연구논문

조사 데이터 분석용 소프트웨어 패키지

Software packages for survey data analysis

성 내 경*
Sung, Nae Kyung

복잡한 확률표본설계를 기초로 수집된 데이터를 적법하게 분석하려면 반드시 조사설계를 고려하여 통계적 추론을 전개해야한다. 만일 설계를 무시하고 분석을 하면 각종 추정량의 분산이 과소추정되며, 이 결과 제1종 오류의 확률이 매우 높아진다. 본고에서는 조사 데이터를 전문적으로 분석하는 소프트웨어 패키지들을 소개하고, 특히 SUDAAN 7.5 판과 SAS 8 판의 분석 능력들에 대한 정보를 요약한다.

In order to make statistically valid inferences from survey data based on complex probability sample designs, survey researchers must incorporate the sample design in the data analysis. If this is not the case, the variance estimates of survey statistics derived under the usual simple random sampling assumptions from an infinite population generally underestimate the true variance, which results in a high Type I error level. In this article we introduce new software packages dedicated to analyze complex survey data. In particular, we summarize analysis capabilities on SUDAAN Version 7.5 and SAS Version 8.

Ⅰ. 개요

여론조사 및 시장조사 등 대규모 표본조사를 효율적으로 수행하려면 조사 예산, 조사 일정 등 주어진 상황을 고려하여 바라는 정확도를 유지할 수 있는 적절한 조사설계가 필수불가결하다. 그런데 주지하다시피 표

* 이화여자대학교 통계학과(nksung@mm.ewha.ac.kr)

본조사는 반드시 표본오차를 수반하며 또 표본오차는 관심있는 표본통계량의 분산으로 수량화할 수 있다. 따라서 표본조사설계의 목적은 주어진 상황에서 관심 통계량의 분산을 최소화하려는 것이라 하겠다. 그리고 이같은 조사설계로부터 비롯된 표본 데이터를 이용하여 도출한 분산 추정값을 기초로 오차의 범위를 정하고 통계적 추론의 유의성을 검증한다. 결국 조사 데이터에 대한 타당한 분석을 하려면 관심있는 표본통계량에 대한 정확한 분산 추정이 선결되어야 한다.

분산 추정은 근본적으로 조사설계 방식에 의존한다. 즉, 어떤 방식으로 조사를 진행하여 데이터를 얻었는가에 따라 분산 추정 공식이 달라지는 것이다. 따라서 조사설계가 다르면 분산 추정 방식 또한 다르다. 그런데 특히 복잡한 조사설계로부터 비롯된 데이터에서 관심있는 표본통계량에 대한 정확한 분산값을 추정하기란 매우 어렵다. 왜냐하면 이런 경우에는 분산 공식 자체를 수리적으로 유도하기조차 쉽지 않기 때문이다.

일반적으로 조사설계의 기본 요소는 임의표집, 층화, 집락화, 비추정 등인 바 대규모 조사에서는 이들 요소들이 단계별로 결합되어 최종 설계가 도출된다. 이렇게 단순한 조사 개념들을 혼합하여 구축한 조사를 복잡한 조사(complex survey)라 한다.

다시 말해서, 표본조사란 확률표본설계에 기초하여 표적 모집단으로부터 표본관측들을 수집하는 과정인 바, 정확성의 향상 및 비용 감소를 목적으로 층화, 집락화 등의 기법들이 사용되는데 이들 때문에 분석에 복잡성이 가해지는 것이다. 층화는 표본분산을 감소시키지만, 집락의 사용, 또는 집락의 부등확률선택은 표본분산을 증가시킨다. 또한 무응답을 보정하기 위하여 고려하는 표집비중의 조정, 사후층화로 인한 가중값의 결정 등은 문제를 더욱 복잡하게 한다. 이러한 복잡한 조사설계에서 추정량에 대한 정확한 표본분산의 계산은 아쉽게도 언제나 가능하지 않다.

