Journal of the Korean Data Analysis Society (February 2016)
Vol. 18, No. 1 (B), pp. 379-393

통계청 생명표에 기초한 100세 이상 사망률 연장에 관한 연구

성주호[1], 최장훈[2]

요 약

우리나라는 저출산과 평균수명의 증가로 인구의 고령화가 세계에 유래 없는 빠른 속도로 진행되고 있다. 이에 따라 장수리스크(longevity risk)가 국민연금의 재정 불균형을 야기할 개연성이 증가할 것으로 전망된다. 재정관리가 사전적으로 이루어지기 위해서는 국민연금 인구추계의 정확성이 한층 더 요구된다. 현재 국민연금의 인구추계에 통계청의 자료를 사용하고 있으나 한계연령(age of limit)이 100세로 제한되어 있어 늘어나는 고령 인구를 제대로 반영하지 못하고 있다. 본 연구에서는 통계청 생명표를 기초로 하여 사망률을 100세 이상으로 연장하기 위하여 감마 커널(gamma kernel)과 극단값 이론(extreme value theory)을 접목한 새로운 모형을 제안하고 사망률을 산출하였다. 전통적 곰퍼츠(Gompertz) 모형과 극단값 이론 모형을 비교 검증대상을 설정하였다. 검증 결과 이들 모형 간에 큰 차이는 없었으나 한계연령이 다소 높고 신뢰구간도 더 넓게 나타났다. 또한 남자와 여자의 한계연령에 차이가 거의 없고 여자의 한계연령은 시간이 지남에 따라 오히려 감소하는 이상 현상이 발생되었다. 마지막으로 본 연구가 향후 초고령층의 사망률 추정 논의에 기여하기를 바란다.

주요용어 : 고령화, 감마커널, 곰퍼츠, 극단값이론, 한계연령.

1. 서론

유엔(UN)이 제시한 기준에 따르면 한 국가의 65세 이상의 인구비중이 전체인구의 7% 이상이면 고령화 사회(aging society), 14% 이상이면 고령 사회(society of advanced age), 그리고 20% 이상이면 초고령 사회(superaged society)로 분류된다.

인구고령화는 OECD 국가에서 일반적인 현상이다. 현재 우리나라의 65세 이상 노령 인구는 다른 국가들과 비교할 때 심각한 수준은 아니지만 문제는 고령화 추세이다. OECD 국가들의 고령화 추세를 보면 Table 1과 같이 고령화 사회에서 고령 사회로 진입하는 데 걸린 시간이 미국은 71년, 독일은 40년, 일본은 24년이 걸렸다. 또한 고령 사회에서 초고령 사회로 진입하는 데 걸린 시간은 미국 27년, 독일 42년, 일본 12년으로 예상된다. 하지만 우리나라의 고령화 추세는 다른 국가들이 경험하지 못한 빠른 속도로 진행되고 있다. 2012년 통계청 자료에 따르면 우리나라의 전체 인구 중 65세 이상의 노인 인구 비율은 2001년에 7.6%로 이미 고령화 사회에 속해 있었고 2017년에는 14%가 되어 고령사회 진입에 걸리는 기간이 17년이 예상되고, 2026년에 노인 인구 비율이 20.8%가 되어 9년이라는 짧은 기간에 초고령 사회로 진입할 것으로 예상된다.

이렇게 인구가 고령화 되는 이유는 출산율 저하와 평균수명의 증가 때문으로 볼 수 있다. 우리나라의 여성 1명이 15~49세 가임기간 동안 낳을 평균자녀수인 합계출산율은 1960년 6.0명에서

[1](교신저자) 02447 서울특별시 동대문구 경희대로 26, 경희대학교 경영대학 교수. E-mail: jhsung@khu.ac.kr
[2]07328 서울시 영등포구 국제금융로6길 38 한국화재보험협회빌딩, 보험연구원 연구위원.
E-mail: james021@kiri.or.kr
[접수 2016년 1월 19일; 수정 2016년 2월 1일, 2016년 2월 17일; 게재확정 2016년 2월 20일]

1983년 2.1명으로 낮아졌고, 2000년대에 들어서는 1.3명 미만으로 추락한 반면에 평균수명은 지속적으로 개선되고 있다.

Table 1. Ageing speed by country

	Aging society->Society of advanced age	Society of advanced age->Superaged society
US	71 years	27 years
Germany	40	42
France	115	43
Italia	61	24
Japan	24	12
Korea	17	9

Source: UN(2010): The world population prospect: the 2010 revision.
Statistics Korea (2012): Population Projections

이에 따라 국민연금의 재정 부담이 급격하게 높아질 것으로 전망되어 재정의 안정적인 관리가 필요하게 되었다. 이를 위하여 국민연금의 재정을 정확하게 추계하는 것이 필요하고 이의 기본 전제인 인구추계의 정확성이 먼저 이루어져야 한다.

보험업을 비롯한 공적연금에서도 공통적으로 고연령 사망률 추정을 위한 데이터의 부족으로 인하여 이를 극복할 수 있는 예측모형의 개발이 필요하다. 현재 대부분의 공적연금제도는 통계청 자료에 근거하여 인구추계를 시행하고 있지만 다음과 같은 근본적 문제가 있다.

첫째, 출산율과 사망률이 너무 높게 추정되었다는 논란이 있다. 이럴 경우 국민연금 재정을 낙관적으로 추계할 개연성이 높다. 둘째, 100세 이상의 고령자에 대한 통계청 경험 자료가 없다. 현재 100세를 상한 연령으로 처리하고 있어 100세 이상 인구를 과소추계하고 있다. 실제로 2010년의 인구실적값은 4,941만명이지만, 인구추계값은 4,887만명으로 약 54만명이 과소 추계되어 있다. 셋째, 통계청의 사망 최고연령이 연도별로 다르게 집계되어 있고, 연령구간도 부족한 상황이다. 통계청의 사망 최고연령은 1970~1992년까지는 80세, 1993년~1998년까지는 85세, 1999~2000년까지는 95세, 그리고 2001년부터 100세로 상향되어 왔다. 또한 각 세별 완전생명표는 1997년 이후에야 비로써 작성되었고 그 이전에는 5세별로 연령구간을 묶은 간이 생명표만 작성되었다. 마지막으로, 2060년 이후의 사망률은 향후 사망률 개선효과를 반영하고 있지 못하다. 국민연금의 재정추계는 매 5년마다 향후 70연간을 추계하므로 인구추계의 정확성이 문제로 제기되고 있다.

그러므로 사망률 전망에 대한 가정과 방법론을 연구하여 정확한 예측모형을 설정하여 통계청 자료보다 현실에 더 가까운 사망률을 산출할 필요가 있다.

