응용통계연구 제 18권 3호, 2005년, pp. 673-687

표본조사에서 설계기반추론과 모형기반추론

김규성[1)]

요 약

표본조사에서 이용하는 모수 추론 방법인 설계기반추론과 모형기반추론을 고찰하였다. 설계기반추론은 확률화 원리에 기초를 두고 있는 반면 모형기반추론은 가정한 모형에서 조건부 원리와 우도 원리에 바탕을 두고 있다. 두 추론은 서로 다른 이론적 근거를 사용하기 때문에 이론적 기초에 관한 논쟁이 오래전부터 있어 왔으며 지금도 진행되고 있다. 이 논문에서는 두 추론 사이에 진행되었던 논쟁의 초점을 살펴보았고 몇 가지 관점에서 두 추론의 장단점을 비교하였다.

주요용어: 대표성 원리, 우도 원리, 유의선정, 조건부 원리, 확률화 원리.

1. 서론

표본조사의 역사는 오래되었지만 근대적인 의미를 같는 표본조사는 백여 년 전에 발표된 Kiaer(1897)에서 그 출발점을 찾아볼 수 있다(Brewer, 1999). 지금은 일반적으로 받아들이는 개념인 대표성 원리(representative principle)는 예전에는 논쟁의 대상이었다. 논쟁의 촛점은 부분으로 전체를 대표할 수 있는가 하는 문제였다. 급증하는 표본조사의 필요성 때문에 표본조사를 하기는 하지만 그 타당성에 대한 의문은 끊임없이 제기되었다. 대표성 원리를 구현하는 방법으로 유의선정(purposive selection)과 확률표집(random sampling)이 제안되었는데 1926년 ISI 보고서는 두 방법이 타당하다고 인정하였으나 어느 방법이 더 나은지는 결론짓지 못하였다. 두 방법의 장단점에 대한 논쟁은 1934년 Neyman에 의해서 일단락된다. Neyman은 유의선정을 이용한 Gini & Galvani의 방법은 모평균 추론에서 불일치성(inconsistency)을 보이는 반면 확률표집을 이용하면 모평균에 대한 최량선형불편추정량을 구할 수 있고 더 나아가 Neyman 배정으로 잘 알려진 표본배정을 하면 추론의 효율을 극대화할 수 있다는 것을 수리적으로 보였다. 따라서 그는 확률표집이 유의선정보다 더 타당하다고 주장하였다. 확률표집을 일반화한 확률화 원리(randomization principle)에 근거를 둔 추론을 설계기반추론(design-based inference)이라고 하는데 1934년 이후에는 설계기반추론이 표본이론의 주류로 자리를 잡게 된다.

Neyman 이후 확률화 원리에 기초를 둔 표집이론들이 많이 개발되었고 총조사와 여러 표본조사에 활용되었다. 설계기반추론에 대한 이론적인 약점은 Godambe(1955, 1966)에 의해 발견되었다. 대표적인 약점 중의 하나는 설계기반추론이 제공하는 선형추정량은 최소분산비편향성을 갖지 못한다는 점이고(Godambe, 1955) 다른 하나는 표집설계에서 만들어진

1) (130-743) 서울시 동대문구 전농동 90, 서울시립대학교 자연과학대학 통계학과, 부교수
E-mail: kskim@uos.ac.kr

우도함수는 표본 안에서 상수값을 갖는다는 점이다(Godambe, 1966). 전자는 설계기반추론의 효율성에 의문을 던지는 것이고 후자는 일반통계학에서 추론의 근거로 많이 사용하는 우도함수가 표본이론에서는 무의미하다는 것을 말하는 것이기 때문에 일반 통계학자에게는 받아들이기 힘든 결과였다. 설계기반추론의 이론 체계를 근본적으로 무너뜨릴 수도 있는 Godambe의 중대한 발견은 설계기반추론의 체계 안에서 설계기반추론의 미비점을 지적한 것이었기 때문에 결과적으로 설계기반추론의 약점을 보완하여 이론 체계를 더욱 탄탄하게 하는 역할을 하였다. 설계기반추론에 대한 다른 본질적인 의문은 Royall(1970, 1971)이 제기하였다. Royall은 표본이론에서도 통계적 추론은 우도원리(likelihood principle)와 조건부 원리(conditionality principle)에 바탕을 두어야 하고 이를 위해서는 유한모집단에서 조사변수를 설명할 수 있는 적절한 모형을 가정하고 가정한 모형에서 모수에 대한 추론을 하는 것이 타당하다고 하였다. 따라서 확률화 원리에 기초를 둔 설계기반추론은 타당하지 않다고 주장하였다. 모형에 기초를 둔 추론을 모형기반추론(model-based inference)이라고 하는데 모형기반추론은 표본이론의 주류로 자리를 잡은 설계기반추론을 근본적으로 부정하는 것이었다. Royall의 주장에 대하여 Hansen et. al. (1983)은 설계기반추론을 옹호하는 적극적인 방어를 하였는데 두 추론에 대해 벌어지는 논쟁은 지금까지도 계속되고 있다.

두 추론에 관한 비교는 여러 학자에 의하여 이루어졌다(예를 들면 Brewer, K.R.W., 1999; Smith, T.M.F., 1976, 1984). 이 논문에서는 두 추론 사이에 벌어진 논쟁의 초점을 구체적으로 살펴보고 두 추론의 타당성과 장단점을 비교하고자 한다.

2. 두 추론의 특징

두 추론이 공통으로 하는 것은 유한모집단 프레임(frame)이다. 크기 N인 유한모집단을 $U = \{1, ..., k, ..., N\}$라고 하고 y_k와 x_k를 k번째 단위에 대응하는 조사변수와 보조변수라고 하자. 조사변수와 보조변수는 일차원 값일 수도 있고 다차원 값일 수도 있다. 모집단의 유한성(finiteness)과 단위의 식별성(identifiability) 그리고 보조변수 값을 주어진 상수로 취급하는 것은 두 추론에서 동일하다. 그러나 조사변수의 경우 설계기반추론에서는 미지의 상수로 간주하고 모형기반추론에서는 확률변수로 간주한다. 상황에 따라서 이러한 가정은 타당할 수도 있고 그렇지 않을 수도 있는데 두 추론에서는 논리 전개를 위하여 조사변수의 상수화 혹은 변수화를 일관되게 전제하고 있다. 두 추론에서 알고자 하는 기본적인 값은 모집단 평균과 모집단 총계이다. 설계기반추론에서는 모평균이나 모집단은 미지의 상수이므로 상수를 추정하는 문제가 되고 모형기반추론에서는 모평균이나 모총계가 확률변수이므로 확률변수를 적합(fitness)하는 문제가 된다. 모평균 혹은 모총계 추정(혹은 적합)을 위하여 유한모집단에서 선정한 표본을 s라고 하자. 이때 표본 s는 표본설계에 의하여 선정된다.

