THE AMERICAN STATISTICIAN
2019, VOL. 73, NO. 1, 56–60: Teacher's Corner
https://doi.org/10.1080/00031305.2017.1311283

Sum of a Random Number of Correlated Random Variables that Depend on the Number of Summands

Joel E. Cohen

The Rockefeller University, New York, NY

ABSTRACT
The mean and variance of a sum of a random number of random variables are well known when the number of summands is independent of each summand and when the summands are independent and identically distributed (iid), or when all summands are identical. In scientific and financial applications, the preceding conditions are often too restrictive. Here, we calculate the mean and variance of a sum of a random number of random summands when the mean and variance of each summand depend on the number of summands and when every pair of summands has the same correlation. This article shows that the variance increases with the correlation between summands and equals the variance in the iid or identical cases when the correlation is zero or one.

ARTICLE HISTORY
Received May 2016
Accepted March 2017

KEYWORDS
Random sum; Random product; Variance; Correlation; Conditional expectation; Wald's equation; Blackwell–Girschick equation

1. Introduction

Vehicles pass a certain point on a highway. The number of vehicles per minute is a random variable, and each vehicle carries a random number of passengers. Because there are rush hours and off-peak hours, the number of vehicles per minute is positively associated with the number of passengers per vehicle and the numbers of passengers in different vehicles are positively correlated. What are the mean and variance of the total number of passengers that pass by per minute?

A random number of fish of a particular species lives under each square meter of the surface of a lake. Each individual fish is infected by a random number of parasites. Nutrient levels and temperatures vary from one region of the lake to another, altering local probability distributions of the number of fish per square meter, which is turn influences the mean and variance of the number of parasites per fish. What are the mean and variance of the total number of parasites that live under each square meter of lake surface?

A random number of stock traders bids for the shares of a particular company's stock each day. Each trader bids for a random number of shares. In bull markets, more traders bid for more shares each. What are the mean and variance of the total number of shares bid for per day?

A random number of tornadoes occurs in the continental United States each day. Each tornado may result in insurance claims, each for a random number of dollars. Because tornadoes occur in outbreaks, and because tornadoes in outbreaks may cause differently distributed claims than an equal number of isolated tornadoes, the number of tornadoes per day influences the distribution of the dollar values of claims, and these values may be correlated on any day. What are the mean and variance of the total dollar value of all tornado claims per day?

These examples from operations research, ecology, stock trading, and insurance ask for the mean and variance of a sum of a random number of random variables. Well-known formulas assume independence between the number of summands and the size of each summand. One extreme case is widely taught: in a "random sum," all the summed random variables are independent and identically distributed (iid). A second extreme case is a less known, though classical: in a "random product," all the summed random variables are identical (with probability 1), so that the sum equals the product of a random number of summands times one random summand. While the mean of a random sum equals the mean of a random product, the variances differ.

Here, we derive formulas for the mean and variance of a sum of a random number of random variables when the number of summands influences the mean and variance of the summands and when the summands are correlated. Special cases of these formulas yield known formulas, including Wald's equation and the Blackwell–Girschick equation, but the main results appear to be new.

These derivations give advanced undergraduate and beginning graduate students an opportunity to see the usefulness and power of conditional expectation and moment-generating functions. Students can follow the proofs because the concepts and methods are elementary (given familiarity with random variables and expectations) and the results have practical interpretations. The results are relevant to the statistical practitioner in applications such as those mentioned above.

2. Notation and Definitions

We model a sum of a random number of random summands when the random summands may be correlated with one

CONTACT Joel E. Cohen cohen@rockefeller.edu Abby Rockefeller Mauzé Professor of Populations, Rockefeller and Columbia Universities, 1230 York Avenue, New York, NY 10065; Department of Statistics, University of Chicago.

another and may have mean and variance that depend on the number of random summands.

Let N be a random variable taking nonnegative integer values $n = 0, 1, 2, \ldots$. We write $E_N(f(N)) = \sum_{n=0}^{\infty} f(n)\Pr\{N = n\}$ for the expectation with respect to N of any function $f(N)$ of N. This expectation may or may not exist. The moment-generating function (mgf) of N with any real argument z is by definition $\phi(z) = E_N(e^{zN}) = \sum_{n=0}^{\infty} e^{zn}\Pr\{N = n\}$. We assume $\phi(z)$ exists for any real z. For any positive integer m, the mth derivative of $\phi(z)$ evaluated at z is $\phi^{(m)}(z) = \sum_{n=0}^{\infty} n^m e^{zn}\Pr\{N = n\}$ and $\phi^{(m)}(0) = E(N^m)$ (e.g., Ross 1997, p. 61). For example, $\phi'(0) = E_N(N)$ is the mean, $\phi''(0) = E_N(N^2)$ is the second moment, and $\operatorname{var}_N(N) = \phi''(0) - [\phi'(0)]^2$ is the variance, which all exist since the mgf $\phi(z)$ exists by assumption. We further assume the mean and variance of N are positive.

