The Korean Journal of Applied Statistics (2017) DOI: https://doi.org/10.5351/KJAS.2017.30.3.363
30(3), 363–376

New composite distributions for insurance claim sizes

Daehyeon Jung[a] · Jiyeon Lee[a,1]

[a]Department of Statistics, Yeungnam University

(Received February 16, 2017; Revised March 31, 2017; Accepted March 31, 2017)

Abstract

The insurance market is saturated and its growth engine is exhausted; consequently, the insurance industry is now in a low growth period with insurance companies that face a fierce competitive environment. In such a situation, it will be an important issue to find the probability distributions that can explain the flow of insurance claims, which are the basis of the actuarial calculation of the insurance product. Insurance claims are generally known to be well fitted by lognormal distributions or Pareto distributions biased to the left with a thick tail. In recent years, skew normal distributions or skew t distributions have been considered reasonable distributions for describing insurance claims. Cooray and Ananda (2005) proposed a composite lognormal-Pareto distribution that has the advantages of both lognormal and Pareto distributions and they also showed the composite distribution has a higher fitness than single distributions. In this paper, we introduce new composite distributions based on skew normal distributions or skew t distributions and apply them to Danish fire insurance claim data and US indemnity loss data to compare their performance with the other composite distributions and single distributions.

Keywords: composite distribution, skew normal distribution, skew t distribution, Danish fire insurance claim data, US indemnity loss data, maximum likelihood estimator, AIC

1. 서론

산업화가 진행되면서 삶의 편의는 꾸준히 높아지고 있지만 그에 따라 불확실성도 높아진 것이 사실이다. 이러한 불확실성을 보완하는 보험산업 역시 성장세를 유지하고 있지만 최근 보험산업의 동향을 보면 이러한 성장세가 계속 이어질지에 대해서는 우려되는 시점이다.

Yun 등 (2015)에 따르면 보험산업의 경우 고령화와 저출산 등과 같은 인구사회적인 요소로 인해 퇴직연금 등 노령인구들을 대상으로 하는 보험 상품을 제외한 다른 보험상품들의 대부분이 성장 동력을 소진하여 명목 경제성장률 수준으로 성장세가 낮아졌다고 한다. 이러한 가운데 판매채널 위주의 양적 경쟁에서 상품서비스 위주의 질적 경쟁 체제로의 전환을 꾀하는 등의 보험산업 경쟁력 강화가 필요하다. 그 일환으로 보험상품의 질을 향상시키면서 동시에 성장 동력을 마련하기 위해서는 보험계리적 계산의 기초가 되는 보험 청구액의 흐름을 잘 설명할 수 있는 분포를 찾아내는 것이 중요하다.

Lee와 Park (2012)은 보험의 특성상 규모가 큰 사건은 청구액이 많은 반면 일어나는 빈도가 낮고 규모가 작은 사건에 대해서는 청구액이 작은 반면에 발생하는 빈도가 높다는 것을 지적하였다. 이러한 특성

[1]Corresponding author: Department of Statistics, Yeungnam University, 280 Daehak-Ro, Gyeongsan, Gyeongbuk 38541, Korea. E-mail: leejy@yu.ac.kr

을 반영할 수 있는 분포들로는 감마(gamma) 분포, 와이블(Weibull) 분포, 로그정규(lognormal) 분포, 파레토(Pareto) 분포 등이 있으며 특히 로그정규분포나 파레토 분포 등이 보험 청구액 데이터에 잘 적합하는 것으로 알려져 있다 (Burnecki 등, 2000).

Azzalini와 Capitanio (2003)은 일반 t 분포의 대칭성(symmetry)을 요란시켜 왜도(skewness)를 가지는 기운 t 분포(skew t distribution)의 특성을 소개하였고, Eling (2012)은 기운 t 분포가 보험 청구액 데이터에 좋은 적합도를 가지는 것을 확인하였다. 한편 McNeil (1997)은 보험 청구액이 적은 경우가 빈번하면서 상대적으로 꼬리가 짧은 경우에는 로그정규분포가 더 적절하고, 파레토 분포는 오른쪽 꼬리가 매우 긴 형태를 가지는 보험 청구액 분포에 더 적절하다는 것을 지적하였다. Cooray와 Ananda (2005)는 앞쪽 머리 부분은 로그정규분포를 따르고 뒷쪽 꼬리 부분은 파레토 분포를 따르는 복합분포(composite distribution)를 소개하면서 McNeil (1997)이 지적한 보험 청구액의 분포 특성을 동시에 반영할 수 있게 하였다. Preda와 Ciumara (2006)은 머리 부분이 와이블 분포를 따르는 와이블-파레토 복합분포를 소개하고 로그정규-파레토 복합분포와 비슷한 특성을 가짐을 보였다. Pigeon과 Denuit (2011)은 두 분포가 결합되는 지점인 경계값(threshold)을 복합분포의 모수로 포함하여 모형을 확장시켰다.

본 논문에서는 단일분포로서 적합도가 높았던 기운 t 분포를 머리 부분으로 이용하여 꼬리 부분의 다른 단일분포와 결합한 다양한 복합분포를 소개하고 선행연구에서 많이 이용되었던 덴마크의 화재보험 청구액 데이터와 미국의 배상금 데이터에 적용하여 다른 청구액 분포와 비교한다. 2장에서는 Klugman 등 (2004)이 제시한 복합분포의 일반적인 형태와 본 논문에서 사용할 복합분포를 위한 몇 가지 조건에 대하여 살펴본다. 3장에서는 기운 t 분포의 특성을 알아보기 위하여 먼저 기운 정규분포(skew normal distribution)에 대하여 설명하고, 기운 t 분포로 확장한다. 또한 기운 정규 복합분포와 기운 t 복합분포의 확률밀도함수를 구하고, 기운 t 복합분포의 모수들의 최대우도추정량(maximum likelihood estimator; MLE)을 구하는 과정을 설명한다. 4장에서는 덴마크의 화재보험 청구액 데이터와 미국의 배상 지불금 데이터에 대해 기운 t 분포를 결합한 복합분포의 적합도가 다른 복합분포와 단일분포와 비교하여 높게 나타남을 확인한다. 5장에서는 본 연구의 성과와 추가적인 연구의 필요성을 제시한다.