2.1. 설계기반추론

유한모집단에서 조사변수 값이 상수이므로 모수 추정에서 나타나는 불확실성 (uncertainty)은 모두 표본설계에서 발생한다. 따라서 추정량의 기댓값과 분산은 주어진 표본설계에서 계산한다.

2.1.1. 설계비편향성

설계기반추론에서는 조사변수와는 무관하게 표본설계에서 추론을 진행하기 때문에 추정량은 조사변수와 무관하게 만들어야 한다. 일반통계학에서 비편향성은 추정량을 평가하는 기준 중에 선호하는 하나의 기준에 불과하지만 설계기반추론에서는 설계비편향성은 반드시 지켜야 하는 거의 절대적인 기준이다. 그 이유는 설계기반비편향 추정량은 조사변수와 무관하게 만들어지기 때문이다. 반면 설계편향 추정량은 표본설계는 물론 조사변수 값에 연관되어 있기 때문에 조사변수 값에 관계없이 표본설계 만으로 추론을 진행하는 설계기반추론에서 설계편향 추정량은 받아들이기 힘든 추정량이다. 따라서 설계기반추론에서 추정량은 설계비편향성을 갖거나 적어도 점근적 설계비편향성(asymptotic design-unbiasedness) 혹은 점근적 설계일치성(asymptotic design-consistency)을 가져야 한다. 설계비편향성을 만족시키기 위한 표본설계는 포함확률이 모두 양수인 설계이다. 만일 포함확률이 0인 조사단위가 있다면 설계비편향성을 만족시키지 못하는 경우가 발생한다. 따라서 설계기반추론에서 포함확률이 양수인 조건은

$$\pi_i > 0,\ i = 1, ..., N \tag{2.1}$$

거의 대부분의 표본설계에 부여되는 조건이 된다.

2.1.2. 타당성

양수의 포함확률을 갖는 표본설계는 앞에서 언급한 대표성 원리를 만족시킬 수 있지만 이러한 표본설계에 근거한 추론이 타당한가 하는 것은 또 다른 문제이다. 설계기반추론의 타당성에 대한 논리적인 언급은 Neyman(1934)으로 거슬러 올라가 찾아볼 수 있다. Neyman은 설계기반추론의 타당성을 신뢰구간으로 설명하였다. Neyman은 모든 가능한 표본에 대하여 다음과 같은 신뢰구간을 만들고

$$\hat{\theta}_s(y) \le \theta(y) \le \hat{\theta}_s(y) \tag{2.2}$$

조사변수 y에 관계없이 모수 $\theta(y)$가 신뢰구간에 들어올 확률이 $1-\epsilon$ 이상이 아니면 추론은 타당하다고 하였다. 여기에서 $1-\epsilon$은 미리 정해진 상수이다. 이와 같은 추론의 타당성에 대한 설명은 설계기반추론의 이론적 기초가 되었다. 따라서 설계기반추론이 타당함을 보이기 위해서는 신뢰구간을 이용하여 신뢰구간의 확률이 미리 정한 명목확률과 근사함을 보여야 한다. 그러나 대개의 경우 고려해야 하는 모든 표본의 수가 너무 많기 때문에 모든 표본에 대하여 신뢰구간을 구하는 것은 거의 불가능하다. 대신 Hansen, et al.(1983)은 주어진 표본설계에서 표본을 반복 추출하고 극한 분포에서 추론의 정확성을 조사변수에 무관하게 언급할 수 있으면 추론은 타당하다고 하였다. 따라서 Hansen, et. al.(1983)에 의하면 설계기반추론이 타당한가를 보이기 위해서는 (i) 표본설계가 양수의 포함확률을 갖고 있고 (ii) 극한분포에서 조사변수에 무관하게 추론할 수 있다는 것을 보여야 한다.

이와 같은 설명은 추론의 타당성에 대한 근거를 극한분포에서 찾고있기 때문에 설계기반추론이 타당하기 위해서는 표본의 크기가 커야함을 알 수 있다. 또한 추정량의 극한 분포를 구해야 함을 요구하고 있다. 통상적으로 모평균이나 모총계 추정량에 대해서는 정규

근사를 이용하는 것이 보통이지만 회귀모집단에서는 모평균에 대한 정규근사가 적용되지 않으므로 정규근사로 모든 경우를 설명하는 것은 무리다. 따라서 개별 추정량에 대해서 설계기반추론이 타당함을 보이려면 추정량의 극한 분포를 구해야 한다.

2.1.3. 효율성과 최적성

설계비편향성 추정량 중에서 최선의 추정량은 최소분산을 갖는 추정량이다. 그러나 Godambe(1955)가 보인대로 그러한 최선의 추정량은 존재하지 않는다. 따라서 설계기반추론에서 설계와 추정량의 최적성은 정의되지 않는다. 최적추정량이 존재하지 않는 이유는 표본수에 비하여 모수의 수가 더 많아서 표본에 속한 단위에 가중치를 유일하게 부여하지 못하기 때문이다. 이러한 최적성의 문제를 풀기 위한 노력이 두 방향으로 전개되었다. 첫째는 유한모집단에서 조사변수 y_k $(k = 1, ..., N)$를 미지의 상수로 고정하고 추정량과 표집설계의 효율성과 최적성을 정의하는 것이다. 두 번째는 조사변수 y_k를 확률변수 Y_k의 조사값으로 간주하고 확률변수 Y_k에 모형을 가정하여 최적성과 효율성을 정의하는 것이다.

