Journal of the Korean
Data & Information Science Society
1998, Vol. 9 No. 2, pp. 337 ~ 344

신뢰성있는 지역통계 추정을 위한 제언

김달호 [1] · **김남희** [2]

요약

통계청에서 조사 발표한 '96가구소비실태조사의 지역결과 자료에 대한 통계적 신뢰도를 분석하며, 전국 혹은 넓은 지역을 대상으로 하는 조사에서 소지역 통계의 오차를 줄여 신뢰도를 높이는 추정방법을 제안한다.

주제어 : 소지역 통계, 소득추정, 신뢰도, 표본설계.

1. 서론

각종 조사의 표본설계시 우선적으로 고려되는 것은 허용오차이며, 허용오차의 설정범위는 중앙정부기관이 주체가 되는 조사에서는 전국을, 광역지방자치단체가 주체가 되는 조사에서는 해당 행정구역 전체를 대상으로 하게 된다. 그런데 이러한 조사의 결과로서 하부 행정단위까지 세분하여 소지역 통계를 작성하고자 할 경우, 소지역 추정치의 오차가 허용오차 범위를 벗어나는 경우가 종종 발생하게 된다. 더구나 이러한 소지역 통계를 검토 없이 받아들여 지역간 순위비교를 하거나 정책적인 판단을 내리는 경우 통계수치에 대한 분쟁을 불러올 수도 있다.

이러한 사례로서, 미국에서는 1980년 인구조사 결과에 의한 기금배분를 두고, 뉴욕주에서 Census Bureau를 상대로 인구조사의 과소집계(undercount)의 수정방법에 대해서 소송을 제기하여 법정논쟁으로 이어지기도 하였다. 이 법정소송 문제는 1990년 인구조사에서도 계속되고 있다. (Ghosh 와 Rao(1994) 참조.)

우리지역에서도 통계의 신뢰성에 관한 논란이 신문지상에 크게 보도된 바 있는데, 경상북도의 소득논쟁이 그것이다. 소득논쟁의 경위를 살펴보면 다음과 같다.

최근 통계청에서 '96년 기준 가구소비실태조사 결과를 공표 하였다. 그 공표내용중 15개 광역자치단체의 연간소득 추정액 부문에서 경상북도가 최하위로 발표된 데에 대하여 지역 정치권이 서로 다른 이견을 보이며, 통계의 신뢰도에 대한 의문을 제기함으로 소득논쟁의

[1](702-701) 대구광역시 북구 산격동 1370번지 경북대학교 자연과학대학 통계학과 조교수
[2](702-702) 대구광역시 북구 산격동 1370번지 경북대학교 자연과학대학 통계학과 박사과정

발단이 되었다. 통계청 발표 직후 자유민주연합 대구경북지부에서는 성명서에서 한나라당 소속 현 지사의 경제 실정(失政)을 신랄하게 비판하였고, 이에 맞서 한나라당 대구경북지부에서는 최근 경상북도가 자체적으로 조사하여 공표한 '97 경북인의 생활과 의식조사의 소득부문 원자료를 재분석한 결과로서 통계청 발표수치에 대한 의문을 제기하였다. 경상북도 가구당 소득을 둘러싼 두 정치권의 논쟁은 당분간 계속될 전망이다. 본문에서는 통계청과 경상북도가 자체적으로 실시한 조사를 비교하며, 또한 통계청이 발표한 지역소득 통계의 신뢰성 문제를 분석하여 보고자 한다.

2. 통계청과 경상북도의 소득조사방법 비교

통계청과 경상북도의 조사내용을 조사대상 및 대상기간, 조사방법 등을 비교하여 (표1)과 같이 정리하였다.

<표 1> 통계청과 경상북도의 소득조사 개요

구분	통계청	경상북도
조사명	'96가구소비실태조사	'97 경북인의 생활과 의식조사
조사 대상	전체 가구중 1인 및 농어가 제외	전체 가구
대상 기간	'95.12. 1 ~ '96.11.30	'96. 1. 1 ~ 12. 31
조사 기간	'96.12. 1 ~ '96.12. 7	'97.10. 6 ~ 10. 10
조사방법	자계식·타계식 면접조사	자계식·타계식 면접조사
표본 설계	층화·계통추출	층화·계통추출
표본조사구수	알려져있지 않음	1,520개 조사구
표본 가구수	전국 21,103가구(경북 924가구)	도내 10,640가구

두 조사는 조사대상 선정에서 차이가 남을 알 수 있다. 통계청의 조사는 농 어가를 제외한 2인 이상 거주가구의 연간 가구소득을 조사한 것이며, 경북도의 조사는 농어가 및 1인 가구까지 포함한 전체가구를 대상으로 한 것이다. 그러므로 경상북도의 10,640 표본가구중 통계청의 조사대상에 해당하는 가구를 추출하여 연간소득에 대한 재분석을 하였다. 분석에 이용된 가구는 다음 (표2)과 같이 추출되었다.

