POISSON CLUMPING HEURISTIC

INTRODUCTION

The ideas of Normal approximation and Poisson approximation, and their formalization (for independent random variables, and for many types of dependent variables) via central limit theorems and laws of small numbers (see ASYMPTOTIC NORMALITY; LIMIT THEOREMS, CENTRAL; POISSON DISTRIBUTION) are one of the oldest ideas in mathematical probability and statistics. While the utility of Normal approximation is universally recognized, the scope of applicability of Poisson and compound Poisson approximation has historically been less appreciated. Perhaps this is because in the independent setting the basic Poisson limit theorem is comparatively elementary, and the result lacks the breadth of statistical application of the Normal result. Aldous [1] describes about 100 problems, covering a wide range of probability models, where compound Poisson approximations are implicitly relevant, and devises a (necessarily rather vague) language in which to present calculations yielding explicit heuristic approximations. Three simple prototypical examples are indicated below.

A restricted version of the *Poisson clumping heuristic* is provided by equation (1) below. Consider a random process of events $(A_i; i \in I)$ which is stationary, and where the index set I is a bounded subset of the integers, or the real line, or their d-dimensional extensions. Write $m(\cdot)$ for the natural measure (counting measure or Lebesgue measure) on I. Suppose we are interested in the probability

$$q = P(\text{ none of the } A_i\text{'s occur }).$$

By stationarity the chance $p = P(A_i)$ of a particular event does not depend on i. When one event occurs there may in fact, by local dependence, be a "clump" of nearby events occurring. Absent some form of long-range dependence, the process of distinct clumps should resemble a Poisson process of some rate λ, and hence $q = P(\text{ no clump intersects } I) \approx \exp(-\lambda m(I))$. The measure $m(\cdot)$ of a typical clump is a random variable C, the "clump size". By considering the mean rate of events we get the identity $p = \lambda EC$, and hence

$$q \approx \exp(-pm(I)/EC). \tag{1}$$

In typical examples the values of p and $m(I)$ are immediate and the only issue in using (1) is to estimate EC.

EXAMPLES

1. Head-runs in coin tossing. *Toss a fair coin 1000 times. What is the chance q of never seeing a run of 9 of more Heads?*

Here we take I to be the the set of integers $[9, 1000]$ and A_i to be the event "tosses $i-8$ through i are not all Heads". So $p = (1/2)^9$ and $m(I) = 992$. If event A_i occurs, then so do $A_{i+1}, A_{i+2}, \ldots, A_{i+N-1}$ where $i+N$ is the first toss after i which comes up Tails. So the clump size $C = N$ and $EC = EN = 2$, because N has geometric$(1/2)$ distribution. Equation (1) now gives $q \approx 38\%$.

2. First-passage time in the ORNSTEIN-UHLENBECK PROCESS. *Standardize the stationary O-U process as $dX(t) = -X(t)dt + \sqrt{2}dB(t)$, so that the stationary distribution is* Normal*(0,1). What is the first passage time*

$$T = \min\{t \geq 0 : X(t) = 3\}?$$

Here we take I to be a time interval $[0, t_0]$ and A_t to be the event "$X_t \in [3, 3+\delta]$" for some small δ. So q is essentially $P(T_3 > t_0)$. Here $p = \phi(3)\delta$, where $\phi(\cdot)$ is the Normal$(0,1)$ density. Given $X_t = 3$, the process has local drift rate -3 and hence EC, the mean time spent in an interval of length δ, is $\delta/3$. Equation (1) gives $P(T > t_0) \approx \exp(-\phi(3)\delta t_0/(\delta/3))$. In other words, T has approximately the exponential distribution with mean $\frac{1}{3\phi(3)} \approx 75$.

3. Empty squares in random scatter. *Place 2,500 particles uniformly randomly in a 50-by-50 square. What is the chance we can find some 3-by-3 subsquare containing no particles?*

Here we take A_x to be the event "the 3-by-3 subsquare with lower left corner $x = (x_1, x_2)$ contains no particles". So $I = [0, 47] \times [0, 47]$. Because the particles follow approximately a Poisson scatter of rate 1, we have $p \approx e^{-9}$. The estimation of EC in this example is too hard to give here, but it turns out (see [1] Example H4) that $EC \approx 1/9$. Then equation (1) gives $q \approx \exp(-e^{-9}47^2/(1/9)) = 9\%$, and we want the chance $1 - q \approx 91\%$.

SCOPE

Other applications involve Markov chain hitting times, extrema in probabilistic combinatorics, maxima of random fields, increments of Brownian motion and assorted coverage problems. Asymptotic justification of these approximations can be provided in many contexts: in particular for extremes of stationary processes (see EXTREMAL PROCESSES). The Stein-Chen method (see [2] and STEIN'S METHOD) provides explicit bounds in certain combinatorial settings.

REFERENCES

[1] Aldous, D.J. (1989). *Probability Approximations via the Poisson Clumping Heuristic.* Springer-Verlag, New York.

[2] Barbour, A.D., Holst, L. and Janson, S. (1992). *Poisson Approximation.* Oxford University Press.

DAVID ALDOUS