복잡한 조사 데이터를 분석할 때 반드시 고려해야하는 몇 가지 공통적인 특징으로 다음과 같은 사항들을 열거할 수 있다.

· 관측들의 부등확률선택을 보정하기 위한 가중값의 사용
· 관측들의 집락화 및 집락내 관측단위들 간에 존재하는 상관관계

· 표집단위들의 층화
· 다단계 표집
· 무응답 처리 및 사후 보정을 위한 가중값
· 비선형 통계량에 대한 분산 추정 공식

그런데 문제는 조사 데이터 분석을 위하여 현재 국내에서 연구자들이 보편적으로 사용하는 SAS 6.12 판, SPSS 10 판, MINITAB 13 판 등과 같은 표준 통계 소프트웨어 패키지들은 근본적으로 조사설계가 함축하고 있는 표본조사 데이터의 특성을 고려하지 않은 채 데이터 분석을 진행하기 때문에, 이러한 통계 소프트웨어 패키지들을 주로 사용하는 조사 연구자들이 잘못된 결론을 얻게될 확률이 매우 높아지는 위험성이 상존하는데 있다.

다시 말해서 거의 모든 통계적 방법론들에서 데이터에 대한 기본 가정은 관측들이 서로 독립이고 동일한 분포를 따르며 각 관측들이 선택될 확률이 같다는 것이다. 즉, 무한 모집단에서 복원추출된 단순임의표본을 데이터로 간주하는 것이다. 그리고 표준 통계 소프트웨어 패키지들 역시 어떤 데이터가 입력되더라도 이러한 이상적인 상황을 가정하고 분석을 수행하는 것이다.

그러나, 조사 데이터에서 이러한 이상적인 상황은 전혀 성립하지 않는다. 심지어 가장 간단한 단순임의표집에서조차 임의표본이란 유한 모집단에서 비복원추출된 데이터를 의미하며 각 관측의 선택확률 또한 같지 않다.

그렇기 때문에 조사 데이터에서 관심 모수에 대한 비편향 추정을 하려면 표집단위에 대한 적절한 가중값, 즉, 표집비중(sampling weight)을 결정해야한다. 표집비중이란 표집단위에 대한 선택 확률의 역수로서 정의하는 바, 이는 곧 표집단위가 대표하는 모집단의 원소수를 의미한다.

그런데, 만일 적절한 조사설계를 채용하여 모든 표집단위들에 대한 표집비중이 동일한 자기가중표본(self-weighting sample)을 얻은 경우라면 가중값이 모두 동일하므로 이런 경우에는 조사 데이터를 SAS나 SPSS의 표준 분석 절차에 직접 적용하여 비가중 분석(unweighted analysis)을 진행

하여도 평균, 분위수 등과 같은 몇 가지 관심 모수들에 대한 점추정값은 정확히 얻을 수 있다.

그러나 다단계의 층화 및 집락을 활용하고 무응답 보정 및 사후 층화를 고려하는 복잡한 조사로부터 비롯된 가중표본에서 모수에 대한 비편향 추정을 하려면 분석용 가중값들을 먼저 결정해야 하는데, 이러한 가중값들은 표집비중을 비롯하여 층화, 집락화 등 조사설계의 성격에 의존한다. 따라서 이같은 경우에 표준 통계 소프트웨어 패키지를 사용하여 비가중 분석을 하면 당연히 편향된 추정값을 얻게 된다. 이를 방지하려면 적절한 가중값들을 입력하여 가중 분석을 수행해야 하며, 다행히 표준 통계 소프트웨어 패키지들에서는 가중 분석을 지원하기 때문에 복잡한 조사라 하더라도 관심 모수에 대한 비편향 점추정값을 얻는데는 어려움이 없다.

그러나 일반적인 통계 소프트웨어 패키지들로 이러한 가중분석을 수행하는 경우에 분산 추정에는 심각한 오류가 발생하며 이 때문에 SUDAAN과 같은 조사 데이터 분석을 전문으로 하는 통계 소프트웨어가 각광을 받고 있다.