2. 복합분포

Klugman 등 (2004)와 Nadarajah와 Bakar (2014)는 k개의 확률밀도함수 $f_1(x), f_2(x), \ldots, f_k(x)$로 구성된 복합분포 혹은 접합분포(spliced distribution)의 확률밀도함수 $f(x)$의 일반적인 형태를 다음과 같이 나타내었다.

$$f(x) = \begin{cases} a_1 f_1^*(x), & c_0 < x < c_1, \\ a_2 f_2^*(x), & c_1 < x < c_2, \\ \quad \vdots & \quad \vdots \\ a_k f_k^*(x), & c_{k-1} < x < c_k, \end{cases} \tag{2.1}$$

여기서 $f_i^*(x)$는 절단된(truncated) 확률밀도함수로서

$$f_i^*(x) = \frac{f_i(x)}{\int_{c_{i-1}}^{c_i} f_i(x)dx}, \quad i = 1, 2, \ldots, k$$

이고, a_i는 0 이상의 혼합 가중값(mixing weight)으로서 $\sum_i a_i = 1$을 만족하고, c_i는 정의역의 구간 경계값이다. 식 (2.1)의 특별한 경우인 2개의 확률밀도함수 $f_1(x)$와 $f_2(x)$가 각각 머리 부분과 꼬리 부분

으로 결합한 복합분포의 확률밀도함수는

$$f(x) = \begin{cases} a_1 f_1^*(x), & -\infty < x < \theta, \\ a_2 f_2^*(x), & \theta < x < \infty \end{cases} \tag{2.2}$$

로 나타낼 수 있으며, 0 이상의 혼합 가중값 a_1과 a_2는 $a_1 + a_2 = 1$이고 $f_i^*(x)$는

$$f_1^*(x) = \frac{f_1(x)}{F_1(\theta)} \quad \text{그리고} \quad f_2^*(x) = \frac{f_2(x)}{1 - F_2(\theta)}$$

가 된다. 여기서 $F_1(x)$과 $F_2(x)$는 확률밀도함수 $f_1(x)$과 $f_2(x)$의 각 분포함수를 나타낸다. Klugman 등 (2004)이 제시한 복합분포 식 (2.2)의 확률밀도함수는 일반적으로 연속성을 만족하지 않는다. 그래서 Bakar 등 (2015)는 식 (2.2)의 복합분포의 확률밀도함수가 연속(continuous)이며 미분가능(differentiable)하기 위해 다음의 조건을 제시하였다.

2.1. 연속성(continuity)

식 (2.2)의 복합분포의 확률밀도함수 $f(x)$의 연속성을 보장하기 위해서 다음의 조건

$$\lim_{x \to \theta-} f(x) = \lim_{x \to \theta+} f(x) \tag{2.3}$$

을 부여한다. 이로서 경계값 θ에서부터는 $f_2(\cdot)$가 복합분포의 꼬리분포가 된다.

2.2. 미분가능(differentiability)

확률밀도함수 $f_1(x)$와 $f_2(x)$가 모든 x에서 미분가능할 때, 경계값 θ에 대해서 다음의 조건을 만족하면 복합분포의 확률밀도함수 $f(x)$도 모든 x에서 미분가능해 진다.

$$\lim_{x \to \theta-} \frac{df(x)}{dx} = \lim_{x \to \theta+} \frac{df(x)}{dx}. \tag{2.4}$$

일반적으로 이 조건으로부터 결합하는 확률분포들의 모수들간의 관계식을 유도하여 복합분포의 모수를 축소하는 효과를 가진다. 이에 대한 설명은 2.4절에서 다룬다.

2.3. 혼합 가중값(mixing weight)

혼합 가중값 a_1과 a_2는 식 (2.3)의 연속성 조건에 의해

$$\delta := \frac{f_1(\theta)[1 - F_2(\theta)]}{f_2(\theta)F_1(\theta)} \tag{2.5}$$

를 이용하여

$$a_1 = \frac{1}{1+\delta} \quad \text{그리고} \quad a_2 = 1 - a_1 = \frac{\delta}{1+\delta}$$

로 나타낼 수 있다. Cooray와 Ananda (2005)는 처음 로그정규-파레토 복합분포를 소개하면서 혼합 가중값을 상수로 둠으로써 모수 2개의 로그정규분포와 모수 2개의 파레토 분포를 결합하여 모수 2개의 로그정규-파레토 복합분포를 생성하였다. Scollnik (2007)는 이를 확대하여 혼합 가중값을 결합하는 확률분포들의 모수들의 함수로 둠으로써 모수 2개의 로그정규분포와 모수 2개의 파레토 분포를 결합하여 모수 3개의 로그정규-파레토 복합분포를 생성하고, Cooray와 Ananda (2005)의 혼합 가중값들이 상수일 때보다 더 좋은 적합도를 가짐을 보였다. 여기서도 혼합 가중값은 식 (2.5)와 같이 확률분포들의 모수들의 함수로 나타낸다.

