

복잡표본조사에서 가중치 산출

2014. 5

한국갤럽조사연구소 자문교수

이계오

목 차

- 1 표본조사에서 가중치 필요성
- 2 가중치 의미와 계산
- 3 총화집락추출에서 가중치 분석 사례
- 4 노동패널조사에서 가중치 계산
- 5 12차년도 노동패널의 가중치 산출
- 6 14차년도 노동패널의 가중치 산출
- 7 결 언

1. 표본조사에서 가중치 필요성

- 대부분의 표본조사는 현실적 환경과 여건으로 복잡표본조사설계로 자료수집
 - 대부분의 통계학 이론은 단순임의추출법을 가정함
 - 단순임의추출법에 의한 설계는 자체가중설계임
 - 조사비용의 제한과 정확한 자료수집
 - 조사내용 특성의 다양한 변인에 의한 영향
 - 모집단에 대한 다양한 정보 수집 및 이용 가능
 - 자료처리 기법과 프로그램의 다양화

1. 표본조사에서 가중치 필요성

● 복잡표본조사 설계(Complex sample design)의 특징

☑ 조사단위별로 상이한 추출확률

- 표본추출방법에 따라 조사단위들이 표본으로 추출될 확률 상이
- 다단계추출법과 다양한 추출법의 적용으로 표본추출확률 상이

☑ 무응답(nonresponse)발생

- 단위 무응답(unit nonresponse)
- 항목 무응답(item nonresponse)

☑ 불완전한 모집단 정보 이용(noncoverage 발생)

- 조사모집단과 목표모집단간의 차이 존재
- 표본추출틀(sampling frame)의 문제
- 표본설계과정에서 고려되지 않은 중요변수의 영향

1. 표본조사에서 가중치 필요성

● 표본조사자료의 오류 발생 요인

☑ 무응답(non response)

- 단위무응답(unit nonresponse)
- 항목무응답(item nonresponse)

☑ 미포함의 문제(noncoverage error)

- 조사모집단과 목표모집단간의 차이 발생
- 시간, 개념, 리스트 관리 등
- 표본추출틀의 영향을 반영
- 외부행정자료를 이용한 포괄성과 일치성 문제 보완

2. 가중치 의미와 계산

● 자체가중설계 사례와 설명

☑ 단순임의추출법 : 100명 회사원 중 20명 SRS로 추출 초과근무 시간 합계 추정

$$\hat{Y} = \sum_{i=1}^{20} W_i y_i = \sum_{i=1}^{20} \frac{100}{20} \cdot y_i = \frac{100}{20} \sum_{i=1}^{20} y_i$$

☑ 비례배분 층화추출법 : H개 층으로 모집단 구성

➤ 각 층에서 모집단과 표본크기 : N_h 와 n_h

➤ 각 층 표본배분은 비례배분 : $n_h = n \cdot N_h / N$

➤ 모총계 추정량 : $\widehat{Y}_{st} = \sum_{h=1}^H \sum_{i=1}^{n_h} W_{hi} y_{hi} = \sum_{h=1}^H \frac{N_h}{n_h} \sum_{i=1}^{n_h} y_{hi} = \frac{N}{n} \sum_{h=1}^H \sum_{i=1}^{n_h} y_{hi}$

2. 가중치 의미와 계산

● 자체가중설계 사례와 설명(계속)

☑ 크기, 비례확률추출법(PPS : probability proportionate to size sampling)

- A 개 조사구로 구성된 시에서 a 개 조사구를 가구수 기준 확률비례추출법 추출 선정된 조사구에서 b 가구 랜덤추출법 선정한 후 가구내의 모든 가구원 조사
- i 조사구내의 가구 수를 B_i 라면 PPS 1단계추출에서 a 개 조사구의 추출확률은 $a \cdot \frac{B_i}{\sum_{i=1}^A B_i}$ 이고 2단계추출에서 선정된 조사구내에서 b 가구의 추출확률은 $\frac{b}{B_i}$ 임
- 전체적인 표본추출확률 : $f = f_1 \cdot f_{2i} = a \cdot \frac{B_i}{\sum_{i=1}^A B_i} \cdot \frac{b}{B_i} = \frac{ab}{\sum_{i=1}^A B_i} = \frac{n}{N}$

2. 가중치 의미와 계산

1) 확률추출, 표본추출률 및 가중치

● 확률추출의 의미

- 모집단의 모든 추출단위들의 추출확률이 사전에 알려져 있으며, 모집단내의 모든 추출단위들이 '0' 이 아닌 확률로 추출된 가능성이 있는 상황에서 추출된다는 의미

● 표본추출률(sampling fraction)

- 모집단에서 표본으로 추출되는 추출단위의 비율 의미, $f=n/N$ (n 은 표본크기, N 은 모집단 크기)
- 단순확률추출에서는 조사단위의 추출률 = 추출 확률
- 표본조사론 관점에서 표본추출률을 높이는 것보다 확률추출법을 이용하는 표본설계가 더 중요

● 표본가중치(sample weight)

- 각 표본의 선택확률(=추출률)로 반영하며, 개별 표본마다 계산
- 유한 모집단 표본의 원소와 연결된 관찰단위의 대표성의 정도를 위해 계산된 값

2. 가중치 의미와 계산

● 표본가중치의 기본개념

- 표본가중치는 개별표본단위의 관찰값과 관련된 수치적인 통계적 측도이며 표본으로 추출된 단위가 모집단을 얼마나 대표하는가 하는 크기를 나타낸 개념
- 표본분포의 불균형을 보정(adjustment) 하는 척도
- 비조정가중치(the unadjusted weight: 기본개념)
 - 관찰단위와 추출단위를 연결해주는 선택확률의 역수
 - 표본추출확률이 p_i 인 개별표본단위는 모집단 원소를 $1/p_i$ 만큼 대표하며, 분석 시 개별 표본 단위마다 $w_i = 1/p_i$ 배의 가중치를 부여한다는 의미
 - 기본 가중치(base weight): 추출률의 역수로 계산, $w = N/n$ (단순확률추출의 경우)

[참고] 가중치의 합 = 모집단의 크기(일반적인 경우)

- 확대승수로도 표현

(예) 추출확률이 1/20인 표본단위는 “모집단에서 해당단위 20개를 대표(represent)를 한다” 고 표현할 수 있음

2. 가중치 의미와 계산

• 가중치를 부여하는 이유

- 불균등 선택확률 (unequal selection probability)을 보정하기 위해
- 무응답(non-response)을 보정하기 위해
- 모집단에서 이미 알려져 있는 특정 변수와 분포 (예 : 성, 연령, 지역 등)와 가중표본결과의 분포를 일치시키기 위해 조정하는 벤치마킹(benchmarking)/ 사후층화(post-stratification) 보정
 - 표본 추정치의 정확도를 향상시키기 위해
 - 무응답 및 미포함(non-coverage) 표본추출틀을 보정하기 위해

2. 가중치 의미와 계산

• 가중치 부여의 장점

- 편향의 감소와 비용의 절감
- 영역별로 서로 다른 추출률로 표본을 추출하여 모집단 규모나 추출률이 작거나 낮은 영역의 추출률을 높여 전체적으로 정확도를 향상
- 중복, 비포괄성 등의 불완전한 추출률 문제 해소
- 조사에서 발생하는 단위 무응답으로 인한 편향 문제를 해소

