한국경제포럼 제15권 제1호 61-96
DOI : 10.22841/kefdoi.2022.15.1.003

『가계동향조사』의 문제와 보정*

김 낙 년**

논문초록

가계동향조사는 매달 표본의 일부가 교체되는 연동표본으로 설계되어 있기 때문에 각 표본이 1년 중에 조사되는 월수는 1-12달(2020년 이후는 1-6달)에 걸쳐 있게 된다. 통계청은 분기 또는 연간 통계를 구할 때 월간 자료에서 해당 월의 조사결과를 단순 평균하는 방식으로 산출하고 있다. 매달 지출이 반복되는 품목이라면 이 방식에 문제가 없지만 그렇지 않은 품목에서는 왜곡이 발생한다. 조사월수와 품목의 구입 빈도에 따라 연간 통계는 실제보다 12배(분기 통계는 3배)까지 과대해지는 편향이 생기는 한편, 조사 월에 구입되지 않은 품목은 조사에서 아예 누락되어 버린다. 전체 평균을 구할 때에는 이러한 과대평가와 누락에 의한 과소평가가 서로 상쇄되는 것으로 나왔다. 그렇지만 가구당 평균은 그렇지 않아 가구 간 소비 격차가 실제 조사된 결과보다 더 벌어진 것으로 나오며, 그러한 왜곡이 최근에 더욱 커졌다. 본고는 2015-16, 2019-20년의 연간 자료에 포함된 각 가구의 품목별 소비의 과대 또는 과소 편향을 합리적 방법으로 보정하였고, 그 결과를 기존의 마이크로 데이터와 비교하여 어디에서 얼마나 왜곡이 나타나는지를 보였다.

핵심 주제어: 가계동향조사, 연동표본, 소비분포
경제학문헌목록 주제분류: D0, D3, R2

투고 일자: 2021. 12. 10. 심사 및 수정 일자: 2022. 1. 7. 게재 확정 일자: 2022. 2. 15.
* 이 연구는 2018년도 대한민국 교육부와 한국학중앙연구원(한국학진흥사업단)을 통해 한국학 세계화 랩 사업의 지원을 받아 수행되었다(AKS-2018-LAB-1250002).
** 동국대학교(서울캠퍼스) 경제학과 명예교수, e-mail: nnkim@dongguk.edu

Ⅰ. 머리말

가계동향조사는 오랜 역사를 가지고 있고 통계청을 대표하는 통계조사라 할 수 있다. 피 조사가구가 매월 소득과 소비에 관한 가계부를 작성하고, 이를 취합하여 분기별 및 연간 통계를 발표하고 있어서 시의성이 요구되는 정책 판단에 널리 이용되고 있다. 특히 가계의 소비지출 통계는 350개가 넘는 소비품목을 매달 조사하는 등 다른 조사로 대체될 수 없는 독보적인 위치를 차지하고 있으며, 그 결과인 품목별 소비지출 통계는 소비자물가지수를 작성할 때 가중치로 이용되고 있다. 그 외에도 가계동향조사의 마이크로 데이터는 많은 연구논문에서 기초통계로 이용되고 있다.

가계동향조사의 소득 항목에 관해서는 과소평가 문제가 제기된 바 있고(예컨대 김낙년 · 김종일, 2013), 통계청은 가계금융복지조사를 대상으로 행정자료로 보정하는 방식으로 이 문제에 대처하고 있다. 다만 연간 통계인 가계금융복지조사는 정확성이 높아진 반면, 조사 시점과 공표 시점의 시차가 커서 시의성을 갖기 어렵다. 따라서 시의성이 요구되는 경우 여전히 가계동향조사의 분기별 조사가 이용되고 있다. 그런데 가계동향조사의 분기 및 연간 소비지출 통계는 월간 자료로부터 산출할 때 적용한 방법이 적절치 않아 실제의 조사결과와 상당히 괴리되어 버렸다는 문제를 제기하고자 한다.

가계동향조사가 후술하듯이 2016년까지는 36개월 연동표본 방식(2019년 이후에는 6-6-6 연동표본)으로 설계되어 있다. 표본을 36개 그룹으로 나누어 매달 표본의 1/36이 교체된다. 36개월을 연속해서 조사한 그룹은 표본에서 빠지고 그를 대신하여 새로운 그룹이 표본에 추가되는 방식이다. 그로 인해 특정 해에 조사된 것으로 한정하면 표본 중에는 12개월 모두 조사된 그룹도 있지만, 1-11개월 조사된 그룹도 존재하게 된다. 통계청이 이들 월간 자료를 연간(또는 분기) 자료로 전환할 때에는 해당하는 월의 조사 결과의 단순평균을 연간(또는 분기[1]) 데이터로 발표하고 있다. 예컨대 연간 자료의 경우 n개의 월이 조사되었다면 n개 월의 단순 평균값을 해당 해의 평균값으로 제시하는 것이다.[2]

---

1) 여기서는 연간 자료의 문제를 중심으로 설명하지만, 분기 자료에 대해서도 타당하다.

2) 통계청이 월 조사 자료에서 분기 또는 연간 통계를 작성하는 방법에 따르면, 소득이나 지출과 같은 금액 통계의 경우 가구별로 조사된 월 자료의 단순 산술평균을 취하고 있다. 이에 대해 가구원 수나 가구주의 산업 또는 직업과 같은 가구 특성 변수의 경우에는 해당 기간(분기나 연간) 중에서 빈도수가 가장 큰 월을 대표 월로 선정하여 그 월의 통계를 싣고 있다. 예컨대 조사월수=7인데, 가구원 수가 3으로 조사된 월이 2개이고 4로 조사된 월이 5개인 경우 가구원 수를 4로

그런데 이러한 방식은 매월 지출이 이루어지는 경우에는 문제가 되지 않을 수 있지만, 내구재와 같이 지출이 빈번하게 일어나지 않는 경우에는 그렇지 않다. 예컨대 자동차(구입가격 2400만원)를 구입한 가구를 예로 들면, 구입한 월에는 2400만원 지출된 것으로 기입되지만 나머지 달의 지출은 0이 될 것이다. 만약 이 가구가 12달 모두 조사되었다면 연간 자료에서 이 가구의 자동차 구입의 월 평균 지출은 200만원(=2400/12)이 된다. 그런데 만약 이 가구가 1달만 조사되었고, 그 달이 우연히 자동차를 구입한 달이라고 해보자. 이 경우 이 가구의 자동차 구입의 월 평균 지출은 통계청의 방식에 따르면 2400만원(=2400/1)이 되어 실제 지출의 12배가 된다. 조사월수가 2, 3달로 늘어나면 월 평균 지출은 1200만원(=2400/2), 800만원(=2400/3)이 되어 실제 지출액의 6배와 4배가 된다. 이처럼 실제 지출액의 최대 12배까지 과대하게 나타난 것은 조사된 월을 단순 평균하는 방식으로는 조사되지 않은 다른 달의 자동차 구입 지출이 0이라는 사실이 반영되지 않기 때문이다.

실제 가계동향조사 연간 자료(마이크로 데이터)에서 자동차(신차구입)의 소비지출이 어떻게 나와 있을까? 〈표 1〉은 2015년의 신차 구입액의 조사월수별 데이터를 정리한 것이다. 전체 표본인 9,709가구에서 자동차를 구입한 것은 223가구인데, 그들의 자동차 월 평균지출을 보면 조사월수에 따라 차이가 크다. 12개월 모두 조사된 자동차의 월평균 지출액은 212만원인데, 이를 기준으로 조사월수에 따른 지출액의 배율을 구해보면 조사월수가 1개월인 경우는 15.8배로 높게 나왔다. 그리고 자동차 사례처럼 구입한 달 이외에는 지출이 0인 경우 통계청의 방식으로 조사월수에 따른 월 평균 지출을 구해 조사월수=12를 기준(=1)으로 제시하면 〈표 1〉의 '비교' 항목과 같다. 양자를 비교하면 차이가 보이는데, 이것은 구입한 자동차의 단가가 조사월수에 따라 균일하지 않기 때문이다. 조사월수가 12개월 미만인 표본수가 적은 것도 한 요인인데 만약 표본수가 더 많았다면 그 배율은 〈표 1〉의 '비교' 항목에 접근하였을 것으로 생각된다. 연동표본 방식에서 조사월수가 몇 개월이 되는가는 우연히 정해지는 것인데, 조사월수에 따른 평균값이 이렇게 큰 차이를 보인 것은 연간 자료를 그대로 이용할 경우 심각한 왜곡이 나타날 수 있음을 시사하고 있다.

조사된 월을 대표 월로 선정한다는 뜻이다. 가구원 수의 평균값을 취할 경우 소수점을 갖게 되는 것을 피하기 위한 것이다. 통계청 홈페이지의 통계설명자료(http://meta.narastat.kr/)에서 "가계동향조사 분기 및 연간통계 작성 방법"을 참조하기 바란다.

〈표 1〉 2015년의 자동차(신차 구입)의 조사월수별 표본 수와 월평균지출

조사 월수	표본수			월 평균지출		
	전체 a	신차구입 b	b/a(%)	천원	배율	비교
1	557	3	0.54	33,629	15.8	12
2	929	3	0.32	8,797	4.1	6
3	666	5	0.75	18,615	8.8	4
4	287	2	0.70	7,883	3.7	3
5	456	7	1.54	3,513	1.7	2.4
6	788	10	1.27	5,801	2.7	2
7	585	16	2.74	3,553	1.7	1.714
8	217	-	-	-	-	1.5
9	407	8	1.97	2,519	1.2	1.333
10	763	25	3.28	2,497	1.2	1.2
11	560	15	2.68	1,950	0.9	1.091
12	3,494	129	3.69	2,123	1	1
total	9,709	223	2.30	2,535		

주: 1) 배율이란 조사월수=12개월인 표본의 월 평균지출을 1로 본 배율이다.
2) 비교(=12/조사월수)란 통계청의 방식으로 구한 배율(조사월수=12를 기준)이다. 신차를 1년에 2번 이상 구입한 경우는 없는 것으로 가정한다.

자료: 통계청, 가계동향조사 마이크로 데이터(RDC로 접근).

자동차의 사례를 이용하여 연간 자료가 갖고 있는 또 다른 편향을 지적할 수 있다. 만약 이 가구에서 조사된 월이 앞의 사례와는 거꾸로 우연히 자동차 구입이 이루어지지 않은 달이라고 해 보자. 그 경우는 실제로는 자동차 구입이 이루어졌음에도 불구하고 자동차의 구입이 없었던 것으로 간주되게 된다. 지출이 빈번하지 않은 내구재의 경우 12개월이 모두 조사되지 않은 한 내구재 소비가 과소하게 파악될 수 있음을 알 수 있다. 다시 〈표 1〉에서 전체 표본 중에서 자동차를 구입한 표본의 비율(=b/a)을 보면 12개월 조사가 이루어진 경우는 3.69%로 나왔지만, 12개월 미만인 경우는 그보다 낮고 조사월수가 적을수록 그 비율도 낮아지는 경향을 보이고 있다. 이것은 조사월수가 12개월 미만인 경우 실제 자동차를 구입했음에도 불구하고 자동차 구입이 없었던 것으로 파악된 표본이 적지 않음을 보여준다. 이것은 자동차 구입의 사례를 든 것이지만, 후술하는 바와 같이 매달 구매가 이루어지는 품목은 오히려 소수이기 때문에 대부분의 품목이 크건 작건 이 사례와 유사한 문제를 안게 된다.

위에서 지적한 두 가지 문제 중에서 앞의 경우는 12개월 모두 조사되지 않은 가구

의 소비를 최대 12배까지 과대평가하는 편향이 있다고 한다면, 후자의 경우는 반대로 소비지출을 과소평가하는 편향을 갖게 된다고 할 수 있다. 이하에서는 전자를 편향A, 후자를 편향B라고 구분해서 지칭하기로 한다. 두 편향이 반대 방향으로 작용하므로 전체 평균을 보면 양자가 서로 상쇄될 수 있다. 그렇지만 연간 자료에서 개별 가구의 소비 구성을 보면 일부 품목은 12배까지 부풀려져 있는 반면 다른 품목은 소비되었음에도 불구하고 아예 누락되어 실제와 달라진다. 가구별 소비액도 후술하듯이 실제의 조사결과와 괴리되기 때문에 가구 간 소비 격차도 왜곡되어 버린다.

현재 가계동향조사의 연간(또는 분기) 자료(마이크로 데이터)에서 각 가구의 품목별 소비지출 통계에는 위의 편향A와 편향B가 포함되어 실태와 상당히 괴리되어 있다. 이하의 제Ⅱ절에서는 36개월 연동표본으로 설계된 2015-16년과 6-6-6 연동표분으로 설계된 2019-20년을 대상으로 이러한 편향이 어떻게 나타나고 있는지, 그리고 이를 바로잡을 수 있는 합리적인 방법이 무엇인지를 모색한다. 제Ⅲ절에서는 이렇게 보정된 결과에 의거하여 구한 가구의 소비 분포를 보정 전과 비교하여 어디에서 얼마나 차이가 생겼는지를 보이고자 한다. 이 조사가 정책과 연구에 널리 활용되고 있는 기초 통계라는 점을 감안할 때 통계청이 현재의 방식으로 제공하고 있는 가계동향조사의 연간(또는 분기) 통계는 개편될 필요가 있다. 제Ⅳ절에서는 본고가 제기한 문제가 가계동향조사에 의거한 기존 연구에 대해 어떤 함의를 갖는지 간단히 언급하기로 한다. 제Ⅴ절에서는 본고에서 밝혀진 사실을 요약하고, 이 문제에 대해 통계청에 어떠한 대응이 요구되는지를 언급한다.

Ⅱ. 가구별 소비지출의 보정

1. 연동표본과 연간 자료의 문제

앞에서 제기한 문제는 가계동향조사가 연동표본으로 설계되어 조사월수가 12개월 미만인 가구가 포함되어 있는데 기인한다. 통계청은 2005년부터 연동표본을 도입하였는데, 거기에는 나름의 이유가 있었다. 그 전에는 가구조사 응답자가 5년간 가계부를 작성하는 부담이 컸고, 표본을 개편할 때마다 전면적인 표본교체가 이루어져 시계열의 단절이 발생했다. 연동표본의 도입은 응답자의 부담을 덜 수 있고 모집단의 변화를 반영하기 용이하고 조사원들의 매너리즘을 예방하는 효과를 기대할 수 있다(윤연옥 외, 2004).