Assume also that $X_1, X_2, \ldots$ are random variables, all with the same distribution, but not necessarily independent, on the same probability space as N such that, for all i, the conditional expectation $E(X_i|N)$ and the conditional variance $\operatorname{var}(X_i|N)$ exist. For real constants g, h, s, t, which do not depend on X_i or N, assume that

$$E(X_i|N) = e^{g+hN}, \quad \text{for all } i, \tag{1}$$
$$\operatorname{var}(X_i|N) = e^{s+tN}, \quad \text{for all } i. \tag{2}$$

The applications in Section 1 all concern sums of nonnegative random variables with positive means and variances. A partial justification for assumptions (1) and (2) is that they guarantee in a natural way that $E(X_i|N) > 0$ and $\operatorname{var}(X_i|N) > 0$ for all N and for all i (but they do not require that X_i be nonnegative). Define the sum S of a random number N of (possibly correlated) random variables X_i as

$$S = 0 \ \text{ if } \ N = 0 \quad \text{and} \quad S = \sum_{i=1}^{N} X_i \ \text{ if } \ N \geq 1. \tag{3}$$

We define the conditional correlation of X_i and X_j given N, which we write as $\rho_{ij}|N$, by

$$\rho_{ij}|N = \frac{\operatorname{cov}(X_i|N, X_j|N)}{[\operatorname{var}(X_i|N) \times \operatorname{var}(X_j|N)]^{1/2}}. \tag{4}$$

If $i = j$, then $\rho_{ij}|N = 1$. Because we assumed $\operatorname{var}(X_i|N) = e^{s+tN} > 0$, for all i, the denominator of the right side of (4) is positive and, by (2), is equal to e^{s+tN}. Then

$$\begin{aligned}\operatorname{cov}(X_i|N,\ X_j|N) &= (\rho_{ij}|N) \times \operatorname{var}(X_k|N) \\ &= (\rho_{ij}|N) \times e^{s+tN} \quad \text{for all } i, j, k. \end{aligned} \tag{5}$$

We define $\{X_i|\ i = 1, 2, \ldots\}$ to be equicorrelated if for each value of N there exists ρ_N such that $-1 \leq \rho_N \leq +1$ and such that, if $i \neq j$, then $\rho_{ij}|N = \rho_N$. We define $\{X_i|\ i = 1, 2, \ldots\}$ to be uniformly equicorrelated if there exists some ρ, $-1 \leq \rho \leq +1$, independent of N, such that, if $i \neq j$, then $\rho_{ij}|N = \rho$.

3. Mean and Variance of Random Sums

We calculate the mean and variance of S and explore some of their properties.

Proposition 1. Under the assumptions above, if $\{X_i|\ i = 1, 2, \ldots\}$ are equicorrelated (not necessarily uniformly equicorrelated), then S defined in (3) satisfies

$$E(S) = e^g\phi'(h), \tag{6}$$
$$\operatorname{var}(S) = e^s\{\phi'(t) + E_N[\rho_N(N^2 - N)e^{tN}]\} + e^{2g}\phi''(2h) - e^{2g}[\phi'(h)]^2. \tag{7}$$

We have not seen these formulas elsewhere.

Proof. To prove (6), we calculate

$$\begin{aligned} E(S) &= E_N[E(S|N)] = E_N\left[E\left(\sum_{i=1}^{N} X_i|N\right)\right] \\ &= E_N\left[\sum_{i=1}^{N} E(X_i|N)\right] = E_N[N \times E(X_1|N)] \\ &= E_N\left[N \times e^{g+hN}\right] = e^g E_N\left[Ne^{hN}\right] = e^g\phi'(h). \end{aligned}$$

