Advances in Science and Technology Research Journal
Volume 8, No. 23, Sept. 2014, pages 88–93
DOI: 10.12913/22998624.1120328

Research Article

APPROXIMATION OF RANDOM SUMS OF RANDOM VARIABLES IN INSURANCE

Paweł Wlaź[1]

[1] Department of Applied Mathematics, Lublin University of Technology, Nadbystrzycka 38, 20-618 Lublin, Poland, email: p.wlaz@pollub.pl

Received: 2014.07.19
Accepted: 2014.08.18
Published: 2014.09.09

ABSTRACT
The paper deals with approximations of random sums. By random sum we mean a sum of random number of independent and identically distributed random variables. Distribution of this sum is called a compound distribution. The model is especially important in non-life insurance. There are many methods for approximating compound distributions, one of the most popular one is approximation with shifted gamma distribution. In this work we show an alternative way – using kernel density, Fast Fourier Transform and numerical optimization methods – for finding shifted gamma approximations and show results suggesting its superiority over classical method.

Keywords: random sums, Fast Fourier Transform, shifted gamma distribution.

INTRODUCTION

One of main problems in actuary science, which can be very clearly interpreted in the area of insurance, is approximation of random sum of random variables [2, 8]. Formally speaking, let:

$$Y_1, Y_2, Y_2, \ldots$$

be a sequence of identically distributed random variables, with distribution function F_Y and let N be a random variable taking values in the set {0, 1, 2, ...}. Moreover, assume that:

$$N, Y_1, Y_2, Y_2, \ldots$$

are defined on the same probability space and are independent. Then we define sum X as:

$$X = \sum_{i=1}^{N} Y_i \tag{1}$$

with the convention that the sum equals 0 whenever $N = 0$.

Now if we think of each Y_i as a possible claim in some insurance framework, and if we treat N as a number of claims in a given time interval, then the distribution of X is the distribution of total claims.

Throughout the rest of the paper, we will assume, that Y is continuous and concentrated on $(0, \infty)$. In that case, X has the same properties.

By distribution function F_Z of a random variable Z we mean a function, that for any real x is calculated as:

$$F_Z(x) = \Pr(Z \le x)$$

It is clear, that F_X, for X described by (1), can be expressed in terms of F_Y and F_N, namely

$$F_X(x) = \sum_{k=0}^{\infty} \big(F_N(k) - F_N(k-1)\big) \cdot F_Y^{*k}(x) \tag{2}$$

where:

$$F_Y^{*k}(x) = \int_0^x F_Y^{*(k-1)}(x-t)\, dF_Y(t) \tag{3}$$

for $k > 0$ and F^{*0} is the distribution of random variable taking value 0 with probability 1.

Formulae (2) and (3) show why it is not in general possible to obtain F_X analytically, even if one has both F_N and F_Y. The real-life situation is even worse, because one has only samples, not F_Y.

Knowing a good approximation of F_X is crucial for insurer calculating optimal premium [12]. Distribution of X is known as *compound distribution*.

PROPERTIES OF COMPOUND DISTRIBUTION

It is generally accepted to assume F_N is known – or at least known is the class of distributions from which F_N is taken [9].

There are many factors suggesting such an assumption; one of them is theoretically and experimentally confirmed property of Poisson distribution as a distribution for "rare" cases, what ideally fits the situation in non-life insurance.

So in this paper we will assume, that N has Poisson distribution. Poisson distribution has only one parameter λ, which can be approximated by average number of claims in unit of time, that is why in practice we can think of F_N as known distribution.

Basic parameters of compound distributions, when N is Poisson with parameter λ are easy to obtain, if moments of F_1 are known. Namely [8]

$$EX = \lambda E Y_1, \tag{4}$$

$$\sigma^2(X) = \lambda \cdot E(Y_1^2), \tag{5}$$

$$\gamma_1(X) = \lambda \cdot E(Y_1^3)/\sigma^3(X), \tag{6}$$

$$\gamma_2(X) = \lambda \cdot E(Y_1^4)/\sigma^4(X), \tag{7}$$

where: γ_1 is skewness and γ_2 kurtosis of random variable.

Example 1. Consider Y_i having continuous uniform distribution in (0, 1) and let N be Poisson with $\lambda = 15$. On Figure 1 we have the density of compound distribution. Details of calculating density functions for compound distributions are given later. We can see that in spite of symmetry of all Y_i, the distribution of X has positive skewness, which is implied by positive skewness of Poisson distribution. In insurance practice usually also Y_i are positively skewed, so in general we expect strong skewness of the distribution of X. In this example the skewness of compound distribution is about 0.33541, whilst $\gamma_1(Y_1) = 0$, $\gamma_1(N) = 1/\sqrt{\lambda} \approx 0.258199$.

