保險開發硏究 第12卷 第3號 2001년 12월

로버스트 신뢰도모형*

강 중 철**

이상재해로 인한 거대손해가 발생할 경우 기존의 신뢰도모형은 몇 가지 문제점을 갖게되는데, 이러한 문제점을 해결하기 위하여 본 논문에서 로버스트성을 갖는 신뢰도모형을 소개하고자 한다. 로버스트 통계적방법은 이상점의 영향에 둔감한 성질을 갖는 방법이다. Huber M-추정량과 Gini의 평균차이에 기초한 추정량이 이상치가 존재할 경우 신뢰도모형에서 좋은 추정량이 될 것이다. 본 연구에서 이러한 로버스트 성질을 갖는 추정량에 기초한 신뢰도모형을 제시하고자 한다.

Ⅰ. 서 론

최근 들어 우리 나라에서도 가격자유화라는 변화의 물결 속에서 보험가격 산출에 대한 관심이 고조되고 있다. 가격자유화의 시대에 있어서는 보험가격은 충분해야 하고 과도하지 말아야 하며 지나치게 차별적이어서는 안된다는 보험가격의 3원칙을 충족시켜 줄 수 있는 적정한 가격 결정이야말로 회사의 존폐를 결정 짓는 중요한 요소가 되고 있다. 또한 보험소비자의 입장에서 보면 지나치게 낮은 보험가격일 경우에는 향후 지급여력의 문제가 발생할 수 있으므로 보험가격의 적정성에 관해서는 보험사업자 뿐만 아니라 보험소비자에게도 중요한 관심사가 되고 있다. 따라서 적정한 보험가격을 결정짓는 제 방법의 과학적인 근거를 마련하여야 한다.

보험가격산출에 있어서의 과학적인 제 방법 중 신뢰도 이론은 가장 기본적이고 중요한 이론 중의 하나로서 미국 및 유럽에서는 1910년대부터 신뢰도모형을 개발, 실무에 적용하여 왔다. 이러한 관점에서 보험가격에 있어서 신뢰도이론은 중요한

* 이 논문은 2001년 동서대학교 학술연구 조성비에 의한 것임.
** 동서대학교 e-비즈니스학부 금융보험학 교수

의미를 가지게 되며, 가격자유화의 시대를 맞이한 국내보험사도 이 문제에 관하여 관심을 가져야 할 때이다. 그러나 이 분야에 대한 연구가 아직까지 미흡한 상태이며 적절한 수리적 모형을 제시하지 못하고 있는 실정이다.

본 논문에서는 신뢰도이론의 제 방법 중 변동제한법모형이 가지고 있는 특성과 함께 몇 가지 문제점을 살펴보고 이를 개선시키고자 하는 새로운 모형을 제시하고자 한다. 즉, 변동제한법모형은 일반적으로 단순 산술평균($\overline{X}$)과 분산(S^2)만을 이용하여 신뢰도 계수를 산출하고 있다. 이와 같은 모형에서는 몇 가지 문제점이 내포하고 있는데, 본 논문에서는 이러한 문제점에서 관하여 살펴보고 이에 대한 대안적인 모형을 제시하고자 하는 것이 본 연구의 주된 목적이다.

즉, 지진, 폭풍 등 자연재해로부터 대규모 손해액이 발생하는 경우에는 손해액의 평균값은 이러한 이상치(outlier)로부터 아주 많은 영향을 받게 되어 평균적인 손해액의 규모를 다소 왜곡시키는 경향이 발생하게 된다. 이러한 경우에는 이상치를 제거시키는 방법보다는 이상치에 덜 민감한 통계량을 사용할 필요가 있다. 또한 위험집단을 분류할 때에는 동질적인 위험집단으로 분류하여야 한다. 그러나 동질적인 위험집단으로 세분화하여 분류할 경우 데이터개수(위험단위수)가 적게 되는 경우가 발생하게 된다. 이러한 경우에는 일반적인 평균과 산포의 추정량(예를들면 표본평균, 표본표준편차)으로는 편의(bias)와 정도(precision)의 측면에서 효율이 떨어질 수 있으며 이러한 추정치에 기초한 신뢰도 모형은 왜곡된 정보를 제시하는 결과를 갖게된다.

본 논문의 주된 연구 대상은 평균과 표준편차에 기초한 변동제한법모형보다는 로버스트(robust)한 통계량을 사용하여 이상치에 민감하게 반응하지 않는 신뢰도 모형을 제시하고자 한다. 또한, 변동제한법 신뢰도모형과 제안된 신뢰도를 비교하기 위하여 Monte Carlo 모의실험을 실시하여 두 신뢰도를 비교 분석함으로 제안된 신뢰도의 우수성을 입증하고자 한다.

본 논문의 구성은 제2장에서는 변동제한법모형에서의 특성 및 문제점을 분석하고, 제3장에서 이러한 문제점을 개선할 수 있는 좀 더 로버스트한 추정량인 M-추정량과 Gini의 평균차이에 기초한 추정량에 대하여 살펴보고, 그리고 제4장에서 이러한 로버스트성을 갖는 추정량에 기초한 신뢰도모형을 제시하고 그 모형의 효율성

을 검토하고자 한다. 제5장에서 결론으로 논문을 마무리 하고자 한다.

Ⅱ. 표본평균과 표준편차에 기초한 신뢰도모형

보험요율을 결정하는데 있어서 통계자료의 보완없이 기존의 통계자료만으로도 충분한 신뢰의 크기를 부여할 수 있다면 그것을 전신뢰도라고 한다. 전신뢰도는 사고건수의 함수로 표시될 수 있으므로 $Z(n)=1$을 만족하는 n의 값을 결정하면 전신뢰도를 부여할 수 있는 사고건수(위험단위수)가 된다. 이러한 값을 결정하는데에는 여러가지 방법이 있다.