범위를 N차원 공간의 한점 $y = (y_1, ..., y_N)'$으로 좁히면 효율성과 최적성은 다음과 같이 정의할 수 있다. 표집설계 p_1과 p_2를 고려하고 대응하는 추정량 $\hat{\theta}_1(y)$과 $\hat{\theta}_2(y)$을 고려했을 때 다음 조건을 만족하면 한 점 y에서 표집전략 $(p_1, \hat{\theta}_1)$이 표집전략 $(p_2, \hat{\theta}_2)$ 보다 효율적이라고 한다(Thompson, 1997, p.21).

$$E_{p_1}(\hat{\theta}_1(y) - \theta(y))^2 \leq E_{p_2}(\hat{\theta}_2(y) - \theta(y))^2 \tag{2.3}$$

위의 효율성에 대한 정의는 정해진 한 점에서만 이루어지므로 매우 제한적이긴 하지만 설계기반추론에서는 모수 $\theta(y)$가 조사변수 값의 함수이므로 모수를 추정하는 과정에서 위의 정의는 유효하다고 할 수 있다. 최적 표집설계와 추정량은 위의 식 (2.3)에서 평균제곱오차를 최소로 하는 것이고 가능한 최소값은 0이다. 그런데 Basu(1971)가 보였듯이 평균제곱오차를 0으로 하는 표집설계와 추정량의 존재는 어렵지 않게 보일 수 있다. 예를 들어 Horvitz-Thompson 추정량에서 관측값과 포함확률이 비례하면 평균제곱오차가 0이 된다. 그렇지만 현실적으로 조사 전에는 관측값을 알 수 없으므로 이에 비례하는 포함확률을 표본 추출에 앞서 지정하는 것은 불가능하다. 즉 활용 불가능한 표집설계인 셈이다. 따라서 한 점이 주어졌을 때 효율성에 대한 정의는 유효하지만 최적성에 대한 정의는 현실성이 없다.

조사변수에 모형 기대값과 모형 분산을 가정하면 설계분산에 대한 모형 기대값을 최소로 하는 추정량을 구할 수 있다(Godambe, 1955; Godambe & Joshi, 1965). 그리고 기대분산의 하한도 얻을 수 있다. 그러나 앞에서 설계기반추론의 타당성은 조사변수와 무관하게 보여야 한다고 하였으므로 조사변수에 모형을 가정하면 사실상 설계기반추론의 범위를 벗어나는 결과가 된다. 이와 관련한 논의는 뒤의 3.1.1 절에서 다시 하기로 한다.

2.2. 모형기반추론

모형기반추론에서는 주어진 표본과 조사변수에 부여된 모형에 근거를 두고 추론을 한다. 이 추론에서는 표본은 주어진 것으로 간주되고 일단 표본이 선정되면 표본추출확률은

추론에서 배제된다. 대신 조사변수에 부여된 모형이 추론의 근거가 된다. 따라서 모형기반 추론의 타당성은 가정된 모형의 타당성으로 설명된다.

2.2.1. 타당성

앞에서 언급한 모형기반추론의 타당성을 구체적으로 살펴보자. 표본설계에 의하여 데이터(data; $\mathcal{D}$)를 얻었다고 하자. 데이터는 표본과 표본에 속하는 단위에 대응하는 조사변수로 구성된다.

$$\mathcal{D} = \{s, (y_i, i \in s)\} \tag{2.4}$$

이때 조사변수는 미지의 상수로 실수값을 갖는다고 하자. 즉 $y = (y_1, ..., y_N)' \in R^N$이다. 하나의 조사변수 값이 주어졌을 때 위의 데이터가 얻어질 확률은

$$\Pr(\mathcal{D}|y) = \begin{cases} p(s), & y \in R^N(y_i; i \in s) \\ 0, & o.w. \end{cases}$$

이다. 여기서 $R^N(y_i : i \in s) = \{y^* \in R^N : y_i^* = y_i, i \in s\}$. 따라서 데이터 $\mathcal{D}$에 대한 조사변수 y의 우도함수는 다음과 같이 표현된다.

$$L(y|\mathcal{D}) = p(s)I(y \in R^N(y_i : i \in s)) \tag{2.5}$$

여기서 I는 지시함수이다. 결국 데이터가 주어졌을 때 조사변수에 대한 우도함수는 상수값이 되어 표본에 속한 조사변수들에게 아무런 정보도 제공하지 못한다. 조사변수 y를 사전정보로 간주하고 y가 사전분포 ξ를 따른다고 가정하면 모총계 T_y의 베이즈 추정량은 사전분포 ξ와 우도함수 $L(y|\mathcal{D})$에 의해서 만들어지는 사후분포에서 사후평균이 된다.

$$\hat{t}_\xi = E_\xi\{T_y|\mathcal{D}\} \tag{2.6}$$

그런데 우도함수가 상수 함수이므로 결과적으로 사후분포는 조사변수에 대한 사전분포와 동일하게 되어 베이즈 추정량 $\hat{t}_{xi}$은 사전분포에서 비편향 추정량이 된다(Godambe, 1982). 조사변수에 대하여 하나의 사전분포를 가정하면 베이즈 추정량을 유일하게 구할 수 있으므로 가정된 모형에 대하여 최소분산 비편향추정량을 유일하게 구할 수 있다.

2.2.2. 최량성

조사변수에 부여된 사전분포의 집합을 $\mathcal{C} = \{\xi\}$라고 하자. 그러면 집합 $\mathcal{C}$에 대하여 모형 비편향성을 만족시키는 모형기반 추정량 중에서 모형분산이 가장 작은 추정량을 최량의 추정량이라고 할 수 있을 것이다. 설계기반추론에서는 설계비편향성을 만족시키는 추정량 중에서 설계분산을 최소로 하는 추정량은 존재하지 않았다. 모형기반추론의 경우는 어떤가? 설계기반추론과 마찬가지로 최소분산 비편향추정량은 존재하지 않는다. 예를 들어 집합 $\mathcal{C}$를 R^N에서 정의된 분포들의 완비 집합이라고 하자. 그러면

$$E_\xi\{\hat{t}_y - T_y\} = 0,\ \xi \in \mathcal{C} \Rightarrow\ [\hat{t}_y - T_y = 0] \tag{2.7}$$

이 되어 전수조사인 경우를 제외하고는 완비 집합 C에서 최적의 모형비편향 추정량은 존재하지 않는다(Godambe, 1982).

