Collision Probability Between Sets of Random Variables

Article *in* Statistics [?] Probability Letters · September 2003

DOI: 10.1016/S0167-7152(03)00168-8 · Source: RePEc

1 author:

Michael Wendl
Washington University in St. Louis
161 PUBLICATIONS **44,243** CITATIONS

Available at WWW.MATHEMATICSWEB.ORG POWERED BY SCIENCE DIRECT®

ELSEVIER

Statistics & Probability Letters 64 (2003) 249–254

STATISTICS & PROBABILITY LETTERS

www.elsevier.com/locate/stapro

Collision probability between sets of random variables

Michael C. Wendl*

Washington University, Box 8501, 4444 Forest Park Blvd., Saint Louis, MO 63108, USA

Received September 2002; received in revised form May 2003

Abstract

We develop the collision probability for a canonical collision problem using a counting procedure based on signed graphs. The result involves Stirling numbers of the second kind and is straightforward to evaluate. Characteristics are discussed in the context of a generalized birthday problem and error of the standard binomial approximation is quantified. The basic solution for two sets is also extended to an arbitrary number of sets.

Keywords: Stirling numbers; Signed graphs

1. Introduction

Investigators have long been interested in stochastic collision processes, many of which can be cast as simple problems having random variations in time or space. For example, consider two airports supporting daily flights between them. If altitudes on the route are assigned to each flight randomly, as is permissible for visual flight rules (VFR) at less than 3000 f, there will be a finite probability of a collision course for simultaneously airborne planes (Patlovany, 1997). Such collision phenomena arise in numerous applications, e.g. celestial objects (Giorgini et al., 2002), transportation (Davis, 1998), etc. In this context, we may also speak equivalently of "matches" between sets of random variables. Notable examples are matches between random length clone fragments in DNA mapping (Soderlund et al., 2000) and dice matches in casino games such as Chuck-A-Luck (Epstein, 1967).

In many cases, the fundamental parameter of interest is an overall probability of success, P_0, i.e. the probability of zero collisions. A wide range of such problems can be cast in the following canonical form. Consider two sets of random variables **A** and **B** comprised of m and n elements, respectively. Each variable in $\{A_1, A_2, \ldots, A_m\}$ and $\{B_1, B_2, \ldots, B_n\}$ can assume any of t discrete values with equally likely probability. A collision occurs when at least one element in **A** is equivalent to at least

* Tel. +1-314-286-1800.
E-mail address: mwendl@watson.wustl.edu (M.C. Wendl).

0167-7152/03/$ - see front matter
doi:10.1016/S0167-7152(03)00168-8

one element in **B**. For example, in the flight problem m and n are the number of planes originating from each of the two airports and t represents the number of discrete altitudes available in the flight envelope.

There is a classical idealization to this problem whereby each potential collision event A_iB_j is assumed independent of all others. This simplification is invoked in many algorithms and calculations, e.g. for DNA clone mapping (Sulston et al., 1988), because brute-force enumeration is impractical. Under this assumption, one can show via elementary methods that $P_0(m,n,t)=(1-1/t)^{mn}$. However, this equation is exact only for the special case where at least one of the sets is limited to a single variable. Trials are not independent for $m \geqslant n > 1$. The present work is motivated by a demand for greater rigor in such calculations. We derive a combinatorially exact expression for $P_0(m,n,t)$ using graph theory to construct a systematic counting procedure for collision events. Error behavior of the binomial approximation is quantified and the result is further generalized to account for an arbitrary number of sets.

2. Probability of success

The following theorem characterizes the collision problem outlined above.

