

## 비대칭도에 대한 영향함수 유도 및 역할 분석<sup>†</sup>

제아라<sup>1</sup> · 이윤희<sup>2</sup> · 김홍기<sup>3</sup>

<sup>1</sup>통계청 · <sup>2,3</sup>충남대학교 정보통계학과

접수 2024년 4월 24일, 수정 2024년 5월 23일, 게재 확정 2024년 5월 30일

### 요약

본 연구에서는 3차 격률과 표준편차의 세제곱에 대한 영향함수를 이용하여 비대칭도에 대한 영향함수를 유도한다. 그리고 표본으로부터 얻어지는 경험적 분포를 기반으로 비대칭도에 대한 경험적 영향함수를 유도하여 한 개의 관측값을 제외하거나 추가하는 경우의 비대칭도의 실제차이와 비대칭도에 대한 경험적 영향함수로 추정한 추정차이의 관계를 살펴본다. 그리고 비대칭도에 대한 영향함수의 타당성을 검증하기 위하여 통계청 조사자료 중 비대칭의 분포를 가지는 데이터를 이용하여 모의실험을 실시하고 그 결과를 살펴본다. 이를 통해 비대칭도의 영향함수의 역할과 그 의미에 대하여 살펴본 점에서 본 연구의 의의를 갖는다.

주요용어: 경험적 영향함수, 비대칭도, 실제차이, 영향함수, 이상치, 추정차이.

### 1. 서론

이상치가 가지는 영향의 정도를 측정하여 데이터에 대한 분석과정에 도움이 되기 위한 방법론이 지속적으로 연구되어 왔으며, Hampel (1974)은 영향함수 (influence function)를 활용하여 이상치를 판별할 수 있는 방법을 가장 먼저 소개하였다. 영향함수는 하나의 관측값이 전체 데이터에 미치는 영향의 정도를 알 수 있는데, 특히 이상치를 판별하는 데 주로 활용되며 Hampel은 영향함수를 활용하여 거의 모든 통계량에서 이상치를 판별할 수 있음을 보여주었다. 이후 Campbell (1978)은 판별분석에서 영향함수를 사용하여 이상치를 탐지하였고, Radhakrishnan과 Kshirsagar (1981)은 다변량 분석에서 여러 가지 모수에 대한 이론적인 영향함수들을 유도하였다. Cook (1977)은 회귀분석에서 영향력 있는 관측값에 대해 연구하였으며, Cook과 Weisberg (1980, 1982)는 회귀분석에서 회귀진단방법으로, Critchley (1985)는 주성분 분석에서 영향력 있는 관측값을 찾아내는 방법으로 적용하였다. Kim (1992, 1994)은 이차원 분할표의 대응분석에서 얻어진 고유치들에 대해 영향함수를 유도하였고, 이를 다차원 분할표의 대응분석으로 확장하였다.

Kim과 Lee (1996), Kim (1998) 등은  $\chi^2$ 통계량에 대한 영향함수, 분할 Table에서의 행 영향함수를 유도하였고, 또한 Kim과 Kim (2005)은  $t$  통계량에 대한 영향함수를, Lee와 Kim (2008)은 변이계수에 대한 영향함수를 유도하는 등 다양한 통계량에 대한 영향함수를 유도해 내는 연구가 활발히 이루어져 왔

<sup>†</sup> 이 논문은 제아라의 박사 논문의 발췌 논문이다. 이 성과는 정부(과학기술정보통신부)의 재원으로 한국연구재단의 지원을 받아 수행되었음(No. NRF-2022M3J6A1084843).

<sup>1</sup> (35208) 대전광역시 서구 청사로 189, 통계청, 통계주무관.

<sup>2</sup> (34134) 대전광역시 유성구 대학로 99, 충남대학교 정보통계학과, 조교.

<sup>3</sup> 교신저자: (34134) 대전광역시 유성구 대학로 99, 충남대학교 정보통계학과, 교수.

E-mail: honggiekim@cnu.ac.kr

다. 최근에는 Kim과 Kim (2017)이 Ghosh & Kim 계수에 대한 영향함수를 유도해냈다. Kim과 Kim (2019)의 모분포의 형태에 따른  $t$ 통계량에 대한 영향함수의 성능에 대한 연구와 Park과 Kim (2019)의  $t$ 통계량에 가장 작은 영향을 미치는 관측값의 위치에 대한 연구, Kang과 Kim (2020, 2021)은 경험적 영향함수와 표본영향함수의 차이 및 보정에 대하여 연구하였고, 이를  $t$ 통계량으로 확장하였다. Kim (2021)은 3차 적률에 대한 경험적 영향함수를 유도하여 그 타당성을 검증하였고, Kang 등 (2022)은 경험적 영향함수를 이용한 표본영향함수 근사들 간의 성능을 비교하는 등 영향함수의 유도와 다양한 분야에서의 활용에 대한 연구가 지금까지도 활발하게 진행되고 있다. 한편 가장 최근의 연구로는 Lee 등 (2024)에 의한  $\sigma^k$ 에 대한 영향함수유도 연구가 있다.

본 연구에서는 3차 적률과  $\sigma^k$ 의 특별한 경우인  $\sigma^3$ 에 대한 영향함수를 이용하여 비대칭도에 대한 영향함수를 유도한다. 그리고 표본으로부터 얻어지는 경험적 분포 (empirical distribution)를 기반으로 경험적 영향함수를 유도하여 한 개의 관측값을 제외하거나 추가하는 경우의 비대칭도의 실제차이와 비대칭도에 대한 경험적 영향함수로 추정한 추정차이의 관계를 살펴보고, 두 차이의 관계로 본 비대칭도에 대한 영향함수의 역할에 대하여 분석해 본다. 2절에서는 영향함수의 정의, 기존에 정의된 평균, 분산 및 표준편차의 영향함수를 이용하여 비대칭도에 대한 영향함수의 유도에 대하여 다룬다. 3절에서는 경험적 영향함수의 정의, 비대칭도에 대한 경험적 영향함수를 유도한다. 4절에서는 타당성 검증을 위하여 모의 실험을 실시한다. 통계청 조사자료 중 실제 비대칭의 분포를 가지는 데이터를 이용하여 그 결과를 살펴본다. 마지막 절에서는 본 연구의 결과를 제시한다. 본 논문의 결과는 일반적으로 비대칭도를 줄이고자 하는 경제학적 자료들 (예를 들어, 불평등한 소득)의 비대칭도를 줄이는 방법으로 사용될 수도 있을 것이다.