결국 일반적인 모형들의 집합에서는 최량의 모형기반 추정량을 찾을 수 없기 때문에 최적의 추정량을 찾기 위해서는 조사변수에 부여되는 모형의 범위를 좁혀야 한다. 통상적으로 연구자들이 주로 이용하는 모형은 회귀모형이다. 표본을 선정한 후에 조사변수를 표본에 속하는 변수와 속하지 않는 변수를 정렬한 후 회귀모형을 표현하면 다음과 같다.

$$\begin{pmatrix} Y_s \\ Y_r \end{pmatrix} = \begin{pmatrix} X_s \\ X_r \end{pmatrix}\beta + \begin{pmatrix} \epsilon_s \\ \epsilon_r \end{pmatrix},\ E\begin{pmatrix} \epsilon_s \\ \epsilon_r \end{pmatrix} = 0,\ Var\begin{pmatrix} \epsilon_s \\ \epsilon_r \end{pmatrix} = \begin{pmatrix} V_s & V_{sr} \\ V_{rs} & V_r \end{pmatrix} \tag{2.8}$$

여기에서 s는 표본을 나타내고 $r = U - s$ 그리고 나머지는 모두 대응하는 벡터 표현이다. 이러한 회귀모형에서 모총계 T_y의 선형추정량은 다음과 같고

$$\hat{t}_y = 1_s' y_s + w_{sr}' y_s \tag{2.9}$$

모형비편향성을 만족시키기 위해서는 다음의 조건을 충족해야 한다.

$$w_{sr}' X_s = 1_r' X_r \tag{2.10}$$

여기에서 $1 = (1, ..., 1)'$, $w_{sr} = (w_1, ..., w_n)'$ 그리고 n은 표본수이다. 결과적으로 모총계에 대한 모형비편향 선형추정량은 다음과 같이 얻을 수 있다.

$$\hat{t}_{BLU} = 1_s' y_s + 1_r'[X_r \hat{\beta}_s + V_{rs} V_s^{-1}(y_s - X_s \hat{\beta}_s)] \tag{2.11}$$

여기에서

$$\hat{\beta}_s = (X_s' V_s^{-1} X_s)^{-1} X_s V_s^{-1} y_s \tag{2.12}$$

이다(Bolfarine & Zacks, 1992, p.24).

2.3. 타당성 비교

두 추론을 비교하기 위하여 다음의 기호를 도입하자. $Y = (Y_1, ..., Y_N)$는 조사변수, $X = (X_1, ..., X_N)$는 보조변수 그리고 $I = (I_1, ..., I_N)$는 표본추출변수라고 하자. 여기서 보조변수 X는 알려진 값이라고 가정한다. 표본설계는 표본추출변수 I에 관련한 것으로 설계기반 추론에서 사용하는 확률화는 보조변수가 주어진 상태에서 표본추출변수의 분포로 나타난다.

$$I|X \sim p(\cdot) \tag{2.13}$$

반면 모형기반추론은 보조변수가 주어진 상태에서 조사변수의 확률화에 기초를 둔다.

$$Y|X, I \sim \xi \tag{2.14}$$

앞에서 설명한 바와 같이 개별 문제에서 각 추론의 타당성을 주장하기 위해서는 추론의 전제조건에 대한 검증이 필요하다. 설계기반추론에서는 표본설계가 양수의 포함확률을 갖는

지를 보이고, 추정량의 극한분포를 구하여 이 분포가 조사변수에 무관함을 보여야 한다. 반면 모형기반추론에서는 조사변수에 부여된 모형이 적절함을 보여야한다. 이와 같은 전제조건이 검증되면 두 추론이 주장하는 대로 두 추론은 모두 타당하다고 할 수 있다.

그러나 본질적으로 어느 추론이 더 옳은가? 이 질문에 답하기 위해서는 두 추론의 장단점을 여러 측면에서 비교해볼 필요가 있다.

3. 두 추론의 비교

두 추론은 논리를 전개하는 근거가 다르기 때문에 한 가지 기준으로 두 추론을 직접 비교하기는 어렵다. 대신 통상적으로 추론을 비교하는데 사용하는 기준인 효율성(efficiency)과 강건성(robustness)을 검토하면 간접적이긴 하지만 두 추론을 비교할 있다. 또한 표본조사에서는 실용성도 중요한 기준이므로 실용성을 두 추론의 비교 기준으로 활용할 수 있다. 이 절에서는 두 추론의 효율성, 강건성 그리고 실용성을 비교, 검토하기로 한다.

3.1. 효율성

두 추론이 제공하는 추정량 중 어느 것이 더 효율적인가? 이 질문에 대답하기 위해서는 다음 사항을 먼저 살펴보아야 한다. 설계기반추론에서는 주로 설계비편향추정량에 관심이 있기 때문에 모든 조사단위의 포함확률이 양수인 표본설계가 관심 대상이다. 반면에 모형기반추론에서는 모형비편향추정량이 주요 관심사이고 표본은 모형 분산을 최소로 하는 것이 선호된다. 이와 같이 두 추론이 추구하는 표본설계와 추정량의 조합이 서로 다르다. 따라서 두 추론을 직접 비교하는 것은 무의미해 보일 수도 있다. 그러나 관심의 대상이 되는 표본설계와 추정량의 조합 중에서 두 추론이 제공하는 최적의 표본설계와 추정량을 찾아서 두 추론을 부분적으로 비교해 볼 수 있다. 이를 위해서는 표본설계와 추정량의 효율을 수량화해야 하고 모형의 불확실성과 표본설계의 불확실성을 동시에 표현하는 측도가 있어야 한다. 두 추론에 동시에 이용할 수 있는 측도로 총분산 AV를 고려하자(Isaki and Fuller, 1982).

$$AV(\hat{t}_y - T_y) = E_{\xi p}[(\hat{t}_y - T_y)^2] - [E_{\xi p}(\hat{t}_y - T_y)]^2 \tag{3.1}$$

만일 표본설계가 무정보설계(non-informative design)이면 모형기대값과 설계기대값을 바꾸어 계산할 수 있으므로 식 (3.1)의 측도는 다음과 같이 표현할 수 있다.

$$AV(\hat{t}_y - T_y) = E_\xi\{E_p[(\hat{t}_y - T_y)^2|y]\} - \{E_\xi E_p[(\hat{t}_y - T_y)|y]\}^2 \tag{3.2}$$

또는

$$AV(\hat{t}_y - T_y) = E_p\{E_\xi[(\hat{t}_y - T_y)^2|s]\} - \{E_p E_\xi[(\hat{t}_y - T_y)|s]\}^2 \tag{3.3}$$

따라서 만일 추정량 $\hat{t}_y$가 설계비편향이면 식 (3.2)로부터 추정량 $\hat{t}_y$의 총분산은

$$AV(\hat{t}_y - T_y) = E_\xi\{E_p[(\hat{t}_y - T_y)^2|y]\} \tag{3.4}$$

가 되고 만일 추정량 $\hat{t}_y$가 모형비편향이면 식 (3.3)으로부터 추정량 $\hat{t}_y$의 총분산은

$$AV(\hat{t}_y - T_y) = E_p\{E_\xi[(\hat{t}_y - T_y)^2|s]\} \tag{3.5}$$

가 된다. 따라서 식 (3.4)의 총분산과 식 (3.5)의 총분산을 비교하는 것은 의미가 있다.

