集落標本資料에 대한 適合度檢定과 獨立性檢定

南宮 坪* · 崔秉洙** · 李柱祿***

要 約

一段集落標本抽出에서 Pearson X^2과 Wald統計量, 標本設計效果에 의한 修正統計量 그리고 縮小因子에 의한 修正統計量을 比較하였다. 適合度 檢定과 獨立性檢定의 경우 시뮬레이션에 의한 결과, Wald 統計量은 Pearson X^2統計量과는 유의한 차이를 나타냈으나 두 개의 修正統計量과는 큰 差異가 없게 나타났다.

1. 서 론

대규모의 표본조사는 층화표본추출이나 집락표본추출 등의 복합표본추출방법을 사용하고 있다. 이러한 표본자료로부터 적합도검정이나 분할표의 독립성검정을 Pearson 통계량 X^2으로 검정한다면 반응치들간에 상관관계가 존재하므로 분석상에 심각한 오차가 수반된다는 것이 많은 문헌에서 지적되고 있는 바 표본자료의 영향을 수정하기 위한 방법이 두 가지 측면으로 진행되어 왔다. 표본설계에 근거한 접근은 자료의 임의성을 표본설계에 두고 이러한 영향을 표본설계 효과에 의해 수정하려는 방법이며(Fellegi(1980), Holt, Scott, Ewings(1980), Rao, Scott(1981, 1984), Bedrick(1983), Landis외(1984), Koch 외(1975), Fay(1985) 남궁평(1986), 박한철(1987)), 모형에 근거한 접근방법은 표본자료의 확률분포로 부터 수정통계량을 제시하는 방법이다. (Altham(1986), Cohen(1976), Brier(1980), Fienberg(1979), Choi Jae Won(1987)).

본 연구는 표본추출이 일단집락표본추출일 때 모형에 근거한 접근방법을 이용하여 적합도와 독립성검정을 위한 시뮬레이션을 통해 Pearson 통계량, Wald 통계량, 표본설계에 의한 수정통계량 그리고 축소인자에 의한 수정통계량 중에서 가장 효율성이 있는 통계량을 구하고자 한다.

2. 모형과 비율추정

유한집단 U는 A 개의 배반적 집락 $U_1, \cdots, U_i, \cdots, U_A$로 구성되고 i 번째 집락 U_i는 B_i 개의 최종요소 $U_i = (U_{i1}, \cdots, U_{ij}, \cdots, U_{iBi})$로 이루어지며 $N = \sum_{i=1}^{A} B_i$는 모집단의 총수이다.

* 成均館大學校 統計學科 敎授, 서울 鍾路區 明倫洞 3街 53
** 漢城大學電算統計學科 敎授, 서울 城北區 三仙洞 2街 392의 2
*** 成均館大學校 統計學科 大學院

이때 다음과 같은 지시변수(indicator variable)를 사용할 수 있다.

$$X_{ijh} = \begin{cases} 1 & \text{요소 } U_{ij} \text{가 } h\text{번째 cell에 떨어질 경우} \\ 0 & \text{그렇지 않은 경우} \end{cases} \qquad (2.\ 1)$$

여기서 $i = 1, \cdots, A$; $j = 1, \cdots, B_i$; $h = 1, \cdots, r$ 이다.

관심의 모수는 $\underline{\pi} = (\pi_1, \cdots, \pi_r)'$ 이고

$\pi_h = N_h/N$, $N_h = \sum_i^A \sum_j^B X_{ijh}$ 이다.

또한, $\pi_h > 0$, $\sum_h^r \pi_h = 1$ 이고

$\underline{X} = (X_1, \cdots, X)'$ 은 모집단벡터이다.