복잡한 조사설계를 기초로 수집한 데이터에 대하여 단순임의표집의 가정을 기초로 유도된 관심 통계량에 대한 분산 추정 공식을 그대로 적용하면 대체로 모분산이 과소 추정된다. 이 결과 신뢰구간의 폭은 비정상적으로 작아지고 또 애초에 설정한 유의수준 이상으로 귀무가설을 기각하는 문제가 생긴다. 이런 문제는 특히 집락표집의 경우에 심각하다(성내경, 1999, 171, 261-264).

따라서 복잡한 조사 데이터를 분석한 조사 보고서를 읽을 때 분석 결과의 타당성을 확인하려면 무엇보다 단순히 원시 자료를 통계 소프트웨어 패키지에 그대로 입력해서 분석한 다음 보고서를 작성했는지 여부를 검토해보아야한다. 그리고 만일 이것이 사실이라면 보고서에 나타난 통계적 추론 결과는 믿기 어렵다. 왜냐하면 이런 결과는 설계를 고려하지 않은 통계적 유의성이기 때문이다. 참고로, 복잡한 조사 데이터 분석시 표준 통계 소프트웨어 패키지의 분석 결과와 조사 데이터 전문 분석 소프트웨어들의 분석 결과들 간의 구체적인 비교 연구에 대해서는 Cohen

(1997), Brogan (1997)을 참조하라.

그런데 복잡한 조사 데이터 분석용 소프트웨어의 필요성이 인지된지는 그리 오래지 않으며, 최근에 와서야 비로소 SUDAAN과 같은 주목할 만한 소프트웨어들이 출현하고 있다. 또한 앞으로 몇 년 내로 대부분의 표준 통계 소프트웨어 패키지들에도 조사 데이터 분석용 절차들이 포함될 것으로 전망된다. 참고로 아직 국내에는 배포되지 않았지만 SAS 8 판에는 이미 본격적인 조사 분석 절차가 삽입되었다.

본 소고에서는 현재 배포되고 있는 조사 데이터 분석용 소프트웨어들을 정리하고 그들의 분산 추정에 관련된 분석 기능들을 비교하여 조사 연구자들에게 유용한 정보를 제공하려한다. 특히 표준 통계 소프트웨어의 대명사인 SAS와 조사 데이터 분석용 소프트웨어의 대표자인 SUDAAN에 대하여 중점적으로 언급한다.

Ⅱ. 표본분산의 근사 추정

복잡한 조사 데이터를 분석할 때 관심 모수에 대한 비편향 추정량을 얻으려면 가중분석을 해야하며, 이러한 비편향 점추정은 적절한 가중값만 산정할 수 있다면 일반 통계 소프트웨어 패키지로도 어렵지 않게 정확한 추정값을 얻을 수 있다. 그러나, 이렇게 얻은 비편향 추정량에 대한 분산을 추정하기란 그리 간단치 않다. 다단계의 집락화, 층화를 하는 경우 정확한 분산 추정 공식을 유도할 수 있다고 하더라도 공식 자체가 매우 복잡하고 여기에 사후층화, 무응답 보정이 가해지면 정확한 추정 공식조차 만들기 어렵다. 특히 비추정, 가중평균 등으로 나타나는 비선형 통계량에 대한 정확한 분산 추정은 거의 대부분의 경우 전혀 가능하지 않다.

따라서 복잡한 조사에서는 점추정량에 대한 근사 분산 추정이 필수적이며, 분산의 근사 추정 방법은 선형화 기법과 복제 기법의 두 가지로 대별한다.

1. 선형화 기법(linearization technique)

선형화 기법(Woodruff, 1971; Binder, 1983), 또는 테일러 급수(Taylor series) 선형화 기법이란 모수들의 평활 비선형 함수로 주어진 관심 모수에 대한 비편향 추정량의 분산을 추정할 때 테일러 급수 전개를 이용하여 비선형 함수를 선형으로 근사시킨 후 분산 추정을 하는 방법이다.

