The Korean Journal of Applied Statistics
2021, Vol. 34, No. 6, 889–904

DOI: https://doi.org/10.5351/KJAS.2021.34.6.889
Print ISSN1225-066X / Online ISSN 2383-5818

Extending the calibration between empirical influence function and sample influence function to t-statistic

Hyunseok Kang[a], Honggie Kim[1,b]

[a]Ministry of Education; [b]Department of Information and Statistics, Chungnam National University

Abstract

This study is a follow-up study of Kang and Kim (2020). In this study, we derive the sample influence functions of the t-statistic which were not directly derived in previous researches. Throughout these results, we both mathematically examine the relationship between the empirical influence function and the sample influence function, and consider a method to approximate the sample influence function by the empirical influence function. Also, the validity of the relationship between an approximated sample influence function and the empirical influence function is verified by a simulation of a random sample of size 300 from normal distribution. As a result of the simulation, the relationship between the sample influence function which is derived from the t-statistic and the empirical influence function, and the method of approximating the sample influence function through the empirical influence function were verified. This research has significance in proposing both a method which reduces errors in approximation of the empirical influence function and an effective and practical method that evolves from previous research which approximates the sample influence function directly through the empirical influence function by constant revision.

Keywords: influence function, outlier, t-statistic, empirical influence function, sample influence function

1. 서론

데이터 분석에서 통계량, 통계적 모형 등에 영향을 미치는 이상치(outlier)의 선별과 이에 대한 적절한 처리는 분석 결과의 신뢰를 높이기 위해 매우 중요한 과정이라고 할 수 있다. 이러한 이상치를 판별하기 위한 도구로써 영향함수(influence function)가 활발하게 활용되고 있다. Hampel (1974)은 영향함수를 활용한 이상치 판별 방법을 가장 먼저 소개하였으며, 대부분의 통계량은 영향함수를 이용해 이상치 판별이 가능함을 보였다. 이후, Campbell (1978)은 판별분석에서 영향함수를 사용하여 이상치를 발견하였고, Radhakrishnan과 Kshirsagar (1981)은 다변량 분석에서 여러 가지 모수에 대해 이론적으로 영향함수를 유도해 냈다. Cook (1977)은 회귀분석에서의 영향력있는 관측값에 대해 연구하였으며, Cook과 Weisberg (1980, 1982)는 회귀 분석에서 회귀진단방법으로서 영향함수를 적용하였다. Critchley (1985)는 주성분분석에서 영향함수를 적용하여 영향력있는 관측치를 찾아내는 방법으로 활용하였다. Kim (1992)은 대응분석(correspondence analysis)에서의 영향함수를 유도하였으며, Kim과 Lee (1996), Kim (1998), Lee와 Kim (2003)은 χ^2통계량에 대한 영향함수, Kim과 Kim (2005)은 t통계량에 대한 영향함수, Lee와 Kim (2008)의 변이계수에 대한 영향함수의 유도에 관한 연구를 진행하였다. 그리고 Kim과 Kim (2019)의 빅데이터에서 모분포의 형태에 따른 t통계량에 대한 영향함수의 성능에 대한 연구와 Park과 Kim (2019)의 통계량에 가장 작은 영향을 미치는 관측값의 위치에 대한 연구에

[1] Corresponding author: Chungnam National University, 99 Daehak-ro, Yuseong-gu, Daejeon 34134, Korea. E-mail: honggiekim@cnu.ac.kr

Published 31 December 2021 / journal homepage: http://kjas.or.kr

이르기까지 다양한 통계량에 대한 영향함수의 유도와 이상치 선별에의 활용 가능성에 대한 연구가 현재까지 활발하게 진행되고 있다.

본 연구에서는 Kang과 Kim (2020)의 연구 방법론을 기반으로 t통계량에 대한 표본영향함수를 직접 유도하고, 경험적 영향함수와 표본영향함수의 차이를 보정하여 각 관측값이 t통계량에 미치는 영향의 정도를 근사적으로 추론할 수 있는 방안을 제안한다. 특히, t통계량은 사회과학 연구에서 활용도가 매우 높으므로 이상치의 선별과 제거 순위를 정하는 데에 의미있는 도움이 될 수 있다. 2장에서는 영향함수의 정의와 평균, 분산, 표준편차, 통계량의 경험적 영향함수 유도와 함께 표본영향함수를 정의한다. 3장에서는 통계량에 대한 표본영향함수를 유도하고, 이를 바탕으로 경험적 영향함수와의 관계를 이론적으로 살펴본다. 4장에서는 모의로 생성한 데이터를 중심으로 3장에서 다룬 추론의 타당성을 검증한 뒤, 5장에서는 실제 자료 분석 과정에 이를 적용한 예를 다룬다. 6장에서는 본 연구의 결론을 제시한다.

2. 영향함수

2.1. 영향함수의 정의

분포함수 F와 실수 c에 대해 $T(F) = C$의 형태로 정의되는 함수 T를 범함수(real-valued function)라 하고, 실수 x에서 확률이 1인 분포함수,

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

를 퇴화분포함수(degenerated distribution function)라고 한다. 분포함수 F에 임의의 관측값 x를 추가할 경우, 분포함수 F와 퇴화분포함수 δ_x의 혼합분포함수 F_ϵ은 다음과 같이 나타낼 수 있다.

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

이때, F_ϵ을 F의 섭동(perturbation)이라고 하며, Hampel (1974)은 관측값 x가 추가됨으로써 $T(F)$에 미치는 영향을 나타내기 위한 방법으로 섭동 F_ϵ를 이용해 영향함수 IF(T, x)를 다음과 같이 정의하였다.

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

그리고 로피탈의 정리를 이용해 식(2.3)은 다음과 같이 정리할 수 있다.

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

2.2. 다양한 통계량에 대한 영향함수 유도

분포함수 $F(t)$의 확률밀도함수를 $f(t)$라 하면 $\partial F(t)/\partial t = f(t)$가 성립하고, 분포함수에 대한 평균 μ와 분산 σ^2을 각각 함숫값으로 갖는 두 범함수 T_1, T_2는 다음과 같이 나타낼 수 있다.

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

이를 이용하여 평균 μ와 분산 σ^2에 대한 영향함수는 다음과 같이 각각 유도할 수 있다 (Hampel, 1974).

$$\begin{aligned} &\text{IF}\,(T_1, x) = \text{IF}\,(\mu, x) = x - \mu, \\ &\text{IF}\,(T_2, x) = \text{IF}\,\left(\sigma^2, x\right) = (x-\mu)^2 - \sigma^2 \end{aligned} \tag{2.6}$$

특히, 범함수 T_3는 $T_3(F) = \sqrt{T_2(F)} = \sigma$를 만족시키도록 정의하면 다음이 성립한다.

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

표준편차 σ에 대한 영향함수는 $\text{IF}(T_3, x) = \text{IF}(\sigma, x) = (1/2\sigma)\cdot\{(x-\mu)^2 - \sigma^2\}$로 유도할 수 있다.

한편, 식(2.4)에서 영향함수는 일차미분계수 형태로 표현되고, 선형 결합(linear combination)에 대해서는 그 성질을 유지하므로,

$$\text{IF}\left(\frac{\mu-\mu_0}{\sigma/\sqrt{n}}, x\right) = \sqrt{n}\left[\text{IF}\left(\frac{\mu}{\sigma}, x\right) - \mu_0 \cdot \text{IF}\left(\frac{1}{\sigma}, x\right)\right], \tag{2.8}$$

이 성립하고, 식(2.6), 식(2.7), 식(2.8)을 이용하여 다음과 같이 정리할 수 있다.

$$\text{IF}\left(\frac{\mu}{\sigma}, x\right) = \left[\frac{\partial\{T_1(F_\epsilon)/T_3(F_\epsilon)\}}{\partial\epsilon}\right]_{\epsilon=0} = \frac{x-\mu}{\sigma} - \frac{\mu}{2\sigma^3}\left\{(x-\mu)^2 - \sigma^2\right\}, \tag{2.9}$$

$$\text{IF}\left(\frac{1}{\sigma}, x\right) = \left[\frac{-\{\partial T_3(F_\epsilon)\}/\partial\epsilon}{\{T_3(F_\epsilon)\}^2}\right]_{\epsilon=0} = -\frac{1}{\sigma^2}\cdot\text{IF}\,(\sigma, x), \tag{2.10}$$

$$\text{IF}\left(\frac{\mu-\mu_0}{\sigma/\sqrt{n}}, x\right) = \sqrt{n}\left[\frac{x-\mu}{\sigma} - \frac{(\mu-\mu_0)\left\{(x-\mu)^2-\sigma^2\right\}}{2\sigma^3}\right] \tag{2.11}$$