3.1.1. 최적 설계기반추정량과 표본설계

모든 조사단위의 포함확률이 양수인 표본설계를 고려하고 추정량 $\hat{t}_y$를 설계비편향이라고 하자. 그리고 조사변수 y_i는 서로 독립이고 유한인 평균 μ_i와 분산 σ_i^2을 갖는다고 하자. 그러면 추정량 $\hat{t}_y$에 대한 기대분산의 하한은 다음과 같다(Godambe & Joshi, 1965, Theorem 6.1).

$$E_\xi V_p(\hat{t}_y) \geq \sum_{i=1}^{N}(\frac{1}{\pi_i} - 1)\sigma_i^2 \tag{3.6}$$

여기에서 첨자 ξ는 모형을 뜻하고 p는 표집설계를 의미한다. 만일 포함확률이 모형 표준편차 σ_i에 비례하면 추정량의 기대분산의 하한은 더 내려갈 수 있다. 이제 다음과 같은 질문을 던져보자. 기대분산의 하한에 도달하는 설계비편향 추정량은 무엇인가?

기대분산을 구하기 위해서는 모형에 대한 가정이 필요하다. 보조정보가 주어졌을 때 보편적으로 사용하는 회귀모형을 고려하자.

$$Y_i = \mu_i + \epsilon_i,\ \mu_i = \beta_{0i} + \beta_1 x_{1i} + ... + \beta_J x_{Ji},\ \epsilon_i \sim (0, \sigma_i^2) \tag{3.7}$$

여기에서 Y_i는 서로 독립이다. 설계기반추론에서 사용하는 일반적인 추정량으로 동질선형추정량과 비동질선형추정량을 들 수 있다.

우선 다음과 같은 동질선형추정량을 고려하자.

$$\hat{t}_{y1} = \sum_s w_{si} y_i \tag{3.8}$$

동질선형추정량의 기대분산의 하한을 구하고 그 하한에 도달하는 추정량을 찾으면 아래와 같다.

$$\begin{aligned} E_\xi V_p(\hat{t}_{y1}) &\geq \sum_U (\frac{1}{\pi_i} - 1)\sigma_i^2 + E_p\{[\sum_s (\frac{\mu_k}{\pi_k})]^2\} - (\sum_U \mu_k)^2 \\ &= E_\xi V_p(\sum_s \frac{y_k}{\pi_k}) \end{aligned} \tag{3.9}$$

따라서 표본 크기가 고정된 n이고 포함확률이 조사변수의 모형기대값에 비례하는 표본설계에서 Horvitz-Thompson 추정량(HTE)

$$\hat{t}_{y1}^* = \sum_s \frac{y_k}{\pi_k},\ \pi_k = \frac{n\mu_k}{\sum_U \mu_j} \tag{3.10}$$

은 동질선형추정량 중에서 기대분산을 최소로 한다(Godambe, 1955).

동질선형추정량을 통하여 구한 최적 추정량 $\hat{t}^*_{y1}$에 관하여 세 가지 사항을 살펴보자. 첫째 식 (3.10)에서 나타나는 포함확률 π_k는 미지의 회귀계수로 계산되기 때문에 비모형 $y = \beta x + \epsilon$ 을 제외하고는 이용이 불가능한 값이다. 따라서 일반적인 회귀모형 (3.7)에서 추정량 $\hat{t}^*_{y1}$을 활용하기 위해서는 회귀계수를 추정하고 뒤이어 포함확률 계산에 대입하는 작업을 해야한다. 물론 이 경우에 기대분산은 하한값에 도달하지 못한다. 그러나 만일 추정한 회귀계수가 미지의 회귀계수에 설계일치성을 가지고 있고 표본수가 충분히 크면 추정량 $\hat{t}^*_{y1}$의 기대분산은 하한값에 도달하리라고 기대할 수 있다. 둘째 어떤 추정량이 최적 추정량이 되기 위해서는 포함확률이 모형의 기대값에 비례하여야 한다. HTE를 생각하면 직관적으로 이해할 수 있다. 그러나 뒤에 나올 비동질선형추정량을 염두에 두면 이 결과는 일반화하기 어렵다. 왜냐하면 최적 비동질선형추정량에서 포함확률은 모형의 표준편차에 비례하기 때문이다. 이와 같은 성질은 설계비편향추정량이 모형의 위치 모수에 대하여 불변성(invariance)을 가지고 있지 않음을 뜻한다. 마지막으로 최적의 설계가 되기 위해서는 고정표본크기를 가져야 한다. 변동표본크기를 가지는 설계는 표본크기의 불확실성으로 인하여 기대분산의 크기가 증가하기 때문이다.

이번에는 비동질선형추정량을 고려하자.

$$\hat{t}_{y2} = w_{s0} + \sum_s w_{si} y_i \tag{3.11}$$

설계비편향성을 만족하는 비동질선형추정량 중에서 최소 기대분산을 갖는 추정량은 모형 (3.7)에서 다음과 같은 추정량이다.

$$\hat{t}^*_{y2} = \sum_s \frac{y_i - \mu_i}{\pi_i} + \sum_U \mu_i \tag{3.12}$$

그리고 최적 설계는 고정표본크기 n과 포함확률 $\pi_i = \sigma_i / \sum_U \sigma_j$을 갖는 설계이다. 추정량 $\hat{t}^*_{y2}$의 기대분산은 기대분산의 하한에 도달한다. 식 (3.12)에 주어진 추정량은 일반화편차추정량(generalized difference estimator, GDE, Cassel et al. 1976)이라고 알려져 있다. 추정량 $\hat{t}^*_{y1}$과 마찬가지로 GDE $\hat{t}^*_{y2}$에 대해서도 세 가지 사항을 언급할 필요가 있다. 첫째 GDE에 포함된 μ_i는 미지의 회귀계수를 포함하고 있기 때문에 직접 사용할 수 있는 추정량이 아니다. 따라서 미지의 회귀계수 대신 회귀계수 추정량을 그 자리에 대입한 추정량을 활용하여야 하는데 이렇게 만들어진 추정량이 잘 알려진 일반화회귀추정량(generalized regression estimator, GREG, Cassel et al. 1976)이다. 당연히 GREG는 기대분산의 하한에 도달하지 못한다. 대신 회귀계수 추정량이 회귀계수에 설계일치성을 가지고 있고 표본의 크기가 충분히 크면 GREG의 기대분산은 기대분산의 하한에 도달하리라고 기대된다. 둘째 포함확률은 모형 표준편차에 비례한다. 앞의 $\hat{t}^*_{y1}$과는 상반되는 결과이다. 그 이유는 비동질선형추정량에서 상수항 w_{s0}이 회귀모형의 평균항 μ을 추정하고 나머지항이 잔차항을 추정하기 때문인데 잔차는 모형분산에 의존하므로 잔차항을 추정하기 위해서는 포함확률을 모형표준편차에 비례하도록 하는 것이 더 효율적인 것이다. 비동질선형추정량을 사용했을 때 생기는 부수적인 장점은 기대분산의 하한을 더욱 낮출 수 있는 것이다. 식 (3.6)에 제시된 기대