예를 들어 $t_1, t_2, \cdots, t_k$를 k 개의 모합이라 하자. $\theta = h(t_1, t_2, \cdots, t_k)$ 이 관심 모수 다. 여기서 h는 t_i들에 대한 평활한 비선형 함수(smooth nonlinear function)다. $\hat{t}_i$ 을 t_i에 대한 비편향 추정량이라 하자. 이 경우 θ 에 대한 추정량은 $\hat{\theta} = h(\hat{t}_1, \hat{t}_2, \cdots, \hat{t}_k)$ 으로 잡을 수 있다. $\hat{\theta}$ 을 θ 근방에서 테일러 급수 전개한 결과는 다음과 같다.

$$\hat{\theta} \approx \theta + \sum_{i=1}^{k} (\hat{t}_i - t_i) \frac{\partial h(t_1, \cdots, t_k)}{\partial t_i}$$

따라서 비록 θ 가 t_i들에 대한 비선형 함수라도 테일러 급수 전개를 시행하면 분산 계산이 그리 어렵지 않은 선형 함수로 바뀐다. 물론 이 경우 2차 이상의 항은 무시한다. 이와 같은 개념 하에 분산을 근사 추정하는 기법을 선형화 기법이라 한다.

선형화 기법은 일반적인 표집설계에 모두 적용 가능한 장점이 있지만 주어진 h 함수마다 일일이 편도함수값들을 산출해야하고, 또 단계별 가중값 적용이 복잡하면 계산이 매우 어려운 단점이 있다.

조사 데이터 분석 소프트웨어들에서 기본적으로 지원하는 분산의 근사추정법이 바로 선형화 기법이다. 관심 모수의 형태에 따라 선형화된 추정량에 대한 개별적인 분산 공식이 필요하나, 자주 사용하는 평균, 비율, 비, 회귀계수 등에 대한 공식들은 대개 소프트웨어에 내장되어있다.

2. 복제기법(replication technique)

복제기법은 수집된 데이터로부터 반복된 부표집(repeated subsampling)을 하여 얻은 복제(replicate) 표본별로 독자적인 모수의 추정값을 구한 후

이들의 표본분산으로 추정량의 표본분산을 구하는 방법이다. 이때 각 복제 표본은 원래의 조사설계를 거울같이 반영하는 축소판이어야한다.

다음과 같은 방법들이 대표적인 복제기법들이다.

- 균형반복복제(BRR: balanced repeated replication), 또는 균형절반표집(balanced half-sampling)
- 잭나이프(jackknife)
- 자육법(自育法: bootstrap)
- 임의그룹법(random group method)

원래 임의그룹법(Mahalanobis, 1946)은 하나의 조사설계를 독립적으로 r (〉1) 번 반복 시행하여 각 반복마다 관심 모수를 추정한 다음, 이러한 추정값들의 표본분산으로 θ 의 분산을 추정하는 기법이다. 그러나, 조사설계를 반복시행하기란 현실적으로 어려우므로, 대신 단 한 번만 조사하여 얻은 표본을 조사설계를 그대로 반영하며 중첩되지 않는 r 개의 그룹으로 임의 분할한 다음, 각 임의그룹마다 독립적으로 추정량을 계산하고, 또 이렇게 얻은 r 개의 추정값들의 표본분산으로 θ 의 분산을 추정한다. 매우 간단한 원리를 사용하기 때문에 분산 계산이 쉬운 반면 분할된 각 임의그룹의 구조가 기본 설계를 거울같이 반영하는 축소판이어야 하므로 이를 만족시키려면 임의그룹들의 수가 작아지는 문제가 나타나며 이 경우 분산 추정이 부정확한 단점이 있다.

