Extending the Poisson Approximation

Louis H. Y. Chen

In his book, *Recherches sur la Probabilité des Jugements en Matière Criminelle et en Matière Civile*, published in 1837, the French mathematician Siméon-Denis Poisson (1781–1840) proved the following limit theorem: Consider n independent events, each of which occurs with probability p. If p decreases to zero as n increases to infinity in such a way that np approaches a fixed positive number λ, then for any nonnegative integer k, the probability that k events will occur approaches the number $e^{-\lambda}\lambda^k/k!$. The limiting distribution, which is given by $P(X = k) = e^{-\lambda}\lambda^k/k!$, where $k = 0,1,2,\ldots$, is called the Poisson distribution with mean λ. In the 150 years since Poisson's work, this distribution has been applied in an enormous range of applications in both the physical and the life sciences.

Despite its utility, there are applications in which Poisson's approximation for independent events is too constraining. In recent years, a method dating back to Stein (*1*) and Chen (*2*) has been developed for the purpose of generalizing the Poisson limit theorem to dependent events under very general conditions. The method also provides a means for bounding the discrepancy between the distribution of the number of occurrences of the events and the Poisson distribution. This generalization toward dependence has proved to be very fruitful as a wide range of important and interesting problems may be phrased in terms of occurrences of possibly dependent events. These problems arise from such fields as spatial statistics, combinatorial probability, random graphs, extreme value theory, and molecular biology.

Bounding the discrepancy between two distributions is stronger than proving a limit theorem. Not only can a limit theorem be deduced this way, but the bound also provides an estimate of the error in approximating one distribution by the other. In the context of Poisson approximation, the discrepancy is given by

$$\sum_{k=0}^{\infty} |b(k) - e^{-\lambda}\lambda^k/k!|$$

where $b(k)$ is the probability of occurrence of k events. This discrepancy is twice the maximum possible error in the approximation and is called the total variation distance.

The method of Poisson approximation involves the solution of a difference equation and works well for dependent events. It is easy to apply and the bounds obtained depend only on the first and second moments.

There are two ways to obtain the bounds: The local approach, which was first used by Chen (*2*) and which is very similar in spirit to the method of normal approximation of Stein (*1*), and the coupling approach of Barbour, Holst, and Janson (*3*). In the local approach, it is assumed that each event may be dependent on a few other events but is independent or almost independent of all the others. The most general upper bound on the total variation distance that has been obtained by this approach is due to Arratia, Goldstein, and Gordon (*4*, *5*).

The coupling approach deals with dependence that is symmetric. To understand this kind of dependence, consider the classical occupancy problem in which r balls are thrown into n boxes such that each ball falls into the boxes with respective probabilities $p_1,\ldots,p_n$. Associate with each box an event. If the ith box is empty, we say the ith event occurs. The number of events that occur is then the number of empty boxes. The relations between the events are symmetric in that the nature of dependence between every two events is the same. The symmetry is even more apparent if we assume that all the p_i's are equal. The coupling approach was systematically developed by Barbour, Holst, and Janson (*3*), where general upper bounds on the total variation distance are also obtained by this approach. The works of both groups (*3*–*5*) contain many vivid examples of application of this method of Poisson approximation in a wide range of fields. I give two examples for illustration.

Siméon-Denis Poisson, French mathematician. [Bettmann Photos]

The first example concerns random graphs. A graph is a mathematical object, represented by a set of vertices (nodes), some or all of which are joined by lines called edges. The theory of random graphs was founded by Erdös and Rényi over 30 years ago. It is the study of graphs by probabilistic methods. The initial objective was to prove the existence of graphs with certain properties. Many results have now been found to have applications in computer algorithms. There is also a connection between random graphs and percolation theory, a mathematical theory of disordered media. The method of Poisson approximation has become an important tool in random graphs. A typical application concerns counting the number of particular configurations of vertices and edges in a random graph.