본 연구에서는 통계청 자료에 기초하여 100세 이상 고연령 사망률을 연장시키는 방법을 연구한다. 본 논문은 다음과 같이 구성되어 있다. 2장에서는 선행연구로 고연령 사망률 추정에 관한 문헌연구들을 소개한다. 3장에서 본 연구에서 사용한 100세 이상 연령의 사망률 추정 방법을 설명하고 추정결과를 제시한다. 마지막으로 4장에서 결과를 요약한다.

2. 고연령 사망률 추정에 관한 문헌연구

사망률은 1년간 발생한 사망자수를 해당 연도의 연앙인구로 나눈 수치로 보통 성별·연령별 사망률로 구분하여 작성한다. 사망률은 대부분의 재정과 정책결정에 필수적이므로 오래전부터 사망률에 관한 다양한 연구가 이루어져 왔다. 이에 관한 해외연구로는 Lee, Carter(1992)를 비롯하여, Lee, Lee(2005), Cairns(2006), Gompertz(1825), Jeon, Kim(2013), Panjer., Russo(1992), Panjer., Tan(1995),

Heligman, Pollard(1980), Coale, Guo(1989), Coale., Kisker(1990), Himes et al.(1994), Li et al.(2008) 등이 있고 국내연구로는 Park, Kim(2011), Baek et al.(2013), Moon(2011), Choi(2015) 등이 있다. 본 장에서는 이러한 연구들 중 본 논문에서 직접 적용되는 연구들에 대해 소개한다.

2.1. Gompertz 모형

Gompertz(1825)가 제안한 모형으로 사망은 연령과 관계없는 우연과 사망에 대항하는 능력의 악화, 이 두 요소에 의해 발생한다는 원리에 의해 만들어진 모형이다. 전자의 경우 사망자수는 연령이 증가함에 따라 등비수열(geometric progression) 형태로 나타난다. 후자는 이것보다 더 빠르게 나타나게 됨을 발견하여 이 두 가지 사망요인에 근거한 사력(force of mortality)을 $\mu_x = BC^x$와 같이 모델화하였다. 연령이 증가함에 따라 사력이 증가하기 때문에 질병사망이 많은 유아 연령대 또는 재해사망이 많은 저연령대에서는 사용이 제한될 수 있다.

사력 $\mu_x = BC^x,\ B>0, C>1$로 표시할 수 있는데 여기서 B는 기준 사망률, C는 연령에 따른 노화 요소, 그리고 x는 연령이다. 사력에 따라 사망확률을 구할 수 있다. 사망확률 q_x는 1년동안 발생산 사망자수를 연초의 인구수로 나눈 값으로 q_x를 유도하면 다음 식(1)과 같다:

$$q_x = 1-e^{-\int_x^{x+1}\mu_x dt} = 1-e^{-\int_x^{x+1}BC^t dt} = 1-e^{-\frac{BC^x(C-1)}{\ln C}} \quad (1)$$

식(1)은 가장 오랜 동안 사용되어 온 사망률 모형 중 하나로 B와 C는 King-Hardy 방법 또는 최소제곱법으로 구할 수 있다. 이 모형은 저연령에서 과소 추정되며, 고연령에서 과대 추정되는 경향이 있는 것으로 알려져 있다(Bongaarts, 2005).

2.2. 극단값 접근법

계리사와 인구통계학자는 고연령 사망률 추정을 위하여 다양한 해법들을 제안하였다. Panjer, Russo(1992) 그리고 Panjer, Tan(1995)은 100세 초과 연령에 대한 생존분포를 연장하기 위해 3차 다항식을 이용하였다. Heligman, Pollard(1980)는 Gompertz 법칙(Gompertz, 1825)의 비연속형을 개발하여 사망률 모형을 수립하였다. Coale, Guo(1989)는 고연령의 사망률이 개선되어 온 효과를 반영한 지역별 새로운 생명표를 작성하였다. 반면 Coale, Kisker(1990)는 미국의 고연령 사망률을 인구조사 자료로부터 직접 계산하면 너무 낮다는 사실을 발견하여 고연령 사망률 추정의 필요성을 제기하고 새로운 추정 절차를 제시하였다. 한편 Himes et al.(1994)은 검증된 국제 데이터를 이용하여 표준사망률을 제안하였고 개인사망률과의 차이는 표준사망률의 이모수 로짓변형에 의해 설명될 수 있음을 제시하였다. 하지만 이 방법들의 단점은 고령 사망률 가정에 대한 통계적 적합성이 부족하다는 것이다. 이를 보완하기 위해 극단값 이론을 사용할 수 있는데 이에 관한 연구로는 Li et al.(2008), Sohn, Liang(2014), Yun(2012), 그리고 Park et al.(2006) 등이 있다. 본 절에서는 Li et al.(2008)의 방법을 소개한다.

Li et al.(2008)은 극단값 이론을 적용한 생명표(threshold life table)를 제안하였다. 이는 극단값 이론(extreme value theory)을 사망률의 매개변수를 포함하는 모형과 구조적으로 통합하는 것이다. 이 이론에 따르면 임계연령(극단값 이론이 적용되는 최소연령: threshold age) 이상의 생존분포를 정확한 사망률 데이터 없이 산출할 수 있고 생명표가 어떤 방식으로 끝맺음을 할 수 있는지를 통계적 분포를 사용하여 결정할 수 있다. 즉, 생명표에서 사망률이 1이 되는 마지막 한계연령 ω를 알 수

있고, 어떤 연령에서 극단값 분포로 바뀌는지를 통계적 수단을 통하여 알 수 있다.

극단값 이론에서 사망시점에서의 누적분포 $F(x)$는 다음 식(2)와 같이 표현된다:

$$F(x)=\begin{cases}1-\exp\left\{-\dfrac{B}{\ln C}(C^x-1)\right\}, x \le N \\ 1-p\left\{1+\gamma(\dfrac{x-N}{\theta})\right\}^{-1/\gamma}, \ x > N.\end{cases} \tag{2}$$

여기서, $p=S(N)=1-F(N)$: 생존함수, N: 임계연령, $B>0,\ C>0,\ \theta>0$.

즉, 임계연령 이전에는 생존함수가 Gompertz 모형을 따르고 이후에는 일반화 파레토 분포(generalized Pareto distribution)를 따른다.