2.3. 경험적 영향함수와 표본영향함수

모집단의 분포함수 F에 대해 $T(F)$의 영향함수 $\text{IF}(T, x)$가 정의되는 경우, 표본분포함수 $\hat{F}$에서 얻은 범함수 $T(\hat{F})$의 영향함수는 모분포의 통계량을 추정하는 추정량을 $\text{IF}(T, x)$에 대입하여 얻는다. 이를 경험적 영향함수(empirical influence function, EIF)라 하며, 식(2.11)에서 모분포가 갖는 평균 μ, 표준편차 σ를 추정하는 추정량을 각각 표본평균 $\bar{x}$, 표본표준편차 s로 대입하면 t통계량의 경험적 영향함수를 다음과 같이 얻을 수 있다.

$$\text{EIF}\,(t, x) = \text{EIF}\left(\frac{\bar{x}-\mu_0}{s/\sqrt{n}}, x\right) = \sqrt{n}\left[\frac{x-\bar{x}}{s} - \frac{(\bar{x}-\mu_0)\left\{(x-\bar{x})^2 - s^2\right\}}{2s^3}\right]. \tag{2.12}$$

한편, 표본의 크기가 n이고, 표본평균이 $\bar{x}$, 표본분포함수가 $\hat{F}$인 표본에서 i번째 관측값 x_i를 제거한 표본의 크기 $n-1$인 표본의 표본평균과 표본분포함수를 각각 $\bar{x}_i$,$\hat{F}_i$ 라 하자. 이때, i번째 관측치를 제거하여 발생하는 범함수의 함숫값 차이에 섭동 ϵ을 $-1/(n-1)$로 고려하여 다시 얻을 수 있는 영향함수를 표본영향함수(sample influence function, SIF)라 한다. Cook과 Weisberg (1982)에 따르면 이는 하나의 관측치가 표본의 통계량에 미치는 영향을 측정하는 도구가 된다.

즉, 범함수 T에 대하여 표본영향함수 $\text{SIF}(T, x_i)$는 $T(\hat{F}_{(i)}) - T(\hat{F}) = \{-1/(n-1)\}\cdot\text{SIF}\,(T, x_i)$ 와 같이 정의한다. 표본의 크기가 n인 표본에서 i번째 관측값 x_i를 제거한 후의 t통계량을 $t_{(i)}$라 하면, 표본영향함수 $\text{SIF}(t, x_i)$는 다음과 같이 표현할 수 있다.

$$\text{SIF}\,(t, x_i) = \text{SIF}\left(\frac{\bar{x}-\mu}{s/\sqrt{n}}, x_i\right) = -(n-1)\{t_{(i)} - t\} \tag{2.13}$$

Kang과 Kim (2020)은 평균, 분산, 표준편차에 대한 표본영향함수를 유도하였다. 그리고 경험적 영향함수와 표본영향함수의 차이를 살펴 표본영향함수에 대해 경험적 영향함수를 활용한 근사적 추론시 엄밀성을 높일 수 있는 보정 방안을 제안하였다. 본 연구에서는 Kang과 Kim (2020)의 연구 방법론을 t통계량까지로 확장시켜 적용해 보기 위해 3장에서 t통계량에 대한 표본영향함수를 유도하기로 한다.

Table 3.1: Summary of a random sample of size 100 from $N(0, 1)$

Min	Mean	Max	Variance	t-statistic*
−2.65351	0.04079	2.24582	0.98289	0.41146

Note: t-statistic when the null hypothesis $H_0 : \mu_0 = 0$ is true.

Table 3.2: Summary of the sample shifted to be $\bar{x} = 2$

Min	Mean	Max	Variance	t-statistic*
−0.6943	2.0000	4.2050	0.98289	20.1733

Note: t-statistic when the null hypothesis $H_0 : \mu_0 = 0$ is true.

3. 표본영향함수의 유도

3.1. t통계량에 대한 표본영향함수 유도

표본의 크기가 n인 표본에서 i번째 관측값 x_i를 제거한 후의 표본평균 $\bar{x}_{(i)}$, 표본분산 $s^2_{(i)}$, t 통계량 $t_{(i)}$는 각각 식 (3.1)과 같이 나타낼 수 있다.

$$\begin{aligned}
\bar{x}_{(i)} &= \frac{1}{n-1}\left(-x_i + \sum_{k=1}^{n} x_k\right),\\
s^2_{(i)} &= \frac{1}{n-2}\left[-\{x_i - \bar{x}_{(i)}\}^2 + \sum_{k=1}^{n}\{x_k - \bar{x}_{(i)}\}^2\right],\\
t_{(i)} &= \frac{\bar{x}_{(i)} - \mu_0}{s_{(i)}/\sqrt{n-1}}.
\end{aligned} \tag{3.1}$$

또한, 식(2.12)의 표현을 정리하면 를 다음과 같이 나타낼 수 있다.

$$\begin{aligned}
\mathrm{EIF}\,(t, x_i) &= \sqrt{n}\left[\frac{x_i - \bar{x}}{s} - \frac{(\bar{x} - \mu_0)\left\{(x_i - \bar{x})^2 - s^2\right\}}{2s^3}\right],\\
&= \frac{\sqrt{n}}{s}(x_i - \bar{x}) - \frac{1}{2s^2}\cdot\frac{\bar{x} - \mu_0}{s/\sqrt{n}}\cdot(x_i - \bar{x})^2 + \frac{1}{2}\cdot\frac{\bar{x} - \mu_0}{s/\sqrt{n}},\\
&= -\frac{t}{2s^2}(x_i - \bar{x})^2 + \frac{\sqrt{n}}{s}(x_i - \bar{x}) + \frac{t}{2}.
\end{aligned} \tag{3.2}$$

식(3.2)에서 $y = \mathrm{EIF}\,(t, x_i)$는 x_i에 대한 이차함수로 $\partial\mathrm{EIF}\,(t, x_i)/\partial x_i = (-t/s^2)\cdot(x_i - \bar{x}) + \sqrt{n}/s = 0$일 때 극댓값을 갖는다. $y = \mathrm{EIF}\,(t, x_i)$는 $x_i = \bar{x} + s\sqrt{n}/t$에서 이차함수의 꼭짓점을 갖고, 이때의 극댓값은 $n/2t + t/2$임을 알 수 있다. 함수 $y = \mathrm{EIF}\,(t, x_i)$가 갖는 이차함수의 성질을 경험적으로 확인해 보기 위해 정규분포 $N(0, 1)$에서 크기가 100인 표본을 임의추출하였고, 귀무가설 $H_0 : \mu_0 = 0$이 참이라는 가정 하에 t통계량을 구했다. 이렇게 구한 표본의 기술 통계는 Table 3.1과 같으며, 생성된 100개 데이터는 Kolmogorov-Smirnov 검정을 통해 정규성을 잘 만족시키는 것으로 확인되었다.

이때 표본의 표본평균인 0.04079를 모든 데이터에서 뺀 뒤, 2만큼을 더하여 $\bar{x} = 2$를 만족시키도록 100개의 데이터 값을 보정(calibration)하였으며, 보정한 뒤의 표본의 기술 통계는 Table 3.2와 같다. 여기서 t통계량은 귀무가설 $H_0 : \mu_0 = 0$이 참이라는 가정 하에 구했다.

$\bar{x} = 2, s = 0.99, n = 100, t = 20.17$ 일 때, 실수 x_i에 대한 함수 $y = \mathrm{EIF}\,(t, x_i)$의 그래프는 Figure 3.1과 같다. $x_i = 2 + 0.99 \times 10/20.17 \approx 2.491$에서 함숫값이 $100/2 \times 20.17 + 20.17/2 \approx 12.56$인 꼭짓점을 갖는 이차함수

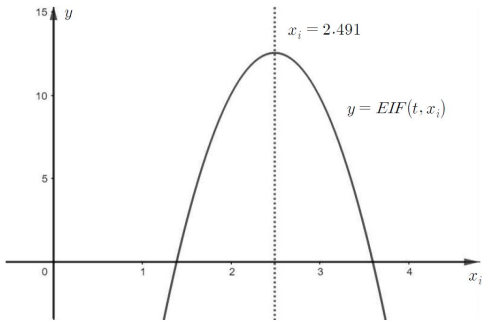

Figure 3.1: *Graph of EIF*(t, x_i) *when* $\bar{x} = 2$, $s = 0.99$, $n = 100$, $t = 20.17$.

