Proceedings of Joint Conference of The Korean Data Analysis Society and Korean Data And Information Science Society, October 17-18, 2003, pp. 139-145

3-모수 카파분포의 모수 추정 방법들의 비교

전유나 · 김연우 · 황영아[1] · 박정수[2]

요 약

본 논문에서는 강수 자료의 예측에 사용되는 3-모수 카파 분포(KD3)에서의 모수 추정 방법을 알아보고 시뮬레이션을 통하여 모수 추정 방법에 따른 성능을 비교해 보았다. 이 분포의 모수 α, β, μ를 추정하기 위하여 적률추정법(MME), L-적률 추정법(LME), 최우추정법(MLE)을 적용하였다. 소표본의 경우뿐만 아니라 대표본의 경우에도 시뮬레이션을 통하여 추정법들의 성능을 비교하였다. 적률 추정법과 L-적률 추정법에서는 제약조건 하에서의 1차원 Newton-Raphson방법을 수정하여 이용하였다. MSE를 기준으로 한 시뮬레이션 결과, KD3의 모수 추정에 있어서 표본의 크기가 100보다 작으면 LME의 적용을 추천하고 표본의 크기가 100이상이면 MLE를 추천한다.

1. 서론

강수량 자료에 사용되어지는 확률분포중의 하나로 카파분포는 Mielke에 의해서 1973에 처음으로 2-모수, 3-모수 카파분포가 소개되었으며 Hosking(1994)은 4-모수 카파분포에 대해 연구하였다. 이러한 카파분포는 감마분포나 로그정규분포와 함께 오른쪽을 긴 꼬리를 갖고 양의 값을 갖는 확률분포로 강수량 자료를 예측하는데 이용되어진다. 또한 카파분포는 감마분포와 로그정규분포에 비하여 분포함수와 백분위함수를 수리적으로 쉽게 구할 수 있어 강수량의 분포를 추정하는데 이점이 있다.

카파분포의 모수를 추정하는 방법으로는 적률 추정법, L-적률 추정법, 최우 추정법에 대해서 시뮬레이션을 통하여 그 성능을 비교하였다. 적률추정법은 적용이 간단하여 많이 사용하는 방법중의 하나이나 분포함수의 형태가 한쪽으로 많이 치우친 경우에는 완전한 추정치를 얻을 수 없고, 고차적률로 갈수록 추정이 부정확해져 왜곡될 가능성이 있다. L-적률 추정법은 순서통계량들의 선형결합으로 이루어진 통계량을 이용하며 기존의 적률추정법에 비해 추정량이 덜 편의되어 있고 자료의 이상치에 대해 보다 덜 민감하다는 특성이 있다. 최우 추정법은 일반적으로 가장 효율적인 추정치를 얻을 수 있고 표본의 크기가 충분히 큰 경우에는 다른 모수 추정 방법과의 효율성을 비교하는 기준으로 사용되어 질 수 있다. 그러나 표본의 크기가 작은 경우에는 대체로 잘 일치하지 않는 결

[1]전남대학교 통계학과 대학원
[2]전남대학교 통계학과 교수

과를 얻을 수 있고 최우 추정치를 구하기 위한 식이 비선형 방정식으로 표현되는 경우가 많아 적률 추정법이나 L-적률 추정법에 비해 모수의 추정치를 구하는데 어려움이 따르기도 한다.

2. 분포의 특성

카파 분포(Kappa distribution)는 Mielke(1973)에 의해 제안되었으며, 누적 확률분포함수 $F(x)$와 확률밀도함수 $f(x)$는 다음과 같다.

$$F_3(x)=\left(\frac{x-\mu}{\beta}\right)\left[\alpha+\left(\frac{x-\mu}{\beta}\right)^{\alpha}\right]^{-1/\alpha}, \tag{2.1}$$

$$f_3(x)=\left(\frac{\alpha}{\beta}\right)\left\{\alpha+\left(\frac{x-\mu}{\beta}\right)\right\}^{-\frac{(\alpha+1)}{\alpha}}, \quad \alpha,\ \beta>0\ , \tag{2.2}$$

