simple as possible, we further assume that the X_i have a density, and we take $S_0 = 0$ to have a clear convention. The convex hull H_n of $S_0, S_1, \cdots, S_n$ is a natural object that turns out to be intriguing both for its probability theory and geometry. A priori one might expect results on H_n to be difficult and incomplete, but—at least as far as first moments go—the theory is surprisingly easy and precise.

The first contributions to the theory of H_n are due to Spitzer and Widom (1961). Their seminal observation was to show that an ancient result of Cauchy could be combined with a purely combinatorial result of Kac (1954) to obtain an exact formula for the expectation of the length L_n of the boundary of H_n:

$$EL_n = 2 \sum_{k=1}^{n} E|S_k|/k.$$

From this beautiful formula one can obtain considerable information about EL_n, and, in particular, one can use it to show that $EL_n \sim c\sqrt{n}$ provided $EX_i^2 < \infty$ and $EX_i = 0$. This observation does not yet put us in the territory of the Poisson law—that comes later—but it does give the first suggestion of a counting law to be discovered.

The second paper to treat the geometry of H_n is due to Baxter (1961). This work shows, among other results, that the number N_n of sides of H_n has a remarkably simple expectation. In fact, it is just twice the nth harmonic number, i.e.,

$$EN_n = 2 \sum_{k=1}^{n} \frac{1}{k}.$$

In Snyder and Steele (1990), a common generalization of these results is given. If we let e_i, $i = 1, 2, \cdots$ denote the lengths of the faces of H_n and if f is any function, then provided both sides make sense we have the identity

$$E \sum_i f(e_i) = 2 \sum_{k=1}^{n} \frac{1}{k} Ef(|S_k|).$$

Naturally, this identity yields that of Spitzer and Widom by taking $f(x) = x$, and we can also get Baxter's identity just by taking $f(x) = 1$.

Now, here is where it may pay to start looking for a Poisson law. If we let f_n denote the indicator of an interval $[a_n, b_n]$, then for any given distribution of the X_i it is not hard to determine a_n and b_n so that for each n the sum $G_n = \sum_i f(e_i)$ satisfies $EG_n = \lambda > 0$. It may be most natural to take $a_n = 0$ in order to focus on the small faces of H_n. The variable G_n is nothing more than a sum of a random number of dependent random variables, the $f_n(e_i)$. Further, these variables do not seem all *that* dependent. Thus, there is a serious possibility of a Poisson approximation to the distribution of G_n.

Still, in this problem the Poisson law seems a long way away. The first moments were obtained through somewhat slippery trickery, and second moments do not seem to be open to more of the same. The Poisson law is honestly in play, yet the Chen–Stein method has far to come to meet the challenge. Can sufficient information be found on the second moments of G_n to complete the Chen–Stein program?

Comment

A. D. Barbour

The Chen–Stein method has added a new dimension to the techniques available for justifying Poisson approximations. In fields such a random graph theory (Bollobás, 1985, Chapter 4), extreme value theory (Smith, 1988; Holst and Janson, 1990) and spatial statistics (Barbour and Eagleson, 1984), where Poisson approximation plays an important role, the Chen–Stein method has already proved to be the best general approach, and its potential has by no means been exhausted. Its strengths are that it makes many sorts of weak dependence easy to handle, it gives explicit estimates of the accuracy of approximation, and it continues to give good results even when the expectation λ is large. The preceding survey illustrates the first two of these aspects admirably, but it gives rather less weight to the third, to which the following comments are addressed. For details and much more about the Chen–Stein method, see the forthcoming book of Barbour, Holst and Janson (1991).

A remarkable feature of the Chen–Stein method is the form of the estimate of Theorem 1. When applied in the simplest setting, that of Theorem 0, it gives an error estimate no greater than $2 \min(1, \lambda^{-1}) \sum_{i=1}^{n} p_{i,n}^2$. Were only an estimate of the form $c \sum_{i=1}^{n} p_{i,n}^2$ required, for some real c, it could be obtained

A. D. Barbour is Professor of Mathematics, Angewandte Mathematik, Universität Zurich, Rämistrasse 74, Zürich CH-8001, Switzerland.

by matching each $X_{i,n}$ separately to a $\mathcal{P}(p_{i,n})$ random variable, as in Serfling's (1975) method. The presence of the factor λ^{-1} shows that something subtler is going on, and the difference, for large λ, is striking: an estimate of order np^2, useful for $p = o(n^{-1/2})$, is replaced by one of order p, useful throughout the range $p = o(1)$. Unfortunately, in Theorem 1, the coefficient of b_3 is (necessarily) only of order $\lambda^{-1/2}$ for λ large: an advantage of the coupling approach, referred to in Section 4.5, is that the analog of Theorem 1 has the magic factor λ^{-1} throughout.