〈표 2〉 36개월 연동표본의 예시

		표본 그룹																																	
년	월	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36
t년	1	1	a	b	c	d	e	f	g	h	i	j	k	l	m	n	o	p	q	r	s	t	u	v	w	25	26	27	28	29	30	31	32	33	34	35	36
	2	1	2	b	c	d	e	f	g	h	i	j	k	l	m	n	o	p	q	r	s	t	u	v	w	25	26	27	28	29	30	31	32	33	34	35	36
	3	1	2	3	c	d	e	f	g	h	i	j	k	l	m	n	o	p	q	r	s	t	u	v	w	25	26	27	28	29	30	31	32	33	34	35	36
	4	1	2	3	4	d	e	f	g	h	i	j	k	l	m	n	o	p	q	r	s	t	u	v	w	25	26	27	28	29	30	31	32	33	34	35	36
	5	1	2	3	4	5	e	f	g	h	i	j	k	l	m	n	o	p	q	r	s	t	u	v	w	25	26	27	28	29	30	31	32	33	34	35	36
	6	1	2	3	4	5	6	f	g	h	i	j	k	l	m	n	o	p	q	r	s	t	u	v	w	25	26	27	28	29	30	31	32	33	34	35	36
	7	1	2	3	4	5	6	7	g	h	i	j	k	l	m	n	o	p	q	r	s	t	u	v	w	25	26	27	28	29	30	31	32	33	34	35	36
	8	1	2	3	4	5	6	7	8	h	i	j	k	l	m	n	o	p	q	r	s	t	u	v	w	25	26	27	28	29	30	31	32	33	34	35	36
	9	1	2	3	4	5	6	7	8	9	i	j	k	l	m	n	o	p	q	r	s	t	u	v	w	25	26	27	28	29	30	31	32	33	34	35	36
	10	1	2	3	4	5	6	7	8	9	10	j	k	l	m	n	o	p	q	r	s	t	u	v	w	25	26	27	28	29	30	31	32	33	34	35	36
	11	1	2	3	4	5	6	7	8	9	10	11	k	l	m	n	o	p	q	r	s	t	u	v	w	25	26	27	28	29	30	31	32	33	34	35	36
	12	1	2	3	4	5	6	7	8	9	10	11	12	l	m	n	o	p	q	r	s	t	u	v	w	25	26	27	28	29	30	31	32	33	34	35	36
t+1 년	1	1	2	3	4	5	6	7	8	9	10	11	12	13	m	n	o	p	q	r	s	t	u	v	w	25	26	27	28	29	30	31	32	33	34	35	36
	2	1	2	3	4	5	6	7	8	9	10	11	12	13	14	n	o	p	q	r	s	t	u	v	w	25	26	27	28	29	30	31	32	33	34	35	36
	3	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	o	p	q	r	s	t	u	v	w	25	26	27	28	29	30	31	32	33	34	35	36
	4	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	p	q	r	s	t	u	v	w	25	26	27	28	29	30	31	32	33	34	35	36
	5	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	q	r	s	t	u	v	w	25	26	27	28	29	30	31	32	33	34	35	36
	6	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	r	s	t	u	v	w	25	26	27	28	29	30	31	32	33	34	35	36
	7	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	s	t	u	v	w	25	26	27	28	29	30	31	32	33	34	35	36
	8	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	t	u	v	w	25	26	27	28	29	30	31	32	33	34	35	36
	9	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	u	v	w	25	26	27	28	29	30	31	32	33	34	35	36
	10	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	v	w	25	26	27	28	29	30	31	32	33	34	35	36
	11	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	w	25	26	27	28	29	30	31	32	33	34	35	36
	12	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36

주: 동일한 숫자(또는 알파벳)은 동일한 표본 그룹을 뜻한다. 예컨대 c로 표시된 표본 그룹은 t년의 3월로 조사가 종료되었고, 4월부터는 4로 표시된 표본 그룹의 조사가 개시되었음을 뜻한다.

〈표 3〉 6-6-6 연동표본의 예시

		표본 그룹																						
년	월	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24
t년	1	1							a	b	c	d	e	f							m	n	o	p	q
	2	1	2							b	c	d	e	f	g							n	o	p	q
	3	1	2	3							c	d	e	f	g	h							o	p	q
	4	1	2	3	4							d	e	f	g	h	i							p	q
	5	1	2	3	4	5							e	f	g	h	i	j							q
	6	1	2	3	4	5	6							f	g	h	i	j	k						
	7		2	3	4	5	6	7							g	h	i	j	k	l					
	8			3	4	5	6	7	8							h	i	j	k	l	m				
	9				4	5	6	7	8	9							i	j	k	l	m	n			
	10					5	6	7	8	9	10							j	k	l	m	n	o		
	11						6	7	8	9	10	11							k	l	m	n	o	p	
	12							7	8	9	10	11	12							l	m	n	o	p	q
t+1 년	1	1							8	9	10	11	12	13							m	n	o	p	q
	2	1	2							9	10	11	12	13	14							n	o	p	q
	3	1	2	3							10	11	12	13	14	15							o	p	q
	4	1	2	3	4							11	12	13	14	15	16							p	q
	5	1	2	3	4	5							12	13	14	15	16	17							q
	6	1	2	3	4	5	6							13	14	15	16	17	18						
	7		2	3	4	5	6	7							14	15	16	17	18	19					
	8			3	4	5	6	7	8							15	16	17	18	19	20				
	9				4	5	6	7	8	9							16	17	18	19	20	21			
	10					5	6	7	8	9	10							17	18	19	20	21	22		
	11						6	7	8	9	10	11							18	19	20	21	22	23	
	12							7	8	9	10	11	12							19	20	21	22	23	24

주: 동일한 숫자(또는 알파벳)은 동일한 표본 그룹을 뜻한다. 예컨대 1로 표시된 표본 그룹은 t년의 1-6월에 조사되고, 6개월 쉰 후 t+1년의 1-6월에 다시 조사가 이루어짐을 뜻한다. 알파벳으로 표시된 표본 그룹은 t년 이전부터 조사가 시작되었으며, 예컨대 a로 표시된 표본 그룹은 t년의 1월에 조사가 종료되었음을 보여준다.

먼저 2016년까지 시행된 36개월 연동표본을 예시하면 〈표 2〉와 같다. 거기에서 동일한 숫자(또는 알파벳)는 동일한 표본 그룹을 뜻한다. 예컨대 1로 표시된 표본 그룹은 t년의 1월부터 조사가 시작되어 3년간 조사된 후 표본에서 사라지게 되며, 2로 표시된 표본 그룹은 1달 늦게 시작하게 된다. 그리고 알파벳으로 표시한 표본 그룹은 t년 이전부터 조사되기 시작하여 예컨대 표본 그룹 a는 t년 1월에, b는 2월에 조사가 종료되었을 뜻한다. 이와 같은 방식으로 전체 표본의 1/36이 매달 교체됨을 알 수 있다. 〈표 2〉에서 t년에 한정해서 보면, 표본 그룹 13에서 36까지는 동일한 표본이 12개월 모두 조사되었지만, 1-12의 표본 그룹에서는 그렇지 못해 조사월수가 1달에서 11월까지 분포되어 있다.[3] t+1년의 경우에도 마찬가지임을 알 수 있다.

2019년 이후에는 가계동향조사의 표본의 설계가 6-6-6연동표본 방식으로 바뀌었다. 표본을 6개월간 연속해서 조사한 후 6개월은 쉬었다가 다시 6개월을 조사한 후 표본에서 빼는 방식이다. 응답자의 조사 부담을 기존의 36개월에서 12개월로 크게 줄였다. 〈표 3〉은 6-6-6 연동표본 방식을 예시하였다. 〈표 2〉와 마찬가지로 동일한 숫자(또는 알파벳)는 동일한 표본 그룹을 뜻하는데, 1로 표시된 표본 그룹을 보면 t년의 1-6월과 t+1년의 1-6월이 조사되기 때문에 동일한 표본으로 전년 동기를 비교할 수 있게 설계되어 있다. 다만 t년으로 한정해서 보면, 1-7또는 f-q로 표시된 표본 그룹은 6개월이 모두 조사되었지만, 8-12 또는 a-e로 표시된 표본 그룹은 1개월에서 5개월까지 조사된 것을 알 수 있다. 이러한 구조는 t+1년의 경우에도 마찬가지다. 〈표 2〉의 36개월 연동표본과 비교하면 모든 표본의 조사월수가 동일하지 않다는 점은 마찬가지지만, 6-6-6 방식에서는 12개월이 온전히 조사된 표본이 없다는 점이 다르다. 후술하듯이 조사월수가 적을 때 생기는 편향은 12개월 모두 조사된 표본을 기준으로 보정할 수 있는데, 이 점에서 6-6-6 방식은 그 편향을 보정하는데 한계가 있다.

그러면 실제로 각 연도의 조사월수에 따라 표본수가 어떻게 분포되어 있을까? 〈표 4〉에 따르면 36개월 연동표본을 적용한 2015-16년의 경우 12개월이 모두 조사된 것은 전체 표본의 36%와 46%로 나온다. 조사월수별 표본수가 일정하지는 않다. 6-6-6 연동표본을 적용한 2020년은 조사월수=6인 표본이 55%를 차지하고 있고 나머지는 조사월수에 따라 비슷하게 분포하고 있다. 다만 2019년은 6-6-6연동 방식을 처음 도입하면서 조사의 안정화를 위해 처음 6개월(즉 2019년 1-6월) 간은 연동교체를

3) 〈표 2〉의 36개월 연동표본은 경제활동인구조사에 적용된 것이고, 가계동향조사에서는 조사 그룹의 수가 달라 〈표 2〉와 다소 차이가 있다. 그렇지만 본고가 문제로 삼고 있는 조사월수가 12개월 미만인 표본이 생긴다는 점은 마찬가지다.

적용하지 않았다. 그로 인해 2019년의 표본에는 6-6-6연동 방식임에도 불구하고 최대 11개월까지 조사된 경우가 포함되어 있다.

〈표 4〉 조사월수별 표본 수와 평균 가중치

조사월수	2015		2016		2019		2020	
	표본수	가중치	표본수	가중치	표본수	가중치	표본수	가중치
1	557	229	188	268	1,029	393	1,191	390
2	929	478	641	489	1,007	797	1,109	769
3	666	663	599	712	885	1,159	1,111	1,126
4	287	982	278	946	927	1,559	1,092	1,509
5	456	1,163	505	1,179	867	1,972	1,015	1,852
6	788	1,368	840	1,398	1,491	2,225	6,624	2,121
7	585	1,703	539	1,616	722	2,738		
8	217	1,906	215	2,129	718	3,169		
9	407	1,994	372	2,141	629	3,472		
10	763	2,230	407	2,312	640	3,713		
11	560	2,380	257	2,903	595	3,955		
12	3,494	2,628	4,106	2,674				
합계/평균	9,709	1,786	8,947	1,969	9,510	2,090	12,122	1,659
평균 조사월수		8.0		8.6		5.5		4.6

자료: 통계청, 가계동향조사 마이크로 데이터(RDC로 접근).

그런데 연동표본에서 월별로 조사된 데이터를 연간(또는 분기) 자료로 바꾸어 줄 때 12개월 모두 조사된 경우는 문제가 없지만, 조사월수가 적은 경우는 조사된 월의 평균이 그 해(또는 그 분기)를 대표한다는 보장이 없다는 전술한 문제가 제기된다. 통계청은 조사된 월의 평균을 그 해(또는 분기)를 대표하는 것으로 보되, 조사월수가 적으면 그에 비례해서 가중치를 낮추어 준다는 점을 강조하고 있다. 〈표 4〉는 조사월수별로 해당 표본의 평균 가중치를 보여주는데, 통계청은 조사월수에 대체로 비례해서 가중치를 부여하고 있음을 알 수 있다. 그 결과 조사월수가 적을수록 편향A(또는 편향B)가 커지지만 적용되는 가중치가 줄어들기 때문에 그로 인해 통계치가 왜곡되는 정도는 그만큼 줄어든다. 그렇지만 조사월수가 적은 가구의 가중치가 0이 되는 것은 아니므로 이러한 편향으로 인한 통계의 왜곡이 없어지지는 않는다.[4] 이하 본고에

4) 〈표 1〉의 신차 구입액에 〈표 4〉의 가중치를 적용하면 그 분포가 어떻게 달라질까? 만약 12개월 모두 조사된 가구만을 대상(이 경우는 편향A와 B가 생기지 않음)으로 구한 신차 구입액 분포의 불평등도를 G_x라고 하고, 조사월수가 12개월 미만(따라서 편향A와 B가 생김)까지 포함한 전체 가구를 대상으로 구한 신차 구입액 분포의 불평등도를 G_y라고 하자. 이 때 만약 조사월수별 가

서는 통계청의 가중치를 그대로 이용하기로 한다.

가계동향조사의 연간 자료에서 12개월 모두 조사된 가구와 12개월 미만으로 조사된 나머지 가구의 두 그룹으로 나눈 다음, 품목별로 두 그룹의 소비지출 평균이 다르지 않다는 귀무가설이 기각되는지 여부를 검증해 볼 수 있다. 전체 소비품목 352개 중에서 이 귀무가설이 유의수준 5% 미만으로 기각되는(즉 통계적으로 다른 것으로 나온) 품목이 2015년에 313개에 달하는 것으로 나왔다(후술하는 〈표 8〉 참조). 각 표본이 어느 그룹에 속하는지는 우연에 따른 것임을 감안하면 이 검증 결과는 조사월수에 따른 편향이 일부 품목에 한정된 것이 아니라 대부분의 품목에서 광범하게 나타나고 있음을 보여준다.

2. 편향A: 조사월수에 따른 평균값의 과대평가

여기서는 소비 품목별로 조사월수가 12개월에 미치지 못한 경우 통계청 방식에 따른 각 품목의 조사된 월의 평균값이 실제의 평균에 비해 얼마나 과대평가되는지(즉 얼마를 보정해 주어야 하는지)를 추정해 보기로 한다. 앞에서 든 자동차의 사례와 달리 만약 매달 규칙적으로 구입되는 품목이라면 조사월수가 적더라도 조사된 월의 평균을 해당 해(또는 분기)의 평균으로 보아도 큰 문제는 없다. 즉 이러한 과대평가의 정도는 구입빈도에 따라 영향을 받기 때문에 각 품목이 1년에 몇 번 구입되는지에 관한 정보가 필요하다.

여기서 이용한 자료는 가계동향조사의 마이크로 데이터이지만, 통상 MDIS에서 다운로드 받을 수 있는 것보다 상세한 정보를 담고 있는 것으로서 통계청의 이용센터 서비스(RDC)로 접근한 것이다. 각 표본이 몇 번 조사되었는지(즉 조사월수)에 관한 정보는 이번에 통계청에 요청하여 이용할 수 있게 되었다.[5] 이와 함께 각 품목의 구

중치의 차이를 두지 않고 분포를 구했다고 하면, 당연히 $G_x < G_y$가 된다. 그런데 〈표 4〉와 같이 조사월수별로 차등한 가중치를 적용하고 조사월수가 12개월 미만을 포함한 모든 가구를 대상으로 구한 신차 구입액 분포의 불평등도를 G_z라고 하자. 그러면 당연히 $G_z < G_y$가 될 것이다. 그렇지만 G_z가 편향을 반영하는 가중치가 낮아졌다고 해도 편향이 포함되지 않은 G_x보다는 높을 것이다. 즉 $G_x < G_z < G_y$가 된다.

