Journal of the Korean Data & Information Science Society
2013, **24**(4), 815–824

http://dx.doi.org/10.7465/jkdi.2013.24.4.815
한국데이터정보과학회지

k-모집단 동질성검정에서 피어슨검정의 오차성분 분석에 관한 연구[†]

허순영[1]

[1]창원대학교 통계학과

접수 2013년 6월 8일, 수정 2013년 7월 3일, 게재확정 2013년 7월 8일

요 약

국가단위의 조사와 같은 대규모 표본조사에서는 표본의 대표성을 확보하기 위해 층화, 집락, 계통, 불균등확률추출 등을 종합적으로 사용하는 복합표본설계가 일반화되어 있다. 이러한 복합표본설계에 기초한 범주형 자료분석에서는 자료의 독립성과 다항분포를 가정하는 전통적인 피어슨검정이 왜곡된 검정결과를 가져올 수 있다. 본 연구는 복합표본설계에 의한 범주형조사자료의 k-모집단 동질성검정에서 설계기반 일치통계량인 Wald 검정통계량을 유도하고, 전통적인 피어슨검정통계량을 사용할 경우 발생할 수 있는 오차요인을 항목별로 분해하여, 분산의 편의에 의한 영향, 추정량의 편의에 의한 영향, 기타 분산의 편의와 추정량의 편의가 교락되어 미치는 영향으로 각각 분해하는 식을 도출하였다. 또한, 도출된 식의 각 항목이 피어슨 카이제곱검정통계량에 미치는 상대적 크기를 경험적으로 확인하기 위해 국민건강영양조사 제4기 2차년도 자료를 이용해 경험분석 하였다. 분석결과, 변수에 따른 차이는 있지만 대체로 분산의 편의가 미치는 영향이 추정량의 편의가 미치는 영향보다 크다는 것을 명확히 확인할 수 있었다.

주요용어: 동질성검정, 범주형자료, 복합표본설계, 설계효과, 왈드검정, 피어슨검정.

1. 서론

전국단위의 조사나 조사범위가 광범위한 조사들은 표본의 대표성 확보를 위해 층화, 집락, 계통, 다단계 추출, 불균등확률추출 등과 같은 여러 가지 표본추출방법을 종합적으로 적용하여 표본을 설계하는데, 이러한 표본추출방법을 복합표본추출법 (complex sampling)이라 한다. 오늘날 이러한 표본설계에 기초한 조사연구를 자주 접할 수 있다 (Kim 등, 2009; Heo 등, 2010). 이러한 복합표본추출법에 의한 표본조사자료는 전통적인 통계이론에서 가정하는 자료들 간의 서로 독립이고 동일한 분포를 따른다 (independent and identically distributed; IID)는 조건을 일반적으로 충족하지 못한다. 표본조사자료의 분석은 표본추출방법이 반영된 분석방법을 사용하여야만 정확한 분석이 이루어질 수 있고, 복합표본추출법에 의한 표본조사자료의 분석 역시 표본추출방법이 반영된 분석이 이루어져야만 정확한 분석결과를 얻을 수 있다.

범주형 자료분석에 자주 이용되는 독립성검정, 동질성검정, 적합성검정은 일반적으로 피어슨 카이제곱검정 (Pearson chi-squared test)을 사용한다. 피어슨 카이제곱검정은 주어진 자료들이 서로 독립이고 동일한 다항분포를 따른다는 가정과 관찰치들이 각 범주에 속할 기대빈도가 충분히 클 때 근사적으로

[†] 이 논문은 2011~2012년도 창원대학교 연구비에 의하여 연구되었음.
[1] (641-773) 경남 창원시 의창구 퇴촌로92, 창원대학교 통계학과, 교수. E-mail: syheo@changwon.ac.kr

카이제곱분포에 따른다는 가정을 동시에 충족시킬 때 정확한 검정결과를 얻을 수 있다. 따라서 IID 가정을 만족하지 못하는 표본조사자료에 피어슨 카이제곱검정을 무리하게 적용할 경우 왜곡된 분석결과를 얻을 수 있다. Heo와 Chung (2012)은 복합표본조사자료의 2-모집단 동질성검정에서 피어슨 카이제곱검정과 Wald 검정을 비교하는 경험분석을 통해 자료분석에 표본설계를 반영하는 것이 중요함을 실증적으로 보여주었다.

복합표본설계에 의한 범주형조사자료분석 과정에서, 전통적 피어슨 카이제곱검정통계량을 계산할 때 모비율 추정치로 단순표본비율을 사용하는 대신 표본가중치를 반영한 설계기반 불편추정량 (design based unbiased estimator)을 사용한 피어슨형 검정통계량에 설계효과 (design effect)를 반영하여 조정하려는 연구들이 활발히 진행되었다 (Holt 등, 1980; Rao와 Scott, 1981, 1984, 1987; Tomas와 Rao, 1987). 그러나 전통적인 피어슨 카이제곱검정통계량을 사용하는 경우, 오차는 분산에 대한 추정량의 편의에서 뿐만 아니라 추정량의 편의에 의해서도 발생한다.