식(2.1)과 식(2.2)는 KD4의 분포함수이며 위치(location) 파라미터 μ와 모양(shape) 파라미터 α, 규모(scale) 파라미터 β를 갖는다. 위 식으로부터 백분위함수는

$$x(F)=\mu+\beta\left[\frac{\alpha F^{\alpha}}{1-F^{\alpha}}\right]^{1/\alpha}, \quad 0<F<1 \tag{2.1.1}$$

가 된다. KD3의 분포형태는 α, β, μ 에 따라 달라지며 다음의 그림(2.1.1)과 그림(2.1.2)를 통해서 알 수 있다. α는 모양 파라미터로 α가 커질수록 모양이 평평해지며, β는 규모 파라미터로 값이 커질수록 퍼지는 특성을 가진다. 또한 μ는 위치 파라미터로 원점을 중심으로 μ의 크기만큼 x축을 이동하는 파라미터이다.

3.1 적률 추정법

임의의 분포에 대한 r차 적률의 일반식은 다음과 같다.

$$E(X^r)=\int_{-\infty}^{\infty}x^r f(x)\,dx=\int_0^1[x(F)]^r dF$$

$$E(X)=\int_0^1[x(F)]dF=\mu+\beta\alpha^{-\frac{\alpha-1}{\alpha}}B\left(\frac{2}{\alpha},\ \frac{\alpha-1}{\alpha}\right), \qquad \alpha>1$$

$$E(X^2)=\mu^2+2\mu\beta\alpha^{\frac{\alpha-1}{\alpha}}B\left(\frac{2}{\alpha},\frac{\alpha-1}{\alpha}\right)+\beta^2\alpha^{-\frac{\alpha-2}{\alpha}}B\left(\frac{3}{\alpha},\frac{\alpha-2}{\alpha}\right),\ \alpha>2$$

$$E(X^3)=\mu^3+3\mu^2\beta\alpha^{-\frac{\alpha-1}{\alpha}}B\left(\frac{2}{\alpha},\frac{\alpha-1}{\alpha}\right)$$
$$+3\mu\beta^2\alpha^{-\frac{\alpha-2}{\alpha}}B\left(\frac{3}{\alpha},\frac{\alpha-2}{\alpha}\right)+\beta^3\alpha^{-\frac{\alpha-3}{\alpha}}B\left(\frac{4}{\alpha},\frac{\alpha-3}{\alpha}\right),\ \alpha>3$$

$$E(X^r) = \int_0^1 [x(F)]^r dF = \sum_{i=0}^{r} \binom{r}{i} \mu^i \beta^{r-i} h_{r-i}(\alpha)$$

여기서, $h_k(\alpha) = \alpha^{-\frac{\alpha-k}{\alpha}} B\left(\frac{k+1}{\alpha}, \frac{\alpha-k}{\alpha}\right), \quad \alpha > k, \quad h_0(\alpha) = 1$

모집단의 적률에 불편추정량이 되게 구한 표본의 r차 적률은 다음과 같으며, 이후부터 쓰이는 m_r은 $E(X^r)$와 같은 의미이다. 적률 추정법을 이용하여 KD3의 모수 μ, α, β 를 추정하기 위해서는 모집단의 적률과 표본의 적률을 같다고 놓고 3차 연립방정식을 풀 수 있다. 그렇지만 3원의 연립방정식의 해를 구하는 것보다 적절한 조작을 통하여 구해진 일원 연립방정식을 만족하는 $\hat{\alpha}$ 을 먼저 구한 다음 구해진 $\hat{\alpha}$ 을 식(3.1.8)에 대입하여 $\hat{\beta}$ 을 구하고 다시 식(3.1.6)에 $\hat{\alpha}, \hat{\beta}$ 을 대입하여 $\hat{\mu}$ 을 구하는 방법이 더 쉬우므로 이를 이용하기로 한다.