## 2. 영향함수와 비대칭도

### 2.1. 영향함수의 정의 및 평균, 분산, 표준편차의 영향함수

$T$ 는 (누적)분포함수  $F$ 에 대해  $T(F) = c$ 와 같이 실수값  $c$ 를 갖는 범함수 (real-valued function)이고,  $\delta_x(t)$ 는 실수 공간의 한 점인  $x$ 에서 확률이 1인 분포함수로

$$\delta_x(t) = \begin{cases} 0, & t < x \\ 1, & t \geq x. \end{cases} \quad (2.1)$$

와 같이 나타낼 수 있으며, 이를 퇴화분포함수 (degenerated distribution function)라고 한다.  $F(t)$ 를  $F$ 로,  $\delta_x(t)$ 를  $\delta_x$ 로,  $F_\epsilon(t)$ 를  $F_\epsilon$ 으로 각각 표기하면 분포함수  $F$ 에 임의의 관측값  $x$ 를 추가함으로써 생기는 분포함수  $F$ 와 퇴화분포함수  $\delta_x$ 의 혼합분포함수  $F_\epsilon$ 는 다음과 같다.

$$F_\epsilon = (1 - \epsilon)F + \epsilon\delta_x, \quad 0 < \epsilon < 1. \quad (2.2)$$

이 때,  $F_\epsilon$ 를  $F$ 의 섭동 (perturbation)이라고 한다.

Hampel (1974)은 관측값  $x$ 가 추가됨으로써 범함수  $T(F)$ 에 미치는 영향을 나타내는 영향함수  $IF(T, x)$ 를 분포함수  $F$ 의 섭동  $F_\epsilon$ 를 이용해 다음과 같이 정의하였다.

$$IF(T, x) = \lim_{\epsilon \rightarrow 0} \frac{T(F_\epsilon) - T(F)}{\epsilon} = \lim_{\epsilon \rightarrow 0} \frac{T[(1 - \epsilon)F + \epsilon\delta_x] - T(F)}{\epsilon}. \quad (2.3)$$

식 (2.3)을 살펴보면 영향함수  $IF(T, x)$ 는 분포  $F$ 에 대한 범함수  $T(F)$ 의 도함수이며 관측값  $x$ 에서 섭동된 범함수  $T(F_\epsilon)$ 에 의한  $T(F)$ 의 순간변화율을 나타낸다. 로피탈의 정리를 이용해 식 (2.3)을 계산하면 다음과 같다.

$$IF(T, x) = \lim_{\epsilon \rightarrow 0} \frac{T(F_\epsilon) - T(F)}{\epsilon} = \lim_{\epsilon \rightarrow 0} \left[ \frac{\partial T(F_\epsilon)}{\partial \epsilon} \right] = \left[ \frac{\partial T(F_\epsilon)}{\partial \epsilon} \right]_{\epsilon=0}. \quad (2.4)$$

한편, 모집단의 분포가 갖는 평균  $\mu$ 와 분산  $\sigma^2$ 에 대한 범함수를 각각  $T_1, T_2$ 라 하고, 모집단의 분포함수  $F(t)$ 의 확률밀도함수를  $f(t)$ 라 하면  $\partial F(t)/\partial t = f(t)$ 가 성립하므로 모집단의 평균과 분산을 다음과 같이 범함수의 형태로 나타낼 수 있다.

$$\begin{aligned} T_1(F) &= \mu = \int t f(t) dt = \int t dF(t), \\ T_2(F) &= \sigma^2 = \int (t - \mu)^2 f(t) dt = \int (t - \mu)^2 dF(t). \end{aligned} \quad (2.5)$$

식 (2.4)와 식 (2.5)를 이용해 평균  $\mu$ 와 분산  $\sigma^2$ 에 대한 영향함수를 유도하면 다음과 같다 (Hampel, 1974).

$$\begin{aligned} IF(T_1, x) &= IF(\mu, x) = x - \mu, \\ IF(T_2, x) &= IF(\sigma^2, x) = (x - \mu)^2 - \sigma^2. \end{aligned} \quad (2.6)$$

$T_3(F) = \sqrt{T_2(F)} = \sigma$ 는 영향함수  $IF(T_2, x)$ 를 이용하면 다음과 같이 정리할 수 있다.

$$\begin{aligned} IF(T_3, x) &= \left[ \frac{1}{2\sqrt{T_2(F_\epsilon)}} \times \frac{\partial T_2(F_\epsilon)}{\partial \epsilon} \right]_{\epsilon=0} \\ &= \frac{1}{2\sigma} IF(T_2, x). \end{aligned} \quad (2.7)$$

즉, 식 (2.7)에 의해  $IF(T_3, x) = IF(\sigma, x) = (1/2\sigma) \cdot \{(x - \mu)^2 - \sigma^2\}$ 로 표준편차  $\sigma$ 에 대한 영향함수를 유도할 수 있다.

