



Some Properties of Similar Pairs

Gunnar Blom; Lars Holst

Advances in Applied Probability, Vol. 21, No. 4. (Dec., 1989), pp. 941-944.

Stable URL:

<http://links.jstor.org/sici?sici=0001-8678%28198912%2921%3A4%3C941%3ASPOSP%3E2.0.CO%3B2-Q>

Advances in Applied Probability is currently published by Applied Probability Trust.

Your use of the JSTOR archive indicates your acceptance of JSTOR's Terms and Conditions of Use, available at <http://www.jstor.org/about/terms.html>. JSTOR's Terms and Conditions of Use provides, in part, that unless you have obtained prior permission, you may not download an entire issue of a journal or multiple copies of articles, and you may use content in the JSTOR archive only for your personal, non-commercial use.

Please contact the publisher regarding any further use of this work. Publisher contact information may be obtained at <http://www.jstor.org/journals/apt.html>.

Each copy of any part of a JSTOR transmission must contain the same copyright notice that appears on the screen or printed page of such transmission.

JSTOR is an independent not-for-profit organization dedicated to creating and preserving a digital archive of scholarly journals. For more information regarding JSTOR, please contact support@jstor.org.

SOME PROPERTIES OF SIMILAR PAIRS

GUNNAR BLOM,* *University of Lund*

LARS HOLST,** *Royal Institute of Technology, Stockholm*

Abstract

In a given set, the elements are compared pairwise. The number W of similar pairs is studied, that is, the number of pairs with a certain property in common. Under certain conditions, W has, approximately, a Poisson distribution. Examples are considered connected with the birthday problem and with a circle problem involving DNA breakages.

BIRTHDAY PROBLEM; DISSOCIATED STATISTICS; DNA BREAKAGES; MULTINOMIAL DISTRIBUTION; POISSON APPROXIMATION

1. Introduction and summary

Let $\{A_1, A_2, \dots, A_n\}$ be a set of *elements*. The elements A_i and A_j are said to form a *similar pair* if they are related in a given way; more briefly, they are then said to be *similar*. For example, the A 's may be coloured balls which are called similar if they have the same colour.

Introduce the indicator random variables I_{ij} , where $I_{ij} = 1$ if A_i and A_j are similar, and $I_{ij} = 0$ otherwise. We are interested in the total number of similar pairs

$$W = \sum_{i < j} I_{ij}.$$

The sum consists of $M = \binom{n}{2}$ terms.

The elements are assumed to be generated by some chance mechanism. We shall consider the following situation:

- (a) The indicator random variables I_{ij} have common mean p .
- (b) If the random variables I_{ij} and I_{kl} have no indices in common, they are independent.
- (c) If I_{ij} and I_{kl} have exactly one index in common then $\text{Cov}(I_{ij}, I_{kl}) = c$.

Hence the random variables I_{ij} are *dissociated*; cf. Barbour and Eagleson (1984).

Let X be a random variable assuming integer values $0, 1, \dots$. Set

$$\Delta_X = \frac{\text{Var}(X)}{E(X)} - 1.$$

This quantity can be positive, negative or zero. If Δ_X is near zero, X can often be approximated by a Poisson distribution. Such cases will be encountered in this paper.

In Section 2, we derive the mean and variance of W and prove our main theorem which concerns the variational distance between W and a Poisson random variable Z with the same mean as W . In Section 3, examples are given concerning the uniform distribution and the multinomial distribution. As special cases we consider a birthday problem and a problem concerning DNA breakages.

Received 20 March 1989; revision received 30 June 1989.

* Postal address: Department of Mathematical Statistics, University of Lund, Box 118, S-221 00 Lund, Sweden.

** Postal address: Department of Mathematics, Royal Institute of Technology, S-100 44 Stockholm, Sweden.

2. Properties of W

It follows from the assumptions (a)–(c) that the mean and variance of W are given by

$$(1) \quad E(W) = Mp; \quad \text{Var}(W) = Mp(1 - p) + 2M(n - 2)c.$$

Hence we obtain

$$(2) \quad \Delta_w = 2(n - 2)c/p - p.$$

Consider the variational distance

$$d(W, Z) = \sup_B |P(W \in B) - P(Z \in B)|$$

between W and Z , where $Z \sim \text{Poisson}(Mp)$.