$$m_1 = \mu + \beta h_1(\alpha), \quad \therefore \hat{\mu} = \widehat{m_1} - \hat{\beta} h_1(\hat{\alpha}) \qquad \text{(3.1.5), (3.1.6)}$$

$$m_2 = (\mu + \beta h_1(\alpha))^2 - \beta^2 h_1^2(\alpha) + \beta^2 h_2(\alpha)$$

$$m_2 - m_1^2 = \beta^2 (h_2(\alpha) - h_1^2(\alpha)), \quad \therefore \hat{\beta} = \left(\frac{\widehat{m_2} - \widehat{m_1^2}}{h_2(\hat{\alpha}) - h_1^2(\hat{\alpha})} \right)^{1/2}, \qquad \text{(3.1.7), (3.1.8)}$$

$$m_3 = m_1^3 + 3m_1(m_2 - m_1^2) + \beta^3 (2h_1^3(\alpha) - 3h_1(\alpha)h_2(\alpha) + h_3(\alpha))$$

$$\therefore m_3 - m_1^3 - 3m_1(m_2 - m_1^2) = \beta^3 (2h_1^3(\alpha) - 3h_1(\alpha)h_2(\alpha) + h_3(\alpha)) \qquad \text{(3.1.9)}$$

여기서, β 를 소거하기 위하여 식(3.1.7)의 양변을 $\frac{3}{2}$ 승하여 식(3.1.9)으로 나눈다.

$$\frac{(m_2 - m_1^2)^{3/2}}{m_3 - m_1^3 - 3m_1(m_2 - m_1^2)} = \frac{\beta^3 (h_2(\alpha) - h_1^2(\alpha))^{3/2}}{\beta^3 (2h_1^3(\alpha) - 3h_1(\alpha)h_2(\alpha) + h_3(\alpha))} \overset{\text{def}}{=} g_M(\alpha)$$

$$\therefore g_M(\alpha) - f(m) = 0, \quad \alpha > 3, \quad f(m) = \frac{(m_2 - m_1^2)^{3/2}}{m_3 - m_1^3 - 3m_1(m_2 - m_1^2)}. \qquad \text{(3.1.10)}$$

식(3.1.10)은 수학적으로는 그 해를 찾기가 어려우므로 Newton-Raphson 반복을 통해 찾는다. 이 방법을 쓰기 위해서는 일차 미분치가 필요한데 이를 위해서는 베타 함수의 미분과 디감마함수를 이용한다. α는 모수표현식에 포함된 Beta 함수 때문에 $\alpha > 3$ 라는 제약조건을 가지게 되므로 Numerical Recipes에 소개된 제약조건 하에서의 One-dimensional Newton-Raphson 방법을 수정하여 이용한다.

3.2 L-적률 추정법

L-적률 추정치는 적률 추정치의 성질을 개선시킨 확률가중 적률 추정치 (Probability Weighted

Moments; PWM)의 특별한 경우로서 순서 통계량들의 선형결합으로 이루어진 통계량이다. L-적률은 기존의 적률보다 추정량이 덜 편의 되어있고, 자료의 이상치에 덜 민감(robust)하며 좀 더 간단하게 적용할 수 있다는 특성이 밝혀져 있다. 이 L-적률을 이용하면 적은 표본으로 치우친 분포에 대한 모수 추정을 더 효율적으로 할 수 있다 (Hosking, 1990). X가 누적분포함수 $F(x)$와 백분위 함수 $x(F)$를 갖는 확률변수라 할 때, $X_{1:n} \le X_{2:n} \le \cdots \le X_{n:n}$은 X의 분포로 나온 크기 n의 확률표본이다. $E(X_{j:r})$을 r개 확률변수 중 j번째 확률변수의 기대값이라 할 때, L-적률은 다음의 식과 같다.

$$E(X_{j:r}) = \frac{r!}{(j-1)!(r-j)!} \int x\{F(x)\}^{j-1}\{1-F(x)\}^{r-j} dF(x)$$

$$\lambda_r \overset{\text{def}}{=} r^{-1} \sum_{k=0}^{r-1} (-1)^k \binom{r-1}{k} EX_{r-k:r} \quad , \quad r=1,2, \tag{3.2.1}$$

$$\lambda_1 = EX = \int_0^1 x(F)dF, \quad \lambda_2 = E(X_{2:2} - X_{1:2})/2 = \int_0^1 x(F)(2F-1)dF,$$

$$\lambda_3 = E(X_{3:3} - 2X_{2:3} + X_{1:3})/3 = \int_0^1 x(F)(6F^2 - 6F + 1)dF,$$

$$\tau_2 = \frac{\lambda_2}{\lambda_1} \quad , \quad \tau_r = \frac{\lambda_r}{\lambda_2} \quad , \quad r=3,4,\cdots. \tag{3.2.2}$$