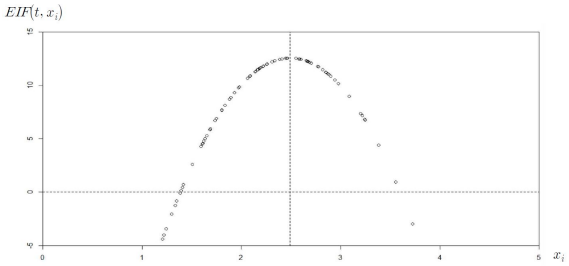

Figure 3.2: *Graph of EIF*(t, x_i) *for a sample of size* 100.

임을 확인할 수 있으며 Table 3.2의 크기가 100인 표본에서 얻은 EIF (t, x_i)의 값들은 Figure 3.1의 그래프 위에 위치하고 있음을 Figure 3.2로 확인할 수 있다.

한편, t통계량에 대한 표본영향함수 유도를 위해 식(3.1)에서 제시한 $t_{(i)}$를 정리하면 다음과 같이 나타낼 수 있다.

$$t_{(i)} = \frac{\sqrt{n-1}\,(\bar{x}_{(i)} - \mu_0)}{s_{(i)}} = \frac{\sqrt{n-1}\left(\frac{-x_i + \sum_{k=1}^{n} x_k}{n-1} - \mu_0\right)}{\sqrt{\frac{-\{x_i - \bar{x}_{(i)}\}^2 + \sum_{k=1}^{n}\{x_k - \bar{x}_{(i)}\}^2}{n-2}}} \tag{3.3}$$

식(3.1)과 식(3.3)를 바탕으로 정리하면 다음의 식이 성립함을 알 수 있다.

$$t_{(i)} = \frac{-\frac{\sqrt{(n-1)(n-2)}}{n-1}(x_i - \bar{x}) + st\sqrt{\frac{(n-1)(n-2)}{n}}}{\sqrt{-\frac{n}{n-1}(x_i - \bar{x})^2 + (n-1)s^2}} \tag{3.4}$$

식(3.4)을 이용해 t 통계량에 대한 SIF(t, x_i)를 유도하면 다음과 같다.

$$\begin{aligned}
\text{SIF}\,(t, x_i) &= -(n-1)(t_{(i)} - t), \\
&= -(n-1)\left\{\frac{-\frac{\sqrt{(n-1)(n-2)}}{n-1}(x_i - \bar{x}) + st\sqrt{\frac{(n-1)(n-2)}{n}}}{\sqrt{-\frac{n}{n-1}(x_i - \bar{x})^2 + (n-1)s^2}} - t\right\}, \\
&= \frac{\sqrt{(n-1)(n-2)}(x_i - \bar{x}) - st(n-1)\sqrt{\frac{(n-1)(n-2)}{n}}}{\sqrt{-\frac{n}{n-1}(x_i - \bar{x})^2 + (n-1)s^2}} + (n-1)t.
\end{aligned} \tag{3.5}$$

식(3.5)에서 표본영향함수 SIF(t, x_i)는 x_i에 대한 함수이고 분모가 0이면 함숫값이 정의되지 않는다. $\{-n/(n-1)\} \cdot (x_i - \bar{x})^2 + (n-1)s^2 = 0$을 만족시키는 $x_i = \bar{x} \pm (n-1)s/\sqrt{n}$에서 함수 SIF$(t, x_i)$는 서로 다른 두 점근선을 가지며, $\bar{x} - (n-1)s/\sqrt{n} < x_i < \bar{x} + (n-1)s/\sqrt{n}$에서만 함수가 정의된다. 한편, 표본영향함수 SIF(t, x_i)의 극댓값을 구해보기로 한다.

$$\begin{aligned}\frac{\partial \text{SIF}\,(t, x_i)}{\partial x_i} &= \sqrt{(n-1)(n-2)}\left\{(n-1)s^2 - \frac{n}{n-1}(x_i - \bar{x})^2\right\}^{-\frac{3}{2}} \\ &\quad \times \left\{-\frac{n}{n-1}(x_i - \bar{x})^2 + (n-1)s^2 - st\sqrt{n}(x_i - \bar{x}) + \frac{n}{n-1}(x_i - \bar{x})^2\right\} \end{aligned} \tag{3.6}$$

이고, $\partial \text{SIF}\,(t, x_i)/\partial x_i = 0$일 때, SIF$(t, x_i)$는 극값을 가지므로 분자가 0일 때 극값을 갖는다.

$$\begin{aligned}\text{분자} &= \sqrt{(n-1)(n-2)} \\ &\quad \times \left\{-\frac{n}{n-1}(x_i - \bar{x})^2 + (n-1)s^2 - st\sqrt{n}(x_i - \bar{x}) + \frac{n}{n-1}(x_i - \bar{x})^2\right\} \\ &= \sqrt{(n-1)(n-2)}\left\{(n-1)s^2 - st\sqrt{n}(x_i - \bar{x})\right\}, \end{aligned} \tag{3.7}$$

이므로 $(n-1)s^2 - st\sqrt{n}(x_i - \bar{x}) = 0$일 때인 $x_i = \bar{x} + s(n-1)/(t\sqrt{n})$에서 SIF$(t, x_i)$는 극값을 유일하게 가지며, 부호변화 확인을 통해 이 값이 극댓값임을 확인할 수 있다.

3.2. t통계량에 대한 경험적 영향함수와 표본영향함수의 차이

t 통계량에 대한 경험적 영향함수와 표본영향함수를 비교해 보면 경험적 영향함수 EIF(t, x_i)는 $x_i = \bar{x} + s\sqrt{n}/t$에서 극댓값을 갖지만 표본영향함수 SIF$(t, x_i)$는 $x_i = \bar{x} + s(n-1)/(t\sqrt{n})$에서 극댓값을 갖는 차이가 있다. $\bar{x} = 2, s = 0.99, n = 100, t = 20.17$ 일 때, 연속적인 x_i에 대한 일반적인 함수 SIF(t, x_i)의 그래프는 Figure 3.3과 같다. 이차함수 EIF(t, x_i)와 유사한 경향을 갖기는 하지만, $x_i = 2 + 0.99 \times 99/10 \approx 11.801$와 $x_i = 2 - 0.99 \times 99/10 \approx -7.801$에서 서로 다른 두 점근선을 갖는 그래프로 $y = \text{EIF}\,(t, x_i)$와 $y = \text{SIF}\,(t, x_i)$는 다른 함수라고 할 수 있다. 특히, $x_i = 2 + 0.99 \times 99/(20.17 \times 10) \approx 2.486$에서 극댓값을 갖는 것도 확인할 수 있다. 크기가 100인 표본에서 얻은 SIF(t, x_i) 의 값들은 Figure 3.3의 그래프 위에 위치하고 있음을 Figure 3.4에서 기하적으로도 확인할 수 있으나, Figure 3.3과 Figure 3.4에서 제시한 그래프가 서로 달라 보이는 이유는 사용된 100개의 데이터 값들이 정의역 −1과 5사이에 속하는 값들로 Figure 3.4가 Figure 3.3의 중앙 일부만을 반영하기 때문이다. Figure 3.3은 함수 SIF(t, x_i)의 그래프가 갖는 점근선의 존재를 표현해 중앙 부분에서는 이차함수와 비슷하지만 전체적으로는 이차함수가 아닌 점을 나타내고 있다.

지금까지 살펴본 바와 같이 경험적 영향함수 EIF(t, x_i)와 표본영향함수 SIF(t, x_i)는 이차함수의 그래프와 유사한 개형을 갖지만 $x_i = \bar{x} + s\sqrt{n}/t$와 $x_i = \bar{x} + s(n-1)/(t\sqrt{n})$에서 각각 극댓값을 갖는 차이를 가지고 있다.

