Probability in the Engineering and Informational Sciences, **18**, 2004, 473–484. Printed in the U.S.A.

COMPOUND RANDOM VARIABLES

EROL PEKÖZ
School of Management
Boston University
Boston, MA 02215
E-mail: pekoz@bu.edu

SHELDON M. ROSS
Epstein Department of Industrial and Systems Engineering
University of Southern California
Los Angeles, CA 90089
E-mail: smross@usc.edu

We give a probabilistic proof of an identity concerning the expectation of an arbitrary function of a compound random variable and then use this identity to obtain recursive formulas for the probability mass function of compound random variables when the compounding distribution is Poisson, binomial, negative binomial random, hypergeometric, logarithmic, or negative hypergeometric. We then show how to use simulation to efficiently estimate both the probability that a positive compound random variable is greater than a specified constant and the expected amount by which it exceeds that constant.

1. INTRODUCTION AND SUMMARY

Let $X_1, X_2, \ldots$ be a sequence of independent and identically distributed (i.i.d.) positive random variables that are independent of the nonnegative integer-valued random variable N. The random variable $S_N = \sum_{i=1}^{N} X_i$ is called a compound random variable. In Section 2, we give a simple probabilistic proof of an identity concerning the expected value of a function of a compound random variable; when the X_i are positive integer-valued, an identity concerning the probability mass function of S_N is obtained as a corollary. In Section 3, we use the latter identity to provide new

 0269-9648/04 $16.00

https://doi.org/10.1017/S0269964804184039 Published online by Cambridge University Press

derivations of the recursive formulas for the probability mass function of S_N when X_1 is a positive integer-valued random variable, and N has a variety of possible distributions. For other derivations of the applications of Section 3, the reader should see the references.

Sections 4 and 5 are concerned with finding efficient simulation techniques to estimate

$$p = P\{S \le c\} \quad \text{and} \quad \theta = E[(S-c)^+],$$

where c is a specified constant and the X_i need not be discrete. Because

$$E[(S-c)^+] = E[S-c|S>c](1-p)$$

and

$$\begin{aligned} E[N]E[X] - c &= E[S-c] \\ &= E[S-c|S>c](1-p) + E[S-c|S \le c]p, \end{aligned}$$

it follows that estimating p and θ will also give us estimates of $E[S-c|S>c]$ and $E[c-S|S \le c]$. Although our major interest is when the X_i are positive, in Section 5 we show how an effective simulation can be performed when this restriction is removed.

2. THE COMPOUND IDENTITY

Consider the compound random variable

$$S_N = \sum_{i=1}^{N} X_i.$$

Let M be independent of $X_1, X_2, \ldots$ and such that

$$P\{M=n\} = \frac{nP\{N=n\}}{E[N]}, \qquad n \ge 0.$$

The random variable M is called the sized bias version of N. (If the interarrival times of a renewal process were distributed according to N, then the average length of a renewal interval containing a fixed point would be distributed according to M.)

THEOREM 2.1 (The Compound Identity): *For any function h,*

$$E[S_N h(S_N)] = E[N]E[X_1 h(S_M)].$$

https://doi.org/10.1017/S0269964804184039 Published online by Cambridge University Press

PROOF:

$$E[S_N h(S_N)] = E\left[\sum_{i=1}^{N} X_i h(S_N)\right]$$

$$= \sum_{n=0}^{\infty}\sum_{i=1}^{N} E[X_i h(S_N)|N=n]P\{N=n\}$$

$$= \sum_{n=0}^{\infty}\sum_{i=1}^{n} E[X_i h(S_n)]P\{N=n\}$$

$$= \sum_{n=0}^{\infty} nE[X_1 h(S_n)]P\{N=n\}$$

$$= E[N]\sum_{n=0}^{\infty} E[X_1 h(S_n)]P\{M=n\}$$

$$= E[N]E[X_1 h(S_M)]$$ ■

COROLLARY 2.1: *If X_1 is a positive integer-valued random variable with $\alpha_i = P\{X_1 = i\}$, then*

$$P\{S_N = k\} = \frac{1}{k} E[N] \sum_{i=1}^{k} i\alpha_i P\{S_{M-1} = k-i\}.$$

PROOF: For an event A, let $I(A)$ equal one if A occurs and let it equal zero otherwise. Then, with $h(x) = I(x = k)$, the compound identity yields that

$$P\{S_N = k\} = \frac{1}{k} E[S_N I(S_N = k)]$$

$$= \frac{1}{k} E[N]E[X_1 I(S_M = k)]$$

$$= \frac{1}{k} E[N]\sum_i E[X_1 I(S_M = k)|X_1 = i]\alpha_i$$

$$= \frac{1}{k} E[N]\sum_i iP\{S_M = k|X_1 = i\}\alpha_i$$

$$= \frac{1}{k} E[N]\sum_i iP\{S_{M-1} = k-i\}\alpha_i.$$

3. SPECIAL CASES

Suppose that X_1 is a positive integer-valued random variable with $\alpha_i = P\{X_1 = i\}$.

