불완전Ⅱ형자료에 기초한 파레토분포의 파라메터들에 대한 통계적추론

리광선, 한류경

경애하는 최고령도자 **김정은**동지께서는 다음과 같이 말씀하시였다.

《과학기술력은 국가의 가장 중요한 전략적자원이며 사회발전의 강력한 추동력입니다.》

(《조선로동당 제7차대회에서 한 중앙위원회사업총화보고》단행본 38페지)

파레토분포는 경제적수익성을 비롯한 사회경제현상을 연구하는데서와 물리와 생물현상, 도시인구, 증권가격, 보험위험, 강수량, 림상실천 등을 연구하는데 많이 쓰이고있다.

지난 시기에는 주로 완전자료인 경우의 추론문제를 많이 연구하였다.

관측자료가 불완전자료로 주어진 경우에는 주로 지수분포, 웨이불분포, 로그정규분포의 파라메터추정문제를 연구하였다.[3－9]

선행연구[6]에서는 불완전Ⅱ형자료인 경우 지수분포의 파라메터추정문제를 연구하였으며 선행연구[1, 2, 5]에서는 완전자료인 경우 파레토분포의 파라메터추정문제를 취급하고 여러가지 추정알고리듬을 제기하였다.

론문에서는 불완전Ⅱ형자료로 주어진 경우 파레토분포의 중요한 파라메터들인 척도파라메터와 형태파라메터들의 통계적추론문제를 취급하였다.

T는 파레토분포에 따르는 우연량이라고 하자.

그러면 T의 분포함수는

$$F(t)=1-\left(\frac{\theta}{t}\right)^{\alpha}\ (t\geq\theta>0,\alpha>0)$$

이다. 여기서 θ는 척도파라메터이고 α는 형태파라메터이다.

그러면 T의 밀도함수는 $f(t)=\alpha\theta^{\alpha}t^{-(\alpha+1)}\ (t\geq\theta>0,\alpha>0)$이고 생존함수는

$$R(t)=1-F(t)=\left(\frac{\theta}{t}\right)^{\alpha}\ (t\geq\theta>0,\alpha>0)$$

이며 위험률함수는

$$\lambda(t)=\frac{f(t)}{1-F(t)}=\frac{\alpha}{t}\ (t\geq\theta>0,\alpha>0)$$

이다.

이제 $T_1, T_2, \cdots, T_n$은 파레토분포에 따르는 크기가 n인 표본이라고 하자. 이때 이러저러한 원인으로 하여 $T_i(i=\overline{1,\ n})$가 관측되지 못하고 단지 $X_i=\min(T_i, C_i)$, $\delta_i=I(T_i\leq C_i)$ $(i=\overline{1,\ n})$만이 관측되였다고 하자. 여기서 $C_1, C_2, \cdots, C_n$은 불완전정보를 표시하는 관측값들이다. 즉 $\delta_i=I(T_i\leq C_i)=1$인 경우에는 T_i가 관측된 경우이고 $\delta_i=I(T_i\leq C_i)=0$ 즉 $T_i\geq C_i$인 경우에는 T_i가 관측되지 못하고 C_i가 관측되였다는것을 의미한다. 이때에는 T_i가

C_i보다 크다는 정보밖에 모른다. 즉 관측된 자료는

$$(X_1,\ \delta_1),\ (X_2,\ \delta_2),\ \cdots,\ (X_n,\ \delta_n) \tag{1}$$

형태로 표시된다. 이제 $X_1,\ X_2,\ \cdots,\ X_n$의 순서통계량을 $X_{(1)},\ X_{(2)},\ \cdots,\ X_{(n)}$이라고 하자. 이때 r개의 관측값($r<n$)만이 관측되고 나머지 $n-r$개의 관측값은 $X_{(r)}$보다 크다는것밖에는 모른다고 하자. 즉

$$X_{(1)} < X_{(2)} < \cdots < X_{(r)} < X_{(r+1)},\ \cdots,\ X_{(n)} \tag{2}$$

이라고 하자. 여기서 r는 사전에 규정한 개수이다.