3.3. t통계량의 경험적 영향함수를 이용한 표본영향함수의 근사적 추론

경험적 영향함수 EIF(t, x_i)를 이용해 표본영향함수 SIF(t, x_i)를 유도 및 추론하기 위해서는 두 함수의 차이를 보정(calibration)하는 방법에 대해 생각해 볼 필요성이 있다. 경험적 영향함수 EIF(t, x_i)와 표본영향함수 SIF(t, x_i)는 전혀 다른 함수이지만 그 경향성이 매우 유사하므로 극댓값이 일치하도록 EIF(t, x_i)를 이동해 SIF(t, x_i)를 설명할 수 있는 방법을 살펴보기로 한다. 두 함수의 정의역이 매우 상이하므로 모든x_i 값에 대하여 두 함수를 일치시킬 수는 없지만, 두 함수의 경향의 유사성을 이용해 극댓값을 일치시키는 방법은 두 함수의 차이를 줄이는 하나의 방법론이 될 수 있다. 먼저, 두 함수는 앞서 살펴본 바와 같이 극댓값을 갖는 x_i의 값이 다르다. 이

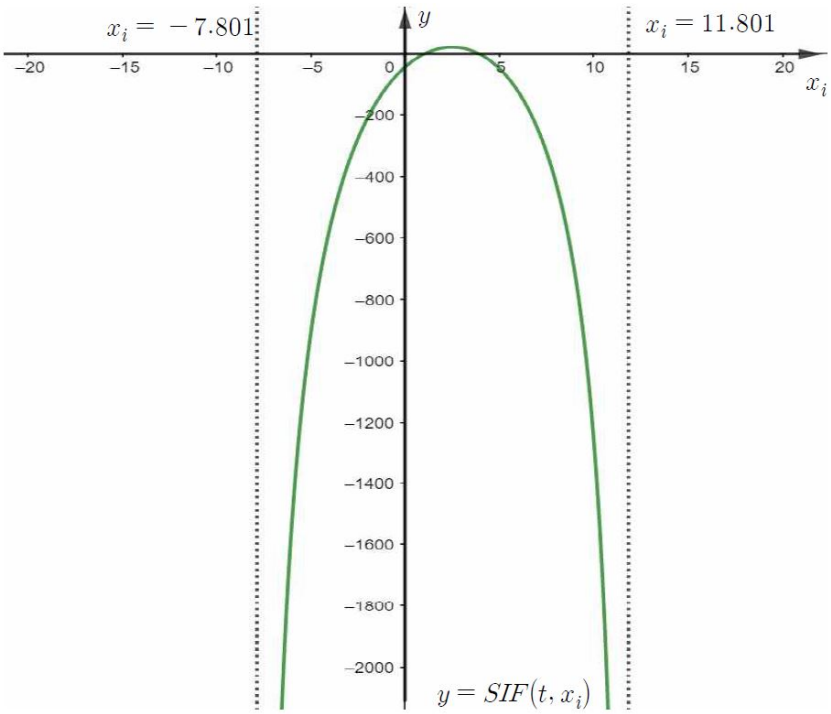

Figure 3.3: *Graph of SIF*(t, x_i) *when* $\bar{x} = 2$, $s = 0.99$, $n = 100$, $t = 20.17$.

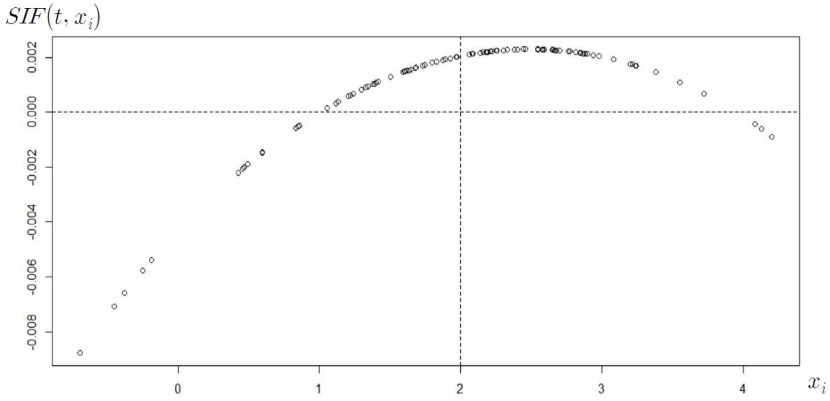

Figure 3.4: *Graph of SIF*(t, x_i) *for a sample of size* 100.

차이를 정리하면 다음과 같다.

$$\left(\bar{x} + \frac{s\sqrt{n}}{t}\right) - \left(\bar{x} + \frac{s(n-1)}{t\sqrt{n}}\right) = \frac{s\sqrt{n}}{t} - \frac{s(n-1)}{t\sqrt{n}} = \frac{s}{t\sqrt{n}}. \tag{3.8}$$

한편, $x_i = \bar{x} + s(n-1)/(t\sqrt{n})$에서 표본영향함수 SIF$(t, x_i)$의 극댓값은 $x_i - \bar{x} = s(n-1)/(t\sqrt{n})$임을 이용해

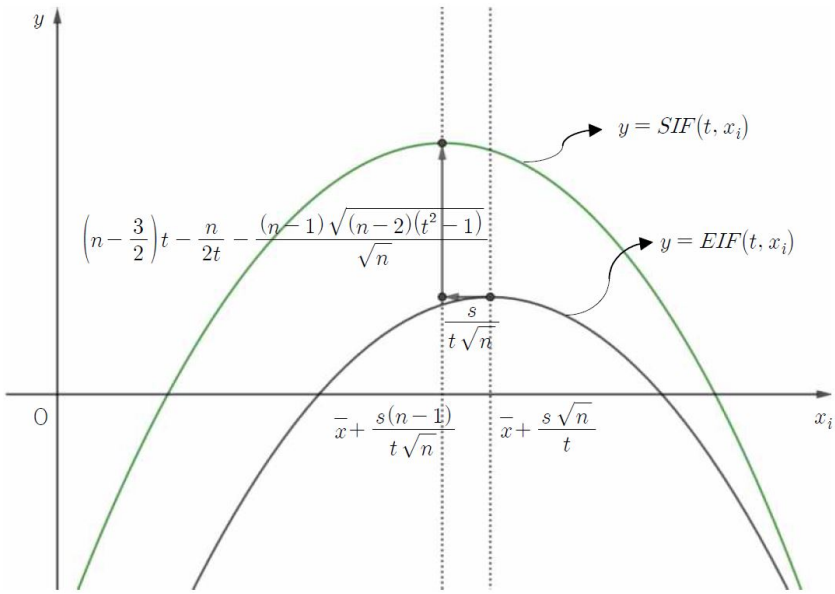

Figure 3.5: *The difference between* $y = EIF(t, x_i)$ *and* $y = SIF(t, x_i)$.

다음과 같이 구할 수 있다.

$$\begin{aligned}
&\mathrm{SIF}\left(t, x_i = \bar{x} + \frac{s(n-1)}{t\sqrt{n}}\right) \\
&= \frac{\sqrt{(n-1)(n-2)} \cdot \frac{s(n-1)}{t\sqrt{n}} - st(n-1)\sqrt{\frac{(n-1)(n-2)}{n}}}{\sqrt{-\frac{n}{n-1} \cdot \frac{s^2(n-1)^2}{nt^2} + (n-1)s^2}} + (n-1)t, \\
&= (n-1)t - \frac{(n-1)\sqrt{(n-2)(t^2-1)}}{\sqrt{n}}.
\end{aligned} \tag{3.9}$$

경험적 영향함수 $\mathrm{EIF}(t, x_i)$의 극댓값은 $\mathrm{EIF}(t, x_i = \bar{x} + s\sqrt{n}/t) = n/2t + t/2$이었으므로, 이를 이용해 표본영향함수 $\mathrm{SIF}(t, x_i)$와 경험적 영향함수 $\mathrm{EIF}(t, x_i)$의 극댓값의 차이를 구해보면 다음과 같다.

$$\begin{aligned}
&\mathrm{SIF}\left(t, x_i = \bar{x} + \frac{s(n-1)}{t\sqrt{n}}\right) - \mathrm{EIF}\left(t, x_i = \bar{x} + \frac{s\sqrt{n}}{t}\right) \\
&= (n-1)t - \frac{(n-1)\sqrt{(n-2)(t^2-1)}}{\sqrt{n}} - \frac{n}{2t} - \frac{t}{2}, \\
&= \left(n - \frac{3}{2}\right)t - \frac{n}{2t} - \frac{(n-1)\sqrt{(n-2)(t^2-1)}}{\sqrt{n}}.
\end{aligned} \tag{3.10}$$

즉, 식(3.8)과 식(3.10)을 이용해 경험적 영향함수 $\mathrm{EIF}(t, x_i)$를 축으로 $-s/(t\sqrt{n})$만큼 평행이동하고 $\mathrm{EIF}(t, x_i)$의 극댓값과 $\mathrm{SIF}(t, x_i)$의 극댓값의 차이만큼 상수항을 더하면 $\mathrm{EIF}(t, x_i)$로 $\mathrm{SIF}(t, x_i)$를 근사해 낼 수 있다. 이를 좌표평면에 그래프로 도식화하면 Figure 3.5와 같다.