Herzog(1999)는 신뢰도 이론의 접근방법을 크게 변동제한법(limited fluctuation approach), 빌만법(Buhlmann's approach), 베이지안법(Bayesian approach)으로 나누고 있다. 변동제한법은 미국의 Mowbray가 1913년부터 최초로 연구를 시작하였다고 하여 보통 미국식 신뢰도(American credibility)라고 말한다. 빌만법은 포트폴리오의 이질성의 정도를 반영하기 위한 모델로서 포트폴리오를 구성하고 있는 등급간의 이질성이 크면 클수록 등급별 경험치에 더 많은 신뢰도가 부여된다. 베이지안 이론은 20세기 후반부터 본격적으로 연구가 되기 시작하였으며 이는 변동제한법 또는 빌만법과 달리 모수를 하나의 확률변수로 취급하여 이에 대한 주관적인 판단 또는 사전적인 정보를 이용하여 모수를 추정하는 방법이다. 본 장에서는 이 세 가지 방법 중에서 변동제한법모형이 갖는 특성 및 문제점을 살펴보고자 한다.

변동제한법을 전개하기 위해 다음과 같은 기본가정들이 필요하다. 첫째, 어떤 주어진 기간동안(예를 들면, 보험기간 1년) 포트폴리오내에서 발생하는 총사고건수를 나타내는 확률변수를 N이라 하면, 이 확률변수 N은 포아송분포(Poisson distribution) 또는 이항분포(Binomial distribution), 음이항분포(Negative Binomial distribution) 등 여러가지 분포를 가정할 수 있겠으나, 본 논문에서는 확률변수 N이 포아송분포를 따른다고 가정한다. 즉, 확률변수 N의 평균과 분산은 동일한 것으로 평균과 분산이 각각 $E(N) = n$, $Var(N) = n$을 가정한다. 둘째, 주어진

보험 포트폴리오내에서 발생하는 사고건에 대한 i번째 손해액의 크기(amount of the ith claim)를 나타내는 확률변수 X_i (i = 1, 2, 3, ······, N)는 모평균이 μ이고 모분산이 σ^2인 동일한 (즉, identically distributed with expectation μ and variance σ^2) 분포를 따른다고 가정한다. 셋째, 주어진 두개의 확률변수들은 서로 독립(independent)이라는 가정이다. 즉, 총사고건수의 확률변수 N과 손해액 크기의 확률변수 $X_1, X_2,$ ······, X_N은 서로 독립이다. 이같은 기본가정하에서 보험포트폴리오내에서 일정기간(예를 들면, 보험기간 1년) 동안에 발생하는 총손해액은 $T = X_1 + X_2 + \cdots\cdots + X_N$로 나타낼 수 있으며, 총손해액 T의 모평균과 모분산은 각각 $E(T) = \tau = n\mu$과 $Var(T) = n(\sigma^2 + \mu^2)$가 된다.

이때 통계적 추정의 정확도의 기준으로 다음의 기준을 사용할 수 있다.

$$\Pr[|T-\tau| \leq k\tau] \geq 1-\alpha \qquad (2.1)$$

(2.1)식 의미하는 것은 총손해액을 T와 T의 평균의 차이가 T의 평균의 100k%이하일 확률이 $1-\alpha$ 이상이어야 한다는 것이다. 이를 다시

$$\Pr\left[-\frac{k\tau}{\sqrt{Var(T)}} \leq \frac{T-\tau}{\sqrt{Var(T)}} \leq \frac{k\tau}{\sqrt{Var(T)}}\right] \geq 1-\alpha \qquad (2.2)$$

으로 바꿔 쓸 수 있다. 따라서 중심극한정리에 따라 T의 분포가 근사적으로 평균이 $n\mu$, 분산 $n(\sigma^2 + \mu^2)$인 정규분포를 따른다고 할 수 있으므로 (2.2)식의 조건은 결국

$$\frac{k\tau}{\sqrt{Var(T)}} = \frac{kn\mu}{\sqrt{n(\mu^2+\sigma^2)}} \geq Z_{1-\alpha/2}$$

을 얻을 수 있고 이것을 n에 대한 조건으로 쓰면

$$n \geq \left(\frac{Z_{1-\alpha/2}}{k}\right)^2\left[1 + \left(\frac{\sigma}{\mu}\right)^2\right]$$

따라서 전신뢰도의 조건을 만족하는 n의 최소값은 다음과 같다.

$$n_F = \left(\frac{Z_{1-\alpha/2}}{k}\right)^2\left[1 + \left(\frac{\sigma}{\mu}\right)^2\right] \qquad (2.3)$$

(2.3)식은 사고건의 손해액의 분산이 클수록, 평균값이 작을수록 별도의 위험단위로 구별되기 위해 필요한 최소위험단위수는 n_F의 값이 증가함을 보여준다. 이 식에서 α, k의 값은 조정할 수 있는 값으로서 일반적으로 $\alpha = 0.1$ 또는 0.05를 사용하며 이에 대응하는 $Z_{1-\alpha/2}$의 값은 각각 1.645, 1.96이며 k의 값으로는 일반적으로 0.1, 0.3, 0.05 등의 값을 이용한다. 참고로 $\frac{Z^2_{1-\alpha/2}}{k^2}$의 값은 k와 $(1-\alpha)$의 값에 따라 아래와 같이 구해진다

〈표 1〉 전신뢰도를 만족시키는 최소사고건수

$1-\alpha$ \ k	0.3	0.2	0.1	0.05	0.01
0.90	30	68	271	1,082	27,060
0.95	43	96	384	1,537	38,416
0.99	74	166	664	2,654	66,358
0.999	120	271	1,083	4,331	108,274

(2.3)식을 사용하기 위해서는 몇 가지 문제점을 안고 있다. 첫째, 전신뢰도를 만족시키는 사고건수를 계산하기 위해서는 (2.3)식에서 보는 바와 같이 범위모수 k와 신뢰수준 p를 가격결정자의 주관에 의해 결정해야만 한다. 즉, 동일한 포트폴리오라고 하더라도 범위모수 k와 신뢰수준 p를 어떻게 결정하느냐에 따라 전신뢰도에 필요한 사고건수의 크기가 달라진다. 둘째, 손해액의 분산 σ과 평균값 μ을 추정시 소수의 고액사고금액에 대하여 상당히 크게 영향을 받게 된다는 점이다. 특히 우리나라의 자동차보험의 대인배상의 경우는 무한보상이기 때문에 사고금액분포 형태를 고려치 않고 일반적인 정규분포를 가정하여 손해액의 평균값과 분산을 추정하는 것은 타당치 못한 방법이라 할 것이다.

