The Korean Journal of Applied Statistics (2014) DOI: http://dx.doi.org/10.5351/KJAS.2014.27.3.501
27(3), 501–512

A Study on the Optimal Cut-off Point in the Cut-off Sampling Method

Sang Eun Lee[a] · Min Ji Cho[a] · Key-Il Shin[b,1]

[a]Department of Applied Information Statistics, Kyonggi University
[b]Department of Statistics, Hankuk University of Foreign Studies

(Received April 17, 2014; Revised May 22, 2014; Accepted June 9, 2014)

Abstract

Modified cut-off sampling is widely used for highly skewed data. A serious drawback of modified cut-off sampling is the difficulty of adjustment of non-response in take-all stratum. Therefore, solutions of the problems of non-response in take-all stratum have been studied in various ways such as substitute of samples, imputation or re-weight method. In this paper, a new cut-off point based on minimizing MSE being used in exponential and power functions is suggested and it can be reduced the number of take-all stratum. We also investigate another cut-off point determination method with underlying distributions such as truncated log-normal and truncated gamma distributions. Finally we suggest the optimal cut-off point which has a minimum of take-all stratum size among suggested methods. Simulation studies are performed and Labor Survey data and simulated data are used for the case study.

Keywords: Take-all stratum, truncated log-normal distribution, truncated gamma distribution, response rate.

1. 서론

절사표본추출법(cut-off sampling method)은 층화추출법(stratified sampling)의 하나로 관심 있는 모집단의 특성치가 한 쪽으로 편중되어 있거나 일부 표본에 대한 신뢰성 있는 표본틀이 존재하지 않는 경우에 흔히 사용된다. 특히 사업체 조사의 고용·생산·판매 등과 같이 조사변수의 분포가 비대칭적인 모집단인 경우에 유용하게 쓰인다. 이러한 절사표본추출법은 시간과 비용 측면에서 경제적이며, 동일한 허용오차범위 내에서 최소의 표본규모를 사용하는 장점을 갖고 있다. 이에 관한 자세한 내용은 Han (2004)와 Namkung (2008)을 참조하기 바란다.

Hidiroglou (1986)는 주어진 허용오차에서 최소의 표본 수를 만족하는 절사점을 제공하는 수정절사법을 제안하였으며 이후 Lavellee와 Hidiroglou (1988)은 전수층, 여러 개의 표본층 그리고 절사층 등으로 구성된 절사법을 위한 알고리즘을 개발하였고, 최근 Baillargeon과 Rivest (2011)는 절사법 사용을 위한 R 코드를 개발하여 절사법은 이제 쉽게 사용할 수 있게 되었다. 따라서 전수층에 무응답이 없고 조사에 어려움이 없다면 기존의 절사표본추출법은 매우 효과적인 방법이다.

This paper was supported by the Hankuk University of Foreign Studies research fund(2014).
[1]Corresponding author: Department of Statistics, Hankuk University of Foreign Studies, Gyeonggi-do 449-791, Korea. E-mail: keyshin@hufs.ac.kr

그러나 최근 조사환경이 열악해지고 정보공개를 꺼리는 사업체, 특히 대기업들의 무응답이 증가하면서 절사표본설계 사용에 어려움이 발생하고 있다. 즉 절사표본설계의 큰 장점인 전수층의 존재는 전수층에서 발생한 무응답으로 인하여 더 이상 전수층 역할을 하지 못하게 되고, 이로 인하여 모수 추정의 정도는 나빠지게 되었다.

따라서 전수층에서 발생하는 무응답 비율을 줄이기 위해 표본설계 단계에서부터 허용오차를 만족하는 조건 하에서 전수층 크기를 최소화하는 연구가 이루어지고 있다. Shin과 Lee (2013)는 절단 로그정규 분포와 절단 감마분포를 이용하여 최적의 절사점을 구하는 방법을 제안하였으며 Lee (2009)와 Lee (2011)는 Hidiroglou (1986)가 제시한 방법을 이용하여 표본크기를 계산한 후, 계산된 표본크기를 고정한 상태에서 MSE를 최소로 하는 절사점을 수치 해석(numerical analysis)적으로 구하는 방법을 제안하였다.

본 연구에서는 주어진 허용오차에서 표본규모를 최소로 하여 절사점을 정하는 기존의 Hidiroglou (1986) 방법을 확장하여 Hidiroglou 방법에서 계산된 표본규모를 고정한 상태에서 MSE를 최소로 하는 절사점 선택 방법을 제안하였다. 이때 MSE를 최소화하는 방법으로 모집단 자료의 분포에 특정 함수를 적합한 후, 적합된 함수를 이용하여 최적 절사점(optimal cut-off point)을 찾는 방법을 사용하였다.

본 논문의 구성은 다음과 같다. 먼저 2절에서는 절사점 결정방법으로 기존의 Hidiroglou 방법과 함수를 적용하여 결정하는 새로운 방법을 소개하였다. 또한 3절에서는 개발된 방법과 이미 사용되고 있는 방법의 비교를 위한 모의실험을 실시하였으며, 4절에서는 실제 사례를 통하여 각 방법별로 계산된 전수층 규모를 비교 분석하였다. 마지막으로 5절에 결론이 있다.

2. 절사점 방법론

이 절에서는 Hidiroglou가 제안한 방법과 본 연구에서 제안한 지수함수(exponential function)와 멱함수(power function)를 이용한 절사점 결정 방법을 소개하였다.