2. 가중치 의미와 계산

- **확률표본을 이용한 추론에서는 표본을 선택한 표본설계의 내용을 반드시 반영하여 추정 및 분석을 수행**

- 변동선택 확률(varying selection probabilities, 불균등추출확률)
- 층화(stratification)
- 집락추출(cluster sampling)

[참고] 확률표본자료의 분석 시 고려해야 할 사항

- 기본단위의 추출확률이 다른 경우 가중치를 반영하여 추정해야 편향이 제거될 수 있음
(전문통계분석 패키지 사용해야 함)

[참고] 등확률추출(EPSEM)인 경우는 일반적인 통계패키지로 분석해도 추정치에 대해 편향이 없는 결과를 도출
(분산은 과소 추정 : 편향)

- 표본추출법에 따라 추출단위 추출확률이 다를 수 있으며, 추출확률이 다른 경우는 자료분석을 위한 전문통계분석패키지를 이용하는 것이 바람직함. (SUDAAN, Strata, SAS survey procedure, SPSS complex 등)

2. 가중치 의미와 계산

2) 표본가중치의 계산(부여 과정)

- 가중치 계산 시 고려해야 할 사항

1. 표본설계 및 추출과 관련된 정보

2. 모집단의 중요한 부차그룹(subgroup)에 대한 응답률(response rates) 자료

3. 표본설계 과정에서 고려되지 않은 중요 인구사회경제학적 변수의 부차그룹에 대한 모집단 자료

2. 가중치 의미와 계산

- 가중치의 표현과 부여과정
- 불균등 선택확률의 가중치
 - 표본추출단위와 관련된 실제표본의 선택과정을 반영
- 표본의 무응답 조정을 위한 가중치
 - 무응답으로 인한 원표본(original sample)과의 불균형을 조정하기 위해 가중 총 조정치(weighting class adjustment: WCA)를 계산하여 반영
- 모집단의 무응답 및 비포함(non-coverage) 조정을 위한 사후층화 가중치
 - 중요한 의미를 지닌 특정변수에 대한 모집단의 분포와 표본분포의 불균형을 수정하기 위해 사후층화조정(PCA) 가중치를 계산하여 반영
- 전체가중값 $W = W_1 \cdot W_{2 \cdot 1} \cdot W_{3 \cdot 21}$

2. 가중치 의미와 계산

• 무응답 보정을 위한 가중치

- 표본의 부차그룹마다 가중응답률 계산
- 가중응답률: (적합한 표본이 완료된 가중 응답자 수) / (표본으로 적합한 가중 표본 수)
- 무응답 보정을 위해 부차그룹의 응답률 역수를 가중치로 사용

➤ 분석 가중값: $W_a = W_{1a} \cdot W_{2a}$

여기서 W_{1a} 는 불균등 선택확률에 의한 가중치

· W_{2a} 는 무응답 보정 가중치

2. 가중치 의미와 계산

• 사후층화 보정 가중치(벤치마킹 가중치)

- 모집단 분포의 변화를 반영한 가중치
 - 목표 모집단 크기와 조사 모집단 크기간의 차이를 반영한 가중치
 - 표본 가중합과 모집단 크기와의 차이를 반영한 가중치
- 중요한 인구학적 변수의 모집단 분포와 가중 표본 분포와의 차이를 보정하는 가중치
 - 분석 가중값: $W''_a = W_a \cdot W_{3a}$

여기서 W_a 는 표본의 불균등 선택 확률과 무응답 보정을 반영한 가중치

2. 가중치 의미와 계산

- Example : 불균등 선택확률에 대한 가중치
- 240 가구의 모집단에서 6가구를 추출했을 경우에는 추출률 $f = 1/40$ 인 epsem 표본으로 선정한 것으로 간주하고, 추출된 가구에서 1명의 성인을 랜덤으로 추출하는 표본설계를 가정하여 취업여부 Z_a 와 Y_a 근무경력을 조사하는 경우
 - 추출률 및 가중값(W_a)

$$p(\alpha\beta) = p(\alpha)p(\beta|\alpha) = \left(\frac{1}{40}\right)\left(\frac{1}{\beta_\alpha}\right) = 1/w_a$$

여기서 β_α 는 α 가구에서 성인수

2. 가중치 의미와 계산

가상자료

표본가구	성인수 (B_a)	W_a	y_a (경력)	Z_a (취업)	$W_a y_a$	$W_a Z_a$
1	2	80	7	1	560	80
2	1	40	9	1	360	40
3	3	120	6	0	720	0
4	2	80	8	1	640	80
5	3	120	4	0	480	0
6	4	160	3	0	480	0
합계	15	600	37	3	3,240	200

2. 가중치 의미와 계산

추정결과

➤ 가중값을 반영하지 않은 추정 결과

• 근무경력 평균: $\bar{y} = \frac{\sum y_a}{6} = \frac{37}{6} = 6.17$

• 취업여부 가구 비율: $\bar{z} = \frac{\sum z_a}{6} = \frac{6}{3} = 0.5$

• 전체 취업자 수: $\hat{\tau} = N \cdot \bar{z} = ((6 \cdot 40) \cdot (15/6)) \cdot 0.5 = 300$

➤ 가중값을 반영한 추정 결과

• 근무경력 평균: $\bar{y}_w = \frac{\sum w_a y_a}{\sum w_a} = \frac{3240}{600} = 5.4$

• 취업여부 가구 비율: $\bar{z}_w = \frac{\sum w_a z_a}{\sum w_a} = \frac{200}{600} = 0.333$

• 전체 취업자 수 : $\hat{\tau} = \sum w_a z_a = 200$

2. 가중치 의미와 계산

- **Example : 불균등 선택확률에 대한 가중치 (2)**
- **어느 지역을 도시와 농촌으로 구분하여 층화 다단계로 표본 설계한다고 가정**
 - 도시지역의 추출률 $1/126.3$ 농촌지역의 추출률 $1/252.6$ 으로 도시지역을 2배 많게 추출
 - 선택된 가구에서 성인 1명을 랜덤으로 추출하여 취업여부를 조사
- **가중값**
 - 도시지역 : $p(\alpha\beta) = p(\alpha)p(\beta|\alpha) = \left(\frac{1}{126.3}\right)\left(\frac{1}{\beta_\alpha}\right) = 1/w_\alpha$
 - 농촌지역 : $p(\alpha\beta) = p(\alpha)p(\beta|\alpha) = \left(\frac{1}{252.6}\right)\left(\frac{1}{\beta_\alpha}\right) = 1/w_\alpha$

2. 가중치 의미와 계산

가상자료

총	표본가구	성인수(B_a)	W_a	Z_a	$W_a Z_a$
도시	1	2	252.6	1	252.6
	2	2	252.6	1	252.6
	3	1	126.3	0	0
농촌	1	4	1010.4	0	0
	2	2	505.2	1	505.2
	3	3	757.8	0	0
합계			2904.9	3	1010.4

2. 가중치 의미와 계산

추정결과

➤ 가중값을 반영하지 않은 추정 결과 (층화 효과)

- 취업 여부 비율 : $\bar{Z} = \frac{\sum Z_a}{6} = \frac{6}{3} = 0.5$

$$\bar{Z}_{st} = \frac{\sum N_h z_h}{\sum N_h} = \frac{631.5 \cdot \frac{2}{3} + 2273.4 \cdot \frac{1}{3}}{2904.9} = 0.406$$