그러므로 모형에 근거한 접근방법은 X_{ijh}가 요소 U_{ij}에 대해 확률변수라 가정하면 다음과 같은 수학적 모형을 설정할 수 있다.

$$E(X_{ijh}) = \pi_h$$

$$E(X_{ijh} \cdot X_{i'j'h'}) = \begin{cases} \pi_h \pi_{h'}, & i \neq i' \\ P_{hh'}, & i = i',\ j \neq j' \\ \pi_h, & i = i',\ j = j',\ h = h' \\ 0, & i = i',\ j = j',\ h \neq h' \end{cases} \qquad (2.\ 2)$$

여기서

$$P_{hh'} = \begin{cases} \theta\pi_h + (1-\theta)\pi_h^2, & h = h' \\ (1-\theta)\pi_h\pi_{h'}, & h \neq h' \end{cases}$$

여기서 θ 는 집락내 상관계수이고 개별집락의 크기에는 의존하지 않는다. $P_{hh'}$는 두 요소가 같은 집락으로부터 추출되었을 때 첫번째 식은 첫요소가 확률 π_h로 h cell에 떨어지고 두번째 요소가 $[\theta+(1-\theta)\pi_h]$로 h cell에 떨어지는 확률이다. 또한 두번째 식은 첫번째 요소가 확률 $\pi_{h'}$로 h cell에 떨어지고 두번째 요소가 확률 $(1-\theta)\pi_{h'}$?! h' cell에 떨어지는 확률이다. 일단집락추출법은 A개의 집락으로부터 a 개의 집락을 선택하고 i 번째 집락 ($i = 1, \cdots, a$) B_i 개의 요소로부터 b 개의 최종요소를 선택하는 방법이다. 이 경우에 추정목적으로 사용되는 x_{ijh} 는 다음과 같다.

$$x_{ijh} = \begin{cases} 1 & (i, j)\text{번째 요소가 } h\text{번째 cell에 떨어질 경우} \\ 0 & \text{그렇지 않은 경우} \end{cases} \qquad (2.\ 3)$$

$$i = 1, \cdots, a,\ j = 1, \cdots, b,\ h = 1, \cdots, r$$

관심의 모수

$\underline{\pi} = (\pi_1, \cdots, \pi_r)'$, $(\sum_h^r \pi_h = 1, \pi_h > 0)$

는 r다항모수이다. $\underline{\hat{\pi}}$의 불편추정량은

$\underline{\hat{\pi}} = (\underline{\hat{\pi}}_1, \cdots, \underline{\hat{\pi}}_r)'$이고

$\underline{\pi}_h$의 분산, 공분산은 다음과 같이 정의된다.

$$Var(\hat{\pi}_h) = E(\hat{\pi}_h^2) - (E(\hat{\pi}_h))^2, \ h = h' \qquad (2.4)$$
$$\mathrm{Cov}(\hat{\pi}_h \ \hat{\pi}_{h'}) = E(\hat{\pi}_h \hat{\pi}_{h'}) - E(\hat{\pi}_h) E(\hat{\pi}_{h'}), \quad h \neq h'$$

임의의 표본요소가 확률 π_h로 h 번째 cell에 떨어진다면

$$E(x_{ijh}) = \pi_h$$
$$E(x_{ijh} \cdot x_{i'j'h'}) = \begin{cases} \pi_{hh'}, & i \neq i' \\ P_{hh'}, & i = i', \ j \neq j' \\ 0, & i = i', \ j = j', \ h \neq h' \\ \pi_h, & i = i', \ j = j', \ h = h' \end{cases} \qquad (2.5)$$

을 가정 할 수 있다. 여기서

$$P_{hh'} = \begin{cases} \theta\pi_h + (1-\theta)\pi_h^2, & h = h' \\ (1-\theta)\,\pi_h \pi_{h'}, & h \neq h \end{cases}$$

이다.