이것이 바로 불완전Ⅱ형자료이다.

여기서는 관측자료가 식 (2)로 주어졌을 때 미지파라메터 α와 θ의 점추정문제와 구간추정문제를 고찰하려고 한다.

1. 파라메터들의 추정량과 그 분포특성

이제 주어진 표본자료 식 (2)에 대한 우도함수는

$$L(\theta,\ \alpha) = p\{(x_1,\ \delta_1),\ ...,\ (x_n,\ \delta_n)\} = \prod_{i=1}^{n} f(x_i)^{\delta_i}\{1-F(x_i)\}^{1-\delta_i} =$$

$$= \prod_{i=1}^{r} \alpha\theta^{\alpha} x_{(i)}^{-(\alpha+1)} \left[\left(\frac{\theta}{x_{(r)}}\right)^{\alpha}\right]^{n-r} = \alpha^r \theta^{\alpha\cdot r} \prod_{i=1}^{r} x_{(i)}^{-(\alpha+1)} \left(\frac{\theta}{x_{(r)}}\right)^{\alpha(n-r)}$$

와 같다. 따라서 로그우도함수는

$$\ln L(\theta,\alpha) = r\ln\alpha + \alpha r\ln\theta - (\alpha+1)\sum_{i=1}^{r}\ln x_{(i)} + \alpha(n-r)\ln\frac{\theta}{x_{(r)}}\ \ (0<\theta\le x_{(1)},\ \alpha>0) \tag{3}$$

이다.

이제 α와 θ의 최대우도추정량을 구하기 위하여 식 (3)을 α와 θ에 관하여 편도함수를 취하고 령으로 놓은 방정식을 풀면 된다. 즉

$$\begin{cases} \dfrac{\partial}{\partial\alpha}\ln L(\theta,\ \alpha) = 0 \\ \dfrac{\partial}{\partial\theta}\ln L(\theta,\ \alpha) = 0 \end{cases} \tag{4}$$

을 풀면 된다. 따라서

$$\frac{\partial}{\partial\alpha}\ln L(\theta,\ \alpha) = \frac{r}{\alpha} + r\ln\theta - \sum_{i=1}^{r}\ln x_{(i)} + (n-r)\ln\frac{\theta}{x_{(r)}} = 0 \tag{5}$$

$$\frac{\partial}{\partial\theta}\ln L(\theta,\ \alpha) = \frac{\alpha r}{\theta} + \alpha(n-r)\frac{1}{\theta} = \frac{\alpha n}{\theta} = 0 \tag{6}$$

이다. 식 (5)로부터

$$\frac{r}{\alpha} = \sum_{i=1}^{r}\ln x_{(i)} - r\ln\theta - (n-r)\ln\frac{\theta}{x_{(r)}}$$

이며 따라서

$$\hat{\alpha}=\frac{r}{\sum_{i=1}^{r}\ln x_{(i)}-(n-r)\ln x_{(r)}-r\ln\theta}$$

이다. 식 (6)으로부터 $\ln L(\theta,\ \alpha)$는 θ에 관하여 증가함수이므로($0<\theta\le x_{(1)}$에서) 최대값점은 $\hat{\theta}=x_{(1)}$인 경우이다. 따라서 α와 θ의 최대우도추정량은

$$\begin{cases}\hat{\theta}=X_{(1)}\\ \hat{\alpha}=\dfrac{r}{\sum_{i=1}^{r}\ln X_{(i)}-(n-r)\ln X_{(r)}-r\ln\theta}\end{cases}\tag{7}$$

이다.

이제 이 추정량들의 분포를 구하기로 하자.