2.4. 경계값(threshold)

연속성 조건 (2.3)과 미분가능 조건 (2.4)에 의해 경계값 θ는 $f_1'(\theta)f_2(\theta) - f_1(\theta)f_2'(\theta) = 0$을 만족한다. 즉, 경계값 θ는 다음 방정식의 해가 된다 (Bakar 등, 2015).

$$\frac{d}{d\theta}\log\left[\frac{f_1(\theta)}{f_2(\theta)}\right] = 0. \tag{2.6}$$

일반적으로 경계값 θ는 모수로 취급하지만 복합분포에서는 식 (2.6)을 만족하는 θ 값이 결합하는 확률분포들의 다른 모수들에 의해 결정되는 특징이 있다. 즉, 복합분포의 혼합 가중값 δ와 경계값 θ는 비록 모수이지만 위의 관계식 (2.5)와 (2.6)을 통해 다른 모수들의 함수로 나타낼 수 있어서 다른 모수들을 먼저 추정한 후에 관계식을 통해 유도할 수 있다. 식 (2.6)을 만족하는 θ를 정확한 함수의 형태로 나타낼 수는 없지만 수치적 방법을 통해 근사적으로 구할 수 있다. 여기서는 뉴튼(Newton) 방법을 이용한 **R**의 `nlm()` 함수를 사용하여 θ를 계산한다. 즉, 경계값 θ는 식 (2.6)을 통해 다른 모수들의 함수로 나타낸 후, 각 모수들의 최대우도추정량을 구한 다음, 각 추정량 값을 식 (2.6)에 대입하여 `nlm()` 함수로 계산되는 근사 해를 θ 값으로 얻는다.

3. 기운 t 분포의 복합분포

이 장에서는 기운 t 분포와의 복합분포를 정의하기 위해 먼저 기운 정규분포를 설명하고 기운 t 분포로 확장한 다음, 복합분포를 구성하는 경우를 살펴본다.

3.1. 기운 정규분포

확률변수 X의 확률밀도함수 $f_X(x)$가

$$f_X(x) = 2\phi(x)\Phi(\alpha x)$$

일 때, 기운 표준정규분포(skew standard normal distribution)를 따른다고 하고 $X \sim SN(0,1,\alpha)$로 나타낸다. 여기서 $\phi(\cdot)$와 $\Phi(\cdot)$는 각각 표준정규분포의 확률밀도함수와 분포함수이고, 실수값 α는 왜도를 나타내는 모수이다 (Azzalini, 1985). 모수 α가 0이면 표준정규분포가 되고, $\alpha \to \pm\infty$이면 절반정규분포(half-normal distribiution)가 된다.

기운 정규분포는 기운 표준정규분포 X의 선형변환 $Y = \xi + \omega X$에 의해 얻어지며, $Y \sim SN(\xi, \omega^2, \alpha)$로 나타낸다. 여기서 $\xi, \omega > 0, \alpha$는 각각 위치(location), 척도(scale), 왜도에 의한 모양(shape)을 나타내는 모수들이다. 그러면 기운 정규확률변수 Y의 확률밀도함수와 분포함수는 각각

$$\begin{aligned} f_Y(x) &= \frac{2}{w}\phi\left(\frac{x-\xi}{\omega}\right)\Phi\left(\frac{\alpha(x-\xi)}{\omega}\right), \\ F_Y(x) &= \Phi\left(\frac{x-\xi}{\omega}\right) - 2T_o\left(\frac{x-\xi}{\omega}, \alpha\right) \end{aligned} \tag{3.1}$$

로 얻어진다. 여기서 $T_o(h,\alpha)$는 다음과 같이 정의되는 오웬(Owen)의 T 함수이다 (Owen, 1956).

$$T_o(h,\alpha) = \frac{1}{2\pi}\int_0^\alpha \frac{\exp\left[-\frac{1}{2}h^2\left(1+x^2\right)\right]}{1+x^2}dx.$$

기운 정규확률변수 Y의 평균과 분산은 각각

$$E[Y] = \xi + \omega\kappa\sqrt{\frac{2}{\pi}} \quad \text{그리고} \quad \mathrm{Var}[Y] = \omega^2\left(1 - \frac{2\kappa^2}{\pi}\right)$$

이다. 여기서 $\kappa := \alpha/\sqrt{1+\alpha^2} \in (-1,1)$이다 (Eling, 2012).

3.2. 기운 t 분포

Azzalini와 Capitanio (2003)는 기운 표준 t 분포(skew standard t distribution)를 다음의 변환을 통해 정의하였다.

$$W = \frac{X}{\sqrt{V/\nu}},$$

여기서 X는 기운 표준정규확률분포 $X \sim SN(0,1,\alpha)$이고 V는 자유도가 ν인 카이제곱 분포 $V \sim \chi^2(\nu)$이며 X와 V는 서로 독립이다. 모수 α가 0이면 자유도가 ν인 t 분포가 되고, $\alpha = 0$이고 $\nu \to \infty$이면 표준정규분포가 된다. 기운 t 분포(skew t distribution)는 기운 표준 t 분포 W의 선형변환 $T = \xi + \omega W$에 의해 얻어지며 $T \sim ST(\xi, \omega^2, \alpha)$로 나타낸다. 여기서 $\xi, \omega > 0, \alpha$는 각각 위치, 척도, 왜도에 의한 모양을 나타내는 모수들이다. 그러면 기운 t 분포의 확률밀도함수는

$$f_T(x) = \frac{2}{\omega} t\left(\frac{x-\xi}{\omega}, \nu\right) T\left(\frac{\alpha(x-\xi)}{\omega}\sqrt{\frac{\nu+1}{\nu+[(x-\xi)/\omega]^2}}, \nu+1\right) \tag{3.2}$$

로 얻어진다. 여기서 $t(\cdot\,,\nu)$와 $T(\cdot\,,\nu)$는 자유도가 ν인 t 분포의 확률밀도함수와 분포함수이다. 기운 t 확률변수 T의 평균과 분산은 각각

$$E[T] = \xi + \omega\tau\kappa \quad \text{그리고} \quad \mathrm{Var}[T] = \omega^2\left(\frac{\nu}{\nu-2} - \tau\kappa^2\right)$$

이다. 여기서 $\kappa := \alpha/\sqrt{1+\alpha^2}$이고 $\tau := \sqrt{\nu/\pi}\,\Gamma((\nu-1)/2)/\Gamma(\nu/2)$이다.