본 논문에서는 이에 대한 대안적인 방법으로 이들의 추정치가 이상치(outlier)의 영향에 둔감한 추정치를 사용할 것을 제안하는 것이다. 즉, 표본평균과 표본분산에 기초한 신뢰도모형이 이상점(outliers)에 민감하기 때문에 이에 대한 단점을 보완한 로버스트 신뢰도모형(robust credibility)를 제안하고자 한다. 본 논문에서 제안하고자 하는 신뢰도모형은 평균의 로버스트한 추정량인 M－추정량에 기초하고, 기존의 신뢰도에서 산포의 추정법으로 흔히 사용되는 표준편차 대신에 Gini의 평균차이(mean difference)에 기초한 추정량이 사용된다.

Ⅲ. M－추정량과 Gini의 평균차이에 기초한 산포추정량

전통적인 통계적 추론법은 모집단에 대하여 정규분포를 가정하는 정규성이론에 의하여 개발되고 연구되어 왔다. 그러나 정규성이론에 의한 통계적 방법들은 대부분이 정규성의 가정에 매우 민감하다. 실제 현실에서 얻은 자료들은 정규성의 가정을 할 수 없는 경우가 많다. 즉, 자료들 가운데 이상점(outlier)이 있거나 자료들이 두터운 꼬리를 갖는 분포를 따르는 경우가 많다. 또한 한쪽으로 치우친 분포를 갖는 자료도 흔히 접하게 된다. 특히 보험에서 접하는 자료의 특성들은 대부분 이러한 특성을 갖는 자료들이다. 이와 같이 정규성의 가정이 만족되지 않을 때는 정규이론에 근거한 추론법은 이론적인 오차보다도 더 많은 오류의 가능성을 내포할 수 있다.

로버스트 통계적 방법들은 이와 같은 문제점을 보완하기 위하여 개발된 기법이다. 즉 로버스트 통계적 방법은 정규성의 가정하에서도 합리적으로 좋은 효율을 유지하고, 넓은 범위의 분포에서 합리적으로 잘 적용될 수 있으며, 이상점의 영향에 둔감한 성질을 갖는 방법이다.

예를들어 항공보험의 손해율이 다음과 같다고 하자.

$$60\%,\quad 65\%,\quad 70\%,\quad 75\%,\quad 80\%$$

이들 자료의 평균과 메디안은 다음과 같다.

$$\text{평균} = 70\%, \qquad \text{메디안} = 70\%$$

그런데 거대손해로 인하여 손해율이 80%에서 800%로 되었다고 하자, 그러면 평균과 메디안은 다음과 같다.

평균 = 214%,   메디안 = 70%

이와 같이 평균 $\bar{x}$ 보다는 메디안이 훨씬 좋은 중심위치를 나타내는 측도가 된다. 즉 메디안은 이상점인 800%에 영향을 받지 않으나, 정규이론에 의해 얻어진 표본평균은 이상점에 심각한 영향을 받고 있음을 알 수 있다.

$x_1, \cdots, x_n$ 을 연속분포인 $F(x-\theta)$에서 랜덤표본이라 할 때, 이들 표본에 기초하여 위치모수 θ를 추정하는 문제를 생각해 보자.

1. M-추정량

Huber(1964)가 M−추정량을 제안한 이래, 많은 연구자들([1], [3], [9])에 의하여 M−0추정량의 표본평균에 대한 로버스트성 및 효율성에 관한 연구가 이루어졌다. 일반적으로 다음과 같은 일표본 위치문제(one-sample location) 모형에서,

$$x_n = \mu + \varepsilon_i \ , i = 1, 2, \cdots, n,$$

위치모수에 대한 Huber의 M−추정량은 목적함수(object function)

$$Q(\mu) = \sum_{i=1}^{n} \rho(x_i - \mu)$$

를 최소화하는 $\hat{\mu}$, 또는 음방정식(implicit equation)

$$\sum_{i=1}^{n} \psi(x_i - \mu) = 0$$

해인 $\hat{\mu}$으로 정의된다. 여기서 ρ가 볼록함수이고 미분가능할 때는 $\rho' = \psi$가 된다. 예를들어 $\rho(x) = x^2 / 2$이면 $\psi(x) = x$이 되고, 이때 대응되는 μ의 추정량은 $\bar{x}$가 된다. M−추정량은 ρ−함수 또는 대응되는 ψ−함수의 형태에 따라 추정량이 달라지며, Huber가 제안한 ψ−함수의 형태는 다음과 같다.

$$\psi(x) = \begin{cases} x, & |x| \le c \\ c \cdot sign(x), & |x| > c \end{cases}$$

여기서 c는 조율상수(tuning constant)로서 $c = 1.5$는 많은 모의실험을 통하여 합리적인 값으로 사용되는 상수 중의 하나이다. c의 값이 작아지면 로버스트 추정량(robust estimator)을 구할 수 있으나 효율이 떨어지게 되고, 반면 큰 c의 값에 대하여 효율은 높아지지만 이상점의 영향을 제어하기 힘들어진다. Huber가 제안한 M−추정량의 동기는 일반적으로 오염된 자료(contaminated data), 또는 길고 두터운 꼬리를 가지는 분포(long and heavy-tailed distribution)에서의 표본에서 표본평균 $\bar{x}$가 갖는 비로버스트성(non-robustness)에 있다. 즉, 표본에 이상점(outliers)이 있을 경우 표본평균은 이상점에 심각하게 영향을 받는다. 그러나 ψ−함수의 구조에서 볼 수 있듯이 Huber의 M−추정량에서는 이상점의 영향이 감소되므로 이상점에 대하여 둔감한 로버스트 추정량을 구할 수 있다.