그러므로 $\hat{\pi}_h$의 분산, 공분산은 식 (2.4), (2.5)로부터 다음과 같이 표현할 수 있다.

$$\mathrm{Var}(\hat{\pi}_h) = \frac{\pi_h(1-\pi_h)}{n} - \frac{n + \theta \sum_i^a b_i(b_i-1)}{n} \qquad (2.6)$$

$$\mathrm{Cov}(\hat{\pi}_h \hat{\pi}_{h'}) = \frac{-\pi_h \pi_{h'}}{n} - \frac{n + \theta \sum_i^a b_i(b_i-1)}{n}$$

여기서 $G = n + \theta \sum_i^a b_i(b_i-1)$

로 정의하면

$$\mathrm{Var}(\hat{\pi}_h) = \frac{G}{n} - \frac{\pi_h(1-\pi_h)}{n} \qquad (2.7)$$

$$\mathrm{Cov}(\hat{\pi}_h \hat{\pi}_{h'}) = \frac{G}{n} - \frac{-\pi_h \pi_{h'}}{n}$$

이다.

3. 적합도 검정

적합도검정의 귀무가설 $H_0 : \underline{\pi} = \underline{\pi_0}$ 하에서 $E(\hat{\pi}_h) = \pi_h$ 이고 $\underline{\hat{\pi}}$ 의 공분산행렬을 V라고 하면 중심극한정리에 의해 점근적으로 다음이 성립한다.

$$(\underline{\hat{\pi}} - \underline{\pi_0}) \rightarrow N(\underline{0}, V) \tag{2. 8}$$

여기서 $N(\underline{0}, V)$는 점근적으로 평균벡터 $\underline{0}$, 공분산행렬 V의 다변량정규분포를 따르는 것을 의미한다. V의 어떤 통계량 V을 이용하여 새로운 통계량 X_w^2을

$$X_w^2 = (\underline{\hat{\pi}} - \underline{\pi})' \; V^{-1}(\underline{\hat{\pi}} - \underline{\pi}) \tag{2. 9}$$

와 같이 정의할 수 있으며 이를 일반화 Wald 통계량이라 하며 H_0하에서 일반화 Wald통계량(generalized Wald statistic)은 점근적으로 자유도 r−1인 X^2분포를 따른다

$$X^2 = \sum_{h}^{r} n \frac{(\hat{\pi}_h - \pi_h)^2}{\pi_h} \tag{2. 10}$$

은 Pearson 통계량이며 식 (2. 7)에 의해 Wald 통계량은

$$X_w^2 = \frac{n}{G} X^2$$

이다. 여기서

$$\frac{n}{G} \text{ 은 } \frac{1}{n} < \frac{n}{G} \leqq 1 \text{ 이므로}$$

Pearson 통계량 X^2을 축소시키며 이를 축소인자(reduction factor)라 부른다.

집락내 관련이 깊은 자료로부터 구해지는 X^2 통계량은 이 축소인자 n/G을 곱하므로 해서 수정통계량을 만들 수 있다.

공분산행렬 V는

$$V = \frac{n}{G} P \tag{2. 11}$$

로 표시할 수 있으며, 여기서

$$P = (D_\pi - \underline{\pi}\underline{\pi}')/n \tag{2. 12}$$

이며 이는 표본자료를 단순임의추출로부터 얻은 자료라고 가정할 때 얻어지는 $\underline{\hat{\pi}}$의 공분산행열이고 D_π는 $\underline{\pi}$의 원소에 의해 이루어지는 대각행렬이다.

Pearson 통계량 X^2의 행렬표현은 다음과 같다.

$$X^2 = (\hat{\underline{\pi}} - \underline{\pi_0})' P_0^{-1} (\hat{\underline{\pi}} - \underline{\pi_0}) \tag{2. 13}$$

여기서 P_0는 $\underline{\pi}=\underline{\pi_0}$일 때 P의 값이다.