한편, $\mathrm{EIF}(t, x_i)$를 x_i축으로 $-s/(t\sqrt{n})$만큼 평행이동한 식은 식(3.8)을 이용해 정리하면,

$$\mathrm{EIF}\left(t, x_i + \frac{s}{t\sqrt{n}}\right) = -\frac{t}{2s^2}\left\{\left(x_i + \frac{s}{t\sqrt{n}}\right) - \bar{x}\right\}^2 + \frac{\sqrt{n}}{s}\left\{\left(x_i + \frac{s}{t\sqrt{n}}\right) - \bar{x}\right\} + \frac{t}{2} \tag{3.11}$$

Table 4.1: Summary of a random sample of size 300 from $N(0, 1)$

Min	Mean	Max	Variance	t-statistic*
−2.30923	0.05943	2.78710	0.91174	1.07806

Note: t-statistic when the null hypothesis $H_0 : \mu_0 = 0$ is true.

Table 4.2: Summary of the sample shifted to be $\bar{x} = 4$

Min	Mean	Max	Variance	t-statistic*
1.631	4.000	6.728	0.912	72.558

Note: t-statistic when the null hypothesis $H_0 : \mu_0 = 0$ is true.

이고, 따라서 표본영향함수 SIF(t, x_i) 는 다음과 같이 근사시킬 수 있다.

$$\text{SIF}\,(t, x_i) \approx \text{EIF}\left(t, x_i + \frac{s}{t\sqrt{n}}\right) + \left(n - \frac{3}{2}\right) - \frac{n}{2t} - \frac{(n-1)\sqrt{(n-2)(t^2-1)}}{n}. \tag{3.12}$$

결론적으로 식(3.12)를 바탕으로 , $T(F) = (\mu - \mu_0)/(\sigma/\sqrt{n})$, $T(\hat{F}) = (\bar{x} - \mu_0)/(s/\sqrt{n})$이라 하면 t통계량에 대한 $T(\hat{F}_{(i)}) - T(\hat{F}) = t_{(i)} - t$는 EIF$(t, x_i)$ 를 이용해 다음과 같이 근사하여 예측 가능하다.

$$\begin{aligned} t_{(i)} - t &= -\frac{1}{n-1}\text{SIF}\,(t, x_i), \\ &\approx -\frac{1}{n-1}\text{EIF}\left(t, x_i + \frac{s}{t\sqrt{n}}\right) + \frac{t^2+n}{2t(n-1)} + \frac{\sqrt{(n-2)(t^2-1)}}{\sqrt{n}} - t. \end{aligned} \tag{3.13}$$

t 통계량에 대한 표본영향함수를 경험적 영향함수로 근사시켜서 추론하기 위해서는 EIF(t, x_i) 의 평행이동과 상수항의 보정이 필요하다는 것을 알 수 있다. 이러한 추론으로 EIF(t, x_i)를 이용해 SIF(t, x_i)를 근사시키는 방법은 SIF(t, x_i)가 $\bar{x} - (n-1)s/\sqrt{n} < x_i < \bar{x} + (n-1)s/\sqrt{n}$에서만 정의되는 함수이므로 어느 정도 한계를 갖는다. 정의되는 구간 이외의 정의역에서는 표본영향함수와 이차함수 형태로 표현되는 경험적 영향함수의 차이가 x_i의 양 끝 쪽의 값으로 갈수록 점점 커질 수 있다는 것도 미리 짐작해 볼 수 있다.

4. 모의실험을 통한 경험적 영향함수와 표본영향함수의 관계 확인

4장에서는 3장에서 이론적으로 접근한 내용을 모의실험을 통해 경험적으로 확인하고자 한다. 모의실험을 진행하기 위해 R 통계 패키지에서 정규분포 $N(0, 1)$을 따르는 임의추출한 크기가 300인 표본을 사용하였고, 이렇게 추출된 표본의 기술 통계는 Table 4.1과 같다.

크기가 300인 표본의 표본평균이 0.05943 이므로 일괄적으로 300개 데이터에서 0.05943을 빼고, 다시 4만큼을 더해서 $\bar{x} = 4$를 만족시키도록 보정한 300개 데이터에 대해 모의실험을 진행하였다. 연구의 목적상 t 통계량이 적절히 큰 값으로 나와 $t_{(i)} - t$의 값 변화 관찰이 용이할 수 있어야 하므로 표본평균 $\bar{x}$가 각각 $2, 3, 4, 5$가 되도록 시프트해 본 후, t통계량이 연구 과정에 부합하는 값이라고 판단된 $\bar{x} = 4$인 경우를 선택하여 시프트를 실시하였다. 초기에 생성한 데이터에서 $\bar{x} = 4$를 만족시키도록 시프트하여 다시 생성한 데이터의 기술 통계는 Table 4.2와 같다.

SIF$(T, x_i) \approx$ EIF (T, x_i)에 의한 $T(\hat{F}_{(i)}) - T(\hat{F}) \approx \{-1/(n-1)\} \cdot$ EIF (T, x_i) 근사로 $T(\hat{F}_{(i)}) - T(\hat{F})$ 를 예측한 경우를 단순 근사(simple approximation, SA)라 하고, 단순 근사에 사용된 $\{-1/(n-1)\} \cdot$ EIF (T, x_i)의 값을 SA_T로 표현하기로 한다. 또한, 단순 근사 과정에서 생긴 $T(\hat{F}_{(i)}) - T(\hat{F})$의 값과 $\{-1/(n-1)\} \cdot$ EIF (T, x_i)의 값의 차이를 단순 예측 차이(simple prediction difference, SPD)라 하여 이 값을 로 표현한다. 3장에서 유도한 EIF(T, x_i)에 실수배, 상수항의 합 보정을 한 식으로 SIF(T, x_i)에 대입 혹은 근사시켜 $T(\hat{F}_{(i)}) - T(\hat{F})$를 예측한 경우를 보정된

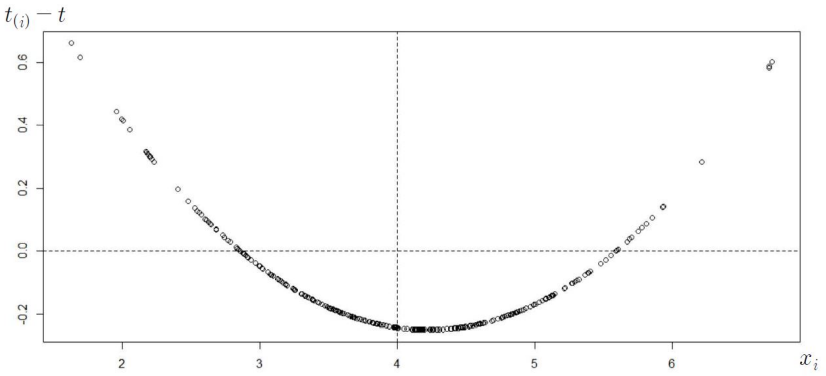

Figure 4.1: *Graph of* $t_{(i)} - t$.