3.3. 복합분포

이번 절에서는 식 (2.2)에서 머리 부분의 $f_1(x)$가 기운 정규분포 또는 기운 t 분포의 확률밀도함수이고, 꼬리 부분의 $f_2(x)$는 다른 단일분포의 확률밀도함수일 때 복합분포를 구하는 경우를 소개한다.

3.3.1. 기운 정규 복합분포 $f_1(x)$가 식 (3.1)의 기운 정규분포의 확률밀도함수이고, $f_2(x)$와 $F_2(x)$가 다른 단일분포의 확률밀도함수와 분포함수일 때 두 분포의 복합분포를 구한다. 혼합 가중값을 위한 δ는 식 (2.5)를 사용하면

$$\delta = \frac{2\phi((\theta-\xi)/\omega)\Phi(\alpha(\theta-\xi)/\omega)[1-F_2(\theta)]}{\omega[\Phi((\theta-\xi)/\omega) - 2T_o((\theta-\xi)/\omega,\alpha)]f_2(\theta)}$$

로 얻어진다. 따라서 복합분포의 확률밀도함수 $f(x)$는 식 (2.2)와 (3.1)에 의해

$$f(x) = \begin{cases} \dfrac{2\phi((x-\xi)/\omega)\Phi(\alpha(x-\xi)/\omega)}{\omega(1+\delta)[\Phi((\theta-\xi)/\omega) - 2T_o((\theta-\xi)/\omega,\alpha)]}, & -\infty < x \le \theta, \\ \dfrac{\delta f_2(x)}{(1+\delta)[1-F_2(\theta)]}, & \theta < x < \infty \end{cases} \tag{3.3}$$

가 된다. 여기서 경계값 θ는 정확한 형태를 알 수는 없지만, 식 (2.6)의 해로서 수치적 방법을 이용하여 다른 모수들의 함수로 나타낼 수 있다. 그러므로 이 복합분포의 모수는 기운 정규분포의 모수 ξ, ω, α와 $f_2(x)$의 모수들로 구성된다.

3.3.2. 기운 t 복합분포 $f_1(x)$가 식 (3.2)의 기운 t 분포의 확률밀도함수이고, $f_2(x)$와 $F_2(x)$가 다른 단일분포의 확률밀도함수와 분포함수일 때 두 분포의 복합분포를 구한다. Azzalini와 Capitanio (2003)은 기운 t 분포의 분포함수 $F_T(x)$는 알려져 있지 않은 이변량 t 분포의 분포함수의 함수이기 때문에 정확한 수식은 주어지지 않고, 몬테칼로 방법(Monte-Calro method) 등 수치적 방법을 통해 구해야 하는 점을 지적하였다. 따라서 기운 t 복합분포의 혼합 가중값을 위한 δ와 확률밀도함수 $f(x)$는 다음과 같이 표현할 수 밖에 없는 한계가 있다.

$$\delta = \frac{2}{\omega} t\left(\frac{\theta-\xi}{\omega}, \nu\right) T\left(\frac{\alpha(\theta-\xi)}{\omega}\sqrt{\frac{\nu+1}{\nu+[(\theta-\xi)/\omega]^2}}, \nu+1\right) \frac{1-F_2(\theta)}{f_2(\theta)F_T(\theta)}, \tag{3.4}$$

$$f(x) = \begin{cases} \dfrac{2t\left(\frac{x-\xi}{\omega}, \nu\right) T\left(\frac{\alpha(x-\xi)}{\omega}\sqrt{\frac{\nu+1}{\nu+[(x-\xi)/\omega]^2}}, \nu+1\right)}{\omega(1+\delta)F_T(\theta)}, & -\infty < x \le \theta, \\ \dfrac{\delta f_2(x)}{(1+\delta)[1-F_2(\theta)]}, & \theta < x < \infty. \end{cases} \tag{3.5}$$

여기서 경계값 θ는 기운 정규 복합분포와 마찬가지로 정확한 형태를 알 수는 없지만, 식 (2.6)의 해로서 수치적 방법을 이용하여 다른 모수들의 함수로 구할 수 있다. 그러므로 이 복합분포의 모수는 기운 t 분포의 모수 ξ, ω, α, ν와 $f_2(x)$의 모수들로 구성된다.

3.3.3. 기운 t 복합분포의 모수 추정 기운 t 복합분포의 모수들 ξ, ω, α, ν와 꼬리분포 $f_2(x)$의 모수들의 최대우도추정량을 구하는 과정을 살펴본다. $x_1, x_2, \ldots, x_n$이 기운 t 복합분포 (3.5)의 표본 데이터이고 $\boldsymbol{\lambda} = (\lambda_1, \lambda_2, \ldots, \lambda_q)$가 꼬리분포 $f_2(x)$의 모수들이라고 할 때 혼합 가중값 δ와 경계값 θ는 각각 식 (3.4)과 (2.6)에 의해 나머지 모수들의 함수로서

$$\delta = \delta(\xi, \omega, \alpha, \nu, \boldsymbol{\lambda}) \quad \text{그리고} \quad \theta = \theta(\xi, \omega, \alpha, \nu, \boldsymbol{\lambda})$$