2. M − 추정량의 수치해법

표본평균 $\bar{x}$와는 달리 M−추정량은 위치−척도−등가변환(location-scale-equivariant)의 조건을 만족하지 못한다[1]. 즉, 일반적으로 M−추정량에서는 다음의 조건이 성립하지 않는다.

$$\hat{\mu}(bx_1 + a, \cdots, bx_n + a) = b\hat{\mu}(x_1, \cdots, x_n) + a$$

따라서 척도에 무관하게 정의되어 사용될 수 있는 척도불변(scale invariant) M−추정량은 다음과 같은 음방정식의 해 $\hat{\mu}$로 정의될 수 있다.

$$\sum_{i=1}^{n} \psi\left(\frac{x_i - \mu}{s}\right) = 0 \tag{3.1}$$

여기서 s는 척도(σ)의 로버스트한 추정량으로서, 절대편차중앙값(median absolute deviation; MAD)에 기초한 추정량이 흔히 사용된다. 즉,

$$s = 1.483 \cdot median_i \{|x_i - median_j (x_j)|\}$$

여기서 1.483은 정규분포에서 s가 σ의 일치추정량이 되도록 하는 상수이다.

식 (3.1)의 해는 μ의 초기추정치(initial estimate) $\hat{\mu}_0$에 대하여 일차 테일러 급수전개(first-order Taylor series expansion)를 한 다음 반복법에 의하여 얻을 수 있으며, 초기추정치로는 중앙값이 흔히 사용된다. 즉, M−추정량을 얻기 위한 알고리즘은 다음과 같다.

$$\hat{\mu}_m = \hat{\mu}_{m-1} + s \cdot \frac{\sum \psi[(x_i - \hat{\mu}_{m-1})/s]}{\sum \psi'[(x_i - \hat{\mu}_{m-1})/s]}, \; m = 1, 2, \cdots$$

한편, 적절한 조건하에서 M−추정량($\hat{\mu}$)은 정규분포로 근사될 수 있는데, 표본크기 n이 충분히 클 경우 다음과 같은 점근 정규성을 갖는다.

$$\sqrt{n}(\hat{\mu} - \mu) \xrightarrow{d} N(0, \sigma^2)$$

여기서 σ^2은 $\sqrt{n}\hat{\mu}$의 점근분산으로, 일단계 M −추정량(one-step M-estimator)을 사용할 경우, 다음과 같이 추정될 수 있다.

$$\hat{\sigma}^2 = s^2 \cdot \frac{\frac{1}{n}\sum \psi^2\left(\frac{x_i - \hat{\mu}_0}{s}\right)}{\left[\frac{1}{n}\sum \psi'\left(\frac{x_i - \hat{\mu}_0}{s}\right)\right]^2}$$

3. 몬테칼로 모의실험을 통한 비교

M−추정량의 로버스트성을 표본평균 $\bar{x}$ 와 비교해보기 위하여 다양한 상황에서 몬테칼로(Monte Carlo) 모의실험을 실시하였으며, 모의실험에서 고려한 모형은 다음과 같다.

$$x_{ij} = 10.0 + \varepsilon_{ij} + \delta; \quad j = 1, 2, \cdots, 5, \quad i = 1, 2, \cdots, 2000$$

여기서 δ는 평균의 변화를 의미하는 값으로서 다음과 같이 변화시키면서 측정하였다.

$$\delta = a \times \frac{\sigma}{2}\ ; \qquad a = 0, 1, 2, \cdots, 6$$

모의실험에서 고려한 오차항의 분포는 오염된 정규분포(contaminated normal distribution; $CN(\alpha, \sigma)$)로서 분포함수 $F(x)$는 다음과 같다. 즉, α 를 표준정규분포에서의 오염비율(contaminated rate)이라 하고, $\Phi(x)$를 표준정규분포의 분포함수라 하면

$$F(x) = (1-\alpha)\,\Phi(x) + \alpha\,\Phi(x/\sigma)$$

으로 나타낼 수 있다. 또한, 모의실험에서 사용된 각 부분군에서 평균의 추정량은 일단계 M−추정량(one-step M-estimator)이며 초기추정량으로 중앙값(median)을, 조율상수 (c)로서 1.5를 사용하였다. 한편, 모의실험의 모든 계산은 S-PLUS(Statistical Sciences〔12〕)를 이용하여 이루어 졌다.

〈표 2〉 $\bar{x}$ 와 M−추정량의 비교

오차분포		$N(0, 1)$	$CN(0.1, 3)$	$CN(0.2, 3)$	$CN(0.1, 5)$	$CN(0.2, 5)$	$CN(0.1, 4)$
Mean	$\bar{x}$	9.9891	9.9914	10.0089	9.9831	9.9991	10.0213
	M	9.9912	9.9899	10.0065	9.9851	10.0026	10.0058
MSE	$\bar{x}$	0.1997	0.3433	0.5166	0.6961	1.2098	0.5186
	M	0.2035	0.2756	0.3586	0.3714	0.4824	0.3260

〈표 2〉은 모의실험(2000번 반복)에서 얻어진 표본평균($\bar{x}$) 및 M−추정량(M)들의 평균(Mean)과 평균제곱오차(MSE)를 계산한 결과이다. 〈표 2〉에서 목표값은 10.0이며, 편의(bias)의 측면에서 보면 표본평균과 M−추정량은 비슷한 정도의 편의를 가지고 있음을 알 수 있다. 그러나 평균제곱오차의 측면에서 보면 $N(0, 1)$ 경우를 제외한 모든 경우에서 M−추정량의 MSE가 표본평균의 그것보다 훨씬 작은 값을 가짐을 알 수 있으며, 이로부터 M−추정량의 로버스트한 성질을 알 수 있다.

4. Gini의 평균차이(mean difference)에 기초한 산포의 추정

Gini의 평균차이(Gini's mean difference)는 지니(Gini(1912))에 의하여 제안된 산포의 추정방법으로 다음과 같이 나타낼 수 있다.

