스노우볼 샘플링 (Snowball Sampling)

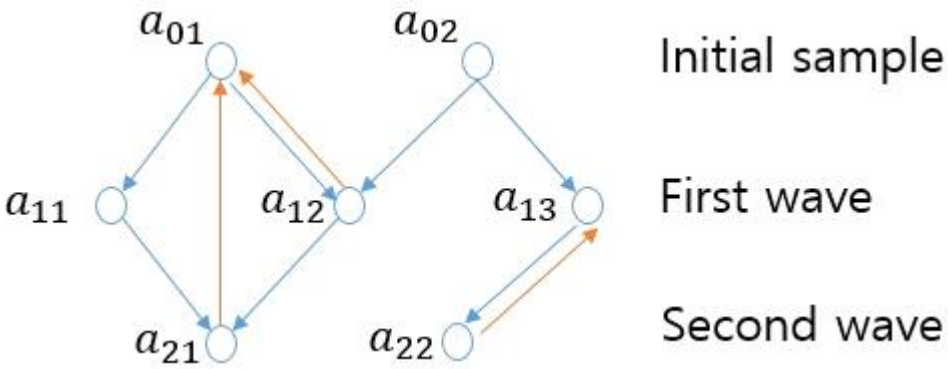
1. 기본 내용

❍ 스노우볼 샘플 (“snowball sample”)이라는 용어와 기본 방법론 아이디어는 Merton (1949)과 Katz and Lazarsfeld (1955)에 소개되어 있지만, 체계적인 조사 방법론으로서의 스노우볼 샘플링 방법에 대한 시초는 Coleman (1958)의 연구로 받아들여지고 있음.

❍ Biernacki and Waldorf (1981)에 의하면, 표본추출틀이 없거나 조사가 어려운 집단 (hard-to-reach population) 조사를 위하여 1차적으로 표본을 뽑고 해당 표본을 중심으로 순차적으로 추가 표본을 뽑는 link-tracing sampling이 유효하다고 하며 그 중에서 스노우볼 샘프링이 가장 대표적이라고 함. 마리화나 흡연자 조사 (Becker, 1963)에 도입된 후에 지금까지 많은 사회 조사에 사용되고 있음.

❍ 스노우볼 샘플링 (종종 chain-referrel sampling이라고도 불림)의 기본 아이디어는 확률 혹은 비확률로 소표본을 뽑은 후에 해당 표본으로부터 그 다음 표본을 순차적으로 생성하는 방법을 의미함. 예를 들어, 1차적으로 표본 a_{01}를 뽑은 후에 설문을 받고 그 다음 설문 대상자 a_{11}를 추천 받아 다음 설문을 진행하는 것임.

