

운영 위험 관련 손실 분포 - 퍼지 히스토그램의 효과

박노진¹

단국대학교 정보통계학과

접수 2009년 3월 10일, 수정 2009년 6월 12일, 게재확정 2009년 6월 20일

요약

히스토그램이 활용의 간편성과 자료의 전체적 구조를 한 눈에 볼 수 있는 정보량을 제공하지만 히스토그램의 계급 구간의 설정에 따라 그 표현이 달라 질 수 있는 문제가 있다. 이러한 문제를 해결하기 위해 퍼지 개념을 활용한 히스토그램이 제안되었고 그 효과가 제시되었다 (Loquin과 Strauss, 2008). 히스토그램이 다양한 분야에서 사용되지만 요즘 운영 위험과 관련된 손실 분포를 추정함에 있어서 유용하게 사용되고 있다. 그런데, 임계치를 활용한 극단치 확률 함수 추정에 사용함에 있어 임계치의 선택에 따른 히스토그램의 모양 변화는 그 활용을 어렵게 하는 경향이 있다. 본 연구는 퍼지 히스토그램을 손실에 대한 극단치 분포를 추정에 사용할 경우 임계치의 선택에 따른 전체적 모양의 차이가 일반적인 히스토그램 보다 크지 않아 상대적으로 안정된 분포를 추정할 수 있음을 보였다.

주요용어: 손실분포, 운영위험, 퍼지함수.

1. 서론

BIS (Bank for International Settlement)의 2003년 보고서에 의하면 운영위험 (operational risk)이란 부적절하거나 잘못된 내부 절차, 시스템, 직원 또는 외부 사건으로 인해 발생하는 손실의 위험을 의미한다. 금융기관의 위험에 대한 국제적 협의가 이루어지는 바젤위원회에서는 각종 보고서를 통해서 ‘소매금융’, ‘도매금융’ 등 8가지 영업 부문에 대하여 ‘내부사취’, ‘외부사취’ 등 7가지 사고에 따른 운영상의 위험을 정의하고 있다. 따라서 금융기관들은 각 영업 부문에 대한 사고로 인한 손실들을 추정하여 예상치 못한 손실에 대해 어느 수준에 비례하여 자기 자본을 확보해야 한다. 운영위험과 관련된 손실 분포는 지수분포 또는 로그정규분포와 같이 꼬리가 긴 확률분포를 따르는 것으로 알려져 있으며, 손실의 규모가 극단적인 경우가 존재하여 분포의 중심부분 보다는 꼬리부분을 정확히 추정하는 것이 필수적이다. 극단적인 위험 분석에 있어 분포의 꼬리부분만 분리하여 추정하는 극단치 이론 (EVT: Extreme Value Theory)를 이용하는 것이 요구된다. 극단치 이론이 생물학적 사건 (Kang, 2005; Han 등, 2007) 등 자연 과학적 현상을 설명하는데 유용하게 사용되고 있으나 경제관련 분야에서도 유용하게 사용될 수 있음을 보였다.

¹ (488-701) 경기도 용인시 수지구 죽전동 126, 단국대학교 정보통계학과, 교수.
E-mail: rjpak@dankook.ac.kr

1.1. 극단치 분포

극단치 이론은 극단적인 사건에 대한 확률적 성질을 연구하는 이론으로서 그 방법론은 일반화 극단치 분포 (GEV: Generalized Extreme Value Distribution)를 추정하는 것과 일반화 파레토 분포 (GPD: Generalized Pareto Distribution)를 추정하는 방법이 있다. 본 논문에서는 GPD의 추정에 관하여 간단히 서술하겠다.

확률변수 X 의 분포함수 F 의 오른쪽 끝점을 $x_F = \sup \{x \in R : F(x) \leq 1\} \leq \infty$ 라고 하자. 임의의 임계치 u 에 대하여 임계치 초과 분포함수를 다음과 같이 정의한다.

$$F_u(x) = P(X - u \leq x | X > u) = \frac{F(u+x) - F(u)}{1 - F(u)}, \quad 0 \leq x < x_F - u.$$

즉, 임계치 초과 분포 함수는 확률변수 X 가 주어진 임계치 u 를 초과할 때 초과크기에 대한 조건부 분포 함수를 의미한다. Balkema와 Haan (1974)과 Pickands (1975)에 의하면 확률변수 X 가 특정한 성질을 갖는다면, $F_u(x)$ 는 아래의 G_{ξ, σ_u} 로 수렴한다고 한다.

$$G_{\xi, \sigma_u}(x) = \begin{cases} 1 - \left(1 + \xi \frac{x}{\sigma_u}\right)^{-1/\xi}, & x \geq 0, 1 + \xi \frac{x}{\sigma_u} > 0, \xi \neq 0, \\ 1 - \exp\left(\frac{-x}{\sigma_u}\right), & x \geq 0, \xi = 0, \end{cases}$$

여기서 $\sigma_u = \sigma + \xi u$.