- 전체 취업자 수 : $\hat{t} = N \cdot \bar{Z} = 2904.9 \cdot 0.5 = 1452.45$
 $\hat{t}_{st} = \sum N_h \bar{Z}_h = 631.5 \cdot \frac{2}{3} + 2273.4 \cdot \frac{1}{3} = 1178.8$

➤ 가중값을 반영한 추정 결과

- 취업 여부 가구 비율 : $\bar{Z}_w = \frac{\sum W_a Z_a}{\sum W_a} = \frac{101.4}{2904.9} = 0.348$

- 전체 취업자 수 : $\hat{t} = \sum W_a z_a = 1010.4$

2. 가중치 의미와 계산

- 비례가중값 (Proportional weights)
- 모집단 총합계를 추정하기 위해 표본단위의 선택확률 역수로 가중값을 부여
- 모평균과 모비율 추정을 위해 표본단위의 선택확률 역수에 비례한 가중값을 부여
 - 도시지역 추출률 : $p(\alpha\beta) = p(\alpha)p(\beta|\alpha) = \left(\frac{1}{126.3}\right) \cdot \left(\frac{1}{\beta_a}\right) = f_1 \cdot \left(\frac{1}{B_a}\right)$
 - 농촌지역 추출률 : $p(\alpha\beta) = p(\alpha)[p(\beta|\alpha)] = \left(\frac{1}{252.6}\right) \cdot \left(\frac{1}{B_a}\right) = \left(\frac{f_1}{2}\right) \cdot \left(\frac{1}{B_a}\right)$
 - 가중값 : 도시지역 = B_a 농촌지역 = $2B_a$

2. 가중치 의미와 계산

가상자료 : 비례 가중

총	표본가구	성인수(B_a)	W_a	Z_a	$W_a Z_a$
도시	1	2	2	1	2
	2	2	2	1	2
	3	1	1	0	0
농촌	1	4	8	0	0
	2	2	4	1	4
	3	3	6	0	0
합계			23	3	8

추정결과

가중값을 반영한 추정 결과

- 취업 여부 비율 : $\bar{z}_w = \frac{\sum W_a z_a}{\sum W_a} = \frac{8}{23} = 0.348$
- 전체 취업자 수 : $\hat{t} = \sum \frac{W_a z_a}{f_1} = 8 \cdot 126.3 = 1010.4$

2. 가중치 의미와 계산

- 표준화 가중값 (Standardized or Normalized weights)
- 표본크기와 일치 ($\sum W_a = n$) 하도록 가중값을 표준화시킨 가중값
- 표준화 가중값의 계산

$$W_a' = W_a \cdot \left(\frac{n}{\sum W_a} \right) \text{ so that } \sum W_a' = n$$

예 : 비례가중값의 자료에서 가중합이 230이고, 표본크기가 60이므로 표준화 가중값을 이용하기 위해 6/23을 곱하여 계산

- 모집단 총합계 추정에 부적절하나 모평균과 모비율 추정에서는 동일한 결과 제공

2. 가중치 의미와 계산

가상자료 : 표준화 가중

층	표본가구	성인수(B_a)	W_a	W'_a	Z_a	$W'_a Z_a$
도시	1	2	2	0.522	1	0.522
	2	2	2	0.522	1	0.522
	3	1	1	0.261	0	0
농촌	1	4	8	2.087	0	0
	2	2	4	1.043	1	1.043
	3	3	6	1.565	0	0
합계			23	6.00	3	2.087

$$\bar{Z}_w = \frac{\sum W'_a Z_a}{\sum W'_a} = \frac{2.087}{6} = 0.348$$

2. 가중치 의미와 계산

- 무응답 보정을 위한 가중값
- 표본의 부차그룹마다 가중응답률 계산
 - 도시지역 추출률 : (적합한 표본이 완성된 가중응답수) / (표본으로 적합한 가중표본 수)
- 무응답 보정을 위해 부차그룹의 응답률 역수를 가중값(W_{2a})으로 사용
 - 분석 가중값 : $W_a = W_{1a} \cdot W_{2a}$

여기서 W_{1a} 는 불균등 선택 확률에 대한 가중치

2. 가중치 의미와 계산

- Example :무응답 보정을 위한 가중값
- 어느 지역을 도시와 농촌 및 북부와 남부로 구분하여 층화 다단계 표본으로 설계하여 950가구를 선택
 - 북부(N)의 추출률 1/100, 남부(S)의 추출률 1/200으로 북부지역을 2배 많이 추출
 - 도시(U)의 응답률이 농촌(R)의 응답률보다 낮음
 - 목표 표본크기 n_h , 조사 표본크기 n'_h , 가스난방에 대한 응답가구 수를 γ'_h 이라고 할 때, 무응답 보정을 고려한 가스 난방 가구비율을 추정 문제를 가정
- 무응답 보정을 위한 가중값

$$W_h = W_{1h} \cdot W_{2h} = W_{1h} \cdot \left(\frac{n_h}{n'_h}\right)$$

2. 가중치 의미와 계산

가상자료 : 무응답 보정 가중값

층	n_h	n'_h	γ'_h	W_{1h}	W_{2h}	W_h	$W_h n'_h$	$W_h \gamma'_h$
NU	100	70	60	100	1.43	143	10,010	8,580
NR	300	270	140	100	1.11	111	29,970	15,540
SU	150	120	110	200	1.25	250	30,000	27,500
SR	400	380	120	200	1.05	210	79,800	44,100
전체	950	840					149,780	95,720

➤ 가스 난방 가구 추정 : $\hat{t} = \sum W_h \gamma_h^1 = 95,720$

➤ 가스 난방 가구 비율 추정 : $\gamma = \frac{\sum w_h \gamma'_h}{\sum w_h n'_h} = \frac{95,720}{149,780} = 0.639$

2. 가중치 의미와 계산

- 사후층화 보정가중값 (모집단 가중값)

- 모집단 분포 변화를 반영한 가중값 : W_{3a}

- 목표 모집단 크기와 조사 모집단 크기의 차이를 반영한 가중값

- 표본 가중 합과 모집단 크기와의 차이를 반영한 가중값

(모집단 크기) / (표본크기 합(=모집단 크기의 추정값))

- 중요한 인구사회학적 변수의 모집단 분포와 가중 표본 분포와의 차이를 조정하는 가중 값

- 분석 가중값: $W_a = W_{1a} \cdot W_{2a} \cdot W_{3a} = W_h \cdot W_{3a}$

2. 가중치 의미와 계산

- 가상자료 : 무응답 보정 가중값 ($W_a'' = W_h \cdot W_{3a}$)

층	n'_h	γ'_h	W_h	$W_h n'_h$	N_h	W''_h	$W''_h n'_h$	$W''_h \gamma'_h$
NU	70	60	143	10,010	44,080	157.7	11,039	9,462
NR	270	140	111	29,970		122.4	33,048	17,136
SU	120	110	250	30,000	107,930	245.7	29,484	27,027
SR	380	120	210	79,800		206.4	78,432	43,344
전체	840			149,780			152,003	96,969