분산의 하한은 포함확률이 모형표준편차에 비례할 때 다음과 같이 낮아진다.

$$\sum_{i=1}^{N}(\frac{1}{\pi_i}-1)\sigma_i^2 \geq \frac{1}{n}(\sum_U \sigma_i)^2 - \sum_U \sigma_i^2 \tag{3.13}$$

셋째 동질선형추정량에서와 마찬가지로 최적 설계가 되기 위해서는 고정표본크기를 가지는 설계이어야 한다.

이제까지 식 (3.6)에 제시된 기대분산의 하한에 이르는 추정량의 형태를 동질선형추정량과 비동질선형추정량의 관점에서 고찰하였다. 그러나 식 (3.6)에 나타난 기대분산의 하한은 더 큰 추정량의 집합에도 적용된다. 비편향동질선형추정량과 비편향비동질선형추정량 뿐 아니라 설계비편향추정량의 집합에서도 기대분산의 하한은 식 (3.6)의 그것과 같다(Godambe & Joshi, 1965).

3.1.2. 최적 모형기반추정량과 표본설계

일반적인 회귀모형에서의 최량선형추정량은 앞의 2.2절에서 언급한 바 있다. 이번 절에서는 설계기반추정량과 비교하기 위하여 모형 (3.7)에서 조사변수가 서로 독립인 경우만 고려하자. 그러면 식 (2.9)에 나타난 선형추정량 중에서 모형비편향성을 만족시키면서 모형분산을 최소로 하는 추정량은 아래의 추정량이다.

$$\hat{t}_{y3}^{*} = \sum_{i\in s} y_i + \sum_{i\in r} x_i'\hat{\beta}_s \tag{3.14}$$

여기에서 $x_i = (1, x_{1i}, ..., x_{li})'$ 그리고 $\hat{\beta}_s = (\sum_s x_i x_i'/\sigma_i^2)^{-1}\sum_s x_i' y_i/\sigma_i^2$ 이다. 추정량 $\hat{t}_{y3}^{*}$의 모형분산은 다음과 같다.

$$E_\xi[(\hat{t}_{y3}^{*} - T_y)^2|s] = \sum_r \sigma_i^2 + t_{xr}'\{\sum_s \frac{x_i x_i'}{\sigma_i^2}\}^{-1} t_{xr} \tag{3.15}$$

여기에서 $t_{xr} = (\sum_r x_{1i}, \sum_r x_{2i}, ..., \sum_r x_{Ji})'$이다. 따라서 추정량 $\hat{t}_{y3}^{*}$의 기대분산을 최소로 하는 표본설계는

$$E_p E_\xi[(\hat{t}_{y3} - T_y)^2|s]\} = \sum_{all\ s} p(s)[\sum_r \sigma_i^2 + t_{xr}'\{\sum_s \frac{x_i x_i'}{\sigma_i^2}\}^{-1} t_{xr}] \tag{3.16}$$

을 최소로 하는 설계이다. 여기에서 r은 표본에 포함되지 않은 단위들의 집합을 뜻한다. 그런데 식 (3.15)에는 모형분산과 보조정보가 동시에 나타나기 때문에 일반적인 경우에 식 (3.16)을 최소화 하는 표본설계를 찾기는 쉽지 않다. 그러나 식 (3.16)에서 우변의 괄호 안에 있는 항들은 표본 s가 주어지면 계산되는 값들이므로 비록 최적의 유의표본 s^*를 찾기는 어렵지만 식 (3.16)의 최소값은 하나의 유의표본 s^*에서 발생한다는 사실은 알 수 있다. 유의표본에서 기대분산의 하한이 얻어진다는 사실은 설계기반추론과 비교하면 매우 다른 결과이다.

간단한 예로써 비모형과 단순회귀모형을 살펴보자. 다음의 비 모형에서

$$y = \beta x + \epsilon, \quad \epsilon \sim (0, \sigma_i^2) \tag{3.17}$$

모총계에 대한 모형비편향추정량은 다음과 같고

$$\hat{t}^*_{y3} = \sum_s y_i + \hat{\beta}_s \sum_r x_i,\ \hat{\beta}_s = \sum_s \{\frac{x_i' y_i}{\sigma_i^2}\} / \sum_s \frac{x_i^2}{\sigma_i^2} \tag{3.18}$$

추정량 $\hat{t}^*_{y3}$에 대한 모형분산은 다음과 같이 얻어진다.

$$E_\xi(\hat{t}^*_{y3} - T_y)^2 = \sum_r \sigma_i^2 + \sum_r x_i (\sum_s \frac{x_i^2}{\sigma_i^2})^{-1} \sum_r x_i \tag{3.19}$$

따라서 식 (3.19)의 모형분산을 최소로 하는 표본설계를 얻기 위해서는 모형분산 σ_i^2과 보조변수 x_i의 관계를 살펴보아야 한다. 있다. 모형분산을 $\sigma_i^2 = v(x_i)\sigma^2$라고 하고 $v(x_i)$는 x의 함수로써 다음과 같은 성질이 있다고 하자.