로 나타낼 수 있다. 따라서 대수우도함수(log-likelihood function)는

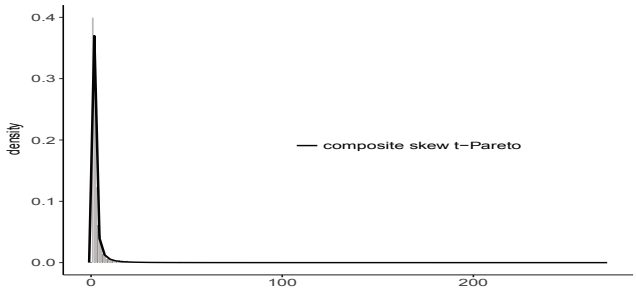
$$\begin{aligned} \log L(\xi, \omega, \alpha, \nu, \boldsymbol{\lambda}) = &-n\log(1+\delta) + \sum_{x_i \le \theta} \log f_T(x_i) - M \log F_T(\theta) \\ &+ \sum_{x_i > \theta} \log f_2(x_i) - m\log[1-F_2(\theta)] + m\log\delta \end{aligned}$$

이다. 여기서 $M = \sum_{i=1}^n I\{x_i \le \theta\}$이고, $m = \sum_{i=1}^n I\{x_i > \theta\}$이다. 일반적으로 대수우도함수를 최대화하는 $(\xi, \omega, \alpha, \nu, \boldsymbol{\lambda})$의 최대우도추정량을 정확한 함수의 형태로 얻을 수는 없으나 $-\log L(\xi, \omega, \alpha, \nu, \boldsymbol{\lambda})$를 최소화하는 값을 **R**의 `nlm()` 함수를 이용하여 비선형 최소화 방법으로 근사적으로 유도할 수 있다.

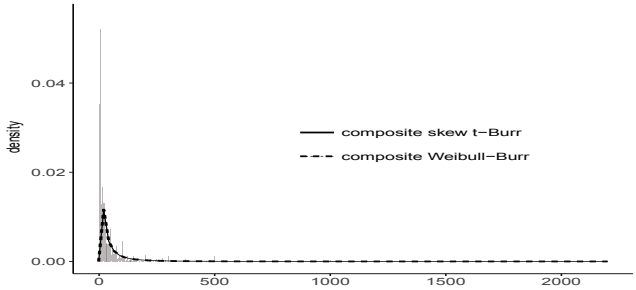
4. 실증분석

4.1. 데이터 소개

실증분석을 위해 보험분야에서 유명한 두 개의 데이터 세트를 이용하고자 한다. 첫 번째 데이터 세트는 **R**의 `SMPracticals` 패키지 (Davison, 2013)에 포함되어 있는 덴마크의 화재보험 청구액 데이터 (Embrechts 등, 1997; Davison, 2003)로서 1980년부터 1990년까지 덴마크 코펜하겐의 재보험회사에서 수집한 보험청구액 데이터로 단위는 백만 덴마크 크로네(Danish Krone)이고 총 2,492개로 이루어져 있다. 특히 이 데이터는 보험통계 분야에서 새롭게 개발되는 다양한 통계방법들을 확인하고자 할 때



(a) Danish fire insurance data



(b) US indemnity data

Figure 4.1. Histogram of data and fitted composite density curves.

Table 4.1. Descriptive statistics for Danish fire insurance data

n	Mean	Min	Max	Median	1st quartile	3rd quartile	Skewness
2492	3.063	0.313	263.300	1.634	1.157	2.645	19.884

Table 4.2. Descriptive statistics for US indemnity data

n	Mean	Min	Max	Median	1st quartile	3rd quartile	Skewness
1500	41.210	0.010	2174.000	12.000	4.000	35.000	9.155

자주 활용되고 있다. Scollnik (2007)와 Nadarajah와 Bakar (2014)는 이 데이터를 로그정규-파레토 복합분포에 적합하여 비교하였으며, Bakar 등 (2015)는 와이블-파레토 복합분포에 적합하였다. Eling (2012)는 단일분포로서 기운 t 분포를 이 데이터에 적용하여 적합도가 높음을 확인하였다.

두 번째 데이터 세트는 미국 Insurance Service Office (ISO)가 무상으로 제공하는, 미국 보험회사들의 배상 지불금 데이터에서 무작위로 선택된 총 1500개의 데이터로서 **R**의 `copula` 패키지 (Hofert 등, 2013)에 포함되어 있다. 단위는 천 미화(USD)이다. Free와 Valdez (1998)는 배상금 데이터와 함께 제공되는 비용 데이터 간의 종속관계를 코풀라(copula)를 이용하여 설명하는 예제 데이터로 활용하였고, Eling (2012)는 기운 t 분포가 이 데이터에 비교적 적절한 적합도를 가지는 단일분포임을 확인하였다.

덴마크의 화재보험 청구액 데이터와 미국의 배상 지불금 데이터의 히스토그램은 Figure 4.1에 나와 있고, 그 기초통계량은 각각 Tables 4.1과 4.2와 같다. 제 1사분위수와 제 3사분위수가 최대값과 크게 차이가 나는 등 매우 치우친 분포로서, 많은 빈도의 작은 청구액과 작은 빈도의 매우 큰 청구액이 공존하는 보험 청구액 분포의 특징을 잘 보여주고 있다.