2.1. Hidiroglou의 절사점

우선 표본틀의 추출단위를 값이 큰 순서대로 다음과 같이 나열한다.

$$y_{(1)}, y_{(2)}, \ldots, y_{(N)}, y_{(i)} \leq y_{(i+1)}, \quad i = 1, 2, 3, \ldots, N-1.$$

그러면 총계 Y, $Y = \sum_{i=1}^{N-t} y_{(i)} + \sum_{i=N-t+1}^{N} y_{(i)}$이 되며 여기서 N은 전체 모집단 수, $N-t$는 표본층의 모집단 수, t는 전수층 모집단 수이다. 또한 총계 추정량 $\hat{Y}$은 다음과 같이 표현된다.

$$\hat{Y} = \frac{N-t}{n(t)-t} \sum_{i=1}^{n(t)-t} y_i + \sum_{i=N-t+1}^{N} y_{(i)}, \tag{2.1}$$

여기서 $y_{(N-t-1)}$에서 $y_{(N)}$는 전수층에 포함된 자료이며 $n(t)$는 추출된 전체 표본크기, $n(t)-t$는 표본층의 표본크기가 된다. 또한 총계 추정량의 분산은 다음과 같다.

$$V\left(\hat{Y}\right) = \frac{(N-t)\{N-n(t)\}}{n(t)-t} S^2_{[N-t]}, \tag{2.2}$$

여기서 $S^2_{[N-t]}$는 표본층의 분산으로 $S^2_{[N-t]} = 1/(N-t-1) \sum_{i=1}^{N-t} (y_{(i)} - \bar{Y}_{[N-t]})^2$이고 $\bar{Y}_{[N-t]} = 1/(N-t) \sum_{i=1}^{N-t} y_{(i)}$이다.

절사표본추출을 위한 표본설계의 핵심 내용은 전체 표본 수 $n(t)$와 전수층 크기 t를 결정하는 것이다. Hidiroglou가 제안한 방법은 t를 하나씩 증가해 가면서 식 (2.3)의 전체 표본크기 $n(t)$를 구하는 것이다. 여기서 E를 허용오차라 하면 $n(t)$는 다음과 같다.

$$n(t) = t + \frac{(N-t)^2 S_{[N-t]}^2}{E^2 \hat{Y}^2 + (N-t) S_{[N-t]}^2}. \tag{2.3}$$

또한 최종적으로 주어진 전체 표본크기, $n(t)$에 대응되는 전수층 크기 t를 절사점 t_H로 결정한다. 자세한 내용은 Hidiroglou (1986)을 살펴보기 바란다.

2.2. 함수를 이용한 절사점

Lee (2009)는 전체 표본크기 $n(t)$가 주어진 상태에서 수치해석적으로 MSE를 최소로 하는 절사점 선택 방법에 관하여 연구하였다. 그러나 수치해석적 방법을 사용하는 경우에는 전수층 크기에 따른 분산 값이 자료의 구조에 따라 비정상적인 형태를 보일 수 있다. 이러한 단점을 보완하기 위해 본 연구에서는 표본층 분산 $S_{[N-t]}^2$을 전수층 크기 t의 미분 가능한 함수로 표현하였다. 이 방법을 사용하게 되면 수치해석적인 방법을 이용하여 최소값을 찾을 때 발생할 수 있는 비정상적 분산값의 영향력을 제거할 수가 있으며 이론적인 전수층 최소값이 얻어진다. 본 연구에서 사용되는 식 (2.1)의 총계 추정량, $\hat{Y}$은 불편 추정량이기 때문에

$$\mathrm{Min}_t \left[\mathrm{MSE}\left(\hat{Y}\right)\right] = \mathrm{Min}_t \left[V\left(\hat{Y}\right)\right]$$

이 된다. 또한 본 연구에서는 총계 추정량의 분산으로 식 (2.2)에서 정의된 것과 같은 $V(\hat{Y})$을 사용하였다. 이에 본 연구에서는 표본층의 분산 $S_{[N-t]}^2$에 다음의 두 가지 함수를 적합하여 MSE를 최소로 하는 새로운 절사점 $t_{Optimal}$을 제안하였다. 실제로 표본층의 분산 $S_{[N-t]}^2$은 정수인 전수층 모집단 수 t의 함수이지만 본 연구에서는 연속인 두 함수를 사용하였다.

2.2.1. 지수함수(exponential function) 본 연구에서 사용한 함수 중 첫 번째 함수는 지수함수로 그 형태는 다음과 같다.

$$\text{Model1}: \ S_{[N-x]}^2 = ae^{-bx}. \tag{2.4}$$

따라서 MSE를 최소로 하는 함수는 다음과 같이 표현할 수 있다.

$$\mathrm{Min}_x \,\mathrm{MSE} = \mathrm{Min}\left[\frac{(N-x)\{N-n^H\}}{n^H - x}\left(ae^{-bx}\right)\right], \tag{2.5}$$

여기서 n^H는 주어진 총 표본 수이다. 이때 식 (2.5)를 x에 대해 미분하여 총계 추정량의 분산 $V(\hat{Y})$를 최소로 하는 절사점 t_{Exp}를 결정할 수 있으며 그 결과는 다음과 같다.

$$t_{Exp} = \left(\left(N+n^H\right)b \pm \left[\left(\left(N+n^H\right)b\right)^2 + 4b\left(N - n^H - bNn^H\right)\right]^{\frac{1}{2}}\right)(2b)^{-1},$$

여기서 모집단 크기 N은 이미 결정되어 있으며 전체 모집단 N에서 추출한 전체표본의 크기 n^H는 Hidiroglou 방법으로 계산된 값을 사용한다. 그러므로 지수함수를 적용하여 산출되는 절사점 t_{Exp}은 b에 의해 결정된다. 이때 b는 식 (2.4)를 이용하여 추정된 값을 사용하면 된다. 물론 $\ln(S_{[N-x]}^2) = \ln(a) - bx$를 이용하여 b를 추정할 수도 있으나 본 연구에서는 b값으로 SAS NLIN Procedure의 최소제곱추정량(LSE; least squares estimator)을 통해 추정 된 $\hat{b}$을 사용하였다.