5) 통계청(MDIS)이 제공하는 가계동향조사의 마이크로 데이터는 3가지 방식으로 이루어진다. 첫째, 해당 사이트에서 누구나 다운로드 받을 수 있는 것인데 소비지출 품목 수에 제한이 있다. 둘째, 분기와 연간 자료의 세부 품목까지 알 수 있는 데이터는 원격접근(RAS)을 통해 가능하다. 셋째, 월간 자료에 대한 접근은 통계청의 이용센터(RDC)에 직접 가서 이용할 수 있다. 이

입빈도에 관한 정보는 가계동향조사의 월간 자료에서 얻을 수 있는데, 이 또한 RDC에서만 접근 가능하다. 거기에는 각 표본의 월별 소비지출에 관한 정보가 담겨 있기 때문에 이를 이용하면 각 가구별로 해당 품목의 구입이 이루어진 월 수를 구할 수 있다.[6] 물론 가구에 따라 구입빈도에 차이가 있겠지만 그 평균을 해당 품목의 구입빈도로 보았다.

그 결과를 〈표 5〉로 제시하였다. 월간 자료에서 12개월 모두 조사된 가구를 대상으로 하여 각 품목이 한번 이상 구입된 경우의 그 평균 월수(즉 구입빈도)를 구한 것이다.[7] 가구별 평균이므로 소수점 이하 값을 반올림하였다. 그에 따르면 구입 빈도가 5개월 이하가 91%(2개월 이하는 66%)로 나와 매달 구입되는 품목이 의외로 적다.[8] 내구재의 구입빈도가 낮은 것은 당연하지만 통상 비 내구재로 생각되는 것도 빈도가 낮은 것이 많았다. 예컨대 과일이나 채소와 같이 여러 품목을 포괄하는 항목의 경우 구입빈도가 11 전후로 거의 매달 구입된 것으로 나오지만, 개별 품목(예컨대 배 복숭아 포도 또는 배추 무 당근 등)으로 세분해서 보면 구입빈도가 1에서 6까지 분포하는

번에 필자의 요청으로 통계청이 분기와 연간 자료에서 각 표본의 조사월수 정보를 제공하기로 하였지만, RDC에서만 이용할 수 있게 제한을 둔 것은 유감스럽다. 마이크로 데이터는 대부분 앞의 두 방식으로 이용되고 있는데, 거기에 조사월수에 관한 정보를 제공하지 않으면 데이터에 포함된 왜곡을 바로잡을 수 없기 때문이다.

6) 월간 자료를 이용하여 가구별로 연간 통계를 만들어보면 통계청의 연간 자료와 일치하지 않는다. 통계청의 설명에 따르면 월간 자료에서 가구id(제공용 가구key)로 되어 있는 것은 단순한 일련번호이며 다른 여러 가구들이 뒤섞여 동일한 id로도 묶일 수 있다고 한다. 즉 월간 자료에서는 실제의 가구id를 알 수 없게 되어 있어 가구별 통계를 구하기는 어렵다. 다만 각 품목이 1년에 구입되는 평균 월수(즉 구입 빈도)를 구하는 것은 어느 정도 가능하다고 본다. 각 품목의 평균 구입 빈도는 그 품목의 성격(예컨대 내구재 또는 비 내구재)에 따른 것이어서 가구 구분(예컨대 사무직, 생산직, 자영업자, 무직 가구 등)이 달라진다고 해서 큰 차이가 날 것으로 생각되지 않기 때문이다. 다만 이러한 자료상의 제약으로 본고가 분석에 이용한 구입 빈도는 실태와 다소 괴리될 수 있으며, 이에 관해서는 후술한다. 한편 본고가 제기한 문제는 소비뿐 아니라 소득에 대해서도 마찬가지로 적용되며, 이를 위해 소득이 발생하는 빈도(소비의 구입 빈도에 대응)를 구할 필요가 있다. 그렇지만 월간 자료의 가구 id의 문제로 예컨대 근로자와 자영업자 또는 무직 가구가 뒤섞인 경우가 생기면 소득의 발생빈도는 소비와 달리 실태와 크게 괴리되어 버린다. 본고가 소비에 초점을 맞출 수밖에 없었던 것은 이러한 자료상의 제약 때문이다.

7) 예컨대 자동차(신차구입)를 사례로 들면 2016년에 신차를 구입한 월수가 1인 경우는 101가구, 2인 경우는 4가구로 나오며 이때의 구입빈도는 (101*1+4*2)/(101+4)=1.04로 구했다. 따라서 구입빈도는 1에서 12끼지로 분포하게 된다.

8) 예컨대 구입 빈도가 10을 넘는 것으로는 전기료, 방송수신료, 상하수도, 이동전화, 외식비 등이다. 우유, 식빵, 두부, 돼지고기 등과 같이 상하기 쉬운 식품도 구입빈도가 비교적 높은 8로 나왔다. 이들에 대한 지출이 매달 이루어지는 가구가 많겠지만, 지출 실적이 1달밖에 되지 않는 가구까지 포함한 평균이라는 점에 유의할 필요가 있다.

것으로 크게 낮아진다. 품목의 계절성이 구입빈도에 영향을 미치고 있음을 알 수 있다. 구입빈도가 낮은 것을 내구재의 성격이 강한 것으로 생각했지만, 비 내구재의 경우에도 품목이 세분되면 구입빈도가 낮아짐을 알 수 있다. 〈표 5〉에서 품목이란 가장 세분된 말단의 항목(즉 과일이 아니라 배 복숭아 포도 등)을 말한다. 그리고 〈표 5〉에는 소비 비목별 품목 수의 분포와 평균 구입빈도도 제시하였다. 음식숙박, 주거수도광열, 통신의 순으로 평균 구입빈도가 높고 내구재가 많은 가정용품이 가장 낮게 나왔다. 여기서는 2016년의 수치를 제공하였지만, 2015년과 큰 차이가 없다.

〈표 5〉 구입빈도별 소비품목 수와 소비비목별 평균 구입빈도(2016년)

구입빈도	품목수	비중(%)	소비 비목	품목수	평균 구입빈도
1	120	34.1%	식료품 등	118	3.06
2	114	32.4%	주류 담배	7	2.50
3	43	12.2%	의류 신발	20	1.87
4	32	9.1%	주거수도광열	15	4.56
5	11	3.1%	가정용품 등	49	1.52
6	11	3.1%	보건	13	2.78
7	7	2.0%	교통	23	2.30
8	7	2.0%	통신	7	4.41
9	2	0.6%	오락문화	41	2.03
10		0.0%	교육	23	1.95
11	3	0.9%	음식숙박	5	5.04
12	2	0.6%	기타	31	2.62
합계	352	100.0%	합계/평균	352	2.60

자료: 통계청, 가계동향조사 월간자료(RDC 접근)에서 산출.

앞의 자동차의 예시에서 보았듯이 구입빈도가 1인 경우는 조사월수에 따라 평균값이 과대해지는 편향(편향A)을 간단히 계산할 수 있다. 그렇지만 각 품목의 구입된 평균 월수(즉 구입빈도)는 1에서 12까지 분포할 터인데, 각 경우에 편향A는 어떻게 나타날까? 이하에서는 혼란을 피하기 위해 구입빈도와 구입횟수를 구분해서 사용하기로 한다. 구입빈도는 전술한 해당 품목의 연평균 구입 월수를 말하며, 12개월 모두 조사되었을 때의 값이다. 해당 품목에 대한 지출이 얼마나 자주 이루어지는가를 보여준다. 이에 대해 해당 품목에 대한 조사월수가 12개월 미만이 되면 실제 구입한 것으로 관찰되는 횟수는 구입빈도보다 적을 터인데 이를 구입횟수라고 칭하기로 한다. 즉 구입빈도=n인 품목이라면 12개월 모두 조사된 경우는 정의상 구입횟수=구입빈도=n이

된다. 그렇지만 조사월수가 적어지면 실제 관찰되는 구입횟수는 n보다 적어진다.

〈표 6〉 조사월수별 편향A와 편향B의 추정(구입빈도=3인 경우)

조사월수	구입횟수별 경우의 수								편향A의 추정			편향B의 추정	
	산식				옆 산식의 계산 결과				평균 구입횟수	평균 구입횟수	편향A (조사월수12=1)	누락 확률	편향B
	구입횟수				구입횟수								
n	0	1	2	3	0	1	2	3	a	a*(12/n)		b	1-b
1	$_9C_1$	$_3C_1$			9	3			1	12	4	0.75	0.25
2	$_9C_2$	$_3C_1$*$_9C_1$	$_3C_2$		36	27	3		1.1	6.6	2.2	0.545	0.455
3	$_9C_3$	$_3C_1$*$_9C_2$	$_3C_2$*$_9C_1$	$_3C_3$	84	108	27	1	1.213	4.853	1.618	0.382	0.618
4	$_9C_4$	$_3C_1$*$_9C_3$	$_3C_2$*$_9C_2$	$_3C_3$*$_9C_1$	126	252	108	9	1.341	4.024	1.341	0.255	0.745
5	$_9C_5$	$_3C_1$*$_9C_4$	$_3C_2$*$_9C_3$	$_3C_3$*$_9C_2$	126	378	252	36	1.486	3.568	1.189	0.159	0.841
6	$_9C_6$	$_3C_1$*$_9C_5$	$_3C_2$*$_9C_4$	$_3C_3$*$_9C_3$	84	378	378	84	1.650	3.300	1.100	0.091	0.909
7	$_9C_7$	$_3C_1$*$_9C_6$	$_3C_2$*$_9C_5$	$_3C_3$*$_9C_4$	36	252	378	126	1.833	3.143	1.048	0.045	0.955
8	$_9C_8$	$_3C_1$*$_9C_7$	$_3C_2$*$_9C_6$	$_3C_3$*$_9C_5$	9	108	252	126	2.037	3.056	1.019	0.018	0.982
9	$_9C_9$	$_3C_1$*$_9C_8$	$_3C_2$*$_9C_7$	$_3C_3$*$_9C_6$	1	27	108	84	2.260	3.014	1.005	0.005	0.995
10		$_3C_1$*$_9C_9$	$_3C_2$*$_9C_8$	$_3C_3$*$_9C_7$		3	27	36	2.500	3	1	0	1
11			$_3C_2$*$_9C_9$	$_3C_3$*$_9C_8$			3	9	2.75	3	1	0	1
12				$_3C_3$*$_9C_9$				1	3	3	1	0	1

주: 1) 구입횟수가 0-3일 때의 각 경우의 수를 조합의 산식과 그 계산결과로 제시하였다.

2) 평균 구입횟수(a)란 구입횟수가 1보다 클 때의 경우의 수를 가중치로 하는 평균 구입횟수를 말한다. 예컨대 조사월수=3일 때의 평균 구입횟수는 1.213=(1*108+2*27+3*1)/(108+27+1)로 구해진다.

3) 누락확률(b)란 구입되었음에도 불구하고 구입되지 않은 것으로 조사될 확률을 말하는데, 구입횟수(0-3)별 경우의 수 합계에서 구입횟수=0의 비율로 구했다.

4) 편향A와 편향B는 모두 조사월수=12를 기준(=1)으로 하였을 때의 값이다.

여기서는 조사월수가 달라짐에 따라 구입횟수가 어떻게 나타나는지를 보이고자 하는데, 양자의 조합으로 이루어지는 경우의 수를 이용하여 접근하고자 한다. 〈표 6〉은 구입빈도가 3인 경우를 예로 들어 편향A와 편향B를 설명한 것이다. 12개월 중에서 구입이 이루어진 달을 1, 그렇지 않은 달을 0이라 하면, 구입빈도=3이란 1이 3개 0이 9개로 이루어진 경우가 된다. 이로부터 조사월수가 늘어남에 따라 어떠한 경우의 수가 가능한지를 〈표 6〉의 처음 4열이 조합(combination)의 산식으로 보여준다. 먼저 조사월수=1이면서 구입횟수=0인 경우의 수는 $_9C_1$(즉 9개의 0에서 하나를 뽑는 경우의 수)로 구할 수 있으며 그 값은 9가 된다. 조사월수= 2 또는 3이면서 구입횟수=0인 경우의 수는 각각 $_9C_2$와 $_9C_3$로 구하며 각각 36과 84가 된다. 한편 조사월수=1이면서 구

입횟수=1인 경우의 수는 $_3C_1$로 구할 수 있으며 그 값은 3이 된다(조사월수=1일 때에는 구입횟수는 1을 넘을 수 없다). 조사월수=2일 때에는 구입횟수가 1 또는 2가 될 수 있는데, 각각의 경우의 수는 $_3C_1*_9C_1$과 $_3C_2$로 나타낼 수 있다. $_3C_1*_9C_1$은 $_3C_1$과 $_9C_1$의 곱인데, $_3C_1$은 3개의 1에서 하나를 뽑는 경우의 수이고, $_9C_1$는 9개의 0에서 하나를 뽑는 경우의 수를 뜻한다. 전자가 3이고 후자가 9이므로 그 값은 27이 된다. 이에 대해 $_3C_2$는 3개의 1에서 둘을 뽑는 경우의 수이며 그 값은 3이 된다. 조사월수=3일 때에는 구입횟수도 1, 2, 3까지 가능하며, 각각의 경우의 수는 $_3C_1*_9C_2$, $_3C_2*_9C_1$, $_3C_3$으로 나타낼 수 있다. $_3C_1*_9C_2$는 3번의 조사월수에서 1이 한 번 0이 두 번 뽑히는 경우의 수이고, $_3C_2*_9C_1$는 1이 두 번 0이 한 번, $_3C_3$는 1이 세 번 뽑히는 경우의 수를 말하며 그 값은 각각 108, 27, 1이 된다. 이를 일반화해서 조사월수=n이고 구입횟수=k일 때의 경우의 수는 $_3C_k*_9C_{n-k}$로 나타낼 수 있다.