균형반복복제(MaCarthy, 1966; 1969) 기법은 합, 또는 분위수의 함수로 주어진 추정량의 분산 계산시 테일러 급수 선형화 기법과 거의 유사한 결과를 준다. 잭나이프와 자육법에 비하여 계산량이 적은 장점이 있다. 그러나, 균형반복복제는 기본적으로 균형된 층화설계에서만 적용 가능하며 층크기가 작으면 분산이 과대 추정되는 경향이 있다.

잭나이프(Tukey, 1958; Lipsitz 외, 1994) 기법은 중첩을 허용하는 임의그룹법이라 할 수 있다. 또 균형반복복제기법처럼 적용이 가능한 설계에 제한이 있지도 않다. 모합의 함수로 주어진 관심 모수에 대한 추정량의 분산 추정시 일치성과 같은 좋은 성질을 보유한다. 그러나 자육법은 복원추출 설계에만 적용할 수 있고, 분위수 관련 추정시 성능이 좋지 않으

며 아직까지 이론적으로 잭나이프 기법의 성능에 대하여는 연구가 그리 많지 않다.

자육법(Shao and Tu, 1995)은 표본을 모집단으로 간주하여 재표집하는 대표적인 방법으로 거의 모든 형태의 모수 추정 및 추정량의 분산 추정에 사용할 수 있다. 그러나 계산량이 매우 많고 복잡한 표본설계에 적용한 자육법의 성능에 대해서는 이론적 연구가 별로 없다.

이상의 복제기법들 중 가장 사용빈도가 높은 방법은 균형반복복제와 잭나이프 기법이다. 복제기법을 활용하면 테일러 급수 전개 선형화와는 달리 분산 공식의 구체적인 유도는 불필요하다. 참고로 SUDAAN 7.5 판에서는 선형화, 균형반복복제, 잭나이프 기법들을 지원한다.

Ⅲ. 소프트웨어 패키지

2000 년 6월 현재 복잡한 조사 데이터에서 관심 모수에 대한 추정량의 분산 추정을 지원하는 전문 소프트웨어 패키지들의 목록과 개략적인 성능 비교 등은 미국통계협회(ASA: www.amstat.org/sections/) 내 조사연구방법분과(Section on Survey Research Methods: www.stat.ncsu.edu/info/srms/srms.html)에서 제공하는 '조사분석 소프트웨어 요약' 보고서(www.fas.harvard.edu/~stats/survey-soft/survey-soft.html)를 참조하라. 그런데 여기에 요약된 패키지들에 대한 사항은 가장 최신의 정보는 아니다. 따라서 여기에 보고된 소프트웨어 패키지들을 비롯하여 최근에 개발된 소프트웨어 패키지들의 이름과 지원 운영체제, 가용한 분산 추정 방법들을 〈표 1〉에 정리하였다.

〈표 1〉에서 운영체제는 PC를 기준으로 Windows 9*x*/NT/2000 시스템과 DOS 시스템의 둘로 구분하였다. 특히 IVEware와 GES는 SAS 기반의 마크로(macro) 프로그램으로 컴퓨터에 SAS 시스템이 설치되어 있어야만 작동한다. SUDAAN은 독립적으로 운영할 수도 있지만, SAS 시스템의 부속 절차로서 운영이 가능하다. 또한 모든 패키지들이 SAS 데이터의 직접적인 입력을 지원한다.

〈표 1〉 조사 데이터 분석 소프트웨어 패키지 및 가용한 분산 추정법

소프트웨어	운영체제	요구 시스템	분산추정법				
			선형화	BRR	잭나이프	임의그룹	자육법
SUDAAN	Windows	(SAS 6.12, 8)	√	√	√		
SAS	Windows		√				
VPLX	Windows			√	√	√	
WesVarPC	Windows			√	√		√
Stata	Windows		√		√		√
IVEware	Windows	SAS 6.12	√		√		
GES	Windows	SAS 6.12	√		√		
CSPro	Windows		√				
PC CARP	DOS		√				
Epi Info	DOS		√				
CLUSTERS	DOS		√				

각 소프트웨어 패키지의 인터넷 사이트 또는 접촉 주소는 다음과 같다.

