균형할당표본추출

Balanced Quota Sampling

허 명 회 · 황 진 모

고려대학교 통계학과 교수 , 고려대 통계학과 석사과정

개 요

평일에는 인구사회적 속성에 따른 재택률 차이가 심하므로 전화조사의 지역*성*나이대 할당표본추출은 심각한 응답자선택편향을 가질 수 있다.

조사시간대를 할당변수로 추가한 '균형할당표본추출'(balanced quota sampling) 방법을 제안한다.

통계청 2004년 생활시간조사 원자료에 적용하여 그 유효성을 살펴본다.

한국조사연구학회 학술대회 2006.6.16

1. 서론

- **전화조사에서의 할당표본추출**
  - 우리나라 대부분 민간조사전문기관에서 시행하고 있는 방법
    - 지역*성*나이대 총 할당표를 준비한다.
    - 시작시각 (오후 1시)부터 전화번호부 데이터베이스에서 임의 추출한 번호에 전화를 건다.
    - 전화를 받은 가구에서 응답자가 조사에 응하면 해당 지역*성*나이대 칸에 여분이 있는 경우 면접에 들어가고 그 칸에서 1을 감한다.
    - 다음 번호에 전화를 건다. 모든 칸에서 남은 숫자가 0이 되거나 마감시각 (오후 10시)이 되면 조사를 종료한다.
    - 칸 가중법(cell weighting)으로 각 조사개체에 가중치를 부여한다. 할당표에 맞게 표본이 확보된 경우 이 과정은 필요 없다.

1. 서론

- 전화조사에서의 할당표본추출
  - 장점과 단점
    - 조사를 신속하게 진행할 수 있다.
    - 낮 시간 재택자 중심의 조사 자료가 만들어진다.
    - 직업, 교육수준 등 사회경제적 측면에서 조사 자료가 편향될 가능성이 있다.
    - 늦은 시간대에서는 조사 효율성이 떨어진다.
    - ❖ 조성겸 (1997), 허명회 외 2인(2003)
  - 임의표본추출: 원칙
    - 많은 재통화(call-back)가 필요하다.
    - 가구내 응답자 선택(respondent selection)이 쉽지 않다.
    - ❖ 양승목 외 2인 (1991), 노규형 외 2인(2002)

1. 서론

- 제안: 균형할당표본추출
  - 전화조사에서 기존의 할당표본추출을 개선하고자 한다.
  - 조사시간대를 새로운 할당변수로 추가한 방법이다.
  - 조사 효율성이 기존의 할당표본추출에 비해 더 높거나 비슷하다.
  - 응답자선택편향이 기존의 할당표본추출에 비해 작다.
  - 통계청의 2004년 생활시간조사에 적용, 몬테칼로 결과를 제시한다.

2. 균형할당추출법

- 표기 및 정의
  - 낮 시간: 1시-6시 vs 저녁 시간: 6시-10시
  - 응답자 i 의 시간대별 재택성향:
    - 낮 h_i = 응답자 i가 낮 시간대에 집에 머문 시간 / 총 낮 시간
    - 저녁 k_i = 응답자 i가 저녁 시간대에 집에 머문 시간 / 총 저녁 시간
  - 조사규칙: 낮 시간대에 1회 전화하고 성공하지 못한 경우 저녁 시간대에 재통화를 시도.
  - 응답자 i 접촉 확률 $s_i = h_i + (1-h_i) * k_i$
  - 시간대별 조건부 확률 → 시간대별 할당비율
    - 낮 $r_i = h_i / s_i$ → $r = average(r_i)$
    - 저녁 $1-r_i = (1-h_i) * k_i / s_i$ → $1-r = average(1-r_i)$
  - 통계청의 2004년 생활시간조사자료로 r과 1-r을 추정.