Next, $\operatorname{var}(S) = E(S^2) - [E(S)]^2 = E(S^2) - [e^g\phi'(h)]^2$. Using (5) and the definition of equicorrelated, $\rho_{ij}|N = \rho_N$, we have

$$\begin{aligned} E(S^2) &= E_N[E(S^2|N)] = E_N\left[E\left\{\left(\sum_{i=1}^{N} X_i|N\right)^2\right\}\right] \\ &= E_N\left[E\left\{\sum_{i=1}^{N}(X_i|N)^2 + 2\sum_{i=1}^{N-1}\sum_{j>i}^{N} X_i\,|N \cdot X_j|\,N\right\}\right] \\ &= E_N\left[\sum_{i=1}^{N} E\left[(X_i|N)^2\right] + 2\sum_{i=1}^{N-1}\sum_{j>i}^{N} E\left(X_i\,|N \cdot X_j|\,N\right)\right] \\ &= E_N\left[\sum_{i=1}^{N}\left\{\operatorname{var}(X_i|N) + [E(X_i|N)]^2\right\}\right. \\ &\quad \left. + 2\sum_{i=1}^{N-1}\sum_{j>i}^{N}\left\{\operatorname{cov}\left(X_i\,|N,\ X_j|\,N\right) + E(X_i|N)\,E\left(X_j|N\right)\right\}\right] \\ &= E_N\left[N\left\{\operatorname{var}(X_1|N) + [E(X_1|N)]^2\right\}\right. \\ &\quad \left. + 2\sum_{i=1}^{N-1}\sum_{j>i}^{N}\left\{\left(\rho_{ij}|N\right)\operatorname{var}(X_1|N) + [E(X_1|N)]^2\right\}\right] \\ &= E_N\left[N\left\{e^{s+tN} + \left(e^{g+hN}\right)^2\right\}\right] + 2E_N\left[\operatorname{var}(X_1|N)\right. \\ &\quad \left. \times\left(\frac{N(N-1)}{2}\right)\rho_N + \left(\frac{N(N-1)}{2}\right)\left(e^{g+hN}\right)^2\right] \\ &= E_N\left[Ne^{s+tN}\right] + e^{2g}E_N\left[Ne^{2hN}\right] + E_N\left[N(N-1)e^{s+tN}\rho_N\right] \\ &\quad + E_N\left[N(N-1)\left(e^{g+hN}\right)^2\right] \\ &= e^s\phi'(t) + e^{2g}\phi'(2h) + e^s E_N\left[N(N-1)e^{tN}\rho_N\right] \\ &\quad + e^{2g}\phi''(2h) - e^{2g}\phi'(2h). \end{aligned}$$

Hence

$$\begin{aligned} &\operatorname{var}(S) \\ &\quad = E(S^2) - [E(S)]^2 = e^s\phi'(t) + e^{2g}\phi'(2h) \\ &\qquad + e^s E_N\left[N(N-1)e^{tN}\rho_N\right] + e^{2g}\phi''(2h) - e^{2g}\phi'(2h) - \left[e^g\phi'(h)\right]^2 \\ &\quad = e^s\left\{\phi'(t) + E_N\left[\rho_N\left(N^2 - N\right)e^{tN}\right]\right\} + e^{2g}\phi''(2h) - e^{2g}\left[\phi'(h)\right]^2. \end{aligned}$$

This proves (7) and Proposition 1. □

Under the assumptions of Proposition 1, $E(S)$ and $\text{var}(S)$ are obviously strictly increasing functions of each of the parameters in (1) and (2), g, h, s, and t, all else held constant. If $h \neq 0$ and $t \neq 0$, then (1) and (2) are strictly convex functions of N. Jensen's inequality gives

$$E_N[E(X_1|N)] > E(X_1|E_N[N])\,,$$
$$E_N[\text{var}(X_1|N)] > \text{var}(X_1|E_N[N])\,.$$

Both inequalities are strict because, since $\text{var}_N(N) > 0$, N gives positive probability mass to at least two integers. When $h \neq 0$ and $t \neq 0$, then the unconditional mean and variance of every X_i strictly exceed the mean and variance that would be obtained from (1) and (2) if the random variable N were replaced by its mean $E_N[N]$.

The variance of S is easily computed from (7) in three special cases (8), (9), (10). Let $0 \leq \rho \leq 1$. (The exclusion of negative values of ρ in this assumption is explained in Proposition 3.) Then

$$\rho_N = \frac{\rho}{N} \Rightarrow \text{var}(S) = e^s\{\phi'(t) + \rho[\phi'(t) - \phi(t)]\} + e^{2g}\phi''(2h) - e^{2g}[\phi'(h)]^2, \tag{8}$$

$$\rho_N = \frac{\rho}{N^2 - N} \Rightarrow \text{var}(S) = e^s\{\phi'(t) + \rho\phi(t)\} + e^{2g}\phi''(2h) - e^{2g}[\phi'(h)]^2, \tag{9}$$

$$\begin{aligned}\rho_N = \rho \Rightarrow \text{var}(S) &= e^s\{\phi'(t) + \rho[\phi''(t) - \phi'(t)]\} + e^{2g}\phi''(2h) - e^{2g}[\phi'(h)]^2 \\ &= e^s(1-\rho)\phi'(t) + e^s\rho\phi''(t) + e^{2g}\phi''(2h) - e^{2g}[\phi'(h)]^2.\end{aligned} \tag{10}$$

Proposition 2. Under the assumptions of Proposition 1 and the additional assumptions that $\{X_i|\ i = 1, 2, \ldots\}$ are uniformly equicorrelated with correlation ρ and $\phi'(t) \geq 1$, then $\text{var}(S)$, which is given by (10), is an increasing function of ρ in $0 \leq \rho \leq 1$, all else held constant.