L-적률의 의미를 살펴보면 확률표본을 크기 순으로 나열하여 이항 계수로 가중치를 주고 순서 통계량의 기대값의 평균을 구한 것이다. 이러한 형태의 L-적률은 적률과 비슷하고 식(3.2.2)로 표현되는 L-적률비(L-moments ratios)도 적률비와 비슷하다. 그러나 이상치들의 측정 오차들이나 표본의 변동에 적률보다 덜 민감하다는 장점이 있다. 특별히 λ_1은 평균으로서 L-위치, λ_2는 산포의 측정량으로서 L-규모, τ_2는 L-변이계수(L-CV), τ_3는 L-왜도, τ_4는 L-첨도이라고 부른다(Hosking, 1990). 그리고, 모집단의 L-적률 λ_r의 불편추정량으로서 표본 L-적률은 다음의 식 (3.2.3)과 같다.

$$l_1 = n^{-1}\sum x_i, \quad l_2 = 1/2\binom{n}{2}^{-1} \sum_{i>} \sum_{j} (x_{i:n} - x_{j:n}),$$

$$l_3 = 1/3\binom{n}{3}^{-1} \sum_{i>} \sum_{j>} \sum_{k} (x_{i:n} - 2x_{j:n} + x_{k:n}),$$

$$l_r = \binom{n}{r}^{-1} \sum_{l \le i_1} \sum_{\le i_2} \cdots \sum_{\le i_n} r^{-1} \sum_{k=0}^{r-1} (-1)^k \binom{r-1}{k} x_{i_{r-k}:n}, \; r=1,2,\cdots n.$$

$$t_2 = l_2/l_1, \quad t_r = l_r/l_2, \quad r=3,4,\cdots$$

여기서 t_r는 r번째 표본의 L-적률비이고, t_3은 표본의 L-왜도, t_4는 표본의 L-첨도라고 부른다. PWM을 이용한 L-적률 표현식은 다음과 같다.

$$\lambda_1 = \mu + \beta\alpha^{-\frac{\alpha-1}{\alpha}} B\left(\frac{2}{\alpha}, \frac{\alpha-1}{\alpha}\right) \overset{\text{def}}{=} \mu + \beta k_1(\alpha) \tag{3.2.6}$$

$$\lambda_2 = 2\beta\alpha^{-\frac{\alpha-1}{\alpha}} B\left(\frac{3}{\alpha}, \frac{\alpha-1}{\alpha}\right) - \beta\alpha^{-\frac{\alpha-1}{\alpha}} B\left(\frac{2}{\alpha}, \frac{\alpha-1}{\alpha}\right) \overset{\text{def}}{=} \beta[2k_2(\alpha) - k_1(\alpha)]$$

$$\lambda_3 = 6\beta\alpha^{-\frac{\alpha-1}{\alpha}} B\left(\frac{4}{\alpha}, \frac{\alpha-1}{\alpha}\right) - 6\beta\alpha^{-\frac{\alpha-1}{\alpha}} B\left(\frac{3}{\alpha}, \frac{\alpha-1}{\alpha}\right) + \beta\alpha^{-\frac{\alpha-1}{\alpha}} B\left(\frac{2}{\alpha}, \frac{\alpha-1}{\alpha}\right) \overset{\text{def}}{=} \beta[6k_3(\alpha) - 6k_2(\alpha) + k_1(\alpha)]$$

$$\tau_2 = \frac{\lambda_2}{\lambda_1} = \frac{\beta[2k_2(\alpha) - k_1(\alpha)]}{\mu + \beta k_1(\alpha)}$$

$$\tau_3 = \frac{\lambda_3}{\lambda_2} = \frac{6k_3(\alpha) - 6k_2(\alpha) + k_1(\alpha)}{2k_2(\alpha) - k_1(\alpha)} \overset{\text{def}}{=} g_L(\alpha) \tag{3.2.9}$$

