한국통계학회논문집
8권 기념호, 2001
pp. 81-90

조사통계 이론의 발전과 전망

신민웅[1)]

1.머리말

조사통계는 여러 분야에서 필요성이 높아져 매우 빠른 성장을 하고 있다. 조사통계의 역할은 통계 이용자들이 요구하는 통계자료들을 생산, 제공하는 것이라고 할 수 있다. 21세기에 들어 지식, 정보화 사회로의 사회변화가 급진전되면서 통계기능도 크게 달라져 오늘날의 통계는 정부, 기업, 개인의 자유로운 판단과 의사결정을 뒷받침하는 사회의 정보인프라의 핵심적인 구성요소의 하나로 인식되고 있다. 이와같이 통계조사는 그 중요성이 높아가고, 규모도 급속히 확대되어 가고 있다. 국가기관에서 실시하고 있는 각 종 통계 조사, 조사업체에서 마케팅 조사, 사회단체에서 시행하는 사회조사, 공업분야에서 실시하는 품질조사가 있는데, 그 외에도 셀수 없이 많이 표본조사(sample survey)가 이루어 지고 있다. 조사통계 업무의 발전 방안으로 과거와 현재를 돌아 보면서 미래에 유의하여야 할 문제점을 지적하고자 한다. 우리가 논의하고자 하는 주제는 크게 나누어, 자료의 수집방법, 표본조사 실시방법, 자료의 분석 방법이다.

간략히 말하여 자료의 수집방법은,

(1)표본조사 (2)직접측정(direct measurement)
(3)관찰(observation) (4)이차적 정보(secondary information)
등이 있다.

표본조사 실시 방법으로는,

(1) 우편 조사 (2) 면접 조사 (3) 전화 조사 (4) 인터넷 조사
등이 있다.

앞으로 많이 활용되어야 할 자료의 분석 방법으로는,

(1) 무응답에 대한 분석
(2) 복합조사(complex survey)에서 분산 추정 (variance estimation)
(3) 복합조사(complex survey)에서 범주형 자료분석
(4) 복합조사(complex survey)에서 회귀 분석
등이 있고

1) 한국외국어 대학교, 정보통계학과 교수
E-mail : mwshin@stat.hufs.ac.kr

표본조사로는

(1) 패널조사(panel survey)
(2) 국제조사(international survey)
(3) 소지역 추정(small-area estimation)
등이 있다.

2. 조사통계의 현황과 발전

조사방법론에 있어서 설문지 설계, 표본조사 기법, 자료수집 방법, 자료처리 과정 그리고 조사분석의 발전이 이루어 지고 있다. 특히 컴퓨터가 조사방법의 발전에 큰 원동력이 되었다. 컴퓨터의 발전으로 계산력이 커지므로서 자료의 수집은 물론 표본설계에서도 많은 정교한 기법이 개발되고 전에는 할수 없었던 통계분석도 가능해졌다.

조사 정보수집에도 컴퓨터를 이용한 CASIC(computer assisted survey information colledtion)방법이 발달되고 있다. 즉

(1) CATI (computer assisted telephone interviewing)
(2) CAPI (computer assisted personal interviewing)
(3) CSAQ (computerized self-administered data collection)
(4) CASI (computer assisted self interviewing)
(5) TDE (tdlephone data entry)
(6) VRE (voice recognition entry)
등이 실시되고 있다.

온라인 조사는
(1) 회원 조사 (member survey)
(2) 방문자 조사 (vistor survey)
(3) 전자우편 조사 (e-mail survey)
(4) 전자 설문 조사 (electronic survey)
로 구분될수 있으며 정보통신망을 매개로 하는 온라인 사회조사는 자료를 수집하는 방법으로 그 활용이 늘어나고 있다.

무응답율은 증가하는 추세로 응답률을 높이려는 노력이 필요하게 되었다. 또한 무응답이 생겼을 때 무응답을 대체(inputation)하여 통계분석을 하는 기법이 발전되고 있다. 내용심사(editting)는 대체법과 같이 앞으로의 연구에 있어서 중요한 주제가 될 것이다. 전화조사는 특별히 가중(weighting)과 조정(adjustment)을 필요로 한다. 무응답 자료를 가중-크래스로 나누는 데는 더 많은 보조변수가 필요하고, 따라서 컴퓨터의 계산력의 도움을 받아야 한다. 컴퓨터의 도움으로 표본설계도 더 효율적으로 할 수 있다.