정리 1 $\hat{\theta}$, $\hat{\alpha}$이 식 (7)과 같이 표시되는 파레토분포의 최대우도추정량이라고 하면

1) $\hat{\theta}$은 파라메터가 θ, $n\alpha$인 파레토분포에 따르게 된다.

2) $\hat{\alpha}^{-1}$은 파라메터가 $r-1$, $r\alpha$인 감마분포 $G(r-1,\ r\alpha)$에 따르게 된다.

증명 1) $\hat{\theta}=X_{(i)}$ 즉 $\hat{\theta}$은 표본 $X_1,\ X_2,\ \cdots,\ X_n$의 1차순서통계량이므로 그 밀도함수는

$$f_{\hat{\theta}}(t)=n[1-F(t)]^{n-1}f(t)$$

이다. 여기서 $F(t)$와 $f(t)$는 파레토분포의 분포함수와 밀도함수이다. 따라서

$$f_{\hat{\theta}}(t)=n\left[1-\left(1-\left(\frac{\theta}{t}\right)^{\alpha}\right)\right]^{n-1}\alpha\theta^{\alpha}t^{-(\alpha+1)}=n\alpha\theta^{n\alpha}t^{-(n\alpha+1)}$$

이며 $\hat{\theta}$은 파라메터가 θ, $n\alpha$인 파레토분포에 따르게 된다.

2) 이제

$$\begin{aligned}z_1&=(n-1)[\ln X_{(2)}-\ln X_{(1)}]\\ z_2&=(n-2)[\ln X_{(3)}-\ln X_{(2)}]\\ &\vdots\\ z_i&=(n-i)[\ln X_{(i+1)}-\ln X_{(i)}]\end{aligned}$$

로 놓자. 그런데 X가 파라메터가 θ, α인 파레토분포에 따르게 되면 $\ln X$는 파라메터가 α인 지수분포에 따르게 된다. 그리고

$$z_1+z_2+\cdots+z_{r-1}=\sum_{i=1}^{r}\ln X_{(i)}-(n-r)\ln X_{(r)}-r\ln X_{(r)}$$

가 성립한다. 그런데 $z_1,\ z_2,\cdots,\ z_{r-1}$은 서로 독립이고 지수분포 $e(\alpha)$에 따르므로 $\sum_{i=1}^{r-1}z_i$는

감마분포 $G(r-1,\alpha)$ 에 따르게 된다.

따라서 $\dfrac{\sum_{i=1}^{r-1} z_i}{r}$ 은 감마분포 $G(r-1,r\alpha)$ 에 따르게 된다. 즉 $\hat{\alpha}^{-1}$ 은 파라메터가 $r-1$, $r\alpha$ 인 감마분포 $G(r-1,\ r\alpha)$ 에 따른다는것이 증명된다.(증명끝)

다음으로 이 추정량들의 특성값들을 고찰하자.

따름 $\hat{\theta},\hat{\alpha}$ 을 식 (7)과 같이 표시되는 파레토분포의 최대우도추정량이라고 하면 그 수학적기대값과 분산은 다음과 같다.

1) $E\hat{\theta}=\dfrac{n\alpha}{n\alpha-1}\theta\ \left(n>\dfrac{1}{\alpha}\right)$

$$\mathrm{Var}\hat{\theta}=\frac{n\alpha}{(n\alpha-1)^2(n\alpha-2)}\theta^2\ \left(n>\frac{2}{\alpha}\right)$$

2) $E\hat{\alpha}=\dfrac{r}{r-2}\alpha\ (r>2)$

$$\mathrm{Var}\hat{\alpha}=\frac{r^2}{(r-2)^2(r-3)}\theta^2\ (r>3)$$

따름으로부터 알수 있는바와 같이 $\hat{\theta}$, $\hat{\alpha}$ 은 θ,α 의 불편추정량이 되지 않는다.

따라서 θ, α 의 불편추정량을 구하면 다음과 같다.