위의 G_{ξ, σ_u} 를 사용하기 위해서 ξ, u 를 우선 결정해야한다. 임계치를 결정하는 방법은 대표적으로 평균 초과 그림 (mean excess plot)과 힐 그림 (Hill plot)을 이용하는 두 가지 방법이 있다. 먼저 평균 초과 그림을 이용하는 방식은 평균 초과 분포 함수의 그래프를 이용하는데, 만일 $F_u(x)$ 가 GPD함수 $G_{\xi, \sigma_u}(x)$ 이고 $\xi < 1$ 이면

$$\begin{aligned} e_X(u) &= E(X - u | X > u) = \int_0^{x_F - u} dF_u(x), \quad u < x_F, \\ &= \frac{\sigma_u}{1 - \xi} = \frac{\sigma + \xi u}{1 - \xi}, \quad \sigma + \xi u > 0 \end{aligned}$$

으로 정의된다. 실제로는 $(u, \hat{e}_X(u)) : x_{1:n} \leq u \leq x_{n:n}$, $\hat{e}_X(u) = \sum_{i=1}^{n_u} y_i / n_u$ 으로 이루어진 산점도에서 u 의 값을 변화시킴에 따라 양의 기울기를 갖는 직선형태의 영역이 시작되는 점을 임계점으로 정하게 된다 (Hill, 1975).

Hill 그림을 이용하는 경우는 $(k, H_{k,n}) : k = 2, \dots, n$, $H_{k,n} = (1/k) \sum \ln X_{i,n} - \ln X_{k,n}$, 으로 이루어진 산점도에서 $H_{k,n}$ 이 안정된 모양을 보이기 시작되는 값을 임계치 u 의 추정치로 정한다. 여기서 그리고 $X_{i,n}$ 은 n 개의 관측치들을 크기에 따라 내림차순으로 배열하였을 때 i 번째에 관찰치를 표시한다.

1.2. 히스토그램과 퍼지 히스토그램

히스토그램은 단변량 자료의 분포를 설명하는 간단하지만 가장 널리 쓰이는 도구이다. 특별히 운영위험과 관련하여 발생하는 손실의 분포를 표현하기 위해 많은 관련 연구자들이 적극적으로 사용하고 있다 (조하현 등, 2004; Chernobai 등, 2007). 그러나 히스토그램을 사용함에 있어 계급 구간의 수 또는 폭에 따라 히스토그램의 전체적인 모양새가 달라 질수 있으므로 계급의 개수와 계급 구간의 적당한 폭을 결정

해야하는 어려움이 있다. 이러한 문제를 다소나마 해결하기 위해 구간의 폭을 결정함에 따른 애매모호함을 퍼지 개념을 사용하여 반영한 퍼지 히스토그램이 있다 (Loquin과 Strauss, 2008). 사전 연구들은 퍼지 히스토그램이 구간의 폭을 결정함에 수반되는 다소간의 차이에도 자료를 올바르게 표현함을 증거하고 있다.

주어진 표본 공간을 구간 $A_k, k = 1, \dots, p$,로 나누고 각 구간이 같은 너비 h 를 갖는다고 하자. 각 구간 위에서 정의된 특성 함수 $1_{A_k} (1_{A_k}(x) = 1$ 만일 $x \in A_k$ 또는 0 만일 $x \notin A_k$)에 대하여

$$\hat{f}_{hist}(x) = \frac{1}{nh} \sum_{i=1}^n 1_{A_k}(X_i), \quad x \in A_k$$

와 같이 히스토그램이 정의된다.