각 표본의 품목별 소비통계를 보면 구입이 이루어진 경우와 구입=0인 경우로 나눌 수 있는데, 편향A는 구입이 이루어진(즉 구입횟수>=1인) 경우를 대상으로 한다. 구입=0인 경우는 다시 실제 구입이 이루어졌지만 조사되지 않은 경우와 실제로 구입=0인 경우로 나뉘는데, 전자가 편향B에 해당한다. 여기서는 먼저 편향A의 추정을 설명한다. 〈표 6〉의 '평균 구입횟수'란 조사월수별로 각 구입횟수(여기서는 1, 2, 3)를 각각에 대응하는 경우의 수를 가중치로 하여 가중 평균한 것이다. 조사월수=12일 때에 평균 구입횟수는 3이 되어 구입빈도와 일치함을 알 수 있다. 조사월수=1일 때의 평균 구입횟수가 1로 나온 것은 구입횟수=0인 경우는 편향B에 해당하기 때문에 여기서는 제외하고 구입횟수>=1인 경우로 한정하였기 때문이다. 그 사이의 예컨대 조사월수=6일 때 조사에서 관찰될 것으로 기대되는 평균 구입횟수는 1.65회임을 알 수 있다. 조사월수가 1에서 12로 늘어남에 따라 평균 구입횟수도 점차 늘어나 3으로 수렴하고 있다. 이것은 구입빈도=3인 품목의 경우 조사월수가 1에서 12로 늘어남에 따라 조사에서 실제 관찰될 것으로 합리적으로 기대할 수 있는 구입횟수라 할 수 있다.[9]

이에 대해 〈표 6〉의 '통계청의 연 환산' 항목은 통계청이 월간 자료로부터 연간 자료를 만들 때 이용한 환산 방법을 적용한 결과를 보인 것이다. 그 방식이란 조사된 월의 평균을 연 평균으로 간주하는 것이므로 조사월수=1일 때 해당 품목의 구입이 이루어졌다면 매달(=12/1) 동일한 구입이 이루어진 것이 된다. 조사월수=2이면 그 두

9) 여기서 구입되는 품목의 가격을 모두 1이라고 보면, 구입횟수는 곧 구입가액으로 해석할 수 있다. 즉 구입횟수가 n배 많다는 것은 구입가액이 n배가 된다는 뜻이 된다.

달의 평균 구입이 매 2달마다(즉 1년에 6회=12/2) 이루어진 것이 되며, 일반화하면 조사월수=n이면 그 n달의 평균 구입이 매 n달마다(즉 1년에 12/n회) 이루어진 것이 된다. 조사월수=12일 때에는 문제가 없지만, 조사월수(n)가 줄어들면 통계청의 연간 자료에 나와 있는 연 평균 구입횟수는 실제 관찰되는 구입횟수(a)의 12/n배가 된다. 〈표 6〉의 '통계청의 연 환산' 항목은 조사월수에 따라 실제 나타날 것으로 합리적으로 추정되는 평균 구입횟수에 12/n을 곱해서 구한 것으로 통계청의 연간 자료의 수치를 보여준다. 이를 조사월수 당 실제 관찰된 구입횟수(=a/n)에 12개월을 곱해서 구한 것으로 표현할 수도 있다.

그 다음 열의 '편향A'는 '통계청의 연 환산'을 조사월수=12인 경우가 실제의 값이므로 그것을 기준(=1)로 하여 조사월수가 줄어듦에 따라 그로부터 얼마나 벗어나는지를 보인 것이다. 이것이 구입빈도=3인 경우의 편향A가 되는데, 실제보다 최대 4배까지 과대평가되고 있음을 알 수 있다. 통계청의 연간 자료에서 나타난 이 편향을 바로잡기 위해서는 구입빈도가 3인 품목의 경우 예컨대 조사월수=1, 2, 3일 때의 연간 자료의 구입액을 각각 4, 2.2, 1.618로 각각 나누어 주면 된다. 따라서 이 비율을 구입빈도=3인 경우 조사월수가 줄어듦에 따라 점차 과대해지는 편향A를 바로잡는 보정비율이라 할 수 있다.

이상은 구입빈도=3인 경우를 예시한 것이지만, 구입빈도가 달라지면 보정비율도 달라진다. 앞에서 구입빈도=3일 때 조사월수=n이고 구입횟수=k일 때의 경우의 수는 ${}_3C_k * {}_9C_{n-k}$로 나타낼 수 있다는 점을 언급했다. 구입빈도=4인 경우는 구입한 달인 1이 4개이고 구입이 없었던 달인 0이 8개인 경우에 해당하며, 구입횟수(k)가 4까지 늘어날 수 있게 된다. 이를 산식으로 나타내면 ${}_4C_k * {}_8C_{n-k}$가 된다. 구입빈도=p라고 하고 일반화해서 나타내면 ${}_pC_k * {}_{(12-p)}C_{n-k}$가 된다.

〈표 7〉은 위의 방법을 이용하여 구한 구입빈도(1-12)와 조사월수(1-12)에 따른 보정 비율(이는 곧 편향A)을 제시하였다. 그에 따르면 구입빈도=12인 경우(즉 매달 지출이 이루어지는 경우)는 조사월수에 상관 없이 보정 비율이 1로 나왔다. 반대로 조사월수=12인 경우에도 구입빈도에 상관 없이 보정 비율이 1로 나왔다. 〈표 7〉에서 대각선 이하의 보정 비율이 모두 1로 나왔는데, 그것은 구입빈도와 조사월수의 합계가 13 이상이면 보정할 필요가 없음을 뜻한다. 이에 대해 구입빈도와 조사월수의 합계가 12 이하면 보정 비율이 1보다 크며, 구입빈도와 조사월수가 적어질수록 보정 비율이 커져 최대 12배까지로 나왔음을 알 수 있다.

〈표 7〉 구입 빈도와 조사월수에 따른 보정 비율(편향A)

조사	구입 빈도											
월수	1	2	3	4	5	6	7	8	9	10	11	12
1	12	6	4	3	2.4	2	1.714	1.5	1.333	1.2	1.091	1
2	6	3.143	2.2	1.737	1.467	1.294	1.179	1.1	1.048	1.015	1	1
3	4	2.2	1.618	1.341	1.189	1.1	1.048	1.019	1.005	1	1	1
4	3	1.737	1.341	1.165	1.076	1.031	1.010	1.002	1	1	1	1
5	2.4	1.467	1.189	1.076	1.027	1.008	1.001	1	1	1	1	1
6	2	1.294	1.1	1.031	1.008	1.001	1	1	1	1	1	1
7	1.714	1.179	1.048	1.010	1.001	1	1	1	1	1	1	1
8	1.5	1.1	1.019	1.002	1	1	1	1	1	1	1	1
9	1.333	1.048	1.005	1	1	1	1	1	1	1	1	1
10	1.2	1.015	1	1	1	1	1	1	1	1	1	1
11	1.091	1	1	1	1	1	1	1	1	1	1	1
12	1	1	1	1	1	1	1	1	1	1	1	1

이상은 36개월 연동표본의 경우(2015-16년)를 상정하고 있는데, 6-6-6 연동표본으로 설계된 2019-20년의 경우에는 어떻게 보정할 수 있을까? 그 경우는 모든 표본의 조사월수가 6개월(2019년은 11개월)을 넘지 않는다. 따라서 2015-16년의 경우에 월간 자료를 이용하여 구한 품목별 구입빈도를 구할 수 없다. 다만 구입빈도는 각 품목에 대한 지출이 얼마나 자주 이루어지는가의 특성을 보여주는 것이기 때문에 연도에 따라 크게 변하지 않을 것으로 볼 수 있다. 실제 2015년과 2016년의 구입빈도를 비교해 보면 352개 소비품목 중에서 구입빈도가 일치하는 품목이 94%로 나왔고 차이가 나는 나머지 6%도 반올림으로 빈도가 달라지는 정도였다. 여기서는 가장 인접한 2016년 구입빈도 통계를 2019-20년에 적용하기로 한다. (2016년에 비해 2019년 이후 품목이 세분된 경우가 있는데, 그 경우 기존 품목의 구입빈도를 적용하였다). 2020년은 조사월수가 6개월(2019년은 11개월)을 넘는 경우가 없지만, 〈표 7〉을 그대로 이용할 수 있다. 즉 2019-20년의 보정비율 표를 따로 제시하지 않았지만, 〈표 7〉에서 조사월수가 1-6개월(2019년은 1-11개월)에 해당하는 보정비율이 2020년(2019년)의 보정비율이 된다. 따라서 2015-16년이나 2019-20년의 마이크로 데이터(연간 자료)의 소비지출액은 모두 〈표 7〉의 보정 비율을 이용하여 수정되었다.[10]

10) 〈표 7〉의 보정비율은 다음과 같은 방법으로 각 가구의 소비지출에 적용된다. 각 가구는 조사월수에 따라 12개로 나누고, 그 가구가 소비한 모든 품목은 구입빈도에 따라 다시 12개로 나누어 도합 144(=12*12)개로 분류된다. 그리고 각 품목의 소비지출을 표의 해당하는 셀의 보정비율로 나누면 편향A가 보정된다. 그런데 이 경우 구입빈도가 반올림한 값이어서 오차가 생길 수 있어

이렇게 보정된 결과가 얼마나 실태에 더 근접하게 되었는지를 검토해 보자. 〈표 8〉은 보정되기 전과 후의 두 데이터를 이용하여 2015-16년의 경우 조사월수=12인 가구와 12개월 미만인 가구로 나눈 다음, 품목별로 두 그룹의 소비지출 평균값이 동일한지의 여부를 t-test한 결과를 보인 것이다. 그에 따르면 2015년은 전체 352개 소비품목 중에서 보정 되기 전에는 그 89%(313개 품목)가 5%의 유의수준에서 두 그룹이 통계적으로 다른 것으로 나왔는데, 보정 후에는 동 비율이 25%(88개)로 줄어들었다. 2016년은 동 비율이 89%에서 28%로 줄어들었다. 위 방식의 보정으로 인해 통계가 실태에 상당히 근접하게 되었음을 알 수 있다. 그리고 이를 소비 비목별로 나누어 보면 이러한 편향A의 문제는 모든 비목에 걸쳐 대체로 유사한 양상을 보이고 있다.

〈표 8〉 조사월수가 달라도 각 품목의 소비지출 평균값이 동일한가 여부를 t-test한 결과

	2015			2016			2019			2020		
	품목수	비중(p<0.05)		품목수	비중(p<0.05)		품목수	비중(p<0.05)		품목수	비중(p<0.05)	
		보정 전	보정 후		보정 전	보정 후		보정 전	보정 후		보정 전	보정 후
식료품 음료	118	95.8%	26.3%	118	96.6%	35.6%	118	98.3%	63.6%	118	98.3%	64.4%
주류 담배	7	100.0%	28.6%	7	85.7%	14.3%	7	100.0%	28.6%	7	100.0%	57.1%
의류 신발	20	90.0%	55.0%	20	100.0%	30.0%	20	100.0%	45.0%	20	100.0%	40.0%
주거수도광열	15	66.7%	33.3%	15	60.0%	46.7%	16	87.5%	56.3%	16	81.3%	50.0%
가정용품 등	49	93.9%	12.2%	49	91.8%	24.5%	49	91.8%	24.5%	49	95.9%	26.5%
보건	13	100.0%	15.4%	13	84.6%	7.7%	13	92.3%	38.5%	13	100.0%	53.8%
교통	23	95.7%	30.4%	23	87.0%	17.4%	24	100.0%	37.5%	24	95.8%	41.7%
통신	7	85.7%	42.9%	7	85.7%	42.9%	7	85.7%	28.6%	7	71.4%	42.9%
오락문화	41	85.4%	9.8%	41	85.4%	17.1%	41	90.2%	48.8%	41	82.9%	29.3%
교육	23	65.2%	30.4%	23	78.3%	26.1%	23	73.9%	56.5%	23	60.9%	39.1%
음식숙박	5	80.0%	20.0%	5	80.0%	0.0%	6	100.0%	66.7%	6	83.3%	33.3%
기타	31	77.4%	29.0%	31	80.6%	35.5%	33	81.8%	33.3%	33	78.8%	60.0%
전체 소비지출	352	88.9%	25.0%	352	88.9%	28.4%	357	92.7%	47.9%	357	90.5%	48.0%

주: 2015-16년의 경우는 조사월수=12인 그룹과 조사월수<12인 그룹 사이에 소비지출 평균값이 동일하다는 귀무가설을 모든 품목에 대해 검증한 결과 5%의 유의수준에서 귀무가설이 기각된 품목의 비중이다. 2019-20년은 조사월수>=6인 그룹과 조사월수<6인 두 그룹에 대해 검증한 결과이다.

자료: 통계청, 가계동향조사(RDC 접근)에서 산출.

다음과 같이 보완하였다. 예컨대 조사월수=2이고 구입빈도가 3인 품목의 보정비율은 2.2인데, 그 품목의 반올림하기 전의 실제 구입빈도가 3.2인 경우를 생각해 보자. 그 경우 구입빈도=4의 보정비율이 1.737이므로 보다 정확한 보정비율은 구입빈도 3과 4의 사이의 위치를 감안한 가중평균 값인 2.2*(4-3.2)+1.737*(3.2-3)=2.107로 구했다.

2019년 이후는 6-6-6 연동표본 방식으로 바뀌면서 12개월 모두 조사된 표본이 없어 비교 그룹을 달리 설정하였다. 2020년은 조사월수=6개월인 가구와 그 미만으로 조사된 가구의 두 그룹으로, 2019년은 조사월수 6개월 이상과 그 미만의 두 그룹으로 나누었다. 그 결과를 보면 2019년은 보정 되기 전에는 전체 357개 품목 중에서 93%가 5%의 유의수준에서 두 그룹이 통계적으로 다른 것으로 나왔는데, 보정 후 동 비율은 48%로 줄었다. 2020년에는 동 비율이 91%에서 48%로 줄어들었다. 2015-16년에 비해 보정 효과가 상대적으로 작게 나왔다. 그것은 조사월수=12인 가구가 비교기준이 된 2015-16년과는 달리 2019-20년은 비교 기준이 되는 그룹에 편향A가 다소 포함되어 있기 때문이다.

위의 방식으로 연간 자료의 편향A를 상당히 보정할 수 있지만, 보정이 불충분한 품목도 적지 않게 나왔다. 그 실태의 일단을 보이기 위해 〈표 9〉는 몇 가지 품목에 대해 보정 이전과 이후에 평균값과 t-test의 p값이 어떻게 변했는지를 예시하였다. 먼저 딸기를 보면 구입 실적이 있는 표본 중에서 조사월수=12인 그룹1과 그 미만인 그룹2는 각각 2,883개와 3,403개이며, 각 그룹의 월 평균 구입액은 보정 전에 각각 3,733원과 6,312원으로 나와 차이가 컸다. 보정 후 그룹2의 평균 구입액은 4,092원으로 떨어져 양자의 갭이 줄어들었지만 보정이 불충분하여 두 그룹은 여전히 통계적으로 다른 것으로 나왔다. 굴비의 경우에는 두 그룹의 평균 구입액은 보정 전에는 5,202원과 8,230원으로 나왔지만 보정 후 격차가 줄어들어 통계적으로 다르지 않은 것으로 나왔다. 이것은 굴비의 평균 구입빈도인 1.32를 적용한 결과인데, 만약 반올림한 구입빈도인 1을 적용했을 때에는 양자의 격차가 커져서 통계적으로 다른 것으로 나왔다. 즉 구입빈도의 소수점까지 고려하면 추계의 오차를 줄일 수 있음을 알 수 있다.