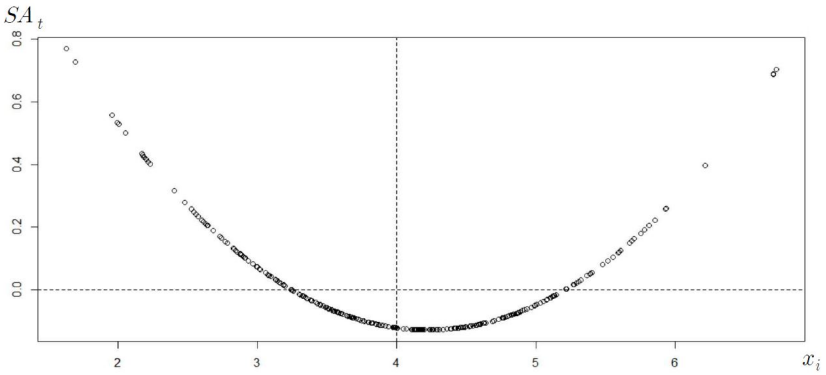

Figure 4.2: *Graph of* $-\frac{1}{n-1}EIF\,(t, x_i)$.

근사(calibrated approximation, CA)라 하고, CA_T로 표현한다. 보정된 근사에 의한$T(\hat{F}_{(i)}) - T(\hat{F})$ 와 근사한 식의 차이를 보정한 예측 차이(calibrated prediction difference, CPD)라 하고 CPD_T로 나타내겠다.

4.1. $t_{(i)} - t$ 의 근사

t 통계량에 대한 $T(\hat{F}_{(i)}) - T(\hat{F})$는 식(4.1)로 정리할 수 있다.

$$T\left(\hat{F}_{(i)}\right) - T\left(\hat{F}\right) = t_{(i)} - t = -\frac{1}{n-1}\text{SIF}\,(t, x_i). \tag{4.1}$$

식(4.1)에서 의 그래프는 Figure 4.1과 같다.

SIF ≈ EIF(t, x_i)로 $t_{(i)} - t$를 예측한 그래프는 Figure 4.2이고, 식(3.13)을 이용해 $t_{(i)} - t$를 예측한 그래프는 Figure 4.3이다. 두 그래프 개형 모두 유사한 경향을 갖지만 SIF(t, x_i) ≈ EIF(t, x_i)의 단순 근사를 통한 $\{-1/(n-1)\} \cdot \text{EIF}\,(t, x_i)$의 그래프는 실제 $t_{(i)} - t$의 그래프보다 최솟값과 최댓값이 더 크게 예측되고 있음을 알 수 있다. 반면, 식(3.13)를 이용해 $t_{(i)} - t$를 예측한 그래프는 최댓값과 최솟값을 비롯한 전반적인 값이 실제 $t_{(i)} - t$의 그래프와 일치하고 있다는 것도 확인할 수 있다. $t_{(i)} - t$의 그래프와 $\{-1/(n-1)\} \cdot \text{EIF}\,(t, x_i)$의 그래프를 Figure 4.4와 같이 한 평면에 그려보면 두 함수의 차이가 일정하게 보이는 경향성을 기하적으로도 관찰할 수 있다. 즉, t통계량에 대해서는 $t_{(i)} - t$를 $\{-1/(n-1)\} \cdot \text{EIF}\,(t, x_i)$로 단순 근사하여 예측하게 되면 그 예측의 엄밀성이 저해될 수 있다.

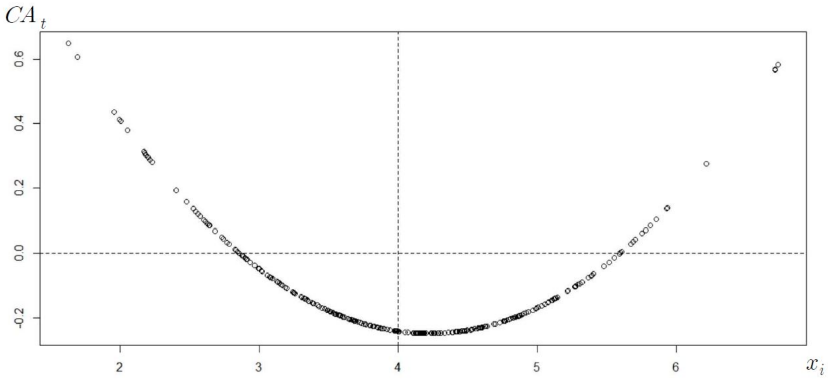

Figure 4.3: *Graph of* $-\frac{EIF\,(t,x_i+\frac{s}{t\sqrt{n}})}{n-1}+\frac{t^2+n}{2t(n-1)}+\frac{\sqrt{(n-1)(t^2-1)}}{\sqrt{n}}-t$.

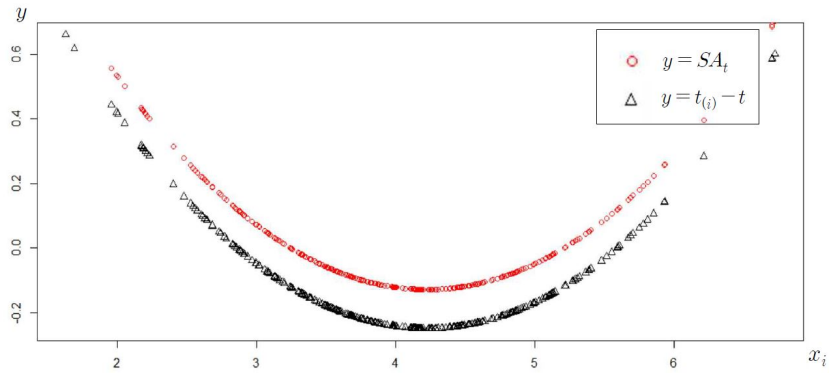

Figure 4.4: *Graph of* $y = t_{(i)} - t$ *and* $y = SA_t$.

한편, $\{-1/(n-1)\}\cdot \mathrm{EIF}\,(t, x_i + s/t\sqrt{n}) + (t^2+n)/\{2t(n-1)\} + \sqrt{(n-2)(t^2-1)}/\sqrt{n} - t$의 그래프와 $t_{(i)} - t$의 그래프도 한 평면에 나타내보면 Figure 4.5와 같이 대부분의 값이 일치하는 모습을 확인할 수 있으며, x_i의 양 끝 쪽의 값에서는 조금씩 일치하지 않는 모습도 관찰할 수 있다. 이는 3장에서 언급한 바와 같이 표본영향함수 $\mathrm{SIF}(t, x_i)$와 경험적 영향함수 $\mathrm{EIF}(t, x_i)$의 근본적인 함수적 차이에 의해 근사 과정에서 발생하는 필연적인 오차임을 밝힌다. 하지만 이러한 필연적 오차를 최대한 줄일 수 있는 측면에서 $\mathrm{SIF}(t, x_i)$에 대한 단순 근사(SA_t)보다는 보정된 근사(CA_t)가 효율성과 정확성 측면에서 우세하다고 판단할 수 있다.

$t_{(i)} - t$의 값의 단순 예측 차이 SPD_t 와 식(3.13)에 의해 보정한 예측 차이 CPD_t는 다음과 같이 정리할 수 있다.

$$\mathrm{SPD}_{\,t} = \{t_{(i)} - t\} - \left\{-\frac{1}{n-1}\mathrm{EIF}\,(t, x_i)\right\}, \tag{4.2}$$

$$\mathrm{CPD}_{\,T} = \{t_{(i)} - t\} - \left[-\frac{\mathrm{EIF}\,\left(t, x_i + \frac{s}{t\sqrt{n}}\right)}{n-1} + \frac{t^2+n}{2t(n-1)} + \frac{\sqrt{(n-2)(t^2-1)}}{\sqrt{n}} - t\right]. \tag{4.3}$$

SPD_t의 그래프는 Figure 4.6과 같다. SPD_t는 x_i의 양 끝 쪽으로 갈수록 그 크기가 증가하는 경향이 있지만, 이보다 더 큰 문제는 SPD_t의 평균값인 -0.12007, 그리고 대부분의 x_i에 대해 SPD_t값이 -0.1209근방에서 일정하게 분포하는 경향으로 $\mathrm{SIF}(t, x_i)$를 $\mathrm{EIF}(t, x_i)$로 단순 근사시켜 예측하게 되는 경우 일정해 보이는 이 오차 때문에 $t_{(i)} - t$의 정확한 예측이 어렵다.