The next point concerns the philosophy of Section 3.1, and the marked point process approximation, which is proposed as a way of reaching an approximation to a compound Poisson distribution. If λ is moderate, the procedure may well be reasonable, but it is in principle inefficient as soon as λ becomes large. Take, for example, the setting of Theorem 0 again, and suppose that $p_{1,n} = \frac{1}{2}$. A process approximation of the kind suggested cannot be good, because of the problem at $j = 1$: yet, for $\lambda = 1 + \sum_{j\geq 2} p_{j,n}$ large and $\max_{j\geq 2} p_{j,n}$ small, $W = \sum_{j=1}^{n} X_{j,n}$ is close to having a Poisson distribution.

The essential drawback is that Theorem 2 for process approximation suffers in comparison with Theorem 1 from having no factor involving a negative power of λ in the estimates. The example just given already shows that this would be impossible. On the other hand, suppose that a type is independently assigned from the uniform distribution on [0, 1] to each i with $X_{n,i} = 1$ and that only the resulting point process of types on [0, 1] is of interest, so that the information about which of the original indices i gave rise to points is not involved. Then the distance from the homogeneous Poisson process with rate λ on [0, 1] is the same as that of W from Poisson λ, as observed by Michel (1988), and therefore incorporates the factor λ^{-1} of Theorem 1. This indicates that, in exchange for some loss of information, better process estimates might be achieved, and this turns out to be indeed the case. Note in particular that, for compound Poisson approximation, information about which indices contribute to the sum is irrelevant, so that better results can be expected than those derived from Theorem 2.

As an illustration of the above remarks, take the joint distribution of the numbers of short cycles, Example 4.6. The framework of 4.6.2 is well suited to the coupling approach, and the coupling analog of Theorem 2 gives an estimate of at most $\sum_{k=1}^{f(n)} 2k^{-1}d_k$ for

$$\Delta = \|(W_1, \cdots, W_{f(n)}) - (Z_1, \cdots, Z_{f(n)})\|,$$

where d_k is the expected distance in the Manhattan metric between a realization of $(W_1, \cdots, W_{f(n)})$, with the conditional distribution given $X_{1k} = 1$ but not counting this k-cycle, and a coupled realization of $(W_1, \cdots, W_{f(n)})$ with its unconditional distribution. To obtain a suitable coupling, observe that, starting with element 1, the integers 1 to n can be ordered using π by next listing the remaining elements of the cycle containing 1 in the order of the cycle, then choosing a new element at random from those still remaining and listing its cycle in order, and so on. If the rth element in the ordering is the last in a cycle, set $Y_{n+1-r} = 1$, and set $Y_{n+1-r} = 0$ otherwise. Then the cycle type of π is determined by knowledge of the Y_j's. However, it is easy to see that $\{Y_j, 1 \leq j \leq n\}$ have the joint distribution of the positions of the records in n iid trials and are hence *independent*, with $P[Y_j = 1] = 1/j$. So couple by first realizing the Y_j's and then, to get the correct cycle type for the conditional distribution, just define

$$Y'_{n+1-j} = 0, \quad 1 \leq j < k;$$

$$Y'_{n+1-k} = 1;$$

$$Y'_{n+1-j} = Y_{n+1-j}, \quad k < j \leq n.$$

It is immediate from this coupling that

$$d_k \leq \sum_{j=1}^{k} EY_{n+1-j} + 2\sum_{j=k+1}^{k+f(n)} EY_{n+1-j}$$

$$\leq 3f(n)\{n - 2f(n)\}^{-1},$$

giving an upper estimate for Δ of about $6 \log f(n) \cdot f(n)/n$.