(i) $v(x)$는 x에 대한 비감소(non-decreasing) 함수

(ii) $v(x)/x^2$는 x에 대한 비증가(non-increasing) 함수

그러면 위의 두 성질로부터 식 (3.19)를 최소화하는 표본설계는 모집단 조사단위 중에서 가장 큰 x값들을 갖는 조사단위들을 표본으로 취하는 설계임을 알 수 있다(Royall, 1970).

$$p^*(\cdot)\ :\ p^*(s^*) = 1,\ \max_s \sum_s x_i = \sum_{s^*} x_i \tag{3.20}$$

이번에는 단순회귀모형을 고려하자.

$$y = \alpha + \beta x + \epsilon,\quad \epsilon\ \sim\ (0, \sigma^2) \tag{3.21}$$

모총계에 대한 모형비편향추정량은 다음과 같다.

$$\hat{t}^*_{y3} = \sum_s y_i + (N-n)\hat{\alpha}_s + \hat{\beta}_s \sum_r x_i \tag{3.22}$$

여기에서

$$\hat{\beta}^*_s = \sum_s x_i y_i / \sum_s x_i^2,\ \alpha^*_s = \bar{y}_s - \hat{\beta}^*_s \bar{x}_s \tag{3.23}$$

이고 식 (3.22)에서 추정량 $\hat{t}_{y3}$의 기대분산을 최소로 하는 표본설계는 아래의 식을 최소로 하는 표본 s^*을 확률 1로 선정하는 표본이다(Royall, 1970).

$$(\frac{1}{N}\sum_{i=1}^N x_i - \frac{1}{n}\sum_s x_i)^2 / \sum_s (x_i - \frac{1}{n}\sum_s x_i)^2 \tag{3.24}$$

비교적 간단한 두 모형에서 유도된 결과 (3.20)과 (3.24)로부터 우리는 다음의 사항을 알 수 있다. 첫째 모형기반추론에서 추정량의 기대분산을 최소화하는 표본은 유의표본이다. 둘째 비모형과 회귀모형의 결과에서 보듯이 모형이 바뀌면 최적의 표본은 바뀔 수 있다. 첫번째 결과인 유의표본의 최적성은 이 글의 서론에서 언급한 대표성 원리에서 유의선정도 논리적으로 타당할 수 있다는 근거를 제공함으로써 Neyman 이후 잠잠했던 모형기반추론이 부활하는 계기가 된다(Royall, 1970). 그러나 두번째 결과에서 보듯이 모형에 따라 최적표본이 바뀌는 것은 가정된 모형이 틀리면 추론도 틀릴 수 있다는 것을 보여주는 것이다. 따라서 모형기반추론을 제대로 활용하기 위해서는 모형가정에 크게 영향을 받지 않는 추론의 강건성(robustness)에 대한 보완이 필요하다.

3.2. 강건성

강건성에 관한 두 추론의 입장 차이는 현격하다. 먼저 설계기반추론에서는 조사변수에 모형 가정을 하지 않을뿐 아니라 표본설계가 조사변수와 무관하고, 추정량의 극한분포에서 추론하기 때문에 추론이 본래부터 강건한 성질을 가지고 있다고 한다. 그러나 모형기반추론에서는 가정된 모형이 적절하지 않을 때 모형기반추정량은 모형의 변화에 민감하게 반응하고 따라서 강건성을 잃게 된다고 한다(Hansen, et. et., 1983).

가정된 모형이 옳지 않을 때 추론의 근거를 상실하게 된다는 사실은 Neyman 이후에 모형기반추론 옹호자들에게는 중대한 고민거리였다. 이러한 고민거리에 대한 부분적인 답이 Royall과 그의 공동연구자들에 의해서 제안되었다. 식 (3.7)과 같은 회귀모형을 가정하고 식 (3.14)와 같은 최량 추정량을 사용했을 때 모형의 일탈에 대한 대비를 어떻게 할 것인가? Royall이 제시하는 답은 이렇다(e.g., Royall, 1988). 회귀모형을 가정했을 때 모형이 가정에서 벗어날 부분은 평균 $\mu = \sum_j \beta_j x_j$과 분산성분 $\sigma_j^2 = v(x_j)\sigma^2$으로 나누어 볼 수 있다. 평균에 대한 가정이 틀리면 앞 절에서도 설명했듯이 최적 표본은 현저하게 다른 형태를 띤다. 따라서 이러한 문제를 해결하는 방안으로 균형표본(balanced sample)을 사용할 것을 Royall은 제안한다. 균형표본을 사용하면 가정된 모형의 변화에 추정량이 민감하게 반응하지 않게 된다. 다음으로 분산성분에 대한 가정에서는 강건한 분산성분 추정량을 사용할 것을 제안하고 있다(이에 대해서는 Royall의 일련의 논문을 참고).

균형표본에 대한 설계기반추론 옹호자들의 반론은 다음과 같다. 첫째 보조변수의 수가 많으면 구체적으로 균형표본을 구하기가 쉽지 않다. 둘째 균형표본을 사용하면 실질적으로 추정량의 형태가 표본평균 형태로 변해가고 또한 표본도 확률표본에서 제공하는 표본의 일종이 되어간다. 따라서 균형표본이 확률표본이 되는 경향을 나타나 확률화 원리를 암묵적으로 동의하는 결과가 된다. (층화추출에서 효율성과 강건성에 대한 비교는 Brewer(1999)를 참고).

3.3. 실용성

설계기반추론이 모형기반추론 보다 더 실용적이라는 사실은 대부분의 통계관련 연구자들이 인정하는 바이다. 또한 대부분의 조사관련 기관이 설계기반추론의 방법론을 채택하여 조사를 수행하고 있는 것도 보통의 현상이다. 앞에서 논의한 여러 쟁점들은 대부분 이론적인 측면의 쟁점인데, 이론적인 논쟁과는 별개로 현실에서는 설계기반추론이 실제 조사에 적용되어 조사의 저변을 확대해 가고 있었던 것이다. 따라서 이론적인 취약점이 있음에도 불구하고 설계기반추론이 제공하는 방법론은 현실 문제에 잘 적용되고 있다고 볼 수 있다.

그럼에도 불구하고 모형기반추론을 선호하는 학자들은 설계기반추론에 대하여 다음과 같은 질문을 계속 던지고 있다. 왜 조사 통계학자들은 일반 통계학자들과는 다른 관점을 가지고 있는가? (Smith, T.M.F., 1976). 이에 대한 이론적인 논의는 앞에서 설명한 바와 같고 관점을 바꾸어 실용적인 측면에서 설계기반추론 옹호자가 던지는 답은 이렇다. 표본조사와 일반 통계학, 특히 실험계획은 처한 환경이 근본적으로 다르다는 것이다. 먼저 실험계획에서는 관심변수의 수가 적은 반면 표본조사에서는 조사변수의 수가 일반적으로 많다. 그리고 조사 목적도 다양하다(multivariate and multi-purpose). 추론에서 모형의 필요성은 대부

분의 조사 통계학자들도 인정하지만 조사변수의 수가 많으면 모형을 설정하기가 현실적으로 어렵고 또한 모형의 강건성을 유지하기도 어렵다. 따라서 조사변수가 많으면 설계기반추론이 훨씬 용이한 추론 방법이 된다. 둘째 표본조사에서는 표본 크기가 큰 반면 일반 실험계획에서는 표본 크기가 작다. 설계기반추론에서는 설계비편향성을 갖는 추정량을 사용하기 때문에 표집분산만 조절하면 되고 표본의 크기가 크면 오차를 일정 수준 이하로 조절할 수 있다. 그러나 모형기반추론에서는 가정된 모형이 틀리면 편향이 발생하기 때문에 표본의 수가 증가한다 하더라도 오차를 일정 수준 이하로 조절하기는 쉽지 않다. 따라서 표본의 크기가 큰 조사에서는 설계기반추론이 더 적절하다. 셋째 표본조사에서 관심 모수는 주로 유한모집단 평균 혹은 유한모집단 총계와 같은 기술 통계량(descriptive statistic)이다. 반면 일반 통계학에서 추론의 대상인 모수는 무한 모집단 모평균 혹은 모분산 등 무한모집단 모수들이다. (See, Kalton's discussion in Smith, 1976.)