➤ 모집단 가중 값

$$W_{1N} = \frac{N_N}{W_N n'_N} = \frac{44080}{39980} = 1.103$$

$$W_{1S} = \frac{107930}{109800} = 0.983$$

- ### ➤ 가스 난방 가구 비율 추정 : $\gamma = \frac{\sum W_h'' \gamma'_h}{\sum W_h'' n'_h} = \frac{96,969}{152,003} = 0.638$

2. 가중치 의미와 계산

- 가중값의 영향 : 추정 정도의 손실
- 표본크기 n 으로 모평균을 추정할 때 가중 변수로 인해 야기되는 영향 (by kish, 1965)

$$\text{Deff}_w = 1 + \left(\frac{n-1}{n} \right) \frac{s_w^2}{w^2} \cong 1 + cv_w^2$$

2. 가중치 의미와 계산

- 총 분산이 거의 비슷한 총화추출에서 비비례배분 (disproportionate allocation)에 의한 손실

$$L = \frac{(\sum_h w_h k_h)}{\sum_h \frac{w_h}{k_h}}$$

여기서 w_h 는 부모집단 비율, k_h 는 h 층의 가중값

- 계산방법 2 $L = n' \left(\frac{\sum w_i^2}{(\sum w_i)^2} \right) = n' \left(\frac{\sum n'_h w_h^2}{(\sum n'_h w_h)^2} \right)$
- 유효 표본크기 $n_{eff} = \frac{(\sum w_i)^2}{\sum w_i^2} = n' / deff(L)$

2. 가중치 의미와 계산

가상자료 : 가중값으로 인한 정도의 손실

층	n'_h	W'_h	$W'_h n'_h$	$W'^2_h n'_h$
NU	70	157.7	11,039	1,740,850
NR	270	122.4	33,048	4,045,075
SU	120	245.7	29,484	7,244,219
SR	380	206.4	79,432	16,188,365
전체	840		153,003	29,218,509

➤ 정도의 손실

$$L = n' \left(\frac{\sum W_i^2}{(\sum W_i)^2} \right) = 840 \times \left(\frac{29,218,509}{(153,003)^2} \right) = 1.0623$$

➤ 효율적인 표본크기

$$n_{eff} = \left(\frac{(153,003)^2}{29,218,509} \right) = 791 = 840 / 1.06226$$

2. 가중치 의미와 계산

극단 가중값의 절사 효과 및 영향

- 지나치게 크거나 작은 극단 가중값이 존재하는 경우 극단 가중값은 추정에 많은 영향을 미치게 되며, 그로 인해 정도의 손실이 크게 발생

➤ 가중값의 변동계수가 증가

$$Def_{fw} = 1 + \left(\frac{n-1}{n}\right) \frac{S_w^2}{W^2} \cong 1 + cv_w^2$$

- 극단 가중값은 절사하는 것이 바람직

2. 가중치 의미와 계산

- Example : 극단 가중값의 영향(1)
- 가상자료

W_h	f_h	$f_h W_h$	$f_h W_h^2$
1.00	500	500	500
1.25	240	300	375
1.50	160	240	360
2.00	130	260	520
3.00	100	300	900
5.00	25	125	625
10.00	45	450	4500
합계	1,200	2,175	7,780

➤ 정도의 손실 크기

$$L = n'(\sum_i w_i^2 / (\sum_i w_i)^2) = 1200 \cdot (7780 / 2175^2) = 1.974$$

2. 가중치 의미와 계산

● 변동계수를 이용한 손실 크기의 계산

$$\text{Def}f_w = 1 + \left(\frac{n-1}{n} \right) \frac{s_w^2}{\bar{w}^2} \cong 1 + cv_w^2$$

- 가중값의 평균 $\bar{w} = \Sigma f_k w_k / \Sigma f_k = 2175 / 1200 = 1.8125$
- 가중값의 분산 $V(w) = \frac{1}{n} \Sigma f_k (w_k - \bar{w})^2 = 3.198$
- 가중값의 변동계수 $CV_w = \sqrt{V(w)} / \bar{w} = 0.974$
- 손실크기 $\text{Def}f_w \cong 1 + cv_w^2 = 1 + 0.974^2 = 1.974$

2. 가중치 의미와 계산

- Example : 극단 가중값의 영향(2)

- 가상자료

W_h	f_h	$f_h W_h$	$f_h W_h^2$
1.00	500	500	500
1.25	240	300	375
1.50	160	240	360
2.00	130	260	520
3.00	100	300	900
5.00	70(25)	350(125)	1,750(625)
10.00	45	450	4,500
합계	1,200	1,950(2,175)	4,405(7,780)

$$L = n' \left(\sum_i w_i^2 / \left(\sum_i w_i \right)^2 \right) = 1200 \cdot \left(\frac{4405}{1950^2} \right) = 1.39$$

$$L^* = 1200(7780/2175^2) = 1.974$$

3. 층화집락추출에서 가중치 분석 사례

- **Example : Stratified Cluster Sample Design**

- **Goal : Researcher want to know how much students spend weekly for ice cream, on the average, and what percentage of students spend at least \$10 weekly for ice cream**
- **Population a junior high school with a total of 4000 students in grades 7,8 and 9**

Grade	Study Group	Students	Sample
7	608	1,824	8
8	252	1,025	3
9	403	1,151	5
total	617	4,000	16

- **Sampling Design : Stratified Cluster Sample Design**
 - **Stratification : Grade**
 - **Cluster : Study Group (psu)**
 - **Sampling Unit : Study Group**
 - **All students in each selected study group are interviewed**

3. 층화집락추출에서 가중치 분석 사례

• Data : SAS file

```
data IceCreamStudy;
    input Grade StudyGroup Spending @@;
    if (Spending<10) then Group='less';
    else Group='more';
datalines;
7 34 7 7 34 7 7 412 4 9 27 14 7 34 2 9 230 15 9 27 15 7 501 2 9 230 8
9 230 7 7 501 3 8 59 20 7 403 4 7 403 11 8 59 13 8 59 17 8 143 12 8 143 16
8 59 18 9 235 9 8 143 10 9 312 8 9 235 6 9 235 11 9 312 10 7 321 6 8 156 19
8 156 14 7 321 3 7 321 12 7 489 2 7 489 9 7 78 1 7 78 10 7 489 2 7 156 1
7 78 6 7 412 6 7 156 2 9 301 8
;
data studygroups;
    input grade _total_ @@ ;
    datalines;
    7 608 8 252 9 403
    ;
data iceCreamStudy;
    set IceCreamStudy;
    if Grade=7 then prob=8/608; if Grade=7 then prob1=20/1824;
    if Grade=8 then prob=3/252; if Grade=8 then prob1=9/1025;
    if Grade=9 then prob=5/403; if Grade=9 then prob1=11/1151;
    weight=1/prob; weight1=1/prob1;
run;
```

3. 층화집락추출에서 가중치 분석 사례

Output : proc surveymeans

```
title 'Stratified Clustered Sample Design';  
proc surveymeans data=IceCreamStudy total=StudyGroups;  
    strata Grade / list;  
    cluster studygroup;  
    var Spending Group;  
    Weight weight;  
run;
```

The SURVEYMEANS Procedure

Data Summary

Number of Strata	3
Number of Clusters	16
Number of Observations	40
Sum of Weights	3162.6