https://doi.org/10.1017/S0269964804184039 Published online by Cambridge University Press

3.1. Poisson Case

If

$$P\{N = n\} = \frac{e^{-\lambda}\lambda^n}{n!}, \qquad n \geq 0,$$

then

$$P\{M - 1 = n\} = P\{N = n\}, \qquad n \geq 0.$$

Therefore, the corollary yields the well-known recursion

$$P\{S_N = k\} = \frac{1}{k}\,\lambda \sum_{i=1}^{k} i\alpha_i\, P\{S_N = k - i\}.$$

3.2. Negative Binomial Case

For a fixed value of p, we say that N is an $NB(r)$ random variable if

$$P\{N = n\} = \binom{n + r - 1}{n} p^r (1 - p)^n, \qquad n \geq 0.$$

Such a random variable can be thought of as being the number of failures that occur before a total of r successes have been amassed when each trial is independently a success with probability p.

If M is the size-biased version of an $NB(r)$ random variable N, then

$$P\{M - 1 = n\} = \frac{n + 1}{r(1 - p)/p}\binom{n + r}{n + 1} p^r (1 - p)^{n+1} = \binom{n + r}{n} p^{r+1}(1 - p)^n;$$

that is, $M - 1$ is an $NB(r + 1)$ random variable.

Now, for N an $NB(r)$ random variable, let

$$P_r(k) = P\{S_N = k\}.$$

The corollary now yields the recursion

$$P_r(k) = \frac{r(1 - p)}{kp} \sum_{i=1}^{k} i\alpha_i\, P_{r+1}(k - i).$$

For instance, starting with

$$P_r(0) = p^r,$$

https://doi.org/10.1017/S0269964804184039 Published online by Cambridge University Press

the recursion yields

$$\begin{aligned} P_r(1) &= \frac{r(1-p)}{p}\,\alpha_1 P_{r+1}(0) \\ &= rp^r(1-p)\alpha_1, \\ P_r(2) &= \frac{r(1-p)}{2p}\,[\alpha_1 P_{r+1}(1) + 2\alpha_2 P_{r+1}(0)] \\ &= \frac{r(1-p)}{2p}\,[\alpha_1^2(r+1)p^{r+1}(1-p) + 2\alpha_2 p^{r+1}], \\ P_r(3) &= \frac{r(1-p)}{3p}\,[\alpha_1 P_{r+1}(2) + 2\alpha_2 P_{r+1}(1) + 3\alpha_3 P_{r+1}(0)], \end{aligned}$$

and so on.

3.3. Binomial Case

If N is a binomial random variable with parameters r and p, then

$$\begin{aligned} P\{M-1=n\} &= \frac{n+1}{rp}\binom{r}{n+1}p^{n+1}(1-p)^{r-n-1} \\ &= \binom{r-1}{n}p^n(1-p)^{r-1-n}, \qquad 0 \le n \le r-1; \end{aligned}$$

that is, $M-1$ is a binomial random variable with parameters $r-1$ and p.

For a fixed p, let

$$P_r(k) = P\{S_N = k\}.$$

The corollary then yields the recursion

$$P_r(k) = \frac{rp}{k}\sum_{i=1}^{k} i\alpha_i P_{r-1}(k-i).$$

3.4. Hypergeometric Case

Let $N = N(w, r)$ be a hypergeometric random variable having the distribution of the number of white balls chosen when a random sample of r is chosen from a set of w white and b blue balls; that is,

$$P\{N=n\} = \frac{\binom{w}{n}\binom{b}{r-n}}{\binom{w+b}{r}}.$$

https://doi.org/10.1017/S0269964804184039 Published online by Cambridge University Press