Table 4.3. Estimated parameters, log-likelihood and AIC for composite distributions for Danish fire insurance data

Composite distribution	Parameters	Log-likelihood	AIC
Skew normal-Pareto	$\xi = 1.017$ $\omega = 0.123$ $\alpha = 0.004$ $\text{scale}' = 0.476$ $\text{shape}' = 1.614$ $\theta = 1.043$ (16% quantile)	-3835.43	7680.87
Skew normal-Burr	$\xi = 1.004$ $\omega = 0.120$ $\alpha = 0.008$ $\text{shape1}' = 0.340$ $\text{shape2}' = 3.594$ $\text{rate}' = 1.215$ $\theta = 1.488$ (44% quantile)	-3831.38	7674.76
Skew t-Pareto	$\xi = 0.844$ $\omega = 0.788$ $\alpha = 30.505$ $\nu = 1.011$ $\text{scale}' = 0.110$ $\text{shape}' = 1.460$ $\theta = 1.921$ (61% quantile)	-3784.91	7581.82
Skew t-Burr	$\xi = 0.841$ $\omega = 0.961$ $\alpha = 40.528$ $\nu = 0.776$ $\text{shape1}' = 0.341$ $\text{shape2}' = 4.198$ $\text{rate}' = 1.126$ $\theta = 0.946$ (9% quantile)	-3785.51	7585.01
Lognormal-Pareto	$\text{location} = 0.182$ $\text{scale} = 1.145$ $\text{scale}' = 0.363$ $\text{shape}' = 1.563$ $\theta = 8.045$ (95% quantile)	-3860.47	7728.94
Lognormal-Burr	$\text{location} = 0.178$ $\text{scale} = 1.093$ $\text{shape1}' = 0.347$ $\text{shape2}' = 4.109$ $\text{rate}' = 1.189$ $\theta = 0.00881$ (0% quantile)	-3857.83	7725.65
Weibull-Pareto	$\text{scale} = 0.969$ $\text{shape} = 15.343$ $\text{scale}' = 0.560$ $\text{shape}' = 1.653$ $\theta = 0.972$ (11% quantile)	-3823.70	7655.40
Weibull-Burr	$\text{scale} = 0.949$ $\text{shape} = 16.203$ $\text{shape1}' = 0.395$ $\text{shape2}' = 3.646$ $\text{rate}' = 1.182$ $\theta = 0.947$ (9% quantile)	-3817.57	7645.14

The prime symbol($'$) refers to parameters for the tail distribution f_2. AIC = Akaike information criterion.

4.2. 분석방법 및 분석결과

복합분포의 모수들의 최대우도추정량을 구하기 위해 기운 정규 복합분포의 경우는 식 (3.3)의 기운 정

Table 4.4. Estimated parameters, log-likelihood and AIC for single distributions for Danish fire insurance data

Single distribution	Parameters	Log-likelihood	AIC
Skew normal	$\xi = 0.571$ $\omega = 8.355$ $\alpha = 68.162$	−7109.85	14225.70
Skew t	$\xi = 0.843$ $\omega = 0.825$ $\alpha = 29.512$ $\nu = 1.159$	−3788.55	7585.10
Weibull	scale = 2.952 shape = 0.948	−5270.47	10544.94
Lognormal	location = 0.672 scale = 0.732	−4433.89	8871.78
Gamma	shape = 1.258 rate = 0.411	−5243.03	10490.05
Cauchy	location = 1.455 scale = 0.493	−4563.49	9130.98
Burr	shape1 = 0.088 shape2 = 14.926 rate = 1.086	−3835.12	7676.24
F	df1 = 11468.635 df2 = 3.735	−4650.73	9305.46
Normal	mean = 3.063 sd = 7.975	−8710.20	17424.39
t	df = 1.164	−7078.32	14158.65
Pareto	scale = 11.900 shape = 5.169	−5051.91	10107.81
Logistic	location = 2.115 scale = 1.463	−6384.42	12772.84
χ^2	df = 2.880	−5264.62	10531.24

AIC = Akaike information criterion.

규 복합분포의 확률밀도함수를 이용하고, 기운 t 복합분포의 경우는 식 (3.5)의 기운 t 복합분포의 확률밀도함수와 함께 기운 t 분포의 누적분포함수 $F_T(x)$를 위해 **R**의 `sn` 패키지 (Azzalini, 2016)를 사용하였다. 한편 로그정규 복합분포와 와이블 복합분포에 대한 적합도 분석은 **R**의 `gendist` 패키지 (Bakar 등, 2016)를 사용하였는데 Scollnik (2007), Nadarajah와 Bakar (2014), Bakar 등 (2015)에도 덴마크의 화재보험 청구액 데이터의 적합도 결과들이 정리되어 있다. 다른 단일분포들의 적합도 분석을 위해서는 **R**의 `nlm()` 함수를 사용하여 음의 대수우도함수값을 최소화하는 각 모수의 최대우도추정량을 구하였다. 여러 분포들의 적합도를 비교하기 위해 최대 대수우도값과 Akaike (1974)가 제시한 Akaike information criterion (AIC)를 사용하였다. 물론 최대 대수우도값은 그 값이 클수록, AIC의 값은 작을수록 분포에 대한 적합도가 높은 것을 나타낸다.

Table 4.3은 덴마크의 화재보험 청구액 데이터에 여러 복합분포를 적용하여 얻은 적합도와 모수 추정결과이다. 최대 대수우도값과 AIC를 고려했을 때, 8개의 복합분포 중 기운 t-파레토 복합분포가 가장 좋은 적합도를 보였다. 추정된 모수의 기운 t-파레토 복합분포의 확률밀도곡선을 Figure 4.1(a)에 데이터의 히스토그램과 함께 나타내었다. 여기서 기운 t-파레토 복합분포의 경계값 θ는 1.921로 얻어졌다.