Li et al.(2008)은 최대가능도추정법(maximum likelihood estimation)을 통하여 임계연령 N과 B, C, γ, θ를 추정하였다. 100세 이상의 데이터가 존재하는 것으로 가정하고 65세 데이터를 시작점으로 하면 가능도함수는

$$L(B,C,\gamma,\theta;N)=\left\{\prod_{x=65}^{99}(\frac{S(x)-S(x+1)}{S(65)})^{d_x}\right\}(\frac{S(100)}{S(65)})^{l_{100}}$$

로 나타낼 수 있고, 로그가능도함수는 다음과 같이 두 로그가능도함수의 합이 된다:

$$\ln\{L(B,C,\gamma,\theta;N)\}=\ln\{L_1(B,C;N)\}+\ln\{L_2(\gamma,\theta;N)\}.$$

구체적으로 살펴보면

$$\ln\{L_1(B,C;N)\}=\sum_{x=65}^{N-1}[d_x\ln\{S(x)-S(x+1)\}]+l_N\ln\{S(N)\}-l_{65}\ln\{s(65)\}$$

여기서 d_x는 x세와 $x+1$세 사이의 사망자수, l_x는 x세에서 생존자 수를 각각 의미한다. 그리고

$$\ln\{L_2(\gamma,\theta;N)\}=\sum_{x=N}^{99}\left[d_x\ln\left\{\frac{S(x)-S(x+1)}{S(N)}\right\}\right]+l_{100}\ln\left\{\frac{S(100)}{S(N)}\right\}$$

가 된다.

따라서 주어진 N에서 $\ln(L_1)$과 $\ln(L_2)$가 각각 최대가 되도록 하는 $\hat{B}(N)$, $\hat{C}(N)$, $\hat{\gamma}(N)$, $\hat{\theta}(N)$을 구하고 로그가능도함수 $\ln\{L(\hat{B}(N),\hat{C}(N),\hat{\gamma}(N),\hat{\theta}(N);N)\}$을 최대가 되도록 하는 N을 구할 수 있다.

2.3. 새로운 감마커널밀도추정량(new Gamma Kernel density estimator; GKDE)에 의한 사망률 추정

커널밀도추정량(Kernel density estimator; KDE)은 각 데이터 포인트에 커널 밀도(kernel density)라는 작은 핵(bump)을 위치시켜 원 데이터의 밀도를 추정하는 방법으로써 다양한 형태의 분포에 적합이 잘 되는 장점이 있다. KDE의 대표적인 형태는 다음 식(3)과 같이 나타낼 수 있다:

$$\hat{f}(x)=\frac{1}{n}\sum_{i=1}^{n}k_b(x-X_i)=\frac{1}{n}\sum_{i=1}^{n}\frac{1}{b}k\left(\frac{x-X_i}{b}\right). \tag{3}$$

여기서 $K_b(\cdot)=K(\cdot/b)/b$, K는 kernel 함수, 그리고 b는 띠폭(bandwidth)이고 양수이다. 하지만 이러한 kernel은 음수에도 양수에서와 같은 비중을 주기 때문에 사망률과 같은 양의 값만을 갖는 데이터에는 적용시키기가 어렵다. 이러한 문제를 해결하기 위해 데이터를 변형시킨 후 Gaussian kernel과 같은 대칭 kernel에 적용할 수 있지만 비대칭적이고 양의 값을 갖는 데이터에 직접 적용할 수 있는 감마 커널(gamma kernel)을 사용하는 방법을 고려할 수 있다. Chen(2000)은 다음과 같은 Gamma 분포 밀도 함수를 제안하였다. 즉,

$$G_{a,b}(t)=\frac{t^{a-1}e^{-t/b}}{\Gamma(a)b^a}.$$

여기서 a,b, 그리고 t는 모두 양수이다. 이 밀도 함수를 kernel로 하여 b를 띠폭으로 하는 감마커널 밀도추정량(gamma kernel density estimator; GKDE)은 아래 식(4)와 같이 나타낼 수 있다:

$$\widehat{f_{CH}}(x)=\frac{1}{n}\sum_{i=1}^{n}G_{x/b+1,b}(X_i)=\frac{1}{n}\sum_{i=1}^{n}\frac{X_i^{x/b}e^{-X_i/b}}{b^{x/b+1}\Gamma(x/b+1)}. \tag{4}$$

Chen(2000)은 GKDE는 언제나 양의 값을 갖고 양의 값을 갖는 KDE 중에서 최적의 수렴속도를 나타냄을 보였다.

하지만, Jeon, Kim(2013)은 GKDE는 표본의 크기가 무한히 커지게 되면 실제 밀도에 수렴하게 되지만 유한한 표본에서는 밀도 함수가 될 수 없고 그렇기 때문에 표본의 크기가 작은 경우 심각한 문제가 발생할 수 있다는 점과 표준분포 값을 정확하게 구할 수 없는 단점이 있음을 지적하였다.

이러한 단점을 보완하고자 Jeon, Kim(2013)은 새로운 감마커널밀도추정량(new gamma kernel density estimator; GKDE)을 고안하였다. 이를 식으로 나타내면 아래 식 (5)와 같다.

$$\widehat{f_G}(x)=\frac{1}{n}\sum_{j=1}^{n}G_{x_j/b+1,b}(x)=\frac{1}{n}\sum_{j=1}^{n}\frac{x^{x_j/b}e^{-x/b}}{\Gamma(X_j/b+1)b^{X_j/b+1}} \tag{5}$$

가 된다. X_j는 j번째 관찰된 수치이다. 그리고 이 모델의 분포함수는

$$\widehat{F_G}(y)=\int_0^y\widehat{f_G}(S)ds=\frac{1}{n}\sum_{j=1}^{n}\Gamma(x_j/b+1;y/b)$$

가 된다. 여기서 $\Gamma(a;y)$는

$$\Gamma(a;y)=\frac{1}{\Gamma(a)}\int_0^y s^{a-1}e^{-s}ds, a>0, y>0$$

로 나타낼 수 있다.

이 모델의 특징은 다음과 같다.

(1) Chen(2000)의 GKDE와 다르게 어떠한 유한한 표본에 대해서도 유효한 밀도함수가 된다.
(2) 따라서 모먼트(moment) 등과 같은 통계값은 정확하게 구할 수 있어서 계산이 용이하다.
(3) 이 모델의 적합 알고리듬이 간단하고 빠르다. 그 이유는 결정해야 할 매개변수가 띠폭 하나뿐이기 때문이다.

띠폭은 유한한 표본의 경우에 추정된 밀도의 평활성(smoothness)을 결정한다. 만약 띠폭이 작아지면 추정된 밀도는 평활성이 적게 이루어지고 커지면 많이 이루어지게 된다. 따라서 추정된 kernel 밀도가 실제 밀도를 잘 나타낼 수 있도록 적절한 평활을 이루도록 띠폭을 결정하는 것이 중요하다.