본 연구에서는 복합표본설계에 의한 범주형조사자료의 k-모집단 동질성검정에서 설계기반 일치추정량인 Wald 검정통계량을 유도하고, 전통적 피어슨 카이제곱검정통계량의 오차 요인을 성분별로 분해하여, 분산의 편의에 의한 영향, 추정량의 편의에 의한 영향, 분산의 편의와 추정량의 편의가 교락되어 미치는 영향의 크기를 각각 구분하는 식을 도출하였다. 도출된 식의 경험분석을 위해 국민건강영양조사 제4기 2차년도 (2008) 자료를 사용하였다. 2절에서는 Wald 검정통계량과 피어슨 카이제곱검정통계량의 오차요인을 구분하는 식을 각각 유도하였다. 3절에서는 경험분석을 통해 각 요인별 상대크기를 비교하였고, 4절에 결론을 제시하였다.

2. k-모집단 동질성검정

k개의 모집단으로부터 크기 $n_1, n_2, \cdots, n_k$인 표본을 독립추출한 후, 각 표본의 표본원소들이 상호배반인 c개의 범주 $A_1, A_2, \cdots, A_c$에 속할 확률 p_{ij} $(i = 1, 2, \cdots, k;\ j = 1, 2, \cdots, c)$이 k개 모집단에서 모두 동일한지를 검정하기 위한 동질성검정 (homogeneity test)의 귀무가설은

$$H_0 : p_1 = p_2 = \cdots = p_k (= a)$$

이다. 여기서, $p_i = (p_{i1}, p_{i2}, \cdots, p_{i,c-1})^T$, $a = (a_1, a_2, \cdots, a_{c-1})^T$이고, $\sum_{j=1}^{c} p_{ij} = 1$, $\sum_{j=1}^{c} a_j = 1$이다.

2.1. 피어슨검정통계량

각 표본의 표본단위들이 c개의 범주에 속하는 관측도수를 $(n_{i1}, n_{i2}, \cdots, n_{ic})$ $(i = 1, \cdots, k)$라 할 때, k-모집단 동질성검정을 위한 종래의 피어슨검정통계량은

$$\chi^2_P = \sum_{i=1}^{k}\sum_{j=1}^{c} \frac{(n_{ij} - n_i \widehat{a}_{pj})^2}{n_i \widehat{a}_{pj}} = \sum_{i=1}^{k}\sum_{j=1}^{c} n_i \frac{(\widehat{p}_{ij} - \widehat{a}_{pj})^2}{\widehat{a}_{pj}}$$

이고, 여기서 $\widehat{p}_{ij} = n_{ij}/n_i$이고 $\widehat{a}_{pj} = \sum_{i=1}^{k} n_{ij}/n = \sum_{i=1}^{k} n_i \widehat{p}_{ij}/n$이다. 이 통계량을 이차형식으로 표현하면

$$\chi^2_P = n(\widehat{p} - \widehat{a}_P)^T \widehat{B} (\widehat{p} - \widehat{a}_P) \tag{2.1}$$

으로 나타낼 수 있다 (Holt 등, 1980). 여기서 $\widehat{p}$과 $\widehat{a}_P$은 각각 $k(c-1)\times 1$ 열벡터로 $\widehat{p}^T = (\widehat{p}_1^T, \cdots, \widehat{p}_k^T)$와 $\widehat{a}_P^T = (\widehat{a}_p^T, \cdots, \widehat{a}_p^T)$이고, $\widehat{p}_i$와 $\widehat{a}_p$은 각각 $\widehat{p}_i = (\widehat{p}_{i1}, \cdots, \widehat{p}_{i,c-1})^T$, $\widehat{a}_p = (\widehat{a}_{p1}, \cdots, \widehat{a}_{p,c-1})^T$이다. 또한,

$$\widehat{B} = \widehat{F} \otimes \widehat{P}^{-1}$$

으로, 여기서 $\bigotimes$는 행렬연산의 직적 (direct product)을 나타내고 $\widehat{F} = diag(\widehat{f}) - \widehat{f}\widehat{f}^T$이고 $\widehat{P} = diag(\widehat{a}_p) - \widehat{a}_p\widehat{a}_p^T$이며, $\widehat{f}_i = n_i/n$이고 $\widehat{f} = (\widehat{f}_1, \cdots, \widehat{f}_k)^T$이다. 이 때, $\widehat{B}$의 계수 (rank)는 $(k-1)\times(c-1)$이다.

각 표본이 모집단으로부터 단순임의복원추출되었고, i번째 표본의 관측도수 $(n_{i1}, n_{i2}, \cdots, n_{ic})$가 다항분포 $(n_i, p_{i1}, \cdots, p_{ic})$를 따른다는 가정을 만족할 때, χ^2_P은 근사적으로 자유도 $(k-1)(c-1)$인 카이제곱분포 $\chi^2_{(k-1)(c-1)}$를 따른다. 그러나 각 모집단으로부터 표본을 추출하는 과정에서 층화, 집락, 불균등확률추출 등이 복합적으로 사용된 복합표본의 경우에는 더 이상 이러한 가정을 충족하지 못하게 되고 그 결과 χ^2_P는 $\chi^2_{(k-1)(c-1)}$를 따르지 않는다.