## 2.2. 비대칭도에 대한 영향함수

분포의 비대칭의 정도를 나타내는 비대칭도 (skewness)는 평균에 대한 3차 적률  $E(X - \mu)^3$ 을  $\sigma^3$ 으로 나눈 형태인

$$skewness = \frac{E(X - \mu)^3}{\sigma^3}$$

으로 주어지며, skewness에 대한 영향함수 유도를 위해 3차 적률과  $\sigma^3$ 에 대한 영향함수 유도가 선행되어야 한다. Kim (2021)은 분포함수  $F$ 의 3차 적률

$$M^3(F) = \int (t - \mu)^3 dF(t)$$

의 영향함수를 유도한 바 있으며,  $M^3(F)$ 를  $M^3$ 로 표기하여 간단히 정리하면

$$\begin{aligned}
 IF(M^3, x) &= \lim_{\epsilon \rightarrow 0} \frac{T(F_\epsilon) - T(F)}{\epsilon} \\
 &= \lim_{\epsilon \rightarrow 0} \frac{T[(1-\epsilon)F + \epsilon\delta_x] - T(F)}{\epsilon} \\
 &= \lim_{\epsilon \rightarrow 0} \frac{\int (t-\mu)^3 d[(1-\epsilon)F + \epsilon\delta_x](t) - \int (t-\mu)^3 dF(t)}{\epsilon} \\
 &= \lim_{\epsilon \rightarrow 0} \frac{(1-\epsilon) \int (t-\mu)^3 dF(t) + \epsilon(x-\mu)^3 - \int (t-\mu)^3 dF(t)}{\epsilon} \\
 &= \lim_{\epsilon \rightarrow 0} \frac{-\epsilon \int (t-\mu)^3 dF(t) + \epsilon(x-\mu)^3}{\epsilon} \\
 &= - \int (t-\mu)^3 dF(t) + (x-\mu)^3 \\
 &= (x-\mu)^3 - M^3
 \end{aligned}$$

이며,  $\sigma^3$ 에 대한 영향함수는 Lee (2024)의 결과에 따라

$$IF(\sigma^3, x) = \frac{3}{2}\sigma\{(x-\mu)^2 - \sigma^2\}$$

이 됨을 알 수 있다.

위의 두 영향함수들을 이용하여 비대칭도에 대한 영향함수는 다음과 같이 유도할 수 있다. 본 연구에서는 *skewness*를 *skew*로 표기하며,  $T_4(F) = M^3$ ,  $T_5(F) = \sigma^3$ 으로 정의 후 유도한다.

$$\begin{aligned}
 IF(skew, x) &= IF\left(\frac{M^3}{\sigma^3}, x\right) \\
 &= IF\left(\frac{T_4(F)}{T_5(F)}, x\right) \\
 &= \left[ \frac{\partial \left\{ \frac{T_4(F_\epsilon)}{T_5(F_\epsilon)} \right\}}{\partial \epsilon} \right]_{\epsilon=0} \\
 &= \left[ \frac{\frac{\partial T_4(F_\epsilon)}{\partial \epsilon} \cdot T_5(F_\epsilon) - T_4(F_\epsilon) \cdot \frac{\partial T_5(F_\epsilon)}{\partial \epsilon}}{\{T_5(F_\epsilon)\}^2} \right]_{\epsilon=0} \\
 &= \frac{\{(x-\mu)^3 - M^3\} \cdot \sigma^3 - M^3 \cdot \frac{3\sigma}{2}\{(x-\mu)^2 - \sigma^2\}}{\sigma^6} \\
 &= \frac{1}{\sigma^3} \left[ \frac{(x-\mu)^2}{\sigma^2} \left\{ \sigma^2(x-\mu) - \frac{3}{2}M^3 \right\} \right] + \frac{1}{2} \cdot \frac{M^3}{\sigma^3} \\
 &= \frac{(x-\mu)^2}{\sigma^2} \left\{ \frac{x-\mu}{\sigma} - \frac{3}{2} \cdot \frac{M^3}{\sigma^3} \right\} + \frac{1}{2} \cdot \frac{M^3}{\sigma^3} \\
 &= z^2 \left( z - \frac{3}{2}skew \right) + \frac{1}{2}skew. \tag{2.8}
 \end{aligned}$$

즉, 비대칭도에 대한 영향함수는 다음과 같이 정리할 수 있다.

$$IF(skew, x) = z^2 \left( z - \frac{3}{2} skew \right) + \frac{1}{2} skew. \quad (2.9)$$

위에서 구해진 영향함수를 사용하여 각 관찰치가 자료분포의 비대칭도에 끼치는 영향의 상대적 비교를 할 수가 있으며 이를 사용하여 비대칭도를 줄이는 방향으로 영향을 미치는 관찰치들을 선별해 낼 수 있을 것이다.

### 3. 경험적 영향함수와 비대칭도

#### 3.1. 경험적 영향함수의 정의

모집단의 분포함수  $F$ 에 대한 범함수  $T(F)$ 의 영향함수  $IF(T, x)$ 가 정의된다면 모집단의 분포함수  $F$  대신 표본으로부터 얻어지는 경험적 분포(empirical distribution)의 표본분포함수  $\hat{F}$ 를 사용한 범함수  $T(\hat{F})$ 의 영향함수를 경험적 영향함수(empirical influence function: EIF)라고 한다. 경험적 영향함수는 모집단의 영향함수 식에서 모수 대신 추정량으로 대체하여 다음과 같이 구할 수 있다. 표본분포함수  $\hat{F}$ 에 대한 표본평균, 표본분산, 표본표준편차의 범함수는 다음과 같다.

$$T_1(\hat{F}) = \bar{x}, \quad T_2(\hat{F}) = s^2, \quad T_3(\hat{F}) = s. \quad (3.1)$$

위의 식(3.1)에 대한 경험적 영향함수는 다음과 같다.

$$\begin{aligned} EIF(T_1, x) &= EIF(\bar{x}, x) = x - \bar{x}, \\ EIF(T_2, x) &= EIF(s^2, x) = (x - \bar{x})^2 - s^2, \\ EIF(T_3, x) &= EIF(s, x) = \frac{1}{2s} \{(x - \bar{x})^2 - s^2\}. \end{aligned}$$

표본에서  $i$ 번째 관측값  $x_i$ 를 제외하여 통계량을 구한 경우를  $(-i)$ , 한 번 더 포함하여 통계량을 구한 경우를  $(+i)$ 로 표시하도록 한다. 또한 범함수에서 섭동량  $\epsilon$ 은 관측값이 한 개 제외되는 경우에는  $1/(n-1)$ 로, 관측값이 한 개 추가되는 경우에는  $1/(n+1)$ 로 고려할 수 있다. 이를 바탕으로 관측값의 제거나 추가에 따른 평균, 표본분산, 표본표준편차의 변화에 대한 추정량과 경험적 영향함수의 관계를 다음과 같이 정리할 수 있다.