Theorem 1 (Probability of success). *The probability of no collisions between two sets of random variables* $\{A_1,A_2,\ldots,A_m\}$ *and* $\{B_1,B_2,\ldots,B_n\}$, *the elements of which can assume any of* t *discrete values with equally likely probability is*

$$P_0(m,n,t)=\frac{1}{t^{m+n}}\sum_{i=1}^{m}\sum_{j=1}^{n}S_2(m,i)S_2(n,j)\prod_{k=0}^{i+j-1}t-k, \tag{1}$$

where S_2 *represents Stirling numbers of the second kind.*

The proof can be conceptually formulated via graph theory. For example, the members of **A** and **B** may be cast as the disjoint vertices of a bipartite graph. Each vertex represents a specific value k, where $1 \leqslant k \leqslant t$. The fact that we assign value to vertices means that there will actually be many graphs for a given parameter group (m,n,t). The probability of success is simply the number of graphs that exhibit no collisions divided by the total number of possible graphs. It is convenient to employ the concept of signed edges (Read and Wilson, 1998) to examine individual cases. We take a solid edge to indicate that two vertices have the same value and a dotted edge to mean they differ.

Actually, standard bipartite graphs are somewhat impractical for this problem. The resulting counting method is essentially identical to that associated with standard set theory, so the inclusion–exclusion calculations become rather cumbersome. Instead, we concatenate the vertices into linear graphs, where the edge connecting the two sets is always a mismatch edge. This relationship permits us to formulate a simple system that ensures no vertex in **A** will be equal to any vertex in **B**. An example of such a graph for $m=5$ and $n=3$ is shown in Fig. 1. We call this instance a *type*, i.e. a representation of the set of all graphs having a specified arrangement of matching and mismatching edges. We also define a graph *configuration* in which the order of edges is not considered and note that there are multiple graph types per graph configuration.

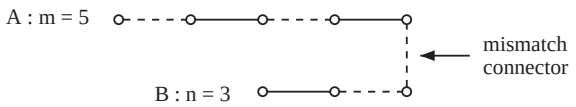

Fig. 1. Example graph type for $m = 5$ and $n = 3$.

Proof of Theorem 1. Let i and j denote the number of uniquely valued vertices in subgraphs representing sets **A** and **B**, respectively. These values are constrained as $1 \leqslant i \leqslant m$ and $1 \leqslant j \leqslant n$. Each subgraph then has $i - 1$ and $j - 1$ mismatch edges. The total number of mismatch edges in the graph, including the connection edge is then $i + j - 1$. Let us pick an arbitrary vertex as a reference point. This vertex is free to take on any of the t admissible values. Any vertex joined to the reference vertex via a solid edge must have the same value. If the edge is dotted, i.e. a mismatch edge, the vertex can only take on one of the remaining $t - 1$ values. The next vertex connected by a dotted edge can have any of $t - 2$ values, etc. Of course, this constraint holds when crossing the connection edge from one subgraph into the other. This process clearly permits us to enumerate the number of successful, i.e. non-colliding graphs for any particular graph type by simply counting the number of mismatching edges. The number of non-colliding graphs is then

$$t \times (t-1) \times \cdots \times (t-i-j+1) = \prod_{k=0}^{i+j-1} t-k. \tag{2}$$

In most cases, there will be more than a single graph type per configuration. For example, the short arm of the graph in Fig. 1 having $n = 3$ with one matching and one mismatching edge can be partitioned in three different ways. Partitions for an arbitrary subgraph are described by Stirling numbers of the second kind (Comtet, 1974), denoted by S_2. Specifically, $S_2(m,i)$ gives the number of ways to partition m vertices among i indistinguishable values, such that each value is represented at least once. Since the two subgraphs are independent of each other, the total number of graph types for each configuration is simply the product of the number of types for each subgraph, i.e. $S_2(m,i)S_2(n,j)$. Multiplying this value by the number of non-colliding graphs per type and then summing over all i and j gives the total number of successful graphs. The success probability is obtained simply by dividing by the total number of possible graphs, t^{m+n}, which follows directly from the observation that each vertex value is independent of all others. □