Table 4.5. Estimated parameters, log-likelihood and AIC for composite distributions for US indemnity data

Composite distribution	Parameters	Log-likelihood	AIC
Skew normal-Pareto	$\xi = 0.019$ $\omega = 0.221$ $\alpha = 13.093$ scale$' = 16.089$ shape$' = 1.233$ $\theta = 0.0629$ (0.3% quantile)	-6571.88	13153.76
Skew normal-Burr	$\xi = 0.133$ $\omega = 4.497$ $\alpha = 25.024$ shape1$' = 10.462$ shape2$' = 0.499$ rate$' = 0.000804$ $\theta = 4.125$ (26% quantile)	-6566.37	13144.75
Skew t-Pareto	$\xi = 0.010$ $\omega = 0.00022$ $\alpha = 314.811$ $\nu = 2.060$ scale$' = 16.152$ shape$' = 1.236$ $\theta = 0.010$ (0.1% quantile)	-6571.11	13154.22
Skew t-Burr	$\xi = 0.031$ $\omega = 4.446$ $\alpha = 100.378$ $\nu = 99.480$ shape1$' = 3.967$ shape2$' = 0.618$ rate$' = 0.008$ $\theta = 3.738$ (24% quantile)	-6563.19	13140.39
Lognormal-Pareto	location $= 2.149$ scale $= 0.494$ scale$' = 16.072$ shape$' = 1.233$ $\theta = 0.989$ (5% quantile)	-6571.56	13151.11
Lognormal-Burr	location $= 2.465$ scale $= 1.639$ shape1$' = 2.688$ shape2$' = 2.683$ rate$' = 2.713$ $\theta = 6692.502$ (100% quantile)	-6566.69	13143.37
Weibull-Pareto	scale $= 4.596$ shape $= 1.199$ scale$' = 14.616$ shape$' = 1.189$ $\theta = 2.134$ (15% quantile)	-6569.04	13146.07
Weibull-Burr	scale $= 3.969$ shape $= 1.302$ shape1$' = 4.148$ shape2$' = 0.609$ rate$' = 0.00756$ $\theta = 3.211$ (22% quantile)	-6559.60	13129.20

The prime symbol($'$) refers to parameters for the tail distribution f_2. AIC = Akaike information criterion.

Table 4.4는 여러 단일분포에 대한 결과로서 기운 t 분포가 가장 높은 적합도를 보이는 것으로 확인되었다. 파레토 분포는 단일분포로는 적합도가 낮았지만 꼬리분포로서 다른 분포와 복합분포를 이루었을 때 그 적합도가 크게 증가하는 특징을 보였다.

Table 4.6. Estimated parameters, log-likelihood and AIC for single distributions for US indemnity data

Single distribution	Parameters	Log-likelihood	AIC
Skew normal	$\xi = 0.010$ $\omega = 110.668$ $\alpha = 1213387.00$	−8148.49	16302.98
Skew t	$\xi = 0.010$ $\omega = 10.748$ $\alpha = 1422774.00$ $\nu = 0.861$	−6594.90	13197.79
Weibull	scale = 26.491 shape = 0.629	−6658.85	13321.70
Lognormal	location = 2.466 scale = 1.638	−6566.77	13137.53
Gamma	shape = 0.506 scale = 81.437	−6766.59	13537.17
Cauchy	location = 8.295 scale = 8.633	−7257.03	14518.07
Burr	shape1 = 1.201 shape2 = 1.012 rate = 0.064	−6572.21	13150.42
F	df1 = 23.679 df2 = 0.729	−7161.64	14327.27
Normal	mean = 41.208 sd = 102.714	−9076.32	18156.65
t	df = 0.312	−8280.55	16563.10
Pareto	scale = 16.229 shape = 1.238	−6572.25	13148.51
Logistic	location = 23.723 scale = 28.678	−8270.46	16544.91
χ^2	df = 12.758	−25814.48	51630.95

AIC = Akaike information criterion.

Tables 4.5와 4.6는 미국의 배상금 데이터에 대한 여러 복합분포와 단일분포들의 적합도 결과로서 이들 중 와이블-버어 복합분포가 가장 좋은 적합도를 보였고 기운 t-버어 복합분포가 그 다음으로 좋은 적합도를 보였다. 이 복합분포들은 결합하는 각 확률분포가 단일분포로서 적합될 때보다 그 적합도가 현저히 향상됨을 확인할 수 있다. 추정된 기운 t-버어 복합분포의 경계값 θ는 3.738로 얻어졌고, 와이블-버어 복합분포의 경계값 θ는 3.211로 추정되었다. 추정된 두 복합분포의 확률밀도곡선을 Figure 4.1(b)에 함께 나타내었다.

5. 결론

본 논문에서는 보험 청구액의 확률분포로서 기운(skewed) 분포를 결합한 새로운 복합분포를 제안하였다. 덴마크 화재보험 청구액 데이터와 미국 배상금 데이터를 통해 기운 t 분포를 머리 부분으로 결합한 복합분포가 단일분포로 사용될 때보다 크게 향상된 적합도를 가짐을 보여주었다. 또한 파레토 분포와 버어 분포는 단일분포보다 복합분포의 꼬리 부분으로 사용될 때 좋은 적합도를 보이는 것을 확인하였다.

저자들이 파악한 바로는 복합분포를 포함하여 현재까지 소개된 보험 청구액의 분포들 중에서 본 논문의 기운 t-파레토 복합분포가 덴마크 화재보험 청구액 데이터에 가장 높은 적합도를 가졌다. 그러나 이 복합분포는 기운 t 분포의 분포함수를 정확히 알 수가 없어 그 형태를 명확하게 표현하는 데에는 한계가 있다. 본 논문에서 제시한 정규분포와 t 분포 외에도 다양한 대칭분포에서 그 대칭성이 깨지고 0이 아닌 왜도를 가지는 기운 분포로의 변형이 가능하다. 좀 더 다양한 형태의 기운 분포들과 기운 분포들의 복합분포를 보험 청구액 분포에 적합하는 지속적인 연구가 필요하다고 판단된다. 더불어 본 논문에서는 복합분포의 경계값을 결합하는 두 분포의 모수들의 함수로 나타내어 결국 복합분포의 모수의 개수는 결합하는 두 분포의 모수의 개수의 합과 같았다. 만약에 미분가능성의 조건을 제거하여 경계값도 모수에 포함시켜 최대우도추정량을 구한다면 모수 추정과 적합도에 어떤 변화가 있을 지 살펴보는 것도 의미있을 것이다.