한편, 퍼지 히스토그램에 대한 개념을 간단히 정리하면 다음과 같다. 표본 공간 $\Omega = [a, b]$ 에 대하여 $a = m_1 < m_2 < \dots < m_p = b$ 가 되는 $p(\geq 3)$ 개의 고정점이 존재한다고 하자. 소속 함수 $\mu_1(x), \dots, \mu_p(x)$ 를 갖는 p 개의 퍼지 부분 집합 A_1, \dots, A_p 가 표본공간을 구성한다고 하자. 아래 조건들을 만족할 때 부분 집합 A_1, \dots, A_p 가 강한 균등 퍼지 분할 (strong uniform fuzzy partition)을 갖는다고 한다.

1. $\mu_{A_k}(m_k) = 1$.
2. 만일 $x \notin [m_{k-1}, m_{k+1}]$, $\mu_{A_k}(x) = 0$.
3. $\mu_{A_k}(x)$ 는 연속형 함수이다.
4. μ_{A_k} 는 $[m_{k-1}, m_k]$ 에서 단조 증가하고 $[m_k, m_{k+1}]$ 에서는 단조 감소한다.
5. 모든 $x \in \Omega$ 에 대하여 $\mu_{A_k} > 0$ 가 되는 k 가 존재한다.
6. 모든 $x \in \Omega$ 에 대하여 $\sum_{k=1}^p \mu_{A_k} = 1$.
7. $k \neq p$ 에 대하여 $h_k = m_{k+1} - m_k = h()$, 즉, $m_k = a + (k - 1)h$.
8. $k \neq 1$ 그리고 $k \neq p$ 에 대하여 $\forall x \in [0, h]$, $\mu_{A_k}(m_k - x) = \mu_{A_k}(m_k + x)$.
9. $k \neq 1$ 그리고 $k \neq p$ 에 대하여 $\forall x \in [m_k, m_{k+1}]$, $\mu_{A_k}(x) = \mu_{A_{k-1}}(x - h) \& \mu_{A_{k+1}}(x) = \mu_{A_k}(x - h)$.

퍼지히스토그램은 주어진 퍼지 함수를 이용하여

$$\hat{f}_{fuzzyhist}(x) = \frac{1}{nh} \sum_{i=1}^n \mu_{A_k}(X_i), \quad x \in A_k.$$

와 같이 정의된다 (Loquin과 Strauss, 2008). 위의 조건들을 만족하는 균등 퍼지 분할을 갖는 대표적인 퍼지 함수는 삼각 퍼지 함수로서, 다음과 같이 정의되며

$$\mu_{A_k}(x) = \begin{cases} (1/h)(m_2 - x)1_{[m_1, m_2]}(x), & k = 1 \\ (1/h)(x - m_{k-1})1_{[m_{k-1}, m_k]}(x) + (1/h)(x - m_{p-1})1_{[m_k, m_{k+1}]}(x), & 1 < k < p \\ (1/h)(x - m_{p-1})1_{[m_{p-1}, m_p]}(x), & k = p, \end{cases}$$

그림 1.1에 $p = 5$ 인 경우의 예시를 수록하였다.

이론적으로 과연 퍼지 히스토그램이 일반적인 히스토그램 보다 뛰어난 적합성을 갖는 지를 확인할 필요가 있다. 모수적 확률함수 추정량으로서 적합도는 IMSE (Integrated Mean Squared Error)를

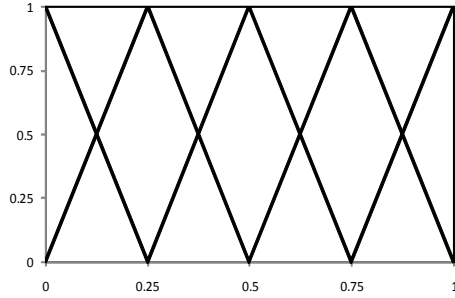


그림 1.1 퍼지 소속함수 예시 ($p = 5$)

통해서 판단할 수 있는데. Waltermann 등 (2005)에 의하면 모델 확률 함수 f 에 대하여 최적 IMSE (IMSE*)는 균등 퍼지 분할을 갖는 소속 함수 μ 에 주어질 때,

$$IMSE^* = \begin{cases} 3\left(\frac{\alpha}{4}R(f')\right)^{1/3}\gamma^{2/3}n^{-2/3} + O(n^{-1}), & \alpha > 0 \\ \frac{5}{4}(\beta R(f''))^{1/5}\gamma^{4/5}n^{-4/5} + O(n^{-1}), & \alpha = 0, \end{cases}$$