표본설계효과에 의한 X^2의 수정통계량은 $P_0^{-1} V_0$의 고유값 λ_{0h}로부터 X^2이 점근적으로 $\sum_{h}^{r-1} \lambda_{0h} Z_h^2$임을 고려하고 있다. 여기서 Z_h^2은 χ_1^2을 따르고 있다. 그러므로 축소인자와 표본설계효과 사이의 관계는

$$P_0^{-1} V_0 = \frac{G}{n} I \tag{2. 14}$$

이고 λ_{0h}는 $I(G/n)$의 양의 고유값이다.

4. 독립성 검정

$R \times C$ 분할표에서 독립성 검정은

$$H_0 : f_{rc}(\underline{\pi}) = \pi_{rc} - \pi_{r+}\pi_{+c} = 0 \tag{2. 15}$$
$$r=1, \cdots, (R-1),\ c=1, \cdots, (C-1)$$

의 귀무가설을 검정하는 문제이다. π_{rc}는 (r, c)번째 모집단비율이고

$$\underline{\pi} = (\pi, \cdots, \pi_{Rc-1})',\ \pi_{r+} = \sum_{c}^{C} \pi_{rc},\ \pi_{+c} = \sum_{r}^{R} \pi_{rc},\ \sum_{c}^{C}\sum_{r}^{R} \pi_{rc} = 1$$

$f(\underline{\pi})$의 선형근사(linear approximation)에 의해

$$f(\hat{\underline{\pi}}) = f(\underline{\pi}) + F(\underline{\pi})\ \ (\hat{\underline{\pi}} - \underline{\pi}) \tag{2. 16}$$

이며 여기서 행렬 F는

$$F = F(\underline{\pi}) = [\partial f_t(\underline{\pi})/\partial \pi_h] \tag{2. 17}$$
$$t=1, \cdots, (R-1)\ (C-1),\ h=1, \cdots, (RC-1)$$

이다.

$f(\hat{\underline{\pi}})$가 $f(\underline{\pi})$의 일치추정량이라면 $f(\hat{\underline{\pi}})$의 공분산 행렬은 FVF'이며 귀무가설하에서

$$(f(\hat{\underline{\pi}}) - f(\underline{\pi})) \rightarrow N(\underline{0}, FVF') \tag{2. 18}$$

를 따른다. FVF' 일치추정량 $\hat{F}\hat{V}\hat{F}'$을 구할 수 있다면 H_0를 검정하기 위하여

$$X_W^2 = f(\hat{\underline{\pi}})' (FVF')^{-1} f(\hat{\underline{\pi}}) \tag{2. 19}$$

의 Wald 통계량을 사용하는 것이 옳을 것이다. 모델에 의한 접근이나 표본설계에 의한 접근이 FVF'의 직접 추정량을 가능하게 한다면 이러한 추정량을 구할 수 있다.

모형하에서 공분산행렬 FVF'과 단순임의표본 공분산행렬 FVF'의 관계는 식 (2. 7)로부터

$$\hat{F}\hat{V}\hat{F}' = \frac{G}{n}\hat{F}\hat{P}\hat{F}' = \hat{V}_f \tag{2. 20}$$

로 나타낼 수 있다.

또한, $\hat{F}\hat{P}\hat{F}' = \hat{P}_f$라 한다면 독립성검정을 위한 Pearson 통계량은

$$X_1^2 = \sum_r^R \sum_c^C \frac{n(\hat{\pi}_{rC} - \hat{\pi}_{r+}\,\hat{\pi}_{+c})}{\hat{\pi}_{r+}\,\hat{\pi}_{+c}} \tag{2. 21}$$

이며 여기서

$$\pi_{r+} = \sum_c^C \pi_{rc}\,,\ \pi_{+c} = \sum_r^R \pi_{rc}$$

이다. Pearson 통계량과 Wald통계량과의 관계는 식 (2. 10)에서 처럼

$$X_W^2 = \frac{n}{G} X_1^2 \tag{2. 22}$$

이다. Person통계량 X_1^2의 행렬표현은 다음과 같다.