<그림 1> 스노우볼 샘플링 모식도



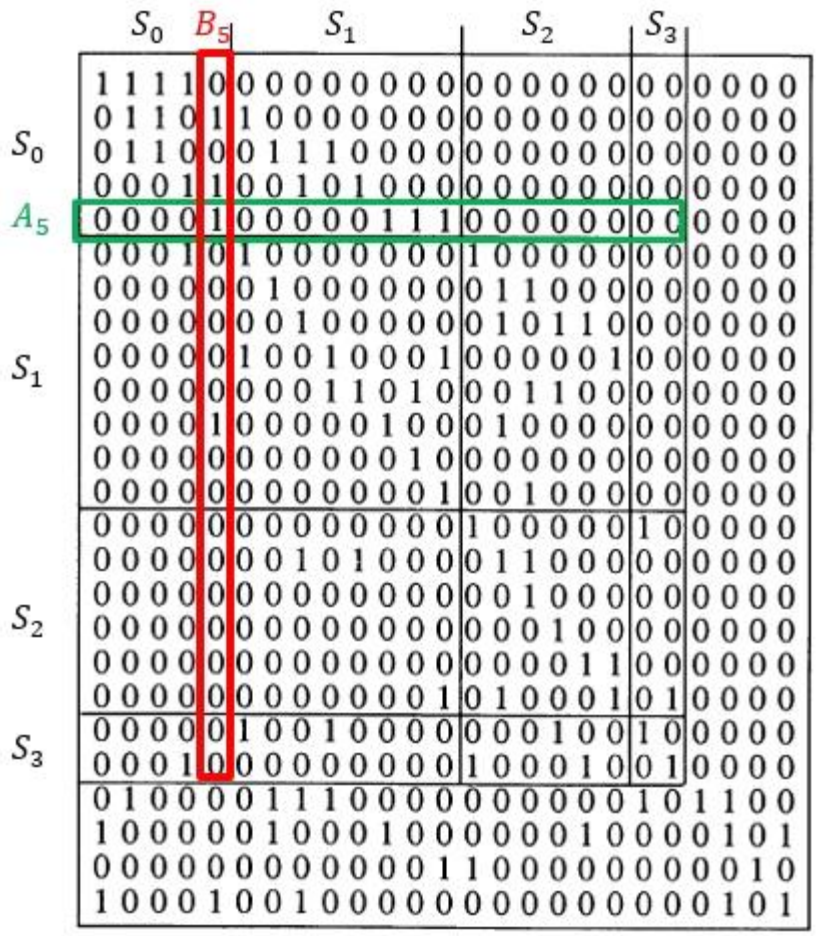
❍ <그림 1>에 스노우볼 샘플링의 표본 수집이 표현되어 있음. 초기 표본으로 a_{01}과 a_{02}를 선택한 후 이들에게 1차 웨이브 표본인 a_{11}, a_{12}, a_{13}를 추천받아 설문을 진행함. 이후 1차 웨이브 표본들에게 2차 웨이브 표본을 추천 받는데, a_{11}가 a_{21}를 소개하였고, a_{12}는 a_{21}과 a_{01}을 소개하였고, a_{13}은 a_{22}를 소개하였음. 이 때, 추천 받은 a_{01}은 이미 초기 표본으로 선택되었기 때문에 전체 표본 크기를 늘리지는 않음.

❍ 2차 웨이브 표본은 a_{01}과 a_{13}을 추천하였는데 모두 이전 단계에서 표집되었음. 이렇게 새로운 표본이 표집되지 않으면 스노우볼 샘플링은 완료되고 순차적으로 표집된 모든 원소를 대상으로 최종 샘플을 구성함. <그림 1> 예시에서 최종 표집된 원소는 $\{a_{01}, a_{02}, a_{11}, a_{12}, a_{13}, a_{21}, a_{22}\}$가 됨.

2. Frank and Snijders (1994)'s method

❍ 스노우볼 샘플링을 활용하여 hard-to-reach 그룹의 크기를 추정하는 방법은 다양함. Frank and Snijders (1994)은 그래프 이론 기반의 추정 방법론을 제한하였고 해당 내용은 응답자 기반 모형 (Respondent-Driven Modeling 에 채택되었음)

<그림 2> 스노우볼 샘플링 인접행렬 모식도 (Frank and Snijders, 1994)



❍ <그림 2>에 소개된 인접행렬을 위해서는 몇 가지 정의가 필요함. 모집단 원소 집합을 $V=\{1,...,v\}$로 표기함. 원소들의 관계를 호(arc)로 정의하며 (i, j)로 표기함. 만약 i=j인 경우를 루프(loop)라고 칭함. 또한, 모든 루프를 포함하는 호의 집합을 $W \in V^2$이라고 정의함.

❍ 초기 표본 (initial sample)을 S_0으로 표기하고 이 후 순차적으로 표집된 표본 집합을 S_1, S_2라고 표기함. 만약에 모집단에 있는 원소 i가 표본으로 표집 되었으면 $x_i = 1$이라고 표기하고 그렇지 않으면 $x_i = 0$이라고 기록함. 만약 원소 i가 j를 알고 있으면 $y_{ij} = 1$이라고 놓고 그렇지 않으면 $y_{ij} = 0$이라고 기록함. 이 때, 일반적으로 y_{ij}와 y_{ji}는 같지 않음. 이러한 가정으로 방향성이 있는 그래프 (directed graph)가 되는 것임. 원소 j가 알고 있는 원소의 집합은 A_j, 원소 j를 알고 있는 원소의 집합은 B_j라고 표기하며 각 각의 사이즈를 a_j와 b_j라고 정의함:

$$A_j = \{i \in V : y_{ji} = 1\}$$
$$B_j = \{i \in V : y_{ij} = 1\}$$

추가적으로 표본 집합 S로부터 정의되는 원소 집합 $A(S)$와 $B(S)$를 다음과 같이 정의할 수 있음.