Class Level Information

Class Variable	Levels	Values
Group	2	less more

3. 층화집락추출에서 가중치 분석 사례

- Output 계속: proc surveymeans

Stratum Information								
Stratum Index	Grade	Population Total	Sampling Rate	N Obs	Variable	Level	N	Clusters
1	7	608	1.32%	20	Spending Group	less	20	8
						more	17	8
							3	3
2	8	252	1.19%	9	Spending Group	less	9	3
						more	0	0
							9	3
3	9	403	1.24%	11	Spending Group	less	11	5
						more	6	4
							5	4

=n/N

Statistics						
Variable	Level	N	Mean	Std Error of Mean	95% CL for Mean	
Spending Group		40	8.923860	0.650859	7.51776370	10.3299565
	less	23	0.561437	0.056368	0.43966057	0.6832130
	more	17	0.438563	0.056368	0.31678698	0.5603394

3. 층화집락추출에서 가중치 분석 사례

Output : proc surveymeans , Without psu

```

title 'Stratified Clustered Sample Design';
proc surveymeans data=IceCreamStudy total=StudyGroups;
  strata Grade / list;
  /*cluster studygroup;*/
  var Spending Group;
  Weight weight;
run;

```

Stratum Information							
Stratum Index	Grade	Population Total	Sampling Rate	N Obs	Variable	Level	N
1	7	608	3.29%	20	Spending Group	less	20
						more	17
2	8	252	3.57%	9	Spending Group	less	9
						more	0
3	9	403	2.73%	11	Spending Group	less	11
						more	6
=학생수/집락수							

Statistics						
Variable	Level	N	Mean	Std Error of Mean	95% CL for Mean	
Spending Group		40	8.923860	0.524566	7.86098915	9.98673108
	less	23	0.561437	0.058262	0.44338747	0.67948611
	more	17	0.438563	0.058262	0.32051389	0.55661253

4. 노동패널조사에서 가중치 계산

4.1 제1차 년도 표본추출방법

- 제1차 년도 노동패널조사의 추출틀은 1997년 고용구조특별조사(이하 고특조사)에 사용되었던 조사구
- 1997년 고특조사의 추출틀은 1995년 인구주택총조사의 표본조사에 사용했던 전체 10% 조사구를 사용
- 조사대상 지역 : 7대 특·광역시와 8개 도의 시지역으로 제한
- 표본크기 : 1,000개
- 15개 시도별 배분은 표본조사의 조사구수를 기준으로 비례배분
- 최종적으로 시도별 할당된 결과와 1997년 고특조사의 도시지역 조사구 규모를 비교·검토하여 15개 시도별 동부/읍면부 별로 표본할당을 결정
 - 예를 들어 서울특별시는 1995년 10%표본조사에 사용한 조사구수는 5,186개로 19,025개의 27.3%에 해당되므로 273개가 할당되었으나, 1997년 고특조사에서 서울의 조사된 조사구는 227개이므로 227개를 서울의 표본규모 결정

4. 노동패널조사에서 가중치 계산

4.1 제1차 년도 표본추출방법

- 1998년 노동패널조사의 시도별 표본할당 결과

시도	서울	부산	대구	인천	광주	대전	울산	경기
고득조사구	227	176	155	150	106	101	70	280
할당조사구	227	97	64	59	32	31	24	167
조사구추출률	4.3772	5.2689	5.2805	5.2867	5.2718	5.2365	5.2362	5.1511
시도	강원	충북	충남	전북	전남	경북	경남	합계
고득조사구	165	159	122	170	152	129	220	2,382
할당조사구	26	22	25	40	25	54	58	951
조사구추출률	5.1587	5.2133	5.1230	5.2840	5.1335	5.2889	5.3016	4.9987

- 추출된 표본조사구내 조사대상가구는 서울 및 경기 읍면부에서는 조사구당 6가구, 기타지역에서는 조사구당 5가구를 랜덤하게 선정

구분	전체모집단		10%표본조사구		표본		추출률(%)	
	조사구	가구수	조사구	가구수	조사구	가구수	조사구	가구수
전국 시부	190,250	11,100,320	19,025	1,110,032	951	5,000	0.4999	0.0450
시의 동부	173,530	10,098,910	17,353	1,009,891	871	4,582	0.5039	0.0454
시의 읍부	16,720	1,001,410	1,672	100,141	80	418	0.04785	0.0417

4. 노동패널조사에서 가중치 계산

4.2 제1차 년도 가중치 계산

- 표본설계 가중치

실제 1차년도의 경우 1995년 인구주택총조사의 10% 표본조사구에서 1997년 고특조사의 조사구를 추출. 하지만 세부적인 내용을 알 수 없으므로 다음과 같은 방법의 추출확률을 계산

➤ 시도별 도시조사구 추출확률 : 시도별 전체 조사구 중 도시조사구 구성비율로 계산

➤ 도시조사구 중 표본조사구 추출확률 : 도시조사구 중 표본조사구의 비로 계산

➤ 시도별 표본조사구의 추출확률은 아래와 같이 계산

- 7대 특광역시 : $0.1 * (\text{해당 시의 표본 조사구수} / \text{해당 시의 전체 조사구수})$

- 도의 동부 : $0.1 * (\text{해당 도 동부 표본 조사구수} / \text{해당 도 동부 전체 조사구수})$

- 도의 시에 속한 읍면부 : $0.1 * (\text{해당 도의 시에 속한 읍면부 표본 조사구수} / \text{해당 도의 시에 속한 읍면부 전체 조사구수})$

- 최종표본가구 추출확률 : $(\text{총 접촉 가구수} / 97 \text{ 고특 응답가구수}) * (97 \text{ 고특응답가구수} / \text{조사구내 전체가구수})$

4. 노동패널조사에서 가중치 계산

4.2 제1차년도 가중치 계산

- 무응답 보정 가중치

KILPS에서 시도별로 무응답자와 응답자의 특성이 같다는 가정을 하지 않고, 조사구내 무응답가구와 응답가구간 차이가 없다는 가정하에서 응답률을 계산

➤ 조사구내 무응답보정 가중치 : 총 접촉 가구수/최종 조사완료된 가구수

- 7대 특광역시 : $1/[0.1 * (\text{해당 시의 표본 조사구수} / \text{해당 시의 전체 조사구수}) * (97 \text{ 고특조사가구수} / \text{조사구내 전체 가구수}) * (\text{총 접촉가구수} / 97 \text{ 고특조사가구수}) * (\text{최종 조사완료가구수} / \text{총 접촉가구수})]$
- 도의 동부 : $1/[0.1 * (\text{해당 도 동부 표본 조사구수} / \text{해당 도 동부 전체 조사구수}) * (97 \text{ 고특조사가구수} / \text{조사구내 전체 가구수}) * (\text{총 접촉가구수} / 97 \text{ 고특조사가구수}) * (\text{최종 조사완료가구수} / \text{총 접촉가구수})]$
- 도의 시에 속한 읍면부 : $1/[0.1 * (\text{해당 도의 시에 속한 읍면부 표본 조사구수} / \text{해당 도의 시에 속한 읍면부 전체 조사구수}) * (97 \text{ 고특조사가구수} / \text{조사구내 전체 가구수}) * (\text{총 접촉가구수} / 97 \text{ 고특조사가구수}) * (\text{최종 조사완료가구수} / \text{총 접촉가구수})]$