표본크기에 관하여 어느 정도의 표본 크기가 설계기반추론의 타당성을 보장해줄 수 있는가에 대해서는 아직도 논쟁의 여지가 있다. 표본의 크기가 크면 설계기반추론을 사용하는 것이 적절하고 표본의 크기가 작으면 모형기반추론을 사용하는 것이 적절하다고 할 때 다음과 같은 질문은 여전히 유효하다. 왜 표본크기에 따라 추론 방법을 바꾸어야 하는가? (Smith, 1976).

4. 결론

이론적인 관점에서 Hansen, et. al.(1983)은 추론은 표본설계에 전적으로 의존해야 한다고 주장하고 Royall은 모형에 기초를 두어야 한다고 주장한다. 두 추론은 근본적으로 다른 기반을 가지고 있지만 추구해야 하는 방향은 대체로 다음과 같다. 설계기반추론에서는 모형화에 대한 관심을 계속 가지고 있어야 하고 모형기반추론에서는 많은 현실적인 문제에 적합한 모형을 개발해야 한다. 대규모 조사에서는 조사변수의 수가 많고 표본수도 많아서 모형을 세우기가 쉽지 않기 때문에 설계기반추론이 보편적으로 사용되는 현상이 나타나는 반면 소지역 추정같이 표본수가 적은 분야나 무응답 처리나 측정오차 처리같이 표집과정이 명시적이지 않은 분야에서는 모형을 도입하여 문제를 풀려고 하는 시도가 활발하고 많은 연구 결과가 나와 있다. 문제에 따라 설계기반추론과 모형기반추론의 접목이 활발하게 이루어지고 있는 셈이다. 향 후 표본조사에 대한 지식이 더욱 누적되고 현실에 적용할 수 있는 확률 모형이 더욱 정교해지면 설계기반추론과 모형기반추론의 거리는 더욱 가까워질 것이다.

참고문헌

Bolfarine, H. and Zacks, S. (1992). *Prediction Theory for Finite Populations*, Springer-Verlag.

Basu, D. (1971). An essay on the logical foundations of survey sampling, part one, In : V.P. Godambe and D.A.Sprott (eds) *Foundation of Statistical Inference*, Holt, Rinehart and Winston, Toronto, 203-242.

Brewer, K.R.W. (1999). Design-based or prediction-based Inference? stratified random vs stratified balanced sampling, *International Statistical Review*, **67**, 35-47.

Cassel, C.M., Sarndal, C.E. and Wretman, J.H. (1976). Some results on generalized difference estimation and generalized regression estimation for finite population, *Biometrika*, **63**, 615-620.

Godambe, V.P. (1955). A unified theory of sampling from finite populations, *Journal of the Royal Statistical Society*, Ser. B **17**, 269-278.

Godambe, V.P. (1966). A new approach to sampling from finite populations I & II, *Journal of the Royal Statistical Society*, Ser. B **28**, 310-328.

Godambe, V.P. (1982). Estimation in survey sampling : robustness and optimality, *Journal of the American Statistical Association*, **77**, 393-406.

Godambe, V.P. and Joshi, V.M. (1965). Admissibility and Bayes estimation in sampling finite populations I, *Annals of Mathematical Statistics*, **36**, 1707-1722.

Hansen, M.H., Madow, W.G. and Tepping. B.J. (1983). An evaluation of model- dependent and probability-sampling inferences in sample surveys, *Journal of the American Statistical Association*, **78**, 776-807.

Iachan, R. (1984). Sampling strategies, robustness and efficiency : the state of art, *International Statistical Review*, **52**, 209-218.

Isaki, C.T. and Fuller, W.A. (1982). Survey design under the regression superpopulation model, *Journal of the American Statistical Association*, **77**, 89-96.

Neyman, J. (1934). On the different aspects of the representative method : the method of stratified sampling and the method of purposive selection, *Journal of the Royal Statistical Society*, Ser A **97**, 558-625.

Royall, R.M. (1970). On finite population sampling theory under certain linear regression models, *Biometrika*, **57**, 377-387.

Royall, R.M. (1971). Linear regression models in finite population sampling theory, In *Foundation of Statistical Inference*, Eds. V.P. Godambe & D.A. Sprott, 259-279. Toronto: Holt, Rinehert and Winston.

Royall, R.M. (1988). The prediction approach to sampling theory, *Handbook of Statistics*, **6**, *Sampling*, 399-413. North-Holland.

Royall, R.M. and Cumberland, W.G. (1981). An empirical study of the ratio estimator and estimators of its variance, *Journal of the American Statistical Association*, **76**, 66-88.

Smith, T.M.F. (1976). The foundation of survey sampling. *Journal of the Royal Statistical Society*, **139**, 183-204.

Smith, T.M.F. (1984). Present position and potential developments, *Journal of the Royal Statistical Society*, **84**, 208-221.

Thompsom, M.E. (1997). *Theory of Sample Surveys*, Chapman & Hall.

[2004년 7월 접수, 2005년 5월 채택]

Design-based and Model-based Inferences in Survey Sampling

Kyu-Seong Kim[1)]

ABSTRACT

We investigate both the design-based and model-based inferences, which are usual inferential methods in survey sampling. While the design-based inference is on the basis of randomization principle, the model-based inference is based on likelihood principle as well as conditionality principle. There have been some disputes between two inferences for a long time and those have not yet been determined. In this paper we reviewed some issues on two inferences and compared their advantages and disadvantages in some viewpoints.

Keywords: Conditionality principle; Likelihood principle; Purposive selection; Randomization principle; Representative principle.

1) Associate Professor, Department of Statistics, University of Seoul, 90 Jeonnong-Dong Dongdaemun-Gu Seoul 130-743, Korea.
E-mail: kskim@uos.ac.kr