실제로 주어진 데이터에 대한 적절한 띠폭을 결정하는 유일한 방법은 존재하지 않는다. kernel 평활을 결정하기 위한 간편한 일반적인 접근법은 교차검증(cross validation) 방법이다. 전형적인 교차검증 방법은 작은 표본의 경우 안정성이 떨어지는 문제가 있는데 이를 해결하기 위해 무작위 분리 교차검증(RSCV: random split cross-validation)을 사용할 수 있다. 이 방법은 다음과 같이 설명할 수 있다.

첫째, 표본 크기 n을 m-1 : 1 의 비율로 임의의 두 세트로 나눈다. 앞의 세트는 조율 세트(training set)로 불리고 뒤의 세트는 테스트 세트(test set)로 불린다. m은 보통 5 또는 10으로 정한다.

둘째, 미리 정해진 일련의 띠폭 값들을 사용하여 각 값에 대해 조율 세트(training set)로부터 GKDE를 구한다.

셋째, 테스트 세트(test set)의 로그가능도들의 절삭(양 끝의 2.5%를 절삭)된 평균을 극대화하는 띠폭 b를 선택한다. 이 띠폭을 $b_{(j)}$로 정한다.

넷째, 위의 제1~3단계를 $j=1,\cdots,\nu$에 대해 반복하여 $b_{(1)}, b_{(2)},\cdots, b_{(\nu)}$를 구한다.

마지막으로, $b^{*}=median(b_{(1)}, b_{(2)}, \cdots, b_{(\nu)})$ 를 띠폭으로 결정한다.

제3단계에서 절삭된 평균을 사용하게 되면 test set의 적합성 검증이 정확해 진다. 왜냐하면 절삭하게 되면 training set와 test set가 많이 다를 경우 종종 발생하는 로그가능도를 지나치게 크게 만드는 데이터의 영향을 완화시켜주기 때문이다.

충분히 큰 표본으로부터 GKDE모형을 사용하면 끝부분을 포함한 데이터 전 범위에서 만족스러운 결과를 얻을 수 있다. GKDE는 주어진 데이터를 Gamma 밀도들의 결합으로부터 결정하는 것이기 때문에 표본이 충분히 크지 않을 경우 끝부분을 구하기 위해서는 Gamma 밀도를 사용하여 추정을 하여야 한다.

3. 100세 이상 사망률 추정

3.1. 추정방법

본 연구에서는 Li et al.(2008)과 Kim(2013)의 극단값 이론을 이용한 고연령 사망률을 Jeon, Kim (2013)의 new namma Kernel density estimator(new GKDE)에 적용하여 추정하였다. new GKDE 모형은 평활을 시키는 기능을 수행하기 때문에 생명표상의 사망률을 유지하면서 임계연령 이후의 사망률과의 연결을 유연하게 이룰 수 있다. 또한 추정해야 할 매개변수가 하나 밖에 없으므로 추정의 정확성을 높일 수 있다. 적용절차는 아래와 같이 요약할 수 있다.

통계청의 생명표에 극단값 이론을 적용하여 임계연령 이전 모형에 new GKDE를 사용하였다. 이미 설명한 바와 같이, 임계연령 이전은 new GKDE 분포를, 임계연령 이후는 일반화 파레토 분포를 따른다고 가정하였다.

Li et at.(2008)과 같이 최대가능도법(maximum likelihood)을 사용하면 로그가능도함수는 다음 식 (6)과 같이 두 로그가능도함수의 합으로 나타낼 수 있다:

$$l(b,\gamma,\theta;N)=l_1(b;N)+l_2(\gamma,\theta;N). \quad (6)$$

임계연령 이전 즉, $x \le N$일 경우에는 아래와 같이 $l_1(b;N)$을 사용하고 임계연령 이후 즉, $x > N$일 경우에는 다음의 $l_2(\gamma,\theta:N)$를 사용한다. 즉,

$$l_1(b;N)=\sum_{x=65}^{N-1}[d_x \ln\{S(x)-S(x+1)\}]+l_N \ln\{S(N)\}-l_{65}\ln\{s(65)\}$$

$$l_2(\gamma,\theta;N)=\sum_{x=N}^{99}\left[d_x \ln\left\{\frac{S(x)-S(x+1)}{S(N)}\right\}\right]+l_{99+1}\ln\left\{\frac{S(99+1)}{S(N)}\right\}$$

여기서

$$S(x)=1-\int_0^x \hat{f}_G(S)ds=1-\frac{1}{n}\sum_{j=1}^{n}\Gamma(x_j/b+1;x/b)$$

$$\frac{S(x)}{S(N)}=(1+\gamma(\frac{x-N}{\theta}))^{(-1/\gamma)},\ \ \theta>0$$

따라서 주어진 N에서 l_1과 l_2가 각각 최대가 되도록 하는 $\hat{b}(N),\hat{\gamma}(N),\hat{\theta}(N)$을 구하고 로그가능도함수 $l(\hat{b}(N),\hat{\gamma}(N),\hat{\theta}(N);N)$을 최대가 되도록 하는 N을 구할 수 있다.

이해의 편의상, 임계값 연령과 매개변수들을 추정하는 절차를 요약하면 다음과 같다.

제1단계: N=98에 대해

(a) l_1을 최대값이 되도록 하는 $\hat{b}$를 결정한다.

(b) l_2를 최대값이 되도록 하는 $\hat{\gamma}$와 $\hat{\theta}$를 결정한다.

(c) $l_1(\hat{b};N)$과 $l_2(\hat{\gamma},\hat{\theta};N)$로부터 $l(\hat{b}(N),\hat{\gamma}(N),\hat{\theta}(N);N)$을 계산한다.

제2단계: N=97, 96, …, 85에 대해 1 단계를 반복한다.

제3단계: $l(\hat{b}(N),\hat{\gamma}(N),\hat{\theta}(N);N)$을 최대값이 되도록 하는 임계연령 N을 결정한다.

극단값 이론에 따르면, 임계연령을 초과한 연령기간은 일반화 파레토 분포를 따른다. 즉, X가 연령, N이 임계연령이라면 $Y=X-N|X>N$의 분포는

$$F_Y(y)=1-(1+\gamma\frac{y}{\theta})^{-1/\gamma}$$

여기서 θ는 항상 양수이고, $\gamma>0$이면 Y는 파레토분포, $\gamma=0$이면 Y는 지수분포 그리고 $\gamma<0$이면 Y는 베타분포(beta distribution)를 각각 따르게 된다. 베타분포일 경우 우측 극대값은 $-\theta/\gamma$가 되므로 이 경우의 최대 생존연령은 $N-\theta/\gamma$가 된다.