5. 12차년도 노동패널의 가중치 산출

5.1 연구배경

- 1998년에 조사를 시작 한국노동패널조사는 2008년 11차까지 조사해오면서 계속 응답한 가구수는 2,607가구로 원표본 5,000가구의 유지율은 52.1%임
- 1998년 1차년도 패널조사대상을 시지역으로 제한하여 패널을 구축. 하지만 패널가구의 시지역에서 군지역으로 이동과 새로 시로 승격 등 모집단 변화를 반영하지 못해 전국을 대표하는데 한계가 존재
- 최근 지역통계 생산의 요구와 세부영역에 대한 분석을 위해 군지역으로 조사지역 확장 및 표본규모 확대가 필요
- 이에 2008년 11차년도에 응답한 6,027가구의 특성을 동부/읍면부별, 16개 시도별 및 가구원수별로 분석하여 현재 시점에서 조사모집단을 대표할 수 있도록 미흡한 부분에 대한 보정방안 연구가 필요

5. 12차년도 노동패널의 가중치 산출

5.2 추가패널 구축 설계

- 추가패널 구축의 규모는 조사비용의 수용성을 감안하여 5,000가구로 함(98년 패널조사와 같은 규모)
- 11차년도의 응답가구와 2005년 인구주택총조사 가구의 16개 시도별 동부/읍면부별 분포를 비교하여 추가 가구수를 계산. 추가표본은 조사구당 5가구 추출 가정에서 추출해야 할 조사구수를 산출

시도	동부			읍면부		
	현행패널	추가패널	조사구	현행패널	추가패널	조사구
서울	1,303	185	37	0	0	0
부산	461	65	13	9	0	0
대구	330	20	4	18	5	1
인천	319	40	8	7	0	0
광주	180	25	5	0	0	0
대전	168	30	6	0	0	0
울산	107	25	5	20	5	1
경기	1,004	240	48	194	50	10
강원	120	25	5	29	60	12
충북	108	30	6	17	70	14
충남	80	30	6	95	85	17
전북	161	25	5	60	30	6
전남	102	15	3	51	125	25
경북	158	50	10	139	65	13
경남	234	40	8	108	85	17
제주	2	55	11	1	20	4
합계	4,837	900	180	748	600	120

● 표본조사구의 추출방법

- 동 부 : 가구원 1~2인 가구수와 단독주택수를 기준으로 확률비례추출법을 적용하여 계통추출법으로 선정
- 읍면부 : 조사구내 가구수를 기준으로 확률비례추출법으로 선정
- 읍면부가 없는 서울, 광주, 대전은 읍면부 추가 표본이 없음. 부산과 인천은 1998년 구축된 원패널 가구들이 잘 유지되어 추가패널이 필요 없음

5. 12차년도 노동패널의 가중치 산출

5.3 12차년도 노동패널의 가중치 산출

- 추가패널의 가중치 산출

- 12년차 패널 = 원패널 + 추가패널
- 12년차 응답가구(6,712)=원패널(3,657)+분가패널(1,649)+추가패널(1,415)
- 추가패널 1,500가구를 조사하였으나 1,415가구만 응답
- 추가패널 16개시도별, 동/읍면부별 확률비례추출법으로 조사구를 선정하고 표본조사구내에서는 조사대상가구는 확률계통추출법으로 추출.
- 추가패널의 가중치는 다음과 같음

$$w_{1hij} = \frac{\sum_{i=1} X_{hi}}{n_h X_{hi}} \cdot \frac{M_{hi}}{m_{hi}} \quad (1)$$

✓ h : 층(시도, 동/읍면부)

✓ i : 조사구

✓ j : 가구

✓ n_h : h 층의 표본크기

✓ X_{hi} : h 층 i 조사구 층도의 크기

✓ M_{hi} : h 층 i 조사구 모집단 가구수

✓ n_{hi} : h 층 i 조사구 표본 가구수

✓ 여기서 사용된 층도의 크기는 동부는 1~2인

가구+단독주택수, 읍면부는 조사구내 가구수

5. 12차년도 노동패널의 가중치 산출

5.3 12차년도 노동패널의 가중치 산출

- 추가패널의 가중치 산출
 - 앞의 식으로 가중치 계산을 위해서는 각 층에서 속도의 크기의 합계를 계산하여 층별로 가중치 합계가 조사모집단의 가구수와 같도록 사후층화보정이 필요
 - 조사모집단의 가구수를 2005년 센서스(1차년도)와 2009년 추계가구수(12차년도)에 맞출 것인지 고려가 필요
 - 기존 원패널에 적용된 가중치는 2009년 추계가구수와 연계되어 있으므로 2009년 추계가구수로 16개시도별 동/읍면부로 산출하여 아래와 같은 식으로 스케일 조정인자 c_h 를 산출하고 식(1)의 w_{1hij} 와 곱해 가중치 산출

$$\sum_{(ij) \in s_1}^{n_{ij(s)}} c_h w_{1hij} = \sum_{(ij) \in s_1}^{n_{ij(s)}} c_h \cdot \frac{\sum_{i=1}^{n_h} X_{hi}}{n_h X_{hi}} \cdot \frac{M_{hi}}{m_{hi}} = M_h, \quad h = 1, \dots, H \quad (2)$$

- 최종가중치는 $c_h * w_{1hij}$ 로 산출하고 w_{1hij}^f 로 표기

5. 12차년도 노동패널의 가중치 산출

5.3 12차년도 노동패널의 가중치 산출

- 통합패널의 가중치 산출

- 12차년도 조사에서 응답한 패널가구는 추가패널에서 응답한 1,415가구와 원패널에서 응답한 5,306가구로 구성. 추가패널의 경우 동부 848가구, 읍면부 567가구 구성되었으며 추가패널의 가중치는 w_{1hij}^f 로 계산
- 원패널 가구의 가중치는 11차년도 조사와 12차년도 조사에 모두 응답한 사람의 개인가중치를 로지스틱 회귀모형으로 응답률을 산출 후 이의 역수를 11차년도 개인가중치에 곱해서 계산한 후 가구에 속한 가구원의 평균으로 계산
- 이렇게 산출한 가중치를 아래 식을 만족하도록 스케일 조정한 가구가중치를 w_{0hij} 라 함

$$\sum_{(ij) \in s_0}^{n_{\bar{j}}(0)} w_{0hij} = M_h, \quad h=1, \dots, H \quad (3)$$

5. 12차년도 노동패널의 가중치 산출

5.3 12차년도 노동패널의 가중치 산출

- 통합패널의 가중치 산출

- 16개시도별 동/읍면부별로 식(2), 식(3)에 주어진 방정식을 이용하여 2009년 추계가구수를 기준으로 산출한 가중치조정계수는 다음과 같다.