<표 2> 표본가구중 1인 및 농어가로 분류하는 기준

구분	분류기준	적용비율(모집단 비율)
1인 가구	가구원수가 1명인 가구	0.156(0.159)
농가	농업매출액이 있는 가구	0.3040 (0.3150)
어가	어업매출액이 있는 가구	0.0081 (0.0081)
제외 가구	1인 가구 혹은 농어업소득이 있는 가구	0.4630 (-)
조사대상 가구		0.5370 (-)

위와 같은 기준에 의해 조사대상외 가구를 분류한 결과, 1인 가구는 전체 표본의 15.6 %, 농가는 30.4 %, 어가는 0.81 %로 집계되었다. 위에서 설명된 분류방법의 타당성을 조사하기 위하여 경북도 전체의 가구특성을 표본집단의 가구특성과 비교한 결과, 1인 가구의 경우 모집단 비율과 표본집단에서의 비율차이는 0.3 %, 농가의 경우는 1.1 %, 어가의 경우는 모집단 비율과 표본집단의 비율이 동일하였다. 특히, 어가는 경북 동해안 연안지역(포항, 경주, 영덕, 울진, 울릉)에만 집중되어 있어, 지역별 분포 차이가 심한데도 불구하고 표본집단에서 어가의 비율이 모집단과 거의 일치하게 추출되었다는 것은 경상북도의 표본가구가 모집단을 대표하는데 무리가 없는 것으로 판단된다. 각 특성별 경상북도 모집단 특성치로는 농림부 농업총조사 결과와, 해양수산부의 어업총조사 및 통계청 인구주택총조사 결과를 이용하였다.

위와 같은 근거에 의하여 경상북도의 원자료에서 통계청 조사대상에 해당하는 5,717개의 기록값만으로 새로운 데이터세트(data set)를 구성한 후 재분석을 실시하였다.

3. 조사 결과의 비교

위와 같은 기준에 따라 새로운 데이터세트을 구성하여 통계청과 경상북도의 소득을 비교한 결과 (표3)과 같은 결과를 얻었다. 연간소득 추정에 이용된 공식은 다음과 같다.

$$\begin{aligned} \bar{y} &= \sum_{i:\text{시군}} w_i \bar{y}_i \\ VAR(\bar{y}) &= \sum_{i:\text{시군}} w_i^2 var(\bar{y}_i) \\ sd(\bar{y}) &= \sqrt{VAR(\bar{y})} \end{aligned}$$

여기에서 $\bar{y}$는 경상북도의 소득평균이며, $\bar{y}_i$는 경북도내 23개 시군별 소득평균을 가리킨다. 또한, $w_i = \frac{f_i n_i}{\sum_{i:\text{시군}} f_i n_i}$이며, f_i는 i지역의 추출률의 역수이며, n_i는 i지역내의 표본가구수이다.

<표 3> 통계청과 경상북도의 가구소득 결과 비교 (단위: 만원, %)

조사기관	연간소득 추정치	표준오차	변동계수
통계청	2,110	37*	1.42*
경상북도	2,335	29	1.2

* 통계청 추정치에 대한 표준오차 및 변동계수는 전국 추정치에 대한 것임.

통계청의 지역 소득추정치에 대해서는 오차가 공개되어 있지 않으므로 통계청 조사결과에 대해서는 경북지역 평균소득과 전국 평균소득에 대한 표준오차 및 변동계수를 실었다.

통계의 정밀성에 관한 정보는 공개하지 않고 단순 점추정치(point estimates)만 발표한 것은 통계 이용자들에게 잘못된 인식과 그릇된 판단을 가져올 수 있으므로 통계를 작성하는 기관에서는 주의를 기울여야 할 부분이라 생각된다.

다음으로 지역 평균소득에 대한 표준오차 $\hat{se}_i$를 추정해 보기로 한다.

(경우 1) 통계청 조사결과에 근거한 지역 표준오차 추정 :

전국과 지역의 가구당 소득분포의 분산은 동일하다고 가정해 보자. 즉, $\sigma^2 = \sigma_i^2$ 라고 하자. σ는 다음과 같이 추정된다.

$$\hat{\sigma} = se^* \cdot \sqrt{n}$$

여기에서 se^*는 전국을 대상으로 한 소득추정치의 표준오차이며, n은 전국 표본가구수 21,103이다. 가정에 따라 위에서 구해진 $\hat{\sigma}$을 σ_i의 추정치로 한다. 그러므로 지역의 소득 추정치의 표준오차 se_i는 다음과 같이 추정된다.

$$\hat{se}_i = \frac{\hat{\sigma}}{\sqrt{n_i}} = se^* \cdot \frac{\sqrt{n}}{\sqrt{n_i}}$$

여기에서 n_i는 시도별 표본가구수이다.