$$\begin{aligned} \bar{x}_{(-i)} - \bar{x} &= -EIF(\bar{x}, x_i) \times \frac{1}{n-1} = -\frac{x_i - \bar{x}}{n-1}, \\ \bar{x}_{(+i)} - \bar{x} &= EIF(\bar{x}, x_i) \times \frac{1}{n+1} = \frac{x_i - \bar{x}}{n+1}, \\ s_{(-i)}^2 - s^2 &\simeq -EIF(s^2, x_i) \times \frac{1}{n-1} = \frac{-\{(x_i - \bar{x})^2 - s^2\}}{n-1}, \\ s_{(+i)}^2 - s^2 &\simeq EIF(s^2, x_i) \times \frac{1}{n+1} = \frac{(x_i - \bar{x})^2 - s^2}{n+1}, \\ s_{(-i)} - s &\simeq -EIF(s, x_i) \times \frac{1}{n-1} = -\frac{1}{n-1} \left\{ \frac{(x_i - \bar{x})^2 - s^2}{2s} \right\}, \\ s_{(+i)} - s &\simeq EIF(s, x_i) \times \frac{1}{n+1} = \frac{1}{n+1} \left\{ \frac{(x_i - \bar{x})^2 - s^2}{2s} \right\}. \end{aligned}$$

즉,  $n$ 개의 표본에서 계산된 추정량과  $i$ 번째 관측값을 제외하거나 추가한 표본에서 계산된 실제 차이는 경험적 영향함수에 섭동량  $\epsilon$ 을 곱한 것으로 예측이 가능하다.

위와 마찬가지로, 3차 적률과  $\sigma^3$ 의 경험적 영향함수는 다음과 같으며

$$\begin{aligned} EIF(T_4, x) &= EIF(\hat{M}^3, x) = (x - \bar{x})^3 - \hat{M}^3, \\ EIF(T_5, x) &= EIF(s^3, x) = \frac{3}{2}s\{(x - \bar{x})^2 - s^2\}. \end{aligned}$$

관측값의 제거나 추가에 따른 두 통계량의 변화에 대한 추정량과 경험적 영향함수와의 관계는 다음과 같다 (Kim, 2021; Lee, 2024).

$$\begin{aligned} \hat{M}_{(-i)}^3 - \hat{M}^3 &= -EIF(\hat{M}^3, x_i) \times \frac{1}{n-1} = \frac{-(x_i - \bar{x})^3 + \hat{M}^3}{n-1}, \\ \hat{M}_{(+i)}^3 - \hat{M}^3 &= EIF(\hat{M}^3, x_i) \times \frac{1}{n+1} = \frac{(x_i - \bar{x})^3 - \hat{M}^3}{n+1}, \\ s_{(-i)}^3 - s^3 &\simeq -EIF(s^3, x_i) \times \frac{1}{n-1} = -\frac{1}{n-1} \cdot \frac{3}{2}s\{(x_i - \bar{x})^2 - s^2\}, \\ s_{(+i)}^3 - s^3 &\simeq EIF(s^3, x_i) \times \frac{1}{n+1} = \frac{1}{n+1} \cdot \frac{3}{2}s\{(x_i - \bar{x})^2 - s^2\}. \end{aligned}$$

### 3.2. 비대칭도에 대한 경험적 영향함수

표본으로부터 구한  $E(X - \bar{X})^3/s^3$ 은 표본이 갖는 비대칭도로써  $\widehat{skew}$ 라고 쓸 수 있으며 식 2.8에 의하여 비대칭도에 대한 경험적 영향함수는 다음과 같다.

$$EIF(\widehat{skew}, x) = z^2 \left( z - \frac{3}{2}\widehat{skew} \right) + \frac{1}{2}\widehat{skew}.$$

하나의 관측값이 제거되거나 포함된 표본에서의 비대칭도는 다음과 같이 나타낼 수 있다.

$$\widehat{skew}_{(-i)} = \frac{\sum_{j \neq i}^n (x_j - \bar{x}_{(-i)})^3 / (n-1)}{s_{(-i)}^3}, \quad (3.2)$$

$$\widehat{skew}_{(+i)} = \frac{\left[ \sum_{j=1}^n \{(x_j - \bar{x}_{(+i)})^3\} + (x_i - \bar{x}_{(+i)})^3 \right] / (n+1)}{s_{(+i)}^3}. \quad (3.3)$$

식 (3.2)과 식 (3.3)에서의 분자부분은  $E(X - \mu)^3$ 의 추정치이며, 분모부분은 각 경우의 표본표준편차의 세제곱이다. 이 두 값에 각각  $\widehat{skew}$ 을 빼주면 표본에서 구한 비대칭도와 하나의 관측값을 제외하거나 추가하는 경우의 비대칭도의 실제차이를 구할 수 있다. 즉  $i$ 번째 관측값을 제외 또는 추가함으로써 생기는 범함수의 합수값의 차이와 섭동량  $\epsilon$ 을 고려하여 나타낸 영향함수의 관계는 다음과 같다.

$$\begin{aligned}\widehat{\text{skew}}_{(-i)} - \widehat{\text{skew}} &\simeq - EIF(\widehat{\text{skew}}, x_i) \times \frac{1}{n-1} \\ &= - \frac{1}{n-1} \left\{ z_i^2 \left( z_i - \frac{3}{2} \widehat{\text{skew}} \right) + \frac{1}{2} \widehat{\text{skew}} \right\}\end{aligned}\quad (3.4)$$

$$\begin{aligned}\widehat{\text{skew}}_{(+i)} - \widehat{\text{skew}} &\simeq EIF(\widehat{\text{skew}}, x_i) \times \frac{1}{n+1} \\ &= \frac{1}{n+1} \left\{ z_i^2 \left( z_i - \frac{3}{2} \widehat{\text{skew}} \right) + \frac{1}{2} \widehat{\text{skew}} \right\}.\end{aligned}\quad (3.5)$$

식 (3.4)과 식 (3.5)는 비대칭도의 실제차이와 경험적 영향함수로 추정한 추정차이와의 관계를 나타낸 식으로, 모의실험을 통해서 그 관계를 확인해 보고자 한다.

