition, very small p-values are almost impossible to determine by simulation, and the Poisson approximation is of much use there.

Comment

Louis H. Y. Chen

The work of Arratia, Goldstein and Gordon is certainly an important contribution to the development of Poisson approximation. I am particularly impressed by their clever treatment of process and compound Poisson approximation and the large number of ingenious applications. Their work comes at about the same time as that of Barbour and Holst (1989) and Barbour, Holst and Janson (1988b), which is another important contribution to the development of Poisson approximation. The latter work and other new results will be reported in a forthcoming monograph by Barbour, Holst and Janson (1991). In considering dependence, Arratia, Goldstein and Gordon take an approach similar to that of Stein (1972) and Chen (1975, 1978), while Barbour and Holst (1989) and Barbour, Holst and Janson (1988b) assume the existence of certain coupling. The possibility of using coupling was also discussed in Stein (1986, pages 92–93). Arratia, Goldstein and Gordon cleverly adapt Poisson approximation to process and compound Poisson approximation, but Barbour chooses to develop new techniques (Barbour, 1988, Barbour and Brown, 1990; Barbour, Chen and Loh, 1990). Both the work of Arratia, Goldstein and Gordon and Barbour and colleagues have significantly advanced the theory and application of Poisson approximation.

I would like to mention a result of Barbour and Eagleson (1983), which has also played an important role in the development of Poisson approximation. Barbour and Eagleson improved significantly the bounds obtained by Chen (1975) on the solution of the difference equation in the Poisson approximation. The improved bounds have helped to ease substantially the task of bounding the error terms in the approximation.

Although it enjoys special attention (due mainly to the work of Arratia, Goldstein and Gordon and Barbour and colleagues), the method of Poisson approximation discussed in the article by Arratia, Goldstein and Gordon is a special case of a general method: Stein's method. In his fundamental paper, Stein (1972) introduced not only a new method of normal approximation but also a whole new way of proving approximation theorems. An exposition of Stein's method in its abstract form is given in the monograph by Stein (1986). For a more recent exposition, see Stein (1990).

CONNECTION WITH POINCARÉ INEQUALITIES

Arratia, Goldstein and Gordon mentioned several connections that the differential equation

$$f'(w) - wf(w) = h(w) - Nh \tag{1}$$

has with other areas. I would like to show here a connection that it has with Poincaré inequalities. For

Louis H. Y. Chen is Professor of Mathematics, National University of Singapore, Lower Kent Ridge Road, Singapore 0511, Republic of Singapore.

any random variable X, let us define the functional $\mathrm{R}(X) = \sup_g \mathrm{Var}[g(X)]/E[g'(X)^2]$, where the supremum is taken over the class of absolutely continuous functions g such that $0 < \mathrm{Var}[g(X)] < \infty$. If $\sigma^2 = \mathrm{Var}(X) < \infty$, define $\mathrm{U}(X) = \mathrm{R}(X)/\sigma^2$. It is proved in Brascamp and Lieb (1976) and Chernoff (1981) that, if X is normally distributed with variance σ^2, then $\mathrm{Var}[g(X)] \le \sigma^2 E[g'(X)^2]$ for all absolutely continuous g with $\mathrm{Var}[g(X)] < \infty$. Since equality holds with $g(x) = x$, this implies that $\mathrm{U}(X) = 1$. Borovkov and Utev (1984) proved the converse that in general $\mathrm{U}(X) \ge 1$ and that if $\mathrm{U}(X) = 1$ then X has a normal distribution. (They also proved that if $\mathrm{R}(X) < \infty$ then the moment generating function of X exists.) This characterizes the normal distribution and, using this characterization, they went on to prove that, if $X_1, X_2, \cdots$ is a sequence of random variables such that $\mathrm{U}(X_n) \to 1$, then the moment generating function of $(X_n - EX_n)/[\mathrm{Var}(X_n)]^{1/2}$ (which must exist by virtue of $\mathrm{R}(X) < \infty$) converges to that of the standard normal random variable on a neighborhood of zero.

We now sketch a simpler proof of the result of Borovkov and Utev concerning the characterization of the normal distribution. Without loss of generality, assume X to have zero mean and unit variance. Suppose $\mathrm{U}(X) = 1$. Then by a variational argument (similar to that in Borovkov and Uter) we obtain

$$E\{f'(X) - Xf(X)\} = 0 \tag{2}$$

for every bounded C^1 function of f. Now let f be the unique bounded solution of (1) with h being bounded and continuous. Then $Eh(X) = Nh$. This implies that X is normally distributed. This argument is used in a more general context in Chen and Lou (1987, Theorem 4.1) and also in Chen and Lou (1989). In the latter, it is proved that for a random variable X with nonvanishing continuous density function on an open interval I (possibly infinite), $\mathrm{R}(X) < \infty$ if and only if there exists a finite constant $c > 0$ and a C^1 function ψ on I with $\psi' > 0$ and $E|\psi(X)| < \infty$ such that

$$E\{c\psi'(X)f'(X) - \psi(X)f(X)\} = 0 \tag{3}$$

for every C^1 function f with compact support in I. In the case $c = 1$ and $\psi(x) = x$, (3) reduces to (2).