Proof. In (10), the coefficient of ρ is $e^s(\phi''(t) - \phi'(t))$, so $\text{var}(S)$ is a (weakly) increasing function of ρ, all else held constant, if and only if $\phi''(t) - \phi'(t) \geq 0$. We show that $\phi''(t) = \sum_{n=0}^{\infty} n^2 e^{tn}\Pr\{N = n\} \geq \phi'(t) = \sum_{n=0}^{\infty} ne^{tn}\Pr\{N = n\}$ if $\phi'(t) \geq 1$. For $n = 0, 1, 2, \ldots$, let $a_n = e^{tn}\Pr\{N = n\} \geq 0$ and, for any positive integer M, let $A_M = \sum_{n=0}^{M} a_n$. Since $\phi'(t) \geq 1$ by assumption, there exists a positive integer M' such that $A_M > 0$ for all $M \geq M'$. For all $M \geq M'$, let $b_{i,M} = a_i/A_M$, $i = 1, \ldots, M$. Then for every $M \geq M'$, $\{b_{i,M}\}_{i=0}^{M}$ is a well-defined probability distribution (a set of nonnegative numbers that sum to 1). For all $M \geq M'$, $E_B(N|M) = \sum_{n=0}^{M} n \times b_{n,M}$ is the expectation of N relative to this probability distribution, $E_B(N^2|M) = \sum_{n=0}^{M} n^2 \times b_{n,M}$ is the expectation of N^2 relative to this probability distribution, and $\text{var}_B(N|M) = E_B(N^2|M) - [E_B(N|M)]^2$ is the variance of N relative to this probability distribution. The variance of any random variable must be nonnegative. Thus $\text{var}_B(N|M) \geq 0$ for all $M \geq M'$ implies that $A_M \times E_B(N^2|M) \geq A_M \times [E_B(N|M)]^2$ which is the same as $\sum_{n=0}^{M} n^2 e^{tn}\Pr\{N = n\} \geq [\sum_{n=0}^{M} ne^{tn}\Pr\{N = n\}]^2$ for all $M \geq M'$. Letting $M \to \infty$ in this inequality gives $\phi''(t) \geq [\phi'(t)]^2$. Since $\phi'(t) \geq 1$ by assumption, $[\phi'(t)]^2 \geq \phi'(t)$. Therefore, $\phi''(t) \geq \phi'(t)$, which proves Proposition 2. □

Obviously, ρ cannot exceed 1. What is the lower limit on ρ for uniformly equicorrelated $\{X_i|i = 1, 2, \ldots, N\}$? We do not have complete freedom in choosing ρ within the interval $-1 \leq \rho \leq +1$ for uniformly equicorrelated summands. To see why, suppose that $X = +2$ with probability 1/2 and $X = 0$ with probability 1/2. Suppose that X_1, X_2, X_3 all have the distribution of X. Then it is impossible that X_1, X_2, X_3 are uniformly equicorrelated with $\rho = -1$. Why? Since $E(X) = 1$, $\text{var}(X) = 1$, we have $\rho_{12} = E(X_1X_2) - E(X_1)E(X_2) = E(X_1X_2) - 1 = -1$ if and only if $X_1 + X_2 = 1$ and $\rho_{13} = E(X_1X_3) - 1 = -1$ if and only if $X_1 + X_3 = 1$. But then $X_2 = X_3$ so X_1, X_2, X_3 are not uniformly equicorrelated with $\rho = -1$.

Proposition 3. If there exists a finite positive integer $n > 1$ such that $\Pr\{N > n\} = 0$, then, for every ρ such that $-1/(n-1) \leq \rho \leq 1$, $X_1, \ldots, X_n$ may be uniformly equicorrelated with correlation ρ. If there exists no finite positive integer $n > 1$ such that $\Pr\{N > n\} = 0$, then $X_1, \ldots, X_n$ may be equicorrelated with correlation ρ for every ρ such that $0 \leq \rho \leq 1$.