한편 2020년은 소주와 구두를 예시하였는데, 그룹1의 평균 구입액이 보정 전에 비해 보정 후에 줄어든 것(소주는 7,630원=>7,212원, 구두는 16,781원=>10,241원)을 알 수 있다. 2015년의 그룹1(조사월수=12)과는 달리 2020년의 그룹1(조사월수=6)은 보정의 영향을 받기 때문이다. 구두의 경우 그룹1과 2의 평균 구입액 격차(16,781원과 34,283원)는 보정에 의해 크게 줄어 양자의 갭은 거의 사라졌다. 그런데 소주의 경우 그룹1과 2의 보정 전 평균 구입액(7,630원과 9,415원)이 보정 후에 역전(7,121원과 5,945원)된 것으로 나와 보정 비율이 과대한 것을 알 수 있는데, 소주의 실제 구입빈도가 여기서 적용된 구입빈도(3.61)보다 더 컸을(따라서 보정비율이 더 낮았을) 가능성을 시사한다. 앞에서 각 품목의 구입빈도를 구할 때 월간 자료의 한계로 인해 오차가 발생할 수 있음을 언급했는데, 이것이 보정의 과대 또는 과소를 낳는 또

하나의 요인이 된다. 그 외에 품목명에 '기타'를 포함한 품목이 적지 않은 것도 한 요인으로 지적할 수 있다. 거기에는 이질적인 것들이 섞여 있을 가능성이 높아 조사월수의 차이를 보정하여도 그룹간 이질성이 남아 있을 수 있기 때문이다.

요컨대, 〈표 7〉의 보정 비율을 적용하면 조사월수의 차이로 인한 평균값의 과대평가(즉 편향A)를 상당히 바로잡을 수 있다. 전체 품목의 90% 전후가 편향A의 문제를 안고 있는데, 이를 25-47% 수준으로 낮추었다. 나머지는 구입빈도의 오차로 인해 보정 비율이 다소 과대 또는 과소해졌지만, 그렇다고 해도 보정하기 전에 비하면 실제의 수치에 상당히 근접시켰다고 할 수 있다.

〈표 9〉 품목의 예시: 보정 전과 후의 t-test 결과 비교

	2015				2020			
	딸기		굴비		소주		구두	
	보정 전	보정 후	보정 전	보정 후	보정 전	보정 후	보정 전	보정 후
구입 표본수								
그룹1	2,883	2,883	755	755	4,114	4,114	1,572	1,572
그룹2	3,403	3,403	742	742	2,520	2,520	751	751
평균 구입액								
그룹1	3,733	3,733	5,202	5,202	7,630	7,212	16,781	10,241
그룹2	6,312	4,092	8,230	4,181	9,415	5,945	34,283	10,709
t-test의 p값	1.2E-55	0.00068	0.0001	0.0580	1.2E-08	0.0000	3.2E-26	0.4380
구입빈도	2.51	(3)	1.32	(1)	3.61	(4)	1.51	(2)

주: 1) 2015년의 그룹1과 그룹2는 조사월수=12인 가구와 그 외의 가구를 말하고, 2020년은 조사월수=6인 가구와 그 외의 가구를 말한다.
2) 평균 구입액이란 해당 품목을 구입한 가구의 월 평균 구입액(단위: 원)을 말한다.
3) 구입빈도에서 ( ) 안의 수치는 반올림한 값이다.
자료: 통계청, 가계동향조사(RDC 접근)에서 산출.

3. 편향B: 조사 누락에 의한 과소평가

다른 한편 실제로는 소비지출이 이루어졌지만 조사에서 누락된 경우(전술한 편향B)가 얼마나 될까? 이에 관해서는 앞의 〈표 6〉에서 편향A를 추정하였던 방식을 응용하여 접근할 수 있다. 거기에서 구입빈도가 3인 경우를 예로 들어 조사월수별로 구입횟수가 0에서 3일 때의 경우의 수를 계산한 바 있다. 거기에서 실제로 구입이 이루어졌음에도 불구하고 조사에서 누락된 경우는 구입횟수=0에 해당한다. 예컨대 조사월수=1일 때에는 구입횟수가 0이거나 1의 두 가지가 가능할 터인데, 각각의 경우의 수

는 $_9C_1$와 $_3C_1$로 구하며 그 값은 9와 3이 된다. 즉 조사월수=1일 때 구입되었음에도 불구하고 누락되는 확률(〈표 6〉의 b)은 9/(9+3)으로 구하며 0.75가 된다. 조사월수=2일 때에는 구입횟수가 0, 1, 2의 세 가지가 가능하며, 각각의 경우의 수는 $_9C_2$, $_3C_1{}^*{}_9C_1$, $_3C_2$가 되며 그 값은 각각 36, 27, 3이 된다. 따라서 이 때 구입횟수가 0이 되어 조사에서 누락될 확률은 36/(36+27+3)으로 구하며 0.545가 된다. 조사월수가 늘어나면 구입횟수=0이 될 확률(즉 〈표 6〉의 누락 확률)은 점차 낮아져서 조사월수가 10을 넘으면 0가 된다. 이를 편향A와 비교할 수 있도록 조사월수=12를 기준(=1)으로 바꾸면 〈표 6〉의 편향B가 된다. 예컨대 조사월수=1일 때의 누락 확률이 75%라는 것은 기준(=1)에 비해 25%(=1-0.75)가 파악되었음을 뜻한다.

〈표 6〉은 구입빈도가 3인 경우를 예시한 것인데, 편향B를 구입빈도가 1-12인 경우로 확장하여 제시한 것이 〈표 10〉이다. 이것은 편향A를 보여준 〈표 7〉과 여러 가지 점에서 대응한다. 구입빈도와 조사월수의 합이 13보다 큰 대각선 이하가 모두 1로 나온 것이 그러하며 이 경우는 보정할 필요가 없음을 뜻한다. 구입빈도나 조사월수가 줄어들면 과소평가(편향A의 경우는 과대평가)의 편향이 점차 커지는 것도 마찬가지이다. 더욱 흥미로운 것은 편향B는 편향A의 역수라는 점이다. 예컨대 구입빈도와 조사월수가 각각 1인 경우 편향A와 B의 값은 각각 12와 0.083(=1/12)이 되며, 구입빈도=3 조사월수=2인 경우 편향A와 편향B는 각각 2.2와 0.455(=1/2.2)로 역수임을 알 수 있다. 이것은 구입빈도와 조사월수(또는 그 합계)가 동일한 품목의 경우 구입이 이루어지거나(이 경우는 편향A가 발생), 그럼에도 조사에서 누락(이 경우는 편향B가 발생)될 수 있는데, 양자로 인한 편향은 상쇄될 수 있음을 시사한다.

그런데 편향A의 경우는 〈표 7〉의 수치를 보정 비율로 이용하여 마이크로 데이터의 각 표본의 해당 구입액을 수정할 수 있었다. 이에 대해 편향B의 경우는 〈표 10〉을 이용하여 직접 마이크로 데이터의 수치를 바로잡을 수가 없다. 편향B의 경우는 각 품목의 구입액이 0인 경우가 대상인데, 거기에는 실제 구입되었음에도 불구하고 조사에서 누락된 경우(즉 편향B) 뿐만 아니라 실제로 구입되지 않아 구입=0인 경우가 섞여 있어 구별할 수가 없기 때문이다. 즉 편향B가 발생한 표본을 특정할 수가 없다.

여기서는 먼저 품목별로 편향B가 어느 정도의 규모인지를 추정해 보기로 한다. 이를 위해 편향A를 바로잡은 마이크로 데이터를 가지고 출발한다. 그리고 2015-16년의 경우는 조사월수=12인 표본을 대상으로 각 품목의 구입가구 비율을 구한다. 예컨대 맵쌀의 경우 구입 실적이 한번 이상 있는 가구가 전체 가구의 89.5%로 나왔고, 자동차의 동 비율은 3.8%로 나왔다. 맵쌀의 경우 구입빈도가 4이고 자동차는 1인데, 〈표

10〉에서 거기에 해당하는 보정비율을 구할 수 있다. 예컨대 구입빈도=4이면서 조사월수=1인 경우는 편향B가 0.333, 조사월수=2, 3, 4일 때에는 각각 0.576, 0.745, 0.859 등이다. 따라서 맵쌀의 조사월수=1, 2, 3, 4일 때의 구입가구 비율은 89.5%에서 이들 보정비율을 곱한 29.8%, 51.5%, 66.7%, 76.8%가 된다. 그리고 이들 비율과 조사월수=12인 경우의 동 비율인 89.5%와의 차이가 누락된 가구 비율로 볼 수 있다. 이 누락 가구 비율에 해당 조사월수에 대응하는 가구수를 곱하면 실제로는 맵쌀을 구입했지만 조사에서 누락된 가구수를 산출할 수 있다. 그리고 누락된 가구의 평균 구입액이 조사된 가구와 다르지 않다고 가정하면 조사에서 누락된 구입액을 구할 수 있다. 자동차의 경우는 조사월수=12일 때의 구입가구 비율이 맵쌀보다 훨씬 낮은 3.8%이므로 여기에다 〈표 10〉에서 자동차(구입빈도=1)의 보정 비율을 적용하면 된다.[11]

〈표 10〉 구입 빈도와 조사월수에 따른 보정비율(편향B)

조사	구입 빈도											
월수	1	2	3	4	5	6	7	8	9	10	11	12
1	0.083	0.167	0.25	0.333	0.417	0.5	0.583	0.667	0.75	0.833	0.917	1
2	0.167	0.318	0.455	0.576	0.682	0.773	0.848	0.909	0.955	0.985	1	1
3	0.25	0.455	0.618	0.745	0.841	0.909	0.955	0.982	0.995	1	1	1
4	0.333	0.576	0.745	0.859	0.929	0.97	0.99	0.998	1	1	1	1
5	0.417	0.682	0.841	0.929	0.973	0.992	0.999	1	1	1	1	1
6	0.5	0.773	0.909	0.97	0.992	0.999	1	1	1	1	1	1
7	0.583	0.848	0.955	0.99	0.999	1	1	1	1	1	1	1
8	0.667	0.909	0.982	0.998	1	1	1	1	1	1	1	1
9	0.75	0.955	0.995	1	1	1	1	1	1	1	1	1
10	0.833	0.985	1	1	1	1	1	1	1	1	1	1
11	0.917	1	1	1	1	1	1	1	1	1	1	1
12	1	1	1	1	1	1	1	1	1	1	1	1

그런데 이 방식으로는 품목별로 조사에서 누락된 편향B의 규모를 구할 수 있지만, 그 편향이 어느 가구에서 발생했는지를 알 수가 없다. 여기서는 다음의 방식으로 편

11) 2015-16년의 경우 마이크로 데이터로부터 각 품목의 조사월수별 구입가구 비율을 직접 구할 수 있으며, 그 결과를 보면 조사월수가 줄어들면 구입가구 비율이 떨어지는 양상을 보인다. 다만 자동차의 경우처럼 구입가구 수가 적은 경우(〈표 1〉 참조)에는 그 비율이 본고의 방식(〈표 10〉)과 다소 괴리될 수 있다. 그리고 2019-20년과 같이 조사월수가 12개월에 미치지 못하는 경우에는 〈표 10〉을 이용하여 조사되지 않은 월수의 구입가구 비율을 추정할 수 있는 장점이 있다. 이러한 점을 고려하여 본고는 〈표 10〉에 의거하여 조사월수별 구입가구 비율을 구했다.

향B가 발생할 확률이 높은 가구를 추정하기로 한다. 이를 위해 가구 j의 품목 i의 구입 여부(구입한 경우를 1 구입하지 않은 경우를 0이라 하자)를 보여주는 더미변수인 P_{ij}를 종속변수로 하는 식 (1)의 probit 모형을 이용한다. 거기에서 Pr는 확률을, Φ는 표준 정규분포의 누적분포함수(CDF)를 말하고, Y_j는 가구 j의 가처분소득[12]을, Z_j는 가구 j의 인구변수와 주거정보 등을 가리킨다. 그 중에는 가구주의 연령과 그 연령의 제곱, 가구원수, 취업인원수, 아동 수, 그리고 가구주의 성별, 종사상지위, 교육수준, 배우자 유무, 도시 거주 여부, 입주형태(자가 거주와 전세 및 월세 등), 거처형태(단독주택, 아파트, 연립주택 등)의 더미 변수가 포함되었다. 각 가구의 해당 품목의 추정 값이 1에 가까울수록 해당 품목이 구입되었을 확률이 높아지고 0에 가까울수록 낮아진다. 여기서는 조사월수에 따라 누락 가구 비율이 달라진다는 점을 고려하여 조사월수별로 나누어 접근하기로 한다.

$$\Pr(P_{ij}=1|Y_j,Z_j)=\Phi(\alpha+\beta \cdot \ln(Y_j)+\delta Z_j) \quad (1)$$

그런데 한 가지 더 고려할 점이 있다. 즉 추정식에 포함된 가처분소득(Y_j)은 각 품목의 구입 여부를 추정하는데 빠뜨릴 수 없는 변수이지만, 다른 한편 그로 인해 소득이 높은 가구의 구입 확률을 실제보다 더 과장되게 만드는 요인이 되기도 한다. 이를 확인하기 위해 먼저 식 (1)을 이용하여 추정된 조사에서 누락된 가구와 해당 품목을 구입한 것으로 조사된 가구의 각 10분위 분포를 구해 비교해 보았다. 내구재를 포함한 많은 품목에서 식 (1)로 추정된 누락 가구가 10분위 분포의 상위 구간에 상대적으로 더 집중된 것으로 나타났다. 여기서는 이러한 쏠림을 바로잡기 위해 모든 품목에 대해 편향B로 인해 조사에서 누락된 가구의 소득 10분위 분포가 해당 품목을 구입한 것으로 조사된 가구의 소득분포와 다르지 않다는 가정을 추가하기로 한다. 어느 가구가 어느 품목을 구입한 월이 조사월에 해당하는지 아닌지는 우연에 불과하므로

12) 2015-16년의 연간자료에는 소득이 포함되어 있지만, 2019-20년에는 분기 자료에는 포함되어 있는 반면 연간 자료에는 빠져 있다. 여기서는 Stata의 duplicates의 명령어를 이용하여 분기와 연간 자료를 가구별로 matching하여 동일 가구를 식별해 내고, 그들의 분기 자료의 소득 평균을 연간 자료의 소득으로 impute하였다. 동일 가구 여부를 판단하는 기준으로는 가구주 및 가구원의 성, 연령, 학력과 같은 인적 속성 이외에도 주택정보(월세평가액, 주거 평수, 보증금이나 월세 등)를 이용하였다. 가구원의 경우 조사기간 중에 연령이나 취업 여부가 변경되는 경우가 있어 이를 제외하였다. Matching이 되지 않거나 중복 matching되는 경우가 약간 나오지만 문제될 정도는 아니다.