2.2. Wald 검정통계량

각 모집단으로부터 복합표본설계에 의한 표본이 추출되었고, $\widehat{\pi}_i$은 표본설계를 반영한 i번째 모집단의 모비율 p_i에 대한 일치추정량 (consistent estimator)이라고 하자. Holt 등 (1980)은 $\widehat{\pi}_i$을 사용한 피어슨형 검정통계량의 편의를 조정하기 위해 조정된 통계량을 제안하였는데, 여기서는 Holt 등 (1980)이 사용한 정의를 사용한다. 즉, $n = n_1 + \cdots + n_k$가 증가할 때 n_i들이 함께 증가하여 $n \to \infty$일 때 $\widehat{f}_i = n_i/n \to f$ $(0 < f < 1; i = 1, \cdots, k)$이고, $n_i \to \infty$일 때

$$\sqrt{n_i}(\widehat{\pi}_i - p_i) \xrightarrow{L} N(0, V_i)$$

이라 하자. 그 때, Slutsky's 정리 (Casella와 Berger, 1990)에 의해 위 식은 $\sqrt{n}(\widehat{\pi}_i - p_i) \xrightarrow{L} N(0, V_i/f_i)$가 된다. 따라서 서로 독립인 두 모집단 (i, k)에 대해, 귀무가설 하에서

$$\sqrt{n}(\widehat{\pi}_i - \widehat{\pi}_k) \xrightarrow{L} N(0, V_i/f_i + V_k/f_k)$$

$(i = 1, \cdots, k-1)$이고, 서로 독립인 k개의 모집단에 대해서는

$$\sqrt{n}(\widehat{\pi} - \widehat{\pi}_{k0}) \xrightarrow{L} N(0, V_0) \tag{2.2}$$

가 된다. 여기서 $\widehat{\pi}$와 $\widehat{\pi}_{k0}$은 $\widehat{\pi}^T = (\widehat{\pi}_1^T, \cdots, \widehat{\pi}_{k-1}^T)$와 $\widehat{\pi}_{k0}^T = (\widehat{\pi}_k^T, \cdots, \widehat{\pi}_k^T)$로 각각 $(k-1)(c-1)\times 1$인 열벡터이고

$$V_0 = HV_FH^T$$

이며, 이 때

$$H = [I_{(k-1)\times(k-1)} \vdots -1_{(k-1)}]_{(k-1)\times k} \otimes I_{(c-1)\times(c-1)}$$이고 $V_F = \oplus_{i=1}^k V_i/f_i$

이다. 여기서 $\bigoplus$은 행렬연산의 직합 (direct sum)이고 $1_{(k-1)}$은 모든 원소가 1이고 길이가 $k-1$인 열벡터이다. 식 (2.2)로 부터

$$n(\widehat{\pi} - \widehat{\pi}_{k0})^T V_0^{-1}(\widehat{\pi} - \widehat{\pi}_{k0}) \xrightarrow{L} \chi^2_{(k-1)(c-1)}$$

가 된다. $\widehat{V}_i$이 표본설계를 반영한 V_i의 일치추정량일 때, V_F에 $\widehat{V}_i$와 $\widehat{f}_i$을 대입하여 얻은 $\widehat{V}_0$을 위 식에 대입하면 Wald 검정통계량

$$\chi^2_W = n(\widehat{\pi} - \widehat{\pi}_{k0})^T \widehat{V}_0^{-1}(\widehat{\pi} - \widehat{\pi}_{k0}) \tag{2.3}$$

을 얻는다. 귀무가설이 참일 때, χ^2_W은 Slutsky's 정리에 의해 근사적으로 $\chi^2_{(k-1)(c-1)}$을 따르고, χ^2_W을 이용하여 자유도 $(k-1)(c-1)$인 카이제곱분포에 기초하여 검정한다.

2.3. 피어슨검정통계량의 분해

복합표본설계 하에서는, 2.1절에서 정의한 모비율 p_i의 점추정량 $\widehat{p}_i$은 일반적으로 편의추정량으로 식 (2.1)의 $(\widehat{p}-\widehat{a}_P)$은

$$\widehat{p}-\widehat{a}_P=(\widehat{\pi}_F-\widehat{a}_\Pi)+(\widehat{p}-\widehat{\pi}_F)-(\widehat{a}_P-\widehat{a}_\Pi) \tag{2.4}$$

으로 분해될 수 있다. 여기서 $\widehat{\pi}_F^T=(\widehat{\pi}_1^T,\cdots,\widehat{\pi}_k^T)$와 $\widehat{a}_\Pi^T=(\widehat{a}_\pi^T,\cdots,\widehat{a}_\pi^T)$로 각각 $k(c-1)\times 1$인 열벡터이고, $\widehat{a}_\pi=(\widehat{a}_{\pi 1},\cdots,\widehat{a}_{\pi,c-1})^T$이고 $\widehat{a}_{\pi j}=\sum_{i=1}^k n_i\widehat{\pi}_{ij}/n$로 $\widehat{\pi}_{ij}$는 $\widehat{\pi}_i$의 j번째 원소이다.