지방자치제의 발달과 특수 지역의 요구에 따라서 소지역 추정이 중요한 연구 과제가 되고 있다. 복잡한 표본설계에서 필요한 분산추정도과 다중대체도 컴퓨터의 도움으로 더욱 효률적으로 처리

할수 있게 되었다.

3. 조사통계의 발전을 위한 통계기법

3.1 자료의 수집 및 표본조사방법

표본조사에서 응답자가 응답 자체를 안하거나 어떤 질문 항목에 대하여 응답을 하지 않는 경우에 무응답 문제는 표본설계나 분석에 큰 영향을 준다. 무응답을 미리막고, 무응답을 측정하고 그 영향을 줄이는 것은 매우 중요한 일이다.

무응답을 다루는 가장 좋은 방법은 무응답이 일어나지 않게 미리 예방 수단을 쓰는 것이다. 무응답자는 응답자들과 본질적으로 다를 수 있다. 만일 무응답을 무시할 수 없다면 응답자들만을 근거로 하는 추론은 심각한 결점이 될 수 있다.

3.1.1 질문지 설계(questionnaire design)

질 높은 조사를 위해서는 질문지 설계가 큰 역할을 한다. 조사통계는 질문지 설계부터 시작된다고 할 수 있고, 질 높은 조사를 위해서는 질문지 작성이 잘 되어야 한다.최근에도 질문지 설계,내용심사(editting)에 대한 연구가 활발히 이루어 지고 있다.

3.1.2 컴퓨터 보조 방법(computer assisted methods)

최근에 컴퓨터 보조방법(computer assisted method)의 이용이 활성화 되고 있다. 즉, CAPI와 CATI가 자료 수집에 대한 프로그램이 발전되고 있다 .인터넷 상에서 조사자료는 수집하는 방법도 발전되고 있다. 이 방법은 인터넷을 사용하고 또 접근 가능한 사람들로 구성된 모집단의 조사에는 실제적으로 매력적이다. 이메일(email)로 질문지를 보내거나, 웹 사이트에 질문지를 띄워서 조사한다. 그러나 샘플링 후레임이 미비하거나 응답률이 낮을 때에는 인터넷 조사가 적절하지는 않다.

3.1.3 비표본오차를 줄이기 위한 조사 설계

잘못된 표본설계의 공통 특징은 표본설계를 하는 시간과 무응답을 추적 조사하는데 드는 시간이 부족하다는 것이다. 조사를 처음하는 많은 사람들은 단순히 자료 수집과정에서 야기될 수 있는 잠재적인 문제들을 고려하지 않고 바로 자료수집부터 시작한다. 보통 우편 설문지를 발송하고 회수된 자료를 분석하는데, 이러한 조사는 낮은 응답률을 보인다.

● 수용할 수 있는 응답율은 어느정도인가?

연구자들이 종종, "이 조사에서는 60%의 응답률을 기대한다"고 말한다. 그것을 받아들일 수 있는가, 또 그 결과가 타당한가? 그 질문에 대한 답변은 무응답의 특성에 의존한다. 만일 무응답자가 MCAR이면 무응답을 무시하고 응답자들을 모집단 표본을 대표하는 것으로 사용할 수 있다.

만일 무응답자들이 응답자들과 다른 경향을 보인다면 응답자들만을 사용해서 얻은 결과로 인한 편의는 조사 전체를 가치없게 할 수있다. 많은 문헌들에는 응답률의 수용성에 관한 조언들이 있다. 예를들어, Babbie(1973)는 다음과 같이 말했다. " 나는 적어도 50%의 응답률이 분석이나 보고에 적절하다고 생각한다. 적어도 60%의 응답률은 좋다. 그리고 70%의 응답률은 매우 좋다". 이러한 절대적인 가이드 라인은 위험하며 많은 조사 연구자들을 무응답에 대해서 자기만족에 그치게 할 수 있다. 70%의 응답률을 보인 조사의 많은 경우, 그 결과에 오점이 있을 수 있다. NCVS(National crime victimization survey)에서는 95%의 응답률을 보일지라도 무응답 편향에 관한 수정이 필요하다.