왜냐하면 Y_j $(j=1,2,\cdots,n)$가 임계연령을 초과한 연령 기간들이고 Y_j가 베타분포를 따르면 $M_n=\max\{Y_j, j=1,2,\cdots,n\}$의 분포는

$$F_{M_n}(y)=\begin{cases}0 & \text{if } y<0\\ [F_Y(y)]^n & \text{if } 0\le y\le-\theta/\gamma\\ 1 & \text{if } y\ge-\theta/\gamma\end{cases}$$

가 된다. 따라서 n이 무한대로 커지면, M_n의 분포는

$$\lim_{n\to\infty}F_{M_n}(y)=\begin{cases}0 \text{ if } y<-\theta/\gamma\\ 1 \text{ if } y\ge-\theta/\gamma\end{cases}$$

가 된다. 그러므로 임계연령 이상에서 살아있는 사람 수가 많아지면 한계연령 ω는 $N-\frac{\theta}{\gamma}$로 수렴하게 된다.

극단값 이론을 적용한 경우 한계연령 ω의 신뢰구간을 계산할 수 있다. 먼저 ω의 추정값인 $\hat{\omega}$의 점근적분산(asymptotic variance)은 델타 방식(delta method)을 사용하여 다음과 같이 산출할 수 있다.

$$Var(\hat{w}) = \left[\frac{\partial w}{\partial \gamma}\ \frac{\partial \omega}{\partial \theta}\right][I(\gamma,\theta)]^{-1}\begin{bmatrix}\frac{\partial \omega}{\partial \gamma}\\ \frac{\partial \omega}{\partial \theta}\end{bmatrix}.$$

여기서 $I(\gamma,\theta)$는 정보행렬(information matrix)이다. ω가 정규분포를 따른다고 가정하면 ω의 95% 신뢰구간은

$$\left[\hat{\omega}-1.96\sqrt{Var(\hat{\omega})}\,,\hat{\omega}+1.96\sqrt{Var(\hat{\omega})}\,\right]$$

가 된다.

3.2. 추정 결과

100세 이상 사망률의 추정 결과를 보여주기 전에 먼저 Gompertz 모형과 new GKDE 모형의 특성을 알아보도록 한다. 두 모형의 비교를 위해 일반 사용이 허용된 통계청 마이크로데이터(2012)의 여자 사망자수를 사망확률로 전환하여 사용하였다. 각 모형의 모수값을 변화시켜 모수값의 변화에 따른 적합성 정도를 비교하면 Figure 1과 같다. Gompertz 모형은 모수값들의 작은 변화에도 곡선의 위치는 크게 달라지지만 곡선의 형태에는 큰 변화가 없다. 반면에 new GKDE 모형은 모수값이 커질수록 평활이 많이 이루어지고 모수값이 작아질수록 실제 데이터 형태로 적합이 잘 이루어짐을 알 수 있다.

이제 실제 사망확률 추정결과를 살펴보도록 한다. 실제 추정에는 마이크로데이터는 사용하지 않고 일반 사용이 허용된 데이터인 통계청 생명표를 사용하였다. 본 연구와의 비교를 위해 Li et al.(2008)이 사용한 모형(Gompertz+Pareto)으로도 추정을 하였다.

본 연구의 모형인 “new GKDE+Pareto”와 “Gompertz+Pareto”와의 차이를 2012년을 기준으로 보면, Figure 2와 Table 2와 같이 2012년 한계연령(age of limit)은 남자의 경우 ″Gompertz+Pareto″ 모형이 108.23세, 그리고 ″new GKDE+Pareto″ 모형이 110.87세로 나타났다. 또한 델타방식을 이용하여 계산된 한계연령 ω의 95% 신뢰구간은 “Gompertz+Pareto” 모형이 (107.17, 109.29) 그리고 “new GKDE+Pareto” 모형이 (108.34, 113.39)로 나타났다. 비슷한 방법으로 여자의 경우는 “Gompertz+Pareto” 모형이 109.28세 그리고 “new GKDE+Pareto” 모형이 110.52세로 나타났다. ω의 95% 신뢰구간은 ″Gompertz+Pareto″ 모형이 (108.62, 109.93), ″new GKDE+Pareto″ 모형이 (107.42, 113.63)으로 나타나 신뢰구간이 Gompertz+Pareto 모형보다 약간 넓게 나타났다.

Park, Kim(2011)의 결과와 비교하면, 남자의 한계연령은 PRS 모형(HPC 모형에 모수 2개 추가)의 경우는 115세를 초과하고, S-WEIB 모형(WEIB 모형에서 모수 1개 추가)의 경우는 107세, CUB 모형(연령에 관한 3차 함수)의 경우는 110세이므로 본 연구의 결과는 CUB 모형의 결과와 유사한 것으로 나타났다. 또한 여자의 한계연령은 PRS 모형의 경우 남자와 마찬가지로 115세를 초과하고, S-WEIB 모형의 경우는 106세, CUB 모형의 경우는 113세이므로 본 연구의 결과는 S-WEIB 모형과 CUB 모형의 중간 정도로 나타났다.