모집단의 L-적률 표현식들을 표본의 L-적률들과 같게 놓고 $\hat{\alpha}$, $\hat{\beta}$, $\hat{\mu}$ 을 구할 수 있다. N-R 법에 의해 식(3.2.10)을 만족하는 $\hat{\alpha}$ 을 구한 다음 식(3.2.11)에 대입하여 $\hat{\beta}$을 구한다. 그런 다음에 구해진 $\hat{\alpha}$, $\hat{\beta}$ 을 식(3.2.12)에 대입하여 $\hat{\mu}$ 을 구한다.

$$g_L(\alpha) - t_3 = 0, \qquad \alpha > 1 \tag{3.2.10}$$

$$l_2 = \beta[2k_2(\alpha) - k_1(\alpha)], \quad \hat{\beta} = \frac{l_2}{2k_2(\hat{\alpha}) - k_1(\hat{\alpha})} \tag{3.2.11}$$

$$l_1 = \mu + \beta k_1(\alpha), \quad \hat{\mu} = l_1 - \hat{\beta}k_1(\hat{\alpha}) \tag{3.2.12}$$

3.3 최우 추정치

최우 추정법은 추출된 표본자료가 나올 수 있는 확률이 최대가 되도록 모수를 추정하는 방법이다. $-\ln L(\alpha, \beta, \mu)$ 를 최소화시키는 것은 우도함수 $L(\alpha, \beta, \mu)$ 가 최대가 되도록 하는 것과 같은 의미이며 $-\ln L(\alpha, \beta, \mu)$ 는 식(3.3.2)와 같다.

$$-\ln L(\alpha, \beta, \mu) = \frac{\alpha+1}{\alpha} \sum_{i=1}^{n} \ln\left[\alpha + \left(\frac{x_i - \mu}{\beta}\right)^{\alpha}\right] - n\ln\left(\frac{\alpha}{\beta}\right) \tag{3.3.2}$$

식(3.3.2)를 최소화시키는 해를 수학적으로 얻기 어려우므로 Quasi-Newton방법을 통하여 찾아낸다. KD3의 파라미터 β는 규모파라미터로 scale-equivariance이고 파라미터 μ는 위치 파라미터로 location-equivariance이다. 따라서 $b\beta$ 를 추정하고자 한다면 $\beta=1$ 의 추정치에 b 를 곱하고, $a+\mu$ 를 추정하고자 한다면 $\mu=0$ 으로 놓고 추정하여 그 추정치에 a 만큼을 더해서 추정할 수 있다. 또한 적률 추정법, L-적률 추정법, 최우추정법에서 β는 규모 파라미터로 scale-equivariance를

만족하고, μ는 위치 파라미터로 location-equivariance를 만족한다. 이는 아래 시뮬레이션에서 이용된다.

4. 시뮬레이션

일반적으로 대표본의 경우 최우 추정법의 분산이 가장 작으므로 최우 추정치가 다른 추정치와의 효율성을 비교하는 기준으로 사용되어 진다. 이러한 경우 주로 소표본에 있어서의 각 추정법에 대한 성능비교가 시뮬레이션의 주요한 목적이 된다. 그러나 본 논문에서는 최우 추정치의 Regularity Condition이 만족되지 않아 이론상으로 표본의 크기가 큰 경우에도 최우 추정치의 성능이 가장 좋다고 말하기 어렵다. 따라서 이 장에서는 소표본의 경우뿐만 아니라 대표본의 경우에도 시뮬레이션을 통하여 각 추정법의 성능을 비교하여 보고 가장 좋은 추정법을 알아보고자 한다. 시뮬레이션의 설정은 다음과 같다.