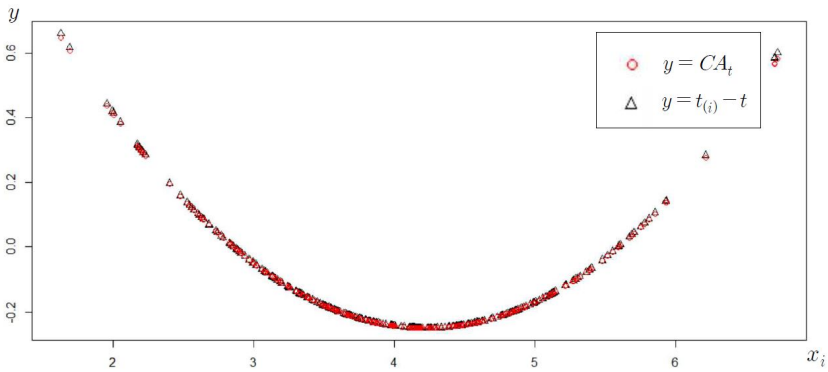

Figure 4.5: *Graph of* $y = t_{(i)} - t$ *and* $y = CA_t$.

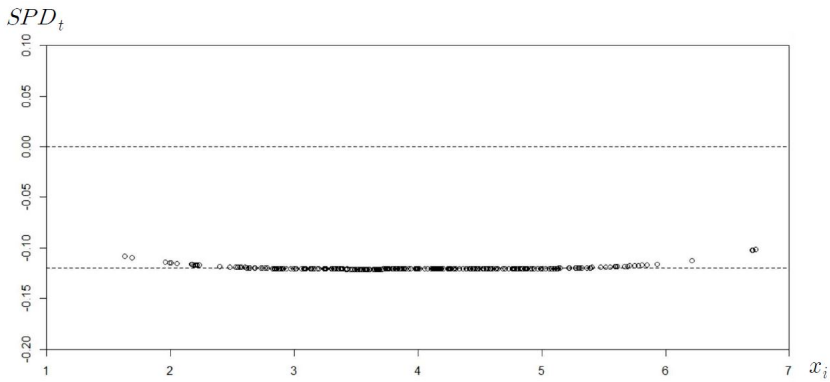

Figure 4.6: *Graph of* SPD_t.

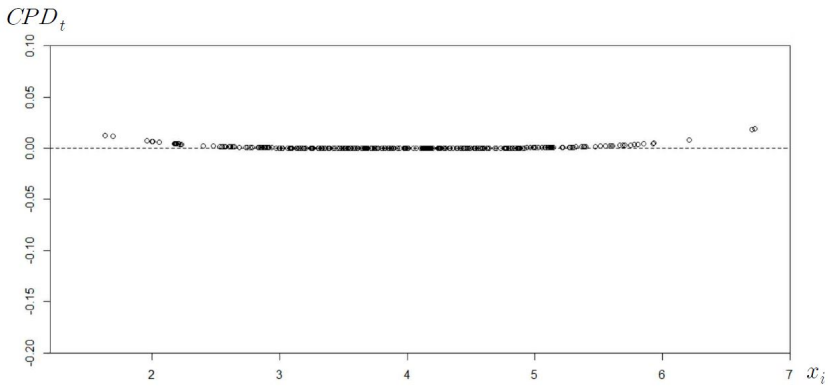

Figure 4.7: *Graph of* CPD_t.

반면, CPD_t의 그래프는 Figure 4.7과 같고, 대부분 CPD_t의 값이 0 근방에 분포하고 있는 모습을 확인할 수 있다. CPD_t의 평균은 0.00086이지만 대부분의 x_i에 대해 CPD_t의 값이 0근방에서 일정하게 분포하는 경향을 갖고 있으며, SPD_t에 비해 $t_{(i)} - t$ 예측 과정에서의 오차가 현저하게 줄어들었음을 알 수 있다. 한

Table 4.3: Comparing differences in the approximation of $t_{(i)} - t$

ID	x_i	$t_{(i)} - t$	EIF(t, x_i)	SA$_t$	SPD$_t$	CA$_t$	CPD$_t$
31	2.77011	0.03445	−46.21982	0.15458	−0.12013	0.03340	0.00105
32	2.78415	0.02899	−44.59876	0.14916	−0.12017	0.02798	0.00101
33	2.82730	0.01253	−39.71471	0.13283	−0.12030	0.01166	0.00087
34	2.83635	0.00915	−38.70911	0.12946	−0.12031	0.00830	0.00085
35	2.84676	0.00528	−37.56119	0.12562	−0.12034	0.00446	0.00082
				⋯			
101	3.57024	−0.19168	21.13440	−0.07068	−0.12100	−0.19170	0.00002
102	3.57436	−0.19240	21.34897	−0.07140	−0.12100	−0.19242	0.00002
103	3.58411	−0.19408	21.85271	−0.07309	−0.12099	−0.19410	0.00002
104	3.58622	−0.19444	21.96029	−0.07345	−0.12099	−0.19446	0.00002
105	3.59702	−0.19627	22.50737	−0.07528	−0.12099	−0.19629	0.00002
				⋯			
181	4.20198	−0.24905	38.31958	−0.12816	−0.12089	−0.24905	0.00000
182	4.23865	−0.24912	38.34181	−0.12823	−0.12089	−0.24912	0.00000
183	4.24454	−0.24909	38.33541	−0.12821	−0.12088	−0.24909	0.00000
184	4.25115	−0.24906	38.32493	−0.12818	−0.12088	−0.24906	0.00000
185	4.25873	−0.24900	38.30863	−0.12812	−0.12088	−0.24900	0.00000
				⋯			
261	5.08407	−0.15105	9.18032	−0.03070	−0.12035	−0.15141	0.00036
262	5.10051	−0.14724	8.04999	−0.02692	−0.12032	−0.14763	0.00039
263	5.10622	−0.14590	7.65203	−0.02559	−0.12031	−0.14630	0.00040
264	5.10685	−0.14575	7.60776	−0.02544	−0.12031	−0.14615	0.00040
265	5.12053	−0.14250	6.64351	−0.02222	−0.12028	−0.14292	0.00042

*Note : Empirical influence function (EIF); Simple approximaton (SA); Simple prediction difference (SPD); Calibrated approximation (CA); Calibrated prediction difference (CPD).

편, CPD$_t$의 그래프는 x_i의 중앙값 관측 위치를 기준으로 0의 값 또는 0에 수렴하는 매우 작은 값을 대부분 갖다가 양 끝 쪽으로 갈수록 차이가 조금씩 더 커지는 경향이 발생한다. CPD$_t$의 값이 x_i의 양 끝에서 미미하게 증가하는 현상은 이차함수인 EIF(t, x_i)로 초월함수(transcendental function)의 형태를 갖는 SIF(t, x_i)를 근사하는 과정에서 생기는 필연적 오차라고 할 수 있다. 하지만 Table 4.2와 같이 생성한 300개 데이터의 t 통계량이 72.558임을 감안한다면 이는 t통계량이 갖는 척도에 비해 무시할 수 있을 정도의 값이다. 결국 3장에서 유도한 식(3.13)에 의해 $t_{(i)} - t$ 의 근사에 대한 효율성과 정확도가 매우 향상되었다고 판단할 수 있다. $x_i, t_{(i)} - t, \text{EIF}(t, x_i), \text{SA}_t, \text{SPD}_t, \text{CA}_t, \text{CPD}_t$ 의 값은 각각 Table 4.3과 같다.

5. 실제 자료 분석 과정에의 적용

t 통계량에 대한 이상치 제거 순위 결정 과정에 본 연구의 내용을 적용해 본다. 2020년 대전 지역의 한 고등학교 3학년 학생 226명의 수학 점수 자료를 활용하였으며, 기술 통계는 Table 5.1과 같다. 그리고 이 중 20명의 점수를 임의추출하였다. 적절히 t통계량의 값을 키우기 위해 검정통계량의 값을 $\mu_0 = 45$으로 설정하였으며, 본 연구에서 제안한 근사 방법으로 얻은 값의 절댓값인 |CA$_t$|으로 각 관측값이 t 통계량에 미치는 영향의 정도를 얻었다. |CA$_t$|값이 클수록 t통계량에 미치는 영향이 큰 것으로 판단할 수 있으므로, 이 값이 높은 순으로 제거 순위를 부여하였다. t통계량에 미치는 영향이 큰 관측값의 제거 순위는 Table 5.2와 같다.