$$X_1^2 = f(\hat{\underline{\pi}})'\ P_f^{-1}\, f(\hat{\underline{\pi}}) \tag{2. 23}$$

또한

$$X_1^2 = \sum_{t=1}^{(R-1)(C-1)} \delta_{to}\ \omega_t^2$$

이며 ω_t^2은 독립적인 χ_1^2을 따르고 δ_{to}들은 귀무가설하에서 $P_f^{-1}\,V_f$의 양의 고유 값이다. 이제 V_f를 $P_f(G/n)$로 대체하면

$$P_f^{-1}\,V_f = P_f^{-1}\,P_f(G/n) = I(G/n) \tag{2. 24}$$

이 되므로

$$\delta_{to} = (G/n)$$

이다.

Pearson통계량 X_1^2은 축소인자 n/G에 의해 Wald 통계량 X_{1W}^2이 된다.

여기서 (n/G)는

$$G = n + \theta \sum_{i}^{a} b_i (b_i - 1)$$

이 경우에 $b_i = b$로 동일하다면

$$\frac{n}{G} = \frac{1}{1+\theta(b-1)}$$

이다. 그러므로 축소인자 (n/G)는 집락내 상관계수(intracluster correlation coefficient) θ에 의존한다. θ의 추정량으로 Cohen(1960, 1968), Fleiss외(1971)는 Kappa 추정량이 효율적임을 밝히고 있다. Kappa 추정량은

$$\hat{k}_h = 1 - \frac{b \sum_{i} \hat{\pi}_{ih}(1-\hat{\pi}_{ih})}{n\hat{\pi}_h(1-\pi_h)}$$

이고, 여기서 π_{ih}는 i 번째 집락의 요소가 h번째 Cell에 떨어지는 요소의 비율에 대한 추정량이다. k_h의 가중평균에 의한 θ의 추정량은

$$\hat{\theta} = \frac{\sum_{h} \hat{\pi}_h(1-\hat{\pi}_h)\, k_h}{\sum_{h} \hat{\pi}_h (1-\pi_h)}$$

이다.

5. 시뮬레이션

5. 1 적합도검정에 대한 시뮬레이션

적합도검정을 위한 가상모집단은 3개의 집락으로 구성되고 각 집락은 4,000개의 요소로 이루어진다. 각 집락에서의 4개의 cell에 대한 모집단비율과 전체모집단비율은 〔표 1〕과 같고

〔표 1〕 시뮬레이션을 위한 모집단비율

집락 Cell	1	2	3	4
1	0.15	0.35	0.2	0.3
2	0.1	0.2	0.3	0.4
3	0.05	0.05	0.4	0.5
전 체	0.1	0.2	0.3	0.4

확률난수에 의해 발생된 12,000개의 모집단요소의 요약결과는 〔표 2〕와 같다.

〔표 2〕 모집단 구성

집락 Cell	1	2	3	4	합 계
1	603	1,368	856	1,173	4,000
2	405	794	1,187	1,614	4,000
3	177	189	1,635	1,999	4,000
합 계	1,185	2,351	3,678	4,786	12,000

표본추출과정은 일양난수에 의해 1단계에서 100개의 집락을 선택하고 선택된 집락내에서 5개의 요소를 복원추출하여 500개의 표본자료를 구성한다.

귀무가설 $H_0 : \underline{\pi} = \underline{\pi_0}$ 하에서

500개의 요소로부터 피어슨통계량(X^2), Wald 통계량(X_W^2), 표본설계에 의한 수정통계량(X_C^2), 본 연구의 수정통계량(X_C^2)을 계산한다. 이들 통계량이 자유도 3인 χ^2분포의 5% 기각역 $\chi_{0.05}^2 = 7.815$보다 큰가를 조사한다. 이러한 과정을 1,000번 반복했을 때 5%기각역에 속한 각 통계량들의 횟수는 〔표 3〕과 같다.