1) 반복계산 횟수: 1000번,
2) sample size: 20,30,40,50,75,100,200,300,500
3) 모수추정 · α : 3.001, 3.5, 4, 5, 6, 7, 10, 15
   · β : Scale-equivariant 이므로 모두 1로 설정
   · μ : Location-equivariant 이므로 모두 0로 설정
4) Quentile 추정 : .9 .95 .99

모수 α, β, μ 값에 따라 표본의 크기를 달리하여 적률 추정법, L-적률 추정법, 최우 추정법에 따른 모수의 추정치와 각각의 모수추정방법의 성능을 비교하기 위한 값을 제시한다(표 생략). 모수 추정 방법의 성능을 비교하기 위해서 반복 계산하여 구해진 추정치들의 MBias, Variance, MSE를 사용하였다. 여기서 추정하고자 하는 모수의 참값과 추정된 모수값 사이의 차의 제곱을 추정된 모수의 수로 나누어서 계산된 MSE는 주로 각각의 추정법의 성능비교에 사용되고 MSE가 작을수록 성능이 좋다고 할 수 있다. 또한 MBias가 0에 가까울수록 추정법의 성능이 좋다고 말 할 수 있다. (결과표 생략)

MSE를 이용하여 성능을 비교해 보면 다음과 같은 결론을 내릴 수 있다. α의 추정에 있어서 α가 작고 표본의 크기가 100보다 작으면 L-적률 추정법이 성능이 좋고, 표본의 크기가 100이상이면 최우 추정법의 성능이 좋다고 볼 수 있다. 최우 추정법에서 μ의 추정의 경우에는 $\mu \leq \min(x)$라는 제약조건이 μ의 범위로 지정되어는데 $\hat{\mu}$이 $\min(x)$로 추정되어짐으로 그 성능이 가장 좋다는 결과를 보인다. 모수 α의 값과 표본의 크기를 변화시켰을 때 각각의 모수 추정법에 따른 0.9, 0.95, 0.99 Quantile을 추정하고 성능을 비교하기 위한 값을 제시한다(결과표 생략). Quantile추정은 분포의 꼬리부분 값에 대한 추정 성능을 비교하기 위함이며 이 경우에도 MSE를 통하여 성능을 비교하여 보았다.(결과표 생략) KD3의 Quantile 추정을 위해서는 α가 작고 표본의 크기가 100(or 200)보다

작으면 L-적률 추정법(적률 추정법)을, 표본의 크기가 100(or 200)이상이면 최우 추정법을 적용할 것을 추천한다.

5. 결론

본 논문에서의 KD3는 자료의 전체적인 이동이 있고 오른쪽으로 긴 꼬리를 갖는 한쪽으로 치우친 분포로 강수량 자료를 예측하는데 사용되어 진다. 또한 KD3는 분포함수나 백분위 함수를 수리적으로 쉽게 구할 수 있다는 이점이 있다. 이에 KD3의 모수 추정을 위하여 적률 추정법, L-적률 추정법, 최우 추정법의 성능을 비교해 보았다.

모수 α, β, μ 의 추정에 있어서는 표본이 100보다 작은 경우에는 L-적률 추정법(적률 추정법)을, 표본이 100보다 큰 경우에는 최우 추정법을 적용할 것을 추천한다. 또한 KD3의 Quantile 추정을 위해서는 α가 작고 표본의 크기가 100(or 200)보다 작으면 L-적률 추정법을, 표본의 크기가 100(or 200)이상이면 최우 추정법을 적용할 것을 추천한다.

참고문헌

오은선, Kappa 분포의 모수 추정과 강수 자료에의 적용, 전남대학교 대학원 통계학과 석사학위 논문, 2001.

Greenwood J. A., Landwehr P. W., Matalas N. C., and Wallis J. R., Probability weighted moments : Definition and Relation to parameters of several distributions expressed in inverse form, *Water Resources Research,* Vol.15, No.5, pp. 1049 ~ 1064, 1979.

Hosking J. R. M., L-Moments : Analysis and Estimation of Distributions using Linear Combinations of Order Statistics, *J. R. Stat. Soc.,* Ser. B, Vol.52, No.2, pp. 105 ~ 124, 1990.

Mielke P. W., Another Family of Distributions for Describing and Analyzing Precipitation Data, *J. Appl. Meteorol,* 12, pp. 275 ~ 280, 1973.

Mielke P. W., Johnson E. S., Three-Parameter Kappa Distribution Maximum Liklihood Estimates and Likelihood Ratio Tests, *Monthly Weather Review,* vol.101, No.9, pp. 701~707, 1973.

Press W. H., Teukolsky S. A., Vetterling W. T., Flannery B. P., *Numerical Recipes in Fortran 77,* Second Edition, Cambridge University Press, 1992.