As an example, consider a complete graph K_n with n vertices, that is, a graph in which every two vertices are joined by an edge. If we delete the edges such that each edge has a probability $1 - p$ of removal (independently of the other edges), we then get a random graph $K_{n,p}$. Now, we want to know how many complete graphs with r vertices there are in $K_{n,p}$, where r is a fixed integer less than n. Associate an event with each complete graph with r vertices in K_n. If the complete graph is in $K_{n,p}$, then we say the event occurs. The number of events that occur is then the number of complete graphs with r vertices that are in $K_{n,p}$. By the coupling approach, it can be proved that the number of complete graphs with r vertices that lie in $K_{n,p}$ has approximately the Poisson distribution with mean

$$\lambda = \binom{n}{r} p^{\binom{r}{2}}$$

provided that $pn^{2/(r-1)}$ is large and $pn^{2/(r+1)}$ is small. The error in this case is at most a fixed constant multiple of

$$n^{r-2}p^{\binom{r}{2}-1}$$

which is small. A variety of examples of application of the method of Poisson approximation to random graphs are presented in Barbour, Holst, and Janson (*3*).

The second example concerns DNA sequence matching. A strand of DNA can be represented as a long string of letters from the alphabet {A,C,G,T}. When two DNA sequences show strong similarity in a region, this may have biological significance. It is therefore relevant to determine whether the similarity could be attributable to chance alone. Smith and Waterman (*6*) proposed the following method of scoring for comparing two DNA sequences. For each pair of segments I and J taken from two given sequences $x = x_1, x_2,\ldots, x_m$, and $y = y_1, y_2,\ldots, y_n$, the letters in the two seg

The author is in the Department of Mathematics, National University of Singapore, Lower Kent Ridge Road, Singapore 0511.

ments are aligned in all possible ways. For each alignment, a score is obtained by counting +1 for a match, $-\mu$ for a mismatch, and $-\delta$ for a letter inserted or deleted (a gap).

For example, AGCACT and AGGT can be aligned as

AGCACT
AG–G–T

to receive score $S = 3 - \mu - 2\delta$. They can also be aligned as

AGCACT
AGGT– –

to receive score $S = 2 - 2\mu - 2\delta$. The score $M_{m,n}(x,y)$, which is defined to be the maximum of all the scores obtained for all possible pairs of segments I and J, is then calculated by an algorithm whose computing time is proportional to the product of m and n.

To calculate the probability of those large values of $M_{m,n}(x,y)$ for which the similarity is significant, one has to know, at least approximately, the distribution of $M_{m,n}(x,y)$ under the assumption that x and y are unrelated. That is, the letters $x_1,\ldots, x_m$, $y_1,\ldots, y_n$ are independently chosen with the same distribution from the alphabet {A,C,G,T}. Karlin and Altschul (*7*) obtained approximations for the probabilities of large values of $M_{m,n}(x,y)$ for the case $\delta = \infty$ (that is, without gaps), assuming the expected score of two letters to be negative. Arratia, Gordon, and Waterman (*8*) considered the score $M_{m,n}(t)$, which is the maximum of the scores obtained by considering only those pairs of segments I and J of a given length t, for the case $\mu = 0$. They established approximations for $M_{m,n}(t)$ under certain mild conditions by the method of Poisson approximation.

The set of all values of the parameters (μ,δ) can be divided into two regions, S_1 and S_2, such that for $m = n$, the growth of $M_{n,n}(x,y)$ is proportional to n in S_1 and the growth of $M_{n,n}(x,y)$ is proportional to $\log n$ in S_2. The cases considered by Karlin and Altschul (*7*) and Arratia, Gordon, and Waterman (*8*) are in the logarithmic region. The work of the latter (*8*) has provided a basis for Waterman and Vingron (*9*) to use the Poisson clumping heuristic of Aldous (*10*) to calculate the probabilities of large values of $M_{m,n}(x,y)$ in the entire logarithmic region.

Let us see how the method of Poisson approximation is applied in the problem of Arratia, Gordon, and Waterman (*8*). Because $\mu = 0$, the score for each pair of the segments I and J of a given length t is just the number of matches. Let s be a given positive integer. Associate an event with each pair of I and J. If the score is at least s for a particular pair of I and J, we say that the associated event occurs. The number of events that occur, say, U is the number of those scores which are at least s. Therefore, $P[M_{m,n}(t) < s] = P(U = 0)$.