사용되는 응답률이 어떻게 정의되는 가에 따라 응답률에 대한 매우 다른 결과들이 생긴다. 다음은 조사에서 사용되는 응답률이다.

① $\frac{\text{완전히 조사된 수}}{\text{표본에 있는 개체 수}}$

② $\frac{\text{완전히 조사된 수}}{\text{접촉한 개체 수}}$

③ $\frac{\text{완전히 조사된 수 + 부적격 개체}}{\text{접촉한 개체 수}}$

④ $\frac{\text{완전히 조사된 수}}{\text{접촉한 개체 수 − (부적격 개체)}}$

⑤ $\frac{\text{완전히 조사된 수}}{\text{접촉한 개체 수 − (부적격 개체) − 거부자수}}$

처음 식을 사용하여 응답률을 계산하는 것보다 마지막 식을 이용하여 계산하는 것이 분모가 더 작기 때문에 "응답율"이 훨씬 커진다.

Statistics Canada(1993)과 Hidiroglou 등(1993)은 응답률을 보고하는데 가이드라인을 제공하였다. 그들은 다음 사항이 포함되는 조사에서 여러 다른 응답률을 보고하도록 제안하였다.

- 비접촉율(no-contact rate) : 접촉되지 않을(no-contacts) 개체수의 비
- 거부율(refusal rate) : 범위-내 개체 수에 대한 거부한 개체 수의 비
- 무응답율(nonresponse rate) : 무응답을 한 개체 수의 비

다음 조언들은 Genzalez 등(1994)이 기록한 것으로, U.S. Office of Management and Budget's Federal Committee on Statistical Methodology 에 의한 것인데 도움이 된다.

조언 1. 조사 스탭들은 시간에 따라 일정한 형식으로 응답률을 계산하여야 한다.

조언 2. 반복적인 조사를 하는데 있어 표본설계의 변경이나 비용등을 서류화할 때에 조사 스탭들은 응답률 요소들(거부, 집에 없음, 범위-밖, 주소불명, post-master returns 등)을 시간에 따라 모니터(monitor)해야한다.

조언 3. 응답률 요소들은 조사 보고서에 실어야 한다.

3.1.4 응답률과 자료의 정확성에 영향을 주는 요인들

(1) 조사 내용
약물 사용이나 경제 관련 문제에 대한 조사는 거부하는 경우가 많다.
(2) 조사 시점
휴가기간에는 가구조사가 좋지 않은 시기이다.
(3) 조사자 (interviewer)
조사자에 따라 응답률간에 변이가 크다.
(4) 자료 수집 방법
전화와 우편조사는 면접 조사보다 응답률이 낮다.
(5) 설문지 설계
우편조사에서 응답자에게 잘 설계된 설문지는 자료의 정확도를 높힌다.
(6) 조사 소개 (survey introduction)
응답자들은 자료가 어떤 목적으로 사용 되는지를 알아야 한다.
(7) 보상 (incentives)
보상은 응답률을 증가시킨다.
(8) 추적 조사 (follow-up)
조사 예산은 조사설계나 무응답의 추적조사에 대해 할당하여야 한다.

3.1.5 재방문 (callback)

실질적으로 표본조사는 처음에 조사되지 못한 사람들로부터 응답을 얻기 위하여 다시 전화를 걸거나 재방문에 의존한다. 재방문 자료의 분석으로부터 무응답자들로부터 나올 수 있는 편의에 관한 정보를 얻을 수 있다.
재방문이 사용되는 조사가 설계될 때에 초기 조사는 대부분 우편조사에 의해 행해진다. 그러나 추적조사는 개인 면접법과 같이 비용이 많이 드는 조사법을 이용한다.