$\hat{\sigma}_G$는 척도(σ)의 로버스트 추정량으로서,

$$\hat{\sigma}_G = \frac{x}{{}_nC_2} \sum_{l=2}^{n} \sum_{j=1}^{l-1} | x_i - x_l | \qquad (3.2)$$

여기서 $x = \sqrt{\pi}/2$를 주로 쓰며, 이는 데이터가 정규분포에서의 표본일 경우 추정량의 불편성을 담보한다. 일반적으로 분포함수 $F(x)$를 갖는 연속확률변수 X에 대한 평균차이 Δ_R 은 다음과 같이 표현될 수 있으며,

$$\Delta_R = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} |x - y| dF(x)\, dF(y)$$

Δ_R 의 적률방식에 의한 추정량(method of moment estimator)은 다음과 같이 나타낼 수 있다.

$$\hat{\Delta}_R = \frac{1}{{}_nC_2} \sum_{l=2}^{n} \sum_{j=1}^{l-1} | x_i - x_l |$$

즉, 지니(Gini)의 평균차이는 이에 기초한 산포의 추정량으로서, 표본의 크기가 작을 때 데이터의 재사용(reuse) 또는 반복사용을 통한 정보활용도를 높여 산포를 추정하기 때문에 추정의 정도(precision) 또는 효율성을 높일 수 있으며, 이상점(outliers)에 덜 민감하여 산포 σ에 대한 로버스트 추정법으로 고려될 수 있다.

한편, 식(3.2)의 추정량은 다음과 같이 표현할 수 있는데,

$$\hat{\sigma}_G = \frac{x}{n(n-1)} \sum_{i=1}^{n} (2i - n - 1)\, X_{(i)} \qquad (3.3)$$

여기서 $X_{(i)}$는 순서통계량(order statistics)으로서 다음이 성립한다.

$$X_{(1)} \leq X_{(2)} \leq \cdots \cdots \leq X_{(n)}$$

식 (3.3)를 이용하여 보다 간편하게 추정량을 계산할 수 있으며, 또한 이에 기초하여 데이터가 정규모집단에서의 표본일 때, Gini의 평균차이에 기초한 산포추정량에 대한 표준오차는 다음과 같이 계산되어질 수 있다.(David(1968), Lomnicki(1952), Nair(1936))

$$\text{S.E. } (\hat{\sigma}_G) = \sigma \sqrt{\frac{1}{n(n-1)} \left\{ n(\pi/3 + 2\sqrt{3} - 4) + (6 - 4\sqrt{3} + \pi/3) \right\}}$$

이는 $c_4(n)$ 또는 $d_2(n)$과 같은 복잡한 형태의 계산식을 필요로 하지 않으며, 또한 이를 위한 계산표(table)도 필요하지 않은 분명한 형태로 표현된다.

한편, Gini의 평균차이에 기초한 산포의 추정법은 그동안 계산의 번거로움 때문에 기피되어온 면이 없지 않다. 그러나 오늘날과 같이 컴퓨터의 사용 및 사용자의 인터페이스가 강화된 프로그램 환경에서 계산알고리즘의 복잡성 또는 번거로움은 더 이상 설득력을 얻기 힘들 것으로 생각된다.

5. 몬테칼로 모의실험을 통한 비교

본 절에서는 앞에서 소개된 Gini의 평균차이를 이용한 산포 추정량들의 편의 및 효율성을 비교하기 위하여 다양한 분포군에서 모의실험을 시행해 보고자 한다. 우선 이를 위하여 다음과 같은 모형을 고려해 보자.

$$x_i = 10.0 + \varepsilon_i \ ; \qquad i = 1, 2, \cdots, n \qquad n = 5, 10, 25$$

여기서 ε_i는 오차항으로서 $E(\varepsilon_i) = 0$, $Var(\varepsilon_i) = \sigma^2$ 이다.

모의실험에서 고려한 분포는 오염된 정규분포(contaminated normal distribution; $CN(\alpha, \sigma)$)로서 분포함수 $F(x)$는 다음과 같다. 즉, α를 표준정규분포에서의 오염비율(contaminated rate)이라 하고, $\Phi(x)$를 표준정규분포의 분포함

수라 하면

$$F(x) = (1-\alpha)\,\Phi(x) + \alpha\,\Phi(x/\sigma)$$

으로 나타낼 수 있다. 한편, 모의실험의 모든 계산은 S-PLUS(Statistical Sciences[12])를 이용하여 이루어 졌다.

〈표 3〉은 모의실험(5000번 반복)에서 얻어진 산포(σ)의 추정량들에 대한 평균(MEAN)과 평균제곱오차(mean squared error; MSE)를 계산한 결과이다. 〈표 3〉에서 파라메타의 목표값(σ)은 분포에 따라 다른데 산포를 표에 계산해 두었다. 여기서 평균제곱오차는

$$\text{MSE} = \frac{1}{5000}\sum(\text{산포의 추정치} - \text{산포}(\sigma))^2$$

으로 계산된다.

한편, 〈표 3〉에서 MSE의 측면을 고려할 때, 비교의 용이성을 위하여 효율(efficiency)을 구하여 괄호(　)속에 표시하였는데, 각 분포 및 표본의 크기에서 효율은 다음과 같이 구하여 진다.

$$\text{효율} = \frac{\text{최소 MSE}}{\text{MSE}} \times 100(\%)$$

〈표 3〉에서 편의의 측면에서 보면 범위를 이용한 추정법이 상대적으로 작은 편의를 보이고 있으나 그 차이가 큰 것은 아니다. MAD를 이용한 산포의 추정은 소표본(small sample)에서 편의 및 MSE의 측면에서 좋지 못한 성능을 가지고 있다.

한편, MSE의 측면에서는 $N(0,1)$ 경우를 제외한 모든 고려된 분포에서 Gini 추정법에 의한 산포추정량이 가장 작은 값을 가지고 있어서 산포를 안정적으로 추정하고 있음을 보여주고 있다.

또한, Gini의 평균차이에 기초한 추정법은 정규모집단에서의 표본일 경우 최적의 추정량으로 알려진 표본표준편차 또는 표본표준편차에 기초한 추정량에 대하여 98%이상의 높은 효율을 유지하고 있어서 로버스트 추정량으로서의 상당한 장점을 갖고 있다고 볼 수 있다. 반면 시뮬레이션 결과에 의하면 Gini의 평균차이에 기초한 추정량은 편의의 측면에서도 큰 차이를 보이지 않고 안정적이며, 정규분포를 제외한 모든 분포에서 비교된 추정량들 중 가장 뛰어난 효율을 갖고 있는 것으로 나타났다.