〔표 3〕 적합도 검정을 위한 5% 기각역에 속한 통계량의 횟수

통 계 량	X^2	X_W^2	X_C^2	X_C^2
5%기각역에 속한 횟수	96	49	54	59

〔표 3〕으로 부터 Wald 통계량은 χ^2 분포를 따른다는 것을 알 수 있고 Pearson 통계량은 χ^2분포에서 이탈한 것으로 나타난다. 대체적으로 표본설계효과에 의한 수정통계량과 본 연구의 수정통계량은 Wald통계량에 가까워지므로 Pearson통계량의 좋은 수정통계량이라고 할 수 있다. 또한 수정통계량은 계산과정이 간단하므로 표본설계에 대한 정보가 부족할 때 효율적으로 사용할 수 있다.

5. 2 독립성검정에 대한 시뮬레이션

독립성검정을 위한 가상모집단도 적합도검정처럼 3개의 집락으로 구성되고 각 집락은 4,000개의 요소로 이루어진다. 2×3분할표에서 각 집락 모집단비율과 전체모집단비율은 〔표 4〕와 같고 확률난수에 의해 발생된 12,000개 모집단요소비율은 [표 5]와 같다.

〔표 4〕 독립성 검정의 시뮬레이션을 위한 모집단비율

행	열	집락 1	집락 2	집락 3	전 체
1	1	0.14	0.06	0.04	0.08
	2	0.12	0.10	0.14	0.12
	3	0.12	0.26	0.22	0.20
2	1	0.16	0.12	0.08	0.12
	2	0.12	0.20	0.22	0.18
	3	0.34	0.26	0.30	0.30

〔표 5〕 독립성 검정을 위한 분할표

행	열	집락 1	집락 2	집락 3	전 체
1	1	559	236	147	942
	2	495	391	580	1,466
	3	470	1,040	848	2,358
2	1	626	485	338	1,449
	2	501	761	855	2,117
	3	1,349	1,087	1,232	3,668
계		4,000	4,000	4,000	12,000

적합도검정과 같은 표본추출과정에서 500개의 표본자료를 구성하여 행과 열에 대한 분할표의 독립성검정을 시행한다. 전체모집단의 구성비로 볼 때 행과 열은 서로 독립적이므로 귀무가설 H_0 : 행과 열은 독립이다. 하에서 통계량들은 χ^2 분포를 따라야 한다. 그러므로 자유도 2인 χ^2 분포 5%의 기각역 $\chi^2_{0.05}=5.991$에 속하는 비율에 의해 통계량들이 유효성을 조사할 수 있다. 표본추출과정을 1,000번 반복했을 때 5%의 기각역에 속하는 각 통계량들의 횟수는 〔표 6〕과 같다.

〔표 6〕 독립성검정을 위한 5%기각역에 속한 통계량의 횟수

통 계 량	X_P^2	X_{PW}^2	X_{PC}^2	X_{PG}^2
5%기각역에 속한 횟수	64	50	52	52

〔표 6〕로부터 Wald 통계량은 χ^2분포를 따르고 있고 Pearson 통계량은 대체적으로 χ^2 분포에서 멀어지고 있다. 표본설계에 의한 수정통계량이나 본 연구의 수정통계량이 같은 결과로 나타나 Wald통계량에 가까워지고 있으므로 좋은 수정통계량이라 할 수 있다.

4. 결 론

본 연구에서는 일단집락표본추출에서 Pearson통계량에 대한 축소인자에 의한 수정통계량의 연구결과를 정리하고 시뮬레이션을 통해 Pearson통계량, Wald통계량, 표본설계효과에 의한 수정통계량, 축소인자에 의한 수정통계량을 비교하였다.