동읍면부	표본크기(12차년도)		가중치합계(12차년도)		2009년 가구수	가중치조정계수	
	원표본	추가표본	원표본	추가표본		원표본	추가표본
서울동부	1,155	161	3345737	3044892	3,465,416	1.035771	1.138108
부산동부	416	61	844881.5	942431.7	1,200,492	1.4209	1.273824
부산읍면	8	0	15093.15	0	31,887	2.11268	0
대구동부	272	20	895363.1	910203.6	803,382	0.89727	0.88264
대구읍면	22	5	61950.05	42190.35	54,197	0.87485	1.284583
인천동부	332	40	792307.9	683924.6	872,355	1.10103	1.275513
인천읍면	8	0	16489.93	0	27,330	1.657375	0
광주동부	152	23	441358	411199.8	504,702	1.143521	1.227389
대전동부	155	29	392037.5	378338	521,887	1.331217	1.37942
울산동부	113	24	204220.4	187301.6	304,625	1.491648	1.626388
울산읍면	20	5	33862.87	37459.79	62,101	1.833897	1.657804
경기동부	1,028	229	2641082	2248869	3,094,728	1.171765	1.376126
경기읍면	182	52	457100.5	511356.1	636,015	1.391412	1.243781

5. 12차년도 노동패널의 가중치 산출

5.3 12차년도 노동패널의 가중치 산출

동읍면부	표본크기(12차년도)		가중치합계(12차년도)		2009년	가중치조정계수	
	원표본	추가표본	원표본	추가표본	가구수	원표본	추가표본
강원동부	109	23	364147.9	346813.2	340,132	0.934049	0.980736
강원읍면	20	53	55911.29	71320.07	210,194	3.75942	2.947193
충북동부	105	28	237244.1	232583.7	332,053	1.399626	1.427671
충북읍면	29	63	56474.95	34748.57	216,025	3.825147	6.216803
충남동부	70	26	154274.3	139447.8	277,328	1.797629	1.988759
충남읍면	107	82	232857.4	202025.2	453,874	1.94915	2.246621
전북동부	143	30	349917.6	378330.5	441,656	1.262171	1.167382
전북읍면	60	25	117298.2	89802.96	210,292	1.792799	2.341705
전남동부	100	15	205593.7	206450.2	281,081	1.367167	1.361495
전남읍면	52	122	100840.9	85678.36	397,328	3.940148	4.637437
경북동부	143	47	371068.9	300774.9	504,777	1.360332	1.678255
경북읍면	146	57	326504.7	284650.7	487,270	1.492383	1.711817
경남동부	235	39	461826.2	469090.5	693,574	1.501808	1.478551
경남읍면	104	83	195040.9	208611	438,565	2.24858	2.10231
제주동부	1	53	1748.74	8621.9	135,657	77.57414	15.734
제주읍면	1	20	6991.4	2315.25	50,042	7.157651	21.61408

5. 12차년도 노동패널의 가중치 산출

5.3 12차년도 노동패널의 가중치 산출

- 통합패널의 가중치 산출

- 12차년도 조사에서 응답한 6,721가구는 원패널가구와 추가패널가구로 구분되지만 이를 통합한 통합패널의 가중치는 표본크기를 가중값으로 원패널과 추가패널의 가중평균을 계산할 수 있으며 다음과 같은 식으로 나타낼 수 있다.

$$\frac{n_{h0}}{n_h} \sum_{(ij) \in S_0} w_{0hij} + \frac{n_{h1}}{n_h} \sum_{(ij) \in S_1} w_{1hij} = M_h$$

- 여기서 n_{h0} 와 n_{h1} 은 각각 h 층에서 원표본크기와 추가표본크기를 나타내며 M_h 는 h 층의 2009년 추계가구수를 나타냄
- 통합표본의 가중치는 추가표본 가중치와 원표본가중치에 위의 표에 주어진 가중치조정인자와 추가표본크기와 원표본크기의 구성비를 곱한 값을 곱해 산출한다.

5. 12차년도 노동패널의 가중치 산출

5.3 12차년도 노동패널의 가중치 산출

- 통합패널의 가중치 산출

- 예) 서울시 동부의 층에서 원표본 가중치조정계수와 원표본의 구성비는 각각 1.035772과 0.87766(=1155/1316). 이들의 곱은 0.909053이므로 이 값을 서울시 동부층의 원표본 가구가중치에 곱해서 통합표본의 가중치를 산출
- 통합표본에서 개인종단가중치, 개인횡단가중치도 횡단가구가중치와 동일한 방법으로 산출
- 한국노동패널의 가중치는 원표본과 통합표본에 대한 가중치를 별도로 관리·산출 필요.
- 원패널 종단가중치는 1998년을 원년으로 해서 산출. 1998년을 시점으로 한 종단적 분석을 위한 개인종단가중치만 제공
- 통합패널은 2009년(12차년도)을 원년으로 종단개인가중치와 횡단가구가중치 및 횡단개인가중치를 산출. 전국단위로 횡단적인 분석을 할 경우에는 통합패널의 횡단가구가중치와 횡단개인가중치를 사용.
- 통합패널을 통해 그 동안 문제가 되었던 대표성의 문제는 해결되었다고 볼 수 있음. 그러나 전국단위의 종단적 분석은 2009년 이후인 경우만 유효하다는 아쉬움이 존재

5. 12차년도 노동패널의 가중치 산출

5.3 12차년도 노동패널의 가중치 산출

- 12년차 노동패널 가중치 과제
 - 패널조사 중 대표적인 노동패널은 14차년도(2011년)까지 조사 완료됨. 심층적인 분석을 통해 개선해야 할 사항이 있는지 살펴보는 것이 바람직함
 - 1차년도(1998년)의 가중치 산출과정을 분석하면서 정상적인 표본설계를 통해 패널을 구축했다라면 활용성이 높았을 것으로 예상
 - 가중치 산출과정에서 각 웨이브마다 사후층화보정에 대한 설명이 없고 스케일조정이라는 언급만 존재하는데 무엇을 기준으로 하였는지 알 수 없음. 12차년도(2009년)부터는 종단개인가중치의 합계는 2009년 만15세 이상 인구수가 되어야 하므로 이를 기준으로 종단가중치를 사후층화보정해야 함.
 - 가중치 산출 후 가중치의 분포적 특성을 분석하여 이상치 유무를 확인해야 함.

5. 12차년도 노동패널의 가중치 산출

5.3 12차년도 노동패널의 가중치 산출

- 12년차 노동패널 가중치 과제
 - 가중치를 표준화 변환하여 Z값이 50이상인 경우를 정리하여 웨이브별로 상위 5개 이상 이상치를 정리한 결과는 다음과 같음.

1998년1차			2005년8차			2008년11차		
가구ID	가중값	표준화	가구ID	가중값	표준화	가구ID	가중값	표준화
2548	4514.563	5.443139	4543	13765.08	9.975057	1857	20956.64	14.04812
2549	4514.563	5.443139	2551	12060.72	8.452671	4543	18547.54	12.20635
2550	4514.563	5.443139	1857	10290.98	6.871886	667	13429.08	8.293277
2551	4514.563	5.443139	3842	9847.975	6.476177	2544	12372.47	7.4855
2552	4514.563	5.443139	2084	9690.041	6.335106	2084	11587.9	6.885691

- 가중치 분석에서 이상치인 경우에는 표준화된 값이 30이상이면 경계점의 가중치로 대체한 후 전체적으로 사후총화보정하여 안정적인 가중치로 대체할 것을 제안
- 좀 더 나아가 횡단개인가중치에서는 총별로나 16개시도별, 성별, 연령대별로 모집단구성과 패널의 표본집단의 구성비를 맞추는 사후총화보정을 통해 모수 추정의 편향을 축소하고 표본 대표성을 제고할 수 있음