$$\alpha = \int_{a-1/2}^{a+1/2} \sum_{i \in \mathbb{Z}} (x-i)\mu(x-i)^2 dx, \text{ for } \forall a \in \mathbb{R},$$

$$\beta = \int_{-1/2}^{1/2} \left(\int w^2 \mu(w) dw + \sum_{i \in \mathbb{Z}} (x-i)^2 \mu(x-i)^2 dx \right),$$

$$\gamma = \int_{-1/2}^{1/2} \int \left(\sum_{i \in \mathbb{Z}} \mu(x-i)\mu(w-i) \right)^2 dw dx,$$

$$R(\phi) = \int \phi^2(x) dx$$

이다. 따라서, 일반적인 히스토그램의 IMSE*는 $O(n^{-2/3})$ 으로 감소하나 위에 보듯이 퍼지히스토그램의 IMSE*는 $O(n^{-4/5})$ 로 감소하고 (만일 $\alpha = 0$ 이면), $O(n^{-2/3})$ 로 감소한다 (만일 $\alpha > 0$). 따라서 적절한 소속 함수를 찾으면 일반적인 히스토그램 보다 퍼지히스토그램이 통계적으로 효율적 (efficient) 일 수 있다. 그런데 삼각형 소속 함수의 경우

$$IMSE^* = \frac{5}{8} \left(\frac{7}{30} R(f'') \right)^{1/5} \gamma^{4/5} n^{-4/5} + O(n^{-1})$$

으로 계산되어 일반적인 히스토그램 보다 효율적인 추정이 가능하다.

1.3. 퍼지 히스토그램의 효과

본 연구에서는 삼각 퍼지 함수를 사용하여 퍼지 히스토그램의 효과를 보이도록 하겠다. 퍼지 히스토그램의 장점을 설명하기 위해 일반적인 히스토그램의 특성함수와 퍼지 히스토그램의 삼각 퍼지 함수의

일부분을 그림 1.2에 그려 보았다. 그림 1.2에서 삼각형으로 표시된 5개의 자료 중에서 일반적인 히스토그램을 적용할 경우 구간이 약간 오른 쪽으로 이동해도 2번째와 5번째 자료는 소속될 구간이 변하지만, 퍼지히스토그램의 경우에는 특정 구간에 속할 소속 정도가 거의 변하지 않음으로 히스토그램의 전체적인 모양새가 크게 변하지 않게 된다.

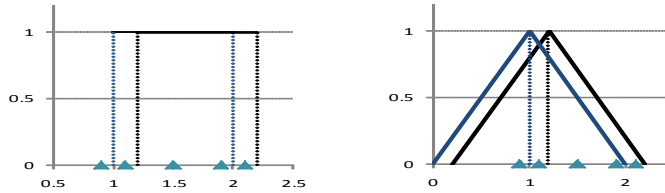


그림 1.2 히스토그램과 퍼지히스토그램에서의 구간변화에 대한 차이

간단한 모의실험을 통해 히스토그램과 퍼지 히스토그램의 모양새를 비교해보자. 모수 $\lambda = 0.1$ 인 지수분포에서 100개의 자료를 무작위 추출하고 90분위수의 근방에 속하는 20과 21보다 큰 자료만 선택한 두 데이터 집합에 대하여 히스토그램과 퍼지히스토그램을 그림 1.3에 그려보았다. 퍼지히스토그램의 경우 20이상인 경우와 21이상인 경우 그 전체적인 모양에서 큰 차이를 보이지 않고 있다. 반면 히스토그램의 경우 우측 꼬리 부분에서 약간의 차이가 있음을 알 수 있다. 실제로 퍼지히스토그램의 두 그림과 히스토그램의 두 그림에 대한 평균 오차 제곱합이 5.14×10^{-4} 과 2.28×10^{-3} 으로 작은 숫자이지만 20배의 차이가 난다. 퍼지히스토그램의 경우 임계값의 차이에 따른 변화가 히스토그램 보다 적음을 알 수 있다.