Because the size of the sample space goes as t^{m+n}, most cases of interest cannot be treated by brute-force counting. For instance, in the DNA clone mapping problem, m and n are of the order of 50 for bacterial artificial chromosome (BAC) clones and $t \approx 236$ for standard measurement resolution (Soderlund et al., 2000). These parameters result in $236^{50+50} \approx 10^{237}$ possible outcomes. Conversely, summation in Theorem 1 goes only as $m \times n$ and Stirling numbers can be calculated using efficient triangular recurrence relationships (Comtet, 1974). Thus, evaluation of P_0 is straightforward. For example, Theorem 1 yields $P_0(50, 50, 236) \approx 5.95 \times 10^{-5}$ in less than 1 s of CPU time on a 400 MHz processor.

Characteristics of P_0 can also be described in the context of a generalized birthday problem. These problems are typically cast as finding the number of people in a party such that there is a specific probability for a common birthday. For example, in the standard version (McKinney, 1966), 23

people are required such that the probability reaches 0.5. Birthdays are assumed to be uniformly distributed over a standard year having $t = 365$ days. Here, we imagine the party divided into two mutually exclusive sub-groups, say m men and n women. The problem is to determine values of m and n that yield a given probability of a shared birthday between at least one man and one woman, i.e. $1 - P_0(m, n, 365)$. Matches exclusively within a sub-group are irrelevant. Unlike the standard case where only the group as a whole is considered, this problem does not have unique solutions for m and n. Although probabilities are symmetric with respect to the two groups, the total number of people required for a specific threshold probability varies considerably. Group size is smallest when the numbers of men and women are exactly, or very nearly equal. For example, $P_0 \approx 0.5$ is obtained for a 32-member party of 16 men and 16 women, as well as a 49-member party of 43 men and 6 women. A first-order estimate can be inferred from the binomial approximation, i.e. $m \times n = C_0$, where C_0 is the constant $[\ln P_0 / \ln(1 - 1/t)]$. Straightforward algebra leads to the following range estimate for the party size: $2\sqrt{C_0} \leqslant m + n \leqslant C_0 + 1$. In this case, the sum $m + n$ will always be greater than its counterpart in the standard birthday problem where all matches are admissible.

3. Error of binomial approximation

As discussed above, binomial approximation has been widely employed for estimating probabilities. Yet, there has been no quantitative assessment of its accuracy. Based on the level of dependence a single event shares with all the others, we would intuitively expect errors to increase with m and n, but to decrease with t. For example, consider a simple game of chance where two players each flip two coins and bet according to the odds of no matches between them. The actual probability is $P_0(2, 2, 2) = 0.125$, but binomial approximation yields 0.0625, a 50% error. If we change the coins to standard six-sided dice, the exact probability is $P_0(2, 2, 6) \approx 0.4861$, while binomial approximation predicts roughly 0.4823, an error of less than 1%.

Although the application spectrum of Theorem 1 covers a broad parameter range, we can outline general error trends for binomial approximation. Fig. 2 shows the special case of $m = n$. The binomial

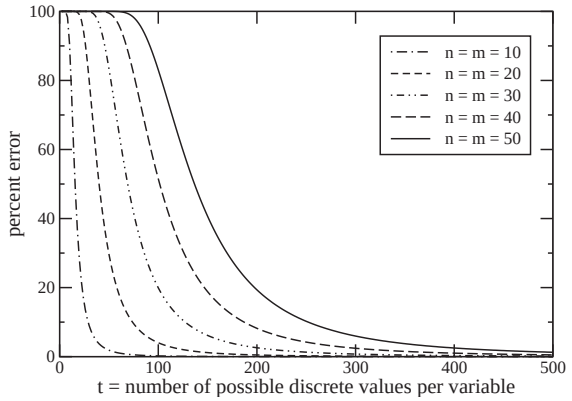

Fig. 2. Under-prediction error of binomial approximation for the special case $m = n$.

model under-predicts true probabilities of collision. Error decreases rapidly as a function of t when the number of variables is small. However, it becomes more persistent as the number of variables rises. Values of (m,n,t) which correspond to the lowest errors are associated with larger values of P_0. In other words, the validity of binomial approximation increases as collision events become less frequent. For cases where collisions are common, the binomial model does not accurately predict the rare non-collision event. Specifically, binomial approximation implies certain collision $P_0 \to 0$, when in fact finite probabilities exist that no collision will occur.