- SUDAAN 7.5: www.rti.org/patents/sudaan/sudaan.html
- SAS 8.0: www.sas.com
- VPLX 1998.09: www.census.gov/sdms/www/vwelcome.html
- WesVarPC 2.12: www.westat.com/wesvar/wesvar.html#download
- Stata 6.0: www.stata.com
- IVEware: www.isr.umich.edu/src/smp/ive/
- CSPro: www.census.gov/ipc/www/cspro/
- GES 4.0: www.statcan.ca/english/IPS/Data/10H0035LHB.htm
- PC Carp: www.statlab.iastate.edu/survey/index6.html
- Epi Info 6.04: www.cdc.gov/epo/epi/epiinfo.htm
- CLUSTERS: vjverma@essex.uk

이상에 제시한 프로그램들에 대한 비교 연구로서 대표적인 것들로는 Lepkowski와 Bowles (1996)의 CENVAR (지금은 CSPro), CLUSTERS, Epi Info, PC Carp, Stata, SUDAAN, VPLX, WesVarPC 간 비교 연구, Cohen (1997)의 Stata, SUDAAN, WesVarPC 간 비교 연구, Carlson (1998)의 SUDAAN, PC Carp, Stata, WesVarPC, VPLX 간 비교 연구 등을 들 수 있다.

이같은 비교 연구 결과들을 종합하면 성능이 가장 뛰어난 것은 SUDAAN이라 할 수 있다. 그런데 최근 전세계적으로 사용자 층이 가장 두터운 통계 소프트웨어 패키지인 SAS 시스템 8 판에 조사 데이터를 전문적으로 분석하는 절차들이 추가되었다.

1. SUDAAN

SUDAAN은 원래 집락설계부터 나오는 상관자료(correlated data)를 전문적으로 분석하는 소프트웨어로 개발되었다(Bieler and Williams, 1997). 그리고 이를 더 확장하여 유한모집단에서 추출된 서로 독립이 아니며 동일한 분포를 따르지 않는 관측들에 대한 체계적 분석을 지원하는 최초의 통계 패키지로 발돋움하였다. SUDAAN에서는 관심 모수들에 대한 비편향 추정량을 계산하는 기술자료분석을 비롯하여, 선형회귀, 로지스틱 회귀, 다항 로지스틱 회귀, 비례위험모형 등의 제반 분석을 지원한다. 덧붙여 SUDAN에서는 제한적인 분포 가정 없이 복잡한 조사설계로부터 도출된 통계량들(평균, 합, 비율, 승산비, 회귀계수 등)에 대한 일치(consistent) 분산 추정값을 계산하며, 또 SUDAN은 분산의 근사 추정 방법으로 테일러 급수 전개를 이용한 선형화 기법, 균형반복복제 기법, 잭나이프 기법들을 동시에 지원한 최초의 소프트웨어 패키지며, 회귀모형 분석시에는 일반화 추정 방정식(GEE: generalized estimating equations)을 이용한 모수추정이 가능하다.

SUDAAN은 거의 모든 조사설계 데이터의 분석을 제공한다. 즉, 층화표집, 집락표집을 포함하는 다단계표집, 크기비례확률표집(pps 표집) 등을 비롯한 제반 등확률 및 부등확률표집, 복원 및 비복원 표집 모두를

인식하고 분석할 수 있는 능력이 있다. 또한 모집단을 분할하여 부분별로 조사설계가 달라도 분석이 가능하다. 이와 같은 다양한 조사설계 방식의 지원은 다른 분석 패키지들이 아직 미치지 못하는 SUDAAN만의 강점이다.

SUDAAN에서 지원하는 분석 절차들에는 다음과 같은 것들이 있다.