3.2 자료의 분석 방법

3.2.1 무응답에 대한 자료 분석

(1) 가중-크래스의 형성

가중 크래스(weighting-class) 방법은 비표본 오차를 보정하려는 방법이다. 표본추출된 모든 원소에 대하여 알려진 변수들은 가중-크래스들을 형성하는 데 사용되고, 같은 가중-크래스 내에서는 무응답자와 응답자가 유사할 것이라고 기대한다.

가중-조정 크래스들은 층으로 간주하여 형성되어야 한다. 크래스들은 각 크래스 내에 있는 원소들이 관심있는 주요 변수들에 관해서 가능한 유사하고, 각 크래스간의 응답율이 크래스마다 다르게 나타나도록 크래스가 형성되어야 한다.

(2) 사후층화 (poststratification)

사후층화는 층의크기 N_h가 가중치를 조정한다는 것을 제외하면 가중-크래스 조정과 유사하다. 표본이 수집된 후 개체들은 대개 성별과 인종같은 인구통계학 변수를 기초로 H개의 다른 사후층으로 분류된다. 즉, 사후층화는 가중-크래스와 유사하지만 h 사후층의 크기가 N_h일 때 N_h가 가중치를 조정한다.

(3) 대체법 (imputation)

조사에서 결측 항목들은 여러 가지 이유로 나타난다. 대체법은 결측 항목에 대체값을 할당한다. 대체법에는 다음과 같은 것이 있다.

[a] 연역적 대체 (deductive imputation)

어떤 값들은 자료를 편집할 때 변수들의 논리적인 관계를 이용하여 대체된다. 연역적인 대체는 반복적인 조사에서 사용될수 있다.

[b] 칸 평균 대체 (cell mean imputation)

응답자들은 가중-크래스 조정에서와 같이 알려져 있는 변수를 근거로 여러 크래스(칸)으로 나뉜다. 그러면, 칸 c에 있는 응답 개체들에 대한 평균값은 각 결측값에 대치된다. 어떤 칸에 있는 응답 개체들에 대한 평균값으로 대치한다.

[c] 핫덱 대체법

핫덱 대체법에서는 평균 대치법이나 가중-조정 방법에서와 같이 표본 개체들을 크래스들로 구분한다. 각 결측 응답에 대해서 같은 클래스에 속하는 응답 개체 값이 대치된다. 핫덱 이라는 명칭은 컴퓨터 프로그램과 자료 집합들을 카드에 구멍을 낼때부터 비롯된다. 분석하고자 하는 자료 집합들을 포함하는 카드들의 데크(deck)는 카드 판독자들에 의해 다루어진다. 따라서 핫덱이라는 용어는 같은 자료 집합에 있는 자료로 대치됨을 의미한다. Fellegi와 Holt(1976)은 자료 편집과 규모가 큰 조사에서의 핫덱 대체법에 대해 논의하였다.

대체하는 개체는 어떻게 선택할 것인가? 이에 대해 여러 가지 방법이 가능하다. 같은 클래스(class)에 속하는 개체 값이 대치된다.

[d] 회귀 대체법

회귀 대체법(regression imputation)은 모든 경우에 관찰된 변수들에서 관심있는 항목에 회귀분석을 이용하여 결측값을 예측하는 방법이다. 확률적 회귀대체법(stochastic regression

imputation)이 있는데, 결측값이 회귀모형에 예측된 값에다 랜덤하게 생성된 오차항이 추가되어 대체된다. 회귀 분석을 이용하여 결측값을 예측하는 방법이다.

[e] Cold-deck 대체법

Cold-deck 대체법에서는 대치되는 값들은 예전의 조사나 다른 정보들 로부터 얻어진다. (대치된는 근거로 사용되는 자료 집합은 현재 컴퓨터에서 작업되지 않기 때문에 "cold"라는 용어를 사용한다.)
대치되는 값들이 과거의 조사나 다른 정보로부터 얻어진다.

[f] 교체 (substitution)

교체(substitution)방법은 cold-deck 대치와 유사하다. 때때로 조사자들은 필드에 있는 동안 대체를 선택할수 있도록 허용된다. 즉, 만일 표본에서 선택된 가구가 집에 있지 않는 다면 그 다음 집을 선택해도 좋다. 옆 가구가 모집단에서 임의로 선택된 가구들보다 더 유사할 때에대체는 무응답 편의를 줄이는데 도움이 되기도 한다. 만일 무응답이 관심있는 특징과 관련된 것이라면 무응답편의가 발생한다. 조사자들은 필드(field)에 있는 동안 대체를 선택할 수 있다.