Theorem 1. The variational distance satisfies the inequality

$$d(W, Z) < (1 - e^{-E(W)})(\Delta_w + 4np).$$

Proof. Consider a sum W of dissociated Bernoulli random variables I_{ij} , $1 \leq i < j \leq n$, identically distributed or not. Set $p_{ij} = P(I_{ij} = 1)$. According to Theorem 1 in Barbour and Eagleson (1984) and Lemma 4 in Barbour and Eagleson (1983) we have

$$d(W, Z) \leq \frac{1 - e^{-E(W)}}{E(W)} \left[\sum p_{ij}^2 + \sum' p_{ij}p_{kl} + \sum' E(I_{ij}I_{kl}) \right].$$

Here \sum denotes summation over $1 \leq i < j \leq n$ and \sum' summation over all pairs of indices (i, j) , (k, l) with exactly one index in common.

Since

$$E(W) = \sum p_{ij}$$

$$\text{Var}(W) = \sum p_{ij}(1 - p_{ij}) + \sum' E(I_{ij}I_{kl}) - \sum' p_{ij}p_{kl}$$

we can rewrite the right-hand side of the inequality in the form

$$(1 - e^{-E(W)})(\Delta_w + R),$$

where

$$R = 2 \left(\sum p_{ij}^2 + \sum' p_{ij}p_{kl} \right) / E(W).$$

Note that Δ_w may be negative, but $\Delta_w + R$ is positive.

In our case, $\sum p_{ij}^2 = Mp^2$, $\sum' p_{ij}p_{kl} = 2M(n - 2)p^2$, $E(W) = Mp$, and hence R reduces to

$$R = 2(2n - 3)p < 4np.$$

Hence the theorem is proved.

As a consequence of Theorem 1, the distribution of W can be approximated by a Poisson distribution if Δ_w and np are both close enough to zero.

Now assume that there is a $\lambda > 0$ such that

$$(C) \quad E(W) \rightarrow \lambda; \quad \text{Var}(W) \rightarrow \lambda \text{ as } n \text{ goes to infinity.}$$

As seen from (1) this requires that p goes to zero as $1/n^2$ and c goes to zero faster than $1/n^3$ as $n \rightarrow \infty$. Condition (C) is equivalent to

$$(C') \quad (n^2/2)P(I_{12} = 1) \rightarrow \lambda; \quad n^3P(I_{12} = I_{13} = 1) \rightarrow 0.$$

As a consequence of (C), or (C'), we have $\Delta_w \rightarrow 0$ and $np \rightarrow 0$ as $n \rightarrow \infty$. Hence Theorem 1 yields the following.

Corollary. If Condition (C), or Condition (C'), holds then W has, in the limit, a Poisson distribution with mean λ .

3. Examples

Example 1. Uniform distribution over the integers 1 to N . Let $\{A_1, A_2, \dots, A_n\}$ be a random sample of n values from a uniform distribution over the integers 1, 2, \dots , N . We find successively

$$\begin{aligned} p &= E(I_{12}) = P(I_{12} = 1) = 1/N, \\ E(I_{12}I_{13}) &= P(I_{12} = I_{13} = 1) = 1/N^2, \\ c &= \text{Cov}(I_{12}, I_{13}) = 0, \\ E(W) &= M/N, \\ \text{Var}(W) &= (M/N)(1 - 1/N). \end{aligned}$$

Since

$$\Delta_w = -1/N; \quad np = n/N$$

we conclude from Theorem 1 that the distribution of W has, approximately, a Poisson distribution with mean M/N if N is large and n/N is small. Further, we infer from the Corollary that, when n and N go to infinity in such a way that $M/N \rightarrow \lambda$, then W has, in the limit, a Poisson distribution with mean λ .

For example, consider 200 random numbers from the set 0000, 0001, \dots , 9999. Then $n = 200$, $N = 10^4$ and so $E(W) = \binom{200}{2}/10^4 = 2.0$. As N is large and $n/N = 0.02$ is small, we may expect that the distribution of W can be approximated by a Poisson distribution with mean 2.0. In fact, we have from Theorem 1 that $d(W, Z) < 0.07$, where $Z \sim \text{Poisson}(2.0)$.

We can also formulate a special case of Example 1 as a *birthday problem*. Consider the birthdays of n persons. Two persons are said to be similar if they have the same birthday. Assuming the $N = 365$ days equally likely as a birthday, the number W of similar pairs among the n persons has mean $\binom{n}{2}/365$. If $n/365$ is small, it follows from Theorem 1 that the distribution of W can be approximated by a Poisson distribution. For example, take $n = 23$ which is an often quoted value since $P(W \geq 1)$ is then slightly greater than 1/2 (in fact, it is equal to 0.5073). Then $E(W) = 0.69$ and Theorem 1 yields $d(W, Z) < 0.12$; this is not very informative. Hence we cannot decide in this way whether the Poisson approximation is good or not. Note, however, that when $n = 23$ the Poisson approximation yields $P(Z \geq 1) = 0.5000$ which is a good approximation of the correct value 0.5073; cf. Schwartz (1988). Remember that the variational distance gives an upper bound for the error of the Poisson approximation of the probability of any event $\{W \in B\}$.