〈표 3〉 산포추정량의 성능비교

분　　포			$N(0, 1)$	$CN(0.1, 3)$	$CN(0.2, 3)$	$CN(0.1, 5)$	$CN(0.2, 5)$
산　포(σ)			1	1.3416	1.6125	1.8439	2.4083
MEAN	n = 5	$\widehat{\sigma}_S$	1.0028	1.2610	1.5088	1.5650	2.0927
		$\widehat{\sigma}_R$	1.0026	1.2669	1.5232	1.5701	2.1091
		$\widehat{\sigma}_G$	1.0029	1.2422	1.4765	1.5018	1.9944
		$\widehat{\sigma}_M$	0.8251	0.9293	1.0566	0.9727	1.2080
	n = 10	$\widehat{\sigma}_S$	1.0028	1.2832	1.5408	1.6590	2.2314
		$\widehat{\sigma}_R$	1.0021	1.3386	1.6302	1.7785	2.4368
		$\widehat{\sigma}_G$	1.0025	1.2386	1.4764	1.5132	2.0160
		$\widehat{\sigma}_M$	0.9141	0.9959	1.1270	1.0367	1.2127
	n = 25	$\widehat{\sigma}_S$	1.0027	1.3160	1.5660	1.7403	2.3006
		$\widehat{\sigma}_R$	1.0059	1.5092	1.8484	2.1788	2.9498
		$\widehat{\sigma}_G$	1.0015	1.2436	1.4642	1.5104	1.9800
		$\widehat{\sigma}_M$	0.9650	1.0565	1.1515	1.0819	1.2074
MSE	n = 5	$\widehat{\sigma}_S$	0.1321(100)	0.4481(89.3)	0.6825(89.5)	1.4130(82.1)	2.2463(86.8)
		$\widehat{\sigma}_R$	0.1397(94.6)	0.4742(84.4)	0.7208(84.7)	1.4270(81.3)	2.3208(84.0)
		$\widehat{\sigma}_G$	0.1347(98.1)	0.4000(100)	0.6107(100)	1.1598(100)	1.9505(100)
		$\widehat{\sigma}_M$	0.2604(50.7)	0.5103(78.4)	0.7911(77.2)	1.1696(99.1)	2.3349(83.5)
	n = 10	$\widehat{\sigma}_S$	0.0568(100)	0.2482(78.4)	0.3663(83.3)	0.9291(70.4)	1.3053(79.6)
		$\widehat{\sigma}_R$	0.0657(86.5)	0.3431(56.7)	0.5059(60.3)	1.2345(53.0)	1.7608(59.0)
		$\widehat{\sigma}_G$	0.0579(98.1)	0.1946(100)	0.3051(100)	0.6545(100)	1.0395(100)
		$\widehat{\sigma}_M$	0.1238(45.9)	0.2635(73.9)	0.4349(70.2)	0.8182(80.0)	1.7138(60.7)
	n = 25	$\widehat{\sigma}_S$	0.0209(100)	0.1156(70.2)	0.1543(83.0)	0.4514(71.7)	0.5707(91.7)
		$\widehat{\sigma}_R$	0.0324(64.5)	0.3309(24.5)	0.4244(30.2)	1.2240(26.4)	1.5731(33.3)
		$\widehat{\sigma}_G$	0.0212(98.6)	0.0812(100)	0.1281(100)	0.3235(100)	0.5234(100)
		$\widehat{\sigma}_M$	0.0553(37.8)	0.1496(54.3)	0.2953(43.4)	0.6525(49.6)	1.5371(34.1)

※ 밑줄: 최소의 편의 및 MSE를 갖는 결과

Ⅳ. 로버스트 추정량에 기초한 신뢰도모형

변동제한법의 신뢰도모형은 이상치가 존재할 때 표본평균과 표준편차에 민감하게 반응하는 단점을 가지고 있다는 사실을 살펴보았다. 이에 대한 대안적인 방법으로 이들의 추정치가 이상치(outlier)의 영향에 둔감한 추정치를 사용할 것을 제안하는 것이다. 즉, 표본평균과 표준편차에 기초한 신뢰도모형이 이상점(outliers)에 민감하기 때문에 이에 대한 단점을 보완한 로버스트 신뢰도모형(robust credibility)를 제안하고자 한다. 본 논문에서 제안하고자 하는 신뢰도모형은 평균의 로버스트한 추정량인 M－추정량에 기초하고, 기존의 신뢰도에서 산포의 추정법으로 흔히 사용되는 표준편차 대신에 Gini의 평균차이(mean difference)에 기초한 추정량이 사용된다. 즉, 기존의 변동제한법 신뢰도모형은 정규분포의 가정하에서 유도된 모형이다. 그러나 이러한 정규성의 가정은 현실적으로 만족되지 않는 가정이며, 특히 이상재해와 같은 거대손해가 발생할 경우에는 정규성의 가정은 완전히 상실되게 된다. 이러한 비현실성의 문제를 개선하고자 평균값의 추정치로서 제Ⅲ장에서 제시한 M－추정량을 사용하고, 분산의 추정치는 Gini의 평균차이에 기초한 추정량을 사용하여 로버스트한 신뢰도모형을 제안한다.

$$n_F = \left(\frac{Z_{1-\alpha/2}}{k}\right)^2 \left[1 + \left(\frac{\hat{\sigma}_G}{\hat{\mu}_m}\right)^2\right]$$

여기서

$$\hat{\mu}_m = \hat{\mu}_{m-1} + \sigma_M \cdot \frac{\sum \psi[(x_i - \hat{\mu}_{m-1}) / \hat{\sigma}_M]}{\sum \psi'[(x_i - \hat{\mu}_{m-1}) / \hat{\sigma}_M]} \quad , m = 1, 2, \cdots$$

$$\hat{\sigma}_G = \frac{x}{{}_nC_2} \sum_{j=2}^{n} \sum_{i=1}^{j-1} | x_i - x_j |$$

$$= \frac{x}{{}_nC_2} \sum_{i=1}^{n} X_{(i)} \times (2i - n - 1), \; X_{(1)} < X_{(2)} < \cdots < X_{(n)}$$

또한, 기존의 신뢰도모형과 제안된 신뢰도모형의 편의 및 효율성을 비교하기 위하여 다양한 분포군에서의 Monte Carlo 모의실험을 실시하고자 한다. 모의실험에서 고려한 모형은 다음과 같다.

$$x_i = 20.0 + \varepsilon_i\ ;\quad i = 1,\ 2,\ \cdots,\ n \qquad n = 10,\ 20,\ 30$$

여기서 ε_i는 오차항으로서 $E(\varepsilon_i) = 0$, $Var(\varepsilon_i) = 25\sigma^2$이다.