2.2.2. 파워함수(power function) 다음에 고려한 함수는 파워함수 형태로 다음과 같다.

$$\text{Model2}: \ S^2_{[N-x]} = cx^{-\gamma}. \tag{2.6}$$

지수함수와 마찬가지로 파워함수를 이용하여 MSE를 최소로 하는 절사점을 구하였다. 즉

$$\text{Min}_x \, \text{MSE} = \text{Min}\left[\frac{(N-x)\left(N-n^H\right)}{n^H - x}\left(cx^{-\gamma}\right)\right]$$

이다. 여기서 n^H는 주어진 총 표본 수이다. 이제 x에 대해 미분하고 정리하면 총계 추정량의 분산 $V(\hat{Y})$를 최소로 하는 새로운 절사점 t_{Power}를 계산할 수 있으며 결과는 다음과 같다.

$$t_{Power} = \left(\left[N - n^H + \gamma\left(N + n^H\right)\right] \pm \left[\left(N - n^H + \gamma\left(N + n^H\right)\right)^2 - 4\gamma\left(\gamma N n^H\right)\right]^{\frac{1}{2}}\right)(2\gamma)^{-1}. \tag{2.7}$$

식 (2.7)에서도 주어진 모집단 크기 N과 Hidiroglou (1986)의 방법으로 계산된 전체 표본크기 n^H 값을 사용한다. 따라서 파워함수를 적용하여 산출되는 절사점인 t_{Power}는 γ에 의해 결정된다. 여기서 γ 추정을 위해 $\ln(S^2_{[N-x]}) = \ln(c) - \gamma \ln(x)$를 이용할 수도 있으나 본 연구에서는 식 (2.6)에 SAS NLIN Procedure를 적용하여 얻어진 최소제곱추정량 $\hat{\gamma}$을 사용하였다.

2.2.3. 최적 함수 선정 표본층의 분산에 함수를 적합하여 산출 된 두 가지 절사점, 즉, 지수함수를 통해 산출된 절사점 t_{Exp}과 파워함수를 통해 산출된 절사점 t_{Power} 중에서 보다 적합한 절사점을 찾기 위한 기준으로 Pseudo-R^2을 사용하였다. 이때 Pseudo-R^2의 정의는 다음과 같다.

$$\text{Pseudo-}R^2 = 1 - \frac{\text{SSerror}}{\text{SStotal(correced)}}.$$

본 연구에서는 Pseudo-R^2를 기준으로 정해진 함수기반 최종 절사점을 t_{Model*}라고 표시하였다. 만약 계산된 두 값이 매우 유사한 경우에는 작은 절사점 값을 주는 모형을 선택하였다.

2.2.4. 최적 절사점 결정 표본층의 분산 $S^2_{[N-t]}$에 함수를 적합하여 얻어진 함수기반 절사점과, Hidiroglou가 제안한 방법에 의해 결정된 절사점을 비교한 후 최종적으로 최적의 절사점을 선택하였다. 즉 본 연구에서는 두 개의 절사점 t_{Exp}, t_{Power} 중 Pseudo-R^2을 기준으로 선택된 절사점 t_{Model*}과, Hidiroglou 방법으로 산출된 절사점 t_H를 비교하여 전수층의 수가 작은 값을 갖는 절사점을 최적 절사점인 $t_{Optimal}$로 결정하였다. 즉

$$t_{Optimal} = \min(t_H, t_{Model*})$$

이다.

2.3. 절단 분포를 이용한 절사점

Shin과 Lee (2013)는 표본층의 분산 $S^2_{[N-t]}$에 절단 분포(truncated distribution)에서 계산된 분산을 적용하여 이론적인 절사점을 계산하였다. 그들은 비대칭적 분포를 갖는 대표적 분포인 감마분포(gamma distribution)와 로그정규분포(log-normal distribution)의 절단 분포를 연구하였으며 이에 본 연구에서도 감마분포와 로그정규분포를 사용하였다. 따라서 본 연구에서 제안한 새로운 절사점과 Hidiroglou가 제안한 수치해석적 절사점, 그리고 이론적 절사점을 비교하였다. 물론 이 경우도 Hidiroglou 방법을 적용하여 구해진 $n(t)$를 적용하며 이때 대응되는 전수층 크기가 최적으로 결정된다. 즉 모든 비교에서 하나의 전체 표본크기 $n(t)$가 사용된다.

2.3.1. 절단 감마분포 절단 감마분포(truncated gamma distribution)의 확률밀도함수($p.d.f$)는 다음과 같다.

$$f\left(x \mid \alpha, \beta, x_0\right) = \frac{x^{\alpha-1}\exp(-x/\beta)}{I(\alpha,\beta,x_0)}, \quad \text{for } 0 < x < x_0,$$

여기서 $I(\alpha,\beta,x_0) = \int_0^{x_0} x^{\alpha-1}\exp(-x/\beta)dx$이고 $\alpha > 0$, $\beta > 0$이므로 절단 감마분포의 절단 1차 및 2차 적률을 구하면 아래와 같다.

$$E(X \mid x_0) = \int_0^{x_0} x^{\alpha}\exp\left(-\frac{x}{\beta}\right)dx / I(\alpha,\beta,x_0) = \frac{I(\alpha+1,\beta,x_0)}{I(\alpha,\beta,x_0)},$$
$$E(X^2 \mid x_0) = \frac{I(\alpha+2,\beta,x_0)}{I(\alpha,\beta,x_0)}.$$

따라서 표본층의 분산 $S^2_{[N-t]}$은 다음과 같이 나타낼 수 있다.