식 (2.4)의 오른쪽 첫 번째 항 $(\widehat{\pi}_F-\widehat{a}_\Pi)$은 표본설계를 반영한 $(p-a_F)$의 일치추정량이고, 두 번째 항과 세 번째 항은 각각 $\widehat{p}$와 $\widehat{a}_P$의 편의에 기인하여 발생되는 양이다. 즉, 세 번째 항의 $\widehat{a}_P^T=(\widehat{a}_p^T,\cdots,\widehat{a}_p^T)$에서 $\widehat{a}_p$의 j번째 원소와 $\widehat{a}_\Pi^T=(\widehat{a}_\pi^T,\cdots,\widehat{a}_\pi^T)$에서 $\widehat{a}_\pi$의 j번째 원소는 각각

$$\widehat{a}_{pj}=\sum_{i=1}^k n_i\widehat{p}_{ij}/n=\sum_{i=1}^k \widehat{f}_i\widehat{p}_{ij}\text{와 }\ \widehat{a}_{\pi_j}=\sum_{i=1}^k n_i\widehat{\pi}_{ij}/n=\sum_{i=1}^k \widehat{f}_i\widehat{\pi}_{ij}$$

이므로, $(\widehat{a}_p-\widehat{a}_\pi)$의 j번째 원소는 $\widehat{a}_{pj}-\widehat{a}_{\pi j}=\sum_{i=1}^k \widehat{f}_i(\widehat{p}_{ij}-\widehat{\pi}_{ij})$이고 $(\widehat{a}_p-\widehat{a}_\pi)=(\widehat{f}^T\otimes I)(\widehat{p}-\widehat{\pi}_F)$가 된다. 여기서 항등행렬 I의 차원은 $(c-1)\times(c-1)$이다. 따라서 식 (2.3)의 세 번째 항은

$$(\widehat{a}_P-\widehat{a}_\Pi)=1_{(k)}\otimes\left[(\widehat{f}^T\otimes I)(\widehat{p}-\widehat{\pi}_F)\right]$$

으로, $(\widehat{p}-\widehat{\pi}_F)$의 함수이다. $(\widehat{p}-\widehat{\pi}_F)$은 $\widehat{p}$의 편의에 대한 추정량이므로 식 (2.4)의 오른쪽 두 번째 항과 세 번째 항은 $\widehat{p}$의 편의로 인해 발생하는 양이라 할 수 있다.

식 (2.4)를 식 (2.1)에 대입하여 정리하면, 피어슨검정통계량은

$$\chi_P^2=n(\widehat{p}-\widehat{a}_P)^T\widehat{B}(\widehat{p}-\widehat{a}_P)=n(\widehat{\pi}_F-\widehat{a}_\Pi)^T\widehat{B}(\widehat{\pi}_F-\widehat{a}_\Pi)+2n(\widehat{\pi}_F-\widehat{a}_\Pi)^T\widehat{B}R_p+nR_p^T\widehat{B}R_p \tag{2.5}$$

가 된다. 여기에서 $R_p=(\widehat{p}-\widehat{\pi}_F)-(\widehat{a}_P-\widehat{a}_\Pi)$이다. 식 (2.5)의 오른쪽 첫 번째 2차형식의 경우, 모비율에 대한 점추정량으로 표본설계를 반영한 일치추정량을 사용하고 있으나 분산은 여전히 다항분포와 IID를 가정한 분산을 사용하고 있어서 일반적으로 분산에 대해 복합표본설계를 반영한 일치추정량을 사용하는 식 (2.3)의 Wald 검정통계량 χ_W^2보다 큰 값이 된다. 따라서 χ_P^2은 다시

$$\chi_P^2=\chi_W^2+\left\{n(\widehat{\pi}_F-\widehat{a}_\Pi)^T\widehat{B}(\widehat{\pi}-\widehat{a_\Pi})-\chi_W^2\right\}+nR_p^T\widehat{V}_F^{-1}R_p+R \tag{2.6}$$

로 나타낼 수 있다. 여기서, $R=2n(\widehat{\pi}_F-\widehat{a}_\Pi)^T\widehat{B}R_p+\left\{nR_p^T\widehat{B}R_p-nR_p^T\widehat{V}_F^{-1}R_p\right\}$이고 $\widehat{V}_F=\oplus_{i=1}^k\widehat{V}_i/\widehat{f}_i$이다. 식 (2.6)의 오른쪽 두 번째 항은 V_i의 일치추정량 $\widehat{V}_i$대신 $\widehat{P}$을 사용함으로 인해 발생하는 양이다. 세 번째 항은 식 (2.5)의 세 번째 항에서 분산의 편의 추정량 대신 일치추정량 $\widehat{V}_i$을 사용하였기 때문에 분산에 의한 편의는 제거되고 비율에 대한 점추정량 $\widehat{p}$와 $\widehat{a}_P$의 편의만이 χ_P^2에 미치는 영향의 크기를 나타낸다. 마지막 항 R은 비율에 대한 점추정량과 분산추정량의 편의가 분리되지 않은 채 교락되어 영향을 끼치는 기타 크기를 나타낸다.