Table 5.1: Summary of 226 math exam scores

Mean	Median	Variance	Standard deviation
55.23	56.70	322.0569	17.9459

Table 5.2: Influence of observation and outlier removal ranking

ID	x_i(score)	$\bar{x}_{(i)}$	$s_{(i)}$	$\lvert t_{(i)} - t\rvert$	$\lvert \mathrm{CA}_t\rvert$	Removal rank (t)
1	94.6	58.8105	17.3581	0.26379	0.27554	11
9	81.1	59.5211	18.5561	0.32079	0.32081	3
23	75.1	59.8368	18.8841	0.30715	0.30805	4
28	75.0	59.8421	18.8886	0.30674	0.30767	5
29	74.9	59.8474	18.8930	0.30633	0.30728	6
38	70.5	60.0789	19.0571	0.28286	0.28457	7
42	70.4	60.0842	19.0601	0.28220	0.28393	8
46	70.2	60.0947	19.0661	0.28087	0.28263	9
53	70.0	60.1053	19.0719	0.27952	0.28130	10
66	66.2	60.3053	19.1591	0.24973	0.25187	12
87	65.2	60.3579	19.1746	0.24059	0.24277	13
122	56.4	60.8211	19.1800	0.13631	0.13850	14
123	56.4	60.8211	19.1800	0.13631	0.13850	14
132	55.8	60.8526	19.1718	0.12758	0.12981	16
141	51.6	61.0737	19.0832	0.06036	0.06329	19
143	51.5	61.0789	19.0804	0.05862	0.06158	20
178	42.3	61.5632	18.6901	0.13101	0.11929	18
180	42.1	61.5737	18.6786	0.13585	0.12375	17
209	25.5	62.4474	17.2295	0.68217	0.57233	2
220	17.2	62.8842	16.0859	1.11437	0.85469	1

*Note: Calibrated approximation (CA)

6. 결론

본 연구에서는 t통계량에 대한 표본영향함수를 유도하고, 이를 바탕으로 경험적 영향함수와 표본영향함수의 차이 및 관계를 이론적으로 확인하였다. 표본영향함수가 갖는 함수적 특성을 고찰하여 표본영향함수를 경험적 영향함수로 근사시켜 엄밀하면서도 효율성을 높일 수 있는 추론 방법을 제안하고, 모의실험을 통해 그 타당성을 검증하였다. 그 결과, 표본영향함수 $\mathrm{SIF}(t, x_i)$와 경험적 영향함수 $\mathrm{EIF}(t, x_i)$가 갖는 함수적 특성을 살펴보고 $\mathrm{EIF}(t, x_i)$의 평행이동과 상수항의 합으로 보정해 $\mathrm{SIF}(t, x_i)$를 근사적으로 추론하는 방법이 $\mathrm{SIF}(t, x_i) \approx \mathrm{EIF}(t, x_i)$의 단순 근사에 비해 정확도 측면에서 뛰어남을 보였다. $\mathrm{SIF}(t, x_i)$와 $\mathrm{EIF}(t, x_i)$가 극댓값을 갖는 x_i의 값의 차이 $s/t\sqrt{n}$과 극댓값의 차이 $(n-3/2)t - n/2t - (n-1)\sqrt{(n-2)(t^2-1)}/\sqrt{n}$를 고려해,

$$\mathrm{SIF}\,(t, x_i) \approx \mathrm{EIF}\left(t, x_i + \frac{s}{t\sqrt{n}}\right) + \left(n - \frac{3}{2}\right)t - \frac{n}{2t} - \frac{(n-1)\sqrt{(n-2)(t^2-1)}}{\sqrt{n}},$$

와 같은 근사식을 수리적으로 유도하였다. 그리고 이 근사식을 바탕으로 $t_{(i)} - t$의 예측을 위하여,

$$t_{(i)} - t \approx -\frac{1}{n-1}\mathrm{EIF}\left(t, x_i + \frac{s}{t\sqrt{n}}\right) + \frac{t^2+n}{2t(n-1)} + \frac{\sqrt{(n-2)(t^2-1)}}{\sqrt{n}} - t,$$

의 근사 방법이 단순 근사 방법에 의한 예측보다 정확성이 높다는 사실을 모의실험으로 확인하여 그 타당성도 검증하였다. 아울러, 본 연구에서 다룬 근사 방안을 실제 자료 분석에 적용해 보며, 이상치 선정 과정에 활용

가능함을 확인할 수 있었다.

References

Campbell NA (1978). The influence function as an aid to outlier detection in discrimination analysis, *Applied Statistics*, **27**, 251–258.

Cook RD (1977). Detection of influential observation in linear regression, *Technometrics*, **19**, 15–18.

Cook RD and Weisberg S (1980). Characterization of empirical influence function for detection influential cases in regression, *Technometrics*, **22**.

Cook RD and Weisberg S (1982). *Residual and Influence in Regression*, Chapman & Hall, New York.

Critchley F (1985). Influence in principal components analysis, *Biometika*, **72**, 627–636.

Hampel FR (1974). The Influence curve and its role in robust estimation, *Journal of the American Statistical Association*, **69**, 383–393.

Kang H 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.

Kim H (1992). Measures of influence in correspondence analysis, *Journal of Statistical Computation and Simulation*, **40**, 201–217.

Kim H (1998). A study on cell influence to Chi-square statistic in contingency tables, *The Korean Communications in Statistics*, **5**, 35–42.

Kim H and Lee H (1996). Influence functions on χ^2 statistic in contingency tables, *The Korean Communications in Statistics*, **3**, 69–76.

Kim H and Kim K (2005). Influence of an observation on the t-statistic, *The Korean Communications in Statistics*, **12**, 453–462.

Kim S and Kim H (2019). A study on the performance of the influence function on the t-statistic depending on population distributions, *Journal of the Korean Data & Information Science Society*, **30**, 573–585.

Lee H and Kim H (2003). The changes in statistic when a row is deleted from a contingency table, *The Korean Communications in Statistics*, **10**, 305–317.

Lee H and Kim H (2008). Influence function on the coefficient of variation, *Communications for statistical applications and methods*, **15**, 509–516.

Park S and Kim H (2019). A study on the location of the observation which has the least effect on the-statistic, *Journal of the Korean Data & Information Science Society*, **30**, 1221–1232

Radhkrishnan R and Kshirsagar AM (1981). Influence functions for certain parameters in multi-variate analysis, *Communications in Statistics*, **10**, 515–529.

Received July 27, 2021; Revised August 20, 2021; Accepted August 30, 2021

경험적 영향함수와 표본영향함수 간 차이 보정의 t통계량으로의 확장

강현석[a], 김홍기[1,b]

[a]교육부; [b]충남대학교 정보통계학과

요 약

본 연구는 Kang과 Kim (2020)의 후속 연구이다. 본 연구에서는 기존 연구에서 직접 유도하지 않았던 통계량의 표본영향함수를 유도한다. 그리고 이 결과를 바탕으로 경험적 영향함수와 표본영향함수는 어떠한 관계를 가지고 있는지 이론적으로 살펴보고, 경험적 영향함수를 통해 표본영향함수를 근사시켜 추정하는 방안에 대해 생각해 본다. 또한, 임의추출한 300개의 데이터를 바탕으로 모의실험을 통해 유도한 함수와 그 관계에 대한 그 타당성도 검증한다. 모의실험 결과 t통계량으로부터 유도한 표본영향함수와 경험적 영향함수와의 관계 및 경험적 영향함수를 통한 표본영향함수의 근사 방안에 대한 타당성도 검증해 냈다. 본 연구는 경험적 영향함수를 이용한 표본영향함수의 근사에서 오차를 줄이기 위한 방안을 제안하고 그 타당성을 검증하였으며, 이를 통해 기존의 연구에서 경험적 영향함수로 표본영향함수를 바로 근사시켰던 연구 방법에 효과적인 근사 방안을 제안한 점에서 의의를 갖는다.

주요용어: 영향함수, 이상치, t통계량, 경험적 영향함수, 표본영향함수

[1]교신저자: (34134) 대전광역시 유성구 대학로 99, 충남대학교 정보통계학과. Email: honggiekim@cnu.ac.kr