4. A generalization

The canonical problem of two sets can be viewed as a special case of a more general problem having an arbitrary number of sets $\mathbf{A}_1, \mathbf{A}_2, \ldots, \mathbf{A}_n$. Each set has an arbitrary number of random variables $m_1, m_2, \ldots, m_n$, all of which are uniformly distributed over t discrete values. We give the generalization of Theorem 1 as

Theorem 2 (General probability). *The probability of no collisions among an arbitrary number of sets of random variables is*

$$P_0(m_1, m_2, \ldots, m_n, t) = \frac{1}{t^M} \sum_{j_1=1}^{m_1} \sum_{j_2=1}^{m_2} \cdots \sum_{j_n=1}^{m_n} \prod_{i=1}^{n} S_2(m_i, j_i) \prod_{k=0}^{J} t - k, \tag{3}$$

where $M = m_1 + m_2 + \cdots + m_n$ *is the total number of random variables and* $J = j_1 + j_2 + \cdots + j_n - 1$.

Proof. The proof can be framed as a straightforward extension of the previous one using concatenated signed graphs. Specifically, the number of mismatching edges, J, on any graph type is one less than the total number of unique vertices over all n subgraphs. The number of non-colliding graphs per graph type is then the product of $t-k$ over $k = 0, 1, 2, \ldots, J$. The number of graph types per configuration is the product of the number of partitions for each subgraph: $S_2(m_1, j_1) \times S_2(m_2, j_2) \times \cdots \times S_2(m_n, j_n)$. Summing over all $j_1, j_2, \ldots, j_n$ gives the total number of non-colliding graphs. The probability is obtained simply by dividing by the total number of possible graphs, t^M. □

Acknowledgements

The author is indebted to colleagues who discussed and reviewed this material, especially Dr. J.W. Wallis and Professor J.D. McPherson of Washington University.

References

Comtet, L., 1974. Advanced Combinatorics. Reidel Publishing, Dordrecht, Holland.

Davis, G.A., 1998. Method for estimating effect of traffic volume and speed on pedestrian safety for residential streets. Transportation Res. Rec. 1636, 110–115.

Epstein, R.A., 1967. The Theory of Gambling and Statistical Logic. Academic Press, New York, NY.

Giorgini, J.D., Ostro, S.J., Benner, L.A.M., Chodas, P.W., Chesley, S.R., Hudson, R.S., Nolan, M.C., Klemola, A.R., Standish, E.M., Jurgens, R.F., Rose, R., Chamberlain, A.B., Yeomans, D.K., Margot, J.L., 2002. Asteroid 1950 DA's encounter with Earth in 2880: physical limits of collision probability prediction. Science 296, 132–136.

McKinney, E.H., 1966. Generalized birthday problem. Amer. Math. Monthly 73, 385–387.

Patlovany, R.W., 1997. US aviation regulations increase probability of midair collisions. Risk Anal. 17, 237–248.

Read, R.C., Wilson, R.J., 1998. An Atlas of Graphs. Clarendon Press, Oxford, UK.

Soderlund, C., Humphray, S., Dunham, A., French, L., 2000. Contigs built with fingerprints, markers, and FPC V 4.7. Genome Res. 10, 1772–1787.

Sulston, J., Mallett, F., Staden, R., Durbin, R., Horsnell, T., Coulson, A., 1988. Software for genome mapping by fingerprinting techniques. Comput. Appl. Biosci. 4, 125–132.