3. 경험분석

본 장에서는 복합표본설계에 기초한 범주형조사자료의 k-모집단 동질성검정에서 전통적인 피어슨검정통계량을 사용하는 경우, 2절에서 제시한 피어슨검정통계량의 오차성분별 크기를 경험분석하기 위해 3단계 층화집락표본추출방법에 의해 표본을 선정한 국민건강영양조사 제4기 2차년도 (2008) (The

fourth Korea national health and nutrition examination survey; KNHANES IV-2) 자료를 이용하였다. KNHANES IV에서는 전국을 29개 층으로 층화한 후, 1차 추출단위인 동·읍·면을 계통추출하였고, 선정된 표본 동·읍·면에서 표본조사구를 추출한 후, 각 표본 조사구 당 계통추출법에 의해 23가구를 선정하는 표본추출방법을 적용하였다 (Korea Centers for Disease Control and Prevention, 2009). KNHANES IV의 표본추출에 대한 자세한 내용은 Lee와 Park (2007)을 참고할 수 있다.

본 연구에서는 경험분석을 위해 KNHANES IV-2 원시자료에 있는 시·도 변수 (변수명: region)를 이용해, 전국 16개 시·도를 2개, 4개 그리고 6개 모집단으로 구분하였는데, Table 3.1은 본 연구에서 사용한 모집단의 정의를 보여준다. 이러한 정의는 본 연구를 위해 임의로 정의한 것으로 KNHANES IV의 층 정의와는 일치하지 않는다.

Table 3.1 Definition of populations

No. of pop	Definition of populations	
2	1. Seoul and 6 metropolitan cities	
	2. 9 Do regions	
4	1. Capital region	Seoul, Incheon, GyeongGi
	2. Gangwon-Gyeongbuk region	Daegu, Gangwon, Gyeongbuk
	3. Central inland region	Deajeon, Chungbuk, Chungnam, Jeonbuk
	4. Jeju-Southern region	Busan, Gwangju, Ulsan,Jeonnam, Gyeongnam, Jeju
6	1. Seoul	
	2. Gyeong-In region	Inchon, GyeongGi
	3. Chungcheong region	Dejeon, Chungbuk, Chungnam
	4. Dongbuk region	Daegu, Gangwon, Gyeongbuk
	5. Dongnam region	Busan, Ulsan, Gyeongnam
	6. Honam region	Gwangju, Jeonbuk, Jeonnam, Jeju

본 연구를 위해 KNHANES IV-2 자료 중에서 일부 변수를 선택하였는데 Table 3.2에서는 본 연구에서 선택된 변수들과 설명을 제시하였다. 범주형자료분석에서 특정범주의 응답자수가 너무 작으면 검정결과의 신뢰성이 떨어지므로 이를 방지하기 위해 근방 범주를 병합하였는데 Table 3.2의 마지막 열에 그 내용을 제시하였다. Table 3.2에 있는 변수들에 대한 보다 자세한 내용은 Korea Centers for Disease Control and Prevention (2009)의 "국민건강영양조사 원시자료 이용지침서 제4기 (2007-2009)"를 참고할 수 있다.

본 연구에서는 Table 3.2에 제시된 각 변수에 대해 Table 3.1에서 정의한 모집단별로 식 (2.5)와 식 (2.6)에서 제시한 피어슨검정통계량의 각 오차성분별 통계량 값을 계산하였다. 그 과정에서 V_i에 대한 일치추정량을 구하기 위해 선형화방법 (linearization method)을 사용하였다. Shao (1996)는 선형화방법을 포함한 V_i의 여러 추정방법의 효율을 비교하였다.

Table 3.2 Variables of KNHANES IV-2 used for empirical analysis

Division	VariableName	Variable Explanation	Recording for analysis
Household survey	ho_incm	Income quartiles(household)	
	live_t	Housing types	Concatenate "Don't Know(DK)" (16 cases) into"etc"category.
Health questionnaire survey	edu	Education	
	tins	Health insurance types	Concatenate "DK (122 cases)" and "Not join (7 cases)"
	D_1_1	Subjective health condition	Record "DK (16 cases)" asmissings.
	BO1	Subjective body type	Record "DK (5 cases)" as missings.
	BP1	Degrees of usual stress recognition	Record "DK (7 cases)" asmissings
Medical checkup	T_Q_HR	Examinee's hearing	Concatenate categories (3,4). Record categories (8.9) asmissings.

Table 3.3은 2절의 식 (2.5)

$$\chi^2_P = n(\widehat{\pi}_F - \widehat{a}_\Pi)^T \widehat{B}(\widehat{\pi}_F - \widehat{a}_\Pi) + 2n(\widehat{\pi}_F - \widehat{a}_\Pi)^T \widehat{B} R_p + n R_p^T \widehat{B} R_p$$

의 피어슨검정통계량 χ^2_P과 오른쪽 세 항을 각각 계산한 결과를 순서대로 제시하였다. 따라서 오른쪽 세 항을 합하면 χ^2_P과 일치한다.