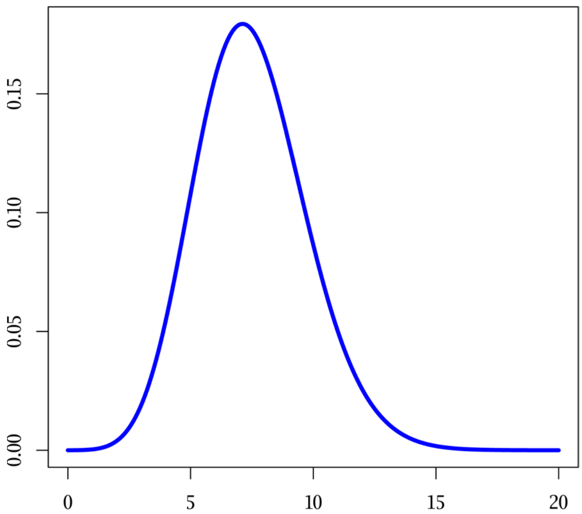

Fig. 1. Density function of compound distribution from example 1

APPROXIMATION WITH SHIFTED GAMMA DISTRIBUTION

Having samples of claims one can easily estimate moments of Y_1 and λ. Together with (4)–(7) basic parameters of X are to be estimated with no effort. Unlike in many other situations, where one-modal continuous distributions are approximated with normal distribution, here normal distribution is not suitable, since normal random variables have skewness equal zero.

To obtain a useful distribution with the same three first moments as X (and therefore the same expected value, standard deviation and skewness) one has to take simple three parametric one-modal distribution and it is common to consider shifted gamma distribution [8].

A random variable is said to have a shifted gamma distribution, if there are real parameters α, β, x_0, for which (Z– x_0) has standard gamma distribution with parameters α, β. In other words (Z– x_0) has density function:

$$f(x) = \begin{cases} \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x} & \text{for } x > 0 \\ 0 & \text{for } x \leq 0. \end{cases}$$

The solution of equations leading to equality of first three moments of distribution of and shifted gamma distributions is as follows:

$$\alpha = \frac{4}{\left(\gamma_1(X)\right)^2} \tag{8}$$

$$\beta = \frac{2}{\gamma_1(X)\sigma(X)} \tag{9}$$

$$x_0 = EX - \frac{2\sigma(X)}{\gamma_1(X)} \tag{10}$$

However, there are many situations when this approximation fails. The main problem is that kurtosis of such an approximation always equals $1.5 \cdot (\gamma_1(X))^2$ which can be very far from real data, moreover, for too big value of skewness gamma distribution does not fit typical compound distribution. Therefore, it is suggested (see [4]) that the approximation with shifted gamma distribution can be applied if:

$$\gamma_1(X) < 1, \tag{11}$$

$$\frac{\gamma_2(X)}{\left(\gamma_1(X)\right)^2} \in \langle 1.2, 1.8\rangle \tag{12}$$

The trouble with the above criteria is that it is very hard to estimate $\gamma_2(X)$, since according to (4)–(7) we have to estimate $E(Y_1^4)$ which – for skewed distributions – cannot be reliably approximated unless sample is extremely big.

Example 2. Consider Y_i log-normal with parameters $\mu = 0$ and $\sigma = 1$, that is $\ln Y_1$ is normal with mean 0 and standard deviation 1. Since for k-th moment of log-normal distribution we have a formula:

$$e^{k\mu+\frac{1}{2}k^2\sigma^2}$$

then we calculate $E(Y_1^4) \approx 2980.96$.

Now generate – in statistical **R** software (see [10]) – 1000 samples from log-normal distribution with:

```
> n = 1000; mu = 0; sigma = 1
> set.seed(1) # to reproduce later
> samples = rlnorm(n, meanlog = mu,
+ sdlog = sigma)
> mean(samples^4)
[1] 4706.422
```

The obtained estimation of 4-th moment is much bigger than the true value. And for another sample of the same size from the same distribution we have:

```
> set.seed(2) # to reproduce later
> samples = rlnorm(n, meanlog = mu,
+ sdlog = sigma)
> mean(samples^4)
[1] 791.0706
```

and the obtained value is much smaller than the true 4-th moment. Thus we see that even a reasonably big sample may fool us about 4-th moment of distribution.