[g] 다중 대체법 (multiple imputation)

다중 대체법(multiple imputation)에서는 각 결측값이 $m(\geq 2)$번 대치된다. 일반적으로 각 대치에서는 같은 확률 모형이 사용된다. 결측이 없는 m개 서로 다른 "자료" 집합들을 만든다. 다중대치에 대한 자세한 내용은 Rubin(1987,1996)에 있다. 각 결측값이 2번 이상 대치된다.

3.2.2 분산 추정 (variance estimation)

모평균과 모총계는 가중(weight)을 써서 쉽게 추정된다. 분산을 추정하는 것은 더 복잡하다. 더 나아가 복합표본조사에서 분산을 구하는 공식이 없는 경우도 있다. 콤플렉스 서베이에서 분산을 구하는 방법은 다음과 같다.

① 선형화(Linearization) 방법

테일러 급수는 오랫동안 통계학에서 분산의 조사치를 구하는데 사용되어 왔다 woodruff(1971)는 복잡한 조사에서 테일러 급수 조사치를 사용했다. 즉, 비선형함수를 테일러 급수(Taylor series)로 전개하여 선형으로 근사시킨 후 분산추정을한다.

② 랜덤 그룹(group)

SRS로 크기 n인 표본이 추출되었다면 크기가 n/R인 R개의 그룹으로 나눈다. 이 R개의 그룹들을 독립적으로 반복 추출된 것으로 간주하여 처리한다.

③ 재표본 추출(resampling)방법

재표본추출 방법은 표본을 그 자체가 모집단인 것 처럼 처리한다. 이 새로운 모집단에서 표본들을 추출하여 분산을 추정하는데 사용한다.

● BRR(balanced repeated replication)

표본조사를 할 때에 각 층들에서 그 psu들이 추출되도록 모집단을 층화시키는 경우를 생각한다.
BRR은 반복절반표본들(replicated half-samples)을 사용하여 분산을 추정한다. 여기서 반복 절반표본들은 균형(balanced)되어 있다.
Mccarthy(1966)을 모두 2^H 개의 균형 절반-표본(BRR) 들을 형성하여, 분산을 추정하였다. BRR은 R개의 반복 절반-표본들로 균형적으로 추출하여 분산을 추정한다.

● 잭나이프(jacknife)

BRR처럼 랜덤 그룹 방법을 확장한 것으로 반복 그룹의 중복을 허용한 다. Quenouille(1949)는 잭나이프를 이용하여 바이어스(bias)를 감소시켰다.잭나이프는 분산추정과 신뢰구간을 구하는 데 이용한다.

● 붓스트랩(Bootstrap)

표본을 모집단으로 간주하여 재표본추출(resampling)하는 방법이다. 이 방법은 신뢰구간을 직접 구하는 데 적합하다. Shao & Tu(1995)는 복합 표본조사에서 붓스트랩에 대한 결과를 요약하였다. 만일 표본이 모집단과 유사하다면 표본에서 재표본추출된 재표본의 분포는 모집단에서 재표본 추출된 표본의 분포와 유사할 것이다.

④ GVF(generalized variance function)

GVF는 많은 조사에서 표본오차를 계산하기 위하여 제공된다. GVF는 년간 발행물의 생산시에 시간을 줄이고 생산의 속도를 빠르게 한다.

3.2.3 범주형 자료 분석

표본이 자체-가중(self-weighting)이 아닌 한 칸 빈도수(cell-frequencies)는 범주들의 상대 돗수를 의미하지 않는다. 표본추출 가중(sampling weight)이 칸 비율을 추정하는 데 사용된다. 만일

표본이 자체-가중이라면 관찰치만으로 칸 비율을 추정한다.