2. 균형할당추출법

<표 1> 평일 낮 시간대 할당비율

서울	20대	30대	40대	50대	60대이상
남	0.29	0.17	0.19	0.26	0.42
여	0.32	0.49	0.43	0.45	0.56

인천/경기	20대	30대	40대	50대	60대이상
남	0.26	0.17	0.20	0.29	0.47
여	0.43	0.46	0.45	0.45	0.56

강원	20대	30대	40대	50대	60대이상
남	0.33	0.17	0.27	0.26	0.51
여	0.28	0.43	0.46	0.49	0.57

대전/충청	20대	30대	40대	50대	60대이상
남	0.35	0.21	0.23	0.26	0.42
여	0.44	0.49	0.42	0.40	0.53

2. 균형할당추출법

<표 1> 평일 낮 시간대 할당비율

광주/전라/제주	20대	30대	40대	50대	60대이상
남	0.30	0.22	0.24	0.33	0.45
여	0.38	0.46	0.41	0.47	0.51

대구/경북	20대	30대	40대	50대	60대이상
남	0.33	0.26	0.22	0.33	0.48
여	0.40	0.53	0.41	0.40	0.50

부산/울산/경남	20대	30대	40대	50대	60대이상
남	0.35	0.22	0.22	0.33	0.49
여	0.36	0.53	0.45	0.45	0.56

전국평균	20대	30대	40대	50대	60대이상
남	0.32	0.20	0.22	0.29	0.46
여	0.37	0.48	0.43	0.44	0.54

총평균 37.8%

2. 균형할당추출법

- 제안방법론: 균형할당추출법
  1) 총할당표(overall quota table)에 <표 1>의 시간대 할당 비율을 곱하여 지역*성*나이대 부할당표(sub-quota table)를 만들고 오후 6시 이전까지는 이를 기준으로 표본추출을 하는 한편 비재택 가구들의 전화번호 리스트를 만든다.
  2) 오후 6시 이후에는 낮 시간대 비재택 가구 리스트에서 전화번호를 순서대로 재추출하여 응답자 접촉에 들어가되 지역*성*나이대 총할당표(overall quota table)를 기준으로 한다.

3. 몬테칼로 연구

- 통계청의 2004년 생활시간조사
  - 전국 12,750가구의 10세 이상 가구원 31,634명의 2일간 시간일지 기록
    - 20세 이상 21,058명.
  - 10분 단위 시간일지(time diary)
  - 9개 대분류, 50개 중분류, 137개 소분류 코딩
  - 자료세트 P1 (가상 모집단):
    - 서울, 인천/경기, 대전/충청, 광주/전라/제주, 대구/경북, 부산/울산/경남 등 7개 지역으로 구분
    - 지역 자료수가 전국 20세 이상 인구분포에 비례하도록 개인의 1일 자료를 총 100,000명이 되도록 부스팅(boosting).

3. 몬테칼로 연구

- 가상(몬테칼로) 전화조사
  - P1을 임의순서로 정렬, 그 순서에 따라 각 응답자에 일정 간격의 접촉 시각 t를 부여.
  - 시각 t에서 해당 응답자가 재택 상태이면 조사 성공, 자료 값을 가져옴.
  - 조사간격 4초를 가정 (시간당 900명 접촉).
  - 기존할당추출: 총 시도 7,504번, 성공 1,000번.
  - 균형할당추출: 총 시도 7,038번, 성공 1,000번.

3. 몬테칼로 연구

- 기존할당추출 (Monte Carlo): 시간대별 실사진행

	1시-	2시-	3시-	4시-	5시-	6시-	7시-	8시-	9시-	합계
총 시도	900	900	900	900	900	900	900	900	304	7504
조사 실패	629	648	728	810	830	830	855	873	301	6504
조사 성공	271	252	172	90	70	70	45	27	3	1000
성공률(%)	30%	28%	19%	10%	8%	8%	5%	3%	1%	13.3%

- 1시부터 6시 이전: 총 시도 4,500건, 조사성공 774건, 조사성공률 = 17.2%
- 6시부터 10시까지: 총 시도 3.004건, 조사성공 145건, 조사성공률 = 4.8%
- 6시 이전 응답자 비율 85.5%

3. 몬테칼로 연구

- 균형할당추출 (Monte Carlo): 시간대별 실사진행

	1시-	2시-	3시-	4시-	5시-	6시-	7시-	8시-	9시-	합계
총 시도	900	900	900	900	278	900	900	900	460	7038
조사 실패	676	778	880	897	276	645	651	784	451	6038
조사 성공	224	122	20	3	2	255	249	116	9	1000
성공률(%)	25%	14%	2%	0%	1%	28%	28%	13%	2%	14.2%