두 그룹의 소득분포에 차이가 날 이유가 없기 때문이다. 그리고 자동차에 관해서는 가구별 소유대수를 알 수 있는데, 자동차 소유가 너무 집중되지 않도록 소유대수가 3대 이상인 가구는 자동차 구입이 없는 것으로 제한하였다. 자동차는 또한 신차와 중고차로 나뉘어 있는데, 동일 가구가 신차와 중고차를 모두 구입한 것으로 추정되기도 한다. 실제로 신차와 중고차를 동시에 구입하는 경우는 거의 없을 것이라 생각되어 양자의 구입 가구가 겹치지 않도록 가정을 추가하였다.

조사에서 누락된 가구는 다음과 같이 선정되었다. 먼저 식 (1)에 의해 품목별로 각 가구의 구입 확률을 구하고, 이를 해당 품목의 구입=0인 가구를 대상으로 구입 확률이 높은 순으로 정렬한다. 이 때 선정되는 누락 가구의 소득분포가 해당 품목을 구입한 것으로 조사된 가구를 대상으로 해서 구한 가처분소득 기준의 소득10분위별 분포와 다르지 않도록 할 필요가 있다. 이를 위해 대상 가구를 조사월수*소득10분위의 소그룹(2020년은 6*10=60개)으로 나누어 구입 확률이 높은 순으로 정렬한 다음 각 그룹별로 산출된 누락된 가구 비율을 적용하였다.[13]

이상의 방법으로 각 품목별로 편향B로 인해 조사에서 누락된 가구를 추정하였지만, 그들이 해당 품목을 구입하는데 얼마를 지출했는가에 관한 정보가 없다. 여기서는 통상의 소비함수인 식 (2)를 이용하여 그 소비지출액을 추정하였다. 거기에서 C_{ij}는 가구 j의 품목 i의 소비지출액을 말하며 그 외의 다른 변수는 식 (1)과 동일하다.

$$C_{ij} = \alpha + \beta \cdot \ln(Y_j) + \delta Z_j + e_{ij} \quad (2)$$

전술한 편향A로 인한 과대평가의 보정에 비해 편향B의 보정은 자료의 제약으로 인해 어느 정도 오차를 각오하지 않으면 안 된다. 무엇보다도 조사에서 누락된 가구를 추정하는데 한계가 있기 때문이다. 다만 후술하는 〈표 11〉에서 볼 수 있듯이 보정 전

13) 앞의 맵쌀의 사례를 가지고 수치로 설명하면 다음과 같다. 조사월수=1, 2, 3일 때의 2015년 맵쌀의 구입가구 비율은 29.8%, 51.5%, 66.7%이고, 이들 비율과 조사월수=12인 경우의 동 비율인 89.5%와의 차이(즉 59.7%, 38%, 22.8%)가 누락된 가구 비율이 된다. 여기서 맵쌀을 구입한 것으로 조사된 가구의 소득10분위 구성비는 예컨대 1, 2, 3분위만 제시하면 9.7%, 9.6%, 9.9%로 나왔다. 따라서 조사월수=1이면서 소득분위=1인 가구의 누락 비율은 59.7%*9.7%가 되고, 조사월수=3이면서 소득분위=3인 가구의 누락 비율은 22.8%*9.9%가 된다. 맵쌀은 분위별 차이가 거의 없지만, 예컨대 자동차의 경우 1, 2, 3분위는 0.6%, 1.8%, 1.1%이고 10분위는 21.1%로 차이가 크다. 이런 방식으로 구성비가 높은(낮은) 분위는 더 많은(적은) 가구가 선정된다.

과 후의 가구당 소비지출액과 함께 거기에 포함된 편향A와 편향B의 추정 값을 제시하였는데, 전체 또는 비목별로 비교한 두 편향의 크기가 대체로 근접한 것으로 나온다. 그것은 전체 또는 비목별 합계를 구할 경우 두 편향의 크기가 같아서 상쇄된다는 본고 논리적 예상과 부합함을 알 수 있다. 이것은 편향B를 보정할 때 생긴 오차가 그렇게 문제가 될 정도는 아니라는 점을 시사한다.

Ⅲ. 보정 이전과 이후의 비교

1. 소비지출의 평균과 구성비

그러면 위에서 추정한 편향A와 편향B가 어느 정도의 규모이고, 시기별로 어떻게 변해 왔는지를 살펴보기로 하자. 〈표 11〉은 위에서 보정된 마이크로 데이터를 이용하여 보정 이전과 이후의 가구당 소비지출액과 함께 편향A와 편향B의 추정 결과를 제시하였다.[14] 2015년의 경우 가구당 월평균 소비지출은 219만원인데, 편향A로 인해 12만원이 과대해졌고, 비슷한 규모의 편향B로 인한 과소평가가 발생해서 보정 후의 소비지출은 보정 전과 크게 다르지 않게 나온 것을 알 수 있다. 이를 소비 비목별로 나누어 보아도 유사하게 나타났다. 교통에서 편향이 가장 큰 것으로 나왔고 음식숙박에서는 낮게 나왔지만, 다른 비목에서도 크고 작은 편향이 발생하였다. 비목별로도 편향A와 편향B의 규모가 비슷해서 상쇄되고 있음을 알 수 있다.

2016년의 경우도 크게 다르지 않다. 편향A(또는 편향B)가 보정 전 전체 소비지출에서 차지하는 비중을 보면 5% 전후로 나왔다. 이에 대해 2019-20년은 양상은 유사하지만 편향의 규모가 상당히 커졌다. 2019년은 보정 전 소비지출액 대비 비중이 13%로 늘어났고 2020년은 다시 18%로 늘어났다. 2015-16년에는 12개월 모두 조사된 표본의 비중이 상대적으로 높은 반면, 표본 설계가 6-6-6 연동표본 방식으로 바뀐 2019년 이후 조사월수가 더욱 줄었기 때문이다. 2015-2020년의 평균 조사월수를 구해 보면(〈표 4〉 참조) 각각 8.0, 8.6, 5.5, 4.6개월로 나온다.

---

14) 식 (1)의 probit 추정식이 종속변수의 분산의 어느 정도를 설명하는가를 보여주는 것이 pseudo R^2인데, 352개 품목의 평균이 2016년에 0.12로 나왔고, 품목별 편차가 커서 0.017에서 0.814 사이에 분포하였다. 2016, 2019-20년의 pseudo R^2의 평균은 각각 0.125, 0.126, 0.119로 나왔다. 식 (2)의 추정식의 경우 adjusted R^2의 전체 품목 평균이 2015-20년에 각각 0.075, 0.087, 0.09, 0.073으로 나왔다. 설명변수의 계수 값은 통계적으로 유의한 값들이 많았지만, 횡단면 자료이므로 R^2가 낮게 나왔다.

〈표 11〉 보정 전과 후의 가구당 소비액과 편향A와 편향B의 추정 값(단위: 천원)

	보정 전	편향A	편향B	보정 후	보정 전	편향A	편향B	보정 후
	2015				2016			
전체 소비지출	2,193	121	-109	2,181	2,165	98	-96	2,164
식료품 음료	307	11	-10	306	301	9	-9	300
주류 담배	29	1	-1	29	31	1	-1	31
의류 신발	137	8	-7	135	134	7	-5	132
주거수도광열	259	9	-6	256	256	7	-6	255
가정용품 등	89	12	-9	86	92	10	-9	91
보건	155	11	-11	155	157	9	-10	158
교통	271	23	-21	269	255	15	-18	258
통신	123	3	-4	124	120	2	-3	120
오락문화	128	13	-13	128	129	11	-12	130
교육	220	18	-16	218	214	16	-14	212
음식숙박	289	2	-2	289	291	2	-2	290
기타	186	10	-9	185	186	7	-7	186
	2019				2020			
전체 소비지출	2,459	335	-354	2,478	2,403	450	-446	2,400
식료품 음료	333	31	-30	332	381	50	-48	379
주류 담배	36	3	-4	36	38	5	-5	38
의류 신발	138	21	-17	133	118	25	-22	115
주거수도광열	279	23	-19	275	289	33	-30	285
가정용품 등	115	35	-32	112	127	54	-50	123
보건	202	33	-38	207	221	52	-59	228
교통	296	63	-60	293	289	89	-80	280
통신	123	12	-12	123	120	15	-18	123
오락문화	180	44	-53	189	140	44	-47	142
교육	205	38	-59	226	159	44	-44	159
음식숙박	346	7	-7	345	319	8	-8	319
기타	206	24	-24	206	304	32	-36	208

주: 1) 1인 가구를 포함한 전체 가구를 대상으로 한 것이다.
2) 가구당 월평균 명목 소비액(단위: 천원)이다.
자료: 통계청, 가계동향조사(RDC 접근)에서 산출.

앞에서 〈표 7〉과 〈표 10〉의 편향A와 편향B가 역수 관계에 있다는 점을 지적한 바 있는데, 〈표 11〉은 실제 추정결과에서도 확인됨을 보여준다. 그것은 통계청이 발표하는 가계동향조사의 통계 중에서 전체 가구를 대상으로 하는 평균 소비지출은 편향A와 편향B에도 불구하고 실제와 크게 다르지 않을 것이라는 점을 시사한다. 비목이나 품목 별로 보더라도 전체 가구를 대상으로 한다면 마찬가지다. 그리고 비목이나 품목별로도 두 편향이 상쇄된다면, 비목이나 품목별 구성비도 실제와 다르지 않게 된다.

이것은 그나마 다행스러운 일이다. 가계동향조사의 품목별 소비액 통계는 소비자물가지수를 산출할 때 가중치 정보로 활용되고 있는데, 만약 이들 편향으로 인해 품목별 구성이 왜곡되었다고 한다면 소비자물가지수에도 영향을 미쳤을 것이기 때문이다. 그렇지만 다행인 것은 여기서 그친다. 전체 가구를 대상으로 하여 소비지출의 평균을 구할 경우에는 두 편향이 상쇄되지만 그 외에는 그렇지 않으며, 가구별 소비지출의 평균이나 분포는 실제 값과 괴리되기 때문이다. 이하에서는 이 점을 좀더 천착해 보기로 한다.

2. 가구당 분포

먼저 개별 품목의 소비 분포를 보자. 앞의 자동차의 사례(〈표 1〉)에서 보았듯이 자동차를 구입한 가구의 소비액은 조사월수에 따라 최대 12배까지 커졌고(즉 편향A), 자동차를 구입하였음에도 불구하고 조사에서 누락된 가구가 있음(편향 B)을 감안하면 자동차 소비의 실제 분포는 〈표 1〉의 결과와는 크게 다를 것임을 알 수 있다. 가계동향조사 원자료(2015년)로 구한 자동차(신차 구입) 소비의 지니계수(구입>0인 가구로 한정)는 0.314였지만, 편향A를 보정하면 0.245로 되고, 편향B까지 보정하면 0.232로 떨어진다. 가계동향조사는 편향A와 B로 인해 소비 분포의 불평등을 상당히 과장하고 있음을 알 수 있다.

자동차 이외의 다른 품목은 어떨까? 먼저 2015년의 전체 349개 품목에서 편향A를 보정하면 311개 품목의 지니계수가 하락(평균 -0.012)하였고, 38개 품목이 상승(평균 0.003)한 것으로 나왔다.[15] 2020년에는 전체 354개 품목에서 279개 품목이 편향A의 보정으로 지니계수가 하락했지만, 75개 품목은 상승한 것으로 나왔다. 전자의 하락 폭의 평균은 -0.013이고 후자의 상승 폭의 평균 0.018로 나왔다. 다른 연도의 양상도 다르지 않다. 조사에서 누락된 편향B를 보정하면 구입=0이었던 가구의 소비지출이 (+)로 바뀌게 되는 것이므로 지니계수가 하락한다. 두 편향을 모두 보정할 경우 극히 일부 품목(2015년은 8개 2020년은 14개 품목)을 제외한 모든 품목에서 지니

15) 편향A의 과대평가를 보정하면 지니계수가 떨어질 것으로 생각되지만 상승하는 경우도 나올 수 있다. 예컨대 소비액이 100과 10인 가구와 소비=0인 가구가 존재한다고 해보자. 만약 편향A가 소비>0인 가구에 걸쳐 분포되었다면 편향A의 보정은 지니계수를 떨어뜨리지만, 만약 편향A가 소비액이 낮은 가구(즉 소비=10인 가구)에 집중된 경우에는 그것을 보정하면 지니계수가 높아진다.

계수가 하락한 것으로 나타났다. 2015-16년의 지니계수의 평균 하락 폭은 각각 -0.032와 -0.026이였지만 2019-20년에는 -0.069와 -0.097로 더욱 커졌다. 앞의 자동차의 사례가 예외적인 것이 아님을 알 수 있다. 가계동향조사에서 나타난 품목별 소비액은 실태와 동떨어져 있어 원자료 그대로 이용하면 그 분포에 관해 오도된 결론이 도출될 수 있다.

그러면 가구당 소비의 분포는 어떨까? 각 가구는 다양한 품목을 소비하고 있는데, 이들 소비 품목의 합계가 해당 가구의 실제 소비액에 근접할까? 앞에서 품목이나 비목 또는 전체의 평균을 구할 때 편향A와 B가 상쇄된다는 것은 전체 가구를 대상으로 할 경우 두 편향이 발생한 가구가 모두 포함되어 서로 대응하고 있기 때문이다. 그렇지만 가구당 소비지출에서는 그러한 상쇄가 일어나지는 않는다. 각 품목에서 편향A가 발생한 가구는 해당 품목의 지출이 있었던 가구이지만, 편향B가 발생한 가구는 해당 품목의 지출이 없는 가구이기 때문이다. 즉 두 편향은 서로 다른 가구에서 발생하기 때문에 동일 가구 내에서는 양자가 상쇄되지 않는다. 그리고 가구별로 소비되는 품목들이 다 다르다. 다만 개별 가구가 소비하는 품목이 많아지면 그 중에는 편향A를 가진 품목이 있는가 하면 편향B로 인해 조사에서 누락된 품목도 있어 가구별로 각 품목의 소비액을 합치면 두 편향이 부분적으로나마 상쇄되는 가능성이 높아진다.