3.2.4 회귀분석

복합 표본조사에서 회귀분석을 시행하는 데는 SAS나 SPSS 통계 프로그램을 사용할수 있다. 서로 다른 추출확률를 가질 수 있다. 추출확률이 반응변수와 관계가 있다면, 이러한 서로 다른 확률을 고려하지 않으면 회귀계수를 추정하는 데 바이어스(bias)가 생길수 있다. 집락추출에서도 SAS나 SPSS로 구한 표준오차가 틀릴 수 있다.

4. 표본조사

4.1 소지역 추정

소지역 추정방법은 첫째, 해당소지역에서 관찰된 표본자료를 이용하여 추정값을 구하는 직접 추정법(DE)과 둘째, 해당소지역에서 추출된 표본의 관찰 자료와 이와 유사한 다른 소지역들의 관찰 자료를 이용하여 추정값을 구하는 간접추정법(IE)이 있다. 간접 추정법인 합성추정량을 사용할 경우 관심변수와 관련이 있는 보조정보가 있어야 하고, 모형가정이 만족되어야 한다. 직접 추정량의 불안정성과 합성추정량의 편의에 대해 서로 보완할 수 있는 방법으로 두 추정량의 가중평균을 사용하는 복합추정량(composite estimator)이 있다.

4.2 패널조사 (panel surveys)

패널조사의 잇점은 오래 전부터 인식되어 왔다. 최근에는 계산과 통계 기법의 발전으로 더욱 유용성이 높아 졌다. 미래에는 더 많은 패널조사가 늘어 날 전망이다. 특히, 생물통계학자에 의하여 longitedinal 분석의 도움을 받을 것이다.

4.3 국제조사 (International surveys)

국제기구등에 의하여 국제조사가 여러분야에서 수행되어 국가간에 비교가 이루어 질 것이다 많은 국가에서 다른 나라 자료와 비교하고자 하는 요구가 높아 지고 있다.

4.4 행정자료와의 연계 (Linkage to administrative data)

자료수집에 있어서 행정업무와 연계는 행정기록 시스템의 확장으로 가능해 진다. 행정자료에 접근은 어려운 점이 있고, 개인의 프라이버시(privacy)를 보호하는 문제점도 있다. 행정자료의 이용으로 경비와 시간이 절약된다.

5. 결론

조사통계의 발전 방안으로는 자료의 수집을 위한 표본설계에서 분석에 이르기까지 각 단계의 자료처리를 통계적이론에 맞게 하여야 한다. 최근에 이르러 통계적인 기법이 컴퓨터의 발전과 더불어 수준이 높아져 이를 잘 적용하여야 질 높은 통계분석을 할수 있다. 조사통계의 발전을 위하여 몇가지 제안을 내고자 한다.

(1) 표본설계, 자료수집, 자료분석에 이르기까지 어느 한 단계라도 완벽하지 않으면 잘못된 결과가 나온다.
(2) 여러기관에서 조사하다가 보면 중복조사가 될 수가 있는데 이를 통합할 장치가 필요하다.
(3) 사이버 공간에서의 조사 기법에 대한 연구의 필요성이 높아가고 있다.
(4) 무응답에 대한 자료분석의 필요성이 높아가고 있다.
(5) 자료수집후에 내용심사(editing)와 통계적 분석기법의 중요성이 높아가고 있다.
(6) 통계수요를 조사통계만으로 충족시키는 것은 어렵다. 행정자료를 이용하여 통계를 작성하는 시스템을 구축해야 한다.
(7) 국제조사의 필요성이 높아 가고 있다.

참고문헌

[1] 조사연구(2000), 한국조사 연구학회
[2] 조사연구의 방법론적 쟁점(2000), 한국조사연구학회
[3] 표본설계(2001), 신민웅, 이상은, 교우사
[4] Babbie.E.R.(1973). Survey research methods.Belmont Calif : Wadsworth.
[5] Gadnathan(2001) Telesurvey Methodologies for household surveys-A review and some thoughts for the future. Survey Methodology.Vol.27. No.1
[6] Kalton.G.(2000). Development in survey research in the past 25 years. Survey Methodology. Vol.26. No.1
[7] Rohr.S.L.(1999). Sampling : Design and Analysis.Duxbury Press.