$$A(S) = \cup_{j \in S} A_j, \qquad B(S) = \cup_{j \in S} B_j.$$

❍ 정의된 내용에 기반하여 <그림 2>를 다음과 같이 해석할 수 있음. <그림 2>는 25개 원소에 대한 인접행렬 (adjacent matrix 혹은 arc matrix)을 나타내고 있음. 초기 표본으로 5개의 원소가 잡혔으며, 1차 웨이브 표본으로 8개 2차 웨이브 표본으로 6개 3차 웨이브 표본으로 2개가 있음. 4개(마지막 컬럼 4줄)는 표집과정에 포함되지 않음.

❍ 초기 표본으로 선택된 5번째 원소가 알고 있는 원소는 본인 자신을 포함하여 4개인 것을 알 수 있으며 (A_5), 5번째 원소를 알고 있는 원소 또한 본인 자신을 포함해서 4개인 것을 알 수 있음 (B_t).

❍ 1차 웨이브 표본의 원소를 알고 있는 원소는 총 16개(4번째 열부터 19번째 열까지)인데, 이 중에서 2개는 초기 표본에 들어가 있고, 8개는 1차 웨이브 표본에 표집된 원소 자체이고, 나머지 6개는 2차 웨이브 표본에 들어가 있음.

❍ <그림 2>와 같이 인접행렬이 정의되고 나면, 식 (2)를 이용하여 숨겨진 모집단 크기를 추정할 수 있음.

$$\hat{v} = n + s(n-1)/r \quad (2)$$

❍ 식 (2)에서 n은 초기 표본의 크기이고, r은 초기 표본에서의 비루프 호(nonloop arcs)의 크기임. 따라서 $r+n$은 초기 표본에서 정의된 모든 호(arcs)의 크기이며 s는 초기 표본이 첫 번째 웨이브 표본의 원소를 알고 있는 경우의 수의 합으로 $s = |W \cap (S_0 \times S_1)|$로 정의됨.

❍ <그림 2>의 초기표본과 1차 웨이브 표본을 이용해서 $r+n$을 구하면 12가 됨 (대각 5개+비대각 7개). s는 초기 표본이 첫 번째 웨이브 표본의 원소를 알고 있는 경우의 수의 합으로 $s = |W \cap (S_0 \times S_1)|$로 정의됨. 본 예제에서는 9의 값을 가짐. 해당 값들을 식 (2)에 대입하면 숨겨진 모집단의 크기는 약 100이 됨. 인접행렬에 들어가는 지시함수값들의 동일한 이항 분포로 정의되기 때문에 가정과 실제와의 괴리도에 따라서 과소 추정되거나 과대 추정될 수 있음.

3. Vincent and Thompson (2017)'s approach

❍ Vincent, K. and Thompson, S. (2017). Estimating population size with link-tracing sample. Journal of the American Statistical Association, 112. 1286-1295.

4. Summary

❍ 핵심 논문을 자세히 읽고 나서는 본인 만의 기준(research topic or interest)에 맞게 정리 및 요약하는 과정이 필요함. 표와 그림으로 표현할 수 있으면 제일 좋음.

❍ 일반적으로 10~20편의 논문을 자세히 읽으면 흐름이 보임. 리뷰 논문이 있다면 리뷰 논문을 먼저 보는 것도 도움이 됨. 그러나 본인 만의 언어로 재정리하는 노력은 꼭 해봐야 함. 방법론을 입체적으로 이해해야 함.

❍ Literature review를 논문으로 옮길 때는 작성한 리뷰의 핵심적인 내용들을 추려서 정리하면 됨. 훈련이 많이 필요함.