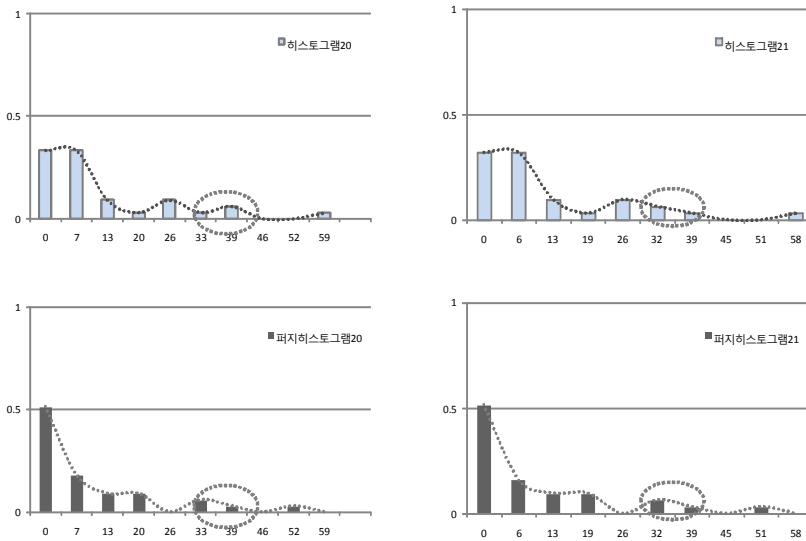


그림 1.3 임계치 (20과 21) 이상의 자료를 이용한 히스토그램과 퍼지히스토그램의 예시

2. 예제

영국의 어떤 금융 기관의 법률 책임에 대한 손실 자료 (Lewis, 2004)를 사용하여 임계점을 구하고 그림 2.2에 히스토그램과 퍼지히스토그램을 그려보았다. 데이터는 총 140개이며 왜도 (skewness)가 2.84이고 첨도 (kurtosis)가 12.81로서 오른쪽으로 꼬리가 긴 분포를 보인다. 편이상 분석의 단위는 1000파운드로 하였다. 그림 2.1의 Hill 그림을 보면 동그라미 친 부분에서 선형의 형태가 처음 나타나고 ME 그림을 보면 동그라미 친 부분에서 처음으로 안정된 형태를 띠게 됨을 볼 수 있다. 그러나, 정확히 어느 시점에서 이러한 변화가 시작되는 지 확실히 지정하기가 어렵다. 시각적으로 판단할 경우, 그 점은 150.95 같기도 하고 151.32 같기도 하여 두 개의 임계점을 이용하여 자료를 분리하여 히스토그램을 그려 보면 아래 그림 2.2에서 보듯이 퍼지 히스토그램의 경우 임계점이 다를지라도 모양새가 크기 차이가 나지 않으나 히스토그램의 경우는 다소간의 차이가 존재함을 볼 수 있다. 서로 다른 두 개의 임계치를 사용하는 경우에 일반적인 히스토그램은 첫 번째와 두 번째 막대기 간에 그리고 네 번째와 다섯 번째 막대기 간에 차이가 퍼지 히스토그램에서의 경우에 비해 차이가 현저히 남을 육안으로도 확인할 수 있다. 실제로 퍼지히스토그램과 히스토그램의 임계치가 다른 두 그림 사이의 평균제곱오차가 각각 1.51×10^{-4} 과 5.91×10^{-4} 으로 비록 숫자는 작지만 5배정도의 차이가 남을 알 수 있다. 따라서 퍼지히스토그램을 사용하는 것이 임계점 선택에 따른 차이를 줄일 수 있고 결국 보다 안정된 추정이 가능함을 알 수 있다.

표 2.1 법률 손실 관련 자료의 통계치

Mean	£ 151,944.04
Median	£ 103,522.90
Standard deviation	£ 170,767.06
Skewness	2.84
Kurtosis	12.81

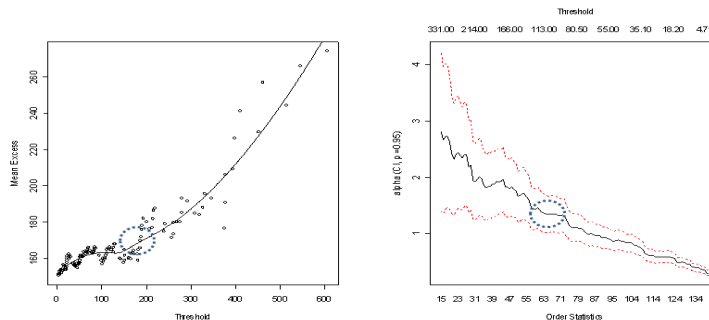


그림 2.1 ME 그림(좌)과 Hill 그림(우)