Proof. If $\Pr\{N > n\} = 0$, let the correlation matrix R be the $n \times n$ matrix with all diagonal elements $R_{ii} = 1$ and all off-diagonal elements $R_{ij} = \rho$, $i \neq j$. By (5), $R \times \text{var}(X) = C$, where C is the covariance matrix of $X_1, \ldots, X_n$. It is well known that a square, real, symmetric matrix is a covariance matrix if and only if it is positive semidefinite (Feller 1971, p. 83), that is, if and only if all its eigenvalues are nonnegative reals. Since the eigenvalues of C are $\text{var}(X)$ times the eigenvalues of R, it suffices to establish the conditions under which all the eigenvalues of R are nonnegative reals.

Let J be the $n \times n$ matrix with all elements equal to 1. As J is a matrix of rank 1, its eigenvalues equal n with multiplicity 1 (corresponding to the eigenvector with all elements equal to 1) and equal 0 with multiplicity $n-1$. Therefore, the eigenvalues of ρJ equal ρn with multiplicity 1 (corresponding to the eigenvector with all elements equal to 1) and equal 0 with multiplicity $n-1$. Let I be the $n \times n$ identity matrix, with all diagonal elements 1 and all off-diagonal elements 0. Then $\rho J + (1-\rho)I = R$ because the diagonal elements of R are 1 and the off-diagonal elements of R are ρ. Adding $(1-\rho)I$ to ρJ simply increases every eigenvalue of ρJ by a constant $1 - \rho$. So the eigenvalues of R equal $\rho n + 1 - \rho = (n-1)\rho + 1$ with multiplicity 1 and $1 - \rho$ with multiplicity $n - 1$. These eigenvalues are real, and they are nonnegative if and only if $-1/(n-1) \leq \rho \leq 1$.

As $n \to \infty$, $-1/(n-1) \to 0$, so if there exists no finite positive integer $n > 1$ such that $\Pr\{N > n\} = 0$, then ρ is constrained to [0, 1] and may fall anywhere in it. □

Returning to the example that precedes Proposition 3, if X_1, X_2, X_3 are equicorrelated with correlation ρ, then Proposition 3 tells us that $\rho \geq -1/2$.

4. Special Cases

In this section, we show that our results lead to known formulas for some special cases, and to possibly new formulas in other special cases.

4.1. Special Cases Leading to Known Results

If $h = 0$, then, for all i, $E(X_i|N) = e^g > 0$, independent of N. In this case, assumption (1) implies no restriction on generality other than the assumption that the mean of X_i is positive, and $\phi'(h) = \phi'(0) = E_N(N)$. Then (6) simplifies to

$$E(S) = E_N(N)E(X_1) \text{ (“the product rule” or Wald's equation)}, \quad (11)$$

regardless of possible pairwise correlations among the summands X_i.

The product rule is well known in the special case that X_i and X_j are independent for all $i \neq j$ and independent of N (e.g., Ross 1997, p. 101). The product rule clearly also holds for non-independent summands whose conditional expectation given $N = n$ is independent of n but most textbooks, like the one by Ross, treat only the case of independent, identically distributed summands. Our calculation shows that the weaker assumptions that X_i and X_j are pairwise uncorrelated (not necessarily independent) or alternatively that their moments are independent of N suffice to yield the product rule.

The product rule also holds if X_i and X_j are perfectly correlated ($\rho = 1$) for all $i \neq j$ and independent of N (e.g., Goodman 1960, p. 709, his eq. (2)). Our calculations show that the product rule holds for the uncorrelated, perfectly correlated, and all intermediate cases if, for all i, $E(X_i|N = n)$ does not depend on n.

If $h = t = 0$, the mean and variance of all X_i are independent of N. Consequently, $e^g = E(X_i)$, $e^s = \text{var}(X_i)$, $\phi'(0) = E_N(N)$, and $\phi''(0) = E_N(N^2)$. In this case, if all X_i are uniformly equicorrelated with correlation ρ, then from (10) we have

$$\begin{aligned}\text{var}(S) = &(1-\rho)\, E_N(N)\text{var}(X_1)\\ &+\rho(\text{var}_N(N) + [E_N(N)]^2)\text{var}(X_1)\\ &+\text{var}_N(N)[E(X_1)]^2. \end{aligned} \quad (12)$$

As a consequence of Proposition 2, if $E_N(N) \geq 1$, then $\text{var}(S)$ is an increasing function of ρ, all else held constant. If $E_N(N) < 1$, then $\text{var}(S)$ need not be an increasing function of ρ, all else held constant.

If $\rho = 0$ in addition to $h = t = 0$, (12) becomes the well-known result

$$\text{var}(S) = E_N(N)\text{var}(X_1) + \text{var}_N(N)[E(X_1)]^2 \quad \text{(Blackwell-Girschick equation)}. \quad (13)$$

Blackwell and Girschick (1947, p. 277, their theorem 2) and Ross (1997, p. 110) derived (13) under the stronger assumption that X_i and X_j are independent for all $i \neq j$. The weaker assumption that X_i and X_j are pairwise uncorrelated suffices.