Then, it is straightforward to check that

$$P\{M-1=n\}=\frac{\binom{w-1}{n}\binom{b}{r-n-1}}{\binom{w+b-1}{r-1}};$$

that is, $M-1$ has the same distribution as N with the modification that w becomes $w-1$ and r becomes $r-1$. Letting

$$P_{w,r}(k)=P\{S_{N(w,r)}=k\},$$

then

$$P_{w,r}(k)=\frac{rw}{k(w+b)}\sum_{i=1}^{k} i\alpha_i P_{w-1,r-1}(k-i).$$

This yields

$$P_{w,r}(1)=\frac{rw}{w+b}\,\alpha_1 P_{w-1,r-1}(0)=\frac{rw}{w+b}\,\alpha_1\,\frac{\binom{b}{r-1}}{\binom{w+b-1}{r-1}},$$

and so on. (We are using the convention that $\binom{n}{k}=0$ if either $k<0$ or $k>n$.)

3.5. The Logarithmic Count Distribution

Suppose that for $0<\beta<1$,

$$P\{N=n\}=C\,\frac{\beta^n}{n},\qquad n=1,2,\ldots,$$

where $C=-1/\ln(1-\beta)$. Then,

$$P\{M-1=n\}=\beta^n(1-\beta),\qquad n\geq 0;$$

that is, $M-1$ has the negative binomial distribution of Subsection 3.2 with $r=1$ and $p=1-\beta$. Thus, the recursion of Subsection 3.2 and the corollary yield the probabilities $P\{S_N=k\}$.

3.6. The Negative Hypergeometric Distribution

Suppose that N has the distribution of the number of blue balls chosen before a total of r white balls have been amassed when balls are randomly removed from an urn containing w white and b blue balls; that is,

$$P\{N=n\}=\frac{\binom{b}{n}\binom{w}{r-1}}{\binom{w+b}{n+r-1}}\,\frac{w-r+1}{w+b-n-r+1}.$$

https://doi.org/10.1017/S0269964804184039 Published online by Cambridge University Press

Using $E[N] = rb/(w+1)$, we obtain

$$P\{M-1=n\} = \frac{\binom{sb-1}{n}\binom{w+1}{r}}{\binom{w+b}{n+r}};$$

that is, $M-1$ has a hypergeometric distribution, implying that the probabilities $P\{S_{M-1}=j\}$ can be obtained from the recursion of Subsection 3.4. Applying the corollary then gives the probabilities $P\{S_N=k\}$.

4. ESTIMATING $P\{S \le c\}$

The raw simulation approach to estimate $p = P\{S \le c\}$ would first generate the value of N, say $N=n$, then generate the values of $X_1,\ldots,X_n$ and use them to determine the value of the raw simulation estimator:

$$I = \begin{cases} 1 & \text{if } \sum_{i=1}^{N} X_i \le c \\ 0 & \text{otherwise.} \end{cases} \tag{1}$$

The average value of I over many such runs would then be the estimator of p.

We can improve upon the preceding by a conditional expectation approach that starts by generating the values of the X_i in sequence, stopping when the sum of the generated values exceeds c. Let M denote the number that are needed; that is,

$$M = \min\left(n : \sum_{i=1}^{n} X_i > c\right).$$

If the generated value of M is m, then we use $P\{N<m\}$ as the estimate of p from this run. To see that this results in an estimator having a smaller variance than does the raw simulation estimator I, note that because the X_i are positive,

$$I = 1 \Leftrightarrow N < M.$$

Hence,

$$E[I|M] = P\{N<M|M\}. \tag{2}$$

Now,

$$\begin{aligned} P\{N<M|M=m\} &= P\{N<m|M=m\} \\ &= P\{N<m\}, \end{aligned}$$

where the final equality used the independence of N and M. Consequently, if the value of M obtained from the simulation is $M=m$, then the value of $E[I|M]$ obtained is $P\{N<m\}$.

https://doi.org/10.1017/S0269964804184039 Published online by Cambridge University Press

The preceding conditional expectation estimator can be further improved by using a control variable. Let $\mu = E[X_i]$, and define the zero mean random variable

$$Y = \sum_{i=1}^{M} (X_i - \mu). \tag{3}$$

Because Y and the conditional expectation estimator $P\{N < M|M\}$ are (strongly) negatively correlated, Y should make an effective control variable.