CALCULATING DENSITY OF COMPOUND DISTRIBUTION WITH FAST FOURIER TRANSFORM

If the distribution of Y_1 is known, then there exist – using nowadays computer technology – quite an easy way to obtain an approximation of the density function of X. Namely, if φ_Z is the characteristic function of Z, that is $\varphi_Z(t) = E(\exp(itZ))$ (where i is the imaginary unit) and P_N is the probability generating function, that is $P_N(x) = E(x^N)$ for N taking only non-negative integer values, then, in terms of previous sections:

$$\varphi_X(t) = P_N\left(\varphi_{Y_1}(t)\right) \tag{13}$$

The equation (13) follows immediately from the fact that characteristic function of sum of independent variables equals the product of characteristic functions of the variables [1].

If we look at the characteristic function as inverse Fourier transform of density function, and if we look at discrete Fourier transform as an approximation of Fourier transform then we obtain the following algorithm (pioneered in [6]), which utilizes computational power of Fast Fourier Transform algorithm [3]:

1. Choose small $h > 0$ and large integer M, calculate vector $\boldsymbol{y} = [y_0, y_1, ..., y_{M-1}]$ being discretization of density function of Y_1 on interval $(0, M \cdot h)$.
2. Calculate $\hat{\boldsymbol{y}} = \texttt{fft}(\boldsymbol{y})$.
3. Calculate $\hat{\boldsymbol{x}} = P_N(\hat{\boldsymbol{y}})$.
4. Calculate $\boldsymbol{x} = \texttt{ifft}(\hat{\boldsymbol{x}})$.

Vector $\boldsymbol{x} = [x_0, x_1, ..., x_{M-1}]$ is an approximation of probability function of discretized version of X. Terms *fft* and *ifft* used here are acronyms for *Fast Fourier Transform* and *Inverse Fast Fourier Transform*, respectively.

Due to computational effectiveness of both *fft* and *ifft*, integer may be as large as millions, what allows for very small and reduces discretization errors of the procedure. Using some additional **R** packages (like [5] for discretization) and knowing that P_N has simple formula:

$$P_N(t) = e^{\lambda(t-1)}$$

for N Poisson with parameter λ, the whole procedure described above may be scripted as one-liner in **R** language.

Example 3. In this example we illustrate effectiveness of approximation with shifted gamma distribution. Say, we have n = 200 samples from log-normal distribution with parameters $\mu = 0$ and $\sigma = 1$, that is $\ln Y_1$ is normal with mean 0 and standard deviation 1.

Let us generate sample with:

```
> n = 200; mu = 0; sigma = 1
> set.seed(1)
> samples = rlnorm(n, meanlog = mu,
+ sdlog = sigma)
```

Let us assume for a moment that we have only these data. Let us assume that from some other considerations we fixed the frequency of claims parameter λ equal 15.

Having samples of claims and parameter λ we can of course calculate estimations of $E(Y_1^k)$, for k = 1, 2, 3, 4 and – using (4)–(7) – we can find estimations of EX, $y_1(X)$ and $y_2(X)$. Then, with equations (8)–(10) we find $\alpha \approx 9.3403$, $\beta \approx 0.32337$ and $x_0 \approx -4.6347$.

Moreover, left side of (11) is about 0.654 and left side of (12) is about 1.27, so both (11) and (12) are fulfilled.

Now suppose we take $p = 0.95$ and we want to know what p-quantile of unknown distribution of X is. This is a common task, as it estimates with high probability maximum of total claims in given period. Of course it is to be found as p-quantile of shifted gamma distribution. In this example it equals about 41.345.

How does the latter figure compare to the p-quantile of real distribution of X? If we knew what is the distribution of Y_1, we could calculate the distribution of X – with *fft* and *ifft*, and then of course calculate p-quantile which turns out to be about 43.905 so relative error of our estimation is about 5.8.

Well, that is only one p-quantile, what about the whole distribution? As usual, an image might be better than numbers. On Figure 2 one can see the density of compound distribution (the blue line, calculated with *fft*, knowing the distribution of Y_1) and the density estimated with shifted gamma distribution based on samples (the red, dashed line).

Of course one set of samples is not enough, one might suspect some sort of coincidence. We repeat experiment analogous to example 3 – for seeds equal 1, 2, ... until we find situations which fit (11)–(12) (during the process about 1/5 sets are rejected) – and calculate mean relative error that we have for different tail p-quantiles (and standard deviations of these errors). We use tail p-quantiles, (p close to 1) as these are the most important in the procedure of premium calculations. The results are summarized in Table 1.