#### 4. 모의실험을 통한 타당성 검증

##### 4.1. 비대칭 분포를 가지는 데이터의 경험적 영향함수

3.2절에서 제시한 식 (3.4)과 식 (3.5)의 타당성을 확인하기 위해 모의실험을 진행하였다. 통계청에서 실시한 2021년 가계금융복지조사의 결과자료에서 관찰된 변수들 중 하나인 ‘자산’데이터를 사용하였으며, 78개 가구의 ‘자산’항목을 종속변수로 한 기술통계는 Table 4.1과 같다.  $\text{skew}$ 의 값이 2.35577인 것으로 보아 어느 정도 비대칭 분포를 가지는 데이터임을 알 수 있으며, 실험의 편의를 위해서 종속변수를 오름차순으로 정렬하여 진행하였다.

Table 4.1 Summary of observations

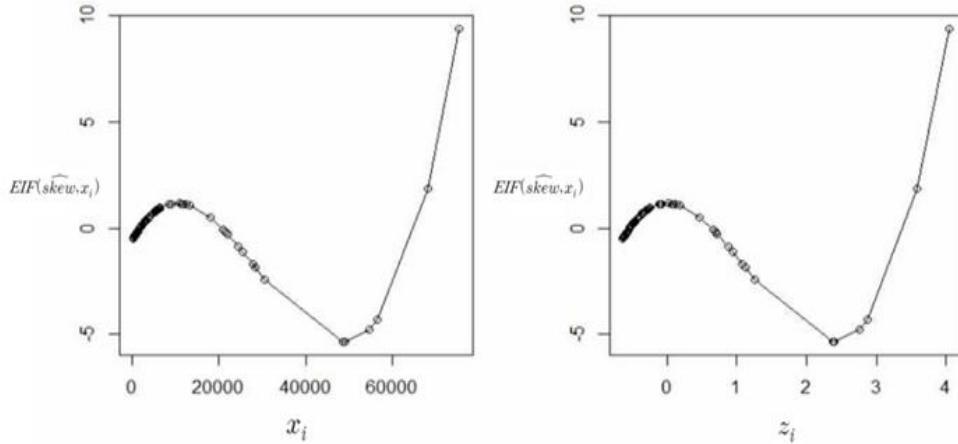
	Min	Mean	Max	Sd	Skewness
	161	10,401	75,480	16,112.98	2.35577

위 78개의 데이터를 바탕으로 표본평균, 표본표준편차, 비대칭도,  $z_i$ 값, 비대칭도에 대한 경험적 영향 함수를 계산한 결과가 아래와 같다.

Table 4.2 Empirical influence function on skewness of data

$i$	$x_i$	$\bar{x}$	$s$	$\widehat{\text{skew}}$	$z_i$	$EIF(\widehat{\text{skew}}, x_i)$
1	161	10401.2308	16112.9829	2.35577	-0.63553	-0.50602
2	180	10401.2308	16112.9829	2.35577	-0.63435	-0.49931
3	305	10401.2308	16112.9829	2.35577	-0.62659	-0.45549
...	...	...	...	...	...	...
21	1640	10401.2308	16112.9829	2.35577	-0.54374	-0.02760
22	1719	10401.2308	16112.9829	2.35577	-0.53883	-0.00453
23	1800	10401.2308	16112.9829	2.35577	-0.53381	0.01886
24	2100	10401.2308	16112.9829	2.35577	-0.51519	0.10324
...	...	...	...	...	...	...
76	56600	10401.2308	16112.9829	2.35577	2.86718	-4.30106
77	68220	10401.2308	16112.9829	2.35577	3.58833	1.88190
78	75480	10401.2308	16112.9829	2.35577	4.03890	9.41980

위의 Table 4.2를 보면  $EIF(\widehat{\text{skew}}, x_i)$ 는 점점 증가하였다가 극대점을 지나며 감소 후, 극소점을 지나 다시 증가하는 3차 함수의 형태를 보이고 있다는 것을 알 수 있다. 그래프로 나타내면 아래의 Figure 4.1과 같다. 영향함수가 0에 제일 가까운 관측값은  $x_{22}$ 와  $x_{23}$ 이며, Figure 4.1의 오른쪽 그래프에서  $z_i = (x_i - \bar{x})/s$  이므로 두 그래프의 개형은 동일하다.

Figure 4.1 Empirical influence function on skewness according to  $x_i, z_i$ 

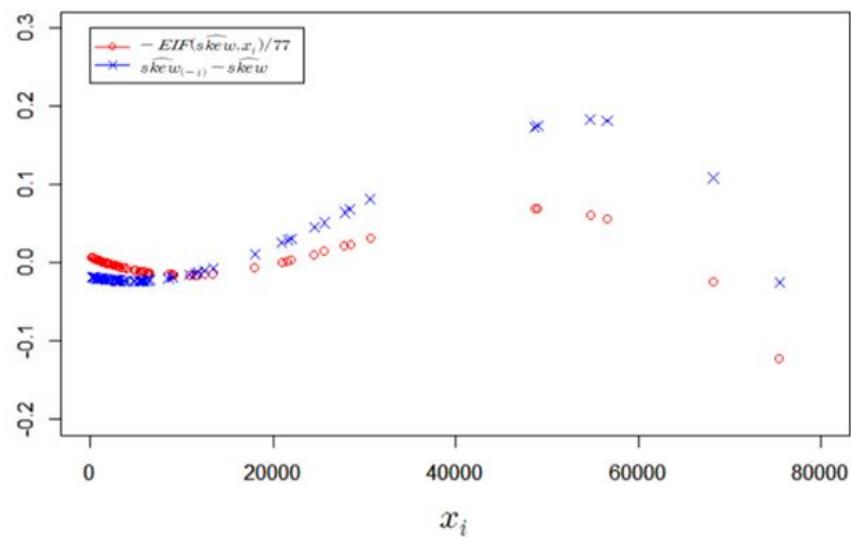
#### 4.2. 한 개의 관측값을 제외하는 경우

$\widehat{\text{skew}}$ 에 대한 경험적 영향함수의 타당성 검증을 위하여 먼저  $-\text{EIF}(\widehat{\text{skew}}, x_i)/(n-1)$ 과  $\widehat{\text{skew}}_{(-i)} - \widehat{\text{skew}}$ 를 비교하도록 한다. Table 4.3은 관측값  $x_i$ 를 제외했을 때의 표본평균, 표본표준편차 및 추정차이를 구하기 위한 값들과 실제차이와 추정차이, 그리고 실제차이와 추정차이의 차이를 구한 표이다.