6. 14차년도 노동패널의 가중치 산출

6.1 원패널(14차년도) 가중치 산출

- 가중치 산출 단계 : 종단개인가중치 산출 $\Rightarrow$ 가구가중치 산출 $\Rightarrow$ 횡단개인가중치 산출
- 종단개인가중치 산출 절차
 - ① 개인별 응답률 계산 : 로짓 회귀모형(종속변수 : 14차년 응답여부, 설명변수 : 13차년도 조사항목(성별, 지역, 학력, 주된 활동, 만 나이, 가구소득)을 적합한 후에 14차년도 응답자를 대상으로 응답확률 계산
 - ② 14차년도 개인종단가중치=13차년도 종단가중치 * 개인별 응답률의 역수
 - ③ 검토사항 : 설명변수 중에서 항목무응답(대체한 후에 로짓 회귀모형을 적합, 항목무응답의 대체방법), 원패널에 속했지만 13차년도 무응답자인데 14차년도 응답자의 종단개인가중치 (0부여, 최근 응답시점의 종단가중치 부여), 원패널에 소속되지 않았으나 결혼/입양/동거 등의 신규 진입자는 0을 부여, 원패널 가구원으로 만15세로 신규 진입자는 속한 가구의 평균가중치

6. 14차년도 노동패널의 가중치 산출

- 가구 가중치 산출 절차

- ① 가구 가중치 계산 : 가구별로 속한 모든 가구원들의 종단개인가중치들의 평균
- ② 14차년도 조사된 모든 가구원이 원패널이 아닌 경우에는 가구가중치는 0부여
- ③ 신규 진입한 가구원(결혼,동거 등)은 평균계산시 분모에는 반영(종단가중치는 0임), 만15세로 신규 진입자는 평균계산에서 제외되지만 종단개인가중치는 존재함

- 횡단개인가중치는 원패널가구원은 개인종단가중치를 부여하고, 신규 진입자는 가구가중치를 부여한 후에 스케일보정(성별*연령대별 경제활동인구조사 통계기준)

- ① 스케일보정계수의 산출근거 : 조사대상지역(95년 센서스의 도시지역의 제한됨), 상주추계인구(성별*연령대별 산출해야 함)

6. 14차년도 노동패널의 가중치 산출

6.2 통합패널(3차년도) 가중치 산출

- 가중치 산출 단계 : 종단개인가중치 산출 $\Rightarrow$ 가구가중치 산출 $\Rightarrow$ 횡단개인가중치 산출
- 종단개인가중치 산출 절차
 - 개인별 응답률 계산 : 로짓 회귀모형(종속변수 : 3차년 응답여부, 설명변수 : 2차년도 조사항목(성별, 지역, 학력, 주된 활동, 만 나이, 가구소득, 추가표본여부)을 적합한 후에 3차년도 응답자를 대상으로 응답확률 계산
 - 3차년도 개인종단가중치 = 2차년도 종단가중치 * 개인별 응답률의 역수
 - 검토사항 : 설명변수 중에서 항목무응답(대체한 후에 로짓 회귀모형을 적합 , 항목무응답의 대체방법), 통합패널에 속했지만 2차년도 무응답자인데 3차년도 응답자의 종단개인가중치(0부여, 최근 응답시점의 종단가중치 부여), 통합패널에 소속되지 않았으나 결혼/입양/동거 등의 신규 진입자는 0을 부여, 통합패널 가구원으로 만15세로 신규 진입자 가구의원의 평균가중치 가구가중치를 개인종단가중치

6. 14차년도 노동패널의 가중치 산출

- 가구 가중치 산출 절차

- ① 가구 가중치 계산 : 가구별로 속한 모든 가구원들의 종단개인가중치들의 평균
- ② 3차년도 조사된 모든 가구원이 통합패널이 아닌 경우에는 가구가중치는 0부여
- ③ 신규 진입한 가구원(결혼,동거 등)은 평균계산시 분모에는 반영(종단가중치는 0임), 만15세로 신규 진입자는 평균계산에서 제외되지만 종단개인가중치는 존재함

- 횡단개인가중치는 통합패널가구원은 개인종단가중치를 부여하고, 신규 진입자는 가구가중치를 부여한 후에 스케일보정(성별*연령대별 2011년 상주추계인구 기준)

6. 14차년도 노동패널의 가중치 산출

6.3 14차년도 노동패널의 가중치 개선사항

- 원패널의 스케일 보정계수의 산출 근거 필요
 - ① 조사대상 지역(67개 시지역)의 가구증가율과 인구증가율의 산출방법(5년간의 평균 증가율의 계산은 산술평균이 아닌 기하평균을 사용 검토)
- 종단개인가중치와 횡단개인가중치의 극단값에 대한 절사가중치 산출방법의 적용을 적극 검토할 것
- 가중치 산출과정에 조사과정에서 수집되는 para data의 정보를 활용하는 방안의 연구검토 (특정지역, 특정연령 또는 별도의 특성을 갖는지 분석한 후에 응답률 산출에 반영 방안)
- 가중치 산출 후 가중치의 분포적 특성을 분석하여 이상치 유무를 확인해야 함.

7. 결 언

- 대규모 통계조사는 대부분 복잡표본조사 설계임
(지역사회 건강조사, 국민건강영양조사, 청소년건강행태조사, 노동패널조사, 특히
패널조사는 대부분 복잡표본조사로 생각하면 됨)
- 많은 표본조사통계학자들은 복잡표본조사설계를 자체 가중설계와 유사하도록 설계하는
방안을 연구하고 있으나 조사과정이나 조사 후에 사후층화보정을 해야 할 경우에는
전체적인 특성은 복잡표본조사가 될 것임
- 복잡표본조사의 자료분석에 앞서 반드시 필요한 작업이 가중치 산출이며 가중치 계산은
이론은 간단하나 실제 계산절차는 탐탁하지 않음
- 가중치가 적용된 표본조사분석은 통상적인 통계패키지로 분석할 수 없으므로 복잡표본조사
데이터를 분석할 수 있는 통계패키지를 사용해야 함
- 복잡표본조사 분석에 많이 사용되는 통계패키지는 SAS surveymeans surveyreg,
STRATA, SPSS complex survey 분석용, SUDAAN, R의 프로그램 등이 있으나
분석목적에 적합한 패키지를 사용해야 함

7. 결 언

- 표본조사 가중치 의미와 산출과정을 설명하였고, 복잡표본조사의 일종인 총화집락추출에
에서 조사데이터 분석 사례를 예시하여 가중치를 사용한 분석방법 이해 증진
(지역사회 건강조사, 국민건강영양조사, 청소년건강행태조사, 노동패널조사 등은
총화집락추출법에 의한 복잡표본조사임)
- ▶ 한국패널조사의 대표격인 노동패널가구조사에서 가중치 계산 사례를 예시
(1998년에 시작한 노동패널가구조사는 노후화와 이탈률의 누적으로 대표성 훼손을
보완하기 위해서 12차년도인 2009년에 1,415가구를 추가패널로 포함)
- ▶ 장기간 운영한 패널조사는 이탈률의 누적과 모집단의 특성변화 등으로 대표성을 보완할
필요가 생길 경우에 노동패널추가 구축방법과 추가구축 후에 가중치 산출과 분석방법에
관한 내용을 참고할 수 있을 것임
- ▶ 표본조사에서 가중치 산출방법도 중요하지만 가중치의 효과와 분석방법도 동시에 고려해야
하며 이에 대한 연구와 분석 및 결과해석은 조사내용과 표본설계에 따라서 상이하므로
전문가의 자문이 필수적임