Still assuming that $h = t = 0$, if $\rho = 1$, then all X_i take identical values. Thus, $S = N \times X_1$. Then (12) becomes (Goodman 1960, p. 709, his eq. (2))

$$\begin{aligned}\text{var}(S) = &[E_N(N)]^2\text{var}(X_1) + \text{var}_N(N)[E(X_1)]^2\\ &+ \text{var}_N(N)\text{var}(X_1). \end{aligned} \quad (14)$$

Thus (12) interpolates $\text{var}(S)$ between (13) (no correlation among summands) and (14) (perfect correlation among summands) for $0 \leq \rho \leq 1$.

In the limit of (12) as $\text{var}_N(N) \to 0$ while $E_N(N)$ remains a positive constant, we have, for some integer $n > 0$, that $N \to n$ in probability and (12) becomes $\text{var}(S) = (1-\rho)\, n\text{var}(X_1) + \rho n^2\text{var}(X_1)$, whence $\text{var}(\bar{X}) = \text{var}(X_1)(\frac{1-\rho}{n} + \rho)$, which is a standard formula for the variance of the sample mean of n identically distributed random variables with average pairwise correlation ρ.

4.2. Number of Summands is Poisson Distributed

Suppose N has the Poisson distribution with mean and variance $\lambda > 0$. Then for any real z, $\phi(z) \equiv E(e^{zN}) = \exp[\lambda(e^z - 1)]$, $\phi'(z) = \lambda e^z \exp[\lambda(e^z - 1)]$ and $\phi''(z) = \lambda e^z(1 + \lambda e^z)\exp[\lambda(e^z - 1)]$ (Ross 1997, p. 62, his Ex. 2.40). From (6),

$$E(S) = \lambda e^{g+h}\exp\left[\lambda\left(e^h - 1\right)\right]. \quad (15)$$

As a check, if $h = 0$, then $\lambda = E_N(N)$ and $e^g = E(X_1)$, independent of N, so (15) becomes the product rule (11).

From (7), assuming $h \neq 0$ and $t \neq 0$,

$$\begin{aligned}\text{var}(S) = &e^s\left\{\lambda e^t \exp\left[\lambda\left(e^t - 1\right)\right] + E_N\left[\rho_N\left(N^2 - N\right)e^{tN}\right]\right\}\\ &+ \lambda e^{2(g+h)}\left\{\left(1 + \lambda e^{2h}\right)\exp\left[\lambda\left(e^{2h} - 1\right)\right]\right.\\ &\left.- \lambda\exp\left[\lambda\left(e^h - 1\right)\right]\right\}. \end{aligned} \quad (16)$$

If the summands are uniformly equicorrelated and $\rho_N = \rho$, then

$$\begin{aligned}\text{var}(S) = &\lambda e^{s+t}\exp\left[\lambda\left(e^t - 1\right)\right]\left\{1 + \rho\left[\left(1 + \lambda e^t\right) - 1\right]\right\}\\ &+ \lambda e^{2(g+h)}\left\{\left(1 + \lambda e^{2h}\right)\exp\left[\lambda\left(e^{2h} - 1\right)\right]\right.\\ &\left.- \lambda\exp\left[\lambda\left(e^h - 1\right)\right]\right\}. \end{aligned} \quad (17)$$

As a check, when $h = t = 0$, (17) reduces to $\text{var}(S) = \lambda e^s\{1 + \rho[(1+\lambda) - 1]\} + \lambda e^{2g}\{(1+\lambda) - \lambda\} = \lambda\text{var}(X_1)\{1 + \rho\lambda\} + \lambda[E(X_1)]^2$, which is identical to (12) when N is distributed as Poisson(λ) and $\lambda = E_N(N) = \text{var}_N(N)$.

5. Conclusions

Two scientific problems mentioned in the Introduction motivated this analysis. First, Lagrue et al. (2015) estimated the numbers of individual host organisms, including fishes, per square meter of lake surface in four lakes in Otago, New Zealand, and the numbers of parasite individuals per square meter of these lakes. Cohen, Poulin, and Lagrue (2016) also estimated the numbers of parasite individuals per host individual (“the parasite load”). They analyzed two extreme models of the correlation of the parasite load between hosts within one square meter: independence and perfect correlation. They found that two models that assumed perfect correlation (across host individuals) of parasite loads within a square meter yielded more realistic variance functions (relations of variance to mean of the number of parasites per square meter) than two models that assumed independence of parasite loads within a square meter. They speculated that correlations less than, but close to, one would further improve the realism of the variance functions of the two more successful models, but they lacked the formulas to test that speculation. The results established here, particularly the monotonic increase of the variance of the sum as a function of the equicorrelation of the summands, confirm their speculation. Direct measurement of the correlation of parasite loads across host individuals within a square meter in the field remains an open empirical challenge.