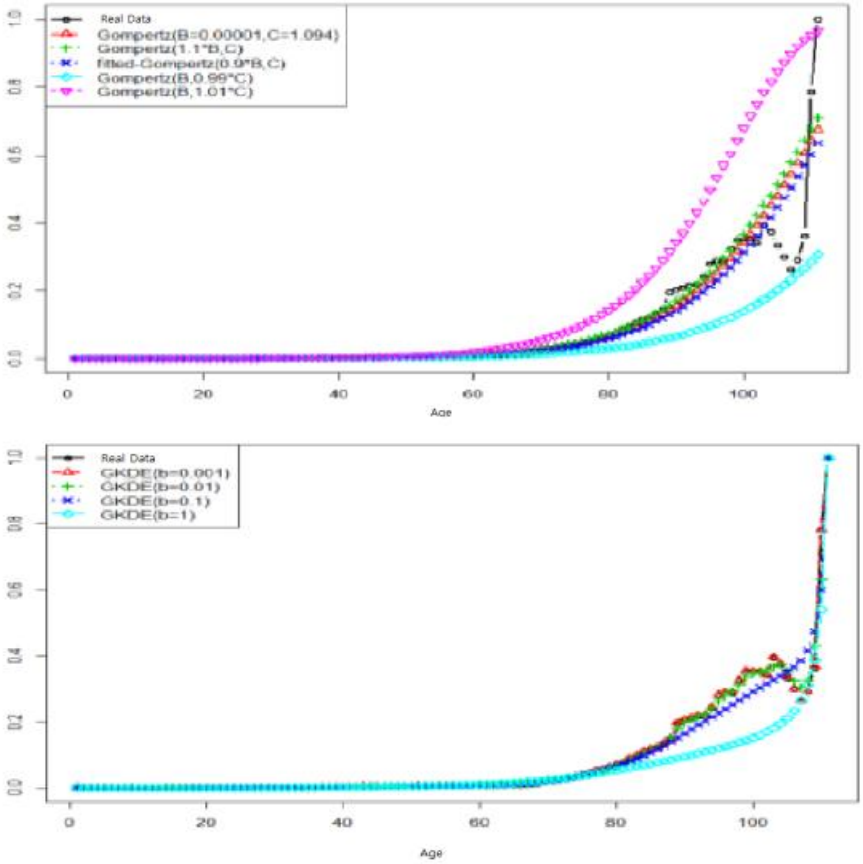

Figure 1. Conformance test: Gompertz vs. new GKDE

Table 2. Estimation results: Gompertz+Pareto vs. new GKDE+Pareto

male (2012)	Gompertz+Pareto	new GKDE+Pareto
threshold age (N)	90 years old	93 years old
age of limit (ω)	108.23	110.87
95% confidence interval of ω	(107.17, 109.29)	(108.34, 113.39)
Sum of squares error (SSE)	0.0010	0.0022
footnote: SSE calculates the difference with Statistics Korea from age 65 65 to 99.		
female (2012)	Gompertz+Pareto	new GKDE+Pareto
threshold age (N)	89 years old	96 years old
age of limit (ω)	109.28	110.52
95% confidence interval of ω	(108.62, 109.93)	(107.42, 113.63)
Sum of squares error (SSE)	0.0343	0.0008
footnote: SSE calculates the difference with Statistics Korea from age 65 65 to 99.		

본 연구의 결과로 특이한 점은 남녀 간 한계연령을 비교하면 일반적으로 여자의 평균수명이 남자보다 높기 때문에 한계연령도 그럴 것이라고 판단할 수 있으나 본 연구의 결과에 의하면 그렇지 않은 것으로 나타났다. 남자와 여자 한계연령의 신뢰구간을 비교해도 신뢰구간이 거의 겹쳐있으므로 성별 한계연령의 차이를 의미있게 해석하기 어렵다. 실제로 통계청 사망확률을 분석해 보면

Table 3과 같이 65세 이상의 경우 여자의 사망확률이 남자의 사망확률보다 낮지만 나이가 많아질수록 사망확률의 증가율은 여자의 경우가 남자보다 높은 것으로 나타났다. 이러한 사실은 본 연구의 결과를 뒷받침해 준다고 볼 수 있다.

2012년의 편차제곱합(SSE)을 비교하면 Gompertz+Pareto 모형과 new GKDE+Pareto 모형에서 남자의 경우는 Gompertz+Pareto모형이, 여자의 경우는 new GKDE+Pareto모형이 우수한 것으로 나타났다.

Table 3. Increasing rates of probability of death by age (Statistics Korea, 2012)

age	prob. of death (male)	increasing rate (male)	prob. of death (female)	increasing rate (female)
65	0.01275	7.7%	0.00469	10.4%
70	0.02222	13.0%	0.00915	15.4%
75	0.03885	12.6%	0.01791	14.9%
80	0.06709	12.2%	0.03727	15.8%
85	0.11400	10.5%	0.07319	13.7%
90	0.17799	8.5%	0.12985	11.1%
95	0.25397	6.6%	0.20538	8.6%
average	0.11545	9.9%	0.08353	12.7%

footnote: average is the average of ages 65~99

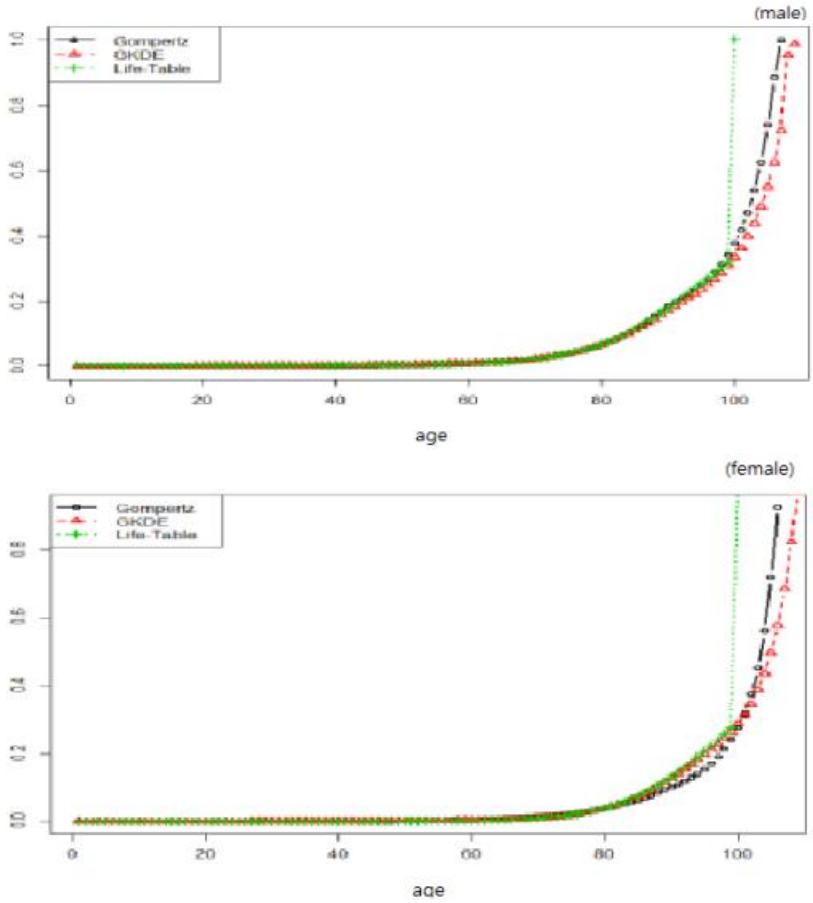

Figure 2. Probability of death (estimation, yr2012)

이와 같이 100세 이상 연장된 사망확률은 일관된 흐름을 보여주지 못하기 때문에 본 연구에서는 2002년~2012년까지 연도별로 사망확률을 100세 이상으로 연장한 후 이들의 평균을 Table 4와 같이 2007년의 연장된 사망확률로 정의하였다. 남자와 여자의 사망확률을 비교한 결과 107세와 108세에서 남녀별 사망확률 역전현상이 발생하여 이러한 현상을 방지하기 위해 “성별 사망확률 조정계수(coordination factor)”를 아래 박스와 같이 정의하고 적용하였다.