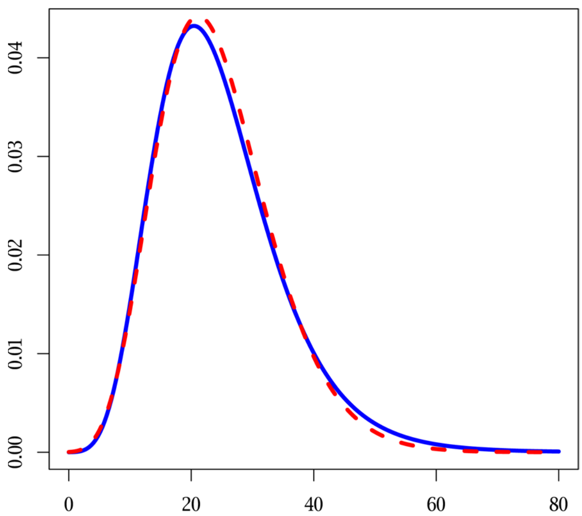

Fig. 2. Comparison of true compound density (blue, solid) with its shifted gamma approximation (red, dashed) based on sample, see example 3

Table 1. Relative errors obtained when p-quantiles of true compound distribution are replaced by their approximations with shifted gamma distribution based on 200-element samples, with (8)–(10). Samples are from log-normal distribution with $\mu = 0$ and $\sigma = 1$, for Poisson with $\lambda = 15$, averages are calculated after considering 1000 situations which fit (11)–(12)

p	Average relative error [%]	St. dev. of relative errors [%]
0.90	8.1	5.8
0.91	8.2	5.9
0.92	8.3	5.9
0.93	8.4	6.0
0.94	8.6	6.1
0.95	8.8	6.2
0.96	9.1	6.3
0.97	9.5	6.5
0.98	10.1	6.8
0.99	11.3	7.4

NON-CLASSICAL SHIFTED GAMMA APPROXIMATION

In this section an alternative method for finding approximation of compound distribution with shifted gamma distribution will be described. The motivation for this method may be found in the following example.

Example 4. Let us assume Y_i are all log-normal with $\mu = 0$ and $\sigma = 1.1$. Let N be Poisson with $\lambda = 15$. The density of X (calculated with *fft*) is illustrated on Figure 3 – solid blue line (partly shadowed by green line).

After calculating EX, $\sigma(X)$ and $y_1(X)$ we can find $\alpha \approx 1.59097$, $\beta \approx 0.0971156$, and $x_0 \approx 11.0866$ for approximation with shifted gamma distribution. The density of shifted gamma distribution for these parameters is pictured with dashed red line on Figure 3. As we see, the lines are far from fitting.

Now look at the green solid line at the same Figure 3. That also is the density of gamma dis-

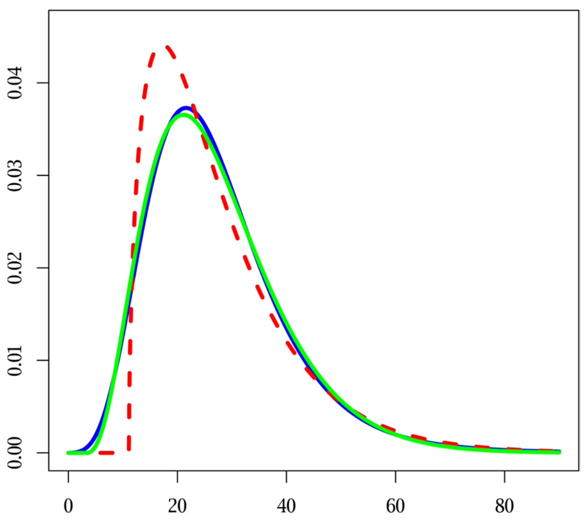

Fig. 3. Motivation for non-classical shifted gamma approximation: blue solid line is the density of compound distribution (Y_i log-normal with $\mu = 0$ and $\sigma = 1.2$, N – Poisson with $\lambda = 15$); red dashed line is the density of shifted gamma approximation obtained with (8)–(10); green solid line is another shifted gamma, apparently better fitting than the red one

tribution, this time with parameters $\alpha \approx 3.88213$, $\beta \approx 0.160105$, and $x_0 \approx 3.08835$. This green line is very close to the line of true density of X.

In the example 4 it was not very hard to obtain better than classical shifted gamma approximation, since we know the exact distribution of X (due to *fft* calculations) and we can minimize maximum of absolute value of differences between distribution functions.