- DESCRIPT - 평균, 합, 비율, 백분율, 기하평균, 분위수 등 모수를 추정하고 분산을 계산한다.
- CROSSTAB - 교차표에 대하여 빈도분포, 백분율 분포, 승산비, 상대위험 등을 추정하고 이들에 대한 신뢰구간을 계산한다. 카이제곱 독립성 검증 및 코크란-만텔-헨첼(Cochran-Mantel-Haenszel)의 층화분석을 수행한다.
- RATIO - 비추정을 하고 표준오차를 계산한다.
- REGRESS - 연속 반응변수에 대한 선형 회귀모형을 적합하고 모수들에 대한 검증을 수행한다.
- LOGISTIC - 이진자료(binary data)에 대하여 로지스틱 회귀모형을 적합하고 모수들에 대한 검증을 수행한다. 승산비의 추정이 가능하며 각 모수에 대한 신뢰구간을 계산한다.
- MULTILOG - 명목 및 순서 범주형자료에 대하여 다항 로지스틱 회귀모형을 적합하고 유의성 검증을 시행한다. 승산비의 추정이 가능하며 각 모수에 대한 신뢰구간을 계산한다. 효율적인 모수 추정을 위하여 GEE 방식의 접근이 가능하다.
- SURVIVAL - 실패시간 자료에 대하여 비례위험모형(또는 콕스(Cox) 회귀모형)을 적합한다. 위험비의 추정이 가능하며 각 모수에 대한 신뢰구간을 계산한다.

SUDAAN에서 제공하는 이상의 절차들은 독립적으로 사용할 수도 있으나, SAS 시스템에 부속시켜 SAS에서 호출 가능(SAS-callable)한 단위 절차처럼 사용할 수 있다. 실제로 SUDAAN의 분석 절차들의 사용법은 SAS 시스템의 절차들의 사용법과 매우 유사하기 때문에 일반 SAS 사용

자들은 SUDAAN을 간편히 사용할 수 있는 이점이 있다. SUDAAN 7.5.3 판은 SAS 6.12 판과 호환되며, SUDAAN 7.5.4 판은 SAS 8 판과 호환된다.

2. SAS

SAS 시스템에서는 7 판부터 표본조사 데이터를 분석하는 절차들이 삽입되었다. 그런데 SAS 7 판은 원래 개발자용 버전이었기 때문에 일반 사용자들은 1999년 말에 출시된 SAS 8 판에서부터 이 절차들을 사용할 수 있게 되었다. 이 절차들의 구체적인 용례에 대한 보고서는 An과 Watts (1998)를 참고하라.

SAS 시스템에는 다음과 같이 확률표본을 추출하는 절차가 존재한다.

- SURVEYSELECT - 표집방법, 표집률 등 설계 모수들을 지정하면 입력된 표집틀로부터 조사설계를 반영하는 임의확률표본을 선택한다. 지원하는 조사설계 방식은 층화표집, 집락표집을 포함하는 다단계표집, 크기비례확률표집 등 제반 등확률 및 부등확률표집, 복원 및 비복원 표집 등으로 SUDAAN에 거의 필적한다. 그리고 출력되는 표집단위마다 선택확률과 표집비중이 자동적으로 계산된다.

그리고 SAS 시스템에서 지원하는 조사 데이터 분석 절차들은 다음과 같다.

- SURVEYMEANS - 모평균, 모합의 추정값을 계산하고 이들에 대한 분산과 신뢰구간을 추정한다. 또한 조사설계를 반영하는 각종 기술통계량들을 산출한다. 분산 추정방식으로 테일러 급수 전개를 이용한 선형화 기법을 이용한다.
- SURVEYREG - 조사 데이터에 대하여 회귀모형을 적합하고, 회귀계수 벡터 및 회귀계수 벡터의 분산 · 공분산행렬을 추정한다. 모수에 대한 유의성 검증을 시행한다.

SAS 시스템에서는 SUDAAN에 비하여 아직 제한적인 분석 절차만을 제공하는 것이 단점이지만, 현재 가용한 절차만으로도 사용자들이 바라는 대부분의 분석 욕구를 충족시킬 수 있다. 또한 확률표본의 선택을 전담하는 SURVEYSELECT 절차의 지원은 당분간 다른 패키지들에서 찾아보기 어려운 강점이 되고있다.