$$\begin{aligned} S^2_{[N-t]} &= \frac{I(\alpha+2,\beta,x_0)}{I(\alpha,\beta,x_0)} - \left[\frac{I(\alpha+1,\beta,x_0)}{I(\alpha,\beta,x_0)}\right]^2 \\ &= \frac{\alpha(\alpha+1)\beta^2 \mathrm{CDF}\,\Gamma(\alpha+2,\beta,x_0)}{\mathrm{CDF}\,\Gamma(\alpha,\beta,x_0)} - \alpha^2\beta^2\left[\frac{\mathrm{CDF}\,\Gamma(\alpha+1,\beta,x_0)}{\mathrm{CDF}\,\Gamma(\alpha,\beta,x_0)}\right]^2, \end{aligned}$$

여기서 $\mathrm{CDF}\,\Gamma(\alpha,\beta)$는 감마분포의 누적분포함수(CDF; cumulative distribution function)를 나타낸다. 또한, $I(\alpha,\beta,x_0) = \int_0^{x_0} x^{\alpha-1}\exp(-x/\beta)dx$이므로 절단 분포를 위한 절단지점 $x_0 = I^{-1}(1-t/N \mid \alpha,\beta)$를 설정함으로써 새로운 절사점 t_{Trun}를 얻을 수 있다.

2.3.2. 절단 로그정규분포 절단 로그정규분포(truncated log-normal distribution)의 k차 적률은 다음과 같다.

$$E\left(X^k \mid x_0\right) = \exp\left(k\mu + \frac{1}{2}k^2\sigma^2\right) \times \frac{\Phi\left(b_0 - k\sigma\right)}{\Phi(b_0)},$$

여기서 x_0는 절단 지점, $b_0 = (\log x_0 - \mu)/\sigma$이며, $\Phi(\bullet)$은 정규분포의 $c.d.f$이다. 또한 절단 로그정규분포(truncated log-normal distribution)의 평균과 분산은 다음과 같다.

$$E\left(X \mid x_0\right) = \exp\left(\mu + \frac{1}{2}\sigma^2\right) \times \frac{\Phi(b_0-\sigma)}{\Phi(b_0)},$$
$$E\left(X^2 \mid x_0\right) = \exp\left(2\mu + 2\sigma^2\right) \times \frac{\Phi(b_0-2\sigma)}{\Phi(b_0)}.$$

따라서 표본층의 분산 $S^2_{[N-t]}$은 아래와 같다.

$$S^2_{[N-t]} = \exp\left(2\mu+2\sigma^2\right) \times \frac{\Phi(b_0-2\sigma)}{\Phi(b_0)} - \left(\exp\left(\mu+\frac{1}{2}\sigma^2\right) \times \frac{\Phi(b_0-\sigma)}{\Phi(b_0)}\right)^2,$$

여기서 $b_0 = \Phi^{-1}(1-t/N)$이므로 이 식을 정리하면 다음과 같다.

$$S^2_{[N-t]} = \exp\left(2\mu+2\sigma^2\right) \times \frac{\Phi\left(\Phi^{-1}\left(1-\frac{t}{N}\right)-2\sigma\right)}{\left(1-\frac{t}{N}\right)} - \left(\exp\left(\mu+\frac{1}{2}\sigma^2\right) \times \frac{\Phi\left(\Phi^{-1}\left(1-\frac{t}{N}\right)-\sigma\right)}{\left(1-\frac{t}{N}\right)}\right)^2.$$

따라서 산출한 표본층의 분산 $S^2_{[N-t]}$를 통해 새로운 절사점 t_{Trun}를 얻을 수 있다.

Table 3.1. Total sample size and cut-off point on gamma distribution

허용오차	모집단크기 N	(α, β)	왜도	전체표본크기 $n(t)$	절사점 t_H
0.03	10,000	(0.050, 30)	9.65	1,106	910
		(0.030, 30)	11.72	808	682
		(0.010, 30)	17.69	359	318
		(0.005, 30)	23.84	197	177
		(0.003, 30)	31.65	133	123
		(0.001, 30)	45.23	50	45
0.05	10,000	(0.050, 30)	9.65	906	706
		(0.030, 30)	11.72	679	554
		(0.010, 30)	17.69	311	262
		(0.005, 30)	23.84	176	155
		(0.003, 30)	31.65	120	107
		(0.001, 30)	45.23	45	42

Table 3.2. Total sample size and cut-off point on log-normal distribution

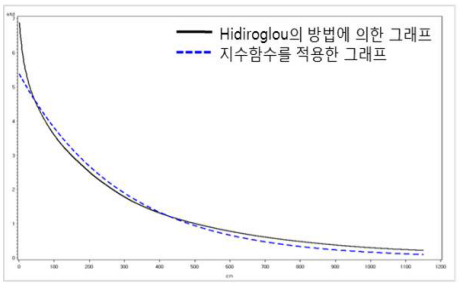
허용오차	모집단크기 N	(μ, σ^2)	왜도	전체표본크기 $n(t)$	절사점 t_H
0.03	10,000	$(10, 1^2)$	8.76	1,775	826
		$(10, 1.25^2)$	18.59	1,821	989
		$(10, 1.5^2)$	33.01	1,728	1,010
		$(10, 1.75^2)$	47.49	1,553	958
		$(10, 2^2)$	58.49	1,336	848
		$(10, 2.25^2)$	65.77	1,102	702
0.05	10,000	$(10, 1^2)$	8.76	1,041	408
		$(10, 1.25^2)$	18.59	1,141	544
		$(10, 1.5^2)$	33.01	1,130	607
		$(10, 1.75^2)$	47.49	1,045	605
		$(10, 2^2)$	58.49	919	558
		$(10, 2.25^2)$	65.77	772	492