그 의미를 좀더 부연하기 위해 조사월수=3(예컨대 12월, 1월, 2월)인 가구를 생각해 보자. 이 가구가 조사된 월에 구매한 것(예컨대 공동주택난방비)의 월 평균 소비액은 다른 달에도 매달 그만큼 소비한 것으로 간주된다(편향A). 그리고 여름철의 과일은 7-8월에 구입되었지만, 그 달에는 조사가 이루어지지 않았기 때문에 이 가구의 해당 과일의 구입은 0이 된다(즉 편향B). 즉 이 가구가 소비한 것으로 된 품목은, 매월 일정하게 구입되는 것이 아닌 한, 조사 월에 구입된 품목은 과대하게 반영되고 조사되지 않은 월에 구입된 품목은 아예 빠져 버리기 때문에 실제의 소비와는 상당히 다르다.[16] 그로 인해 이 가구의 소비 구성은 실제와는 동떨어진 것이 되지만, 소비액의 합계는 어느 정도 근접할 수 있다. 조사 월에 구입되어 과대평가된 품목의 합계와 조사되지 않은 월에 구입되어 과소평가된 품목의 합계가 어느 정도 상쇄되기 때문이다.[17] 따라서 가구당 전체 소비액의 분포는 개별 품목의 분포에 비해 편향이 줄어들

16) 이 가구와 달리 여름에 조사된 가구의 경우는 반대로 난방비의 구매=0이 되고 여름 과일의 월평균 소비액은 다른 달에도 그만큼 구매한 것으로 간주된다. 이런 방식으로 조사 월이 다른 모든 가구의 소비액을 평균하면 실제 값에 접근하지만, 가구당 소비의 분포는 그렇지 않다.

17) 이 가구의 경우 조사월수=3이지만 예시한 공동주택난방비는 구입빈도가 4이므로 실제보다

수 있다.

다만 두 편향의 크기가 일치하는 것은 우연에 불과하기 때문에 편향A가 편향B보다 커서 소비액이 실제보다 과대평가된 (+) 가구와 반대인 (-) 가구가 나타나게 된다. 여기서 편향A, B를 보정하면 (+) 가구의 소비액은 줄고, (-) 가구의 소비액이 늘어나므로 통상은 지니계수가 낮아질 것으로 기대할 수 있다. 다만 (+) 가구가 원래 소비액이 낮은 가구에 집중되었거나 (-) 가구가 원래 소비액이 높은 가구에 집중된 경우라면 보정된 결과 지니계수가 반대로 높아질 수도 있다. 결국 가구당 소비의 분포가 보정 전과 후에 어떻게 달라질 것인지에 관해서는 선험적으로 추론하기는 어렵다.

본고에서 구한 가구당 소비지출의 분포가 보정 전과 후에 어떻게 달라졌는지를 보기로 한다. 〈표 12〉는 가구당 전체(또는 비목별) 소비지출의 분포를 보정 전, 편향A를 보정한 경우, 편향A와 B 모두 보정한 경우로 나누어 각 지니계수를 제시하였다. 예컨대 2015년의 전체 소비지출의 경우 보정 전의 지니계수는 0.346이었는데, 편향A를 보정하면 0.344로, 편향B까지 보정하면 0.344으로 하락한다. 2020년에는 동 지니계수가 0.35 => 0.327 => 0.344로 하락하였다. 비목별로 나누어 보면 양상은 더욱 다양하다. 2020년을 보면 가정용품의 경우 지니계수가 0.657 => 0.611 => 0.595으로 크게 하락하였다. 오락문화의 경우도 그러하다. 이에 대해 음식숙박이나 교육의 경우에는 지니계수의 하락이 미미하였다. 보정에 의해 지니계수가 대체로 하락하였지만 그렇지 않은 경우도 있다. 식료품음료의 경우 0.337 => 0.342 => 0.318로 변해서 편향A의 보정으로 상승했다가 편향B의 보정으로 하락하였다. 교통의 경우는 반대로 0.642 => 0.564 => 0616로 변해서 편향A와 B의 보정으로 인해 하락했다가 상승하기도 하였다. 전체 소비지출의 보정 전과 후의 격차가 비목별로 본 경우에 비해 대체로 작다고 할 수 있지만, 그렇지 않은 경우도 존재한다.

---

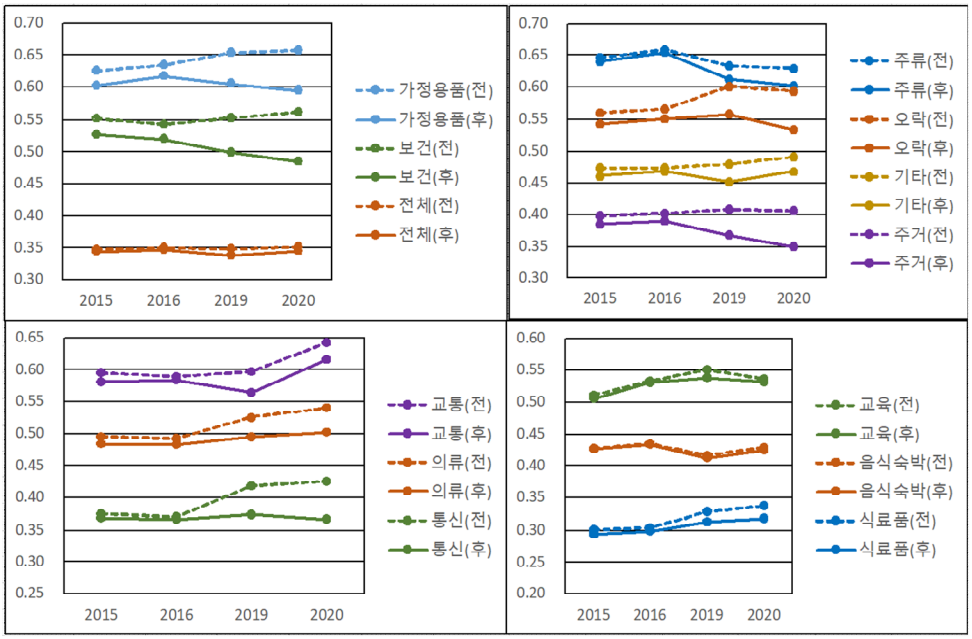
1. 341배(〈표 7〉 참조) 과대평가되어 있고, 여름 과일(수박과 참외)은 구입빈도가 2이므로 실제(이 가구의 여름 과일의 구입 여부는 전술한 방법으로 추정하고 구입한 것으로 추정된 경우의 소비액은 전체 가구의 평균으로 간주)보다 0.455배(〈표 10〉) 과소평가되어 있는 것으로 볼 수 있다. 이를 다른 모든 품목에 대해 적용하여 과대 또는 과소 평가된 금액을 구할 수 있다. 만약 고려되는 품목 수가 적을 경우에는 각 품목의 구입빈도나 구입금액에서 차이가 큰 경우가 생겨 과대 또는 과소 평가된 금액간의 격차가 커질 수 있다. 고려되는 품목 수가 늘어나면 그러한 outlier가 평탄화되어 과대 또는 과소 평가된 금액간의 격차가 줄어들 가능성이 높아진다. 다만 양자가 일치한다는 논리적 근거는 없다.

〈표 12〉 비목별 소비지출 지니계수의 보정 전후 비교

	2015					2016					2019					2020				
	보정 전	A	보정	AB	보정	보정 전	A	보정	AB	보정	보정 전	A	보정	AB	보정	보정 전	A	보정	AB	보정
전체 소비지출	0.346	0.344	0.344			0.349	0.348	0.347			0.347	0.335	0.338			0.350	0.327	0.344		
식료품 음료	0.301	0.306	0.294			0.303	0.307	0.299			0.328	0.334	0.313			0.337	0.342	0.318		
주류 담배	0.646	0.650	0.640			0.659	0.662	0.654			0.633	0.641	0.612			0.629	0.643	0.602		
의류 신발	0.495	0.499	0.484			0.492	0.496	0.483			0.526	0.528	0.495			0.540	0.549	0.502		
주거수도광열	0.398	0.388	0.385			0.402	0.395	0.390			0.408	0.390	0.367			0.406	0.381	0.350		
가정용품 등	0.625	0.611	0.602			0.634	0.625	0.617			0.653	0.627	0.604			0.657	0.611	0.595		
보건	0.551	0.546	0.527			0.542	0.537	0.519			0.552	0.535	0.498			0.562	0.535	0.485		
교통	0.595	0.579	0.582			0.589	0.580	0.584			0.597	0.548	0.564			0.642	0.564	0.616		
통신	0.375	0.365	0.369			0.370	0.363	0.367			0.417	0.374	0.374			0.425	0.368	0.366		
오락문화	0.560	0.555	0.543			0.566	0.559	0.551			0.601	0.580	0.558			0.593	0.561	0.534		
교육	0.511	0.518	0.506			0.533	0.537	0.531			0.551	0.567	0.538			0.536	0.552	0.532		
음식숙박	0.428	0.427	0.426			0.435	0.434	0.434			0.416	0.414	0.412			0.429	0.426	0.424		
기타	0.474	0.465	0.461			0.474	0.468	0.468			0.480	0.461	0.452			0.491	0.462	0.468		

주: A보정이란 편향A를 보정한 결과이고, AB보정은 편향A와 편향B를 모두 보정한 결과이다.
자료: 통계청, 가계동향조사(RDC 접근)에서 산출.

〈그림 1〉 보정 전과 후의 비목별 소비지출 지니계수의 추이



주: 1) 점선은 보정 전, 실선은 편향A, B 모두 보정한 후를 나타낸다. 범례의 순서를 그래프의 순서와 맞추었다.
2) 2016-19년은 3년 구간이며 그 사이에 연동표본 방식의 변경이 있었다는 점에 유의할 것.
자료: 통계청, 가계동향조사(RDC 접근)에서 산출.

〈그림 1〉은 〈표 12〉에 의거하여 보정 전과 후에 비목별 지니계수의 추이가 어떻게 달라졌는지를 보여준다. 점선이 보정 전, 실선이 보정 후를 보여주는데, 대체로 보정 전에 비해 보정 후의 지니계수가 더 낮으며, 2015-16년에 비해 2019-20년로 가면서 격차가 더 벌어졌다. 보정 전과 후의 격차가 큰 비목을 보면 가정용품, 보건, 오락문화 등인데 이들은 내구재이거나 구입빈도가 낮은 서비스 품목이 많아 편향이 상대적으로 컸다. 이에 대해 식료품이나 음식숙박 또는 교육은 보정 전과 후의 격차가 작게 나타났다. 비목별로 나누어 본 경우에는 비목 간 차이가 있지만 대체로 2016년 이전과 2019년 이후에 연동표본 방식의 변경으로 조사월수가 줄어든 영향으로 편향A, B가 더 커졌고 그로 인해 보정 전의 지니계수가 실제보다 과대해진 것으로 나타났다. 이에 비해 전체 소비지출을 보면 보정 전과 이후의 격차가 그만큼 크지 않다. 가구의 소비액을 비목으로 나누어 보는 경우보다 전체 품목을 대상으로 할 경우 편향A, B가 상쇄되는 가능성이 상대적으로 더 커진다는 점을 반영한 것으로 생각된다.

Ⅳ. 기존연구에 대한 함의

이상에서 살펴본 바와 같이 가계동향조사의 연간 자료를 이용하여 전체 가구를 대상으로 하는 소비의 평균을 구하면 편향A와 B가 상쇄되어 조사 결과와 괴리되지 않는다. 그렇지만 개별 가구별로 보면 두 편향은 상쇄되지 않으며, 가구가 소비하는 품목의 구성도 앞의 겨울의 난방비와 여름 과일의 사례에서 보았듯이 실제와는 상당히 달라진다. 가계조사의 이용자는 당연히 개별 가구의 통계치가 조사결과와 일치하며, 따라서 전체 가구는 물론 이를 하위 그룹으로 나눈 경우 그 하위그룹의 통계치 또한 조사된 결과와 일치할 것으로 기대한다. 그런데 가계동향조사의 연간 데이터의 경우에는 그러한 기대가 충족되지 않는다. 여기서는 본고가 제기한 문제가 가계동향조사를 이용하여 이루어진 기존 연구에 대해 어떤 함의를 갖는지 간단히 언급하기로 한다.

먼저 개별 품목을 대상으로 하는 가구별 소비 분포가 어떻게 왜곡되는지는 알기 쉽다. 전술한 가구의 난방비를 사례로 들면, 난방비를 지출한 것으로 되어 있는 가구는 겨울에 조사된 가구로 한정되고, 그것도 실제로 지출한 난방비보다 몇 배(만약 난방이 이루어지는 겨울이 4개월이라면 12/4=3배)가 더 많은 난방비를 지출한 것으로 된다(편향A). 그리고 다른 계절에 조사된 가구는 겨울에 난방비가 지출됨에도 불구하고 난방비 지출=0(편향B)이 된다. 이런 데이터를 그대로 이용하여 난방에 사용되는

연료에 대한 과세를 강화할 경우 가계에 미치는 세 부담의 분포를 추정한다고 해 보자. 그 경우 우연히 겨울에 조사된 가구들만 난방비의 세 부담을 지게 되고(그것도 실제보다 몇 배 많게), 여름에 조사된 가구는 세 부담이 0으로 추정된다. 실제로는 거의 모든 가구가 크든 작든 난방비를 지출하고 있는 현실과 동떨어져 있다. 즉 가계동향조사를 보정하지 않은 채로 품목별 소비 분포를 그대로 이용하여 세 부담을 구하게 되면 그 불평등이 실제보다 상당히 과대평가될 수 있다.

난방비는 소비의 계절성이 큰 경우인데 그렇지 않은 경우는 어떤가? 자동차와 같이 몇 년에 한 번 구입되는 내구재의 경우에는 가구의 조사 월수에 따라 최대 12배까지 과대해지는 편향이 생기고 그 가구별 분포도 실제보다 과장된다는 점은 이미 전술한 바와 같다. 그러면 계절성이 없거나 내구재도 아닌 경우에는 어떤가? 만약 매달 큰 변동 없이 소비되는 품목이 있다면 왜곡이 발생하지 않는다. 그렇지만 앞에서 소비품목의 구입빈도를 보면(〈표 5〉 참조) 매월 구입이 이루어지는 것은 오히려 예외적임을 알 수 있다. 즉 거의 모든 품목에서 소비 패턴의 계절성이나 내구재 성격 여하에 따라 정도의 차이가 있겠지만 위의 난방비나 자동차에서 본 바와 같이 각 품목의 가구당 소비 분포가 실제와 괴리되는 문제가 발생한다.

개별 품목(또는 관련되는 몇 가지 품목)에 초점을 맞추어 그 소비분포를 살펴본 연구는 많지만 여기서는 한 두 연구만을 예시하기로 한다. 대표적인 것으로서 음(-)의 외부효과가 있는 담배, 주류, 유류에 대해서는 높은 세금이 부과되고 있는데, 그에 대한 과세를 강화하는 것이 소득분배에 어떤 영향을 미치는가를 구명한 연구들이다(예컨대 전승훈, 2013). 또는 에너지 세제를 개편하거나 탄소세를 도입할 경우에 가계의 세 부담이 계층별 또는 가구 유형별로 어떻게 달라지는가를 구명한 연구도 그러하다(예컨대 김승래, 2019). 이들 연구는 모두 과세되고 있는 개별 품목(군)의 가구당 소비(그에 대한 세 부담)에 초점을 맞추고 있으며, 가계동향조사를 데이터로 이용하고 있다. 따라서 이들 연구들이 도출한 가구당 세 부담의 분포는 가계동향조사의 연간 자료가 안고 있는 문제를 반영하여 그 불평등이 실제보다 과대하게 평가되었다고 생각된다. 이것은 기초 자료의 왜곡이 학술연구는 물론 정책적 함의를 왜곡할 수 있음을 보여준다.