Second, over recent decades, there has been no long-term trend in the average annual number of tornadoes (of intensity F1 or greater, called F1+, which are those tornadoes sufficiently intense to be reliably observed). However, more F1+ tornadoes have been reported on the days when tornadoes occurred (Brooks et al. 2014; Elsner et al. 2015). An outbreak is defined as a sequence of F1+ tornadoes with initiation times separated by not more than 6 hours; an outbreak may stretch over more than one calendar day. The mean and the variance of the number of F1+ tornadoes per outbreak have increased in recent decades (Tippett and Cohen 2016; Tippett et al. 2016). Counts of tornadoes per day or per outbreak are thus increasingly clustered and overdispersed, despite the absence of any trend in the total number per year of F1+ tornadoes. The results in this article provide a framework for analyses of the mean and variance of the claims per tornado day or per tornado outbreak when summands are correlated and dependent on the number of summands. These results could be used in combination with Markov's inequality to provide bounds on the so-called Value at Risk (VaR) in financial mathematics and insurance (e.g., Resnick 2007).

In historical perspective, our main result (6) on the expected sum is a generalization of Wald's equation (Wald 1945, p. 142, his eq. (4.4)), and our main result (7) on the variance of the sum is a generalization of the Blackwell–Girschick equation (Blackwell and Girschick 1947, p. 277, their theorem 2). Some recent results focus on the asymptotic distribution of the sum when the random number of possibly dependent summands is independent of the summands (Islak 2016). Other recent results on random sums focus on the asymptotic distribution of the sum when the number of summands is independent of the mutually independent but not necessarily identically distributed summands (Sunklodas 2012) or when the summands are iid and the number of summands is independent of the summands (Sunklodas 2014, 2015). Although we made no attempt at an exhaustive survey, we found no prior exact results which assumed, as we did here, that the summands were correlated and dependent on the number of summands.

In principle, the approach used here to calculate the first and second moments of the sum of a random number of correlated summands dependent on the number of summands could be extended to higher moments or correlation structures more elaborate than equicorrelation. It would also be interesting to study the asymptotic distribution of the sum under our assumptions. But the approach used here may prove laborious and more powerful tools may be required. The simple problems solved here might prove useful for a doctoral qualifying exam. The unsolved problems might prove suitable for doctoral research.

Funding

This work was partially supported by U.S. National Science Foundation grant DMS-1225529 and the assistance of Priscilla K. Rogerson.

ORCID

Joel E. Cohen http://orcid.org/0000-0002-9746-6725

References

Blackwell, D., and Girshick, M. A. (1947), "A Lower Bound For The Variance Of Some Unbiased Sequential Estimates," *Annals of Mathematical Statistics*, 18, 277–280. [59,60]

Brooks, H. E., Carbin, G. W., and Marsh, P. T. (2014), "Increased Variability Of Tornado Occurrence In The United States," *Science*, 346, 349–352. [60]

Cohen, J. E., Poulin, R., and Lagrue, C. (2016), "Linking Parasite Populations In Hosts To Parasite Populations In Space Through Taylor's Law And The Negative Binomial Distribution," *Proceedings of the National Academy of Sciences USA*, 114, E47–E56. [59]

Elsner, J. B., Elsner, S. C., and Jagger, T. H. (2015), "The Increasing Efficiency Of Tornado Days In The United States," *Climate Dynamics*, 45, 651–659. [60]

Feller, W. (1971), *An Introduction to Probability Theory and Its Applications* (vol. 2, 2nd ed.), New York: Wiley. [58]

Goodman, L. A. (1960), "On The Exact Variance Of Products," *Journal of the American Statistical Association*, 55, 708–713. [59]

Islak, Ü. (2016), "Asymptotic Results For Random Sums Of Dependent Random Variables," *Statistics and Probability Letters*, 109, 22–29. [60]

Lagrue, C., Poulin, R., and Cohen, J. E. (2015), "Parasitism Alters 3 Power Laws Of Scaling In A Metazoan Community: Taylor's Law, Density-Mass Allometry, And Variance-Mass Allometry," *Proceedings of the National Academy of Sciences USA*, 112, 1791–1796. [59]

Resnick, S. I. (2007), *Heavy-Tail Phenomena: Probabilistic and Statistical Modeling*, New York: Springer. [60]