Table 3.3 Pearson chi-squared test statistics for k-population homogeneity test and its three components in the right-hand side of equation (2.5)

No. of pops	VariableName	No. of cat.	χ^2_P	1st term	2nd term	3rd term
2	ho_incm	4	72.71	22.59	34.89	15.23
	live_t	6	464.96	187.42	146.25	131.29
	edu	4	69.22	53.78	11.34	4.10
	tins	5	17.05	5.70	7.04	4.31
	D_1_1	5	43.26	17.34	16.65	9.28
	BO1	5	32.32	26.86	0.64	4.82
	BP1	4	7.52	6.02	0.48	1.03
	T_Q_HR	3	23.26	7.59	11.30	4.37
4	ho_incm	4	230.53	181.18	40.24	9.11
	live_t	6	1296.77	1189.83	-19.49	126.42
	edu	4	113.49	101.59	3.19	8.70
	tins	5	75.76	70.78	-6.61	11.59
	D_1_1	5	202.71	220.68	-50.39	32.42
	BO1	5	26.58	27.39	-4.29	3.48
	BP1	4	25.58	23.47	-2.04	4.15
	T_Q_HR	3	36.09	13.09	16.62	6.38
6	ho_incm	4	266.82	202.89	41.76	22.17
	live_t	6	1641.49	1501.05	-62.34	202.78
	edu	4	134.57	134.09	-13.06	13.53
	tins	5	88.44	81.07	-7.83	15.19
	D_1_1	5	520.42	565.24	-63.55	18.73
	BO1	5	56.62	57.91	-7.60	6.31
	BP1	4	34.81	44.96	-20.37	10.22
	T_Q_HR	3	24.85	7.90	10.38	6.57

Table 3.4는 식 (2.5)를 오차요인별로 재표현한 식 (2.6)

$$\chi^2_P = \chi^2_W + \left\{ n(\widehat{\pi}_F - \widehat{a}_\Pi)^T \widehat{B}(\widehat{\pi} - \widehat{a}_\Pi) - \chi^2_W \right\} + n R_p^T \widehat{V}_F^{-1} R_p + R$$

에 기초하여 오차성분별로 각각 계산한 결과를 제시하였다.

Table 3.4는 네 번째 열 (χ^2_P)에 피어슨검정통계량, 다섯 번째 열 (χ^2_W)에 Wald 검정통계량, 2nd term 열에 식 (2.6)의 오른쪽 두 번째 항, 3rd term 열에 세 번째 항, 그리고 마지막 열에 R의 값을 각각 제시하였다. 따라서 Table 3.4의 마지막 네 열의 합은 χ^2_P과 일치한다. Table 3.5에서는 Table 3.4의 마지막 네 열의 각 값을 χ^2_P로 나눈 값의 백분율을 제시하였다. 따라서 Table 3.5의 네 열의 합은 100%가 된다.

Table 3.5에서 Wald 검정통계량 값은 피어슨검정통계량 값의 약 2.16%~95.75%에 해당하는 크기의 값을 가진다. 대체로 가구조사 변수인 live_t, ho_incm에서 그 비율이 건강설문조사 변수인 BO1과 BP1에 비해 크게 작다. 즉, 이 변수들에서 피어슨검정통계량은 Wald검정에 비해 크게 과대추정된다. 건강설문조사 변수들 중 BO1과 BP1의 경우, 모집단 수와 상관없이 이 백분율이 모두 큰 값을 갖는다. 특히 BO1의 모집단수가 4인 경우 두 검정통계량이 값이 매우 유사하고 이 백분율은 95.75%이다.

식 (2.6)의 두 번째 항은 검정통계량의 계산에서 비율 추정량에는 일치추정량을 사용하되 분산은 일치추정량 $\widehat{V}_i$대신 $\widehat{P}$를 사용함으로써 추가되는 증가분으로, Table 3.5의 5번째 열은 이 값에 대한 χ^2_P의 백분비를 나타낸다. 이 백분율은 χ^2_W/χ^2_P의 경우와 반대의 경향을 나타낸다. χ^2_W/χ^2_P에서 작은 비율을

나타내었던 가구조사 변수들이 대체로 상대적으로 높은 비율을 나타내어, 건강설문조사 보다 가구조사에서 설계효과 (design effect)가 더 크게 나타남을 알 수 있다. 한편 검진조사 변수인 T_Q_HR은 모든 모집단수에 대해 가장 낮은 비율들을 보이는 경향을 나타낸다. 특히, Table 3.4에서 T_Q_HR의 모집단수가 6인 경우, 식 (2.6) 두번째 항의 값은 -0.72로 음의 값으로 나타난다.

Table 3.4 Pearson chi-squared and Wald test statistics for k-population homogeneity test, and three components in the right-hand side of equation (2.6)

No. of pops	VariableName	No. of cat.	χ^2_P	χ^2_W	2nd term	3rd term	R
2	ho_incm	4	72.71	3.91	18.68	2.98	47.14
	live_t	6	464.96	10.06	177.36	14.72	262.83
	edu	4	69.22	4.62	49.16	1.88	13.55
	tins	5	17.05	1.21	4.48	1.13	10.22
	D_1_1	5	43.26	7.71	9.63	5.37	20.56
	BO1	5	32.32	19.10	7.76	2.69	2.77
	BP1	4	7.52	4.08	1.94	0.71	0.79
	T_Q_HR	3	23.26	4.61	2.98	3.23	12.44
4	ho_incm	4	230.53	36.92	144.26	1.14	48.21
	live_t	6	1296.77	146.17	1043.66	16.57	90.37
	edu	4	113.49	37.11	64.48	3.33	8.57
	tins	5	75.76	25.89	44.89	7.50	-2.52
	D_1_1	5	202.71	82.46	138.22	14.79	-32.76
	BO1	5	26.58	25.45	1.94	2.69	-3.50
	BP1	4	25.58	18.83	4.64	2.63	-0.52
	T_Q_HR	3	36.09	6.10	6.99	3.06	19.94
6	ho_incm	4	266.82	43.04	159.85	4.24	59.69
	live_t	6	1641.49	188.76	1312.29	26.22	114.22
	edu	4	134.57	49.94	84.15	7.02	-6.55
	tins	5	88.44	35.06	46.01	13.94	-6.58
	D_1_1	5	520.42	287.42	277.82	12.22	-57.04
	BO1	5	56.62	43.62	14.29	4.28	-5.57
	BP1	4	34.81	32.52	12.44	6.62	-16.77
	T_Q_HR	3	24.85	8.62	-0.72	5.88	11.07