이에 대해 가구당 소비지출의 불평등에 초점을 맞춘 연구(김대일, 2007; 박기백, 2017)의 경우에는 개별 품목의 소비가 아니라 전체 소비를 대상으로 한다. 이들이 소비에 초점을 맞추어 불평등을 살펴보는 것은 소비는 소득처럼 과소 보고할 가능성이 낮고, 소득보다 소비가 경제적 후생 지표로서 더 적합하다고 보기 때문이다. 그런데

본고에서 지적한 문제를 감안하면 가계동향조사의 가구당 소비지출을 그대로 이용할 경우 실제와 얼마나 괴리되는가를 추가로 고려할 필요가 제기된다. 가구별 소비액은 여러 품목의 지출액이 합쳐진 것이기 때문에 그 중에는 편향A 또는 편향B가 포함된 품목이 섞여 있을 수 있어 합계를 하면 두 편향이 부분적으로 상쇄되는 효과가 있다. 그로 인해 앞의 개별 품목에서 나타난 바와 같은 분포의 왜곡은 상대적으로 덜 할 것으로 생각된다. 그렇지만 전술했듯이 동일 가구에서 두 편향의 크기가 접근하여 상쇄된다는 논리적 근거는 없다. 그리고 박기백(2017)은 비목별로 나누어 소비지출 불평등의 수준이나 추이를 논하고 있지만, 거기에는 본고(〈그림 1〉)에서 살펴본 바와 같이 보정 이전과 이후의 갭이 포함되어 있다. 최근에는 연동표본 방식이 개편되면서 그러한 갭이 더욱 커졌다.

가계조사를 이용할 때 전체 가구의 하위 그룹에 초점을 맞추는 연구가 적지 않다. 예컨대 전체 가구를 소득계층이나 연령 또는 가구 유형별로 나누어 볼 수 있고, 저소득층 가구나 노인 가구에 초점을 맞춘 연구도 있다(예컨대 이현주, 2016). 그런데 가계동향조사를 이용하여 접근할 경우 저소득층이나 노인가구와 같은 하위 그룹으로 한정한 통계치가 실제 조사된 결과와 일치한다는 보장이 없다. 전체 가구를 대상으로 할 때에는 편향A와 B가 서로 대응하여 상쇄될 수 있었지만, 하위 그룹으로 나눌 경우 편향A가 발생한 품목을 소비한 가구와 편향B가 발생한 품목을 소비한 가구가 동일한 하위 그룹에 속하는 것은 우연에 불과하기 때문이다.

결국 가계동향조사의 소비 통계에서 조사결과와 일치할 것으로 기대되는 것은 전체 가구를 대상으로 하는 평균값뿐이다. 개별 가구의 소비 통계가 실제의 조사결과와 괴리되어 있기 때문에 가구별 분포를 구해도 그 또한 실제의 분포와 달라진다. 이것은 가계동향조사를 마이크로 데이터로 이용하는데 커다란 한계라 하지 않을 수 없다. 본고가 시도한 것은 편향 A와 B를 보정함으로써 개별 가구의 소비 통계를 실제의 조사결과에 접근시키고자 한 것이다. 이를 통해 위에서 지적한 가계동향조사 연간 데이터에서 제기되는 문제를 해소하고자 하였다.

이상은 가계동향조사의 소비에 초점을 맞추어 서술해 왔지만, 거기에서 지적된 문제는 소득에 대해서도 마찬가지로 지적할 수 있다. 소득은 매월 발생되는 경우가 많겠지만 그렇지 않은 경우도 적지 않다. 예컨대 근로소득 중의 상여금이 그러하며, 사업소득이나 이전소득 기타 소득 등에서도 그러리라 생각된다. 가구주에 비해 배우자나 기타 가구원의 소득에는 그러한 경향이 더 클 것으로 생각된다. 그 경우 소비에서 지적한 편향A와 편향B가 소득에서도 마찬가지로 발생하게 된다. 본고는 또한 연간

자료를 대상으로 설명해 왔지만, 분기 자료에 대해서도 동일한 문제를 제기할 수 있다. 분기 자료의 경우 조사월수가 1-3월로 연간 자료(1-12월)보다 적어 편향A와 편향B가 최대 3배와 1/3배로 벌어진다는 점에서 연간 자료(최대 12배와 1/12배)와 차이가 있지만, 본고의 설명이 분기 자료에 대해서도 마찬가지로 적용될 수 있다. 그리고 이 문제는 본고가 다룬 연도에 국한되지 않으며, 연동표본으로 바뀐 2005년 이후의 통계뿐만 아니라 앞으로 발표되는 통계에도 해당한다. 따라서 그 동안 가계동향조사의 소득이나 소비 데이터를 기초 자료로 이용한 많은 연구들이 이 문제로부터 자유롭지 않다.

그런데 본고가 소득 문제를 다루지 못한 것은 가계동향조사의 월간 자료를 이용하는데 제약이 있었기 때문이다. 만약 월간 자료를 제대로 이용할 수 있게 되면 본고의 추계를 개선하거나 대체할 수 있는 가능성이 열릴 것으로 생각된다. 예컨대 추계에 이용된 구입빈도는 전술했듯이 월간 자료에서 가져왔지만 가구 식별 정보가 없이 추정된 값인데, 이를 개선할 수 있을 것이다. 그리고 본고에서 편향A의 보정은 실태에 근접할 것으로 생각되지만, 편향B의 보정은 전술했듯이 어느 정도의 오차 발생이 불가피하다. 월간 자료를 이용하게 되면 전술한 편향이 없는 월별 가구 소득이나 소비의 분포를 구할 수 있다. 다만 연동표본으로 인해 매월 이루어지는 샘플의 교체가 분배지표에 미치는 영향을 어떻게 처리할 것인지를 포함해서 월별 지표를 분기 및 연간 지표로 만드는 방식에 관해서는 추가적인 검토가 필요하다.

Ⅴ. 맺음말

통계청이 가계동향조사의 분기 또는 연간 통계를 구할 때 월간 자료에서 해당 월의 조사결과를 단순 평균하는 방식으로 산출하고 있다. 매달 지출이 반복되는 품목이라면 이 방식에 문제가 없지만 그렇지 않은 품목에서는 왜곡이 발생한다. 조사월수와 품목의 구입빈도에 따라 연간 통계는 실제보다 12배(분기 통계는 3배)까지 과대해지는 편향이 생기는 한편, 조사 월에 구입되지 않은 품목은 조사에서 아예 누락되어 버린다. 전체 가구를 대상으로 소비의 평균을 구할 때에는 이러한 과대평가와 누락에 의한 과소평가가 서로 상쇄되는 것으로 나왔다. 그렇지만 가구당 평균은 그렇지 않아 가구 간 소비 격차가 실제보다 더 벌어진 것으로 나오며, 그러한 왜곡이 최근에 더욱 커졌다.

본고는 이러한 편향을 보정하는 시도를 하였고 그 편향이 어느 정도인가를 보였지

만, 통계 이용자의 입장에서는 보정이 필요하다는 것 자체가 여간 불편한 일이 아니다. 결국 통계청이 이 문제의 해소에 적극 나설 수밖에 없다고 생각한다. 여기서는 통계청이 지금 조치할 수 있는 일과 통계조사 방법의 개편까지 포함해서 좀 더 시간을 두고 고려할 과제가 무엇인지 언급해 두고자 한다.

먼저 분기 및 연간 데이터에서 발생한 편향은 결국 월간 데이터로부터 평균을 산출하는 방식에서 온 것이므로 연구자들에게 그러한 편향이 발생하기 이전의 월간 데이터를 직접 이용할 수 있도록 할 필요가 있다. 이를 위해서는 적어도 두 가지 조치가 요망된다. 하나는 월간 데이터의 가구 식별 id가 실제의 가구가 아니라는 전술한 문제이다. 이로 인해 현재의 월간 데이터는 가계조사로서 활용이 어렵다. 이 데이터에는 이미 가구의 익명성을 보장하는 조치가 되어 있기 때문에 이러한 가구 id에 관한 정보를 굳이 제한하는 이유를 알기 어렵다. 또 하나는 월간 데이터에 대한 접근의 장벽을 낮출 필요가 있다는 것이다. 현재 월간 데이터뿐만 아니라 저자의 요청에 의해 제공된 가구별 조사월수에 관한 정보도 통계이용센터(RDC)에서만 이용할 수 있게 제한을 두고 있다. 그렇지만 RDC 이용자가 극소수임을 감안하면 이러한 조치는 사실상 월간 데이터를 이용하지 말라는 것과 다르지 않다. 월간 데이터의 이용은 분기 및 연간 데이터의 문제로 인해 불가피해진 것이므로 이에 대해 적어도 원격접근(RAS)이 가능하도록 완화하는 것이 불가결하다. 가계동향조사의 정보를 보다 투명하게 하고 그에 대한 접근을 더 용이하게 하는 것이 이 통계에 대해 제기될 수 있는 불신을 덜 수 있는 길이라고 생각한다.

가계조사가 연동표본 방식으로 이루어지는 경우 조사월수의 차이로 인해 가구별 통계에 편향이 생기지 않도록 설계를 고안할 필요가 있다. 가계동향조사와 대응하는 미국(BLS)의 Consumer Expenditure Survey의 방식을 보면 가계부 방식(Diary survey)과 면접조사(Interview survey)를 병행하고 있다. 그 중 면접조사의 경우 분기로 나누어 4번 조사하지만 조사 대상을 해당 월을 포함한 앞의 3개월로 하고 있는 점이 주목된다. 이 방식은 조사월수는 적지만 모든 월을 커버하고 있어 본고에서 언급한 편향이 발생하지 않는 장점이 있다. 다만 지난 3개월에 걸친 지출에 관한 회고조사는 내구재와 같이 지출의 규모가 크거나 규칙적으로 지출이 이루어지는 경우에는 가능하지만 소소한 일상적인 지출의 경우에는 개략적인 파악을 넘기 어려운 한계도 있다. 그렇지만 CES의 경우 면접조사가 가계부 방식에 비해 우월한 것으로 평가되고 있다(Bee, Meyer and Sullivan, 2015).

장기적으로는 소비지출에 관한 조사는 결국 신용카드의 지출 정보와 같은 빅 데이

터를 이용하는 방향으로 갈 수밖에 없다고 생각된다. 본고에서 제기한 문제는 연간 자료가 조사된 결과와 괴리되어 있다는 점이며, 조사결과 그 자체가 실태를 얼마나 잘 반영하는지에 관해서는 여기서 문제를 삼지 않았다. 그렇지만 가계부 방식이든 면접조사 방식이든 가계의 소비지출을 정확히 파악하는데 한계가 있으며(Carroll et al., 2015), 앞으로 신용카드의 정보가 이를 보완하는데 활용될 가능성이 높다. 예컨대 통계청의 가계금융복지조사의 소득이나 비 소비지출 데이터가 행정자료에 의해 보정되어 정확성을 높인 것처럼 가계동향조사의 소비지출을 신용카드 정보로 보정하는 것을 생각해 볼 수 있다. 신용카드 정보는 분기별 발표와 같은 시의성이 요구되는 경우에도 대응할 수 있으며, 모든 월을 다 조사하는 것도 가능하기 때문에 전술한 편향을 피할 수 있다. 다만 이러한 접근은 현재에도 기술적으로는 문제가 없지만 개인정보 보호와 같은 제도적으로 해결해야 하는 과제를 안고 있다고 생각된다.

■ 참 고 문 헌

1. 김낙년 · 김종일, "한국 소득분배 지표의 재검토," 『한국경제의 분석』, 제19권 제2호, 금융연구원, 2013, pp. 1-64.
2. 김대일, "불평등도 지표로서의 소득과 소비의 비교," 『노동경제논집』, 제30권 제3호, 금융연구원, 2007, pp. 77-102.
3. 김승래, 『에너지세제 분석을 위한 기초연구: 산업연관표 구축 및 시뮬레이션모형 개발』, 국회예산정책처, 2019년도 연구용역보고서, 2019.
4. 박기백, "소비 · 소비 불평등의 관계 및 소비불평등 분해," 『재정정책논집』, 제19집 제3호, 2017, pp. 149-179.
5. 윤연옥 · 김규영 · 이명호, "통계청 가구부문 조사의 표본설계," 『조사연구』, 제5권 제1호, 2004, pp. 103-130.
6. 이현주, 『저소득층 가구소비 변화와 정책적 함의』, 한국보건사회연구원, 2016.
7. 전승훈, "담배소비세 및 담배기격 인상에 따른 담배소비 및 세 부담 변화: 소득분위별 분석," 『재정정책논집』, 제15집 제4호, 2013, pp. 89-121.
8. Bee, A., B. Meyer, and J. Sullivan, "The Validity of Consumption Data: Are the Consumer Expenditure Interview and Diary Surveys Informative?" in Carroll C., T. Crossley, and J. Sabelhaus, eds., *Improving the Measurement of Consumer Expenditures*, University of Chicago Press, 2015.
9. Carroll C., T. Crossley, and J. Sabelhaus, eds., *Improving the Measurement of Consumer Expenditures*, University of Chicago Press, 2015.
10. U.S. Bureau of Labor Statistics, "Consumer Expenditure Surveys" (https://www.bls.gov/cex/).

Problems and Corrections of the *Household Income and Expenditure Survey*

Nak Nyeon Kim*

Abstract

Since the *Household Income and Expenditure Survey* is designed as rotation sampling in which a certain percentage of the sample is replaced every month, the number of months each sample is surveyed in a year spans 1-12 months (1-6 months after 2020). Statistics Korea calculates quarterly or annual statistics by simply averaging the survey results for the month from monthly data. This is fine for items with similar recurring monthly expenditures, but distortion occurs for items that do not. Depending on the number of months surveyed and the frequency of purchase of items, the annual statistics tend to be exaggerated by up to 12 times (three times in the case of quarterly statistics) than the actual ones, while items not purchased in the survey month are completely omitted from the survey. When calculating the overall average, it was found that such overestimation and underestimation due to omission offset each other. However, the average per household is not, so the consumption gap between households is wider than it really is, and the distortion has grown bigger in recent years. In this paper, the bias of items consumed by each household included in the annual data for 2015-16 and 2019-20 was corrected in a reasonable way, and the results were compared with the existing micro data.

Key Words: *Household Income and Expenditure Survey*, rotation sampling, consumption distribution

JEL Classification: D0, D3, R2

Received: Dec. 10, 2021. Revised: Jan. 7, 2022. Accepted: Feb. 15, 2022.

* Professor Emeritus, Department of Economics, Dongguk University, 30, Pildong-ro 1gil, Jung-gu, Seoul 04620, Korea, Phone: +82-2-2260-3273, e-mail: nnkim@dongguk.edu