3. 모의실험

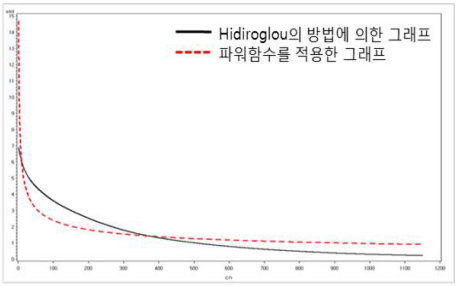
본 모의실험에서 사용된 모집단 분포는 2절에서 연구된 분포인 감마분포와 로그정규분포를 따르는 것으로 설정하였다. 잘 알려진 것처럼 이 분포의 경우에는 모수에 따라서는 왜도가 매우 커질 수 있다. 먼저 모집단 크기 N은 10,000개로 정하고, 95% 신뢰도에 허용오차는 3%와 5%로 구분하여 살펴보았으며, 이때 왜도는 최소 8.76에서 최대 65.77까지 나타나도록 하였다. 참고로 감마분포의 모수 β와 로그정규분포의 모수 μ는 절사점 결정과 무관하기 때문에 본 모의실험에서는 고정된 값을 사용하였다. Table 3.1과 Table 3.2는 Hidiroglou 방법을 이용하여 구한 전체 표본크기와 절사점, t_H이다.

3.1. 함수를 이용한 절사점 산정

3.1.1. 감마분포 Table 3.1에서 α가 0.05이고 β가 30인 감마분포를 따르는 모집단 자료의 결과를 자세히 살펴보면 다음과 같다. 먼저 표본층의 분산 $S^2_{[N-t]}$에 대한 적합성 검정을 살펴본 결과, Pseudo-R^2이 0.9917로 지수함수가 보다 잘 적합한 것으로 나타났다. 이는 x축이 전수층 크기 t이고 y축이 $S^2_{[N-t]}$인 Figure 3.1을 통해서도 쉽게 알 수 있다. 또한 주어진 전체 표본 개수에 해당하는 절사점을 구한다. 따라서 t_{Model*}는 t_{Exp}로 결정된다. 같은 방법으로 각각의 α와 β에 따른 결과는 Table



$Pseudo\ R^2 = 0.9917499$

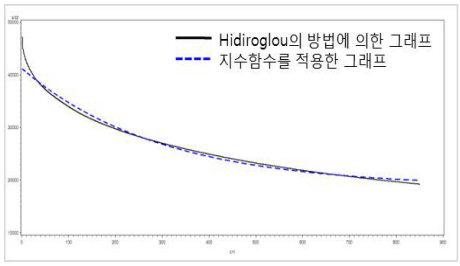


$Pseudo\ R^2 = 0.8646401$

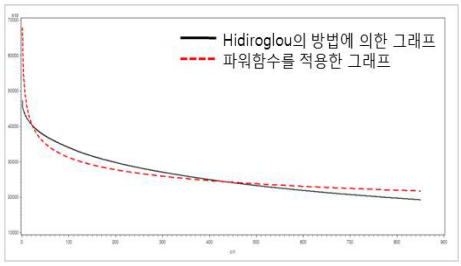
Figure 3.1. Fitted results using Gamma(0.05, 30)

Table 3.3. Cut-off point using fitted function method on gamma distribution

(α, β)	Pseudo-R^2_{Exp}	Pseudo-R^2_{Power}	허용 오차					
			0.03			0.05		
			t_{Exp}	t_{Power}	t_{Model*}	t_{Exp}	t_{Power}	t_{Model*}
(0.001, 30)	0.9944342	0.9204820	45	24	t_{Exp}	40	21	t_{Exp}
(0.003, 30)	0.9897497	0.8885967	119	54	t_{Exp}	106	48	t_{Exp}
(0.005, 30)	0.9975174	0.8475331	170	72	t_{Exp}	149	64	t_{Exp}
(0.010, 30)	0.9963211	0.8250425	306	123	t_{Exp}	258	106	t_{Exp}
(0.030, 30)	0.9918866	0.8456192	638	254	t_{Exp}	509	212	t_{Exp}
(0.050, 30)	0.9917499	0.8646401	829	332	t_{Exp}	628	269	t_{Exp}



$Pseudo\ R^2 = 0.9901938$



$Pseudo\ R^2 = 0.9943299$

Figure 3.2. Fitted results using log-normal(10, 1)

3.3과 같다

3.1.2. 로그정규분포 다음은 로그정규분포를 따르는 모집단 자료를 생성한 후 표본층의 분산에 지수함수와 파워함수를 적합한 모의실험 결과이다.

우선 μ는 10이고 $\sigma = 1$인 로그정규분포를 따르는 모집단 자료의 결과를 살펴보면 Figure 3.2와 같이 표본층 분산 $S^2_{[N-t]}$에 적합된 Pseudo-R^2가 0.9943으로 파워함수가 지수함수보다 더 잘 적합한 것으로 나타났다. 따라서 t_{Model*}는 t_{Power}로 결정하였다. 같은 방법으로 로그정규분포의 결과는 Table 3.4와 같다.

Table 3.4. Cut-off point using fitted function method on log-normal distribution

(μ, σ^2)	Pseudo-R^2_{Exp}	Pseudo-R^2_{Power}	허용 오차					
			0.03			0.05		
			t_{Exp}	t_{Power}	t_{Model*}	t_{Exp}	t_{Power}	t_{Model*}
(10, 1^2)	0.9901938	0.9943299	1,490	295	t_{Power}	755	163	t_{Power}
(10, 1.25^2)	0.9826621	0.9891558	1,561	380	t_{Power}	880	226	t_{Power}
(10, 1.5^2)	0.9821122	0.9880138	1,449	365	t_{Power}	850	228	t_{Power}
(10, 1.75^2)	0.9508287	0.9791979	1,385	421	t_{Power}	876	275	t_{Power}
(10, 2^2)	0.9142857	0.9793084	1,219	441	t_{Power}	802	297	t_{Power}
(10, 2.25^2)	0.8508101	0.9765170	1,024	441	t_{Power}	297	297	t_{Power}