성별 사망확률 조정계수 = 남녀별 평균 사망확률/총 평균 사망확률
- 사망확률 조정계수(남자) = 0.699/0.691 = 1.012
- 사망확률 조정계수(여자) = 0.683/0.691 = 0.988
- 총 평균 사망확률 = (0.699+0.683)/2 = 0.691

연령별 사망확률(남자) = 총 연령별 사망확률 × 성별 사망확률 조정계수(남자)
연령별 사망확률(여자) = 총 연령별 사망확률 × 성별 사망확률 조정계수(여자)

본 연구에서는 서론에서 언급한 인구 과소추계 문제를 해결하기 위하여 통계청의 100세까지 나와 있는 사망확률을 100세 이상으로 연장하는 방법을 소개하였고 그 결과(아래 Table 5 참조)를 산출하였다. 연구결과에 따르면 우리나라의 한계연령은 남자와 여자의 경우 각각 112세와 114세로 나타났다. 실제로 안전행정부 주민등록 인구 통계 자료에 따르면 2014년 7월 기준으로 전국의 100세 이상 인구는 1만 4592명으로 나타났다. 따라서 통계청에서 발표하는 한계연령인 100세는 너무 낮다고 판단된다. 하지만 극단값 이론을 적용하는 방법에서 나타나는 성별, 시기별 사망률 역전 현상은 100세 이상 고연령 사망확률 추정 시 해결해야 할 부분으로 앞으로 추가적인 연구가 필요할 것으로 보인다.

4. 결론

본 연구에서 사용한 극단값 이론을 통한 사망확률 연장 방법은 임계연령을 기준으로 임계연령 이전은 GKDE 모형을 사용하고 임계연령 이후는 Pareto 모형을 사용한 방법이다. 이 방법은 통계적 분포를 통해 한계연령을 결정할 수 있으므로 주관적인 판단을 할 필요가 없는 장점이 있지만 자료의 충분성이 확보되지 않으며 모형의 정확성을 검증할 수 없는 단점 또한 있다. 이러한 단점을 극복한 Jeon, Kim(2013)은 new gamma KDE(GKDE)를 채택하여 우리나라와 같이 고연령 데이터가 부족한 경우 에 적용성이 높음을 보이고 있다.

본 연구의 주요 결과로서 100세 이상 고연령 사망확률을 new GKDE 모형에 극단값 이론을 적용하여 추정하고 2012년 추정결과를 기존의 Gompertz 모형에 극단값 이론을 적용한 경우와 비교하였다.

new GKDE를 사용한 모형(new GKDE+Pareto 모형)과 Gompertz를 사용한 모형(Gompertz+Pareto 모형)의 결과를 비교하면 전체적으로 큰 차이가 없는 것으로 나타났으나 한계연령은 “new GKDE+Pareto” 모형의 경우가 약간 높은 것으로 나타났다.

다음으로 남자와 여자의 한계연령이 비슷한 것으로 나타났지만 여자의 한계연령이 남성보다 낮게 추정되어 보정이 필요하였다. 성별 한계연령 추세의 일관성을 제고하기 위해 2002년부터 2012년까지의 사망확률의 평균을 계산하여 조정계수를 산출하고 적용하였다. 조정한 결과 남성의 한계연령은 112세로 나타났고 여성은 114세로 나타났다. 실제로 주민등록 인구 통계 자료에 따르면 2014년 7월 기준으로 전국의 100세 이상 인구는 1만 4592명으로 나타났다. 향후 통계청 자료는 100세 이상의 고연령으로 연장되어야 할 것이다.

Table 4. Estimation of prob. of death over ages 100 to calculate coordination factors

age (male)	year											avg.
	'02	'03	'04	'05	'06	'07	'08	'09	'10	'11	'12	'07
100	.34	.35	.34	.33	.30	.33	.35	.33	.35	.34	.33	.336
101	.37	.38	.37	.36	.32	.36	.38	.36	.39	.37	.35	.364
102	.40	.41	.40	.39	.34	.39	.41	.39	.43	.40	.39	.397
103	.44	.45	.44	.43	.37	.43	.45	.43	.48	.44	.43	.436
104	.49	.50	.49	.47	.40	.48	.50	.48	.54	.49	.47	.482
105	.54	.55	.54	.52	.43	.53	.56	.55	.61	.55	.53	.539
106	.61	.62	.60	.59	.48	.60	.64	.63	.71	.62	.61	.608
107	.69	.69	.68	.67	.52	.69	.73	.74	.82	.70	.70	.694
108	.80	.78	.78	.78	.58	.80	.83	.87	.92	.80	.82	.796
109	.91	.88	.88	.90	.66	.92	.94	.99	.96	.90	.95	.898
110	.99	.95	.96	.99	.74	1.0	.97	1.0	1	.97	1	.962
111	1.0	.96	.97	1.0	.85	1.0	.97	1	1	.97	1	.974
112	1	1	1	1	.95	1	1	1	1	1	1	.996
113	1	1	1	1	1.0	1	1	1	1	1	1	1.000
114	1	1	1	1	1.0	1	1	1	1	1	1	1.000
avg. of prob. of death												.699
age (female)	year											avg.
	'02	'03	'04	'05	'06	'07	'08	'09	'10	'11	'12	'07
100	.31	.33	.31	.29	.29	.33	.31	.31	.30	.30	.28	.305
101	.33	.35	.34	.32	.31	.37	.34	.34	.34	.34	.30	.334
102	.36	.39	.37	.34	.33	.41	.38	.38	.38	.38	.33	.367
103	.39	.42	.41	.37	.35	.46	.43	.43	.43	.43	.36	.408
104	.42	.47	.45	.41	.38	.53	.49	.50	.50	.49	.40	.458
105	.47	.52	.51	.45	.41	.61	.56	.58	.58	.58	.45	.520
106	.52	.58	.57	.50	.45	.72	.65	.69	.70	.68	.52	.599
107	.58	.66	.65	.56	.50	.85	.78	.83	.84	.82	.62	.698
108	.65	.75	.74	.64	.55	.96	.92	.95	.96	.95	.75	.801
109	.74	.86	.85	.74	.61	.97	.97	.96	.96	.96	.92	.867
110	.85	.94	.95	.85	.69	1	1	1	1	1	1	.936
111	.95	.96	.96	.95	.78	1	1	1	1	1	1	.964
112	.98	1	1	.98	.89	1	1	1	1	1	1	.985
113	1	1	1	1	.97	1	1	1	1	1	1	.997
114	1	1	1	1	.98	1	1	1	1	1	1	.998
avg. of prob. of death												.683

연구결과를 살펴보면 남자와 여자의 한계연령에 큰 차이가 없는 것으로 나타났는데 그 이유는 연령별 사망확률은 남자가 여자보다 높지만 연령이 높아짐에 따른 사망확률 증가속도는 고연령대에서 여자가 남자보다 빠르기 때문으로 볼 수 있다. 또한 여자의 한계연령은 시간이 지남에 따라 오히려 감소하는 경향을 발견할 수 있는데 그 이유는 여자의 경우 연령이 높아짐에 따른 사망확률 증가속도가 시간이 지남에 따라 빨라지기 때문으로 판단할 수 있다. 이러한 추세가 한계연령에 영향을 주므로 향후에도 이러한 추세가 유지될지 지속적인 관찰이 필요하다.