In this estimate, no advantage has yet been taken of the size of λ. However, since $\lambda \sim \log f(n)$, an extra factor of λ^{-1} would improve it to the desired order of $f(n)/n$. Now the information as to which indices mark the short cycles is not involved in the joint distribution of $(W_1, \cdots, W_{f(n)})$, and it turns out that a process approximation theorem incorporating a magic λ-factor can indeed be brought to bear. Unfortunately, the best general analog of the factor λ^{-1} that has yet been obtained for processes is of order $\lambda^{-1}(1 + \log_+\lambda)$, which improves the error estimate here to order $\log\log f(n) \cdot f(n)/n$ but still does not quite achieve $f(n)/n$. Whether the magic factor in such process estimates can be improved to order λ^{-1} is a tantalizing open problem.

The error estimate of order $f(n)/n$ can be attained here by observing that short cycles arise in the main only from the early Y_j's. More precisely, construct the cycle types of permutations of $1, \cdots, n$ and $1, \cdots, M$ simultaneously $(M > n)$ from the same sequence $\{Y_j, j \geq 1\}$. The short cycles in the two permutations

differ only on the set

$$\left(\bigcup_{j=n-f(n)+1}^{n}\{Y_j=1\}\right)\cup\left(\bigcup_{j=n+1}^{M}B_j\right),$$

where

$$B_j=\{Y_j=1\}\cap\left(\bigcup_{r=j+1}^{j+f(n)}\{Y_r=1\}\right),$$

whose probability is no greater than

$$\frac{f(n)}{n}+\sum_{j=n+1}^{M}\frac{1}{j}\cdot\frac{f(n)}{j+f(n)}\le 2\frac{f(n)}{n}.$$

However, by the argument above (and probably by that of Section 4.6.1), the error in approximating the joint distribution of the short cycles in the permutation of $1,\cdots,M$ by the joint Poisson distribution of the Z's is of order M^{-1} for fixed n, and can be made arbitrarily small by choice of M. Hence $\Delta\le 4f(n)/n$.

Comment

Michael S. Waterman

The authors of this article mention that their interest in Poisson approximation was motivated by questions in sequence matching. Sequence matching refers to the comparison of two or more sequences to locate regions that are exceptionally similar. While these questions are of interest in computer science, my own motivation to study sequence matching has been molecular biology. Section 5 of the paper is devoted to a biological example. I will take this opportunity to expand upon some statistical questions of interest to molecular biology.

Biology is embarked on one of the most exciting scientific endeavors of the century. The widely publicized Human Genome Initiative (*Understanding Our Genetic Inheritance. The U.S. Human Genome Project: The First Five Years, FY 1991–1995*, April 1990; DOE/ER-045 2P) has as its goal the analysis of the structure of human DNA and the determination of the location of the estimated 100,000 genes. Other model organisms are included in the Initiative to provide the comparative information necessary for understanding the human and other genomes. Medical doctors and legislators may choose to focus on the understanding and possible consequent treatment of more than 4000 human genetic diseases. Some may well view the project as providing the initial data for a fundamental understanding of the processes of life. In any case, the rate at which information is being gathered is astonishing. International DNA databases began to be formed in 1982. The databases are DDBJ (Japan), EMBL (Europe) and GenBank (US). By 1986, Release #42.0 of GenBank had 6.7×10^6 nucleotides (bases) of sequence data. Release #62.0 cited by the authors had 37.2×10^6 nucleotides, while the most recent release #65.0 in 1990 has 49.2×10^6 nucleotides. New technology promises to accelerate the rate of sequence determination. Molecular biology has been an experimental and empirical science. The flow of sequence information is changing the character of the subject.

Our interest in Poisson approximation began with an early analysis of the DNA database. DNA sequences average 1000 nucleotides in length and have a four letter alphabet adenine (A), guanine (G), cytosine (C) and thymine (T). In 1981, Temple Smith and I devised a method or algorithm for finding similar regions of sequences. Briefly, this method optimized a score for all segments I of sequence $\mathbf{x}=x_1x_2\cdots x_n$ and all segments J of sequence $\mathbf{y}=y_1y_2\cdots y_n$. The score, in its simplest form, counts +1 for a match or identical letter from I and J, counts $-\mu$ for a mismatch or nonidentity and counts $-\delta$ for a letter inserted or deleted from a sequence (an indel). For example AAGTC and AGCC can be arranged or aligned as

```
AAGTC
A–GCC
```

to receive score $S=3-\mu-\delta$. They can also be aligned as

```
AAGTC
AGCC–
```

to receive score $S=1-3\mu-\delta$. The algorithm, based on dynamic programming, provides a straightforward

Michael S. Waterman is Professor of Mathematics and of Molecular Biology, University of Southern California, Los Angeles, California 90089-1113.