모의실험에서 고려한 분포는 오염된 정규분포(contaminated normal distribution ; $CN(\alpha, \sigma)$)로서 분포함수 $F(x)$는 다음과 같다. 즉, α를 표준정규분포에서의 오염비율(contaminated rate)이라 하고, $\Phi(x)$를 표준정규분포의 분포함수라 하면

$$F(x) = (1-\alpha)\,\Phi(x) + \alpha\,\Phi(x/\sigma)$$

으로 나타낼 수 있다.

Case 0 : $5^2\,N(0,1)$

Case 1 : $5^2\,\{0.9N(0,1) + 0.1\,N(0,5^2)\}$

Case 2 : $5^2\,\{0.9N(0,1) + 0.1\,N(0,10^2)\}$

Case 3 : $5^2\,\{0.9N(0,1) + 0.1\,N(0,15^2)\}$

Case 4 : $5^2\,\{0.9N(0,1) + 0.1\,N(0,20^2)\}$

한편, 모의실험의 모든 계산은 S-PLUS(Statistical Sciences[12])를 이용하여 이루어 졌다.

<표 4>은 모의실험(5000번 반복)에서 얻어진 평균(μ) 및 산포 (σ)의 추정량들에 대한 평균(MEAN)과 평균제곱오차(mean squared error; MSE)를 계산한 결과와 전신뢰도에 필요한 사고건수를 기존모형과 로버스트한 모형에서 계산된 결과이다. <표 4>에서 파라메타의 목표값 (μ, σ)은 분포에 따라 다른데 산포를 표에 계산해 두었다. 여기서 평균제곱오차는

$$\text{MSE} = \frac{1}{5000}\sum(\text{평균 및 산포의 추정치} - \text{평균 및 산포}(\mu,\sigma))^2$$

으로 계산된다.

또한, 모의실험에서 사용된 각 분포군에서 평균의 추정량은 일단계 M−추정량(one-step M-estimator)이며 초기추정량으로 중앙값(median)을, 조율상수(c)로서 1.5를 사용하였다.

〈표 4〉에서 평균추정의 측면에서 보면 표본평균과 M−추정량은 비슷한 정도의 편의를 가지고 있음을 알 수 있다. 그러나 평균제곱오차의 측면에서 보면 $N(0, 1)$ 경우를 제외한 모든 경우에서 M−추정량의 MSE가 표본평균의 그것보다 훨씬 작은 값을 가짐을 알 수 있으며, 극단적인 이상치가 존재하는 경우 및 소표본(small sample)에서도 훨씬 작은 값을 가짐을 알 수 있다. 이로부터 M−추정량의 로버스트한 성질을 알 수 있다.

또한, 산포추정에서는 편의의 측면에서 보면 Gini 평균차이에 의한 추정법이 상대적으로 큰 편의를 보이고 있으나, MSE의 측면에서는 $N(0, 1)$경우를 제외한 모든 고려된 분포에서 Gini 추정법에 의한 산포추정량이 작은 MSE값을 가지고 있어서 산포를 안정적으로 추정하고 있음을 보여주고 있다.

한편, 전신뢰도 사고건수 측면에서도 표본평균과 표본분산에 기초한 전신뢰도 사고건수가 이상치에 크게 영향을 받아 변동이 심한 반면에 M−추정량과 Gini 평균차이를 이용한 추정량에 기초한 사고건수는 이상치에 크게 영향을 받지 않고 안정적임을 알 수 있다.

이로서 로버스트한 추정량에 기초한 신뢰도모형은 이상치가 존재하여 정규성의 가정을 만족하지 못한 경우에 안정적인 모형으로서 위험집단에 대한 정확한 정보를 제공해 주는 모형임을 알 수 있다.

〈표 4〉 로버스트 신뢰도모형의 성능비교

분 포				$N(0, 1)$	Case1	Case2	Case3	Case4
산 포(σ)				5.0	9.2195	16.5076	24.1868	31.97655
평균추정	n=10	Mean	$\hat{\mu}_X$	20.0006	19.9270	19.9109	20.0622	20.0807
			$\hat{\mu}_M$	20.0118	19.9281	19.9257	20.0855	20.1521
		MSE	$\hat{\mu}_X$	2.5000	8.6619	27.5525	55.2199	103.9630
			$\hat{\mu}_M$	2.6207	5.4762	12.4281	22.6342	41.0425
	n=20	Mean	$\hat{\mu}_X$	20.0002	19.9720	20.0376	20.1116	20.0390
			$\hat{\mu}_M$	20.0021	19.9811	19.9961	20.1002	20.0715
		MSE	$\hat{\mu}_X$	1.2844	4.2156	13.0882	29.9508	52.3899
			$\hat{\mu}_M$	1.3539	2.5120	5.6141	11.7882	20.3206
	n=30	Mean	$\hat{\mu}_X$	19.9916	19.9966	19.9755	20.0319	19.9447
			$\hat{\mu}_M$	19.9930	19.9943	19.9845	19.9953	19.9565
		MSE	$\hat{\mu}_X$	0.8558	2.8465	9.4292	19.5831	35.7984
			$\hat{\mu}_M$	0.8897	1.5815	3.8611	7.8304	14.1874
산포추정	n=10	Mean	$\hat{\sigma}_S$	4.8660	8.1237	13.0000	17.8252	23.1922
			$\hat{\sigma}_G$	5.0041	7.6180	11.0153	14.2175	17.7592
		MSE	$\hat{\sigma}_S$	1.3262	22.2311	118.6327	290.7302	553.1792
			$\hat{\sigma}_G$	1.4859	16.0221	89.3122	230.3576	442.9586
	n=20	Mean	$\hat{\sigma}_S$	4.9509	8.42681	14.1679	20.6194	26.7082
			$\hat{\sigma}_G$	5.0156	7.48283	10.8803	14.4563	17.8596
		MSE	$\hat{\sigma}_S$	0.6882	13.1780	69.4715	183.2727	344.7295
			$\hat{\sigma}_G$	0.7482	9.3182	59.4866	163.2228	327.9121
	n=30	Mean	$\hat{\sigma}_S$	4.9657	8.7133	14.9778	21.5939	28.8470
			$\hat{\sigma}_G$	5.0092	7.5580	10.9503	14.3725	18.1945
		MSE	$\hat{\sigma}_S$	0.4608	9.5771	51.2631	129.2266	242.3269
			$\hat{\sigma}_G$	0.4971	7.1648	49.6891	141.3351	275.2576
전신뢰도사고건수	n=10		기존모형	31	38	1,553	15,411	31,718
			새로운 모형	32	36	61	705	1,203
	n=20		기존모형	31	34	57	697	12,562
			새로운 모형	32	36	42	111	504
	n=30		기존모형	31	36	54	122	1,757
			새로운 모형	31	34	41	52	78