위의 추정식에 의해 경북지역 소득추정치의 표준오차 및 변동계수를 계산하면 다음 (표4)과 같이 주어진다.

<표 4> 경상북도 평균 가구소득의 추정된 표준오차 및 변동계수 I (단위: 만원, %)

조사기관	소득 추정치	표준오차	변동계수
통계청	2,110	177	8.4

위의 결과로서 경북지역 소득의 95 % 신뢰구간을 구해보면 (1763, 2457)이 된다. 결국 경북소득 2,110만원은 추정값의 오차가 커짐으로 지역소득 대표값으로서의 신뢰도가 떨어진다고 볼 수 있다.

통계청에서 발표된 지역별 소득에 대해서 위의 표준오차 추정방법으로 95 % 신뢰구간을 구하여 (표5)에서 제시하였다.

<표 5> 15개 시도별 연간소득에 대한 95 % 신뢰구간 (단위: 만원)

시도명(순위별)	소득추정값	표본가구수	추정된 표준오차	95 % 신뢰구간
서울	2,911	2,738	103	(2709, 3113)
부산	2,626	2,289	112	(2406, 2846)
충남	2,585	866	183	(2226, 2944)
광주	2,524	1,470	140	(2250, 2798)
제주	2,515	474	247	(2031, 2999)
경남	2,514	1,416	143	(2234, 2794)
대구	2,476	1,491	139	(2204, 2748)
대전	2,438	1,460	141	(2162, 2714)
인천	2,405	1,946	122	(2166, 2644)
경기	2,383	2,414	109	(2169, 2597)
전북	2,349	843	185	(1986, 2712)
전남	2,288	765	194	(1908, 2668)
충북	2,281	970	173	(1942, 2620)
강원	2,279	1,037	167	(1952, 2606)
경북	2,110	924	177	(1763, 2457)

각 지역간 소득격차가 유의한지에 대한 차이검정은 하지 않았지만 추정된 오차로 볼 때, 지역간 소득격차는 유의하지 않을 것으로 짐작된다.

(경우 2) 경상북도 조사결과에 근거한 지역 표준오차 추정 :

통계청과 경상북도는 표본가구수 뿐만 아니라 표본설계에서도 차이가 난다. 통계청에서는 전국을 24개 지역[3] 으로 층화하여 각 지역별로 산업별 종사율, 거처당 평균가구, 주택유형 등의 분류지표에 따라 조사구를 분류한 후, 분류지표 순으로 정렬된 조사구 명부에서 표본조사구를 계통 추출하는 방식을 취한다.

경상북도에서는 23개 시군별로 층화한 후 각 시군 내에서 평균 10 % 표본조사구를 계통추출하며, 10 % 표본조사구내에서 다시 표본가구를 계통 추출하였다.

그러므로 표본가구수에만 의존하여 두 조사의 가구당 소득추정치의 신뢰성을 비교하는 것은 무리가 있다. 그러나 두 조사의 비교를 위해 경상북도의 가구당 소득추정치의 표준오차를 단순임의추출을 가정하여 다시 추정하였다. 통계청 조사에서도 단순임의추출을 가정하였는 데, 이는 분류지표 순서에 의한 조사구 추출의 효과가 없음을 가정한 것이다.

[3]광역시 6개 지역, 9개도의 경우 각각 시부, 군부로 분류함

이러한 가정에 의해 두 조사는 조사방법에서 동일하며, 표본가구수에서만 차이가 난다. 단순임의추출을 가정한 경북의 $\hat{\sigma}_i$은 2,228[4] 만원이었으며, 통계청조사에서 경북지역 가구당 소득의 표준오차를 추정하기 위하여 $\hat{\sigma}_i$을 경북지역에 배정된 표본가구의 제곱근으로 나누어 보았다.

그 결과 계산된 경북지역 소득추정치의 평균값의 표준오차 및 변동계수는 다음 (표6)과 같이 주어진다.

<표 6> 경상북도 가구소득의 추정된 표준오차 및 변동계수 II (단위: 만원, %)

조 사 기 관	연간소득 추정치	표 준 편 차	변 동 계 수
통계청	2,110	73	3.5

위의 결과로서 경북지역 소득의 95 % 신뢰구간을 구해보면 (1967만원, 2253만원)이 된다. 이는 (표4) 의 결과와 비교하여 볼 때, 매우 향상된 결과이기는 하나, 이것 역시 추정값의 오차가 크다고 볼 수 있다.