Ⅳ. 결어

일반적인 표준 통계 소프트웨어 패키지들에서는 데이터 분석시 표본이 무한 모집단에서 단순임의표집되었고, 관측들은 정규분포나 이항분포를 따른다는 분포가정을 하며, 선형 통계량에 대한 추정 위주로 분석 루틴을 구성하고 있다. 그러나, 일반적으로 조사 데이터들은 유한 모집단에서 복잡한 확률표집설계 하에 수집되며 특정한 분포 가정이 어려운 경우가 대부분이며 또한 관심 모수 자체가 비선형 함수로 주어지는 경우가 많다. 따라서 표본조사 데이터를 제대로 분석하려면 분석시 반드시 조사설계를 반영할 수 있도록 고안된 전문적인 소프트웨어 패키지의 활용이 필수적이다. 특히 관심 통계량에 대한 분산 추정시 표준 통계 소프트웨어 패키지들의 결과는 대부분의 경우 과소 추정이 초래되므로 신뢰하기 어렵다. 이러한 관점에서 현재 국내의 조사 연구자들이 가장 유용하게 활용할 수 있는 소프트웨어 패키지는 SUDAAN과 SAS라고 하겠다. 왜냐하면 SAS 시스템은 전세계적으로 통계 분석의 표준 소프트웨어로서 사용자가 가장 많고, 또 SUDAAN은 SAS와 매우 흡사한 명령어 체계를 갖고 있기 때문이다.

참고문헌

성내경. 1999.《표본조사방법론》. 자유아카데미.
An, T. and D. L. Watts. 1998. "New SAS procedures for analysis of sample

survey data." in *Proceedings of the Twenty-Third Annual SAS Users Group International Conference*, Cary, North Carolina; SAS Institute, Inc.

Bieler, G. S. and R. L. Williams. 1997. "Analyzing survey data using SUDAAN Release 7.5." *Joint Statistical Meetings, Continuing education workshop*. Research Triangle Institute.

Binder, D. A. 1983. "On the variances of asymptotically normal estimators from complex surveys." *International Statistical Review*. 51: 279-292.

Brogan, D. J. 1997. "Pitfalls of using standard statistical software packages for sample survey data." in *Encylopedia of Biostatistics*, edited by P. Armitage and Colton, T., Wiley.

Carlson, B. L. 1998. "Software for statisticl analysis of sample survey data." in *Encyclopedia of Biostatistics*, edited by P. Armitage and Colton, T., Wiley.

Cohen, S. B. 1997. "An evaluation of alternative PC-based software packages developed for the analysis of complex survey data." *American Statistician* 51(3): 285-292.

Lepkowski, J. and Bowles J. 1996. "Sampling error software for personal computers." *Survey Statistician*. 35: 10-17.

Lipsitz, S, K. Dear and L. Zhao. 1994. "Jackknife estimators of variance for parameter estimates from estimating equations with applications to clustered survival data." *Biometrics*. 50: 842-846.

Mahalanobis, P. C. 1946. "Recent experiments in statistical sampling in the Indian Statistical Institute." *Journal of the Royal Statistical Society*. 109: 325-370.

McCarthy, P. J. 1966. "Replication: An approach to the analysis of data from complex surveys." *Vital and Health Statistics*. 2(14). National Center for Health Statistics, Public Health Srvice, Washington, D.C.

McCarthy, P. J. 1969. "Pseudoreplication: Half-samples." *Review of the International Statistical Institute*. 37: 239-264.

Shao, J. and D. Tu. 1995. *The jackknife and bootstrap.* Springer-Verlag.

Tukey, J. W. 1958. "Bias and confidence in not quite large samples." *Annals of Mathematical Statistics.* 29: 614.

Woodruff, R. S. 1971. "A simple method for approximating the variance of a complicated estimate." *Journal of the American Statistical Association.* 66: 411-414.