UNBOUNDED FUNCTIONS AND LARGE DEVIATIONS

The article by Arratia, Goldstein and Gordon deals with Poisson approximation for bounded functions. I would like to take this opportunity to discuss briefly Poisson approximation for unbounded functions and large deviations. Let $X_1, \cdots, X_n$ be independent indicators with

$$P(X_i = 1) = 1 - P(X_i = 0) = p_i,$$

$$W = \sum_{i=1}^n X_i,$$

$$W^{(i)} = W - X_i,$$

$$\lambda = \sum_{i=1}^n p_i$$

and let Z be a Poisson random variable with mean λ. We begin with the following identity

$$Eh(W) - Eh(Z) = \sum_{i=1}^n p_i^2 E V_\lambda h(W^{(i)}), \tag{4}$$

where h is not necessarily bounded, $V_\lambda h(w) = f_h(w+2) - f_h(w+1)$ and f_h is a solution of the difference equation

$$\lambda f(w+1) - wf(w) = h(w) - Eh(Z). \tag{5}$$

Let $A_i(k) = P(W^{(i)} = k)/P(Z = k)$. Then the right hand side of (4) can be written as

$$\sum_{i=1}^n p_i^2 E A_i(Z) V_\lambda h(Z).$$

It is proved in Chen (1975c, page 998) that

$$A_i(k) \le S/(1 - p_i),$$

where

$$S = \left\{1 + \lambda^{-1} \sum_{i=1}^n p_i^2/(1 - p_i)\right\}^t$$

and t is the largest integer not exceeding

$$\lambda + 1 + \sum_{i=1}^n p_i^2/(1 - p_i).$$

It is not difficult to see that S is of the order $\exp(\sum_{i=1}^n p_i^2)$ (since $\lambda^{-1}\sum_{i=1}^n p_i^2$ is small). Using this result, we can bound the right hand side of (4) to obtain the following result:

$$\begin{aligned} |Eh(W) - Eh(Z)| &\le S\left(\sum_{i=1}^n \frac{p_i^2}{1-p_i}\right)\Big\{(2\lambda^2)^{-1}EZ(Z-1)|h(Z)| \\ &\qquad + \lambda^{-1}EZ|h(Z)| + \frac{3}{2}E|h(Z)|\Big\} \\ &= S\left(\sum_{i=1}^n \frac{p_i^2}{1-p_i}\right)\Big\{\frac{1}{2}E|h(Z+2)| \\ &\qquad + E|h(Z+1)| + \frac{3}{2}E|h(Z)|\Big\}. \end{aligned} \tag{6}$$

By iterating (4) and using similar arguments as for (6), the following large deviation result is obtained in Chen and Choi (1990): let h be a polynomial, $A \subset \{1, 2, \cdots\}$, $a = \min A$, $H = hI_A$. Suppose $\max_{1 \le i \le n} p_i \to 0$, λ remains bounded and

$$a = o\left(\lambda\left[\sum_{i=1}^{n} p_i^2\right]^{-1/2}\right) \quad \text{as } n \to \infty.$$

Then

$$\frac{EH(W)}{EH(Z)} - 1 \sim -\frac{a^2}{2\lambda^2}\left(\sum_{i=1}^{n} p_i^2\right). \tag{7}$$

Here $X_1, \cdots, X_n$ and $p_1, \cdots, p_n$ are regarded as triangular arrays. By taking $h \equiv 1$ and $A = \{z + 1, z + 2, \cdots\}$ we obtain the following corollary:

$$\frac{P(W > z)}{P(Z > z)} - 1 \sim -\frac{z^2}{2\lambda^2}\left(\sum_{i=1}^{n} p_i^2\right). \tag{8}$$

This approach to large deviations is different from that in Stein's heuristic treatment (1986, Chapter 5). The above result (6), (7) and (8) are improvements of those in Chen (1975c) in the Poisson case, and (6) an improvement of the corresponding ones in Barbour (1987). By using the method (Section 6) in Arratia, Goldstein and Gordon (1989), extension of (6), (7) and (8) to the multivariate case seems straightforward. But it seems less so to extend these results to the dependent case.