Table 3.5. Cut-off point using truncated distribution method

Gamma(α, β)	허용 오차		$N(\mu, \sigma^2)$	허용 오차	
	0.03	0.05		0.03	0.05
	t_{Trun}	t_{Trun}		t_{Trun}	t_{Trun}
(0.001, 30)	24	21	(10, 1^2)	1024	555
(0.003, 30)	93	82	(10, 1.25^2)	1050	612
(0.005, 30)	153	133	(10, 1.5^2)	983	602
(0.010, 30)	313	265	(10, 1.75^2)	864	548
(0.030, 30)	763	633	(10, 2^2)	724	472
(0.050, 30)	1,063	862	(10, 2.25^2)	577	386

3.1.3. 절단 분포를 이용한 절사점 산정 Shin과 Lee (2013)가 제안한 절단 감마분포(truncated gamma distribution)와 로그정규분포(truncated distribution)에서 산출된 절사점 t_{Trun}를 모의실험자료에 적용한 결과를 Table 3.5에 수록하였다. 본 모의실험에서는 알려진 모수값을 사용하였다. 만약 모수가 알려져 있다면 알려진 모수를 사용하면 되고, 만약 모수가 알려져 있지 않으면 MLE 등을 이용하여 모수를 추정한 후 2.3절의 결과를 이용하면 절사점을 구할 수 있다.

3.2. 모의실험 결과 비교

감마분포로부터 10,000개의 모집단 자료를 생성하고 허용오차를 0.03과 0.05로 모의실험을 수행 한 결과를 정리하면 다음과 같다. 비교적 왜도가 크지 않은 경우, 즉 α가 0.01, 0.03, 0.05일 때는 표본층의 분산에 지수함수를 적용하여 산출한 절사점인 t_{Model*}가 최적 절사점으로 나타났고, 분포의 왜도가 큰 경우, 즉 α가 0.001, 0.003, 0.005일 때는 절단 감마분포를 이용하여 산출한 절사점인 t_{Trun}이 최적 절사점으로 나타났다. 그 결과는 Table 3.6과 같다. 마찬가지로 로그정규분포로부터 10,000개의 모집단 자료를 생성하고 허용오차를 0.03과 0.05로 모의실험을 수행 한 결과를 Table 3.7에 정리하였다. 이 경우에는 표본층의 분산에 파워함수를 적용하여 산출한 절사점인 t_{Model*}가 최적 절사점으로 나타났다.

4. 사례연구

본 논문의 사례연구에서는 고용노동정책의 기초자료 활용 및 경기전망 등을 파악하는 것을 목적으로 하는 사업체노동력조사(구: 사업체임금근로시간조사)의 2007년 임금자료를 사용하였다. 사업체노동력조사는 매월 전국의 종사자 1인 이상 민간사업체 및 공공기관 중에서 층화계통추출법으로 선정 된 표본사업체를 대상으로 종사자수, 빈 일자리 수, 입·이직자수, 임금 및 근로시간에 관한 사항을 조사한다. 그 중에서 임의로 선정된 지역에 대한 임금자료를 이용하여 사례연구를 실시하였다. 물론 실제 자료 분석

Table 3.6. Comparison results of cut-off points on Gamma Distribution

N	허용오차	(α, β)	왜도	$n(t)$	t_H	t_{Model*}	t_{Trun}	$t_{Optimal}$
10,000	0.03	(0.001, 30)	45.23	50	45	45	24	t_{Trun}
		(0.003, 30)	31.65	133	123	119	93	t_{Trun}
		(0.005, 30)	23.84	197	177	170	153	t_{Trun}
		(0.010, 30)	17.69	359	318	306	313	t_{Model*}
		(0.030, 30)	11.72	808	682	638	763	t_{Model*}
		(0.050, 30)	9.65	1,106	910	829	1,063	t_{Model*}
	0.05	(0.001, 30)	45.23	45	42	40	21	t_{Trun}
		(0.003, 30)	31.65	120	107	106	82	t_{Trun}
		(0.005, 30)	23.84	176	155	149	133	t_{Trun}
		(0.010, 30)	17.69	311	262	258	265	t_{Model*}
		(0.030, 30)	11.72	679	554	509	633	t_{Model*}
		(0.050, 30)	9.65	906	706	628	862	t_{Model*}

Table 3.7. Comparison results of cut-off points on log-normal distribution

N	허용오차	(μ, σ^2)	왜도	$n(t)$	t_H	t_{Model*}	t_{Trun}	$t_{Optimal}$
10,000	0.03	$(10, 1^2)$	8.76	1,775	826	295	1,024	t_{Model*}
		$(10, 1.25^2)$	18.59	1,821	989	380	1,050	t_{Model*}
		$(10, 1.5^2)$	33.01	1,728	1,010	365	983	t_{Model*}
		$(10, 1.75^2)$	47.49	1,553	958	421	864	t_{Model*}
		$(10, 2^2)$	58.49	1,336	848	441	723	t_{Model*}
		$(10, 2.25^2)$	65.77	1,102	702	359	577	t_{Model*}
	0.05	$(10, 1^2)$	8.76	1,041	408	163	555	t_{Model*}
		$(10, 1.25^2)$	18.59	1,141	544	226	612	t_{Model*}
		$(10, 1.5^2)$	33.01	1,130	607	228	602	t_{Model*}
		$(10, 1.75^2)$	47.49	1,045	605	275	548	t_{Model*}
		$(10, 2^2)$	58.49	919	558	297	472	t_{Model*}
		$(10, 2.25^2)$	65.77	772	492	248	386	t_{Model*}

에서는 관심변수에 대한 모집단 자료가 존재하지 않기 때문에 관심변수와 가장 상관관계가 높다고 판단되는 모집단 자료를 이용하여 표본설계가 이루어진다.

지역 A는 모집단 수 N이 10,615이며, 왜도가 26.8938로 치우침이 심한 분포를 갖고 있다. 이 때 Hidiroglou 방법으로 산정 된 최적 표본크기는 허용오차가 3%일 때 1,536개이고, 허용오차가 5%일 때 1,077개로 나타났다.