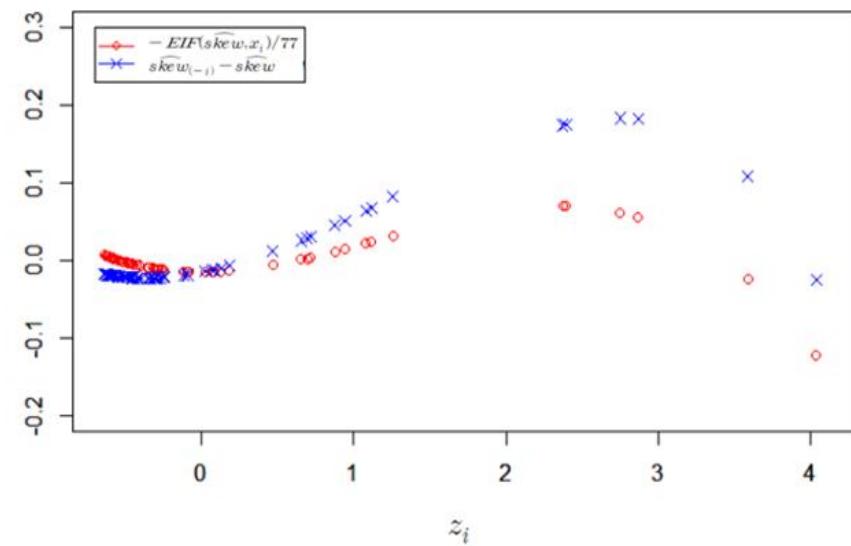
Table 4.3 Actual difference, estimated difference, and difference between the two values, when deleting one observation

$i$	$\bar{x}_{(-i)}$	$s_{(-i)}$ (W)	$\frac{\sum_{j \neq i}^{n-1} (x_j - \bar{x}_{(-i)})^3}{n-1}$ (X)	$\widehat{\text{skew}}_{(-i)}$ (X/W <sup>3</sup> )	$\widehat{\text{skew}}_{(-i)} - \widehat{\text{skew}}$ (Y)	$\frac{-\text{EIF}(\widehat{\text{skew}}, x_i)}{77}$ (Z)	$\text{Diff}_{(-i)}$ (Y - Z)
1	10534.2208	16175.4967	9894000413290	2.33776	-0.01802	0.00657	-0.02459
2	10533.974	16175.6569	9894112079296	2.33771	-0.01806	0.00648	-0.02454
3	10532.3506	16176.7032	9894854148152	2.33743	-0.01834	0.00592	-0.02425
...	...	...	...	...	...	...	...
58	10418.6494	16217.9038	9969556108930	2.33718	-0.01859	-0.01497	-0.00362
59	10395.2727	16218.5565	9987729933312	2.34116	-0.01462	-0.01526	0.00064
60	10386.7013	16218.1287	9994388490390	2.34290	-0.01287	-0.01508	0.00221
...	...	...	...	...	...	...	...
76	9801.24675	15316.5408	9119523324703	2.53799	0.18222	0.05586	0.12636
77	9650.33766	14781.2769	7959058056643	2.46448	0.10871	-0.02444	0.13315
78	9556.05195	14373.3688	6921176950798	2.33080	-0.02498	-0.12234	0.09736

$x_{59} = 10860$ 일 때 두 값의 차이가 0.00064로 0에 가까운 값을 가지며, 이를 중심으로 점점 차이가 커지는 것을 알 수 있다. 이 때,  $x_{59}$ 는 평균인 10401.2308과 가장 가까운 값으로, 제외되는 관측값이 평균과 가까워질수록  $-\text{EIF}(\widehat{\text{skew}}, x_i)/(n-1)$ 과  $\widehat{\text{skew}}_{(-i)} - \widehat{\text{skew}}$ 의 값이 작아지고 이 두 값이 거의 비슷해지는 것을 알 수 있다. Table 4.3에 해당하는 값들을 그래프로 나타낸 결과가 Figure 4.2와 Figure 4.3과 같다.



**Figure 4.2** Actual difference and estimated difference according to  $x_i$  when deleting one observation



**Figure 4.3** Actual difference and estimated difference according to  $z_i$  when deleting one observation

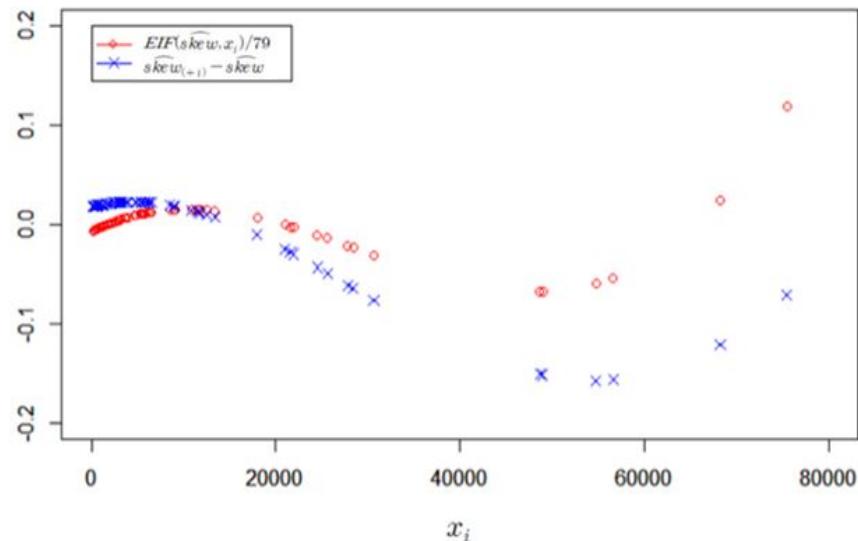
### 4.3. 한 개의 관측값을 추가하는 경우

이 절에서는 한 개의 관측값을 한 번 더 포함시킨 경우로,  $EIF(\widehat{skew}, x_i)/(n + 1)$ 과  $\widehat{skew}_{(+i)} - \widehat{skew}$ 를 비교하도록 한다. Table 4.4는 관측값  $x_i$ 를 한 번 더 추가했을 때의 표본평균, 표본표준편차 및 추정차이를 구하기 위한 값들과 실제차이와 추정차이, 그리고 실제차이와 추정차이의 차이를 구한 표이다.