Table 3.5 Relative sizes of Wald test statistics and three components in equation (2.6) to Pearson test statistics (unit: %)

No. of pops	VariableName	No. of cat.	$\chi^2_W/\ \chi^2_P$	2nd term/ χ^2_P	3rd term/ χ^2_P	$R/\ \chi^2_P$
2	ho_incm	4	5.38	25.69	4.10	64.83
	live_t	6	2.16	38.15	3.17	56.53
	edu	4	6.67	71.02	2.72	19.59
	tins	5	7.11	26.30	6.66	59.94
	D_1_1	5	17.82	22.26	12.41	47.52
	BO1	5	59.10	24.01	8.32	8.57
	BP1	4	54.26	25.80	9.44	10.51
	T_Q_HR	3	19.82	12.81	13.89	53.48
4	ho_incm	4	16.02	62.58	0.49	20.91
	live_t	6	11.27	80.48	1.28	6.97
	edu	4	32.70	56.82	2.93	7.55
	tins	5	34.17	59.25	9.89	-3.32
	D_1_1	5	40.68	68.19	7.30	-16.16
	BO1	5	95.75	7.30	10.12	-13.17
	BP1	4	73.61	18.14	10.28	-2.03
	T_Q_HR	3	16.90	19.37	8.48	55.25
6	ho_incm	4	16.13	59.91	1.59	22.37
	live_t	6	11.50	79.95	1.60	6.96
	edu	4	37.11	62.53	5.22	-4.86
	tins	5	39.65	52.03	15.76	-7.44
	D_1_1	5	55.23	53.38	2.35	-10.96
	BO1	5	77.04	25.24	7.56	-9.84
	BP1	4	93.42	35.74	19.02	-48.18
	T_Q_HR	3	34.69	-2.90	23.66	44.55

식 (2.6)의 세 번째 항은 식 (2.5)의 마지막 항에서 분산의 편의요인을 제거한 후 추정량 편의에 의해 발생하는 값으로 Table 3.5의 6번째 열은 이 값에 대한 χ^2_P의 백분율을 나타낸다. 이 백분율은 범위는 0.49%∼23.66%으로 5번째 열의 -2.90%∼80.48%보다 크게 작다. 따라서 추정량의 편의가 피어슨검정통계량에 기여하는 부분은 분산의 편의가 기여하는 부분보다 상대적으로 작은 것으로 나타난다. Table 3.5의 6번째 열의 값은 χ^2_W/χ^2_P과 유사한 경향을 보인다. 가구조사 변수인 live_t, ho_incm에서 매우 작은 비율을, 그 밖의 변수들에서는 상대적으로 큰 비율을 나타낸다. 따라서 추정량의 편의가 피어슨검정통계량에 미치는 영향은 후자의 변수들에서 더 크게 나타나고 있다. 모집단의 수에 따른 이 비율들의 변화는 변수마다 달라서 일반화하는 데 한계가 있다. Table 3.5의 마지막 열은 앞의 요인들에 의해서 설명되지 않은 분산과 추정량의 편의가 교락되어 피어슨 검정통계량에 영향을 미치는 정도를 나타낸다. 분산의 편의에 의해 영향을 받는 정도가 가장 작았던 BP1에서 가장 작고, 분산의 편의에 의한 영향이 비교적 크고 추정량의 편의에 의해 영향을 받는 정도가 아주 작지 않았던 tins와 ho_incm에서 가장 큰 비율을 보인다.

4. 결론

최근의 많은 조사연구에서는 층화, 집락, 계통, 다단계 추출, 불균등 확률추출 등을 종합적으로 사용하는 복합표본설계에 의한 표본조사를 실시하는 것이 일반적 추세이다. 복합표본조사의 범주형자료는 전통적인 피어슨 카이제곱검정에 필요한 서로 독립이고 동일한 다항분포를 따른다는 조건을 충족시키지 못한다. 본 연구에서는 복합표본설계에 의한 범주형조사자료의 k-모집단 동질성검정에서 설계기반 일치통계량인 Wald 검정통계량을 유도하고, 전통적인 피어슨 카이제곱검정통계량을 사용하는 경우 발생할 수 있는 오차성분들을 분해하는 식을 도출한 뒤, 분산의 편의가 영향을 미치는 크기, 추정량의 편의가 영향을 미치는 크기, 기타 두 영향의 교락에 의한 크기로 각각 구분하였다.