4.1. Improving the Conditional Expectation Estimator

Let M be defined as earlier and write

$$P\{S \le c\} = \sum_j P\{M > j\}P\{N = j\}.$$

The conditional expectation estimator is obtained from the preceding by generating M and using $I\{M > j\}$ as the estimator of $P\{M > j\}$.

We now show how to obtain a more efficient simulation estimator of $P\{M > j\}$. Let F denote the distribution function of X_i and write

$$P\{M > j\} = P\{M > j|X_1 \le c\}F(c).$$

If we now simulate X_1 conditional on the event that it is less than or equal to c, then for this value of X_1, the estimator

$$P\{M > j|X_1\}F(c)$$

is an unbiased estimator of $P\{M > j\}$ having a smaller variance than $I\{M > j\}$. Let $x_1 \le c$ be the generated value. For $j > 1$, we have

$$P\{M > j|X_1 = x_1\}F(c) = P\{M > j|X_1 = x_1, X_2 \le c - x_1\}F(c - x_1)F(c).$$

Hence, generating X_2 conditional on the event that $X_2 \le c - x_1$ gives, when this generated value is x_2, the estimate

$$P\{M > j|X_1 = x_1, X_2 = x_2\}F(c - x_1)F(c).$$

By continuing in this manner it follows that we can obtain, for any desired value n, estimates of $P\{M > j\}, j = 1, \ldots, n$. We can then obtain estimators of the probabilities $P\{M > j\}, j > n$, by switching to an ordinary simulation. With e_j denoting the estimator of $P\{M > j\}$, we obtain their values as follows.

1. $e_0 = 1, s = 0$.
2. $I = 1$.
3. $e_I = F(c - s)e_{I-1}$.
4. Generate X conditional on $X \le c - s$. Let its value be $X = x$.
5. $s \to s + x, I \to I + 1$.

https://doi.org/10.1017/S0269964804184039 Published online by Cambridge University Press

6. If $I \leq n$, go to 3.
7. Generate $X_1, \ldots$ until their sum exceeds $c - s$. Let R denote the number needed; that is,

$$R = \min\{k : X_1 + \cdots + X_k > c - s\}.$$

8. $e_{n+k} = e_n I\{R > k\}$, $k \geq 1$.

The estimator of $P\{S \leq c\}$ from this run is

$$\text{EST} = \sum_j e_j P\{N = j\} \tag{4}$$

and its average over many runs is the overall estimate.

4.2. A Simulation Experiment

In this subsection, we give the numerical results of a simulation study done to evaluate the performance of the techniques 1–4. We let X_i be independent and identically distributed (i.i.d.) uniform (0,1) random variables and let N be Poisson, having mean 10. Table 1 summarizes the standard deviations of the estimators for different values of c. Ten thousand replications were done for each value of c to estimate $P(\sum_{i=1}^{N} X_i \leq c)$. Technique 1 is the raw simulation method; technique 2 is the conditional expectation method; technique 3 is the conditional expectation method along with the control variable (3); technique 4 uses the estimator (4). The raw estimator (technique 1), as expected, performs poorly and the other estimators perform much better.

Next, we let X_i be i.i.d. exponential random variables with mean 1, and, again, let N be Poisson, having mean 10. Table 2 summarizes the standard deviations of the estimators for different values of c. Ten thousand replications were done for each value of c to estimate $P(\sum_{i=1}^{N} X_i \leq c)$.

Table 1. Mean and Standard Deviations of the Estimators for Different Values of c

c		Technique 1	Technique 2	Technique 3	Technique 4
5	Mean	0.5293	0.6333901	0.6332048	0.6332126
	SD	0.4991657	0.1828919	0.06064626	0.1654034
7	Mean	0.8575	0.9096527	0.909443	0.9094954
	SD	0.3495797	0.08442904	0.04381291	0.07792921
10	Mean	0.9924	0.9960367	0.9960445	0.9960576
	SD	0.08685041	0.007884455	0.006071804	0.007213068
15	Mean	0.99999	0.9999976	0.9999977	0.9999977
	SD	0.01	0.00001479114	0.00001409331	0.00001230383

https://doi.org/10.1017/S0269964804184039 Published online by Cambridge University Press

TABLE 2. Mean and Standard Deviations of the Estimators for Different Values of c

c		Technique 1	Technique 2	Technique 4
15	Mean	0.8638	0.9052721	0.9054672
	SD	0.343018	0.1489983	0.1284144
20	Mean	0.9739	0.983253	0.9832735
	SD	0.1594407	0.05296109	0.04602155
25	Mean	0.9976	0.9977583	0.9977649
	SD	0.04893342	0.01600277	0.01322325
30	Mean	0.9996	0.999749	0.9997479
	SD	0.019997	0.003714619	0.003133148

Thus, based on this small experiment, it appears that the reduction in variance effected by technique 4 over technique 2 is not worth the additional time that it takes to do a simulation run. Moreover technique 3, which does not require much more additional time than either technique 1 or technique 2, usually gives an even smaller variance than technique 4.