3. 결론

본 연구는 현재 국제결제은행 (BIS, 2003)의 요구를 만족하기 위한 운영위험의 손실분포 추정에 퍼지 히스토그램을 어떻게 사용하는 지를 구현하였다. 특별히, 임계점을 기준으로 한 극단치들의 분포를 추정함에 있어서 임계점의 다소간의 차이에 대하여 일반적인 히스토그램에 비해 퍼지 히스토그램의 안정

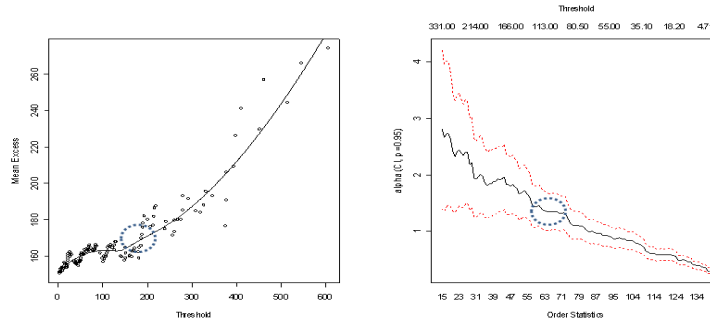


그림 2.2 법률 관련 책임 손실에 대한 히스토그램과 퍼지 히스토그램; 임계치가 150.95와 151.32인 경우

성이 탁월함을 알 수 있었다. 실제로 자료를 분석함에 있어서 임계점을 결정하는 것에 다소의 주관성이 간여함을 고려할 때 임계점 선택에서의 다소간의 차이에도 그 분포의 모양이 크게 변하지 않는 퍼지 히스토그램의 활용성이 기대된다.

참고문헌

- 조하현, 이승국, 김종호 (2004). < 운영리스크-측정과 관리>, 세경사, 서울.
- BIS (2003). *Supervisory guidance on operational risk advanced measurement approaches for regulatory capital*, www.bis.org.
- Balkema, A. and de Haan, L. (1974). Residual life time at great age. *The Annual of Probability*, **2**, 792-804.
- Chernobai, A, Rachev, S. and Fabozzi, F. (2007). *Operational Risk*, John Wiley and Sons, Hoboken, N. J.
- Han, J. T. and Kang, S. B. (2007). Estimation for the generalized extreme value distribution based on multiply type-II censored samples. *Journal of the Korean Data & Information Science Society*, **18**, 317-326.
- Hill, B. M. (1975). A simple general approach to inference about the tail of a distribution. *Annals of Statistics*, **3**, 1163-1173.
- Kang, S. B. (2005). Estimation for the extreme value distribution based on multiply type-II censored samples. *Journal of the Korean Data & Information Science Society*, **16**, 629-638.
- Lewis, N. D. C. (2004). *Operational risk with Excel and VBA*, John Wiley and Sons, Hoboken, N. J.
- Loquin, K. and Strauss, O. (2008). Histogram density estimators based upon a fuzzy partition. *Statistics & Probability Letters*, **78**, 1863-1868.
- Pickands, J. (1975). Statistical inference using extreme order statistics. *Annals of statistics*, **3**, 119-131.
- Walterman, L., Kaymak, U. and Berg, J. van den (2005). Fuzzy histograms: A statistical analysis. *Proceedings of the Joint 4th Conference of the European Society for Fuzzy Logic and Technology*, 605-610.

Fuzzy histogram in estimating loss distributions for operational risk

Ro Jin Pak¹

Department of Information & Statistics, Dankook University

Received 10 March 2009, revised 12 June 2009, accepted 20 June 2009

Abstract

Histogram is the oldest and most widely used density estimator for presentation and exploration of observed univariate data. The structure of a histogram really depends on the number of bins and the width of the bins, so that slight changes on bins can produce totally different shape of a histogram. In order to solve this problem the fuzzy histogram was introduced and the result was good enough (Loquin and Strauss, 2008). In particular, when estimating loss distribution related with operational risk a histogram has been widely used. In this article, instead of an ordinary histogram we try to use a fuzzy histogram for estimating loss distribution and show that a fuzzy histogram provide more stable results.

Keywords: Fuzzy membership function, loss distribution, operational risk.

¹ Professor, Department of Information & Statistics, 126, Dankook University, Chunan 448-701, Korea.
E-mail: rjpak@dankook.ac.kr