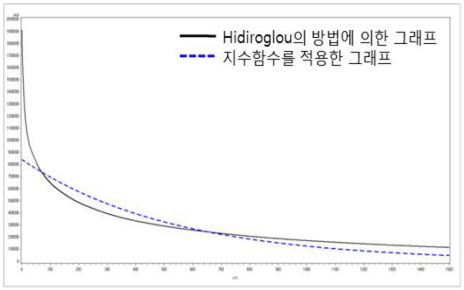
또한 지역 B는 모집단 수 N이 21,230이며, 왜도가 40.5653으로 역시 치우침이 심한 분포를 갖고 있는 것으로 나타났다. 이 경우 Hidiroglou 방법으로 계산된 최적 표본크기는 허용오차가 3%일 때 2,224개이고, 허용오차가 5%일 때 1,537개인 것으로 계산되었다. 마지막으로 지역 C는 모집단 수 N이 65,535이며, 왜도가 28.8098로 역시 치우침이 심한 분포를 갖고 있는 것으로 나타났다. 이 경우 Hidiroglou 방법으로 산정 된 최적 표본크기는 허용오차가 3%일 때 4,820개이고, 허용오차가 5%일 때 3,059개인 것으로 확인되었다. 정리된 결과는 Table 4.1에 수록하였다.

4.1. 사례연구방법

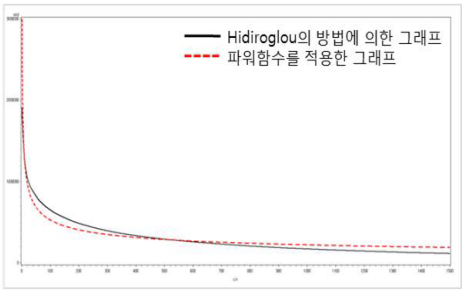
모의실험과 동일한 과정을 통해 세 가지 절사점 즉 기존 Hidiroglou 방법에 의해 산출 된 절사점 t_H, 함수에 의해 산출 된 절사점 t_{Model*}, 절단 분포에 의해 산출 된 절사점 t_{Trun}을 산출하였다. 그러나 모의

Table 4.1. Total sample size with skewness and population size

모집단크기 N	왜도	전체표본크기 $n(t)$	
		허용오차 = 0.03	허용오차 = 0.05
10,615(A)	26.8938	1,536	1,077
21,230(B)	40.5653	2,224	1,537
65,535(C)	28.8098	4,820	3,059



$Pseudo\ R^2 = 0.9438677$



$Pseudo\ R^2 = 0.963659$

Figure 4.1. Fitted results for region A

실험에서의 모집단 자료는 설정된 분포에서 생성된 자료인 반면, 사례연구의 자료는 모집단의 분포를 알지 못하기 때문에 제 2절에서 살펴본 것처럼 절단 분포에 의한 절사점 산출 시 분포에 대한 적합도 검정을 실시하였다. 본 연구에서는 SAS의 UNIVARIATE Procedure를 이용하여 감마분포와 로그정규분포의 P-P plot과 적합도 검정 결과를 바탕으로 적합한 분포를 선택하였다.

4.2. 사례연구결과

모집단 크기가 10,615이고 왜도가 26.8938인 첫 번째 지역 A는 Hidiroglou 방법을 사용한 결과, 허용오차가 3%일 때 최적 전체표본크기가 1,536개이고 전수층의 표본 수는 992개로 나타났다. 또한 허용오차가 5%일 때에는 최적 전체표본크기가 1,077개이고 전수층의 표본 수는 653개인 것으로 나타났다. 또한 표본층의 분산 $S^2_{[N-t]}$에 함수를 적합해보면, 파워함수를 적합했을 경우의 Pseudo-R^2가 0.9636으로 더 크므로 지수함수보다 파워함수가 더 잘 적합된 것을 확인할 수 있다. 이는 x축이 전수층 크기 t이고 y축이 $S^2_{[N-t]}$라고 할 때의 그래프인 Figure 4.1을 살펴보면 쉽게 이해 할 수 있다. 따라서 t_{Power}를 t_{Model*}라고 하면, 허용오차가 3%일 때 기존의 절사점 $t_H = 992$에서 $t_{Model*} = 456$로, 허용오차가 5%일 때 기존의 절사점 $t_H = 653$에서 $t_{Model*} = 312$로 전수층의 크기가 줄어든 것을 확인할 수 있다.

절단 분포로부터 계산 된 표본층의 분산을 이용하기 위해서는 자료의 모집단 분포가 결정되어야 하며 이를 위해 감마분포와 로그정규분포에 대한 적합성 검정과 P-P Plot을 살펴보았다. 그 결과 감마분포보다는 로그정규분포에 더 적합함을 알 수 있었고, 지역 A에 대한 모집단 분포를 로그정규분포로 가정하는데 무리가 없었다. 따라서 평균 μ와 분산 σ^2을 추정한 후 로그정규분포의 절단 분포를 이용하여 산출한 절사점 t_{Trun}를 보면, 허용오차가 3%일 때 기존의 절사점 $t_H = 992$에서 $t_{Trun} = 845$로, 허용오차가 5%일 때 기존의 절사점 $t_H = 653$에서 $t_{Trun} = 563$으로 전수층의 크기가 줄어든 것을 알 수 있다. 따라서 지역 A에 대한 사례연구 결과 주어진 허용오차에서 Hidiroglou 방법을 이용하여 최소의 전체표본크기를 결정하고, 이 표본크기를 고정한 후 전수층의 크기를 최소로 하는 최적 절사점 $t_{Optimal}$은 t_{Model*}임을 알 수 있으며 Hidiroglou가 제안한 전수층 수, t_H에 비해 전수층 표본 수가 약 50% 축소