마지막으로, 본 연구에서 사용한 고연령 추정방법은 다양한 많은 방법들 중 한 가지 방법으로 다른 방법들을 사용하여 추정결과를 산출하여 이 결과들을 비교·분석할 필요가 있을 것이다.

Table 5. Estimated probabilities of death over ages 100

age	male	female
100	0.32453	0.31696
101	0.35308	0.34484
102	0.38671	0.37769
103	0.42688	0.41692
104	0.47560	0.46451
105	0.53569	0.52319
106	0.61082	0.59657
107	0.70424	0.68780
108	0.80809	0.78923
109	0.89316	0.87232
110	0.96003	0.93763
111	0.98060	0.95772
112	1.00000	0.97898
113		0.98658
114		1.00000

References

Baek, H., Noh, J., Lee, H. (2013). Estimation of mortality and actuarial analysis applying Lee-Carter model, *Journal of the Korean Data Analysis Society,* 15(3), 1553-1572. (in Korean).

Baik, C. (2012). A process of old age mortality estimation using the Gompertz mortality function, *Korean Academy of Actuarial Science,* 4(1), 59-77. (in Korean).

Bongaarts, J. (2005). Long-range trends in adult mortality: Models and projection methods, *Demography,* 42(1), 23-49.

Cairns, A. J. G., Blake, D., Dowd, K. (2006). A two-factor model for stochastic mortality with parameter uncertainty: Theory and calibration, *The Journal of Risk and Insurance,* 73(4), 687-718.

Chen, S. (2000). Probability density function estimation using gamma kernels, *Annals of the Institute of Statistical Mathematics*, 52(3), 471-480.

Choi, J. H. (2015). Estimating the benefit-cost ratios by applying life-expectancies of national pension old-age pensioners, *The Korean Journal of Applied Statistics,* 28(4), 621-641. (in Korean).

Coal, A., Guo, G. (1989). Revised regional model life tables at very low levels of mortality, *Population Index,* 55(4), 613-643.

Coal, A., Kisker, E. (1990). Defects in data on old-age mortality in the United States: New procedures for calculating schedules and life tables at the higher ages, *Asian and Pacific Population Forum,* 4, 1-31.

Gompertz, B. (1825). On the nature of the function expressive of the law of human mortality and on a new mode of determining life contingencies, *The Royal Society of London, Philosophical Transactions, Series A,* 115, 513-585.

Heligman, L., Pollard, J. H. (1980). The age of pattern of mortality, *Journal of the Institute of Actuaries,* 107, 437-455.

Himes, C. L., Preston, S. H., Condran, G. A. (1994). A relational model of mortality at older ages in low mortality countries, *Population Studies,* 48(2), 269-291.

Jeon, Y., Kim, J. (2013). A gamma kernel density estimation for insurance loss data, *Insurance: Mathematics and Economics,* 53, 569-579.

Kim, S. (2013). A study on the estimating mortality rates for the old and the highest attained ages using extreme value theory, *Korean Risk Management Society*, 24(1), 129-150. (in Korean).

Lee, R. D., Carter, L. R. (1992). Modeling and forecasting U.S. mortality, *Journal of the American Statistical*

Association, 87(419), 659-671.

Lee, N., Lee, R. (2005). Coherent mortality forecasts for a group of populations: An extension of the Lee-Carter method, *Demography,* 42(3), 575-594.

Li, J. S. H., Hardy, M. R., Tan, K. S. (2008). Threshold life tables and their applications, *North American Actuarial Journal,* 12, 99-115.

Moon, S. (2011). Study of Korean mortality by means of spatial statistics, *Journal of the Korean Data Analysis Society,* 13(1), 221-232. (in Korean).

Panjer, H. H., Russo, G. (1992). Parametric graduation of Canadian individual insurance mortality experience: 1982-1988, *Proceedings of the Canadian Institute of Actuaries,* 23, 378-449.

Panjer, H. H., Tan, K. S. (1995). *Graduation of Canadian individual insurance mortality experience: 1986-1992*, Canadian Institute of Actuaries.

Park, Y., Kim, S. (2011). A method for construction of life table in Korea, *The Korean Journal of Applied Statistics,* 24(5), 769-789. (in Korean).

Park, J., Roh, H., Yoon, S., Son, Y. (2006). VaR estimation using the generalized extreme value distributions, *Journal of the Korean Data Analysis Society,* 8(1), 227-240. (in Korean).

Sohn, K. T., Liang X. (2014). Estimation of return levels of typhoon best track data based on generalized extreme value distribution, *Journal of the Korean Data Analysis Society,* 16(3), 1259-1267. (in Korean).

Yun, S. (2012). Extreme value models in stationary time series with applications to the Dow Jones industrial average index, *Journal of the Korean Data Analysis Society,* 14(5), 2487-2497. (in Korean).

Research on Extending Mortality Over Ages of 100 Based on the Life-Table of Statistics Korea

Joo-Ho Sung[1], *Jang Hoon Choi*[2]

Abstract

The population is rapidly aging due to low-birth rates and longer life expectancies in Korea. In order to maintain financial stability, financial and population projections of the Korean national pension scheme should be more accurate. Currently, the life-tables of Statistics Korea are used for the population projections but they do not reflect the rapidly aging trends because the highest available age in mortality is limited to 100. Our new model combining the gamma kernel with the extreme value theory is designed to extend mortality over ages of 100 based on the life-table of Statistics Korea. It is also compared with the existing model in which the Gompertz is combined with the extreme value theory. The results are not very different between two models, but the ages of limit and their confidence levels are a little higher in the newly designed model. Two unexpected results are occurred. One is that the age of limit of female is not higher than that of male. The other one is that the age of limit in yr 2012 is lower than that in yr 2001 in case of females. Thus the probabilities of death over ages of 100 are generated from yr 2002 to 2012 and the averages of those probabilities of death are defined as those of yr 2007.

Keywords : Aging, GKDE, Gompertz, Extreme value theory, Age of limit.

[1](Corresponding Author) Professor, Dept. of Business Administration, Kyung Hee University, 26, Kyungheedae-ro, Dongdaemun-gu, Seoul, 02447, Korea. E-mail : jhsung@khu.ac.kr
[2]Research Fellow, Korea Insurance Research Institute, 38, Gukjegeumyung-ro-6-gil, Youngdeungpo-gu, Seoul 07328, Korea. E-mail : james021@kiri.or.kr
[Received 19 January 2016; Revised 1 February 2016, 17 February 2016; Accepted 20 February 2016]