정리 2

$$\begin{cases}\hat{\theta}_1=\left[1-\dfrac{r}{n(r-1)\hat{\alpha}}\right]\hat{\theta}=\left[1-\dfrac{\sum_{i=1}^{r}\ln X_{(i)}+(n-r)\ln X_{(r)}-n\ln X_{(i)}}{n(r-1)}\right]X_{(1)}\\ \hat{\alpha}_1=\dfrac{r-2}{r}\hat{\alpha}=\dfrac{r-2}{\sum_{i=1}^{r}\ln X_{(i)}+(n-r)\ln X_{(r)}-n\ln X_{(i)}}\end{cases}$$

는 각각 θ,α 의 불편추정량이다. 그리고 $\hat{\theta}_1,\hat{\alpha}_1$ 의 분산은 다음과 같다.

$$\begin{cases}\mathrm{Var}\hat{\theta}_1=\dfrac{n^4\alpha^4+4n^3\alpha^3r+3n^2\alpha^2+2n\alpha+r-4n^3\alpha^3-n^4\alpha^4r-4n^2\alpha^2r}{n\alpha(r-1)(n\alpha-1)^2(n\alpha-2)}\theta^2\ \left(r>1,\ n>\dfrac{2}{\alpha}\right)\\ \mathrm{Var}\hat{\alpha}_1=\dfrac{1}{r-3}\alpha^2\ (r>3)\end{cases}$$

2. 불완전Ⅱ형자료에 기초한 파레토분포파라메터들에 대한 구간추정

미지파라메터 α 와 θ 의 믿을구간을 구하는 문제를 고찰하자.

정리 3 미지파라메터 α 의 $100(1-\gamma)\%$ 믿을구간은 다음과 같다.

$$\left(\frac{\hat{\alpha}}{2r}\chi_1^2,\ \frac{\hat{\alpha}}{2r}\chi_2^2\right)$$

여기서 χ_1^2, χ_2^2는 $\int_0^{\chi_1^2} k_{2r-2}(x)dx = \int_{\chi_2^2}^{+\infty} k_{2r-2}(x)dx = \frac{\gamma}{2}$인 수표값이다.

정리 4 미지파라메터 θ의 $100(1-\gamma)\%$ 믿을구간은 다음과 같다.

$$\left(\hat{\theta}e^{-\frac{F_2}{(r-1)\hat{\alpha}}},\ \hat{\theta}\,e^{-\frac{F_1}{(r-1)\hat{\alpha}}}\right)$$

여기서 F_1, F_2는 $\int_0^{F_1} f_{2,\ 2r-2}(x)dx = \int_{F_2}^{+\infty} f_{2,\ 2r-2}(x)dx = \frac{\gamma}{2}$인 수표값이다.

참 고 문 헌

[1] B. C. Arnold; Pareto Distribution, International Co－operative Publishing House, 55～87, 1983.

[2] A. M. Hossain et a1.; Communications in Statistics Theory and Methods, 29, 4, 23, 2000.

[3] Y. Wang et a1.; Statistics and Probability Letters, 73, 199, 2005.

[4] N. S. Samindra et a1.; Statistics and Probability Letters, 79, 899, 2009.

[5] G. Qin et a1.; Statistics and Probability Letters, 77, 549, 2007.

[6] Wen Chuan Lee et a1.; Journal of Computational and Applied Mathematics, 231, 648, 2009.

[7] 高旅端 等; 数理统计与应用概率, 10, 83, 2000.

[8] 郑祖康; 应用概率统计, 20, 119, 2004.

[9] 顾益明; 上海师范大学学报, 30, 28, 2001.

주체106(2017)년 12월 5일 원고접수

Statistical Inference for Parameters of Pareto Distribution based on Censored Type Ⅱ Data

Ri Kwang Son, Han Ryu Gyong

In this paper, we study point and interval estimations for the parameters of Pareto distribution based on censored type Ⅱ data.

Key words : Pareto distribution, censored type Ⅱ data, interval estimation