In the real insurance work there is one more problem – we do not have distribution of Y_i, only samples of values of Y_i. The algorithm that we propose in this work is as follows:

1. Find approximation of density of Y_i, basing on samples, let F be distribution function corresponding to that density.
2. Basing on density from step 1, find approximation of distribution of X.
3. If by F_{α,β,x_0} we denote the distribution function of shifted gamma distribution with shape parameter α, rate parameter β and shift parameter x_0, then find $\hat{\alpha}$, $\hat{\beta}$ and $\hat{x}_0$ minimizing:

$$\max_{x \in \mathbb{R}} \left| F(x) - F_{\alpha,\beta,x_0}(x) \right| . \tag{14}$$

And here are some explanations concerning the above three steps.

Density based on samples

To find an approximation of unknown distribution of claim size Y_i based on samples y_1, y_2, ..., y_n we use Gaussian kernel density estimation which approximates unknown density with mean of normal densities $N(\mu = y_i, \sigma = h)$, where parameter h is selected according to method of Sheather and Jones [11].

Compound distribution based on density of claim size

In this step – as it was mentioned at the beginning of the paper – we assume that frequency of claims is Poisson with given parameter λ, and having calculated density of claim size we then proceed with the algorithm using *fft*.

Optimizing parameters

To minimize (14) we use general-purpose Nelder-Mead method [7]. This method needs starting point (in this case three-dimensional). Our proposition is to take a few starting points lying between parameters of classical shifted gamma distribution and parameters of gamma distribution with no shift (taking $x_0 = 0$ – based only on mean and standard deviation) and choose the one giving best result.

Example 5. In this example we fix $\mu = 0$ and $\sigma = 1$, generate $n = 200$ claims from log-normal distribution. If conditions (11)–(12) are not met, repeat generation. After successful generation, taking $\lambda = 15$ as frequency of claims we calculate exact distribution of X, classical shifted gamma distribution based on samples and non-classical gamma distribution based on samples. We then calculate chosen p-quantiles and calculate relative errors. The above procedure is repeated times and average relative errors and their standard deviations are obtained. The results are summarized in Table 2.

CONCLUSIONS

The presented calculations and illustrations show that it makes sense to use an alternative way to find shifted gamma approximations of compound distributions. Of course our novel method is much more cumbersome than the classical method, but calculating premium in real environment does not have to be done in seconds.

The method still needs much more evidence, especially its usefulness for other types of distributions used for modelling claim size should be checked.

Table 2. Classical and non-classical shifted gamma approximation. Comparison of mean relative errors obtained for different p-quantiles and standard deviations of relative errors, 1000 trials described in example 5

p	Average relative error [%]		St. dev. of relative errors [%]	
	classical	non-classical	classical	non-classical
0.90	9.2	9.0	6.4	6.5
0.91	9.4	9.1	6.5	6.6
0.92	9.6	9.2	6.7	6.6
0.93	9.8	9.3	6.8	6.7
0.94	10.1	9.5	6.9	6.7
0.95	10.5	9.7	7.1	6.8
0.96	11.0	10.0	7.3	7.0
0.97	11.7	10.5	7.6	7.2
0.98	12.9	11.2	8.1	7.6
0.99	15.2	12.9	8.8	8.4

REFERENCES

1. Billingsley P. Probability and Measure. 3rd ed. John Wiley & Sons, 1995.
2. Bowers N. et al. Actuarial Mathematics. Society of Actuares, Itasca, Illinois, 1986.
3. Cooley J. and Tukey J. An algorithm for the machine calculation of complex Fourier series. Math. Comput. 19, 1965, 297–301.
4. Daykin D., Pentikainen T., and Pesonen E. Practical Risk Theory for Actuaries. Chapman and Hall, 1994.
5. Dutang C., Goulet V., and Pigeon M. Actuar: An R Package for Actuarial Science. Journal of Statistical Software 25(7), 2008, p. 38.
6. Heckman P. and Meyers G.. The calculation of aggregate loss distributions from claim severity and claim count distributions. Proc. Casualty Actuar. Soc. LXX, 1983, 22–61.
7. Nelder J.A. and Mead R. A simplex algorithm for function minimization". Computer Journal 7, 1965, 308–313.
8. Otto W. Ubezpieczenia majątkowe. WNT, Warszawa 2004.
9. Pitts S. "Nonparametric estimation of compound distributions with application in insurance". Ann. Inst. Statist. Math. 46(3), 1994, 537–555.
10. R Core Team. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing. Vienna, Austria, 2013. url: http://www.R-project.org/.
11. Sheather S.J. and Jones M.C. A reliable data-based bandwidth selection method for kernel density estimation. Journal of the Royal Statistical Society, B 53, 1991, 683–690.
12. Willekens E. Asymptotic approximations of compound distributions and some applications. Bull. Soc. Math. Belg. Ser. B 41, 1989, 51–61.