We would have been done if the distribution of U was approximately Poisson with mean, say, λ^*. For then, we would have had $P[M_{m,n}(t) \geq s] = 1 - P(U = 0) \simeq 1 - e^{-\lambda^*}$. However, this is not the case. The events associated with the I's and the J's occur in clumps. By the Poisson clumping heuristic of Aldous (*10*), it is the number of clumps that is expected to have approximately the Poisson distribution. So we declump and modify the events so as to obtain events which are associated with the clumps. Let W denote the number of clumps that occur. The method of Poisson approximation is then applied with the local approach. The result is that $P[M_{m,n}(t) \geq s] \simeq 1 - P(W = 0) \simeq 1 - e^{-\lambda}$, where λ is the mean of the approximating Poisson distribution.

There are many situations in which occurrences of events happen in clumps. The book by Aldous (*10*) provides many such examples. In these situations, the appropriate approximating distribution is the compound Poisson distribution. One of the new developments relating to the method of Poisson approximation is the extension of the method to compound Poisson approximation by Barbour, Chen, and Loh (*11*). This work extended the associated difference equation to an integral equation. Although Arratia, Goldstein, and Gordon (*5*) also considered compound Poisson approximation, the approach of Barbour, Chen, and Loh (*11*) is different and holds promise for producing better results. Much work is also being done on multivariate or process approximation, which was initiated by Barbour (*12*) and Arratia, Goldstein, and Gordon (*4*) using different approaches. Finally, for approximation for relative errors, which is very useful when the probabilities are small, a new approach was introduced in Chen and Choi (*13*).

References and Notes

1. C. M. Stein, in *Proceedings of the Sixth Berkeley Symposium on Mathematical Statistics, and Probability* (Univ. of California Press, Berkeley, 1972), vol. 2, pp. 583–602.
2. L. H. Y. Chen, *Ann. Probab.* **3**, 534 (1975).
3. A. D. Barbour, L. Holst, S. Janson, *Poisson Approximation* (Clarendon, Oxford, 1992).
4. R. Arratia, L. Goldstein, L. Gordon, *Ann. Probab.* **17**, 9 (1989).
5. ———, *Stat. Sci.* **3**, 403 (1990).
6. T. F. Smith and M. S. Waterman, *J. Mol. Biol.* **147**, 195 (1981).
7. S. Karlin and S. F. Altschul, *Proc. Natl. Acad. Sci. U.S.A.* **87**, 2264 (1990).
8. R. Arratia, L. Gordon, M. S. Waterman, *Ann. Stat.* **18**, 539 (1990).
9. M. S. Waterman and M. Vingron, in preparation.
10. D. Aldous, *Probability Approximations via the Poisson Clumping Heuristic*, vol. 77 of *Applied Mathematical Sciences* (Springer, New York, 1989)
11. A D. Barbour, L. H. Y. Chen, W.-L. Loh, *Ann. Probab.* **20**, 1843 (1992).
12. A. D. Barbour, *J. Appl. Probab.* **25(a)**, 175 (1988).
13. L. H. Y. Chen and K. P. Choi, *Ann. Probab.* **20**, 1867 (1992).

14. I thank M. Waterman for helpful discussions on the molecular biology application and Z. Chen, C. T. Chong, and Y. K. Leong for their helpful comments on the preliminary drafts of this Perspective.

Conformational Flexibility of Enzyme Active Sites

Chen-Lu Tsou

The activity of enzymes is strongly dependent on their conformational integrity. Our laboratory has been interested in the precise relationship between enzyme activity changes and protein unfolding. The observation that, under denaturing conditions, loss of enzyme activity can precede marked changes in protein conformation led us to hypothesize that enzyme active sites may display more conformational flexibility than the enzyme molecules as a whole (*1*, *2*). Here I discuss recent results that support this concept.

The author is at the National Laboratory of Biomacromolecules, Institute of Biophysics, Academia Sinica, 15 Datun Road, Beijing, 100101, China.

Exposure of the enzyme creatine kinase to denaturants such as guanidine hydrochloride (GuHCl) and urea results in an initial phase of rapid inactivation; this inactivation can be conveniently measured by following the substrate reaction with a stopped-flow apparatus (*3*). In parallel, conformational changes induced by the denaturants can be monitored by conventional methods that detect changes in intrinsic fluorescence, absorbance in the ultraviolet, circular dichroism, or exposure of buried thiol groups. Comparison of conformation and activity changes of creatine kinase during denaturation indicates that enzyme inactivation occurs at a much lower concentration of denaturant than is required to