COMPOUND POISSON APPROXIMATION

My final point for discussion concerns an approach to compound Poisson approximation that is different from that of Arratia, Goldstein and Gordon. I would like to discuss compound Poisson approximation using Stein's method directly. This is a natural extension of the Poisson approximation, and in this approach we consider the solution of the integral equation

$$\begin{aligned}\lambda \int tf(w + t)\, d\mu(t) - wf(w) \\ = h(w) - Eh(Z),\end{aligned} \tag{9}$$

where μ is a probability measure with no atom at 0, h is a bounded function and Z has the compound Poisson distribution $e^{\lambda(\mu - \delta_0)}$. In the case μ is the Dirac measure at 1, (9) reduces to (5). An advantage of this approach is that it does not apply only to indicators. Unfortunately, the integral equation (9) is difficult to solve in general, and even if a solution is obtained it is difficult to obtain an effective bound on it. However, I believe that this approach will produce the best results when effective bounds are obtained on the solution of (9).

In Barbour, Chen and Loh (1990), Stein's method is applied to obtain the following result: let $\{X_\alpha : \alpha \in I\}$ be nonnegative random variables. Suppose for each $\alpha \in I$, there exist $A_\alpha \subset B_\alpha \subset I$ with $\alpha \in A_\alpha$ such that X_α is independent of $\{X_\beta : \beta \in A_\alpha^c\}$ and $\{X_\beta : \beta \in A_\alpha\}$ is independent of $\{X_\beta : \beta \in B_\alpha^c\}$. Let

$$W = \sum_{\alpha \in I} X_\alpha,$$

$$Y_\alpha = \sum_{\beta \in A_\alpha} X_\beta,$$

$$\lambda = \sum_{\alpha \in I} EX_\alpha Y_\alpha^{-1},$$

$$p_\alpha = P(X_\alpha > 0)$$

and

$$\xi_\alpha = P\left(\sum_{\beta \in B_\alpha} X_\beta > 0\right).$$

Define the probability measure μ on the Borel subsets of $(0, \infty)$ by

$$\mu(A) = \lambda^{-1} \sum_{\alpha \in I} EX_\alpha Y_\alpha^{-1} I(Y_\alpha \in A).$$

Here we adopt the convention that $0/0 = 0$. Let Z be distributed as $e^{\lambda(\mu - \delta_0)}$. Then

$$\begin{aligned}\|\mathscr{L}(W) - \mathscr{L}(Z)\| &\le 4e^{\lambda} \sum_{\alpha \in I} p_\alpha \xi_\alpha \\ &\le 4e^{\lambda} \sum_{\alpha \in I} \sum_{\beta \in B_\alpha} p_\alpha p_\beta.\end{aligned} \tag{10}$$

In the case the X_α's are indicators, the factor e^{λ} in the bounds in (10) can be improved to $(1 \wedge \lambda_1^{-1})e^{\lambda}$, where $\lambda_i = i^{-1} \sum_{\alpha \in I} EX_\alpha I(Y_\alpha = i)$. If, in addition, $i\lambda_i \downarrow 0$ as $i \to \infty$, then e^{λ} can be replaced by

$$\left(1 \wedge \frac{1}{\lambda_1 - 2\lambda_2}\right)\left[\frac{1}{4(\lambda_1 - 2\lambda_2)} + \log^+ 2(\lambda_1 - 2\lambda_2)\right].$$

The appearance of the factor e^{λ} is due to our inability to bound the solution of (9) effectively.

Suppose X_α, $\alpha \in I$, are indicators satisfying the above dependence assumption. Then by Theorem 1 of Arratia, Goldstein and Gordon, the error bound for the Poisson approximation is

$$2(1 \wedge \lambda^{-1})\left\{\operatorname{Var}(W) - \lambda + 2 \sum_{\alpha \in I} \sum_{\beta \in A_\alpha} p_\alpha p_\beta\right\}.$$

This together with (10) imply that, in the case $(1 \wedge \lambda^{-1})(\operatorname{Var}(W) - \lambda)$ is large, for which Poisson approximation fails, we still have an approximation—compound Poisson approximation, provided $(1 \wedge \lambda_1^{-1})e^{\lambda} \sum_{\alpha \in I} p_\alpha \xi_\alpha$ is small. A consequence of this is that we can avoid declumping in applications, as we shall see below.