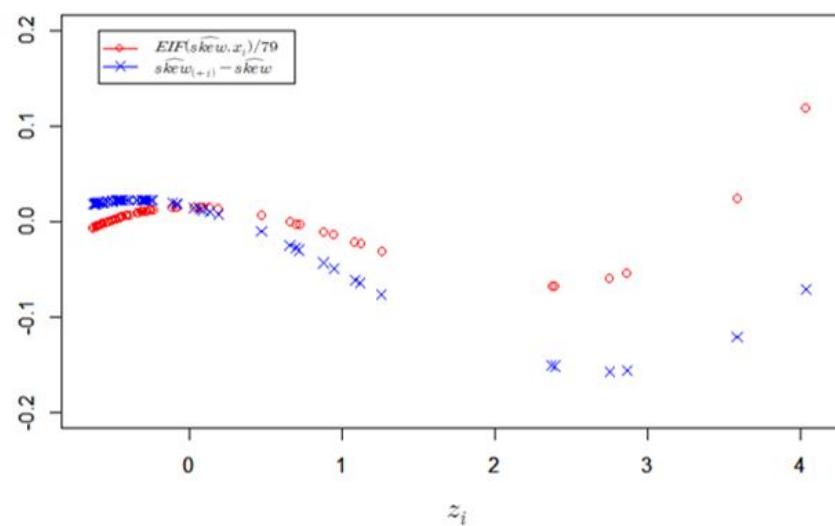
**Table 4.4** Actual difference, estimated difference, and difference between the two values, when adding one observation

$i$	$\bar{x}_{(+i)}$	$s_{(+i)}$	$[\sum_{i=1}^n \{(x_i - \bar{x}_{(+i)})^3\} + (x_i - \bar{x}_{(+i)})^3]/(n + 1)$	$\widehat{skew}_{(+i)}$	$\widehat{skew}_{(+i)} - \widehat{skew}$	$\frac{EIF(\widehat{skew}, x_i)}{79}$	$Diff_{(+i)}$
			( $W'$ )	( $X'$ )	( $X'/W'^3$ )	( $Y'$ )	( $Z'$ )
1	10271.6076	16050.764	9815678883175	2.37374	0.01797	-0.00641	0.02437
2	10271.8481	16050.6107	9815568975155	2.37378	0.01801	-0.00632	0.02433
3	10273.4304	16049.6092	9814839195735	2.37405	0.01828	-0.00577	0.02404
...	...	...	...	...	...	...	...
58	10384.2532	16010.0726	9743213568211	2.37423	0.01846	0.01459	0.00386
59	10407.038	16009.4446	9725946680918	2.37030	0.01453	0.01487	-0.00035
60	10415.3924	16009.8562	9719620162611	2.36857	0.01280	0.01470	-0.00190
...	...	...	...	...	...	...	...
76	10986.0253	16832.0064	10487542505524	2.19921	-0.15657	-0.05444	-0.10212
77	11133.1139	17280.5161	11529297735205	2.23425	-0.12152	0.02382	-0.14534
78	11225.0127	17604.2712	12462513470780	2.28429	-0.07148	0.11924	-0.19072

$x_{59} = 10860$ 일 때 두 값의 차이가 -0.00035로 0에 가까운 값을 가지며, 이를 중심으로 점점 차이가 커지는 것을 알 수 있다. 이 때,  $x_{59}$ 는 평균인 10401.2308과 가장 가까운 값으로, 추가된 관측값이 평균과 가까워질수록  $EIF(\widehat{skew}, x_i)/(n + 1)$ 과  $\widehat{skew}_{(+i)} - \widehat{skew}$ 의 값이 작아지고 이 두 값이 거의 비슷해지는 것을 알 수 있다. Table 4.4에 해당하는 값들을 그래프로 나타낸 결과가 Figure 4.4와 Figure 4.5와 같으며, 두 그래프의 개형이 하나의 관측값을 제외하는 경우의 그래프인 Figure 4.2, Figure 4.3과  $y = 0$ 에 대하여 대칭임을 알 수 있다.



**Figure 4.4** Actual difference and estimated difference according to  $x_i$  when adding one observation



**Figure 4.5** Actual difference and estimated difference according to  $z_i$  when adding one observation

## 5. 결론

본 연구에서는 데이터 분포의 비대칭성을 나타내는 척도인 비대칭도의 영향함수를 유도하기 위하여 먼저 평균,  $\sigma$ ,  $\sigma^2$ 에 이어  $\sigma^3$  및 3차 적률의 영향함수를 유도한 뒤 비대칭도의 영향함수를 다음과 같이 유도하였다.

$$IF(skew, x) = z^2 \left( z - \frac{3}{2} skew \right) + \frac{1}{2} skew.$$

섭동 (Perturbation)을 고려하여 반영한 비대칭도에 대한 경험적 영향함수로 추정한 추정차이와 실제 차이의 관계는 다음과 같으며,

$$\begin{aligned}\widehat{skew}_{(+i)} - \widehat{skew} &\simeq EIF(\widehat{skew}, x_i) \times \frac{1}{n+1} \\ &= \frac{1}{n+1} \left\{ z_i^2 \left( z_i - \frac{3}{2} \widehat{skew} \right) + \frac{1}{2} \widehat{skew} \right\}, \\ \widehat{skew}_{(-i)} - \widehat{skew} &\simeq - EIF(\widehat{skew}, x_i) \times \frac{1}{n-1} \\ &= - \frac{1}{n-1} \left\{ z_i^2 \left( z_i - \frac{3}{2} \widehat{skew} \right) + \frac{1}{2} \widehat{skew} \right\}.\end{aligned}$$