Example 2. Non-uniform distribution. Consider again a distribution over 1 to N , where k now occurs with probability p_k , $k = 1, 2, \dots, N$, $\sum p_k = 1$. Setting $p = \sum p_k^2$, $r = \sum p_k^3$, we find

$$\begin{aligned} E(W) &= Mp; \quad \text{Var}(W) = Mp(1 - p) + 2M(n - 2)(r - p^2), \\ \Delta_w &= 2(n - 2)(r/p - p) - p. \end{aligned}$$

By Theorem 1 we can use the Poisson approximation if Δ_w and np are small. Condition (C') becomes in this case $(n^2/2)p \rightarrow \lambda$, $n^3r \rightarrow 0$. If these limiting relations are satisfied, it follows from the Corollary that W has, in the limit, a Poisson distribution with mean λ .

Example 3. DNA breakages. In Cowan et al. (1987), a model for studying damage to circular DNA is studied. The mathematical model can be described as follows. Let a circle have a circumference of length 1. On the circumference n points are independently plotted using a uniform distribution. Each point is marked 0 or 1, independently by flipping a fair

coin. If two points with different marks are too close, the circle breaks. We seek the probability of this event.

To be more precise, suppose that *elements* $A_i = (P_i, X_i, U_i)$ are generated in the following way: the *points*, P_1, \dots, P_n , are taken from a uniform distribution on the circumference, the *critical distances*, X_1, \dots, X_n , are i.i.d. random variables and the *marks*, U_1, \dots, U_n , are *Bernoulli*($\frac{1}{2}$) random variables. The P 's, X 's and U 's are all independent. Let D_{ij} be the arc-length distance between P_i and P_j and define $I_{ij} = 1$ if U_i and U_j are different and $D_{ij} < \min(X_i, X_j)$, and $I_{ij} = 0$ otherwise. When $I_{ij} = 1$ the elements A_i and A_j are said to form a *similar pair*. The event, 'no breakage occurs', is equivalent to the event $W = \sum_{i < j} I_{ij} = 0$. Thus we have $P(\text{no breakage}) = P(W = 0)$.

The indicators I_{ij} have the structure described in Section 1. Conditional on X_1, X_2 the event $I_{12} = 1$ happens with probability

$$(2 \min(X_1, X_2)) \cdot (1/2) = \min(X_1, X_2).$$

Further, the event $I_{12} = I_{13} = 1$ occurs with probability

$$\min(X_1, X_2) \min(X_1, X_3).$$

Hence we have to take

$$p = E[\min(X_1, X_2)]$$

$$c = \text{Cov}(\min(X_1, X_2), \min(X_1, X_3))$$

in (1).

If Δ_w given by (2) and np are small, Theorem 1 shows that the distribution of W can be approximated by a Poisson distribution. As a consequence we obtain

$$P(W = 0) \approx \exp\left[-\binom{n}{2} E(\min(X_1, X_2))\right].$$

For example, if the X 's are uniformly distributed over the interval $(0, b)$, it is found that $E(X_1) = b/2$ and $E[\min(X_1, X_2)] = b/3$. Also it is seen after some calculation that $c = b^2/45$. Inserting these values in the inequality of Theorem 1 we can judge whether the variational distance is small enough to allow the Poisson distribution to be used. If this is possible, we conclude, finally, that

$$P(W = 0) \approx \exp\left[-\binom{n}{2} \frac{b}{3}\right].$$

This is then, approximately, the probability of no breakage.

For further results on the case $X_i \equiv b/2$, constant, see Cowan et al. (1990) and Holst (1989).

References

BARBOUR, A. D. AND EAGLESON, G. K. (1983). Poisson approximation for some statistics based on exchangeable trials. *Adv. Appl. Prob.* **15**, 585–600.

BARBOUR, A. D. AND EAGLESON, G. K. (1984) Poisson convergence for dissociated statistics. *J. R. Statist. Soc. B* **46**, 397–402.

COWAN, R., COLLIS, C. M. AND CRIGG, G. W. (1987) Breakage of double-stranded DNA due to single-stranded nicking. *J. Theor. Biol.* **127**, 229–246.

COWAN, R., CULPIN, D. AND GATES, D. (1990) Asymptotic results for a problem of DNA breakage. *J. Appl. Prob.* **27** (2).

HOLST, L. (1989) A circle covering problem and DNA breakage. *Statist. Prob. Lett.* **8** (2).

SCHWARTZ, W. (1988) Approximating the birthday problem. *Amer. Statistician* **42**, 195–196.