- 1시부터 6시 이전: 총 시도 3,878건, 조사성공 371건, 조사성공률 = 9.6%
- 6시부터 10시까지: 총 시도 3.160건, 조사성공 629건, 조사성공률 = 19.9%
- 6시 이전 응답자 비율 37.1%

3. 몬테칼로 연구

- 기존할당추출 (Monte Carlo): 결과

변수	n	평균	표준편차	최소값	제1사분위수	중간값	제3사분위수	최대값
수면시간	1,000	442	102	0	380	450	500	880
주업종사	1,000	119	191	0	0	0	245	790
TV시청	1,000	166	128	0	70	140	240	690
재택시간	1,000	1,095	268	240	880	1,160	1,320	1,440

- 비교: 모집단 P1

변수	N	평균	표준편차	최소값	제1사분위수	중간값	제3사분위수	최대값
수면시간	100,034	436	84	0	390	440	480	900
주업종사	100,034	237	234	0	0	230	450	1,190
TV시청	100,034	111	107	0	30	90	160	930
재택시간	100,034	859	290	0	640	790	1,090	1,440

- 수면시간에서는 기존할당추출 442분, 모집단 436분으로 별 차이가 없다.
- 주업종사시간에서는 기존할당추출 119분, 모집단 237분으로 50% 과소.
- TV시청시간에서는 기존할당추출 166분, 모집단 111분으로 50% 과다.
- 재택시간에서는 기존할당추출 1,095분, 모집단 859분으로 27% 과다.

3. 몬테칼로 연구

- 균형할당추출 (Monte Carlo): 결과

변수	n	평균	표준편차	최소값	제1사분위수	중간값	제3사분위수	최대값
수면시간	1,000	442	87	60	390	450	490	790
주업종사	1,000	171	207	0	0	0	390	700
TV시청	1,000	138	111	0	60	120	190	660
재택시간	1,000	986	255	260	770	950	1,210	1,440

- 비교: 모집단 P1

변수	N	평균	표준편차	최소값	제1사분위수	중간값	제3사분위수	최대값
수면시간	100,034	436	84	0	390	440	480	900
주업종사	100,034	237	234	0	0	230	450	1,190
TV시청	100,034	111	107	0	30	90	160	930
재택시간	100,034	859	290	0	640	790	1,090	1,440

- 수면시간에서는 균형할당추출 442분, 모집단 442분으로 별 차이가 없다.
- 주업종사시간에서는 균형할당추출 171분, 모집단 237분으로 28% 과소.
- TV시청시간에서는 균형할당추출 138분, 모집단 111분으로 24% 과다.
- 재택시간에서는 균형할당추출 986분, 모집단 859분으로 15% 과다.

4. 맺음 말

- **균형할당표본추출이 기존할당추출에 비해**
  - 작은 편향을 갖는다 (1/2 수준).
  - 실사효율성이 비슷하거나 좋다: 균형 성공률 14.2% vs 기존 성공률 13.3%
  - 응답률(response rate)이 높다: 균형 응답률 25.8% vs 기존 응답률 13.3%

- **균형할당표본추출에서도 지속되는 문제**
  - 편향이 작아지지만 없어지지 않는다 (→ 휴일에 재통화)
  - 조사성공률이 시간 경과에 따라 떨어진다.
    - 6시대와 7시대 28%, 8시대 13%, 9시대 2%
  - 대책: 8시 이후 지역*성*나이대 할당에 관계없이 응답자와 접촉
    - 8시대 성공률 50%, 9시대 68%
    - 저녁시간대 자료에 칸 가중치(cell weights)를 부여하여 자료를 보정

4. 맺음 말

- **조사 시간대 제어가 갖는 의미**
  - 직업 (학력수준) 분포의 균형화
  - 미디어 노출(TV 시청량) 효과 제어

- **향후 과제**
  - 균형할당추출법의 실용적 버전 개발: 예컨대
    - 낮 시간대과 저녁 시간대로 나누어 지역*성*나이대에 표본 크기를 할당하되
    - 8시 이후에는 할당 제한 없이 실사를 진행하여 조사 효율성을 높이는 방법
  - 실제조사에서 기존할당추출과 균형할당추출을 실험적으로 비교해보는 것.