타당성 검증을 위해 비대칭분포를 가지는 데이터를 이용하여 한 개의 관측값  $x_i$ 를 제외한 경우와 추가한 경우의 실제 비대칭도의 차이와 비대칭도에 대한 경험적 영향함수로 추정한 추정차이의 관계를 살펴보았다. 관측값을 제외하는 경우와 추가하는 경우에서의 추정차이와 실제차이 간 그래프의 개형이  $y = 0$ 에 대하여 대칭이 있으며, 비대칭도에 대한 경험적 영향함수로 추정한 추정차이와 실제차이는 관측값이 평균과 가까울수록 서로 근접하거나 일치하는 것을 확인할 수 있었으나, 데이터의 양 끝의 값으로 갈수록 두 차이값이 커지는 한계점이 있음을 확인하였다.

## References

- Campbell, N. A. (1978). The influence function as an aid to outlier detection in discrimination analysis. *Applied Statistics*, **27**, 251-258.
- Cook, R. D. (1977). Detection of influential observation in linear regression. *Technometrics*, **19**, 15-18.
- Cook, R. D. and Weisberg, S. (1980). Characterization of empirical influence function for detection influential cases in regression. *Technometrics*, **22**, 495-508.
- Critchley, F. (1985). Influence in principal components analysis. *Biometrika*, **72**, 627-636.
- Hampel, F. R. (1974). The influence curve and its role in robust estimation. *Journal of the American Statistical Association*, **69**, 383-393.
- Kang, H. S. and Kim, H. (2020). A study on the difference and calibration of empirical influence function and sample influence function. *The Korean Journal of Applied Statistics*, **33**, 527-540.
- Kang, H. S. and Kim, H. (2021). Extending the calibration between empirical influence functions and sample influence function to t-statistic. *The Korean Journal of Applied Statistics*, **34**, 889-904.
- Kang, H. S., Jeh, A. L. and Kim, H. (2022). Comparison of the performances in approximating sample influence function using empirical influence function. *Journal of the Korean Data & Information Science Society*, **33**, 209-222.
- Kim, M. J. and Kim, H. (2017). Derivation and verifications of Influence function on parameter proposed by Ghosh and Kim. *The Korean Journal of Applied Statistics*, **30**, 529-538.
- Kim, H. (1992). Measures of influence in correspondence analysis. *Journal of Statistical Computation and Simulation*, **40**, 201-217.

- Kim, H. (1994). Influence functions in multiple correspondence analysis. *The Korean Journal of Applied Statistics*, **7**, 69-74.
- Kim, H. (1998). A study on cell influence to chi-square statistic in contingency tables. *The Korean Communications in Statistics*, **5**, 35-42.
- Lee, H. S. and Kim, H. (1996). Influence functions on  $\chi^2$  statistic in contingency tables. *The Korean Communications in Statistics*, **3**, 69-76.
- Lee, Y. H., Yim, M. H., Ko, M. M. and Kim, H. (2024). Derivation of the influence function on the parameter  $\sigma^k$  and its application. *Journal of the Korean Data & Information Science Society*, **35**, 195-205.
- Kim, K. H. and Kim, H. (2005). Influence of an observation on the  $t$ -statistic. *The Korean Communications in Statistics*, **12**, 453-462.
- Kim, S. J. and Kim, H. (2019). A study on the performance of the influence function on the  $t$  statistic depending on population distributions in big data sets. *Journal of the Korean Data & Information Science Society*, **30**, 570-585.
- Lee, Y. H. and Kim, H. (2008). Influence function on the coefficient of variation. *Communications for Statistical Applications and Methods*, **15**, 509-516.
- Park, S., Kang, H., Kim, S. and Kim, H. (2019). A study on the location of the observation which has the least effect on the  $t$ -statistic. *Journal of the Korean Data & Information Science Society*, **30**, 1221-1232.
- Radhkrishnan, R. and Kshirsagar, A. M. (1981). Influence functions for certain parameters in multi-variate analysis. *Communications in Statistics*, **10**, 515-529.
- Cook, R. D. and Weisberg, S. (1982). *Residual and influence in regression*, Chapman ad Hall, New York.
- Kim, J. H. (2021). *A study on induction about the empirical influence function of the 3rd central moment and validity verification*, Master's Thesis, Chungnam National University, Deajon.

## Derivation of influence functions on skewness and analysis of its role<sup>†</sup>

A La Jeh<sup>1</sup> · Yun Hee Lee<sup>2</sup> · Honggie Kim<sup>3</sup>

<sup>1</sup>Statistics Korea

<sup>2,3</sup>Department of Information and Statistics, Chungnam National University

Received 24 April 2024, revised 23 May 2024, accepted 30 May 2024

### Abstract

In this study, the influence function on skewness is derived by using the influence functions on the third central moment and the third power of the standard deviation. Based on the empirical distribution obtained from the sample, we examine the relationship between the actual difference in skewness, when excluding or adding one observation, and the estimated difference. In addition, in order to verify the validity of the influence function on skewness, a close computation is conducted using data with an asymmetric distribution obtained from the National Statistical Office survey data and the results are examined. This study is meaningful in that it examines the role and meaning of the influence function on skewness.

**Keywords:** Actual difference, empirical influence function, estimated difference, influence function, outlier, skewness.

<sup>†</sup> This article is a condensed form of the first author's doctoral thesis from Chungnam National University.  
This work was supported by the National Research Foundation of Korea Grant funded by the Korean Government(NRF-2022M3J6A1084843).

<sup>1</sup> Officer, Statistics Korea, Daejeon 35208, Korea.

<sup>2</sup> Ph.D. assistant, Department of Information and Statistics, Chungnam National University, Daejeon 34134, Korea.

<sup>3</sup> Corresponding author: Professor, Department of Information and Statistics, Chungnam National University, Daejeon 34134, Korea. E-mail : honggiekim@cnu.ac.kr