5. ESTIMATING $\theta = E[(S - c)^+]$

Start by letting $S_j = \sum_{i=1}^{j} X_i$ and note that

$$\theta = E\left[\sum_j (S_j - c)^+ P\{N = j\}\right].$$

To estimate θ, follow the procedure of (2) and generate the sequence $X_1, \ldots$, stopping at

$$M = \min(j : S_j > c).$$

Let

$$A = S_M - c$$

and use the estimator

$$\begin{aligned} E\left[\sum_j (S_j - c)^+ P\{N = j\} | M, A\right] &= \sum_{j \geq M} (A + (j - M)E[X])P\{N = j\} \\ &= (A - ME[X]) \sum_{j \geq M} P\{N = j\} \\ &\quad + E[X]\left(E[N] - \sum_{j<M} jP\{N = j\}\right); \end{aligned}$$

https://doi.org/10.1017/S0269964804184039 Published online by Cambridge University Press

that is, if the generated values of M and A are m and a, then the estimate of θ from that run is

$$(a - mE[X])P\{N \geq m\} + E[X]\left(E[N] - \sum_{j<m} jP\{N = j\}\right).$$

6. WHEN THE X_i ARE UNCONSTRAINED IN SIGN

When the X_i are not required to be positive, our previous methods no longer apply. We now present an approach in the general case. To estimate p, note that for a specified integer r,

$$P\{S \leq c\} = \sum_{j=0}^{r} P\{S_j \leq c\}P\{N = j\} + P\{S \leq c | N > r\}P\{N > r\}.$$

Our approach is to choose a value r and generate the value of N conditional on it exceeding r; if this generated value is g, then simulate the values of $S_1, \ldots, S_r$ and S_g. The estimate of p from this run is

$$\hat{p} = \sum_{j=0}^{r} I(S_j \leq c)P\{N = j\} + I(S_g \leq c)P\{N > r\}.$$

The larger the value of r chosen, the smaller the variance of this estimator. (When $r = 0$, it reduces to the raw simulation estimator.)

Similarly, we can estimate θ by using

$$\theta = \sum_{j=0}^{r} E[(S_j - c)^+]P\{N = j\} + E[(S_j - c)^+ | N > r]P\{N > r\}.$$

Hence, using the same data generated to estimate p, the estimate of θ is

$$\hat{\theta} = \sum_{j=0}^{r} (S_j - c)^+ P\{N = j\} + (S_g - c)^+ P\{N > r\}.$$

Acknowledgment

This research was supported by the National Science Foundation grant ECS-0224779 with the University of California.

References

1. Chan, B. (1982). Recursive formulas for discrete distributions. *Insurance: Mathematics and Economics* 1(4): 241–243.
2. Panjer, H.H. (1981). Recursive evaluation of a family of compound distributions. *ASTIN Bulletin* 12: 22–26.
3. Panjer, H.H. & Willmot, G.E. (1982). Recursions for compound distributions. *ASTIN Bulletin* 13: 1–11.
4. Ross, S. (2002). *SIMULATION*, 3rd ed., San Diego, CA: Academic Press.
5. Schroter, K.J. (1990). On a family of counting distributions and recursions for related compound distributions. *Scandinavian Actuarial Journal* 3/4: 161–175.

6. Sundt, B. (1992). On some extensions of Panjer's class of counting distributions. *ASTIN Bulletin* 22: 61–80.
7. Sundt, B. & Jewell, W.S. (1981). Further results on a recursive evaluation of compound distributions. *ASTIN Bulletin* 12: 27–39.
8. Willmot, G.E. (1993). On recursive evaluation of mixed Poisson probabilities and related quantities. *Scandinavian Actuarial Journal* 2: 114–133.
9. Willmot, G.E. & Panjer, H.H. (1987). Difference equation approaches in evaluation of compound distributions. *Insurance: Mathematics and Economics* 6: 43–56.