또한, 각 항목이 피어슨검정통계량에 미치는 상대적 크기를 경험적으로 확인하기 위해 국민건강영양조사 제4기 2차년도 (2008) 자료 (KNHANES IV-2)를 사용하여 분석하였다. 분석결과를 종합하면, 분석에 사용된 변수들 중, 가구조사 변수들에서 대체로 피어슨검정통계량은 분산의 편의 의해 크게 과대추정되며 추정량의 편의가 기여하는 부분은 매우 작은 것으로 나타나며, 건강설문조사 변수 BP1, BO1과 검진조사 변수인 T_Q_HR은 분산의 편의가 피어슨검정통계량에 미치는 효과는 상대적으로 작으나 추정량의 편의가 미치는 효과는 가구조사 변수들보다 큰 것을 확인할 수 있었다. 피어슨검정통계량이 분산이나 추정량의 편의에 의해 영향을 받는 정도와 모집단수와의 관련성은 변수들마다 차이가 있어 일반화하는 데는 한계가 있다.

본 연구결과 중 실증분석결과는 KNHANES IV-2 자료 중 지극히 소수인 몇 개의 변수에 대해 실시된 것으로 KNHANES IV-2 자료 전체로 일반화하는 데는 어려움이 있지만 이러한 경험분석은 본 연구에서 도출해낸 식의 확장가능성이 있음을 확인하였다.

References

Casella G. and Berger R. L. (1990). *Statistical inference*, Brooks/Cole Publishing Company, California.

Heo. S. and Chang. D. (2010). A sample survey design for service satisfaction evaluation of regional education offices. *Journal of the Korean Data & Information Science Society*, **21**, 671-678.

Heo, S. and Chung, Y. (2012). Effect of complex sample design on Pearson test statistic for homogeneity. *Journal of the Korean Data & Information Science Society*, **23**, 757-764.

Holt, D., Scott, A. J. and Ewings, P. D. (1980). Chi-squared tests with survey data. *Journal of the Royal Statistical Society A*, **143**, 302-320.

Kim, D. H., Cho, K. H., Hwang, J. S., and Jung, K. H. (2009). A sample design for life and attitude survey of Gyeongbuk people. *Journal of the Korean Data & Information Science Society*, **20**, 1165-1167.

Korea Centers for Disease Control and Prevention (2009). *The fourth Korea national health and nutrition examination survey (KNHANES IV-2)*, Ministry of Health & Welfare/Korea Center Disease Control and Prevention, Seoul.

Lee, K. and Park, J. (2007). *Final report of the sample design for the fourth Korea national health and nutrition examination survey*, The Korean Association for Survey Research, Seoul.

Rao, J. N. K. and Scott, A. J. (1981). The analysis of categorical data from complex sample surveys: Chi-squared tests for goodness of fit the independence in two-way tables. *Journal of the American Statistical Association*, **76**, 221-230.

Rao, J. N. K. and Scott, A. J. (1984). On chi-squared test for multiway contingency tables with cell proportions estimated from survey data. *The Annals of Statistics*, **12**, 46-60.

Rao, J. N. K. and Scott, A. J. (1987). On simple adjustments to chi-square tests with sample survey data. *The Annals of Statistics*, **15**, 385-397.

Shao, J. (1996). Resampling methods in sample surveys (with discussion). *Statistics*, **27**, 203-254.

Thomas, D. R. and Rao, J. N. K. (1987). Small-sample comparisons of level and power for simple goodness-of-fit statistics under cluster sampling. *Journal of the American Statistical Association*, **82**, 630-636.

Journal of the Korean Data & Information Science Society
2013, **24**(4), 815–824

http://dx.doi.org/10.7465/jkdi.2013.24.4.815
한국데이터정보과학회지

Error cause analysis of Pearson test statistics for k-population homogeneity test[†]

Sunyeong Heo[1]

[1]Department of Statistics, Changwon National University

Received 8 June 2013, revised 3 July 2013, accepted 8 July 2013

Abstract

Traditional Pearson chi-squared test is not appropriate for the data collected by the complex sample design. When one uses the traditional Pearson chi-squared test to the complex sample categorical data, it may give wrong test results, and the error may occur not only due to the biased variance estimators but also due to the biased point estimators of cell proportions. In this study, the design based consistent Wald test statistics was derived for k-population homogeneity test, and the traditional Pearson chi-squared test statistics was partitioned into three parts according to the causes of error; the error due to the bias of variance estimator, the error due to the bias of cell proportion estimator, and the unseparated error due to the both bias of variance estimator and bias of cell proportion estimator. An analysis was conducted for empirical results of the relative size of each error component to the Pearson chi-squared test statistics. The second year data from the fourth Korean national health and nutrition examination survey (KNHANES, IV-2) was used for the analysis. The empirical results show that the relative size of error from the bias of variance estimator was relatively larger than the size of error from the bias of cell proportion estimator, but its degrees were different variable by variable.

Keywords: Categorical data, complex sample design, design effect, homogeneity test, Pearson test, Wald test.

[†] This research is financially supported by Changwon National University in 2011∼2012.
[1] Professor, Department of Statistics, Changwon National Uviversity, Changwon 641-773, Korea. E-mail: syheo@changwon.ac.kr