위의 결과에 근거하여 경북지역의 표준오차가 전국범위 오차(=37만원)와 같은 수준이 되기 위한 최소표본가구수를 구해 보면 3,626가구가 되는데, 이는 통계청 경북지역 표본의 약 4배에 해당한다.

이러한 결과로 볼 때, 통계청의 지역소득 표준오차는 전국범위의 오차보다 훨씬 더 커지며, 따라서 지역소득 추정치로서 신뢰도가 떨어진다고 볼 수 있다. 더욱이 이러한 표준오차의 고려 없이 점추정치만으로 지역별 소득의 순위를 부여하는 등의 통계이용은 매우 위험한 것이라 할 수 있다.

4. 토 의

다음으로 통계청의 가구소득 표본설계에서 몇 가지 문제점을 들어 개선방법을 제시하고자 한다. 표본설계의 초기단계에서 가장 중요한 작업은 조사기준 시점의 모집단 상황을 최대한 잘 반영할 수 있는 추출틀(sampling frame)을 작성하는 일이다. 만일 설계시에 추출단위들이 실제의 모집단을 제대로 반영하지 못할 경우 처음부터 표본의 대표성에 문제가 생기게 되어, 비표본오차의 일종인 추출틀오차(frame error)가 발생하게 된다. 통계청 '96가구소비실태조사의 표본은 1990년 인구주택총조사 결과를 근거로 추출한 것이므로 '96년의 지역상황을 제대로 반영한다고 볼 수 없으며, 이러한 점에서 통계청의 추정량은 편의(bias)를 가지고 있다고 볼 수 있다.

이상의 결과로 볼 때, 전국단위의 각종 조사에서 소지역까지 추정량의 정도를 고려해 표본설계를 하는 것은 조사비용의 과다와 조사기간의 장기화와 조사표 관리에서 막대한 전문인력을 필요로 하는 등의 이유로 인해 사실상 불가능한 것이다. 또한, 표본추출틀에서 이

[4]층화추출에 의한 추정은 2,193만원 임

미 추출틀오차와 같은 비표본오차가 발생할 수도 있으므로 1회 조사결과만으로 모집단을 직접 추정하는 것은 큰 위험이 따른다.

그러므로 과거의 자료 혹은 추정하고자 하는 통계량과 관련성이 높은 다른 변수가 있다면, 이를 추정식에 포함하는 것이 바람직하다. 또한, 추정하고자 하는 소지역과 유사한 다른 소지역의 정보가 있다면 이들 지역으로부터 정보를 빌려옴으로 ("borrow strength") 소지역 추정치의 정도를 높일 수도 있을 것이다.

Fay 와 Herriot(1979)은 미국의 39,000여개의 군(county)지역의 1인당 소득을 추정하기 위하여 회귀모형에서 경험적 베이즈방법에 근거한 소지역 추정(small area estimation)을 제안하였으며, 센서스결과만을 이용하거나 행정자료만으로 추정한 추정량보다 오차가 더 작아짐을 보였다. 결국 Fay와 Herriot은 행정자료에 의한 1인당소득 회귀추정치와 소지역에 대한 census 결과의 가중평균을 구하여 소지역추정량을 제안하였다.

우리 지역에서도 지역소득추정을 위해 지역의 생산수준(GRDP)이나 재정자립도, 지역의 소비실태 조사결과 등을 고려한 소득 추정모형의 개발을 서둘러야 할 것이다.

참 고 문 헌

1. 농림부(1995). 농업총조사보고서
2. 박홍래(1992). 統計調査論, 영지문화사
3. 통계청(1995). 인구주택총조사보고서
4. 통계청(1996). 가구소비실태조사보고서
5. 통계청(1996). 표본개편 연구회 연구 보고
6. 해양수산부(1996). 어가경제통계
7. Fay, R. E. and Herriot, R. A. (1979). Estimates of income for small places: an application of James-Stein procedures to census data, *Journal of the American Statistical Association*, 74, 269-277.
8. Ghosh, M. and Rao, J. N. K. (1994). Small area estimation: an appraisal, *Statistical Science*, 9, 55-93.

A Suggestion of Reliable Etimation for the Local Area Statistics

Dal Ho Kim [5] and Nam Hee Kim [6]

Abstract

We analyze statistical reliability about local area results of '96 Survey of Family Consumption Behavior by the National Statistical Office. Also, we suggest reliable estimation procedures for reducing statistical errors in local ares estimation problem.

[5]Assistant Professor, Department of Statistics, Kyungpook National University, Taegu, 702-701, Korea.
[6]Department of Statistics, Kyungpook National University, Taegu, 702-701, Korea.