Ross, S. M. (1997), *Introduction to Probability Models* (6th ed.) , San Diego, CA: Academic Press. [57,59]

Sunklodas, J. K. (2012), "On The Normal Approximation Of A Sum Of A Random Number Of Independent Random Variables," *Lithuanian Mathematical Journal*, 52, 435–443. [60]

——— (2014), "On The Normal Approximation Of A Binomial Random Sum," *Lithuanian Mathematical Journal*, 54, 356–365. [60]

——— (2015), "On The Normal Approximation Of A Negative Binomial Random Sum," *Lithuanian Mathematical Journal*, 55, 150–158. [60]

Tippett, M. K., and Cohen, J. E. (2016), "Tornado Outbreak Variability Follows Taylor's Power Law Of Fluctuation Scaling And Increases Dramatically With Severity," *Nature Communications*, 7, 10668. [60]

Tippett, M. K., Lepore, C., and Cohen, J. E. (2016), "More Tornadoes In The Most Extreme U.S. Tornado Outbreaks," *Science* 10.1126/science.aah7393. [60]

Wald, A. (1945), "Sequential Tests Of Statistical Hypotheses," *Annals of Mathematical Statistics*, 16, 117–186. [60]

THE AMERICAN STATISTICIAN
2020, VOL. 74, NO. 2, 212: Correction
https://doi.org/10.1080/00031305.2019.1708461

Correction

Article title: Sum of a Random Number of Correlated Random Variables that Depend on the Number of Summands
Authors: Cohen, J. E. and Philip Turk
Journal: *The American Statistician*
Citation details: Volume 73, Number 1, pages 56–60
DOI: http://dx.doi.org/10.1080/00031305.2017.1311283

P.T. pointed out two mistakes. As a consequence of correcting one of these mistakes, J.E.C. here states and proves a stronger, simpler version of Proposition 2. Correcting the other mistake has no effect on the result claimed.

First, on page 58, left column, 3 lines from the bottom, the right side, namely, $A_M \times [E_B\,(N|M)]^2$, does not equal the right side given in the next line.

The right side should be $(1/A_M) \times \left[\sum_{n=0}^{M} ne^{tn}P\{N = n\}\right]^2$. Re-examination of the remainder of the proof led J.E.C. to a stronger, simpler version of Proposition 2 that omits the unnecessary assumption that $\varphi'(t) \geq 1$ and adds to the conclusion a necessary and sufficient condition for Var (S) to be strictly increasing.

Revised Proposition 2. Under the assumptions of Proposition 1, if $\{X_i | i = 1, 2, \ldots\}$ are uniformly equicorrelated with correlation ρ, then Var (S), which is given by (10), is an increasing function of ρ in $0 \leq \rho \leq 1$, all else held constant. A necessary and sufficient condition for Var (S) to be a strictly increasing function of ρ in $0 \leq \rho \leq 1$, all else held constant, is that there exists $n > 1$ such that $P\{N = n\} > 0$.

Revised proof. *Var* (S) given by (10) is an increasing function of ρ in $0 \leq \rho \leq 1$, all else held constant, if and only if $\varphi''(t) \geq \varphi'(t)$. The left side is

$$\varphi''(t) = \sum_{n=0}^{\infty} n^2 e^{tn} P[N = n]$$

and the right side is

$$\varphi'(t) = \sum_{n=0}^{\infty} ne^{tn} P[N = n].$$

Compare the nth term of $\varphi''(t)$ with the nth term of $\varphi'(t)$. Obviously,

$$n^2 e^{tn} P[N = n] > ne^{tn} P[N = n]$$

unless $n = 0$ or $n = 1$ or $P[N = n] = 0$. In these three cases,

$$n^2 e^{tn} P\,[N = n] = ne^{tn} P[N = n].$$

In any event, taking the sum of both sides of the two immediately preceding formulas, for n from 0 to ∞, yields $\varphi''(t) \geq \varphi'(t)$ and this inequality is strict if and only if there exists $n > 1$ such that $P\{N = n\} > 0$. □

After seeing this improved Proposition 2, P.T. pointed out that on page 59, in the left column, the paragraph directly below Eqn. (12) can be strengthened to: "As a consequence of the Revised Proposition 2, Var (S) is an increasing function of ρ in $0 \leq \rho \leq 1$, all else held constant, and is strictly increasing if and only if there exists $n > 1$ such that $P\{N = n\} > 0$."

Second, on page 58, in the right column, in lines 12 and 13, the statements $X_1 + X_2 = 1$ and $X_1 + X_3 = 1$ should be $X_1 + X_2 = 2$ and $X_1 + X_3 = 2$. The conclusions of the argument are not affected.