Ⅴ. 결 론

본 논문에서는 기존의 변동제한법 형이 갖는 비로버스트성의 문제점을 지적하고 이에 대한 대안으로 로버스트한 추정치에 기초한 신뢰도모형을 즉, 일반적으로 표본평균과 표준편차는 이상점에 대하여 민감하게 반응하므로 평균손해 이상의 거대손해(태풍, 지진, 전쟁)가 발생할 경우에는 좋은 추정치가 되지 못한다. 그러므로 이에 기초한 신뢰도모형은 위험집단에 대한 왜곡된 정보를 제시하는 결과를 갖게 된다. 이러한 문제점을 해결할 수 있는 대안적 방법으로 M－추정량과, Gini 평균차이를 이용한 추정량에 기초한 새로운 신뢰도모형을 제시하였다.

또한 신뢰도모형 분석에 필요한 평균과 산포의 로버스트 추정법으로서 M－추정법과 Gini의 평균차이에 기초한 추정법을 논하고, 이의 유용성을 모의실험을 통하여 입증하였다. 신뢰도모형에서 평균과 산포추정에 관련된 중요한 문제는 이상치가 존재할 경우와 데이터의 개수(위험단위수의 크기)가 아주 작다는 것이다. 이러한 경우에는 일반적인 평균과 산포의 추정량(예를들면 표본평균, 표본표준편차)으로는 편의(bias)와 정도(precision)의 측면에서 효율이 떨어질 수 있으며, 흔히 논의되는 추정량 계산의 간편성은 요즘과 같이 계산지원 소프트웨어가 발달한 상황에서는 설득력이 떨어질 수밖에 없을 것으로 생각된다.

M－추정법은 이상점에 대한 로버스트성이 있으며, Gini의 평균차이를 이용한 산포의 추정은 소표본에서 데이터가 갖는 산포에 관한 정보를 최대한 활용한다고 볼 수 있다. 기존의 표준편차는 n개의 편차(차이)를 이용하여 산포를 추정하고, 범위를 이용한 추정은 1개의 차이(범위)만 이용한다. 그러나 Gini의 평균차이에서는 $n(n-1)/2$개의 차이들을 이용하게 되므로 작은 표본에서 특히 데이터가 갖는 정보의 재사용(reuse)의 측면에서 유용한 산포의 측도로 고려될 수 있다. 시뮬레이션 결과는 Gini 추정법의 이러한 효율성을 입증해 주고 있다.

이러한 로버스트성을 가진 M－추정법과 Gini의 평균차이에 기초한 추정법에 의한 새로운 신뢰도모형은 이상치가 존재할 경우 전신뢰도를 만족하는 위험단위의 수가 안정적임을 또한 알 수 있었다.

참 고 문 헌

〔1〕 송문섭,『로버스트 통계』, 자유아카데미, 1996.

〔2〕 이창수,「신뢰도기법을 이용한 자료의 충분성평가와 보험요율의 조정」,『보험개발연구』, 보험개발원, 1997, 2(21).

〔3〕 Andrews, D.F., Bickel, P.J., Hampel, F.R., Huber, P.J., Rogers, W.H. and Tukey, J.W. *Robust Estimates of Location: Survey and Advances,* Princenton University Press, 1972.

〔4〕 David, H.A., *Gini's mean difference rediscovered,* Biometrika 55, pp.573-575, 1968.

〔5〕 Gini,C., "Variabilita e mutabilita, contributo allo studio delle distribuzioni e delle relazioni statistiche," *Studi Economico-Giuridici della R. Universita di Cagliari.* 3, part 2, 1912, pp.3-159,

〔6〕 Gisler,A. and Reinhard,P., *Robust Credibility,* Astin Bulletin, Vol.23, No.1, 1993.

〔7〕 Herzog, T.N., *Introduction to Credibility Theory,* ACTEX Publications, Inc., 1999.

〔8〕 Huber,P.J.,"Robust Estimation of a Location Parameter," *The Annals of Mathmatical Statistics,* Vol. 35, 1964, pp.73-101,

〔9〕__________,"Robust Statistics: A Review," *The Annals of Mathmatical Statistics,* Vol. 43, 1972, pp.1041-1067,

〔10〕 Lomnicki, Z.A., "The Standard Error of Gini's Mean Difference," *Annals of Mathematical Statistics,* Vol. 23, 1952, pp.635-637,

〔11〕 Nair, U.S., *The Standard Error of Gini's Mean Difference,* Biometrika 28, 1936, pp.428-436,

〔12〕 Statistical Sciences, *S_PLUS for Windows User's* Manual, Siattle: Statistical Sciences, 1994.

Abstracts

This paper attempts to introduce robust methods into the area of credibility. In many actuarial applications excess claims(outliers) which are much bigger than ordinary claims do occur. Excess claims lead to an unsatisfactory behavior of credibility estimators. We suggest in this paper to use robust methods in order to obtain better estimators. The Huber's M-estimator is a well-known robust estimator in sense of distributional robustness. Also, we suggest the use of the Gini's mean difference for the estimation of the σ.

These robust estimators are thus good candidates for dealing with the excess claims in credibility. Our estimator performs well if there are outlying claims and the number of years available is not very small. Therefore, this study aims to look at the credibility estimator based on the robust statistics.