References

Akaike, H. (1974). A new look at the statistical model identification, *Electronic Journal of Probability*, **19**, 716–723.

Azzalini, A. (1985). A class of distributions which includes the normal ones, *Scandinavian Journal of Statistics*, **12**, 171–178.

Azzalini, A. (2016). The R package `sn`: The skew-normal and skew-t distributions (version 1.4-0), `http://azzalini.stat.unipd.it/SN`.

Azzalini, A. and Capitanio, A. (2003). Distributions generated by perturbation of symmetry with emphasis on a multivariate skew t distribution, *Journal of the Royal Statistical Society. Series B*, **65**, 367–389.

Bakar, S., Hamzah, N. A., Maghsoudi, M., and Nadarajah, S. (2015). Modeling loss data using composite models, *Insurance: Mathematics and Economics*, **61**, 146–154.

Bakar, S., Nadarajah, S., Adzhar, Z., Mohamed, I. (2016). Gendist: an R package for generated probability distribution models, *PLOS ONE*, **11**, e0160903.

Burnecki, K., Kukla, G., and Weron, R. (2000). Property insurance loss distribution, *Physica A: Statistical Mechanics and its Applications*, **287**, 269–278.

Cooray, K. and Ananda, M. M. (2005). Modeling actuarial data with a composite lognormal-pareto model, *Scandinavian Actuarial Journal*, **2005**, 321–334.

Davison, A. (2003). *Statistical Models*, Cambridge University Press, Cambridge.

Davison, A. (2013). The R package `SMPracticals`: Practicals for use with Davison (2003) "Statistical Models"(version 1.4-2), `http://CRAN.R-project.org/package=SMPracticals`.

Eling, M. (2012). Fitting insurance claims to skewed distributions: are the skew-normal and skew student good models?, *Insurance: Mathematics and Economics*, **51**, 239–248.

Embrechts, P., Kluppelberg, C., and Mikosch, T. (1997). *Modelling Extremal Events for Insurance and Finance*, Springer, Berlin.

Frees, E. W. and Valdez, E. A. (1998). Understanding relationships using copulas, *North American Actuarial Journal*, **2**, 1–25.

Hofert, M., Kojadinovic, I., Maechler, M., and Yan, J. (2013). The R package `copula`: Multivariate Dependence with Copulas(version 0.999-16), `http://copula.r-forge.r-project.org/`.

Klugman, S. A., Panjer, H. H., and Willmot, G. E. (2004). *Loss Models: From Data to Decisions* (2nd edition), John Wiley and Sons, New York.

Lee, K. and Park, H. C. (2012). Comparisons of the fitness for lognormal, Pareto and composite lognormal-Pareto distribution based on the insurance payments data, *Journal of the Korean Data Analysis Society*, **14**, 163–172.

McNeil, A. I. (1997). Estimating the tails of loss severity distribution using extreme value theory, *ASTIN Bulletin*, **27**, 117–137.

Nadarajah, S. and Bakar, S. (2014). New composite models for the Danish fire insurance data, *Scandinavian Actuarial Journal*, **2014**, 180–187.

Owen, D. (1956). Tables for computing bivariate normal probabilities, *Annals of Mathematical Statistics*, **27**, 1075–1090.

Pigeon, M. and Denuit, M. (2011). Composite lognormal?Pareto model with random threshold, *Scandinavian Actuarial Journal*, **2011**, 177–192.

Preda, V. and Ciumara, R. (2006). Modeling with Weibull-Pareto models, *North American Actuarial Journal*, **16**, 147.

Scollnik, D. P. M. (2007). On composite lognormal-Pareto models, *Scandinavian Actuarial Journal*, **2007**, 20–33.

Yun, S. H. et al. (2015). *Insurance Industry Outlook and Issues 2015*, Korea Insurance Research Institute.

보험 청구액에 대한 새로운 복합분포

정대현[a] · 이지연[a,1]

[a]영남대학교 통계학과

(2017년 2월 16일 접수, 2017년 3월 31일 수정, 2017년 3월 31일 채택)

요 약

보험 시장은 포화되고 그 성장 동력은 소진되어 보험 산업이 저성장에 머물러 있는 가운데 보험사들은 치열한 경쟁 환경에 놓여있다. 이러한 상황에서 보험 상품에 대한 보험수리적 계산의 기초가 되는 보험 청구액의 흐름을 잘 설명할 수 있는 확률분포를 찾아내는 것은 중요한 쟁점이 될 것이다. 보험 청구액의 분포는 일반적으로 두꺼운 꼬리를 가지면서 왼쪽으로 치우친 로그정규분포나 파레토 분포로 잘 설명된다고 알려져 있으나 최근에는 기운 정규분포나 기운 t 분포가 보험 청구액 분포로 적절한 것으로 고찰되었다. Cooray와 Ananda (2005)는 로그정규분포와 파레토 분포의 장점을 모두 가진 로그정규-파레토 복합분포를 제시하고 단일분포보다 더 높은 적합도를 가짐을 확인하였다. 본 논문에서는 기운 정규분포와 기운 t 분포를 머리 부분으로 결합한 새로운 복합분포를 소개하고 덴마크의 화재보험 청구액 데이터와 미국의 배상 지불금 데이터에 적용하여 기존의 다른 복합분포들을 포함하여 여러 단일분포들과 그 성능을 비교한다.

주요용어: 복합분포, 기운 정규분포, 기운 t 분포, 보험 청구액 분포, 덴마크의 화재보험 청구액 데이터, 미국의 배상 지불금 데이터, 최대우도추정량, AIC

[1]교신저자: (38541) 경북 경산시 대학로 280, 영남대학교 통계학과. E-mail: leejy@yu.ac.kr