Consider the example in Section 4.2.1 of Arratia, Goldstein and Gordon with $X_\alpha = C_\alpha C_{\alpha+1} \cdots C_{\alpha+t-1}$. Then the above dependence assumption is satisfied with $A_\alpha = \{1 \vee (\alpha - t + 1), \cdots, \alpha + 2t - 2\}$ and $B_\alpha = \{1 \vee (\alpha - 2t + 2), \cdots, \alpha + 3t - 3\}$. By (10), we obtain

$$\begin{aligned} &\| \mathscr{L}(U) - \mathscr{L}(Z) \| \\ &\quad \le 4e^{\lambda}(1 \wedge ((p + nq)qp^t)^{-1})n(5t - 4)p^{2t}, \end{aligned}$$

where $q = 1 - p$. Note that $W = U$. In order that the distribution of Z be determined, we need to compute λ_i for all i. It can be shown that

$$\lambda_i = i^{-1} \sum_{\alpha=1}^{n} p^t P(V_{t-1} + V'_{\alpha \wedge t-1} = i - 1)$$

where $V_0 \equiv 0$, V_m and V'_m are geometric (p) truncated at m, and V_{t-1} and $V'_{\alpha \wedge t-1}$ are independent. We can either proceed to compute each λ_i explicitly to determine $\mathscr{L}(Z)$ or approximate $\mathscr{L}(Z)$ by $\mathscr{L}(Z^*)$ to obtain the following result:

$$\begin{aligned} &\| \mathscr{L}(U) - \mathscr{L}(Z^*) \| \\ &\quad \le 4e^{\lambda}(1 \wedge ((p + nq)qp^t)^{-1})n(5t - 4)p^{2t} \\ &\qquad + 4(2n - t)p^{2t} + 4q^{-1}p^{t+1}, \end{aligned} \tag{11}$$

where Z^* has the compound Poisson distribution $\exp[\lambda^*(\mu^* - \delta_0)]$ with $\lambda^* = nqp^t$ and $\mu^*(\{i\}) = qp^{i-1}$, $i = 1, 2, \cdots$ ("one plus a geometric (p)"). In approximating $\mathscr{L}(Z)$ by $\mathscr{L}(Z^*)$ we need not calculate λ_i explicitly. For bounded λ^*, the order of the error bound in (11) is the same as that obtained by Arratia, Goldstein and Gordon. Note that $q\lambda^* \le \lambda_1 \le \lambda \le np^t = q^{-1}\lambda^*$. Hence the result (11) not only provides an approximation for $\mathscr{L}(U)$ but also can be used to obtain the asymptotic distribution of R_n, the length of the longest run of heads beginning in the first n tosses of a coin, since $\{R_n < t\} = \{U = 0\}$ and $P(Z^* = 0) = e^{-\lambda^*}$.

In the same way, (11) can also be applied to the biological example in Section 5 of the article by Arratia, Goldstein and Gordon to obtain an approximation result for $\mathscr{L}(\sum_{\alpha \in I} I(S_\alpha \ge s))$ and the asymptotic distribution of $M_n(t_n)$, the largest number of matches witnessed by any comparison of length t_n substrings of two strands of DNA.

Rejoinder

Richard Arratia, Larry Goldstein and Louis Gordon

At least one of us used to speak of the methods we have presented here as the philosopher's stone. None of us make such extravagant claims any longer; the discussants have put their collective fingers on a number of reasons why.

The method as we have presented it works best for dealing with local dependence, corresponding to situations in which b_1 is small and $b_3 = 0$. In these situations, b_2 is small and our approximations are useful if and only if second moments are well behaved. Steele gives an intriguing example having weak long-range dependence that is much harder to deal with. In Steele's problem, even if second moments were well controlled, there would still be difficulties due to the nonlocal dependence captured by b_3. Here is another such related example.

The question is inspired by the important problem of analyzing the expected, as opposed to worst-case, behavior of the simplex method. See Borgwardt (1987) for an exposition. Specifically, one is led to study the number of edges or vertices in the convex hull of n independent and identically distributed points in, say, $\mathbf{R}^2$. For a line segment joining two of the observed points to be an edge of the convex hull, all of the other points must lie on one of the half-planes determined by these points. The usual heuristic applies. There are a large number of pairs of points to serve as a potential edge, and the probability that a given pair is actually an edge in the convex hull is small. Hence, the total number of edges in the convex hull should be approximately Poisson. As with Steele's example, first moments are tractable. Unfortunately, second moments and nonlocal dependence are again a problem. If an edge is indeed in the convex hull, one of its endpoints is also on a second edge of the convex hull, this is reflected in the second moment and b_2. There is also some additional nonlocal dependençe which is part of b_3. This type of behavior reinforcess the issues raised by Steele's example.

In discussing Section 3.1, Barbour gives an example involving a Bernoulli variable with $p_{1,n} = 1/2$. This example shows that no negative power of λ can be