Pearson X^2통계량은 실제 분류된 조사자료에서 적합도검정이나 분할표의 독립성검정에 널리 이용되는 검정 통계량이다. 그러나 조사자료가 단순임의 추출이 아닌 층화추출, 집락추출 또는 이들을 결합한 복합표본추출로부터 발생된다면 X^2 통계량은 χ^2분포를 이탈하게 된다. 그러므로 표본추출이 복합표본추출일 경우 적합도검정이나 분할표의 독립성 검정을 위해 Pearson 통계량을 이용한다면 검정결과는 심각한 오류를 범할 가능성이 있게 된다.

표본추출과정에서 얻을 수 있는 어떤 유효한 정보를 이용하여 X^2 통계량에 적용시킨다

면 X^2 통계량의 효율적인 수정통계량을 만들 수 있다.

본 연구의 시뮬레이션을 통해 이러한 수정통계량이 참값인 Wald통계량에 가까와지고 있음을 알 수 있었다.

그러므로 발표된 자료로 Wald통계량을 구할 수 없는 경우라면 본 연구의 수정통계량처럼 간단한 계산과정을 통해 Pearson 통계량의 수정을 가하는 것이 타당할 것으로 생각된다.

앞으로 이러한 분야의 추가적인 연구는 다단계집락추출이나 복합표본추출에 대해서도 시뮬레이션을 적용시켜 볼 수 있을 것이다.

참 고 문 헌

(1) 남궁 평(1986). 복합표본추출에서 비율의 검정법에 관한 비교연구, 박사학위논문, 성균관대학교 대학원.

(2) 박한철(1987). 집락표본추출에서 비율의 검정법에 관한 연구, 석사학위논문, 성균관대학교 대학원.

(3) Altham, P.M.E.(1976). "Discrete variable analysis for individuals grouped into families", *Biometrika*, 63, 263-9

(4) Bedrick, E.J.(1983). Adjusted chi-squared tests for cross-classified tables of survey data, *Biometrika*, 70, 591-5.

(5) Brier, S.S.(1980). Analysis of contingency tables under cluster sampling, *Biometrika*. 67, 591-6.

(6) Choi, J.W.(1987). A reduction factor in goodness-of-fit and independence tests for clustered and weighted obeservations, *NCHS*.

(7) Choi, J.W.and Landis, J.R.(1987). Estimation of cluster correlation form a two-stage nested random effects model for unbalanced categorical data, Unpublished Revised for Manuscripted(JASA).

(8) Cohen, J.E.(1976). The distribution of chi-squared statistic under clustered sampling from contigency tables, *JASA*, 71, 355, •p, 665-70.

(9) Fellegi, I.P.(1980). Approximate texts of independence and goodness of fit based upon stratified multistage samples, *JASA*, 75, 261-8.

(10) Fienberg, S.E.(1979). The use of chi-squared statistics for categorical data problems, *JRSS*, B, 41, 54-64.

(11) Holt, D., Scott, A.J.and Ewings, P.O.(1980). Chi-squared tests with survey data, *JRSS*, A, 143, 302-20.

(12) Landis, J.R., Lepkowski, J.M., Eklund, S.A., Stehower, S.A.(1984). A statistical methodology for analyzing data from a complex survey : The first National Health and Nutrition Examination Survey, Series 2, No. 92, *NCHS*.

(13) Rao, J.N.K. and Scott, A. J.(1981). The analysis of categorical data from complex sample survey : Chi-squared tests for goodness of fit and independence in two-way tables, *JASA*, 76, 221-30.

(14) Wald, A.(1943). Test of statistical hypothesis concerning several parameters when the number of observation in large, *Trans. Math. Soc.* 54, 426-482.

Goodness of Fit and Independence Tests for Clustered Sample Data

Pyong Namkung, Byung Soo Choi, Joo Lock Lee

Abstract

Modified Pearson X^2 statistic is concerned. Moreover the four statistics(Pearson, Wald, modified sample design effects and reduction factor) are compared in one-stage sampling situation.

In case of categorical of fit and independence tests for sample data above, it is shown that there is a significant behavior between Pearson X^2 and Wald statistic, but minor difference in modified statistics by simulation methods.