Table 4.2. Comparison results of cut-off points on case study

Region	N	왜도	허용오차	$n(t)$	t_H	t_{Model*}	t_{Trun}	$t_{Optimal}$
A	10,615	26.89	0.03	1,536	992	456	845	t_{Model*}
			0.05	1,077	653	312	563	t_{Model*}
B	21,230	40.56	0.03	2,224	1,449	636	1,163	t_{Model*}
			0.05	1,537	949	432	767	t_{Model*}
C	65,535	28.81	0.03	4,820	2,661	1,206	2,414	t_{Model*}
			0.05	3,059	1,624	753	1,457	t_{Model*}

되었다. 같은 방법으로 지역 B와 C를 분석하였으며 그 결과를 Table 4.2에 수록하였다. Table 4.2 결과를 살펴보면 다음과 같다. 먼저 지역 A는 본 연구에서 제안한 새로운 절사점을 사용한다면 허용오차가 0.03일 때에는 전수층 크기가 536개 감소하고, 허용오차가 0.05일 때에는 341개의 전수층 크기가 감소하는 것을 볼 수 있다. 지역 B 또한 본 연구에서 제안한 새로운 절사점 산정 방법을 사용한다면 허용오차가 0.03일 때에는 813개가, 허용오차가 0.05일 때에는 517개의 전수층 크기가 감소한다. 또한 지역 C에서도 같은 결과를 확인할 수 있다.

5. 결론

사업체 조사에서 대기업이 상당 부분을 이루고 있는 전수층 조사는 표본층 조사보다 상대적으로 더 어려운 것이 사실이다. 특히 무응답이 발생할 경우 전수층의 결측값을 대체하는 것이 표본층의 결측값을 대체하는 것보다 더욱 어렵다. 하지만 전수층에서의 대체는 표본층에서의 대체보다 추정결과에 큰 영향력을 주기 때문에 전수층에서의 대체는 매우 중요하다. 따라서 표본층의 분산 $S^2_{[N-t]}$을 모형화하고 평활화(smoothing)하여 최적 절사점 $t_{Optimal}$을 계산함으로써 MSE를 최소화함과 동시에 기존의 전수층 규모보다 감소된 전수층 규모를 결정하는 것은 매우 의미있고 중요하다. 본 연구에서는 두 함수, 지수함수와 파워함수가 사용되었으며 파워함수의 형태가 일찍 평평해 지기 때문에 전수층 크기가 많이 줄어든 것을 확인 할 수 있다. 반면 지수함수를 사용한 경우에는 세 방법이 일정 수준 일치하는 것을 확인할 수 있다. 이는 어떤 함수를 적용하느냐에 따라 전수층의 크기 변화가 크게 될 수 있음을 보여주는 결과이므로 향후 함수 선택에 관한 추가적인 연구가 필요하다.

References

Baillargeon, S. and Rivest L.-P. (2011). The construction of stratified designs in R with the package stratification, *Survey Methodology*, **37**, 53–65.

Han, G. (2004). Estimation of cut-off stratum in the highly skewed population, *Survey Research*, **5**, 93–101.

Hidiroglou, M. A. (1986). The construction of a self-representing stratum of large units in survey design, *The American Statistician*, **40**, 27–31.

Lavellee, P. and Hidiroglou, M. A. (1988). On the stratification of skewed population, *Survey Methodology*, **14**, 33–43.

Lee, K. (2009). Determining the sample size of take-all strata in modified cut-off sampling with nonresponses, Master degree Dissertation, Department of Applied and Information Statistics, Kyonggi University.

Lee, S. (2011). The cut-off point based on MSE in modified cut-off sampling, *Journal of The Korean Official Statistics*, **16**, 82–94.

Namkung, P. (2008). Sample design for materials and components industry trend survey, *Communications of the Korean Statistical Society*, **15**, 883–897.

Shin, K.-I. and Lee, S. E. (2013). Cut-off sampling for right skewed long tail distribution, *Proceedings of ISI 2013*, HongKong

절사표본에서 최적 절사점에 관한 연구

이상은[a] · 조민지[a] · 신기일[b,1]

[a]경기대학교 응용정보통계학과, [b]한국외국어대학교 통계학과

(2014년 4월 17일 접수, 2014년 5월 22일 수정, 2014년 6월 9일 채택)

요 약

상당수의 사업체 조사는 절사표본설계법을 사용하고 있다. 이는 절사표본설계법에서 얻은 전수층이 많은 정보를 포함하고 있어 전체 표본크기를 최소화 할 수 있는 장점이 있기 때문이다. 그러나 최근 전수층에 포함된 사업체들의 무응답률이 높아감에 따라 전수층이 가지고 있는 장점에 한계가 나타나고 있다. 이에 Lee (2011), Shin과 Lee (2013)는 표본설계 단계에서부터 주어진 허용오차를 만족하면서 전수층 규모를 최소화하는 연구를 실시하였다. 본 연구에서는 주어진 허용오차를 만족하고 Hidiroglou (1986)가 제안한 방법으로 산출 된 표본크기를 고정한 상태에서 표본층 분산에 알려진 함수를 적합하여 전수층 크기를 최소화하는 새로운 최적 절사점을 제안하였다. 또한 Hidiroglou (1986)와 Shin과 Lee (2013)가 제안한 절단분포를 이용한 방법과 본 연구에서 제안한 방법을 모의실험과 사례연구를 통해 비교하였다.

주요용어: 전수층, 절단로그정규분포, 절단감마분포, 응답률.

이 논문은 2014년 한국외국어대학교 교내연구비 지원을 받아 수행되었음.

[1]교신저자: (449-791) 경기도 용인시 처인구 모현면 외대로 81, 한국외국어대학교 통계학과.
E-mail: keyshin@hufs.ac.kr